Statistical Methods for Data Analysis

Luca Lista

INFN Napoli

Definition of probability

- Two main different definitions of the concept of probability are possible
- Frequentist
  - **Probability** is the ratio of the number of occurrences of an event to the total number of experiments, in the *limit* of very large number of **repeatable** experiments.
  - Can only be applied to a specific classes of events (repeatable experiments)
  - Meaningless to state: "*probability that the lightest SuSy particle's mass is less than 1 TeV*"
- Bayesian
  - **Probability** measures someone's the degree of belief that a statement is (or will be…) true
  - Can be applied to most of unknown events (past, present, future):
    - *"Probability that Velociraptors hunted in groups"*
    - *"Probability that R.S.C. Anderlecht will win next championship"*

Problems with probability definitions

- Frequentist probability is, to some extent, circularly defined
  - A phenomenon can be proven to be random (i.e.: obeying laws of statistics) only if we observe infinite cases ("converge in probability")
  - F.James et al.: *"this definition is not very appealing to a mathematician, since it is based on experimentation, and, in fact, implies unrealizable experiments ($N\rightarrow\infty$)"*. But a physicist can take this with some pragmatism
  - Frequentist models can be justified from details of poorly predictable underlying physical phenomena
    - Deterministic dynamic but poorly predictable (chaos theory, …)
    - Quantum Mechanics: intrinsically probabilistic…!
  - A school of statisticians state that Bayesian statistics is a more natural and fundamental concept, and frequentist statistic is just a special sub-case

- On the other hand, Bayesian statistics is subjectivity by definition, which is unpleasant for scientific applications.
  - Bayesian reply that it is actually inter-subjective, i.e.: the real essence of learning and knowing physical laws…

- Frequentist approach is preferred by the large fraction of physicists (probably the majority), but Bayesian statistics is getting more and more popular in many application, also thanks to its simpler application in many cases

Axiomatic definition (A. Kolmogorov)

- Axiomatic probability definition applies to both frequentist and Bayesian probability
  - Let (Ω, $F \subseteq 2^{\Omega}$, P) be a measure space that satisfy:
  - 1 $P(E) \geq 0 \quad \forall E \in F$
  - 2 $P(\Omega) = 1$
  - 3 $\forall (E_1, \cdots, E_n) \in F^n : E_i \cap E_j = 0$

$$P(\bigcup_{i=1,\cdots,n} E_i) = \sum_{i=1,\cdots,n} P(E_i)$$

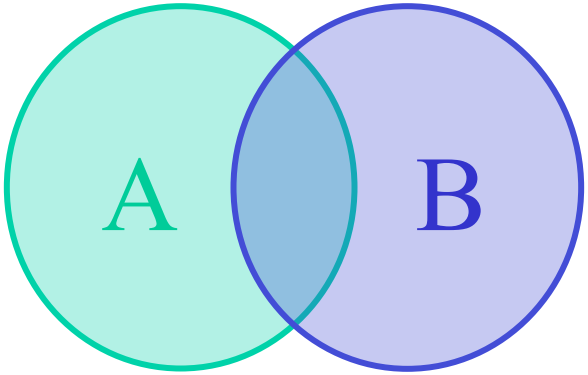
Andrej Nikolaevič Kolmogorov
(1903-1987)

  - Terminology: Ω = sample space, F = event space, P = probability measure
- So we have a formalism to deal with different types of probability

Conditional probability

- Probability of A, given B: $P(A \mid B)$
- i.e.: probability that an elementary experiment, known to belong to set B, is also a member of set A:

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



- Event A is said to be independent on B if the conditional probability of A given B is equal to the probability of A:
  - $P(A \mid B) = P(A)$
- Hence, if A is independent on B:
  - $P(A \cap B) = P(A)\, P(B)$
- → If A is independent on B, B is independent on A

Bayes theorem

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

$$\Downarrow$$

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

$$\Downarrow$$

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

Thomas Bayes (1702-1761)

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

Bayes and likelihood function

- Likelihood function definition: a PDF of the variables $x_1, \ldots, x_n$:

$$L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m) = \left[\frac{\mathrm{d}P(x_1, \cdots, x_n)}{\mathrm{d}x_1 \cdots \mathrm{d}x_n}\right]_{\theta_1, \cdots, \theta_m}$$

- Bayesian posterior probability for $\theta_1, \ldots, \theta_m$:

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

- Where:
  - $P(\theta_1, \ldots, \theta_m)$ is the prior probability.
    - Often assumed to be uniform in HEP papers, but there is no motivation for this choice (and a uniform distribution depends on the parameterization choice!)
  - $\int L(\ldots)\mathrm{P}(\ldots)\, \mathrm{d}^m\theta$ is a normalization factor
- Interpretation:
  - The observation modifies the prior knowledge of the unknown parameters as if L is a probability distribution function for $\theta_1, \ldots, \theta_n$
  - F.James et al.: *"The difference between $P(\theta)$ and $P(\theta \mid x)$ shows how one's knowledge (degree of belief) about θ has been modified by the observation x. The distribution $P(\theta \mid x)$ summarizes all one's knowledge of θ and can be used accordingly."*

Bayesian inference

- Just use the (normalized) product of likelihood function times the prior probability as the posterior PDF for the unknown parameter(s) θ:

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

In literature often the notation π(θ) is used to define prior

- You can evaluate then:
  - The average value of θ
  - The variance of θ
  - The mode (most likely value) and a (68%) coverage interval
  - In many cases, the most likely value and average don't coincide!
- Notice that the Maximum Likelihood estimate is the mode of Bayesian inference with a uniform prior
- Upper limits are easily computed using the Bayesian approach

Counting experiments

- The only information from our measurements is the number of observed events of the kind of interest
- Expected distribution is Poissonian:

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

- Hypotheses test terminology:
  - Null hypothesis (H_0): $s = 0$
  - Alternative hypothesis (H_1): test against a specific value of $s > 0$
- An experiment outcome is a specific value of n: $n = n_{\mathrm{obs}}$
- If we observe zero events we can state that:
  - No background events have been observed ($n_b = 0$)
  - No signal events have been observed ($n_s = 0$)
- Further simplification: let’s assume that the expected background b is negligible: $b \cong 0$

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

Bayesian inference of a Poissonian

- Posterior probability, assuming the prior to be $\pi(s)$, setting $b = 0$ for simplicity:

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

- If is $\pi(s)$ is uniform, the denom. is: $\displaystyle\int_0^\infty \frac{s'^n e^{-s'}}{n!}\mathrm{d}s' = 1$

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

$$\langle s\rangle = n+1 \quad \mathrm{Var}(s) = n+1$$

- We have:
- Most *probable* value: $s^{max} = n$

… but this result depends on the choice of the prior!

Bayesian upper limit

- The posterior PDF for s, assuming a uniform prior, is:

$$f(s|n) = \frac{s^n e^{-s}}{n!} \quad \Rightarrow \quad f(s|0) = e^{-s}$$

For zero observed events

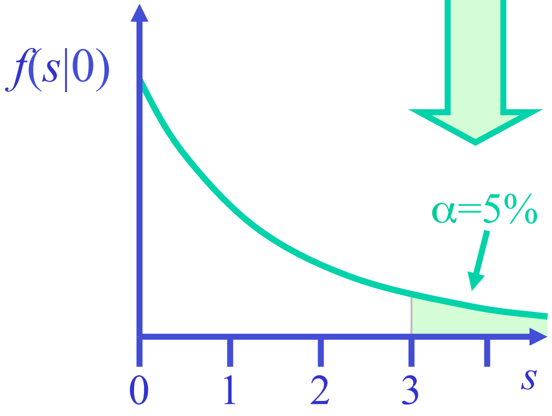
- The cumulative distribution is:

$$F(s|n) = \int_0^s \frac{t^n e^{-t}}{n!} \mathrm{d}t = 1 - e^{-s} \left(\sum_{m=0}^{n} \frac{s^m}{m!} \right)$$

- In particular for $n=0$: $F(s|0) = 1 - e^{-s}$

$$P(s \leq 2.996|n = 0) = 0.95$$
$$P(s \leq 2.303|n = 0) = 0.90$$

- We will see that *by chance* this result is identical also for a frequentist limit:
- But the interpretation is very different!



Frequentist *vs* Bayesian

- Interpretation of parameter errors:
  - $\theta = \theta^{\mathrm{est}} \pm \delta \quad \Rightarrow \quad \theta \in [\, \theta^{\mathrm{est}} - \delta,\ \theta^{\mathrm{est}} + \delta]$
  - $\theta = \theta^{\mathrm{est}}\,{}^{+\delta 2}_{-\delta 1} \quad \Rightarrow \quad \theta \in [\, \theta^{\mathrm{est}} - \delta_1,\ \theta^{\mathrm{est}} + \delta_2]$

- Frequentist approach:
  - Knowing a parameter within some error means that a large fraction (68% or 95%, usually) of the experiments contain the (fixed but unknown) true value within the quoted confidence interval: $[\theta^{\mathrm{est}} - \delta_1,\ \theta^{\mathrm{est}} + \delta_2]$
- Bayesian approach:
  - The posterior PDF for θ is maximum at θ^{est} and its integral is 68% within the range $[\theta^{\mathrm{est}} - \delta_1,\ \theta^{\mathrm{est}} + \delta_2]$

- The choice of the interval, i.e.. δ_1 and δ_2 can be done in different ways, e.g: same area in the two tails, shortest interval, symmetric error, …
- Note that both approaches provide the same results for a Gaussian model using a uniform prior, leading to possible confusions in the interpretation

Problems with Bayesian inference/limits

- Bayesian inference, as well as Bayesian limits, require the choice of a prior distribution
  - This makes estimates somewhat subjective
- Choices frequently adopted in physics are not unique:
  - Uniform PDF as a function of the signal strength?
  - Uniform PDF as a function of the Higgs boson mass?
- In some cases results do not depend strongly on the assumed prior
  - But this usually happens when the statistical sample is sufficiently large, which is not often the case for upper limits

Choosing the prior PDF

- If the prior PDF is uniform in a choice of variable
- Uniform PDF not preserved when applying coordinate transformation
- Given a prior PDF in a random variable, there is always a transformation that makes the PDF uniform
- Harold Jeffreys' prior: chose the prior form that is invariant under parameter transformation
- metric related to the Fisher information (metrics invariant!)

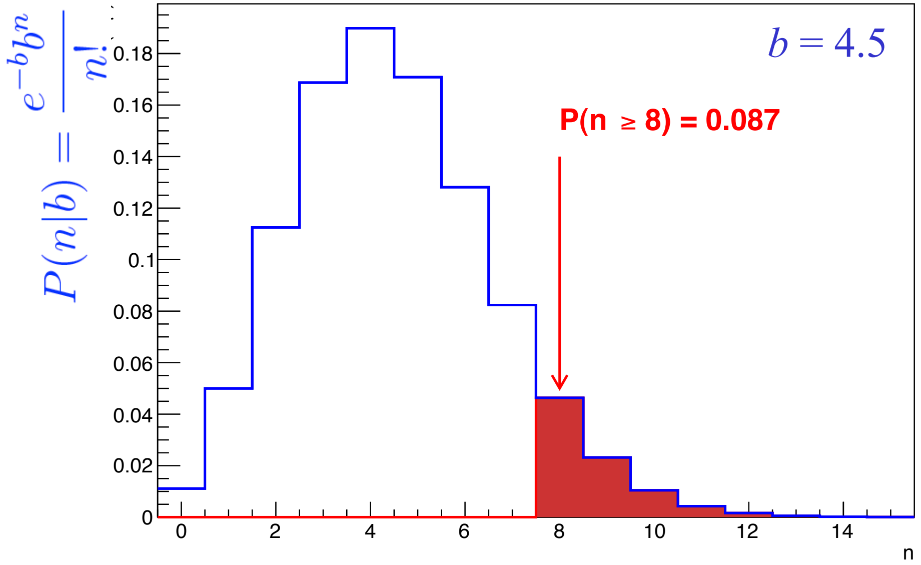
$$p(\vec{\theta}) \propto \sqrt{I(\vec{\theta})} \qquad I(\vec{\theta}) = \det\left[\left\langle \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_i} \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_j} \right\rangle\right]$$

- Some common cases:
  - Poissonian mean: $p(\mu) \propto 1/\sqrt{\mu}$
  - Poissonian mean with background b: $p(\mu) \propto 1/\sqrt{\mu + b}$
  - Gaussian mean: $p(\mu) \propto 1$
  - Gaussian r.m.s: $p(\sigma) \propto 1/\sigma$
  - Binomial parameter: $p(\varepsilon) \propto 1/\sqrt{\varepsilon(1-\varepsilon)}$

- Problematic with more than one dimension!
  - Reference priors….

Frequentist approach: *p*-value

- *p*-value: probability to observe at least $n_{\rm obs}$ events if the null hypothesis H_0 ($s = 0$) is true
- Probability that a background (over)fluctuation gives at least the observed number of events



- If H_0 is true ($s = 0$) the distribution of the *p*-value is uniform if the distribution is continuous. It is approximately uniform in case of discrete distributions
- Remark: the *p*-value is not the probability that H_0 is true: this is probability has meaning only under the Bayesian approach!

Significance and discovery

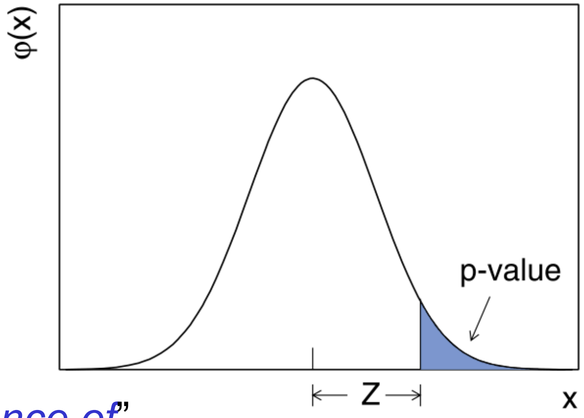
- The *p*-value, which measures the observed "incompatibility" with the background-only hypothesis is often converted into a number of standard deviations ("$n\sigma$") corresponding to a Gaussian distribution

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

or:

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

Φ = cumulative of a normal distribution



- Usually, in literature:
  - If the significance is > 3 ("3σ") one claims "*evidence of*"
  - If the significance is > 5 ("5σ") one claims "*observation*" (discovery!)
    - probability of background fluctuation $p < 2.87\times10^{-7}$
- For a counting (=Poissonian) experiment with a large expected background b, a Gaussian approximation may be accurate enough:

$$Z = \frac{s}{\sqrt{b}} = \frac{n-b}{\sqrt{b}}$$

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 ☺

Excluding a signal hypothesis

- Assuming a given value of $s > 0$ (H$_1$), a corresponding *p*-value can be computed
- in this case the *p*-value measures the probability of a signal underfluctuation ($n \leq n_{\mathrm{obs}}$)
  - Null hypothesis is inverted w.r.t. the discovery case
- The exclusion of a signal hypothesis usually has milder requirements:
  - $p < 0.05$ (i.e.: 95% confidence level): $Z = 1.64$
  - $p < 0.10$ (i.e.: 90% confidence level): $Z = 1.28$
- Discovering a new signal usually requires more stringent evidence than excluding it!

Zero events observed

- The probability to observe $n = n_s$ events expecting s events (H_1), is:

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

  - The probability to observe $n \leq n_{\mathrm{obs}} = 0$ (*p*-value) is:

$$P(0|s) = e^{-s}$$

- We can set an upper limit on the expected signal yield s excluding values of s for which the *p*-value $p = e^{-s}$ is less than 5% or 10%
  - $p = e^{-s} \geq \alpha = 1 - \mathrm{CL}$

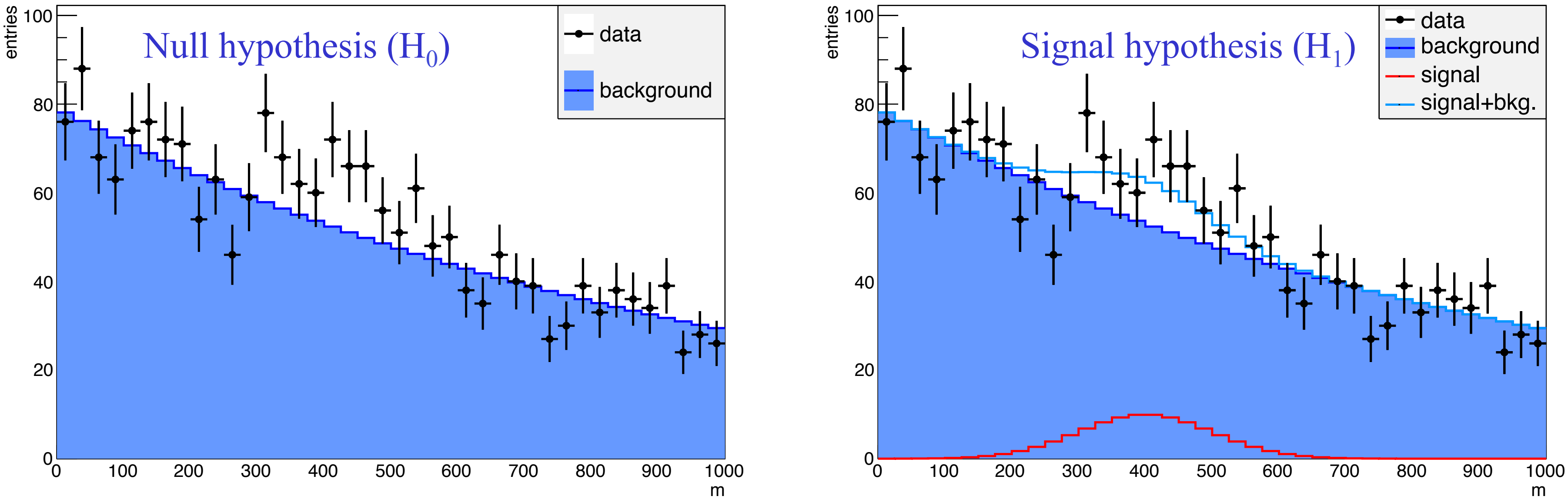
- So: $s \leq -\ln(\alpha) = s^{\mathrm{up}}$. For $\alpha = 5\%$ or 10%:
  - $s \leq \mathbf{2.996}$ @ 95% C.L.
  - $s \leq \mathbf{2.303}$ @ 90% C.L.

Same results as for the Bayesian case, but the interpretation is different!

In general, Bayesian and frequentist limits do not coincide

A more realistic case

- A variable (e.g.: reconstructed invariant mass) is samples for a number of events
- It’s distribution is used to look for a signal



Likelihood model

- The probability to observe the present (unbinned) sample is given by the extended likelihood function:

$$\mathcal{L}(\vec{m}|s, b, \mu, \theta) = \frac{e^{-(b+\mu s)}(b+\mu s)^n}{n!}\prod_{i=1}^{n}(f_b P_b(m_i|\theta) + f_s P_s(m_i|\theta))$$

$$\left.\begin{aligned} f_s &= \frac{s}{s+b} \\ f_b &= \frac{b}{s+b} \end{aligned}\right\} \rightarrow \quad = \frac{e^{-(b+\mu s)}}{n!}\prod_{i=1}^{n}(bP_b(m_i|\theta) + \mu s P_s(m_i|\theta))$$

- μ is usually the “signal strength” (i.e.: σ/σ_{SM}) in case of Higgs search, instead of number of signal events s
- Or, considering just the binned information:

$$\mathcal{L}(\vec{n}|\vec{s}, \vec{b}, \mu) = \prod_{i=1}^{n_{\mathrm{bins}}} \mathrm{Pois}(n_i|b_i + \mu s_i)$$

s_i, b_i = signal/background templates,
μ = signal strength
θ = nuisance parameters

Nuisance parameters

- So called “nuisance parameters” (θ) are unknown parameters that are not interesting for the measurement (μ)
  - E.g.: background rate, detector resolution, background shape modeling, other sources of systematic uncertainties, etc.
- Two main possible approaches:
- Add the nuisance parameters together with the interesting unknown to your likelihood model
  - But the model becomes more complex!
  - Easier to incorporate in a fit than in upper limits
- “Integrate them away” (→ Bayesian)

Nuisance pars. in Bayesian approach

- Notation below:
  - μ = parameter of interest, θ = nuisance parameters
- No particular treatment:

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

- $P(\mu|x)$ obtained as marginal PDF, "integrating out" θ:

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

Frequentist: the test statistics

- A test statistics has to be chosen to discriminate the two hypotheses
- Neyman-Pearson lemma suggests to adopt a likelihood ratio between two hypotheses to achieve the “best” discrimination:

$$t = -2\ln \frac{\max_{\mu,\theta,\in H_1} \mathcal{L}(x|\mu,\theta)}{\max_{\mu,\theta\in H_0} \mathcal{L}(x|\mu,\theta)}$$

- Variations are available on the market
  - Mainly: how to deal with nuisance parameters

Profile Likelihood

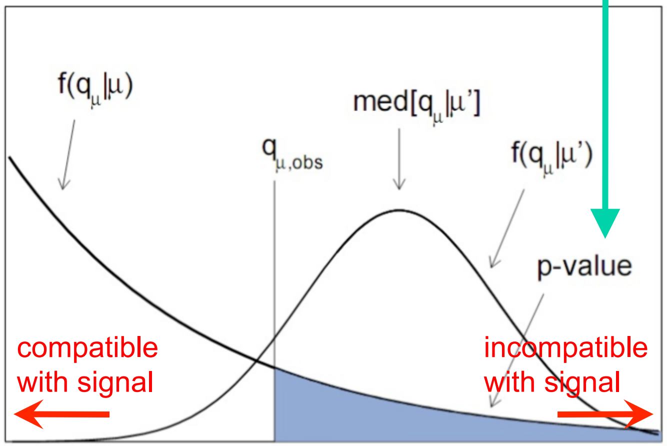
- Adopted test statistics:

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

← Fix μ, fit θ

← Fit both μ and θ

$$p_\mu = \int_{q_\mu^{\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\mathrm{d}q_\mu$$



$$q_\mu = -2\ln\lambda(\mu)$$

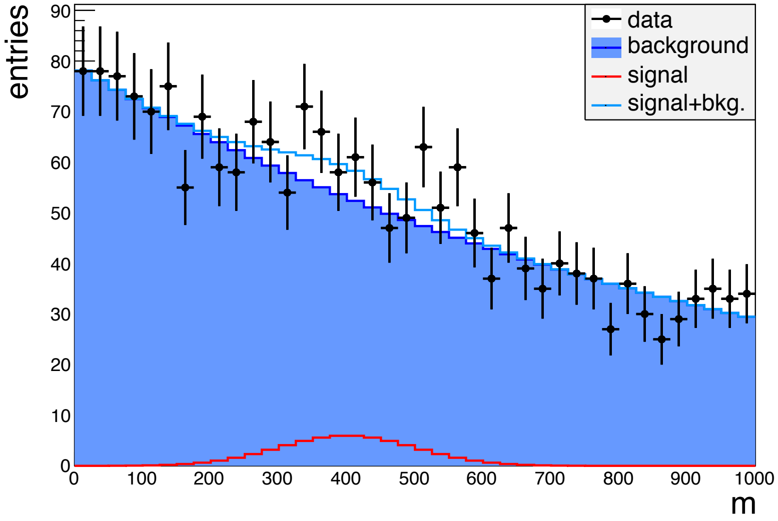
- Profile likelihood shape is broadened by nuisance parameters θ (loss of information)
- Nice asymptotic property: distribution of $q_\mu = -2\ln\lambda(\mu)$ tends to a χ^2 distribution with one degree of freedom due to Wiks' theorem (one parameter of interest = μ)

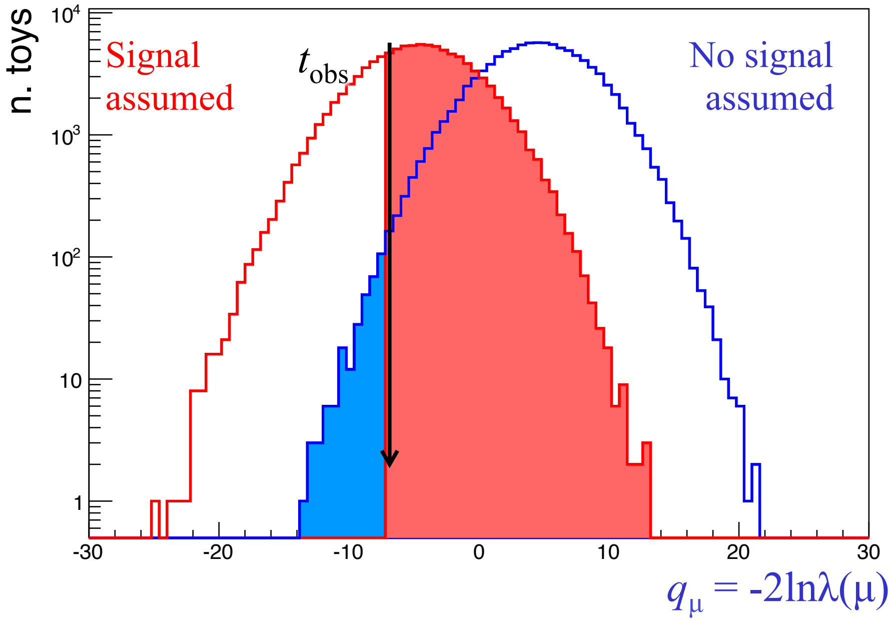
$$\boxed{Z_\mu \simeq \sqrt{q_\mu}}$$

- Different 'flavors' of test statistics, e.g.: deal with unphysical *μ* < 0, …

Bump hunting

- Significance evaluation can be performed by generating random pseudo-experiments ("toy MC")
- *p*-value estimated as the fraction of toys with value of $t = q_\mu$ less than t_{obs}.

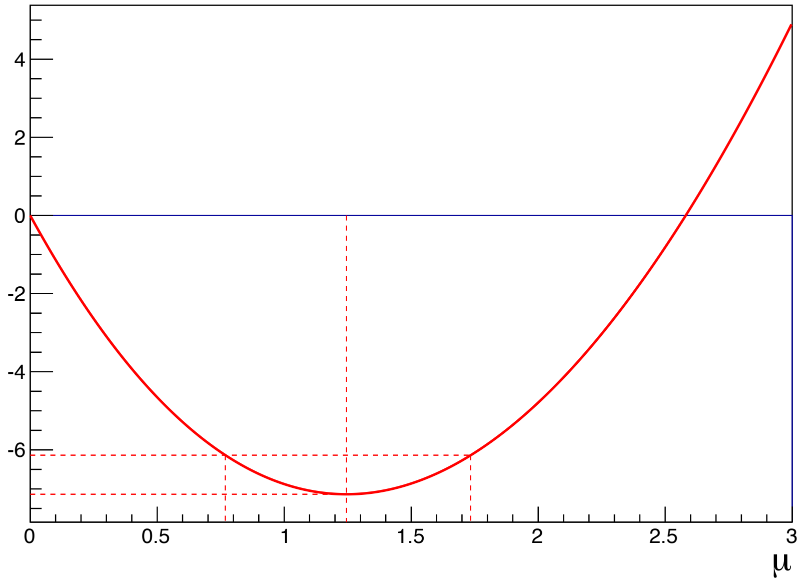




- In the presented case, assuming no nuisance parameter (μ is the only parameter of the model):
- $p = 374/10^5 = 3.7 \pm 0.2\%$
- $Z = 2.7$

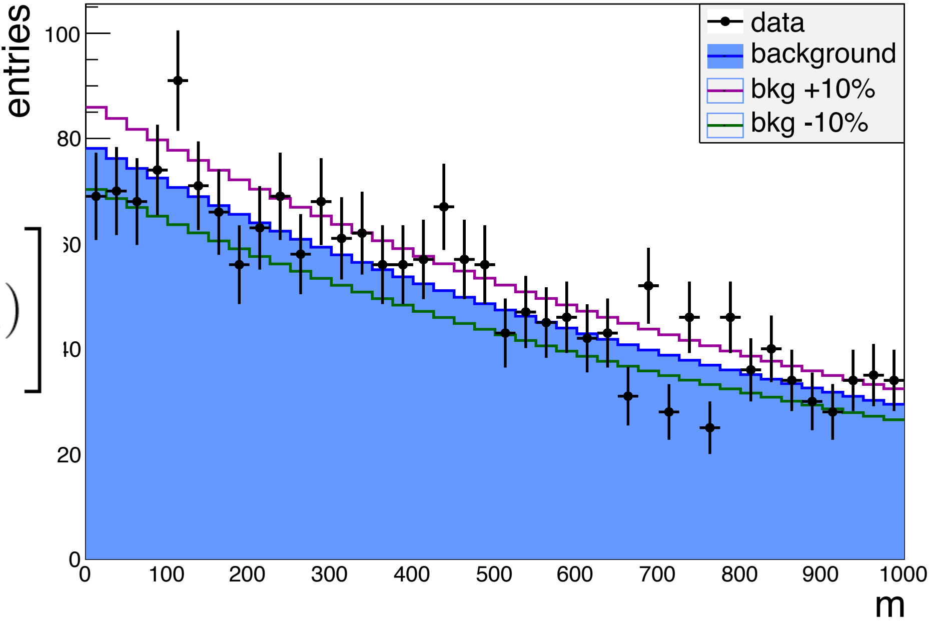
Determining the signal strength

- A scan of the test statistics reveals its minimum at the best parameter value (μ=signal strength here)
- The fit value of μ is the maximum-likelihood estimate
  - $\mu = 1.24^{+0.49}_{-0.48}$
- Using the likelihood ratio instead of just the likelihood: from Wilks' theorem, the likelihood ratio approximately follows a χ^2 distribution in the null hypothesis
  - In presence of signal:
    - Significance $Z \sim \sqrt{-t_{min}}$
    - In this case $Z \sim 2.7$
  - Goodness-of-fit, in case of no presence of signal



Dealing with systematics

- Several assumptions done in the model are affected by systematic uncertainties
- In the present case, the uncertainty on the background rate/shape has large effect on the signal estimate
- The amount of background can be scaled by a factor β, which becomes one of the nuisance parameters (θ) of the model
- More in general, for binned cases, the effect may be modeled by shifting "up" or "down" the signal and background templates corresponding to the uncertainty amount

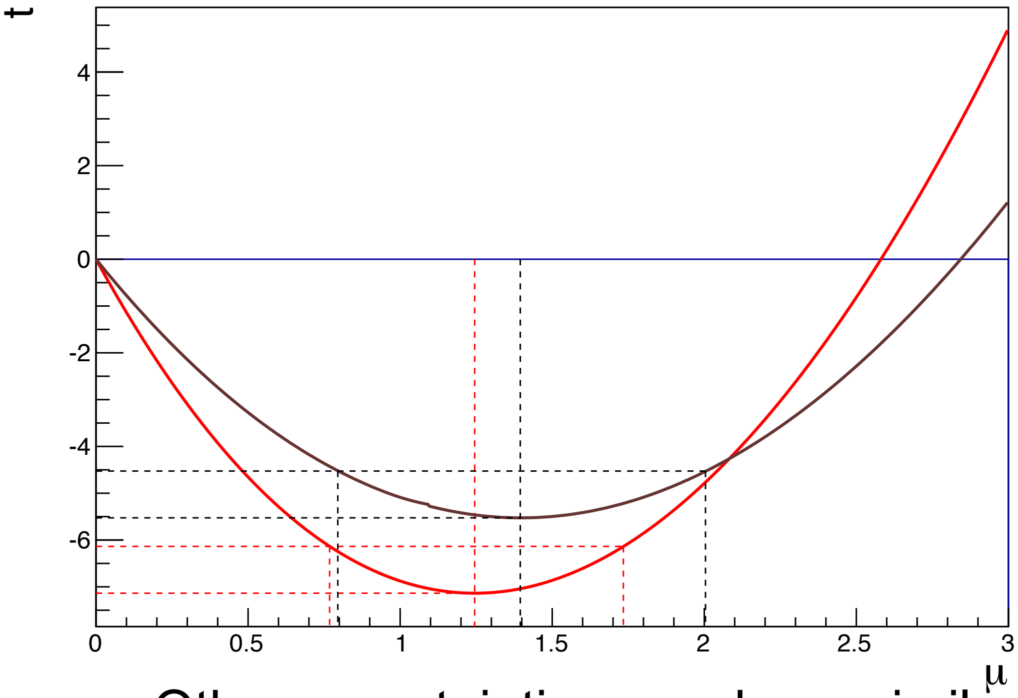


$$\mathcal{L}(\vec{n}|\vec{s},\vec{b},\mu,\beta) = \left[\prod_{i=1}^{n_{\mathrm{bins}}} \mathrm{Pois}(n_i|\beta b_i + \mu s_i)\right]$$

$$t = -2\ln\frac{\max_\beta \mathcal{L}(\vec{n}|\vec{s},\vec{b},\mu,\beta)}{\max_\beta \mathcal{L}(\vec{n}|\vec{s},\vec{b},\mu=0,\beta)}$$

Effect of systematic uncertainty

- Larger uncertainty: $\mu = 1.40^{+0.61}_{-0.60}$
- Smaller significance: $Z \sim 2.3$



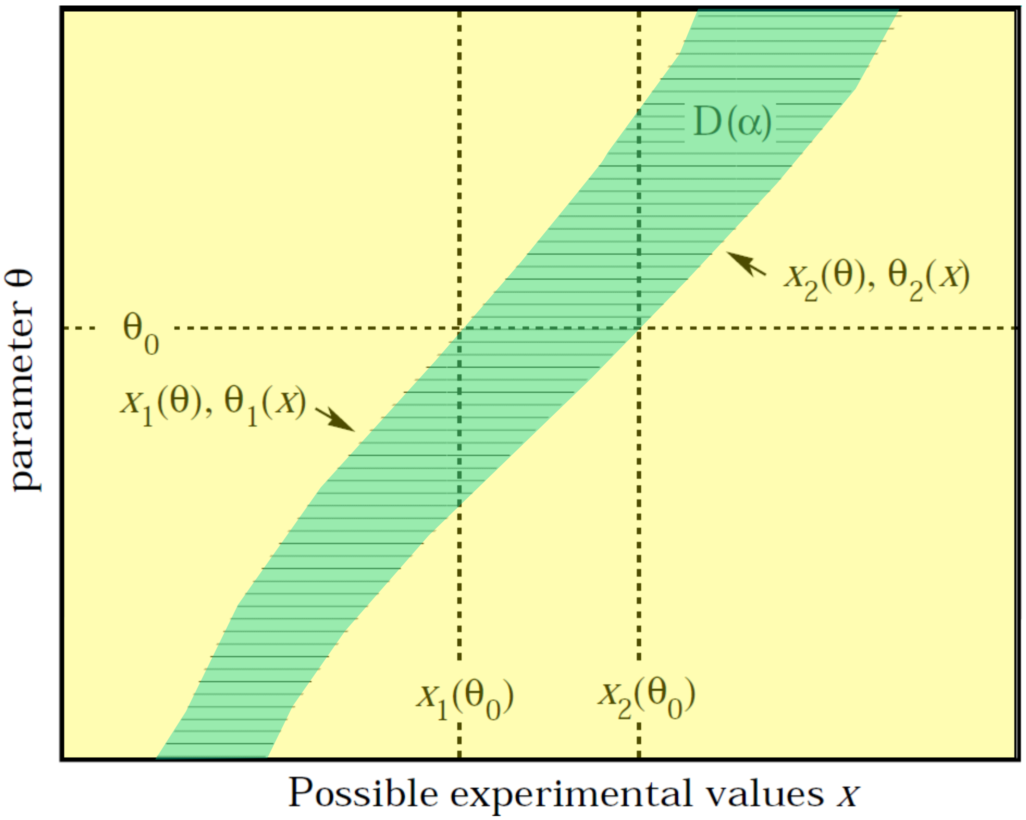
- Other uncertainties may have similar treatment: background shape (affects b_i), signal resolution (affects s_i), etc.

- Practical side effects:
  - Slower generation of toy MC (require a minimization for each extraction)
  - Asymptotic (=Wilks' approximation) evaluations are more convenient in those cases
  - For a counting experiment the following approximation is a valid for large n:

$$Z = \frac{n-b}{\sqrt{b + (\Delta b)^2}}$$

Jerzy Neyman's confidence intervals

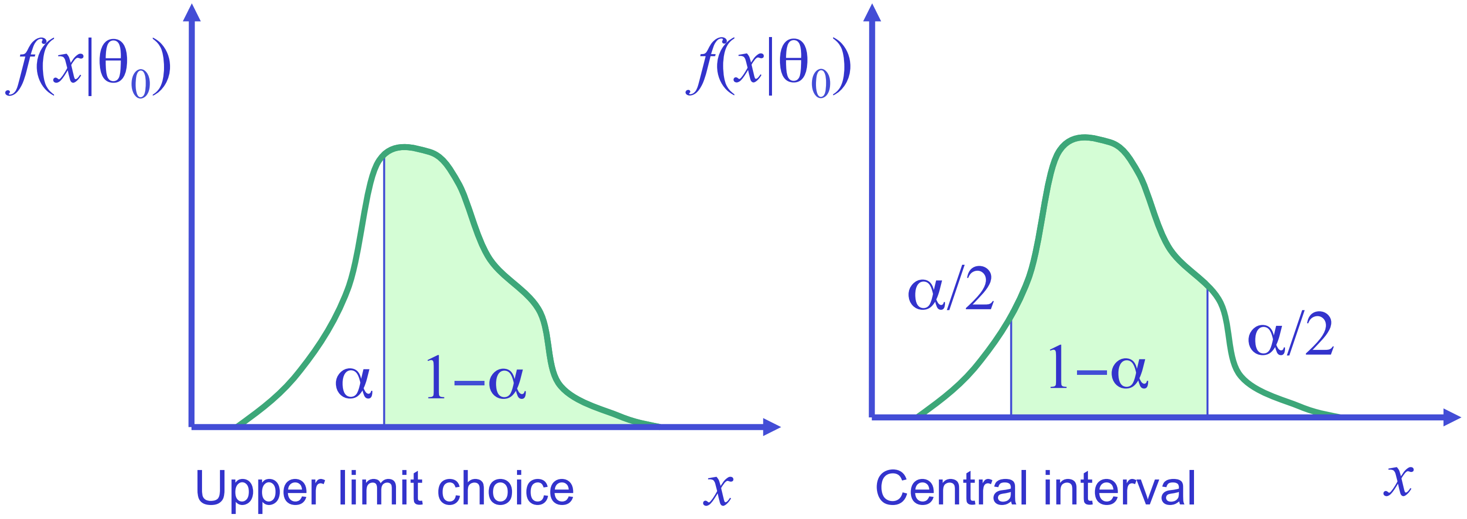
- Scan an unknown parameter θ (or μ) over its range
- Given θ, compute the interval $[x_1, x_2]$ that contain x with a probability $\mathrm{CL} = 1-\alpha$
- Ordering rule is needed!
  - Central interval? Asymmetric? Other?
- Invert the confidence belt, and find the interval $[\theta_1, \theta_2]$ for a given experimental outcome of x
- A fraction $1-\alpha$ of the experiments will produce x such that the corresponding interval $[\theta_1, \theta_2]$ contains the true value of μ (coverage probability)
- Note that the random variables are $[\theta_1, \theta_2]$, not θ



Plot from PDG statistics review

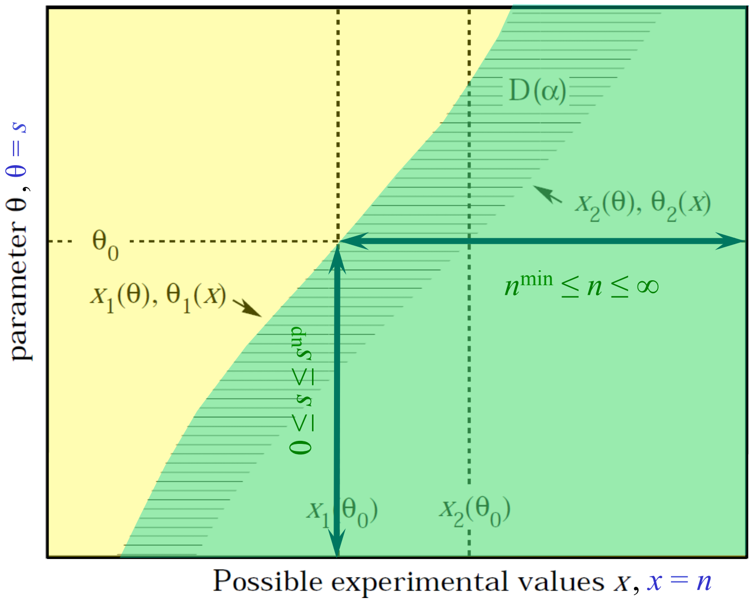
Ordering rule

- For a fixed $\theta = \theta_0$ we different possible intervals choices give the same probability $1-\alpha$



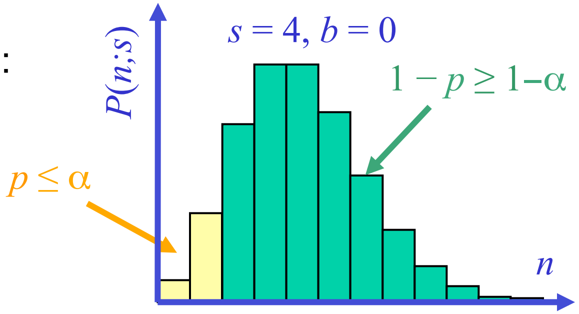
Frequentist upper limits

- Upper limit construction from inversion of Neyman belt with asymmetric intervals
- Building a confidence interval on the observable
  - The observable x is the number of events n for counting experiments
  - Final confidence interval must be asymmetric if we want to compute upper limits:
    - $s \in [s_1, s_2] \Rightarrow s \in [0, s^{\mathrm{up}}]$
  - Upper limit = right-most edge of asymmetric interval
  - Hence, we should have an asymmetric interval on n:
    - $n \in [n_1, n_2] \Rightarrow n \in [n^{\mathrm{min}}, \infty]$


parameter θ, θ = s
D(α)
$x_2(\theta)$, $\theta_2(x)$
θ_0
$x_1(\theta)$, $\theta_1(x)$
$n^{\mathrm{min}} \leq n \leq \infty$
$0 \leq s \leq s^{\mathrm{up}}$
$x_1(\theta_0)$
$x_2(\theta_0)$
Possible experimental values x, $x = n$

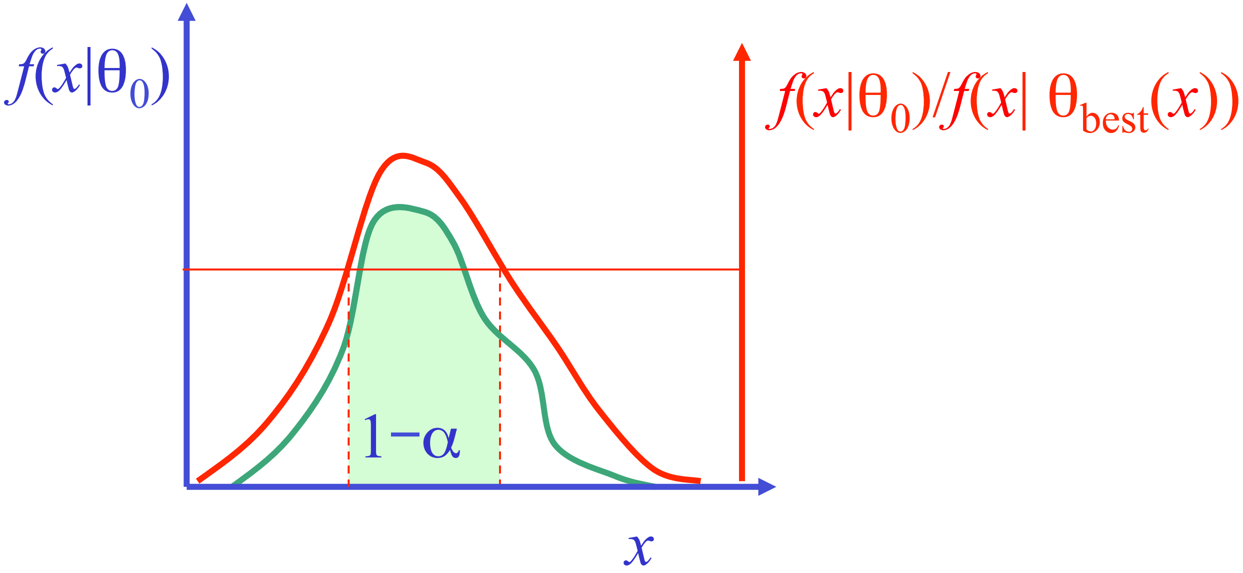
- Poissonian distributions involve discrete values, can't exactly satisfy coverage: produce the smallest overcoverage
  - Use the smallest interval that has at least the desired C.L.:

$$P(s \in [0, s^{\mathrm{up}}]) \geq \mathrm{CL} = 1 - \alpha$$
$$\Leftrightarrow$$
$$P(n \in [n^{\mathrm{min}}, \infty]) = 1 - p \geq \mathrm{CL} = 1 - \alpha$$



Feldman-Cousins ordering

- Find the contour of the likelihood ratio that gives an area α
- $R_\mu = \{x : L(x|\theta) \,/\, L(x|\theta_{\mathrm{best}}) > k_\alpha\}$
- Motivation discussed in next slides



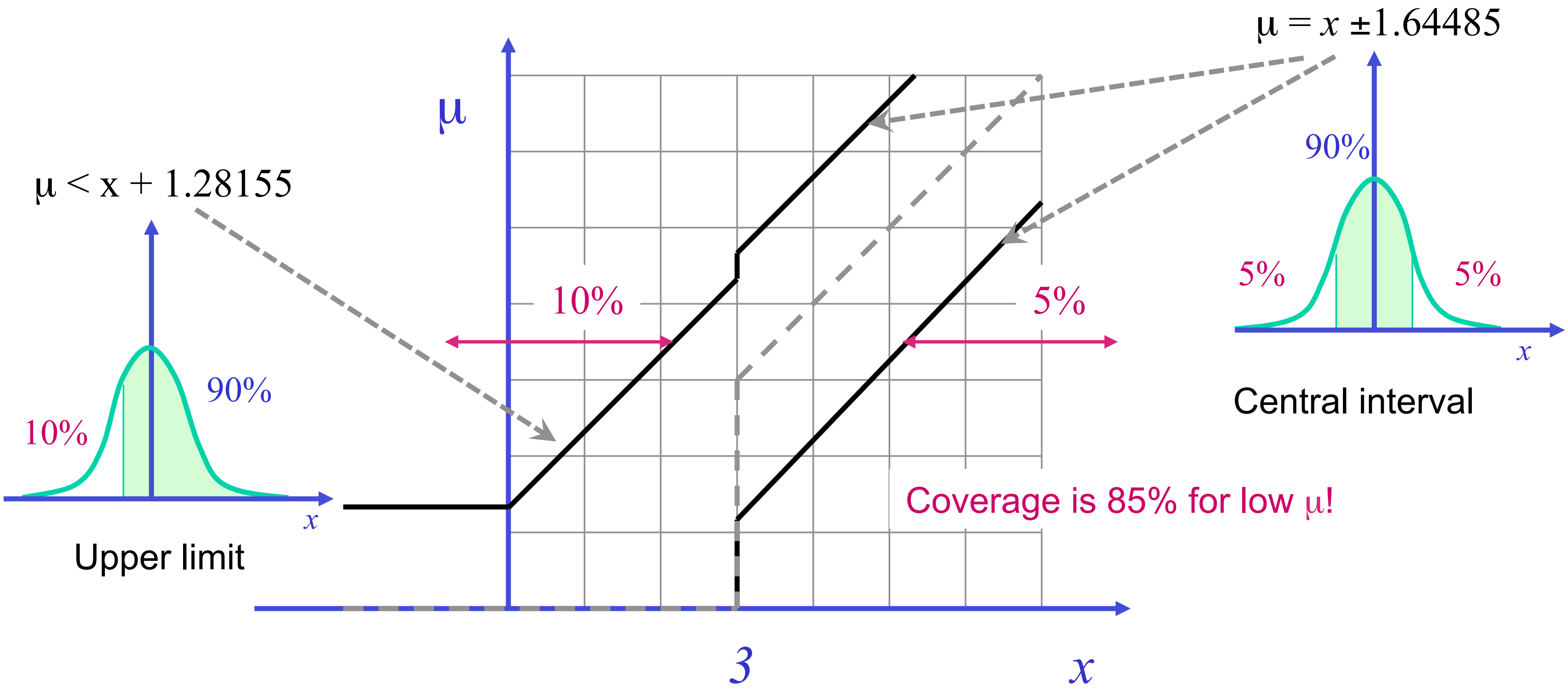
Gary J. Feldman, Robert D. Cousins, Phys.Rev.D57:3873-3889,1998

“Flip-flopping”

- When to quote a central value or upper limit?
- E.g.:
  - *“Quote a 90%* C.L. *upper limit of the measurement is below 3σ; quote a central value otherwise”*
- Upper limit ↔ central interval decided according to observed data
- This produces incorrect coverage!
- Feldman-Cousins interval ordering guarantees the correct coverage

“Flip-flopping” with Gaussian PDF

- Assume Gaussian with a fixed width: σ=1

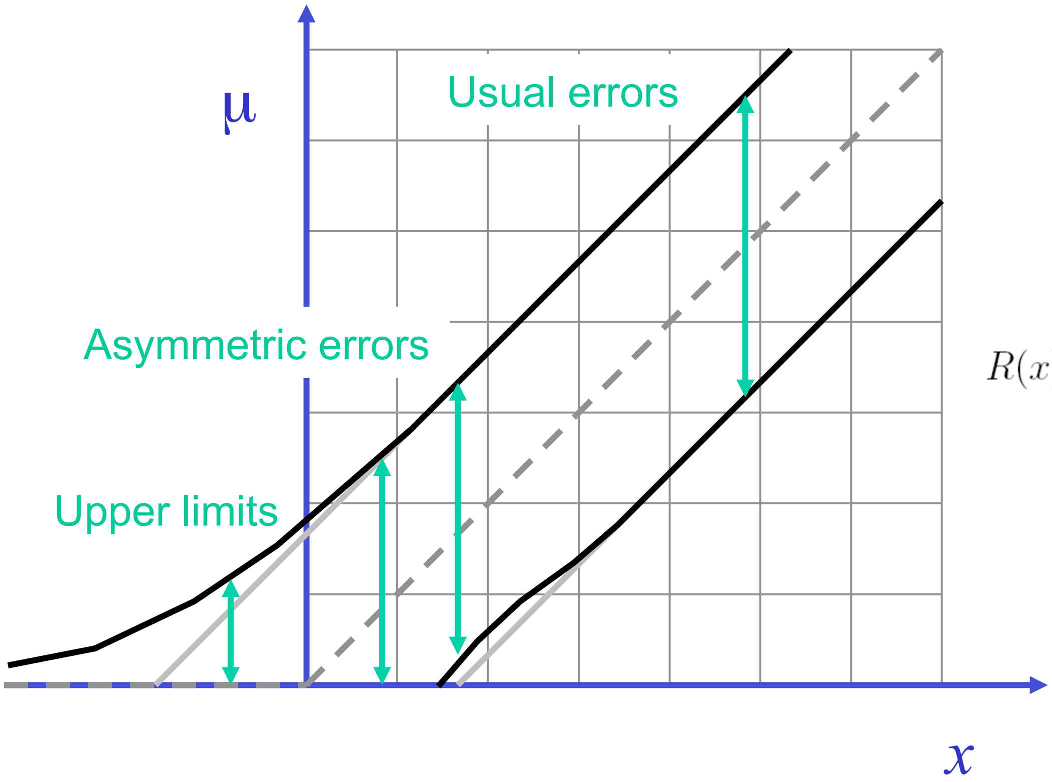


Gary J. Feldman, Robert D. Cousins, Phys.Rev.D57:3873-3889,1998

Feldman-Cousins approach

- Define range such that:
  - $P(x|\mu) \,/\, P(x|\mu_{\text{best}}(x)) > k_\alpha$

$$\mu_{\text{best}} = \max(x, 0)$$



$\mu_{\text{best}} = x$ for $x \geq 0$

$$P(x|\mu_{\text{best}}) = \begin{cases} 1/\sqrt{2\pi}, & x \geq 0 \\ \exp(-x^2/2)/\sqrt{2\pi}, & x < 0. \end{cases}$$

$$R(x) = \frac{P(x|\mu)}{P(x|\mu_{\text{best}})} = \begin{cases} \exp(-(x-\mu)^2/2), & x \geq 0 \\ \exp(x\mu - \mu^2/2), & x < 0. \end{cases}$$

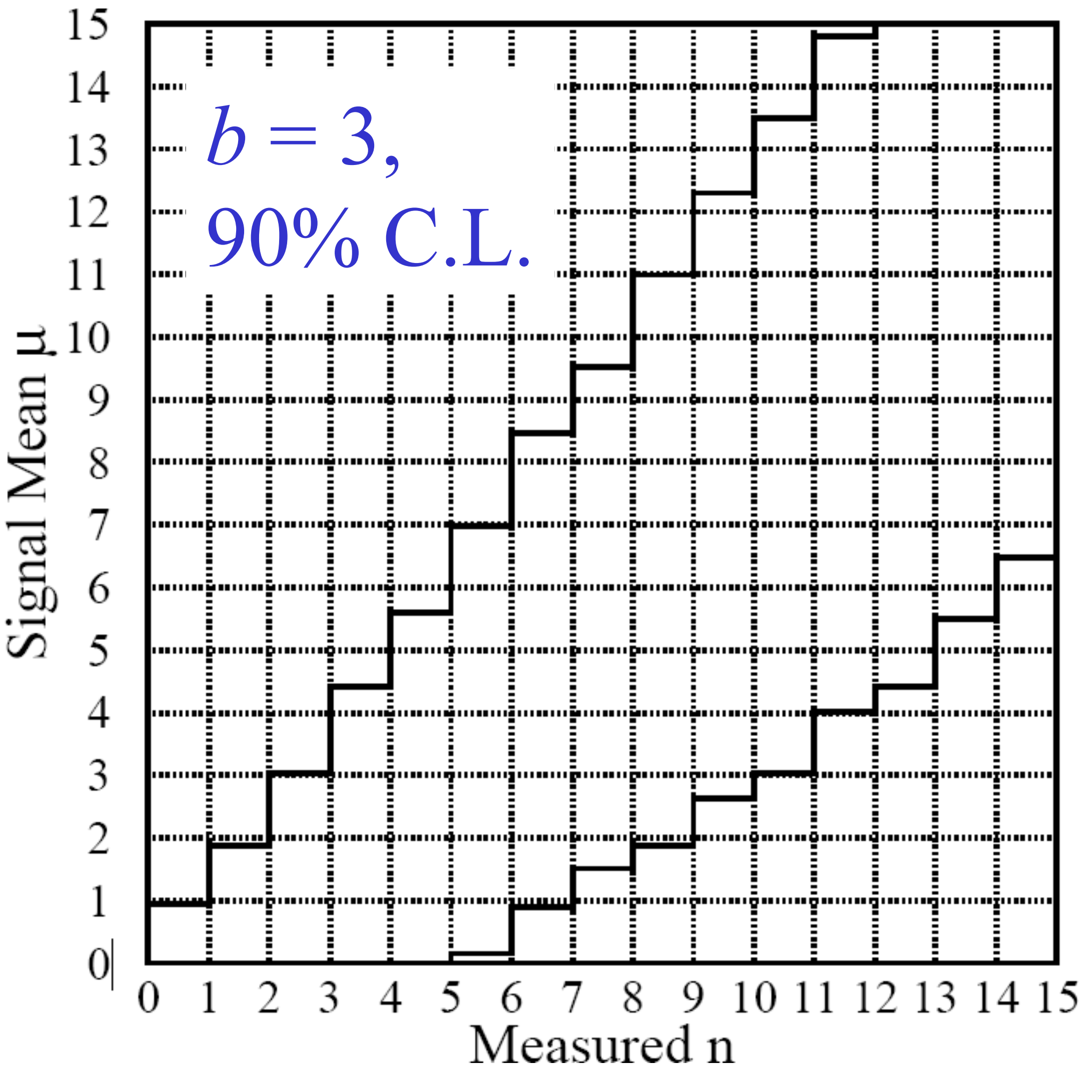
Solution can be found numerically

Feldman-Cousins: Poissonian case

Purely frequentist

ordering based on likelihood ratio

Belt depends on b, of course

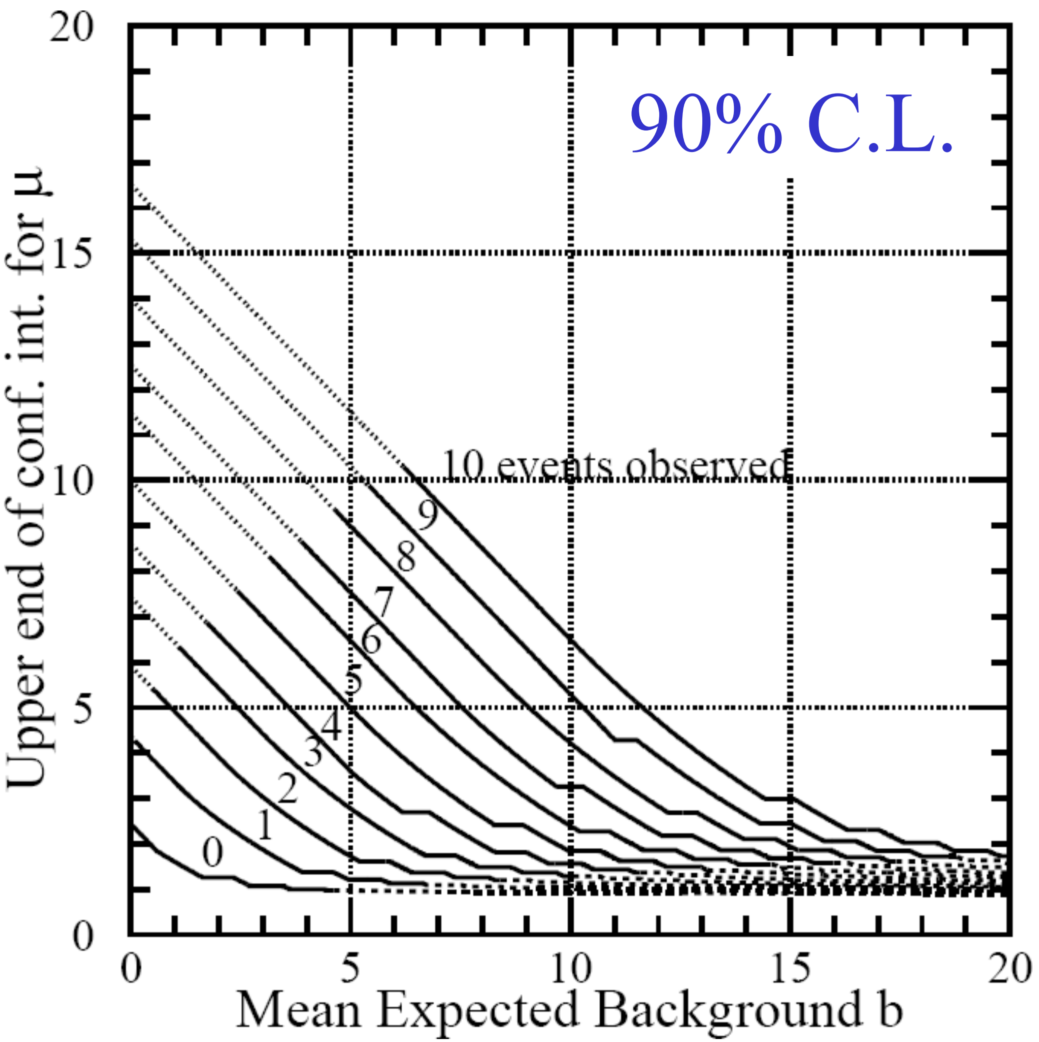


G.Feldman, R.Cousins, Phys.Rev.D,57(1998), 3873

Upper limits with Feldman-Cousins

Note that the curve for *n* = 0 decreases with *b*, while the result of the Bayesian calculation is independent on *b*, at 2.3

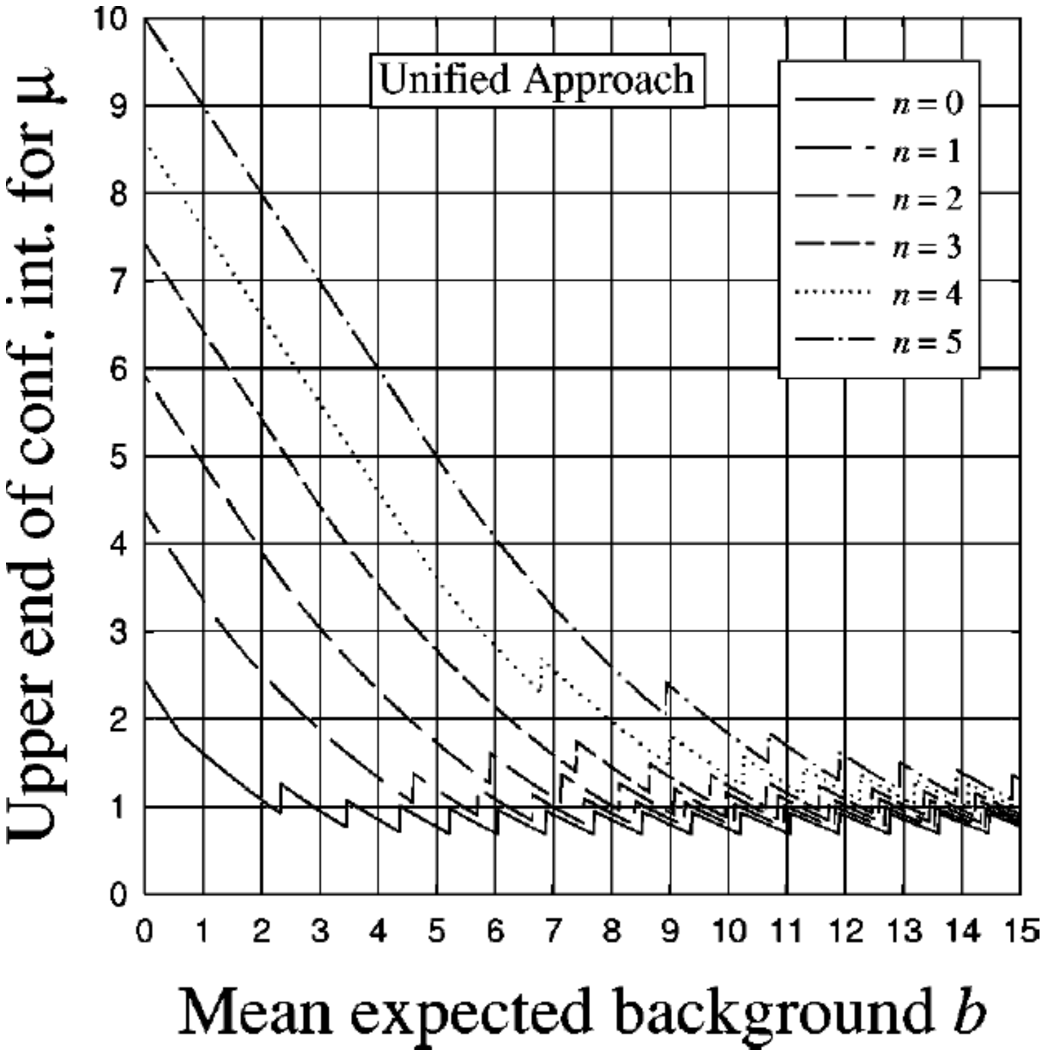
F&C reply: frequentist interval do not express *P*(μ|*x*) !



G.Feldman, R.Cousins, Phys.Rev.D,57(1998), 3873

A Close-up

Note the ‘ripple’ Structure due to the discrete nature of Poissonian statistics



C. Giunti, Phys.Rev.D,59(1999), 053001

Pros and cons of F&C approach

- Pros:
  - Avoids problems with physical boundaries on parameters
    - Important in many cases, e.g.: cross-section measurement, neutrino bounds, …
  - Never returns an empty confidence interval
  - Does not incur flip-flop problems
  - Ensure proper statistical coverage
- Cons:
  - Constructing the confidence intervals is complicated, requires CPU-intensive numerical algorithms, and often large toy Monte Carlo generations
  - Systematic uncertainties are not easily to incorporate
  - Peculiar features with small number of events
  - In case of zero observed events, gives better limits for experiments that expect higher background

From PDG Review…

“The intervals constructed according to the unified procedure 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”

Modified frequentist method: $\mathrm{CL_s}$

- Method developed for Higgs limit at LEP-II
- Using the likelihood ratio as test statistics:

$$Q(m_H) = \frac{L(s+b)}{L(b)}$$

- Confidence levels estimator (→different from Feldman-Cousins):

$$\boxed{\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} = \frac{P(Q_{s+b} \le Q_{\mathrm{obs}})}{P(Q_b \le Q_{\mathrm{obs}})} = \frac{N(Q_{s+b} \le Q_{\mathrm{obs}})}{N(Q_b \le Q_{\mathrm{obs}})}}$$

  - Gives over-coverage w.r.t. classical limit ($\mathrm{CL_s} > \mathrm{CL_{s+b}}$: conservative)
  - Similarities with Bayesian C.L.

- Identical to Bayesian limit for Poissonian counting!

$$1 - \mathrm{CL} = e^{-s^{\mathrm{up}}} \frac{\sum_{m=0}^{n} \frac{(s^{\mathrm{up}}+b)^m}{m!}}{\sum_{m=0}^{n} \frac{b^m}{m!}}$$

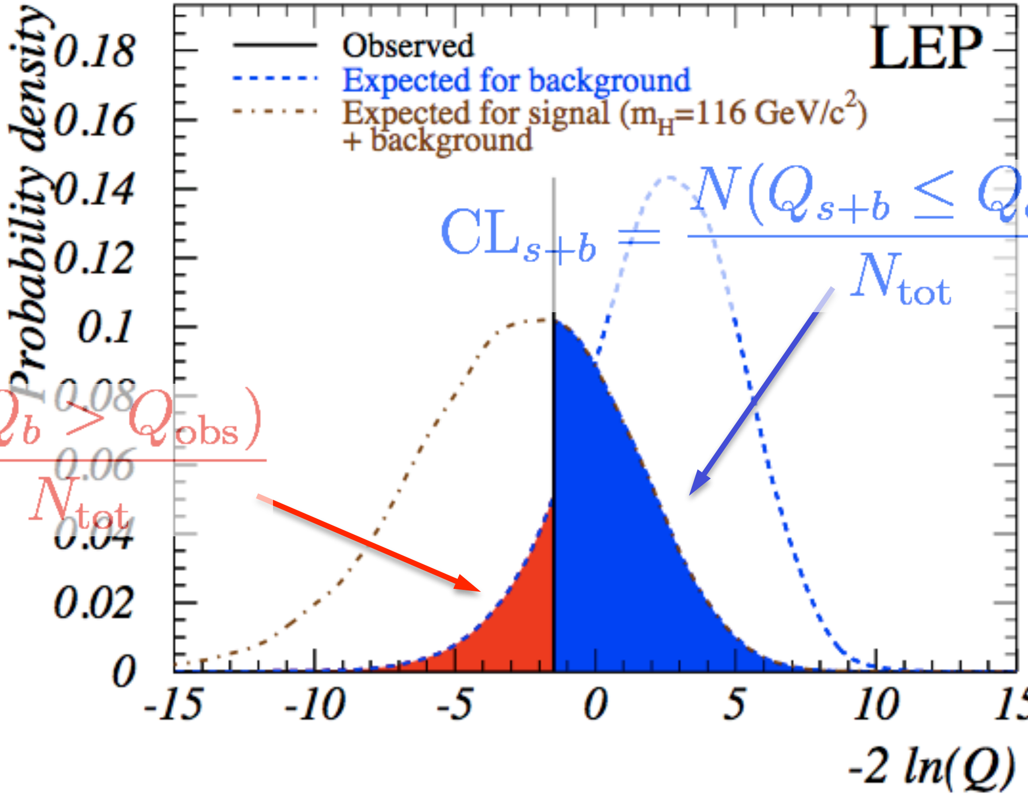
- *"approximation to the confidence in the signal hypothesis, one might have obtained if the experiment had been performed in the complete absence of background"*

- No problem when adding channels with low discrimination

$\mathrm{CL_s}$ with toy experiments

- The actual $\mathrm{CL_b}$ and $\mathrm{CL_{s+b}}$ are computed in practice from toy Monte Carlo experiments

$$\mathrm{CL}_s = \frac{N(Q_{s+b} \le Q_{\mathrm{obs}})}{N(Q_b \le Q_{\mathrm{obs}})}$$

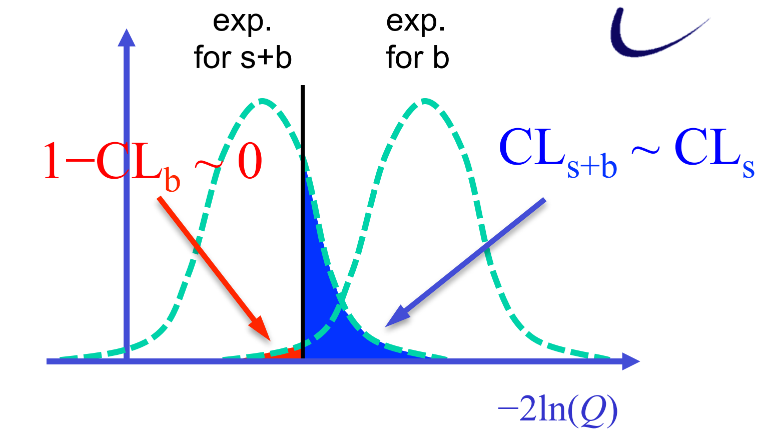


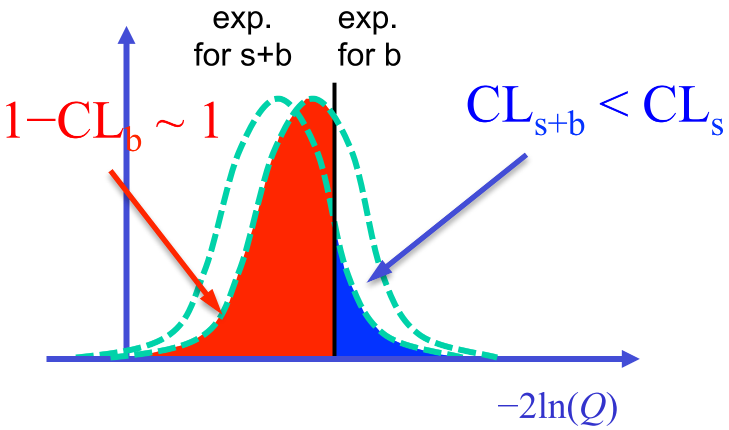
$$\mathrm{CL}_{s+b} = \frac{N(Q_{s+b} \le Q_{\mathrm{obs}})}{N_{\mathrm{tot}}}$$

$$1 - \mathrm{CL}_b = \frac{N(Q_b > Q_{\mathrm{obs}})}{N_{\mathrm{tot}}}$$

Main CL$_s$ features

- CL$_{s+b}$: probability to obtain a result which is less compatible with the signal than the observed result, assuming the signal hypothesis
- CL$_b$: probability to obtain a result less compatible with the signal than the observed one in the background-only hypothesis
- If the two distributions are very well separated than 1−CL$_b$ will be very small ⇒ CL$_b$ ~1 and CL$_s$ ~ CL$_{s+b}$, i.e: the ordinary *p*-value of the s+b hypothesis
- If the two distributions are very close than 1−CL$_b$ will be large ⇒ CL$_b$ small, preventing CL$_s$ to become very small
- CL$_s$ < 1−α prevents to reject where there is little sensitivity





$$\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} = \frac{P(Q_{s+b} \le Q_{\mathrm{obs}})}{P(Q_b \le Q_{\mathrm{obs}})}$$

Observations on 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, 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 control sample (y)

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

- Use likelihood method in more than one dimension
- May be CPU intensive
- Usually leads to results that are very similar to the hybrid Cousins-Highland hybrid method (→ next slide)

Cousins-Highland hybrid approach

- No fully solid background exists on a genera approach to incorporate nuisance parameters within a frequentist approach
- Hybrid approach proposed by Cousins and Highland
  - Integrate(“marginalize”) the likelihood function over the nuisance parameters (Nucl.Instr.Meth.A320 331-335, 1992)

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

- Also called “hybrid” approach, because some Bayesian approach implicit in the integration: “*seems to be acceptable to many pragmatic frequentists*” (G. Zech, Eur. Phys. J. C 4 (2002) 12)
  - Bayesian integration of PDF, but likelihood used in a frequentist way
- Some numerical studies with Toy Monte Carlo showed that the frequentist calculation gives very similar results in many cases
- May undercover in case of high significance

$\mathrm{CL_s}$ with profile likelihood at LHC

- Use profile likelihood as test statistics, then evaluate CLs

$$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(\mathrm{data}|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\mathrm{data}|\hat{\mu},\hat{\theta})} \qquad 0\le\hat{\mu}\le\mu$$

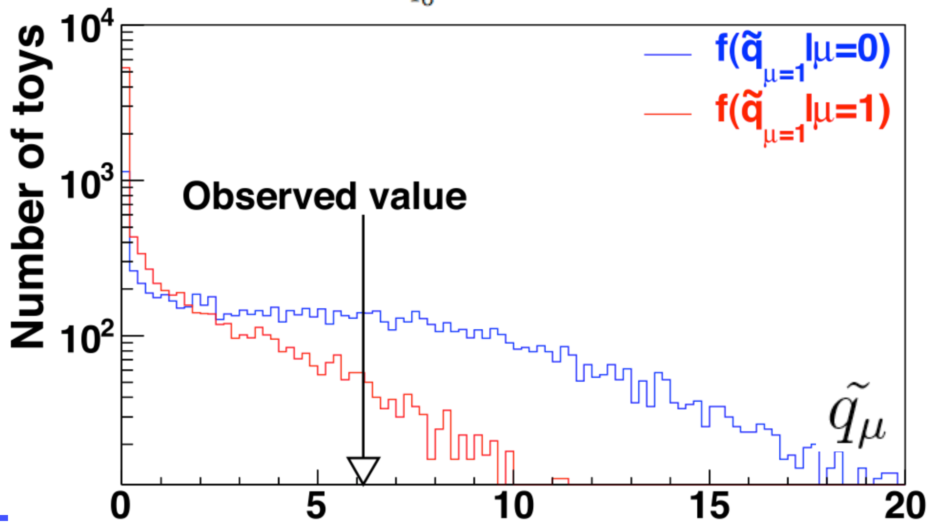
- The constraint $\hat{\mu}\le\mu$ ensures that upward fluctuations of the data such that $\hat{\mu}>\mu$ are not considered as evidence against the signal hypothesis, namely a signal with strength μ

$$p_\mu = P(\tilde{q}_\mu \ge \tilde{q}_\mu^{obs}\,|\,\mathrm{signal{+}background}) = \int_{\tilde{q}_\mu^{\mathrm{obs}}}^{\infty} f(\tilde{q}_\mu|\mu,\hat{\theta}_\mu^{\mathrm{obs}})\,d\tilde{q}_\mu$$

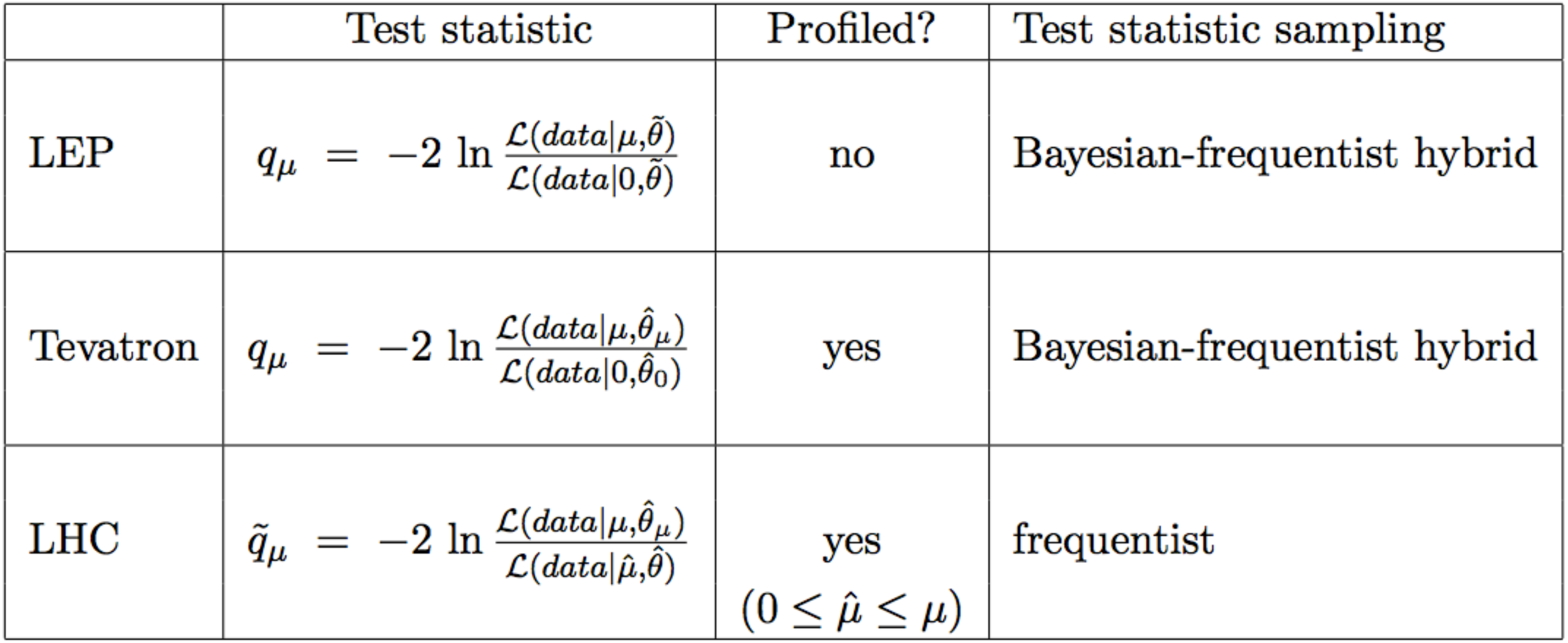
$$1-p_b = P(\tilde{q}_\mu \ge \tilde{q}_\mu^{obs}\,|\,\mathrm{background\text{-}only}) = \int_{q_0^{\mathrm{obs}}}^{\infty} f(\tilde{q}_\mu|0,\hat{\theta}_0^{\mathrm{obs}})\,d\tilde{q}_\mu$$

$$CL_s(\mu) = \frac{p_\mu}{1-p_b}$$

- Agreed estimator between ATLAS and CMS for Higgs search:
  - ATL-PHYS-PUB-2011-11



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

Asymptotic approximations

- The constrain $\hat{\mu} \leq \mu$ imposed on the profile likelihood distorts its from Wilks'[*] asymptotic approximation, so the distribution tends no longer to a χ^2, but:

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2\ , \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu+\mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2\ . \end{cases}$$

- Where: $\sigma^2 = \frac{\mu^2}{q_{\mu,A}}$ and $q_{\mu,A}$ is the test statistics evaluated on the Asimov set (→ next slide)
- Approximations are a valuable way to perform computation quickly
- More details on asymptotic approximations:
  - Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554
  - [*] S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. 9, 60–62 (1938)
  - A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. 54(3), 426–482 (1943)

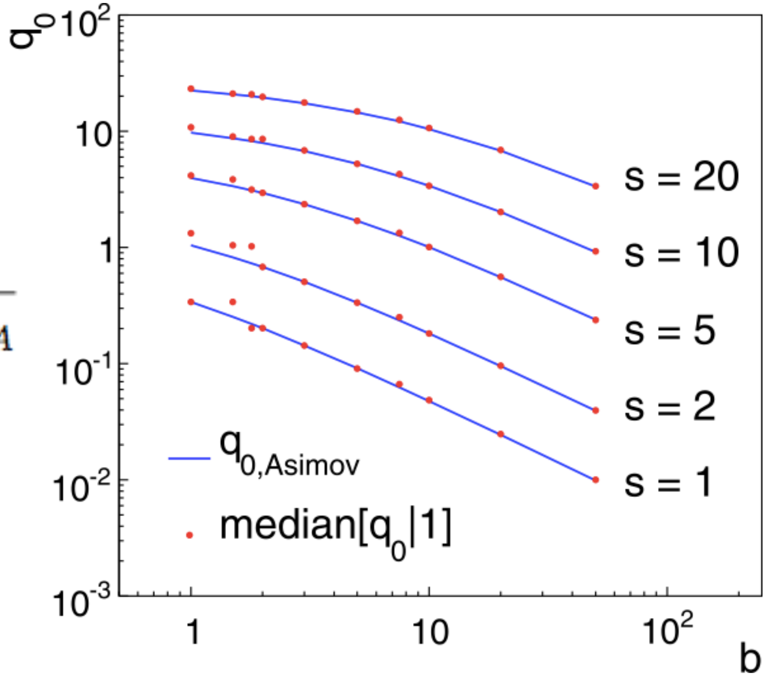
Asimov[*] sets

- Approximate evaluation of expected (median) limits avoiding CPU-intensive generation of toy Monte Carlo samples by using a single "representative set"
- Replace each bin of the observable distribution (e.g.: reconstructed Higgs mass spectrum) by its expectation value
- Set nuisance parameters to their nominal value
- Approximation valid in the asymptotic limit
- Median significance can be approximated with the sqrt of the test statistic, evaluated at the Asimov set:

$$\mathrm{med}[Z_\mu|0] = \sqrt{\tilde{q}_{\mu,\mathrm{A}}}$$

- Uncertainty bands on expected upper limits can also be evaluated using Asimov sets, avoiding large toy MC extractions: $\sigma^2 = \dfrac{\mu^2}{q_{\mu,A}}$
- Mathematical validity and approximations of this approach are discussed by Cowan *et al.* [**]

[*] Asimov, Franchise, in Isaac Asimov: The Complete Stories, vol. 1 (Broadway Books, New York, 1990)
[**] Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554



Look-Elsewhere Effect (LEE)

- Imagine you search for a signal peak over a background distribution that is spread 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 $B_s \rightarrow \mu\mu$
  - 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 *t* under the background-only hyp., possibly convert it to the area of a Gaussian tail:

$$p = \int_{t_{\mathrm{obs}}}^{\infty} f(t|0)\mathrm{d}t \qquad Z = \Phi^{-1}(1-p) \qquad p = 1 - \frac{1}{2}\mathrm{erf}\left(\frac{Z}{\sqrt{2}}\right)$$

LEE (cont.)

- 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:

$$p(m_0) = \int_{t_{\mathrm{obs}}(m_0)}^{\infty} f(t|0)\mathrm{d}t$$

  - Different w.r.t. the (global) probability to find a background fluctuation at least as large as your signal at **any** mass value
  - “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
  - 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:

$$t'_{\mathrm{obs}} = -2\ln\frac{L(s=0)}{L(\hat{s};\hat{m})} \qquad p' = \int_{t'_{\mathrm{obs}}}^{\infty} f(t'|0)\mathrm{d}t'$$

Estimate LEE

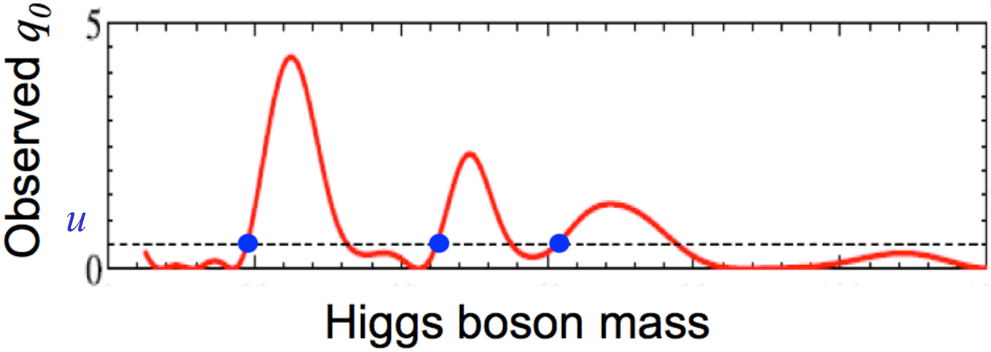
- The effect can be evaluated with brute-force Toy Monte Carlo
  - Run N experiments with background-only, find the largest 'local' significance over the entire search range, and get its distribution to determine the 'overall' significance
  - Requires very large toy Monte Carlo samples: need to go down to ~10^{-7} (5σ: $p = 2.87\times10^{-7}$)
- Approximate evaluation based on local p-value, times correction factors ("trial factors", Gross and Vitells, EPJC 70:525-530,2010, arXiv:1005.1891)

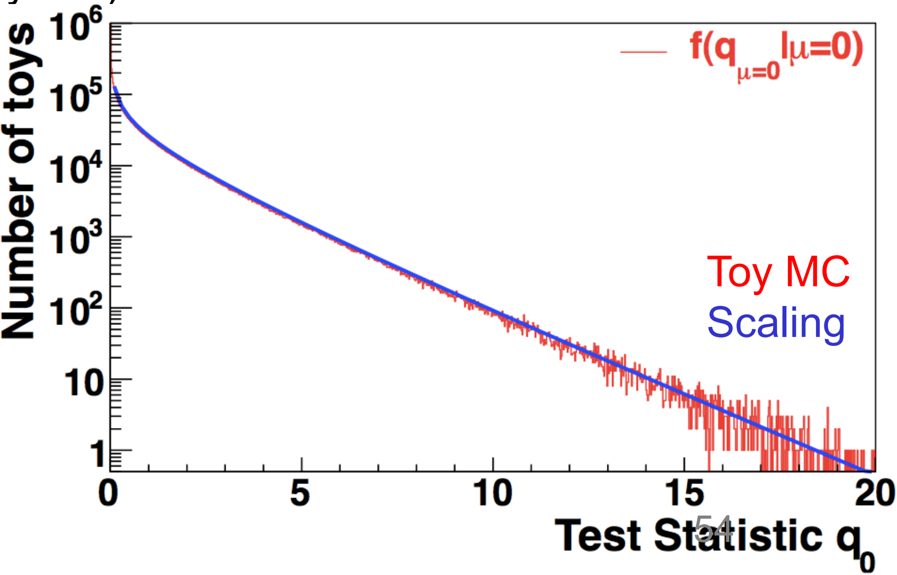
$$p_b^{global} = P(\boxed{q_0(\hat{m}_H)} > u) \leq \langle N_u \rangle + \boxed{\frac{1}{2} P_{\chi^2_1}(u)} \longleftarrow \text{Asympt. limit (Wilk's theorem)}$$

$$q_0(\hat{m}_H) = \max_{m_H} q_0(m_H)$$

- $\langle N_u \rangle$ is the average number of up-crossings of the likelihood ratio scan, can be evaluated at some lower referene level (toy MC) and scaled:

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





In conclusion

- Many recipes and approaches available
- Bayesian and Frequentist approaches lead to similar results in the easiest cases, but may diverge in frontier cases
- Be ready to master both approaches!
- … and remember that Bayesian and Frequentist limits have very different meanings

- If you want your paper to be approved soon:
  - Be consistent with your assumptions
  - Understand the meaning of what you are computing
  - Try to adopt a popular and consolidated approach (even better, software tools, like RooStats), wherever possible
  - Debate your preferred statistical technique in a statistics forum, not a physics result publication!

Backup

General likelihood definition

- The exact definition of the likelihood function depends on the data model “format”. In general:

$$\mathcal{L}(\text{data}\,|\,\mu,\theta) = \text{Poisson}\,(\,\text{data}\,|\,\underset{\substack{\uparrow\\ \text{signal strength}}}{\mu}\cdot s(\theta)+b(\theta)\,)\cdot\underbrace{p(\tilde{\theta}|\theta)}_{\substack{\text{nuisance parameters}\\ \text{PDF, typically Gaussian,}\\ \text{log-normal, flat}}}$$

- Binned case (histogram):

$$\prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!}\, e^{-\mu s_i - b_i}$$

- Unbinned case (signal/background PDFs):

$$\prod_i (\mu S f_s(x_i) + B f_b(x_i)) \cdot e^{-(\mu S + B)}$$

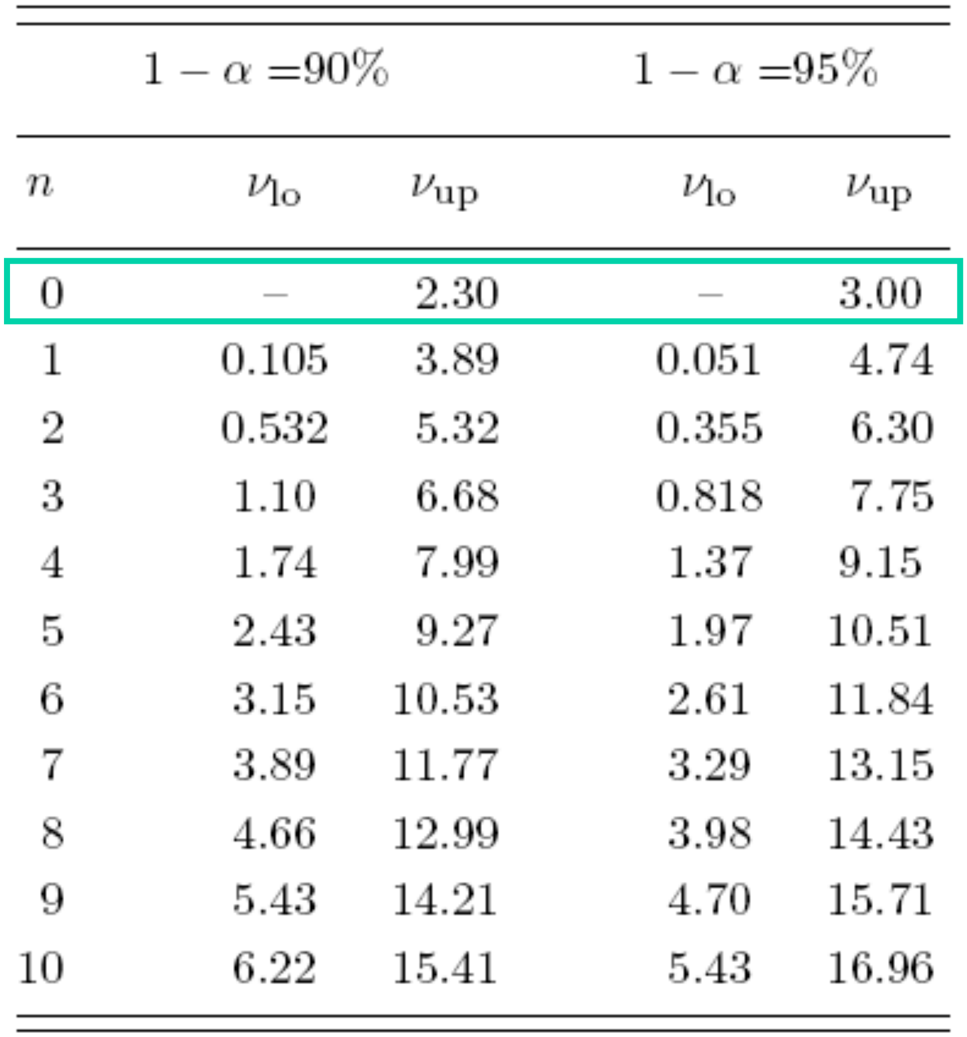
Upper limit with event counting

From PDG

in case of no background

“ *It happens that the upper limit from [central Neyman interval] coincides numerically with the Bayesian upper limit for a Poisson parameter, using a uniform prior p.d.f. for v.* ”

More details on Neyman limits in next slides…



	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	$\nu_{\rm lo}$	$\nu_{\rm up}$	$\nu_{\rm lo}$	$\nu_{\rm up}$
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51
6	3.15	10.53	2.61	11.84
7	3.89	11.77	3.29	13.15
8	4.66	12.99	3.98	14.43
9	5.43	14.21	4.70	15.71
10	6.22	15.41	5.43	16.96

Upper limits with background

- Let's start with the Bayesian approach which has an easier treatment
- A uniform prior, $\pi(s) = 1$, from 0 to ∞ simplifies the computation:

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

- Where, for a fixed b:

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

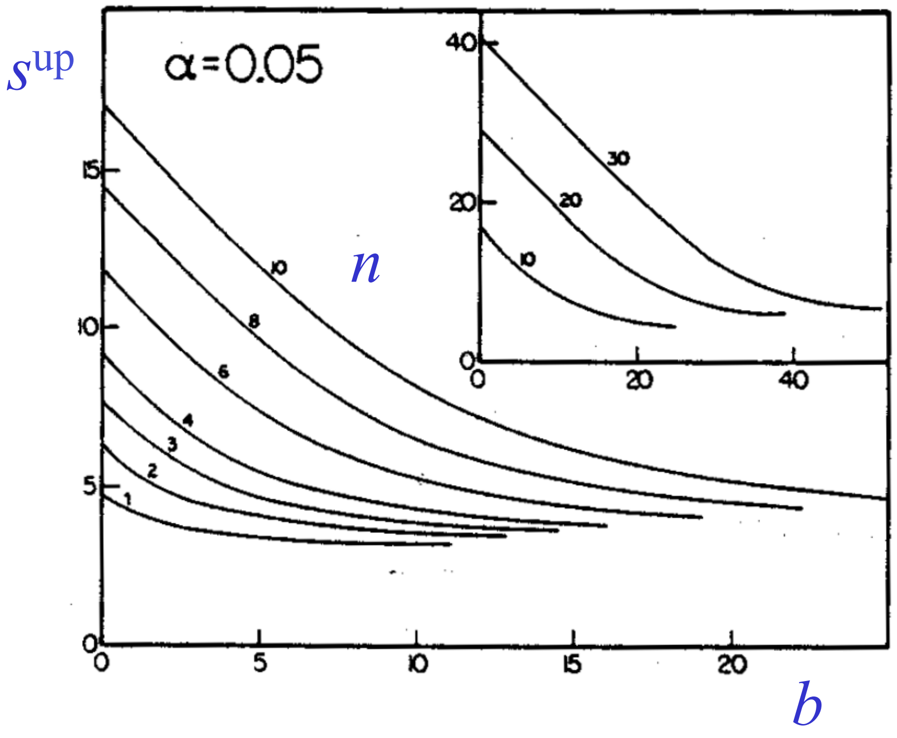
- The limit s^{up} can be obtained inverting the equation:
- The special case with $b = 0$, $n = 0$ gives the previous result

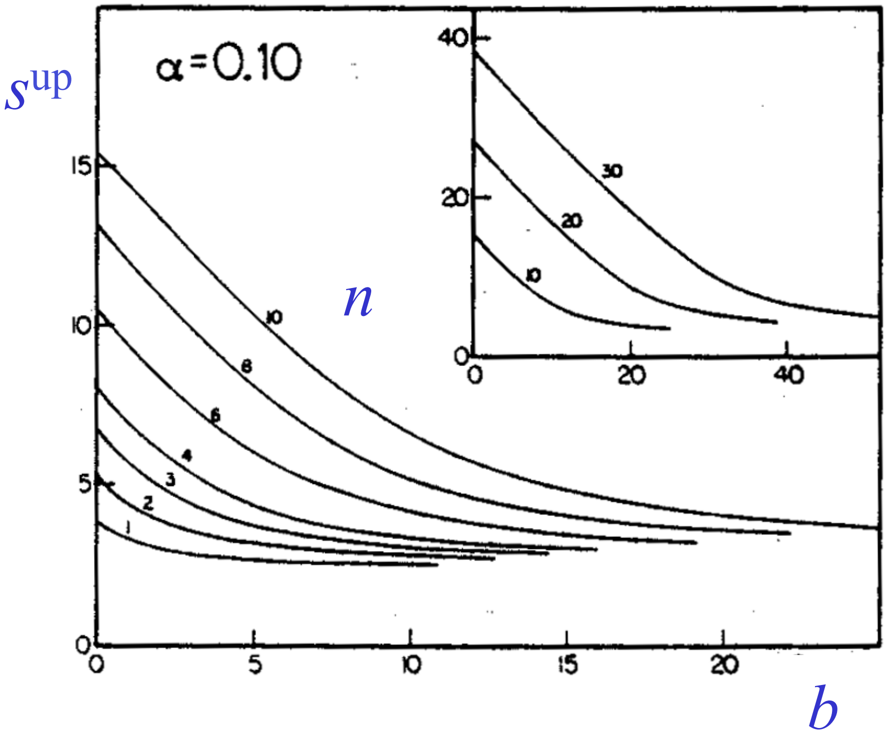
$$1 - \mathrm{CL} = e^{-s^{\mathrm{up}}} \frac{\sum_{m=0}^{n} \frac{(s^{\mathrm{up}}+b)^m}{m!}}{\sum_{m=0}^{n} \frac{b^m}{m!}}$$

Upper limits with background (cont.)

- Graphical view (due to O. Helene, 1983)

O. Helene. Nucl. Instr. and Meth. A 212 (1983), p. 319



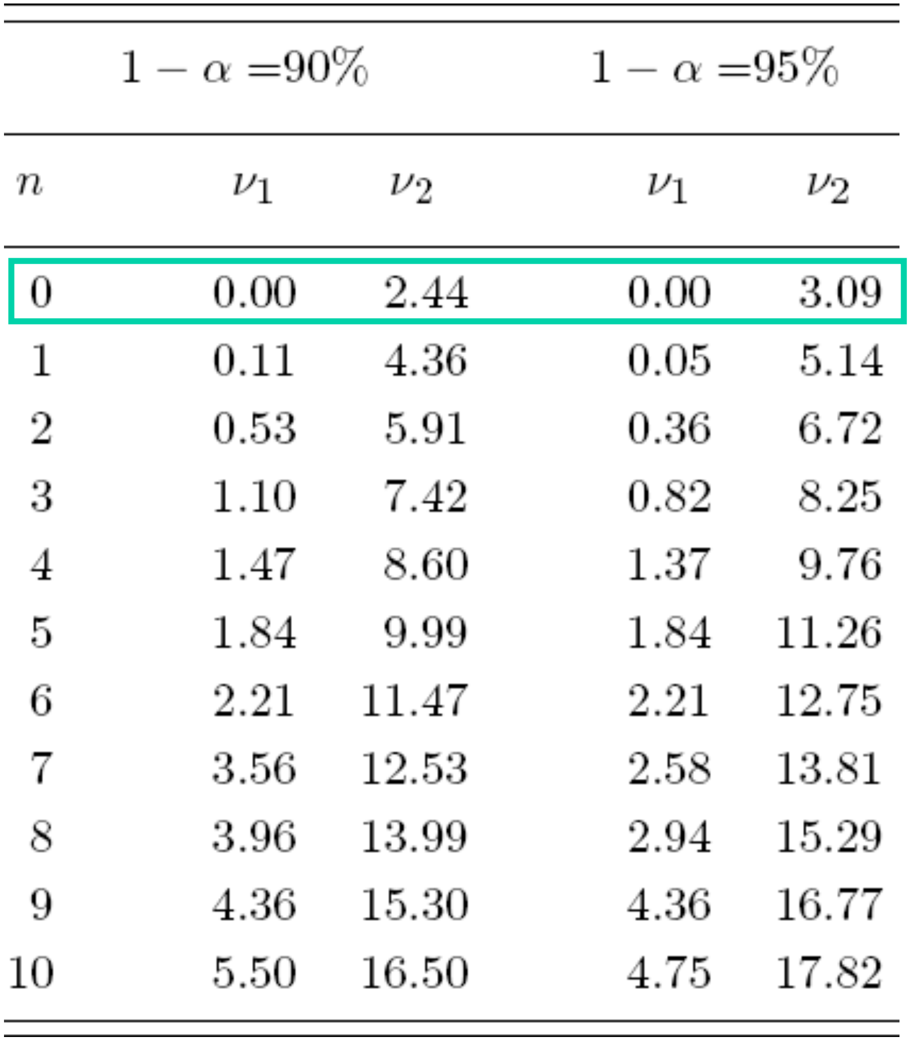


Remember, it's under the Bayesian approach

Limits in case of no background

From PDG

“Unified” (i.e.: Feldman-Cousins) limits for Poissonian counting in case of no background are larger than Bayesian limits



	$1-\alpha=90\%$		$1-\alpha=95\%$	
n	ν_1	ν_2	ν_1	ν_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

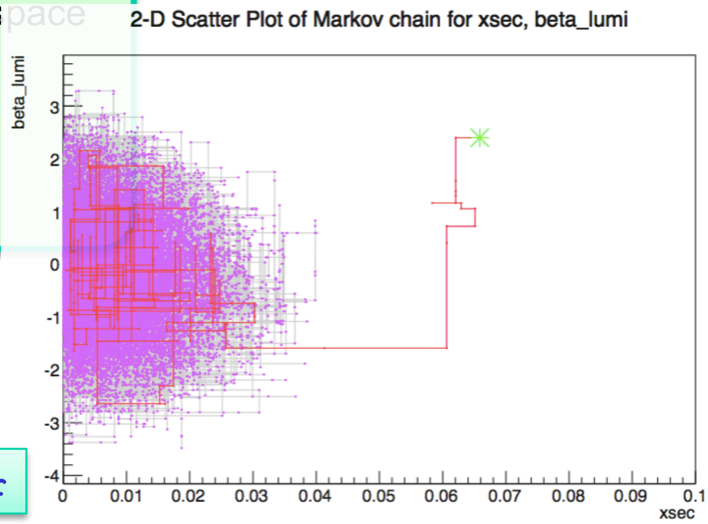
How to compute Posterior PDF

- Perform analytical integration
  - Feasible in very few cases
- Use numerical integration
  - May be CPU intensive
- Markov Chain Monte Carlo
  - Sampling parameter space efficiently using a random walk heading to the regions of higher probability
  - Metropolis algorithm to sample according to a PDF $f(x)$

RooStats::BayesianCalculator

1. Start from a random point, x_i, in the parameter space
2. Generate a proposal point x_p in the vicinity of x_i
3. If $f(x_p) > f(x_i)$ accept as next point $x_i+1 = x_p$ else, accept only with probability $p = f(x_p) / f(x_i)$
4. Repeat from point 2

  - Convergence criteria and step size must be defined



RooStats::MCMCCalculator