Statistical Methods for Particle Physics

Lecture 2: Tests based on likelihood ratios

http://www.pp.rhul.ac.uk/~cowan/stat_orsay.html

Cours d'hiver 2012 du LAL
Orsay, 3-5 January, 2012

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

Lecture 1: Introduction and basic formalism
Probability, statistical tests, confidence intervals.

→ Lecture 2: Tests based on likelihood ratios
Systematic uncertainties (nuisance parameters)
Limits for Poisson mean

Lecture 3: More on discovery and limits
Upper vs. unified limits (F-C)
Spurious exclusion, CLs, PCL
Look-elsewhere effect
Why 5σ for discovery?

A simple example

For each event we measure two variables, $\boldsymbol{x} = (x_1, x_2)$.

Suppose that for background events (hypothesis H_0),

$$f(\mathbf{x}|H_0) = \frac{1}{\xi_1} e^{-x_1/\xi_1} \, \frac{1}{\xi_2} e^{-x_2/\xi_2}$$

and for a certain signal model (hypothesis H_1) they follow

$$f(\mathbf{x}|H_1) = C \, \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2} \, \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x_2-\mu_2)^2/2\sigma_2^2}$$

where $x_1, x_2 \geq 0$ and C is a normalization constant.

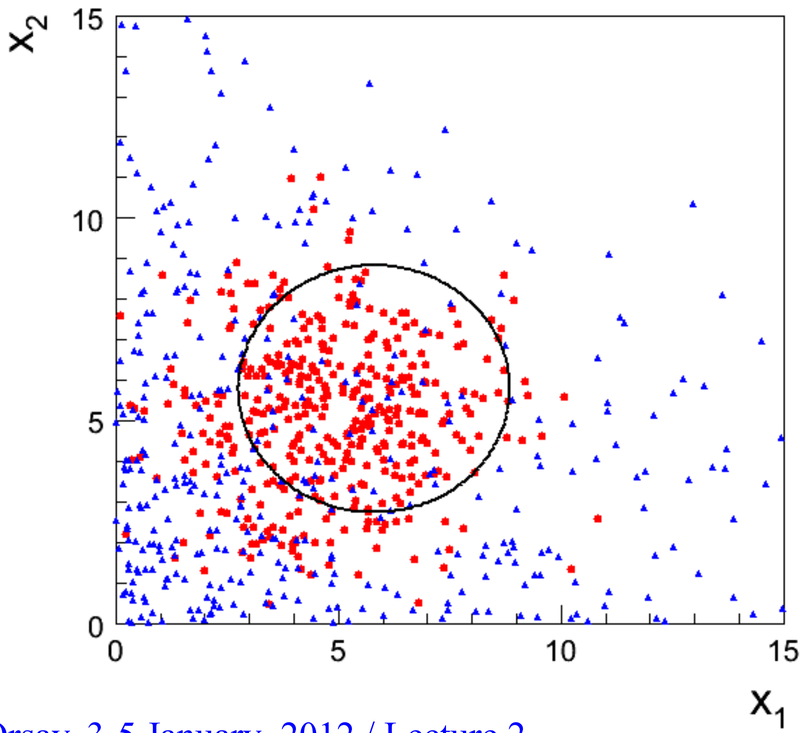
Likelihood ratio as test statistic

In a real-world problem we usually wouldn't have the pdfs $f(\boldsymbol{x}|H_0)$ and $f(\boldsymbol{x}|H_1)$, so we wouldn't be able to evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

for a given observed $\boldsymbol{x}$, hence the need for multivariate methods to approximate this with some other function.

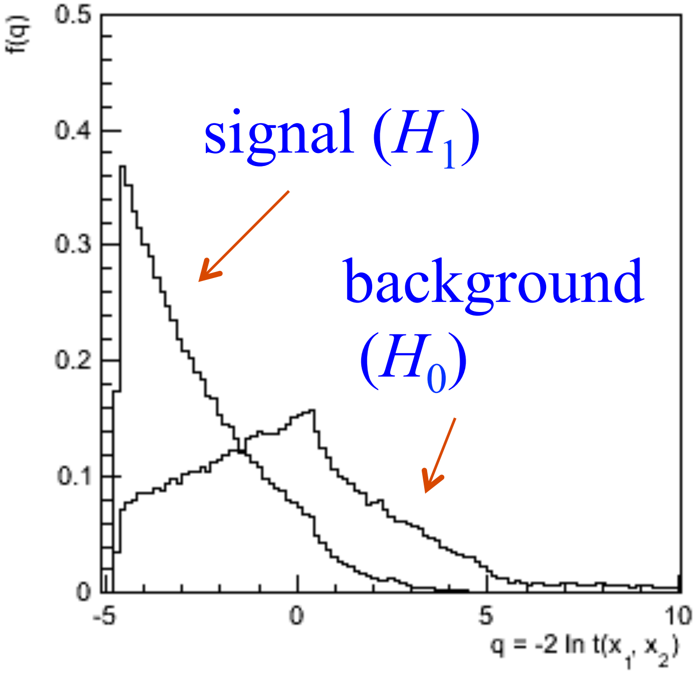
But in this example we can find contours of constant likelihood ratio such as:



Event selection using the LR

Using Monte Carlo, we can find the distribution of the likelihood ratio or equivalently of

$$q = \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - \frac{2x_1}{\xi_1} - \frac{2x_2}{\xi_2} = -2 \ln t(\mathbf{x}) + C$$



From the Neyman-Pearson lemma we know that by cutting on this variable we would select a signal sample with the highest signal efficiency (test power) for a given background efficiency.

Search for the signal process

But what if the signal process is not known to exist and we want to search for it. The relevant hypotheses are therefore

H_0: all events are of the background type
H_1: the events are a mixture of signal and background

Rejecting H_0 with $Z > 5$ constitutes "discovering" new physics.

Suppose that for a given integrated luminosity, the expected number of signal events is s, and for background b.

The observed number of events n will follow a Poisson distribution:

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

Likelihoods for full experiment

We observe n events, and thus measure n instances of $\boldsymbol{x} = (x_1, x_2)$.

The likelihood function for the entire experiment assuming the background-only hypothesis (H_0) is

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^{n} f(\mathbf{x}_i | \mathrm{b})$$

and for the "signal plus background" hypothesis (H_1) it is

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^{n} \left(\pi_{\mathrm{s}} f(\mathbf{x}_i | \mathrm{s}) + \pi_{\mathrm{b}} f(\mathbf{x}_i | \mathrm{b}) \right)$$

where π_s and π_b are the (prior) probabilities for an event to be signal or background, respectively.

Likelihood ratio for full experiment

We can define a test statistic Q monotonic in the likelihood ratio as

$$Q = -2 \ln \frac{L_{s+b}}{L_b} = -s + \sum_{i=1}^{n} \ln \left(1 + \frac{s}{b} \frac{f(\mathbf{x}_i | \mathrm{s})}{f(\mathbf{x}_i | \mathrm{b})} \right)$$

To compute p-values for the b and s+b hypotheses given an observed value of Q we need the distributions $f(Q|\mathrm{b})$ and $f(Q|\mathrm{s+b})$.

Note that the term $-s$ in front is a constant and can be dropped.

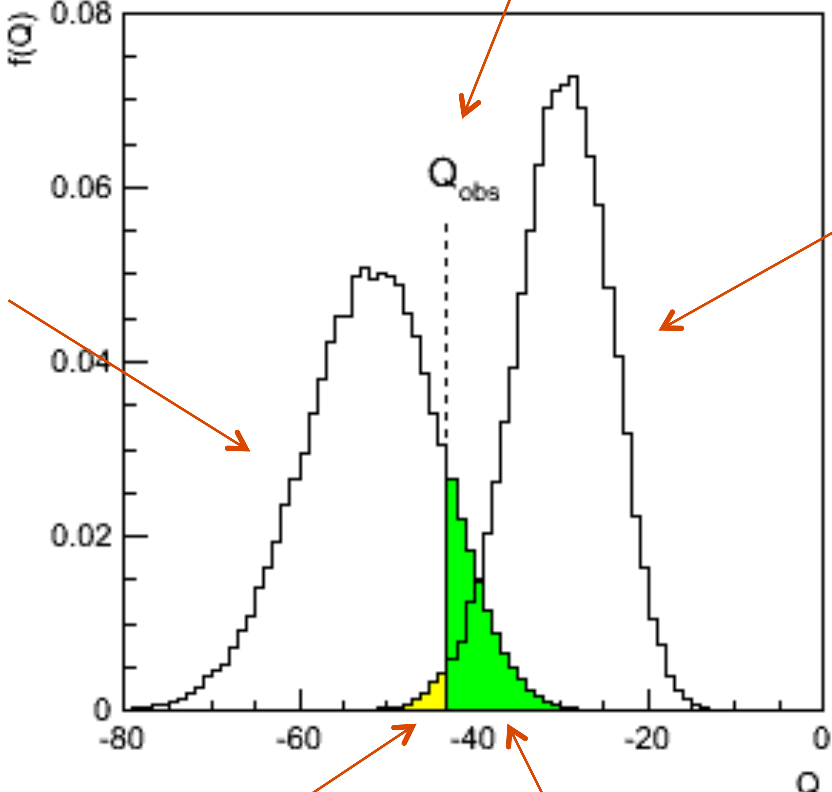
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

Can exploit this to relate distribution of Q to that of single event terms using (Fast) Fourier Transforms (Hu and Nielsen, physics/9906010).

Distribution of Q

Take e.g. b = 100, s = 20.

Suppose in real experiment Q is observed here.



$f(Q|\text{s+b})$

$f(Q|\text{b})$

p-value of b only

p-value of s+b

Systematic uncertainties

Up to now we assumed all parameters were known exactly.

In practice they have some (systematic) uncertainty.

Suppose e.g. uncertainty in expected number of background events b is characterized by a (Bayesian) pdf $\pi(b)$.

Maybe take a Gaussian, i.e.,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_0)^2/2\sigma_b^2}$$

where b_0 is the nominal (measured) value and σ_b is the estimated uncertainty.

In fact for many systematics a Gaussian pdf is hard to defend – more on this later.

Distribution of Q with systematics

To get the desired p-values we need the pdf $f(Q)$, but this depends on b, which we don't know exactly.

But we can obtain the prior predictive (marginal) model:

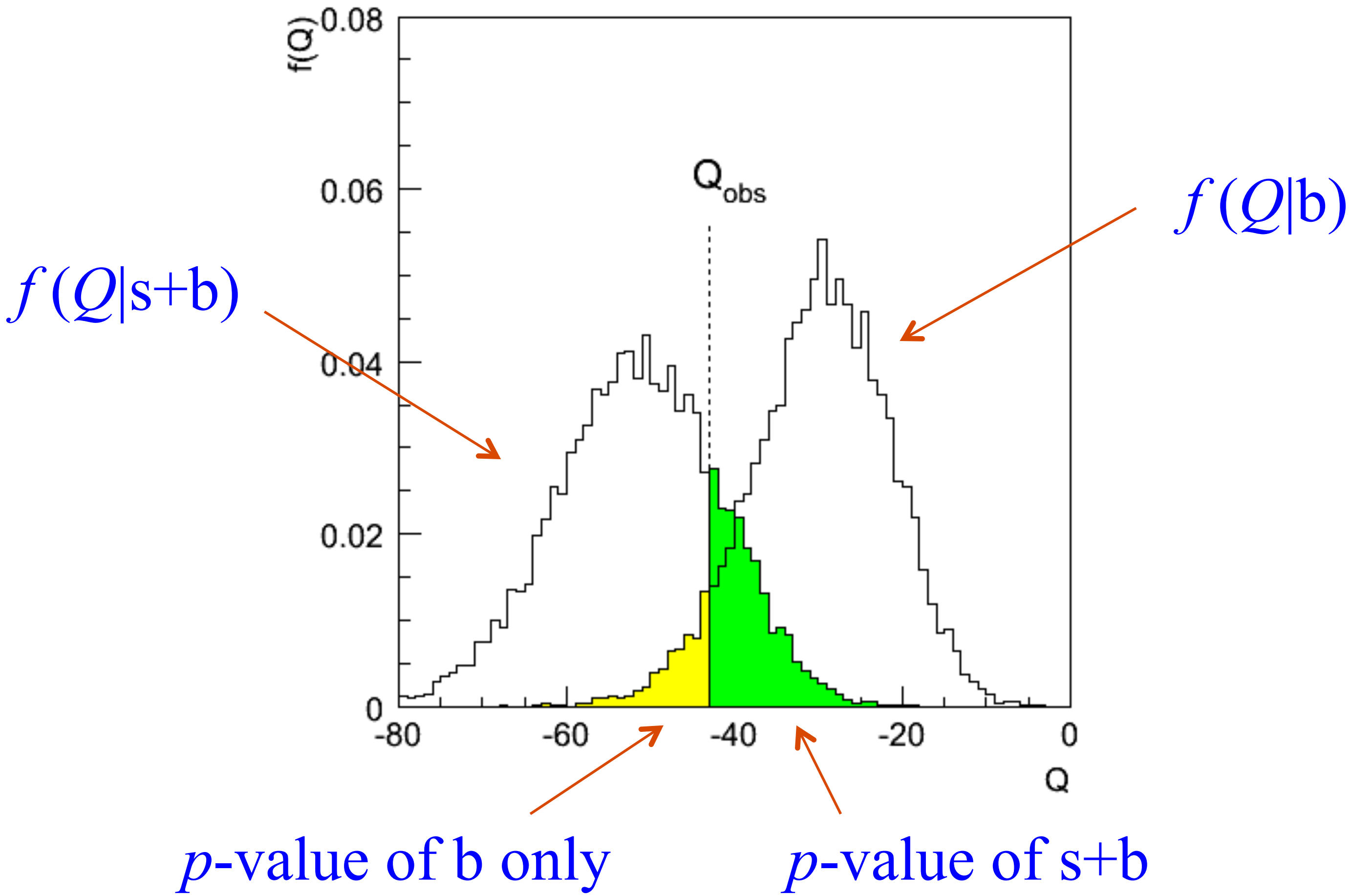
$$f(Q) = \int f(Q|b)\pi(b)\,db$$

With Monte Carlo, sample b from $\pi(b)$, then use this to generate Q from $f(Q|b)$, i.e., a new value of b is used to generate the data for every simulation of the experiment.

This broadens the distributions of Q and thus increases the p-value (decreases significance Z) for a given Q_{obs}.

Distribution of Q with systematics (2)

For $s = 20$, $b_0 = 100$, $\sigma_b = 20$ this gives



Using the likelihood ratio $L(s)/L(\hat{s})$

Instead of the likelihood ratio L_{s+b}/L_b, suppose we use as a test statistic

$$\lambda(s) = \frac{L(s)}{L(\hat{s})}$$

← maximizes $L(s)$

Intuitively this is a measure of the level of agreement between the data and the hypothesized value of s.

low λ: poor agreement
high λ : better agreement
$0 \leq \lambda \leq 1$

$L(s)/L(\hat{s})$ for counting experiment

Consider an experiment where we only count n events with $n \sim \text{Poisson}(s + b)$. Then $\hat{s} = n - b$.

To establish discovery of signal we test the hypothesis $s = 0$ using

$$\ln \lambda(0) = n \ln(b) - b - n \ln n + n$$

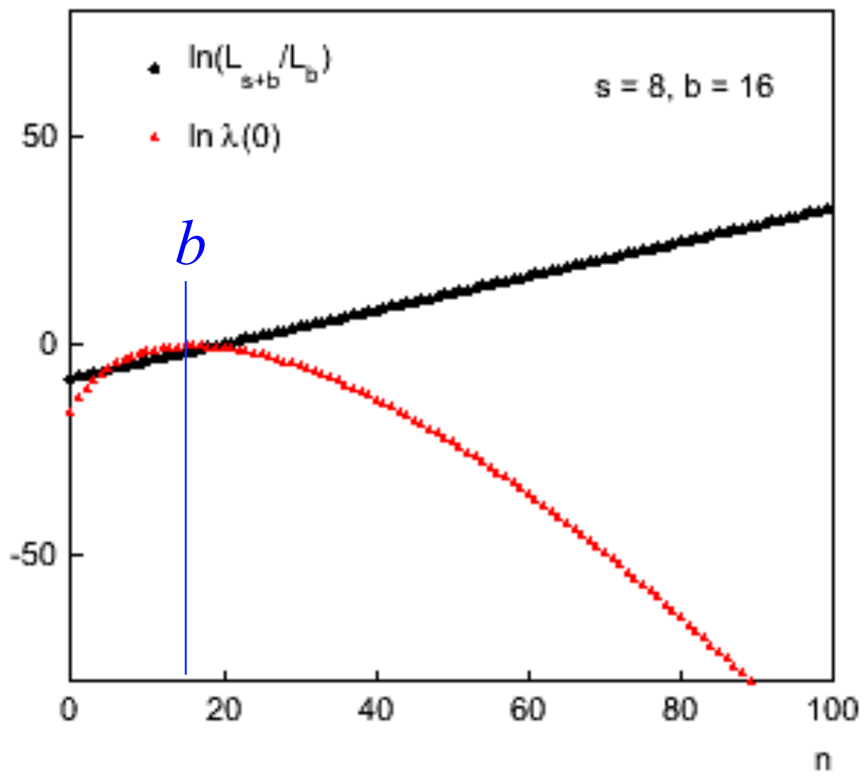
whereas previously we had used

$$\ln \frac{L_{s+b}}{L_b} = n \ln \left(1 + \frac{s}{b} \right) - s$$

which is monotonic in n and thus equivalent to using n as the test statistic.

$L(s)/L(\hat{s})$ for counting experiment (2)

But if we only consider the possibility of signal being present when $n > b$, then in this range $\lambda(0)$ is also monotonic in n, so both likelihood ratios lead to the same test.



$L(s)/L(\hat{s})$ for general experiment

If we do not simply count events but also measure for each some set of numbers, then the two likelihood ratios do not necessarily give equivalent tests, but in practice should be very close.

$\lambda(s)$ has the important advantage that for a sufficiently large event sample, its distribution approaches a well defined form (Wilks' Theorem).

In practice the approach to the asymptotic form is rapid and one obtains a good approximation even for relatively small data samples (but need to check with MC).

This remains true even when we have adjustable nuisance parameters in the problem, i.e., parameters that are needed for a correct description of the data but are otherwise not of interest (key to dealing with systematic uncertainties).

Large-sample approximations for prototype analysis using profile likelihood ratio

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Search for signal in a region of phase space; result is histogram of some variable *x* giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\rm tot} \int_{\mathrm{bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx\,, \qquad b_i = b_{\rm tot} \int_{\mathrm{bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx\,.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$,b_{tot})

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

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

maximizes L for Specified μ

maximize L

The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

The profile LR hould be near-optimal in present analysis with variable μ and nuisance parameters $\boldsymbol{\theta}$.

Test statistic for discovery

Try to reject background-only ($\mu = 0$) hypothesis using

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only regard upward fluctuation of data as evidence against the background-only hypothesis.

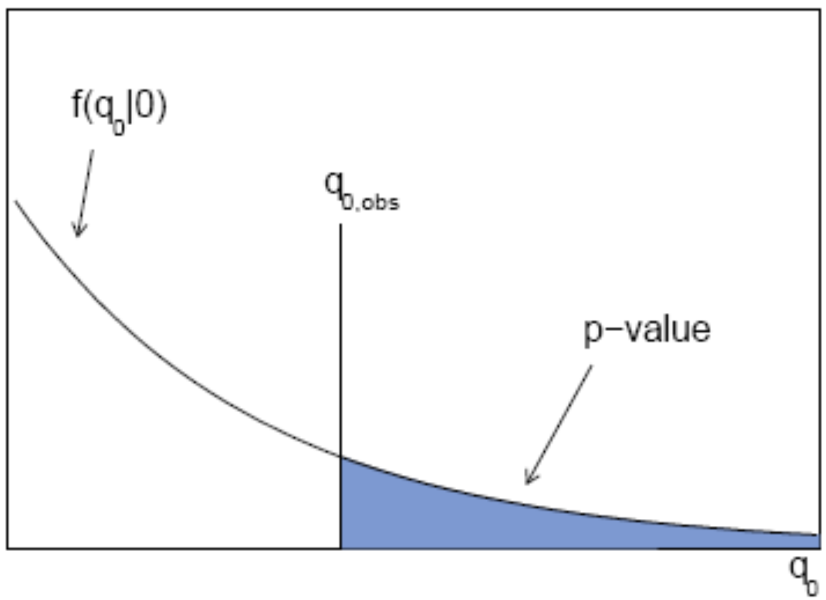
Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

p-value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p-value for an observed $q_{0,\mathrm{obs}}$ is

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

will get formula for this later

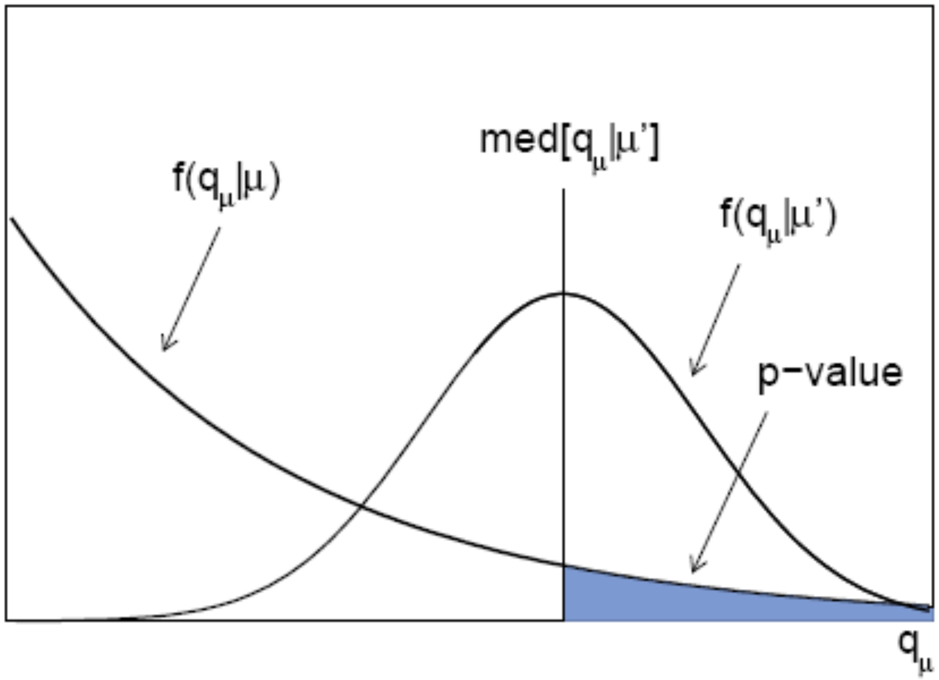


From p-value get equivalent significance,

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

Expected (or median) significance / sensitivity

When planning the experiment, we want to quantify how sensitive we are to a potential discovery, e.g., by given median significance assuming some nonzero strength parameter μ'.



So for p-value, need $f(q_0|0)$, for sensitivity, will need $f(q_0|\mu')$,

Test statistic for upper limits

For purposes of setting an upper limit on μ one may use

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

Note for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ.

From observed q_μ find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

95% CL upper limit on μ is highest value for which p-value is not less than 0.05.

Alternative test statistic for upper limits

Assume physical signal model has $\mu > 0$, therefore if estimator for μ comes out negative, the closest physical model has $\mu = 0$.

Therefore could also measure level of discrepancy between data and hypothesized μ with

$$\tilde{\lambda}(\mu) = \begin{cases} \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\[2ex] \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0 \, . \end{cases} \qquad \tilde{q}_\mu = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Performance not identical to but very close to q_μ (of previous slide).

q_μ is simpler in important ways: asymptotic distribution is independent of nuisance parameters.

Wald approximation for profile likelihood ratio

To find *p*-values, we need: $f(q_0|0), \quad f(q_\mu|\mu)$

For median significance under alternative, need: $f(q_\mu|\mu')$

Use approximation due to Wald (1943)

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

sample size

$$\hat{\mu} \sim \text{Gaussian}(\mu', \sigma)$$

i.e., $E[\hat{\mu}] = \mu'$

σ from covariance matrix V, use, e.g.,

$$V^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

Noncentral chi-square for $-2\ln\lambda(\mu)$

If we can neglect the $O(1/\sqrt{N})$ term, $-2\ln\lambda(\mu)$ follows a noncentral chi-square distribution for one degree of freedom with noncentrality parameter

$$\Lambda = \frac{(\mu - \mu')^2}{\sigma^2}$$

As a special case, if $\mu' = \mu$ then $\Lambda = 0$ and $-2\ln\lambda(\mu)$ follows a chi-square distribution for one degree of freedom (Wilks).

The Asimov data set

To estimate median value of $-2\ln\lambda(\mu)$, consider special data set where all statistical fluctuations suppressed and n_i, m_i are replaced by their expectation values (the "Asimov" data set):

$$n_i = \mu' s_i + b_i$$

$$m_i = u_i$$

$$\rightarrow \quad \hat{\mu} = \mu' \qquad \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$$

$$\lambda_{\rm A}(\mu) = \frac{L_{\rm A}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{\rm A}(\hat{\mu}, \hat{\boldsymbol{\theta}})} = \frac{L_{\rm A}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_A(\mu', \boldsymbol{\theta})}$$

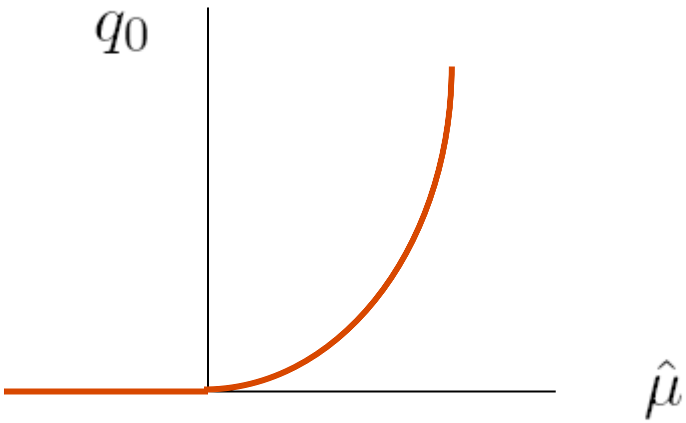
$$-2\ln\lambda_{\rm A}(\mu) = \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda$$

Asimov value of $-2\ln\lambda(\mu)$ gives non-centrality param. Λ, or equivalently, σ.

Relation between test statistics and $\hat{\mu}$

Assuming Wald approximation, the relation between q_0 and $\hat{\mu}$ is

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2 & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$



Monotonic, therefore quantiles of $\hat{\mu}$ map one-to-one onto those of q_0, e.g.,

$$\mathrm{med}[q_0] = q_0(\mathrm{med}[\hat{\mu}]) = q_0(\mu') = \frac{\mu'^2}{\sigma^2} = -2\ln\lambda_{\mathrm{A}}(0)$$

Distribution of q_0

Assuming the Wald approximation, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a "half chi-square" distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The p-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

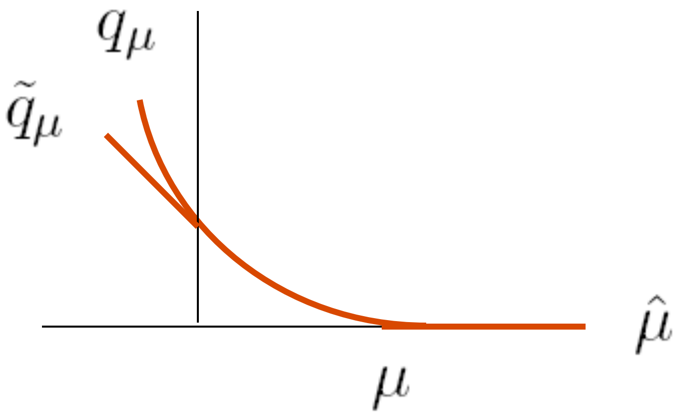
Therefore the discovery significance Z is simply

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$

Relation between test statistics and $\hat{\mu}$

Assuming the Wald approximation for $-2\ln\lambda(\mu)$, q_μ and $\tilde{q}_\mu$ both have monotonic relation with μ.

$$q_\mu = \begin{cases} \frac{(\mu-\hat{\mu})^2}{\sigma^2} & \hat{\mu} < \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$



$$\tilde{q}_\mu = \begin{cases} \frac{\mu^2}{\sigma^2} - \frac{2\mu\hat{\mu}}{\sigma^2} & \hat{\mu} < 0 \\ \frac{(\mu-\hat{\mu})^2}{\sigma^2} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \ , \end{cases}$$

And therefore quantiles of q_μ, $\tilde{q}_\mu$ can be obtained directly from those of $\hat{\mu}$ (which is Gaussian).

Distribution of q_μ

Similar results for q_μ

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{(\mu - \mu')}{\sigma}\right)^2\right]$$

$$f(q_\mu|\mu) = \frac{1}{2}\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}e^{-q_\mu/2}$$

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{(\mu - \mu')}{\sigma}\right)$$

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

Distribution of $\tilde{q}_\mu$

Similar results for $\tilde{q}_\mu$

$$f(\tilde{q}_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(\tilde{q}_\mu)$$

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

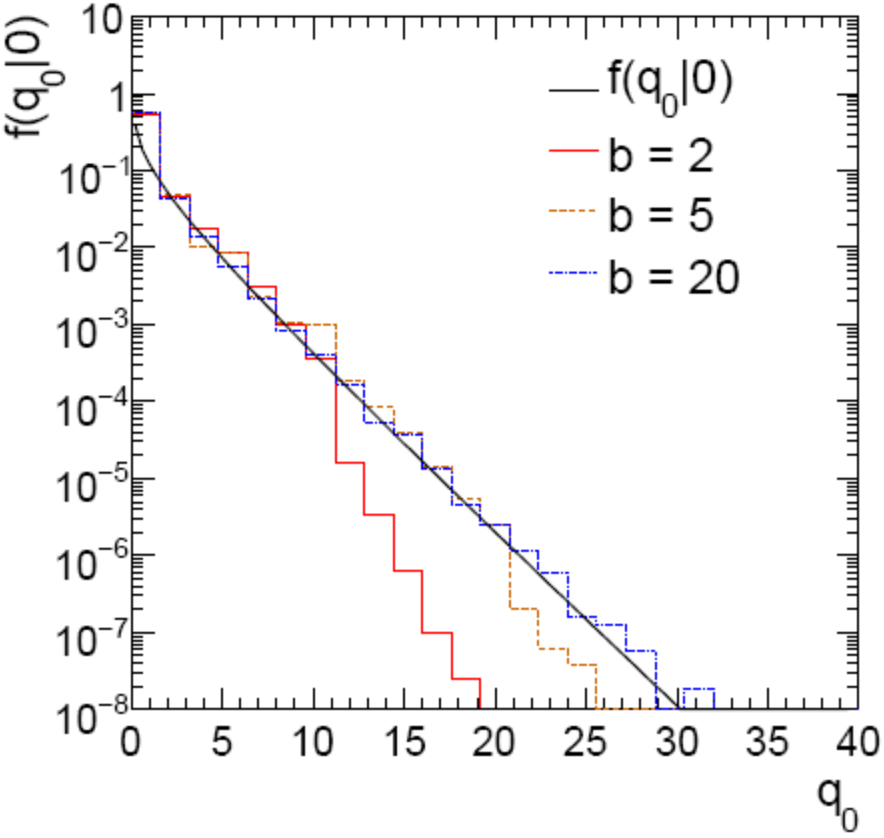
$$F(\tilde{q}_\mu|\mu') = \begin{cases} \Phi\left(\sqrt{\tilde{q}_\mu} - \frac{(\mu-\mu')}{\sigma}\right) & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2 \;, \\ \Phi\left(\frac{\tilde{q}_\mu - (\mu^2 - 2\mu\mu')/\sigma^2}{2\mu/\sigma}\right) & \tilde{q}_\mu > \mu^2/\sigma^2 \;. \end{cases}$$

Monte Carlo test of asymptotic formula

$$n \sim \text{Poisson}(\mu s + b)$$

$$m \sim \text{Poisson}(\tau b)$$

Here take $\tau = 1$.



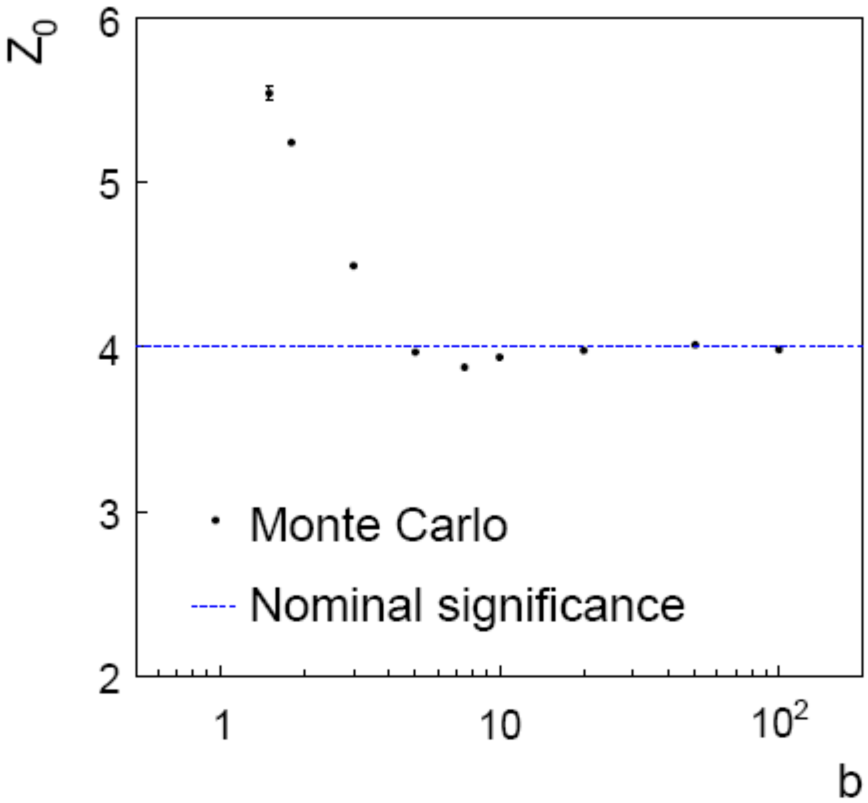
Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.

Monte Carlo test of asymptotic formulae

Significance from asymptotic formula, here $Z_0 = \sqrt{q_0} = 4$, compared to MC (true) value.

For very low b, asymptotic formula underestimates Z_0.

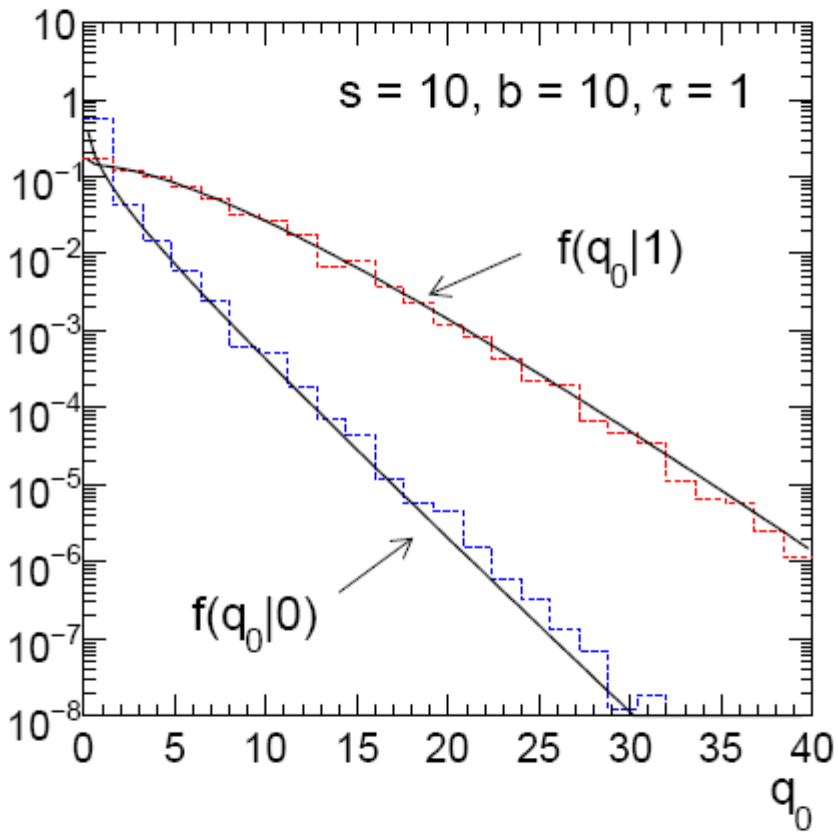
Then slight overshoot before rapidly converging to MC value.

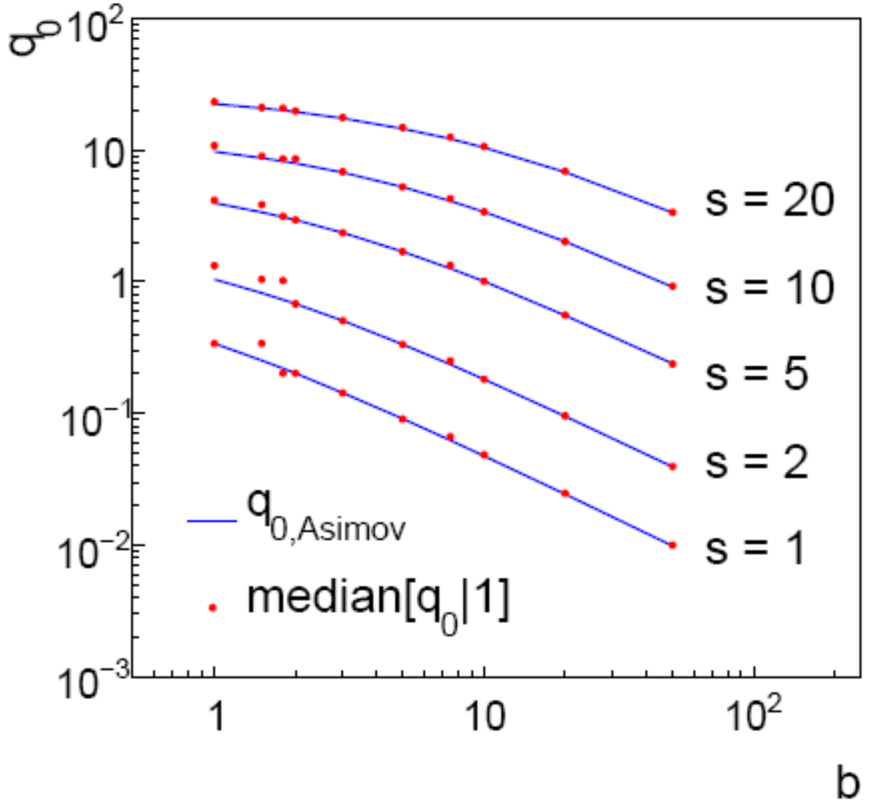


Monte Carlo test of asymptotic formulae

Asymptotic $f(q_0|1)$ good already for fairly small samples.

Median$[q_0|1]$ from Asimov data set; good agreement with MC.





Monte Carlo test of asymptotic formulae

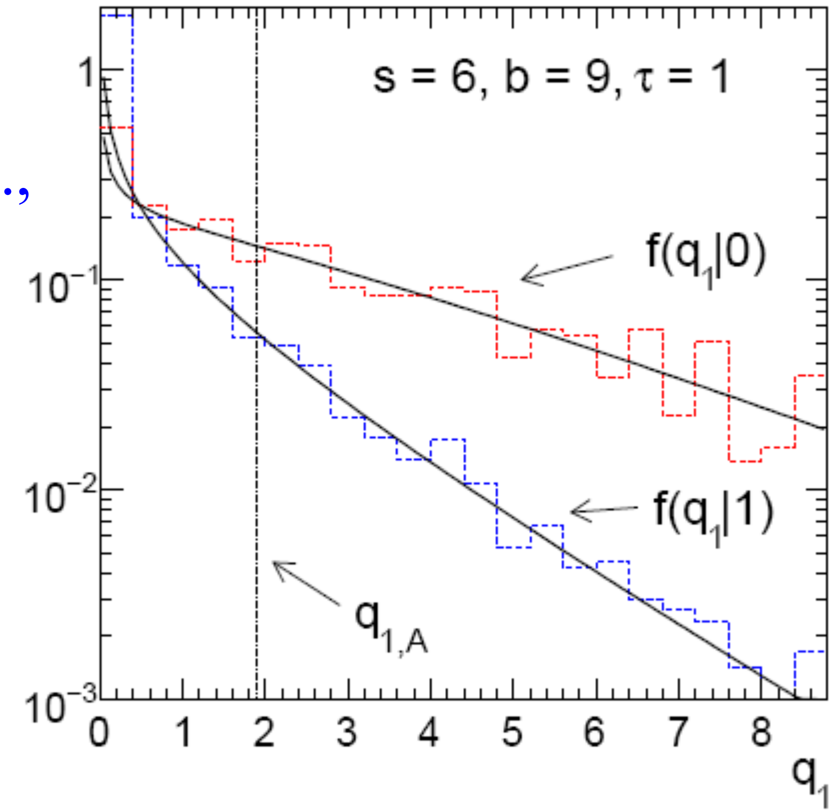
Consider again $n \sim$ Poisson $(\mu s + b)$, $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median$[q_1|0]$ gives "exclusion sensitivity".

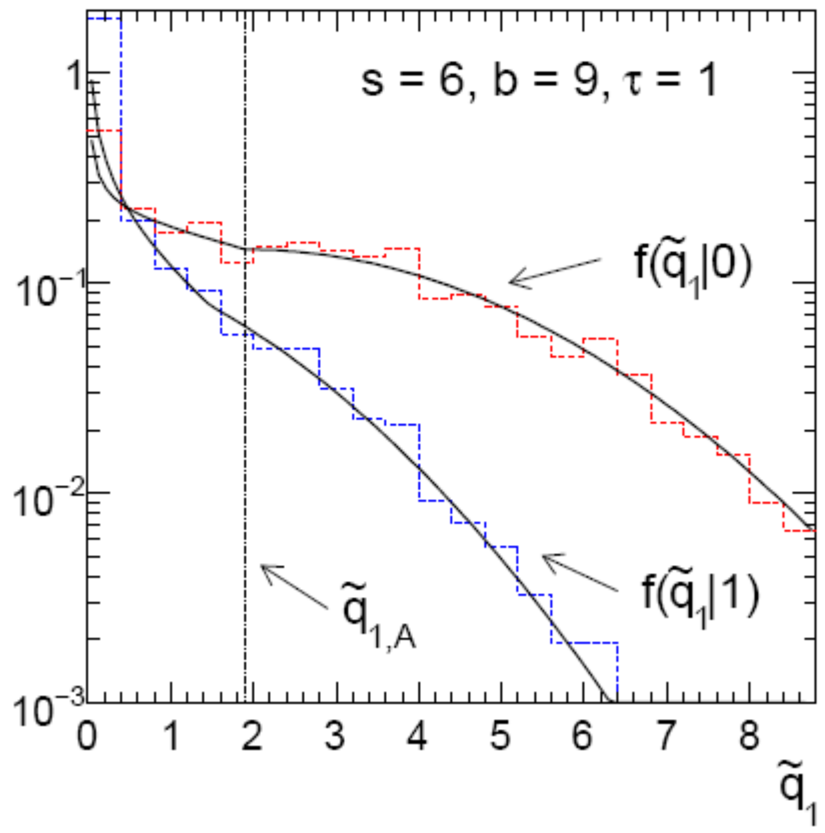
Here asymptotic formulae good for $s = 6$, $b = 9$.



Monte Carlo test of asymptotic formulae

Same message for test based on $\tilde{q}_\mu$.

q_μ and $\tilde{q}_\mu$ give similar tests to the extent that asymptotic formulae are valid.



Setting limits on Poisson parameter

Consider again the case of finding $n = n_s + n_b$ events where

n_b events from known processes (background)
n_s events from a new process (signal)

are Poisson r.v.s with means s, b, and thus $n = n_s + n_b$ is also Poisson with mean $= s + b$. Assume b is known.

Suppose we are searching for evidence of the signal process, but the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{obs} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s.

Upper limit for Poisson parameter

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\text{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\text{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\text{up}} = -\ln \gamma \approx 3.00$$

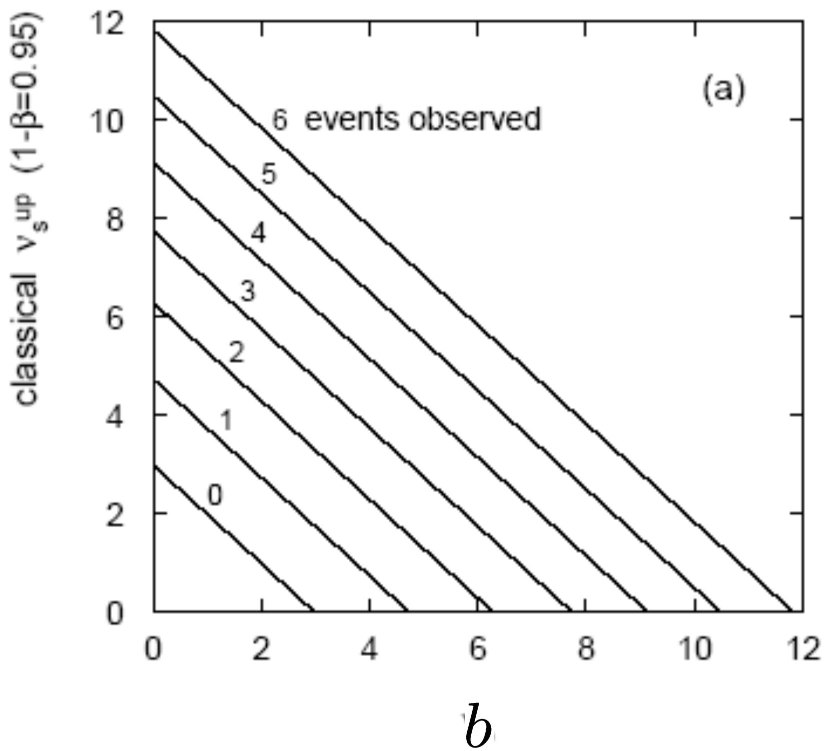
Calculating Poisson parameter limits

To solve for s_{lo}, s_{up}, can exploit relation to χ^2 distribution:

$$s_{\text{lo}} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha; 2n) - b$$

Quantile of χ^2 distribution

$$s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1-\beta; 2(n+1)) - b$$



For low fluctuation of n the formula can give negative result for s_{up}; i.e. confidence interval is empty.

Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for s_{up}

$$s_{up} = -0.197 \quad (\mathrm{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when limit of parameter is close to a physical boundary.

Expected limit for $s = 0$

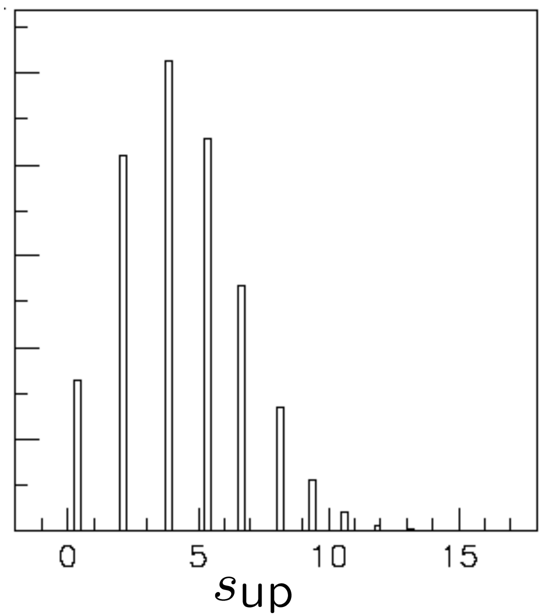
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

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

Integrate posterior pdf $p(\theta \mid x)$ to give interval with any desired probability content.

For e.g. $n \sim$ Poisson($s+b$), 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s_{\rm up}} p(s|n)\,ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for s

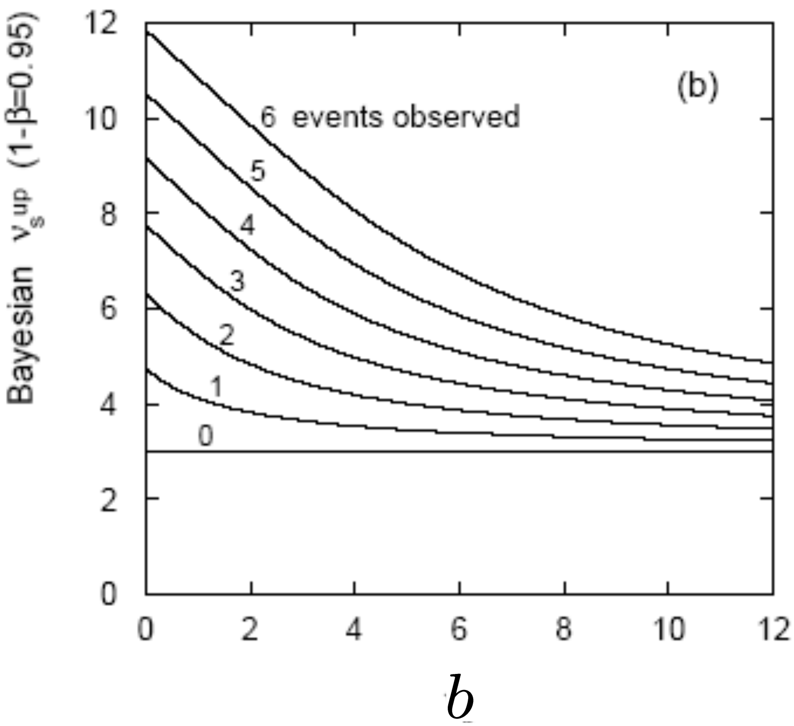
Solve numerically to find limit s_{up}.

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as classical case ('coincidence').

Otherwise Bayesian limit is everywhere greater than classical ('conservative').

Never goes negative.

Doesn't depend on b if $n = 0$.



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called "objective priors"
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In a Subjective Bayesian analysis, using objective priors can be an important part of the sensitivity analysis.

Priors from formal rules (cont.)

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties. For a review see:

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arxiv:1002.1111 (Feb 2010)

Jeffreys' prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln L(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} L(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters.

For a Gaussian mean, the Jeffreys' prior is constant; for a Poisson mean μ it is proportional to $1/\sqrt{\mu}$.

Jeffreys' prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu}$$

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu}$$

$$I = -E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{(s + b)}$, which depends on b. Note this is not designed as a degree of belief about s.

Bayesian limits with uncertainty on b

Uncertainty on b goes into the prior, e.g.,

$$\pi(s,b) = \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)}$$

$$\pi_s(s) = \text{const}, \; \sim 1/\sqrt{s+b}\ldots$$

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over the nuisance parameter b,

$$p(s|n) = \int p(s,b|n)\, db$$

Then use $p(s|n)$ to find intervals for s with any desired probability content.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form,
also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

Google for 'MCMC', 'Metropolis', 'Bayesian computation', ...

MCMC generates correlated sequence of random numbers:
cannot use for many applications, e.g., detector MC;
effective stat. error greater than $\sqrt{n}$.

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an *n*-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$ e.g. Gaussian centred about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathrm{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than naive $\sqrt{n}$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.
If proposed step rejected, hop in place.

More on priors

Suppose we measure $n \sim \text{Poisson}(s+b)$, goal is to make inference about s.

Suppose b is not known exactly but we have an estimate $\hat{b}$ with uncertainty σ_b.

For Bayesian analysis, first reflex may be to write down a Gaussian prior for b,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-\hat{b})^2/\sigma_b^2}$$

But a Gaussian could be problematic because e.g.

$b \geq 0$, so need to truncate and renormalize;

tails fall off very quickly, may not reflect true uncertainty.

Gamma prior for b

What is in fact our prior information about b? It may be that we estimated b using a separate measurement (e.g., background control sample) with

$m \sim$ Poisson(τb) (τ = scale factor, here assume known)

Having made the control measurement we can use Bayes' theorem to get the probability for b given m,

$$\pi(b|m) \propto P(m|b)\pi_0(b) \propto \frac{(\tau b)^m}{m!} e^{-\tau b} \pi_0(b)$$

If we take the "original" prior $\pi_0(b)$ to be to be constant for $b \geq 0$, then the posterior $\pi(b|m)$, which becomes the subsequent prior when we measure n and infer s, is a Gamma distribution with:

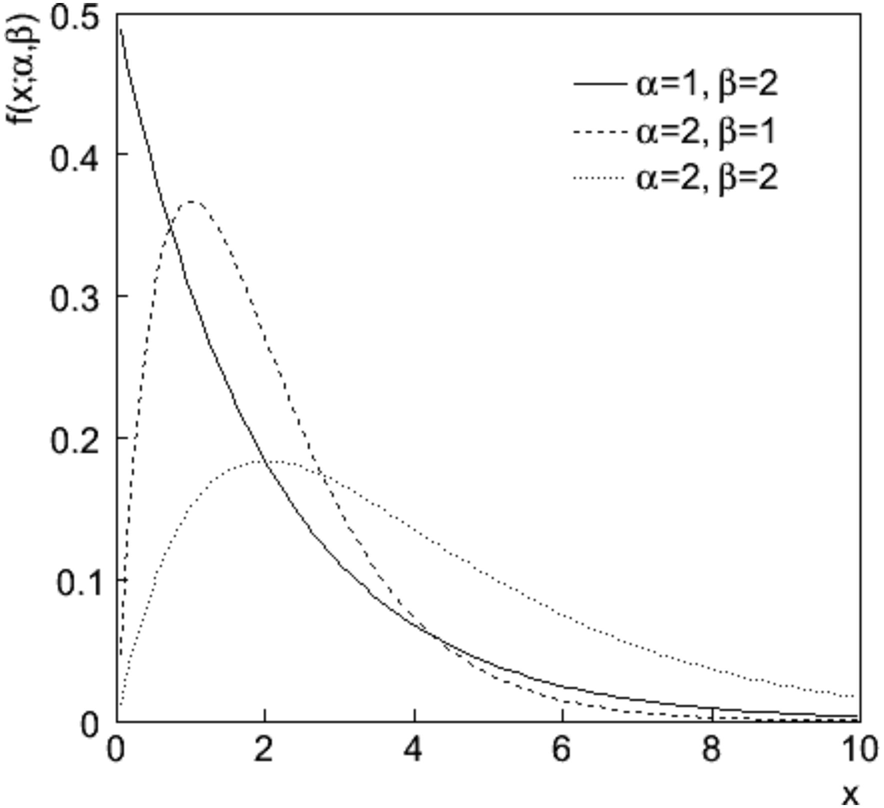
mean = $(m + 1) / \tau$
standard dev. = $\sqrt{(m + 1)} / \tau$

Gamma distribution

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

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$



Frequentist approach to same problem

In the frequentist approach we would regard both variables

$n \sim \text{Poisson}(s+b)$
$m \sim \text{Poisson}(\tau b)$

as constituting the data, and thus the full likelihood function is

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

Use this to construct test of s with e.g. profile likelihood ratio

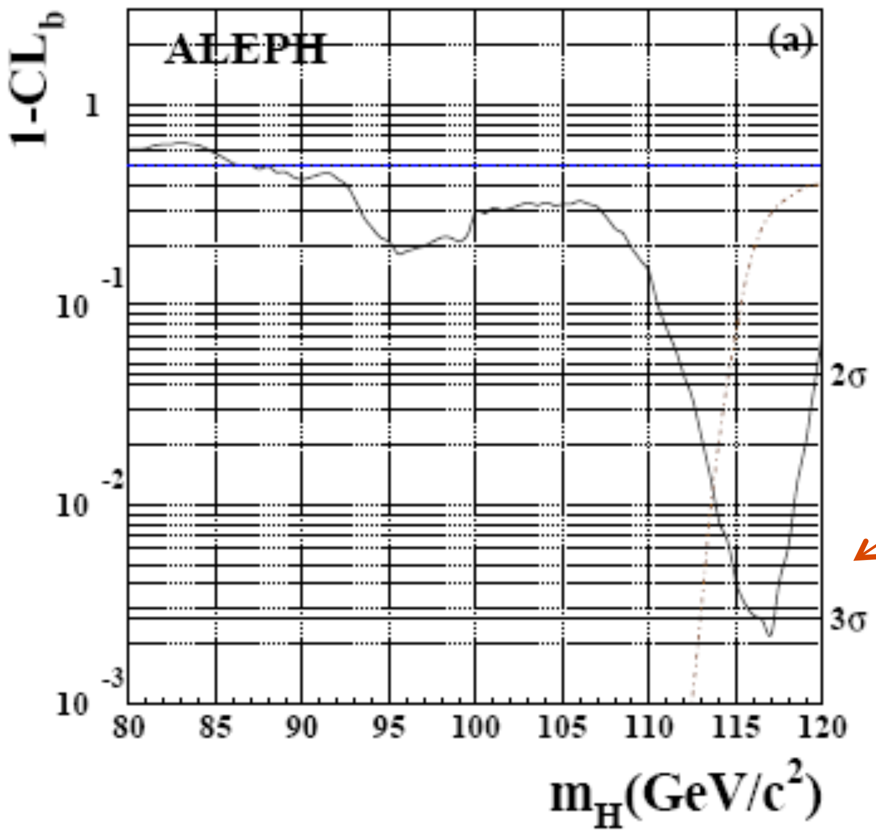
$$\lambda(s) = \frac{L(s, \hat{\hat{b}})}{L(\hat{s}, \hat{b})}$$

Note here that the likelihood refers to both n and m, whereas the likelihood used in the Bayesian calculation only modeled n.

Extra Slides

Example: ALEPH Higgs search

p-value ($1 - CL_b$) of background only hypothesis versus tested Higgs mass measured by ALEPH Experiment

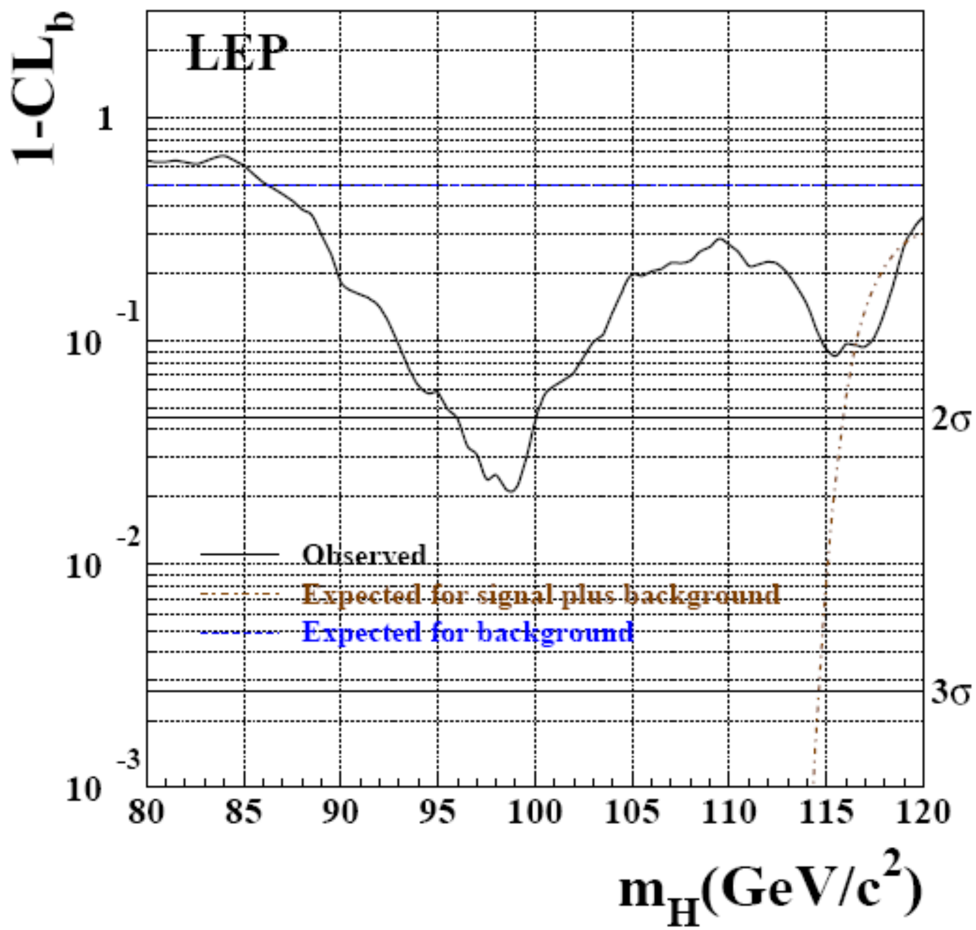


Possible signal?

Phys.Lett.B565:61-75,2003.
hep-ex/0306033

Example: LEP Higgs search

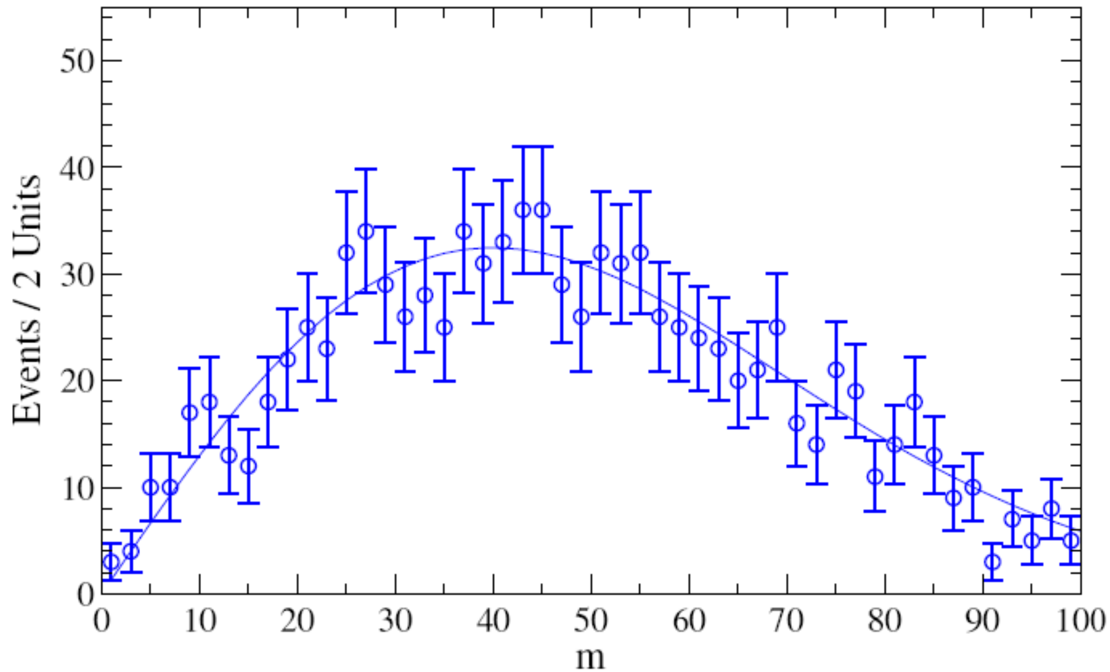
Not seen by the other LEP experiments. Combined analysis gives *p*-value of background-only hypothesis of 0.09 for m_H = 115 GeV.



Phys.Lett.B565:61-75,2003.
hep-ex/0306033

Example 2: Shape analysis

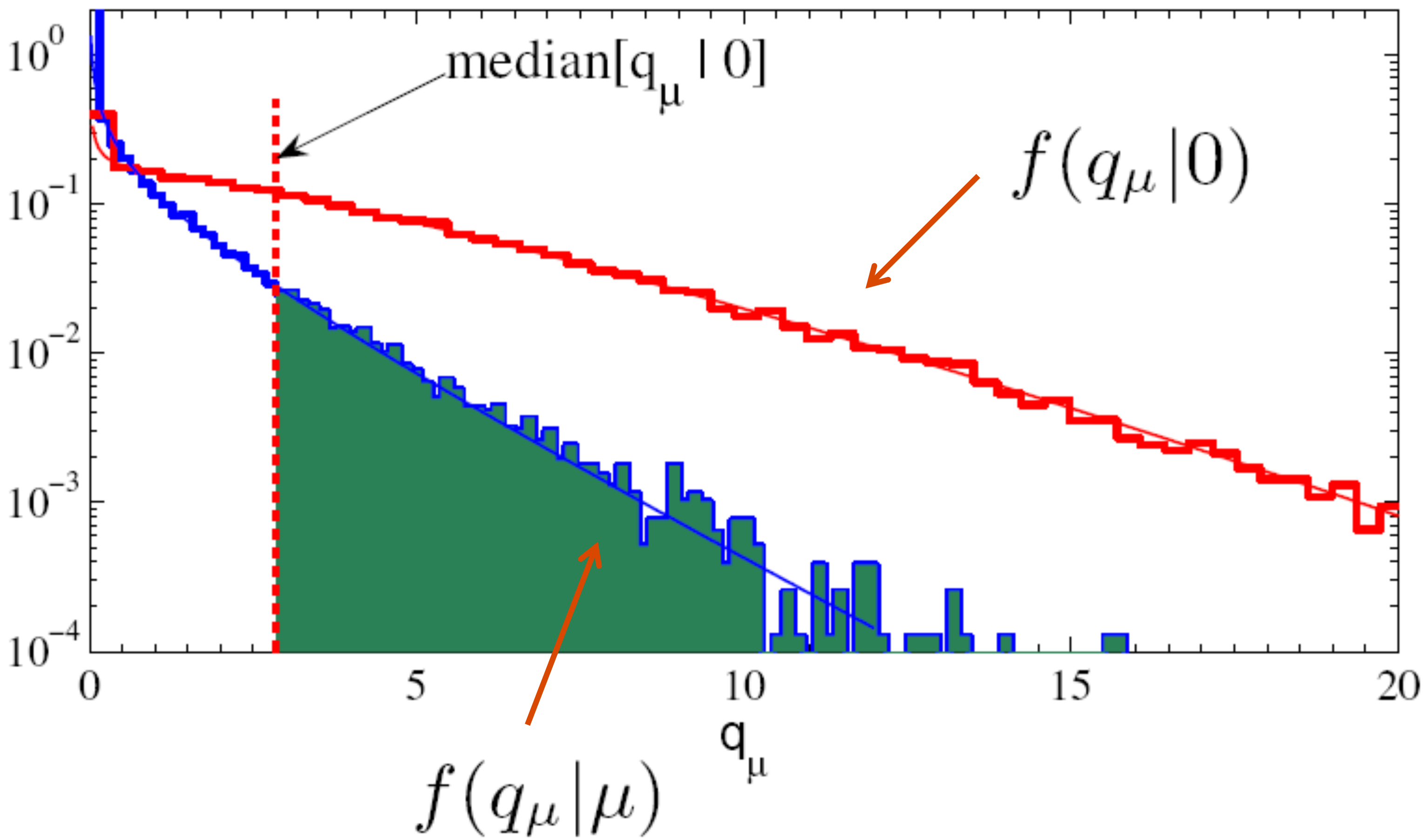
Look for a Gaussian bump sitting on top of:



$$L(\mu, \theta) = \prod_{i=1}^{N} \frac{(\mu s_i + \theta f_{\mathrm{b},i})^{n_i}}{n_i!} e^{-(\mu s_i + \theta f_{\mathrm{b},i})}$$

Monte Carlo test of asymptotic formulae

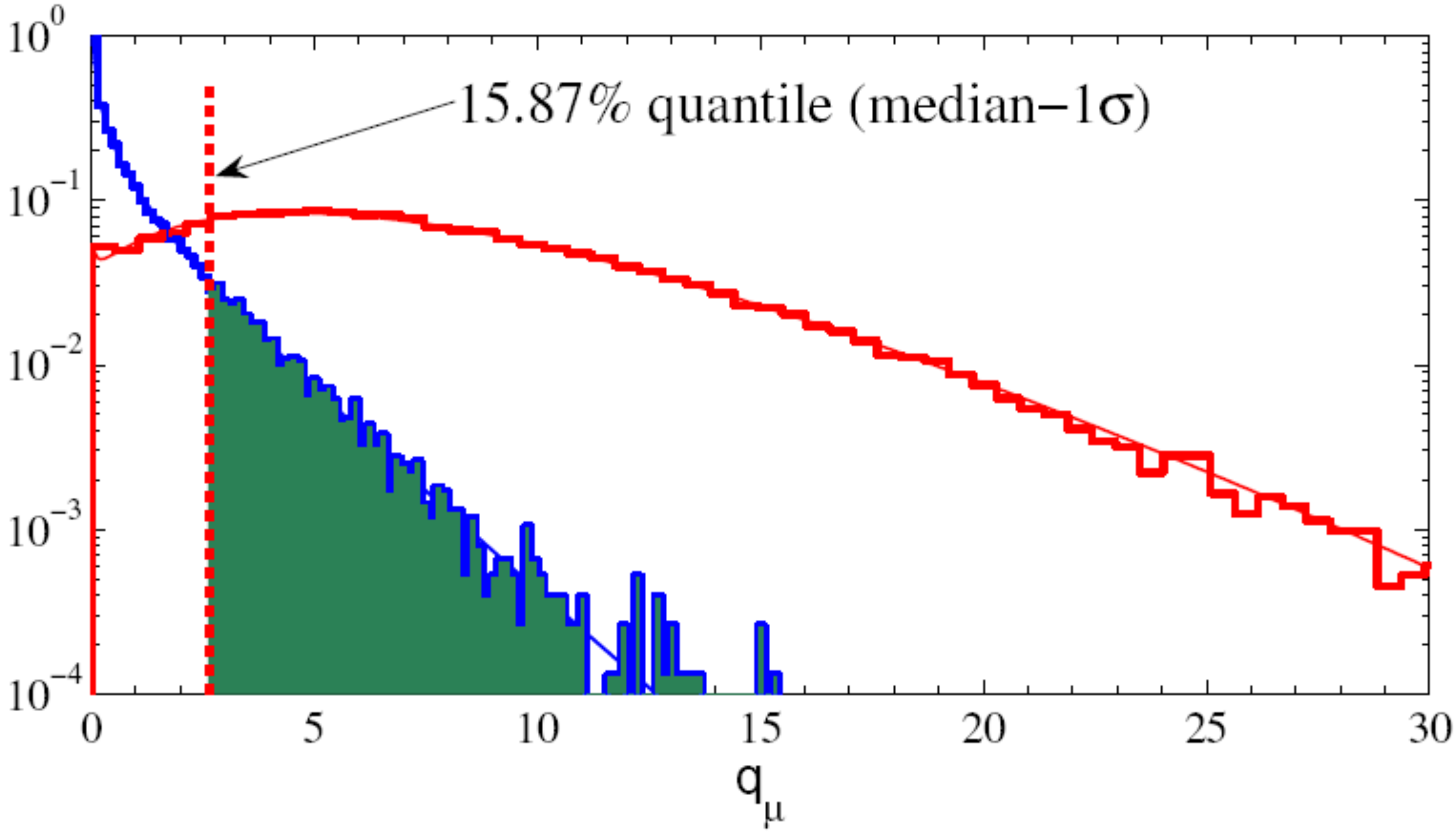
Distributions of q_μ here for μ that gave $p_\mu = 0.05$.



Using $f(q_\mu|0)$ to get error bands

We are not only interested in the median $[q_\mu|0]$; we want to know how much statistical variation to expect from a real data set.

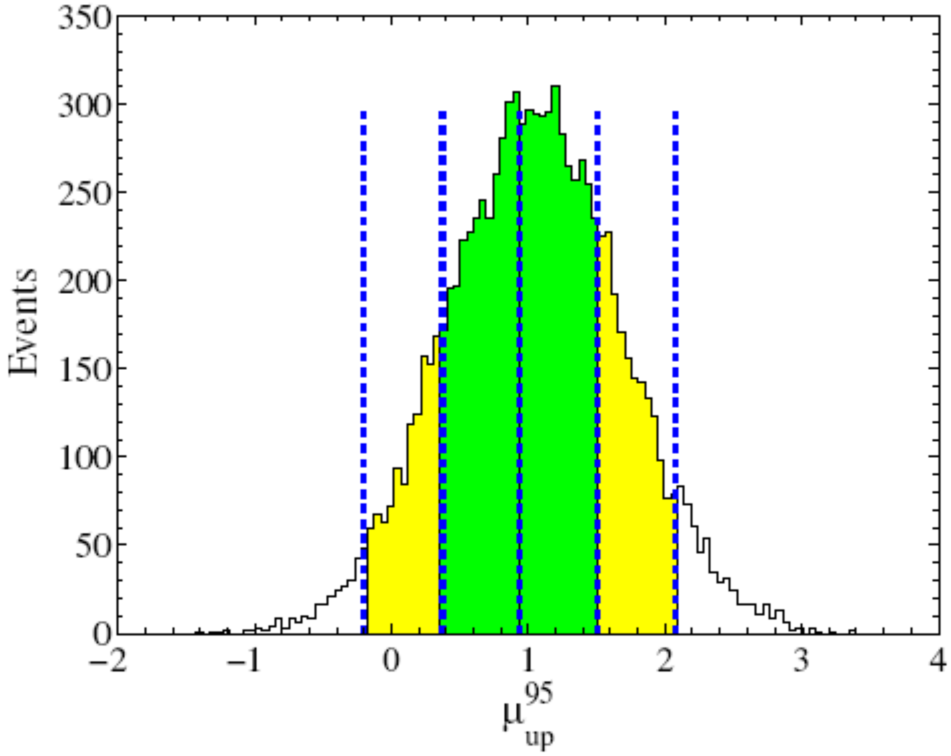
But we have full $f(q_\mu|0)$; we can get any desired quantiles.



Distribution of upper limit on μ

$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from MC;

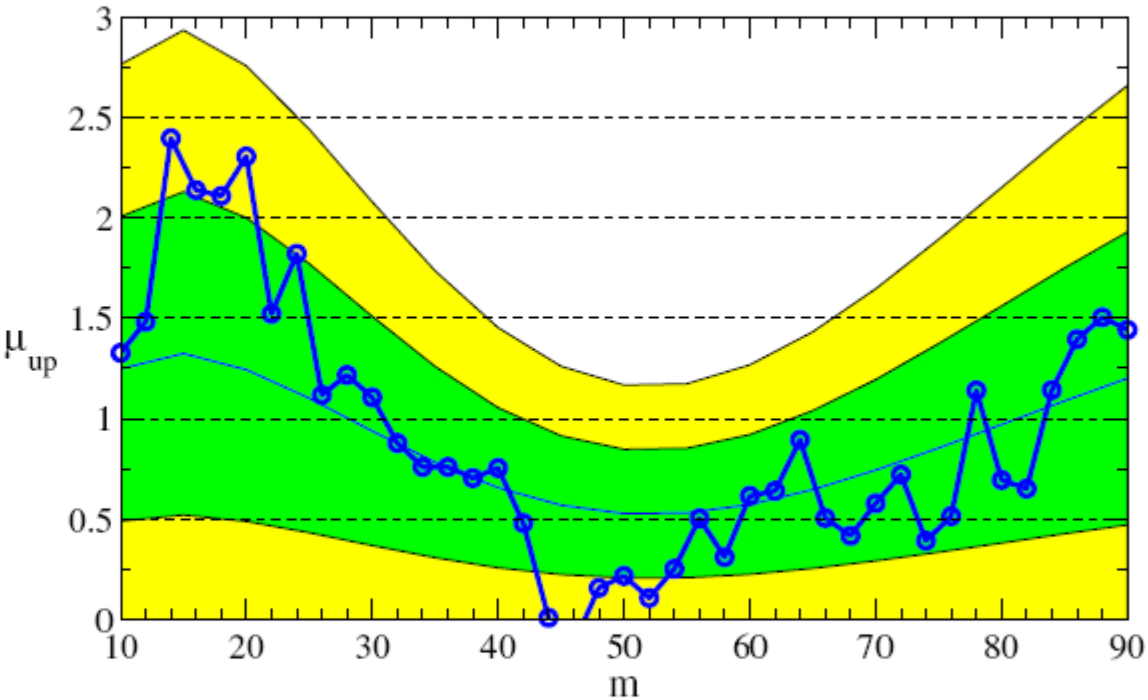
Vertical lines from asymptotic formulae



Limit on μ versus peak position (mass)

$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from asymptotic formulae;

Points are from a single arbitrary data set.



Using likelihood ratio L_{s+b}/L_b

Many searches at the Tevatron have used the statistic

$$q = -2 \ln \frac{L_{s+b}}{L_b}$$

likelihood of $\mu = 1$ model (s+b)

likelihood of $\mu = 0$ model (bkg only)

This can be written

$$q = -2 \ln \frac{L(\mu = 1, \hat{\hat{\boldsymbol{\theta}}}(1))}{L(\mu = 0, \hat{\hat{\boldsymbol{\theta}}}(0))} = -2 \ln \lambda(1) + 2 \ln \lambda(0)$$

Wald approximation for L_{s+b}/L_b

Assuming the Wald approximation, q can be written as

$$q = \frac{(\hat{\mu} - 1)^2}{\sigma^2} - \frac{\hat{\mu}^2}{\sigma^2} = \frac{1 - 2\hat{\mu}}{\sigma^2}$$

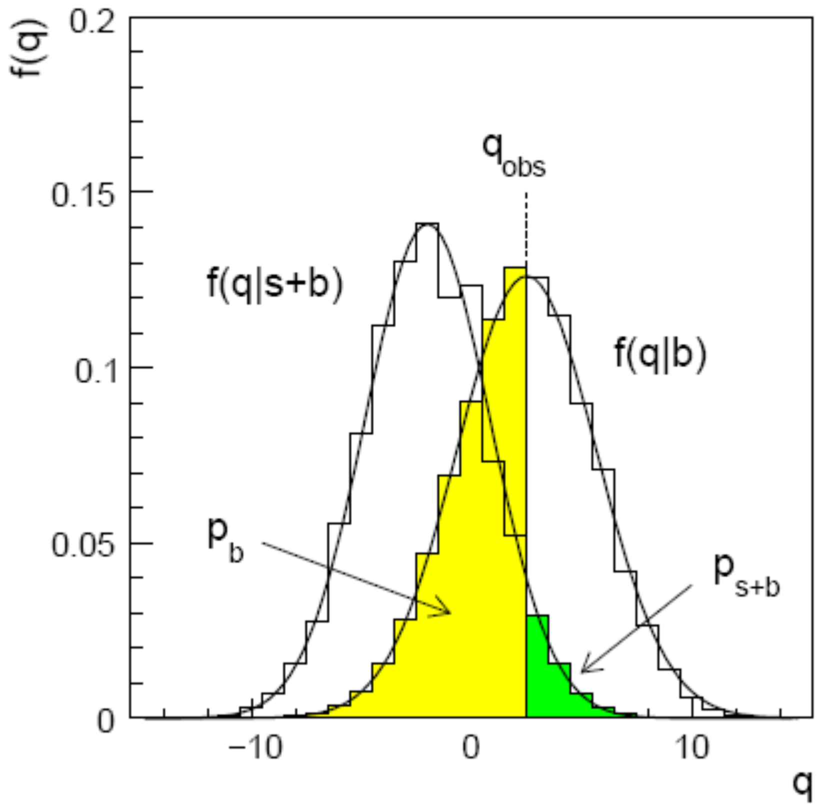
i.e. q is Gaussian distributed with mean and variance of

$$E[q] = \frac{1 - 2\mu}{\sigma^2} \qquad V[q] = \frac{4}{\sigma^2}$$

To get σ^2 use 2nd derivatives of ln$L$ with Asimov data set.

Example with L_{s+b}/L_b

Consider again $n \sim$ Poisson $(\mu s + b)$, $m \sim$ Poisson(τb)
$b = 20$, $s = 10$, $\tau = 1$.



So even for smallish data sample, Wald approximation can be useful; no MC needed.

Discovery significance for $n \sim$ Poisson($s + b$)

Consider again the case where we observe n events , model as following Poisson distribution with mean $s + b$ (assume b is known).

1) For an observed n, what is the significance Z_0 with which we would reject the $s = 0$ hypothesis?

2) What is the expected (or more precisely, median) Z_0 if the true value of the signal rate is s?

Gaussian approximation for Poisson significance

For large $s + b$, $n \to x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob($x > x_{obs} \mid s = 0$),:

$$p_0 = 1 - \Phi\left(\frac{x_{\rm obs} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\rm obs} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\text{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for Poisson significance

Likelihood function for parameter *s* is

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

or equivalently the log-likelihood is

$$\ln L(s) = n \ln(s+b) - (s+b) - \ln n!$$

Find the maximum by setting $$\frac{\partial \ln L}{\partial s} = 0$$

gives the estimator for *s*: $$\hat{s} = n - b$$

Approximate Poisson significance (continued)

The likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left(n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, \ 0 \text{ otherwise}$$

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z_0 \approx \sqrt{q_0} = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b, \ 0 \text{ otherwise}$$

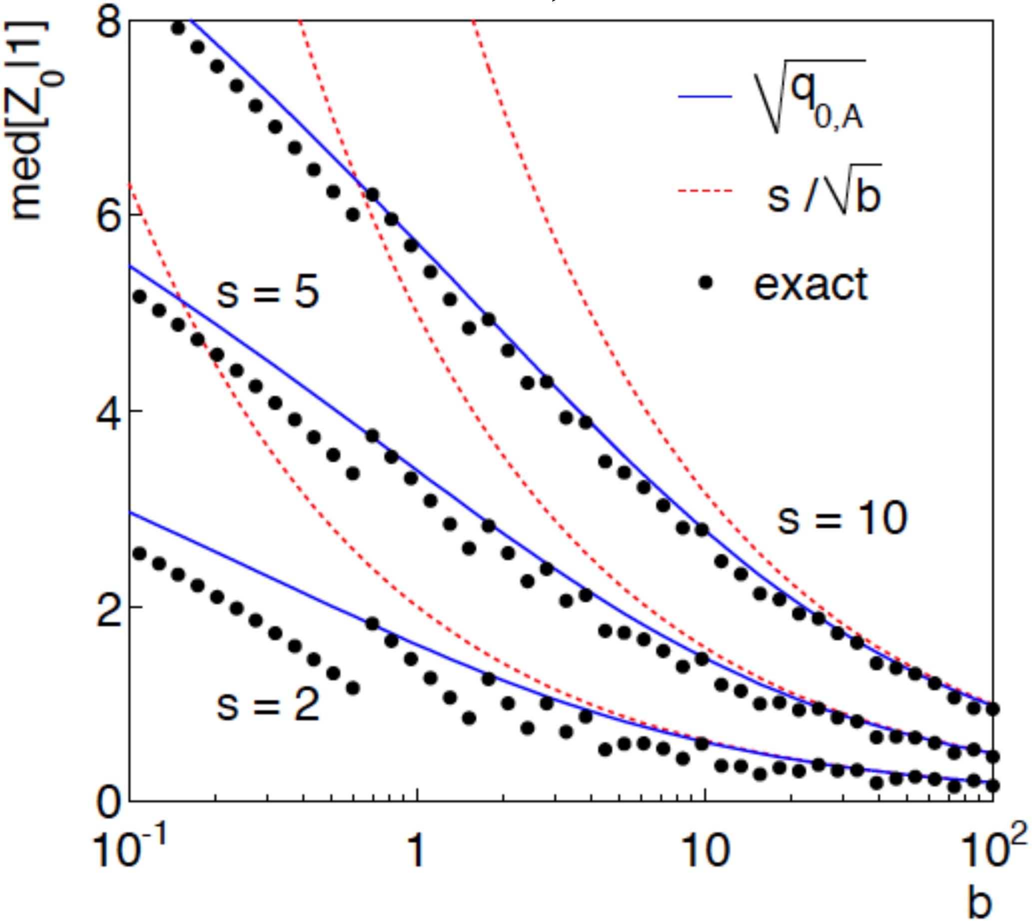
To find median[$Z_0|s+b$], let $n \to s + b$ (i.e., the Asimov data set):

$$\text{median}[Z_0|s+b] \approx \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}$$

This reduces to $s/\sqrt{b}$ for s << b.

$n \sim$ Poisson($\mu s+b$), median significance, assuming $\mu = 1$, of the hypothesis $\mu = 0$

CCGV, arXiv:1007.1727



"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,\mathrm{A}}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.