Simple Analysis Optimization

Alex Read
University of Oslo
26 Janary, 2006

Abstract: A frequently asked question concerning both measurements and searches in particle physics is how to optimize the event selection and the analysis of the selected events. It is well-known by now that using likelihoods for measurements (parameter estimation) or searches (hypothesis testing) increases the sensitivity of the analysis. What is not so well-known is how, apart from brute-force methods, to optimize a measurement or search except in the case of a simple counting experiment where the criteria S/sqrt(B) and S/sqrt(S+B) are often used. A derivation will be shown of a general formula (but still fairly simple) for binned pdf's which reproduces the criteria given above for the limiting case of 1 bin (a counting experiment).

Outline

- Event selection
- Likelihood and likelihood ratio
- Significance and power
- Optimization
- Summary

Selecting events

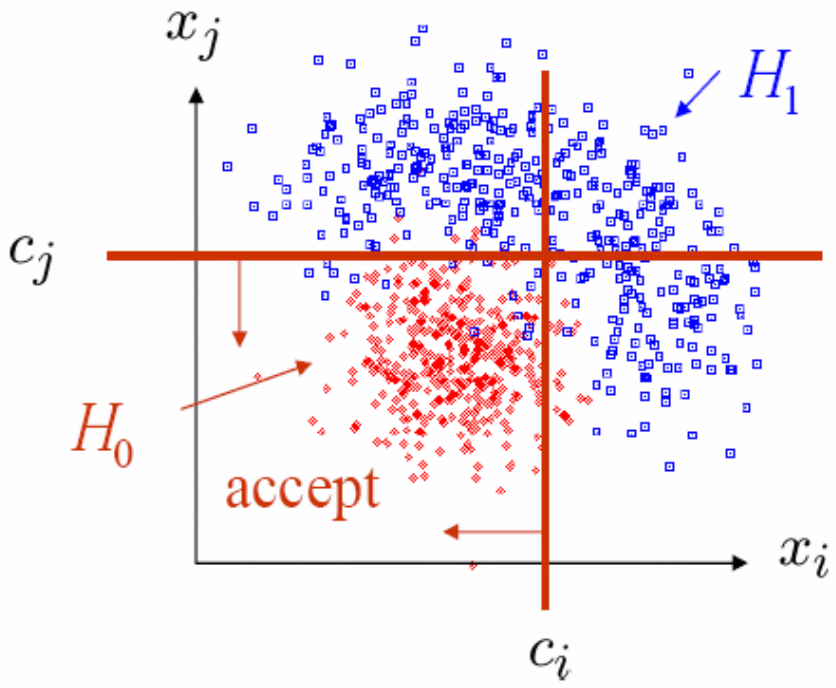
Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_0.

Each event is a point in $\vec{x}$ space. What decision boundary should we use to accept/reject events as belonging to event type H_0?

Probably start with cuts:

$$x_i < c_i$$

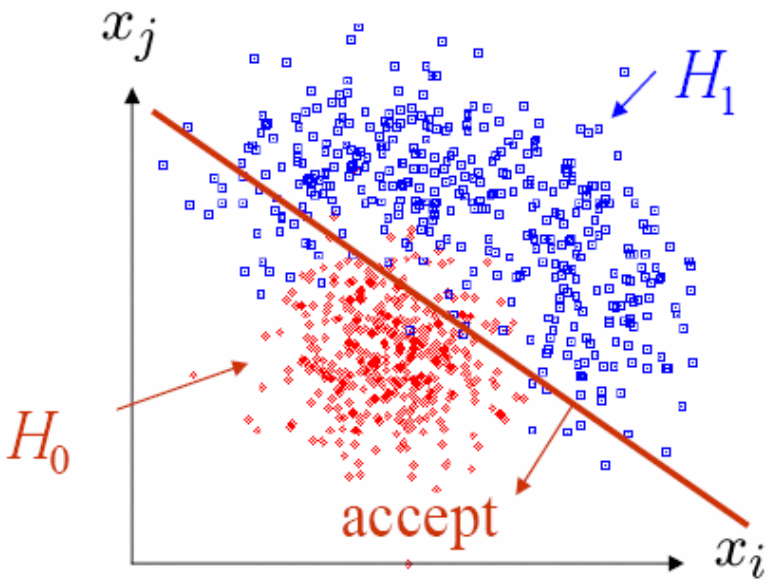
$$x_j < c_j$$



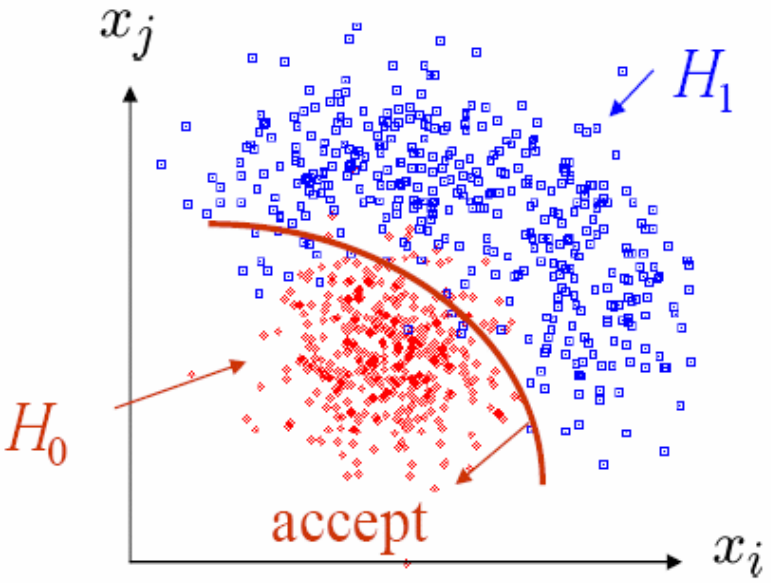
Other ways to select events

Or maybe use some other sort of decision boundary:

linear



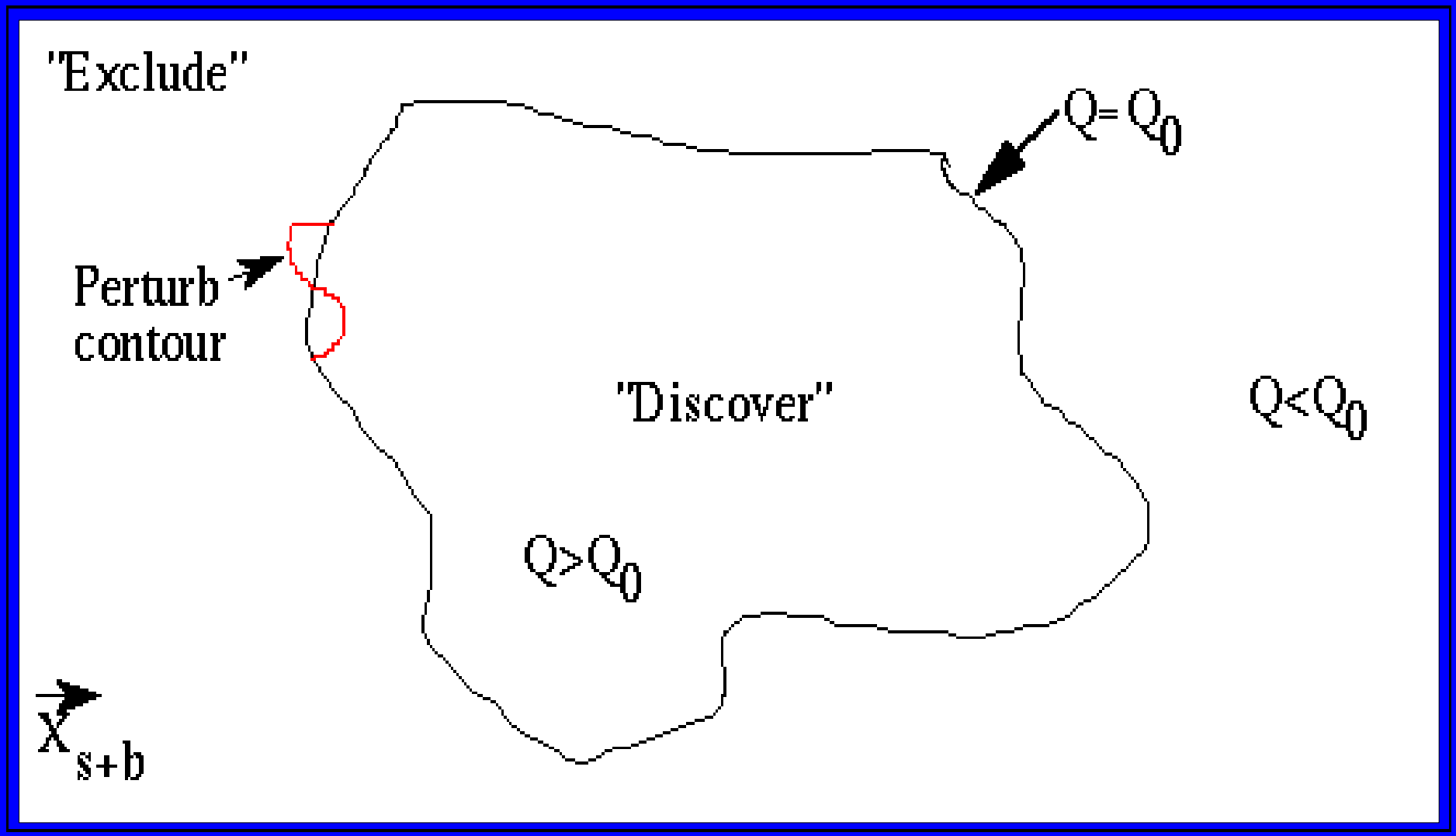
or nonlinear



How can we do this in an 'optimal' way?

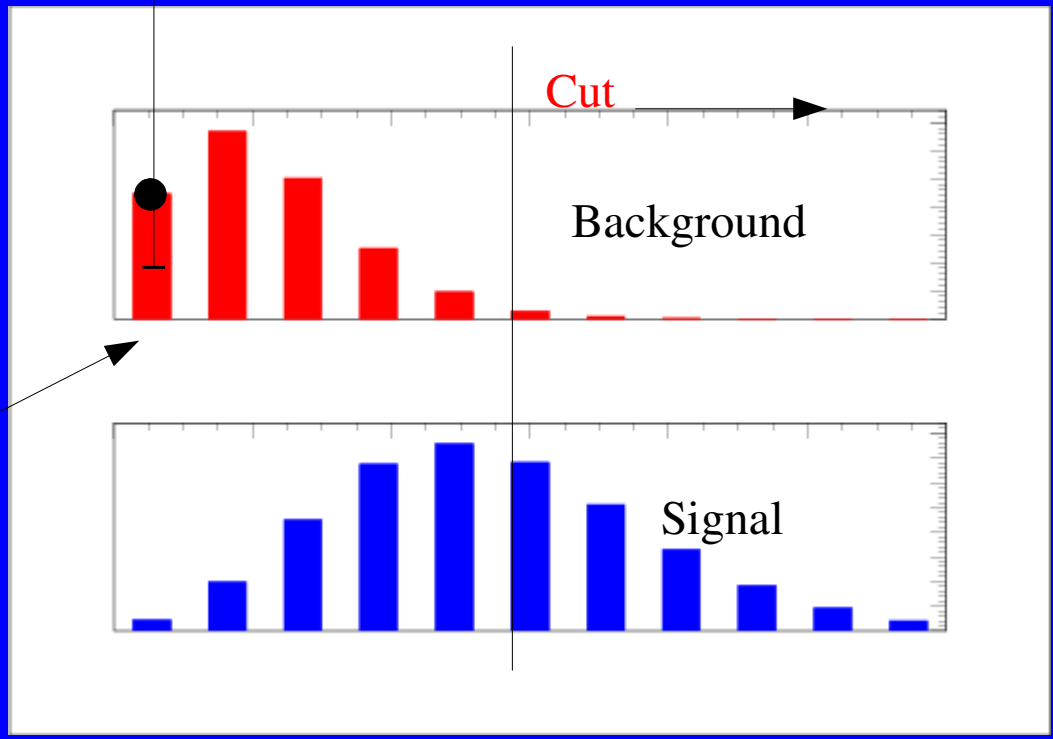
Likelihood ratio
Q=L(H0)/L(H1)
or approximation in case of complexity

Neyman-Pearson lemma



Event selection

- Remove background
  - which contains ~no signal
  - which is highly uncertain (sometimes)
- Usually better not to make the last cut
  - loss of sensitivity
  - single events around the cut unpleasant



Example: ttH in ATLAS

http://documents.cern.ch/cgi-bin/setlink?base=atlnot&categ=Note&id=phys-2003-024

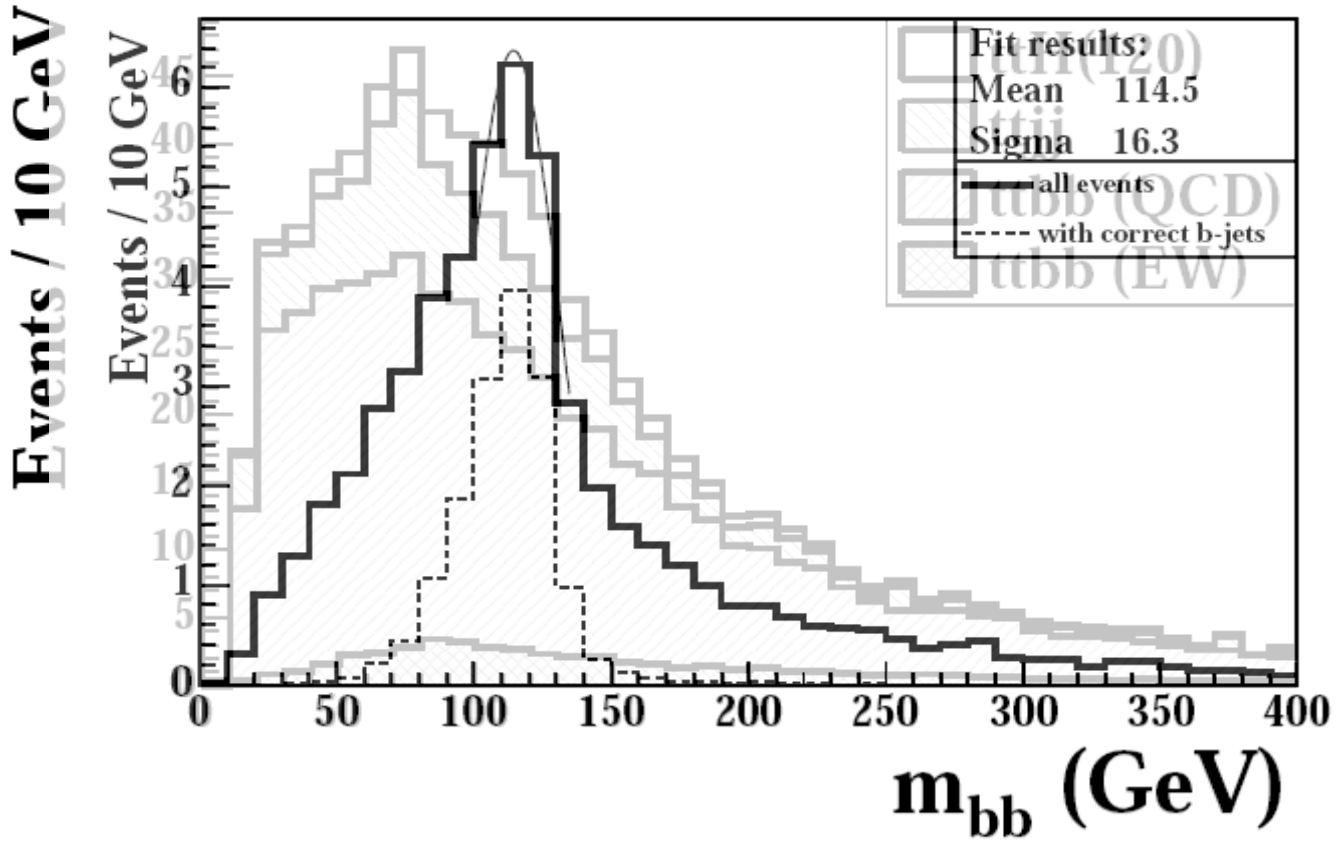


Figure 12: The sum of the reconstructed m_{bb} spectra for signal and background normalized to the rate of expected events for $\mathcal{L} = 30$ fb^{-1}.

Likelihood ratio for binned counts

$$Q_i = \frac{\dfrac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i^{cand}}}{n_i^{cand}!}}{\dfrac{e^{-b_i} b_i^{n_i^{cand}}}{n_i^{cand}!}}$$

- Likelihood ratio for bin i
  - Poisson statistics

$$-2\,lnQ_i = 2\ s_i - 2\ n_i \ln\left(1+\frac{s_i}{b_i}\right)$$

- ...in useful form

$$Q = \prod Q_i$$

- Likelihood ratio for entire distribution (or combination of channels, experiments, etc)

$$-2\ln Q = \sum -2\ln Q_i$$

- ...in useful form

Means and RMS's of -2lnQ$_i$

$$w_i = \ln\left(1+\frac{s_i}{b_i}\right)$$

- LR weight (optimal)

$$\langle -2\ \ln Q_i \rangle_{s+b} = 2\ s_i - 2\ (s_i + b_i)\, w_i$$

- Mean -2lnQ for "s+b" expts.

$$\langle -2\ \ln Q_i \rangle_{b} = 2\ s_i - 2\ (b_i)\, w_i$$

- Mean -2lnQ for "b" expts.

$$\sigma^2_{n_i} = n_i$$

- RMS for Poisson pdf.

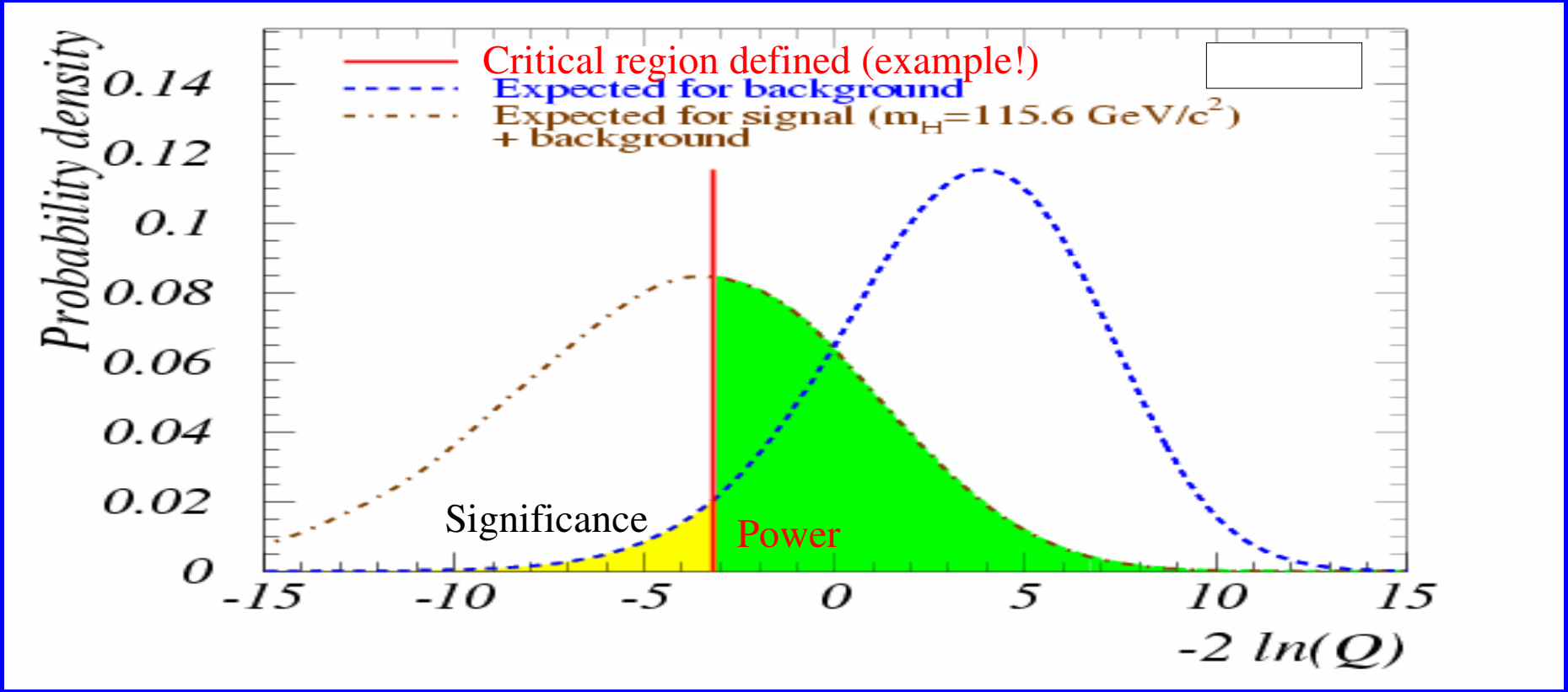
$$\sigma^2_{s_i+b_i} = 4(s_i + b_i)\, w_i^2$$

- RMS for -2lnQ for "s+b"

$$\sigma^2_{s_i+b_i} = 4(b_i)\, w_i^2$$

- RMS of -2lnQ for "b"

Distributions of -2lnQ



- These are pdf's of -2lnQ for 2 hypotheses
  - "s+b" or signal plus background
  - "b" or background-only

Optimize significance of signal

$$\lambda = \frac{\langle -2 \ln Q \rangle_b - \langle -2 \ln Q \rangle_{s+b}}{\sigma_b}$$

$$\lambda = \frac{\sum 2(s_i - b_i w_i) - 2(s_i - (s_i + b_i) w_i)}{\sqrt{\sum 4\, b_i w_i^2}}$$

$$\lambda = \frac{\sum s_i w_i}{\sqrt{\sum b_i w_i^2}}$$

$$\lambda = \frac{S w}{\sqrt{B w}} = \frac{S}{\sqrt{B}}$$

- Number of background sigmas between average signal and average background experiment
- For a single bin (counting)

Optimize power of analysis

$$\kappa = \frac{\langle -2 \ln Q \rangle_b - N\sigma_b - \langle -2 \ln Q \rangle_{s+b}}{\sigma_{s+b}}$$

$$\kappa = \frac{\sum s_i w_i - N\sqrt{\sum b_i w_i^2}}{\sqrt{\sum (s_i + b_i) w_i^2}}$$

$$\kappa = \frac{S - N\sqrt{B}}{\sqrt{(S+B)}}$$

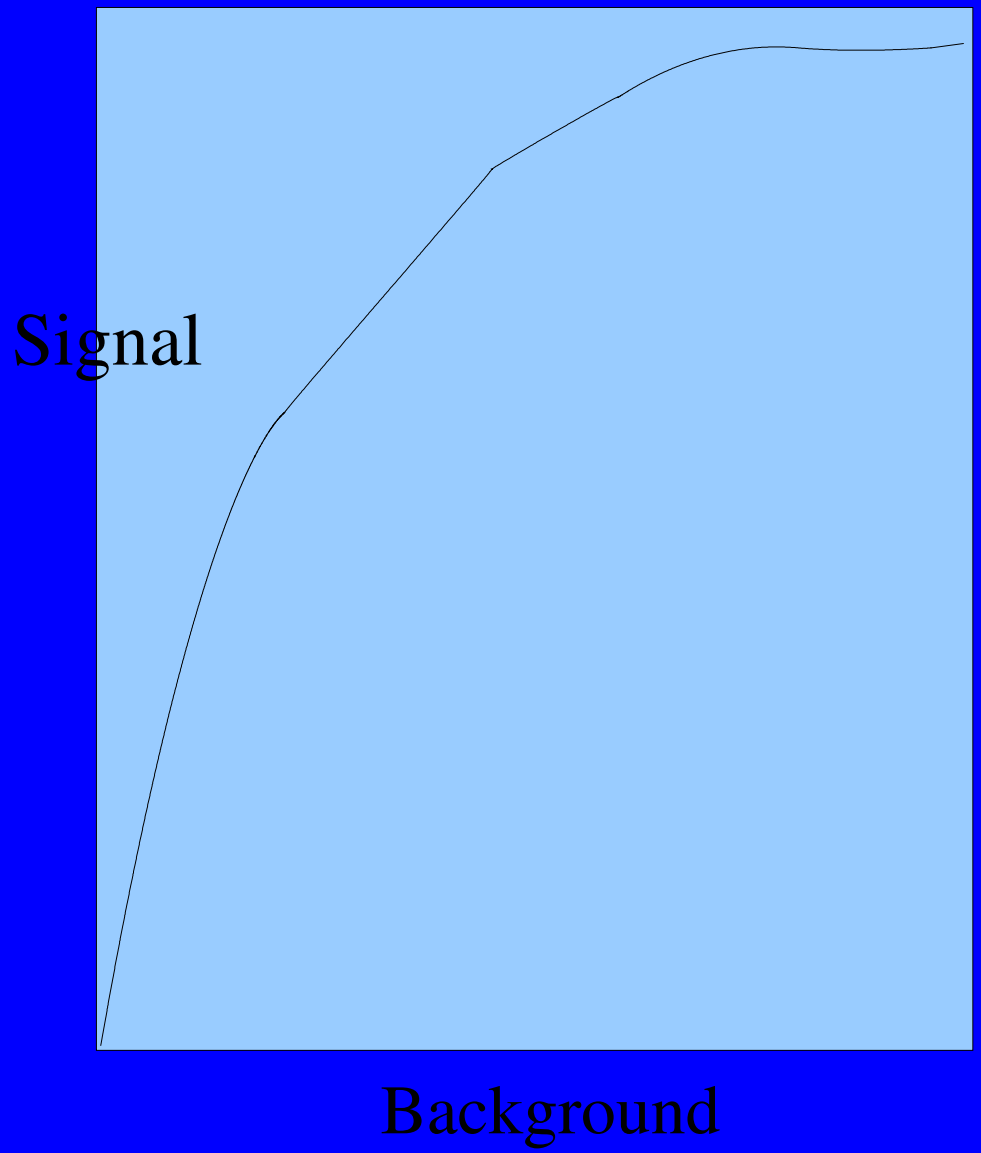
$$\kappa = \frac{S}{\sqrt{(S+B)}}$$

- Number of signal + background sigmas included in critical region (discovery)
- For a single bin (counting)
- For strong signals (measurement)

High statistics

- Correspondence of integrals in tails to Gaussian or normal distribution depends on statistics
  - If not > ~10 events in every bins the correspondence may be poor – and more sophisticated or heavy methods needed
    - alrmc (alrmc++) developed for LEP/DELPHI, but heavy to use
  - For high statistics (like the ttH example) we know the tails quite well and can use tables or lookup functions.

Signal efficiency vs. background



- Efficiency vs. background depends on event selection!

Optimizing

- Since both shapes and rates may depend on event selection there **is** an optimal working point
- The optimizations (exclusion, observation, discovery, measurement) are not identical
  - If not similar consider 2-d likelihood functions and don't make the previous cut

Summary

- Event selection and analysis optimization is a nontrivial exercise but standard statistics practice gives us solid guidelines
- Formulae for significance of average expectations $S/\sqrt{B}$ and power of search or measurement $S/\sqrt{(S+B)}$ are easily generalized for binned likelihood ratio
  - in this case and for high statistics it is relatively easy to make statements about significance and power