

Limit determination

Physics at the Terascale RT1



Miniworkshop on statistical tools



19th June 2008
DESY, Hamburg

Andre Holzner CERN/PH

Outline

- Hypothesis testing
- The questions to ask
- Type I and type II errors
- The method used by the LEP Higgs working group
 - The likelihood ratio and why it is optimal
 - Confidence levels
 - Frequentist calculation of confidence levels
 - CL calculation illustration
 - Interpretation of individual events (candidate weights)
 - Limits on parameters
 - The CL_s method
- Beyond LEP
 - Example: Tevatron Higgs
 - Limit calculations in ROOT
- Summary

Hypothesis testing

- Searches usually do a **two-hypothesis test** (point hypotheses):

H_0 : only background is present in the data

H_1 : signal+background is present in the data

- These are point hypotheses (**not depending on parameters** which could be fitted from the data)
- Note that other hypotheses like:
 - my background description is wrong
 - my detector is not performing as the simulation describes it
 - my reconstruction is not as efficient as I thought

are usually **not explicitly included** although eliminated as much as possible by the experimentalists before calculating confidence levels

The questions to ask

- We usually can easily calculate:

$$P(\text{data} | B) = P(\text{data} | H_0)$$

$$P(\text{data} | S + B) = P(\text{data} | H_1)$$

i.e. the probability that the data originates from a background only or from signal + background processes. However, we usually would like to know:

$$P(S + B | \text{data}) = P(H_1 | \text{data})$$

i.e. the probability that the signal is present in the data

- Note: $P(H_1 | \text{data})$ and $P(\text{data} | H_1)$ are **NOT THE SAME !**

The questions to ask

- The answer lies in Bayes' Theorem:

$$P(H_1 | \text{data}) = P(\text{data} | H_1) \cdot \frac{P(H_1)}{P(\text{data})}$$

- $P(H_1)$ is called **prior probability** (before looking at an experiment's data) for the hypothesis H_1
- The choice of $P(H_1)$ is not unique:
 - Can include information from previous experiments
 - Could be chosen as flat
 - Can be used to exclude non-physical regions (e.g. zero for theories with negative masses)
 - Influences $P(H_1 | \text{data})$!

Type I and type II errors

- In practice, we're faced with four possible situations:

<div style="text-align: center;">Actual situation</div> <div style="text-align: center;">Our conclusion</div>	Background only present in the data	Signal + background present in the data
Claim background only (accept H_0)	(True) exclusion of signal	False exclusion of signal / missed discovery Type I error (α) Typically 5% (2σ)
Claim background + signal (reject H_0 , accept H_1)	False discovery Type II error (β) Typically $5.7 \cdot 10^{-7}$ (5σ)	(True) discovery

The likelihood ratio and why it is optimal

Neyman-Pearson lemma:

- Given:
 - two point hypotheses H_0 and H_1 (no free parameters !)
 - the (fixed) probability α of a type I error α
- Neyman-Pearson states that
 - The most powerful (i.e. with minimal type II error) test is a **likelihood ratio** test:

$$Q = \frac{L(d | H_1)}{L(d | H_0)} > \eta \rightarrow \text{reject } H_0 \text{ (and accept } H_1)$$

where η is determined by the choice of α

False discovery rate

(no free parameters !)

type I error α

Minimal missed discovery rate

minimal type II error

The likelihood ratio and why it is optimal

- Likelihood ratio:

$$Q = \frac{L(d | h_1)}{L(d | h_0)} = \frac{L(d | S + B)}{L(d | B)}$$

where the likelihood is the product of the probabilities of all observations.

For binned distributions, the likelihood is (for background):

$$L(d | B) = \prod_{i \in \text{bins}} \frac{b_i^{d_i} e^{-b_i}}{d_i!}$$

i.e. the product of the Poisson probabilities of observing d_i events when b_i events are expected

The product runs over all bins, channels, experiments etc.

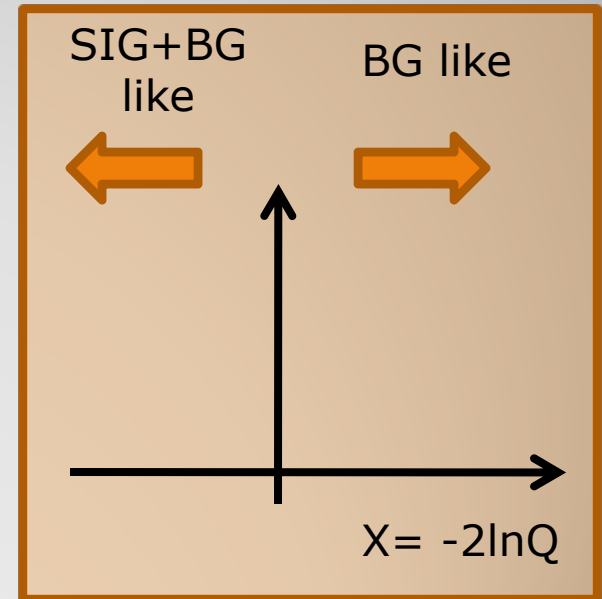
The likelihood ratio and why it is optimal

- Often, the log of the likelihood ratio is used: $X = -2 \ln Q$

side effect: X becomes a $\Delta\chi^2$ in the (Gaussian) limit of large statistics

- Properties of Q and X :

Type of experimental outcome	Values of Q	Values of $X = -2 \ln Q$
Very background like	Much smaller than one	Very positive
Very signal+background like	Much larger than one	Very negative



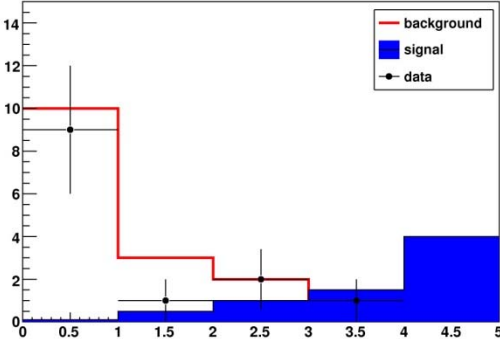
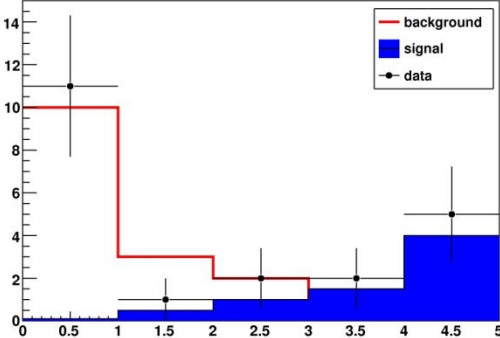
The likelihood ratio and why it is optimal

- Example:
 - Counting experiment:
 - Expected background: 100 events
 - Expected signal: 50 events

Number of observed events	Likelihood ratio Q	$X = -2\ln Q$
100	$1.63 \cdot 10^8$	18.9
150	$4.00 \cdot 10^{-10}$	-21.6
80	$1.80 \cdot 10^{15}$	35.1
170	$3.62 \cdot 10^{-17}$	-37.9

The likelihood ratio and why it is optimal

- Example: (background: 16.5 events, signal: 7.1 events)
 - Binned distribution (Q from individual bins combined)

observed	Observed Distribution	Likelihood ratio Q	X = -2ln Q
13		$5.92 \cdot 10^{-3}$	10.3
21		$8.92 \cdot 10^2$	-13.6

Confidence levels

- We now have an optimal ordering rule (i.e. according to the likelihood ratio) of experimental outcomes
- We now can ask the question:

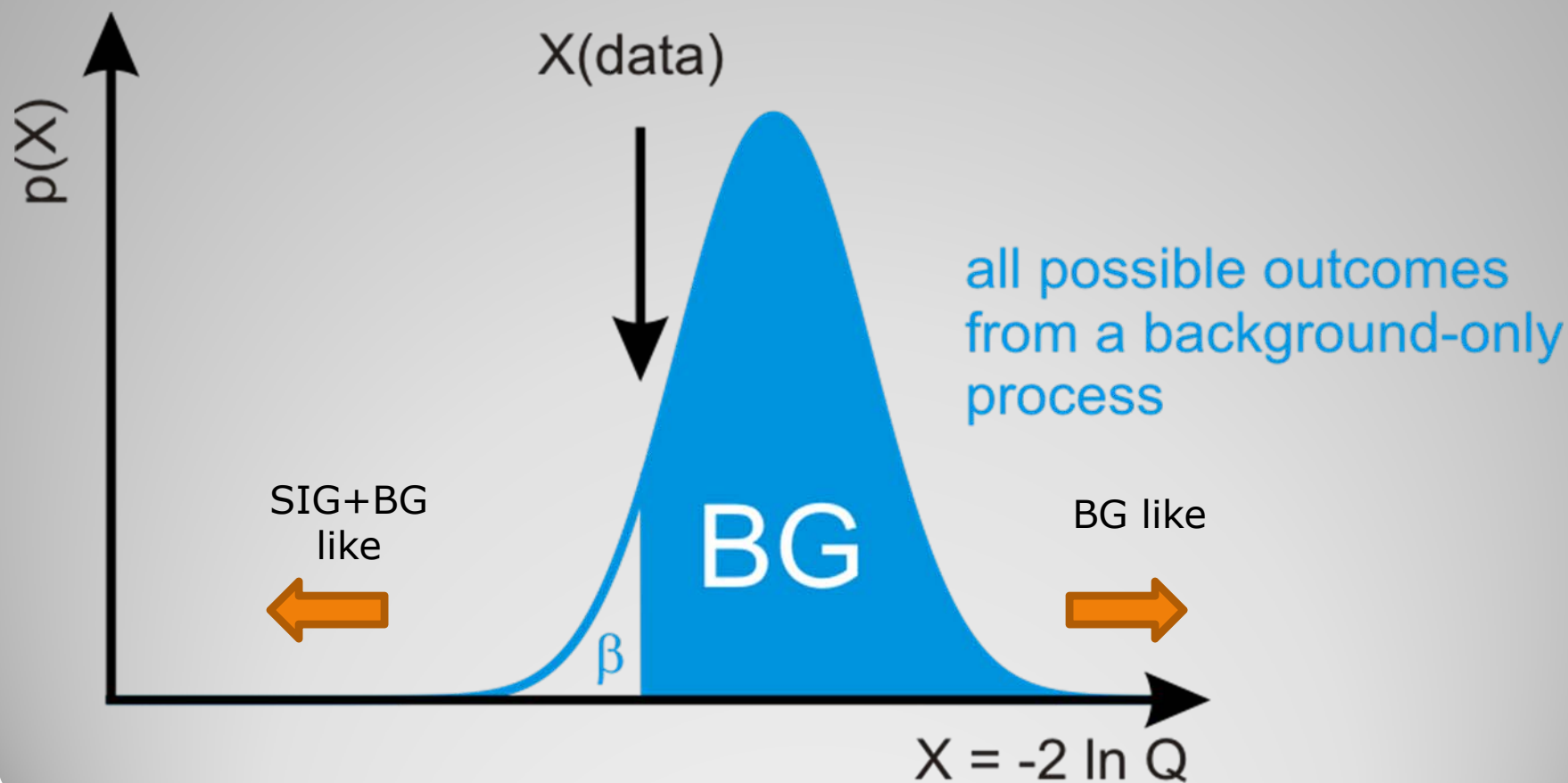
How significant is the observation ?

Or more concrete:

What is the probability that a **background only** process generates a fluctuation that is **more signal+background** like than the data ?

Confidence levels

What is the probability that a **background only** process generates a fluctuation that is **more signal+background like** than the data ?



Confidence levels

- Terminology used by the LEP Higgs WG:

CL_b : probability that a background only experiment yields an outcome which is **as S+B like or less** S+B like as the data

$1-CL_b$: probability that a background only experiment is **more S+B like** than the data

In terms of counting experiments: probability that one observes **more events than the data** in a background only experiment

$1-CL_b = 0.5$ for the median background outcome

$1-CL_b$ **small for signal+background like outcomes**
e.g. $1-CL_b = 5.7 \cdot 10^{-7}$ corresponds to a significance of 5σ

$1-CL_b$ is **used to discover signals**

Confidence levels

- Terminology used by the LEP Higgs WG:

CL_{s+b} : probability that a signal+background experiment yields an outcome which is **as B like or less** B like as the data

In terms of counting experiments: probability that one observes the number of events seen in the data **or less than that** in a signal+background experiment

(how often does signal+background **underfluctuate** such that it looks like my data ?)

$CL_{s+b} < 5\%$ → it's quite unlikely that signal+background produces **so few events** as I saw in the data

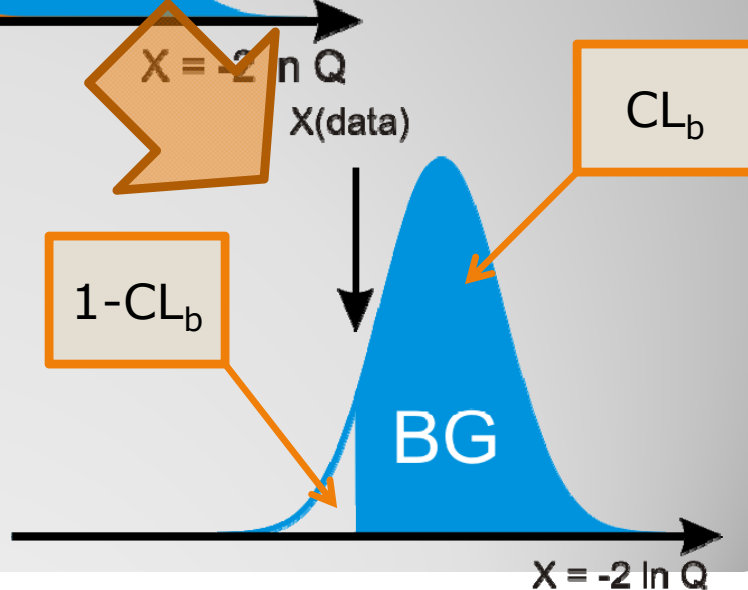
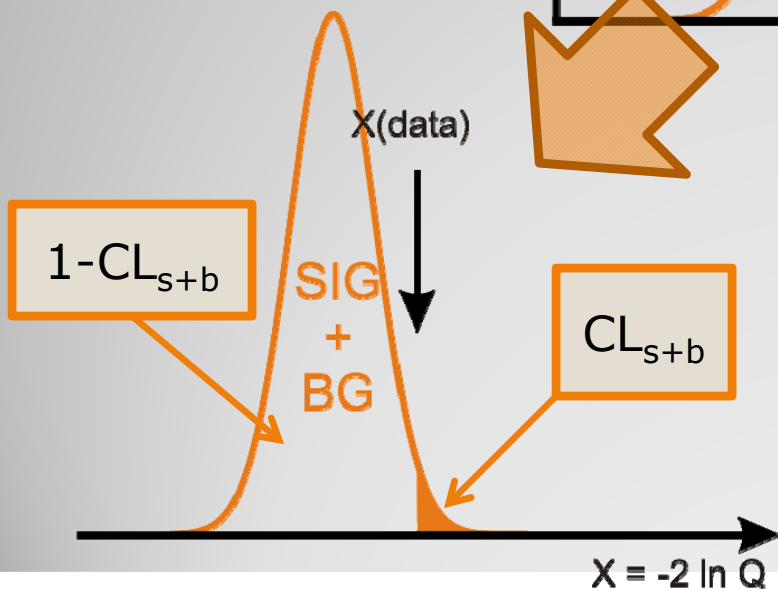
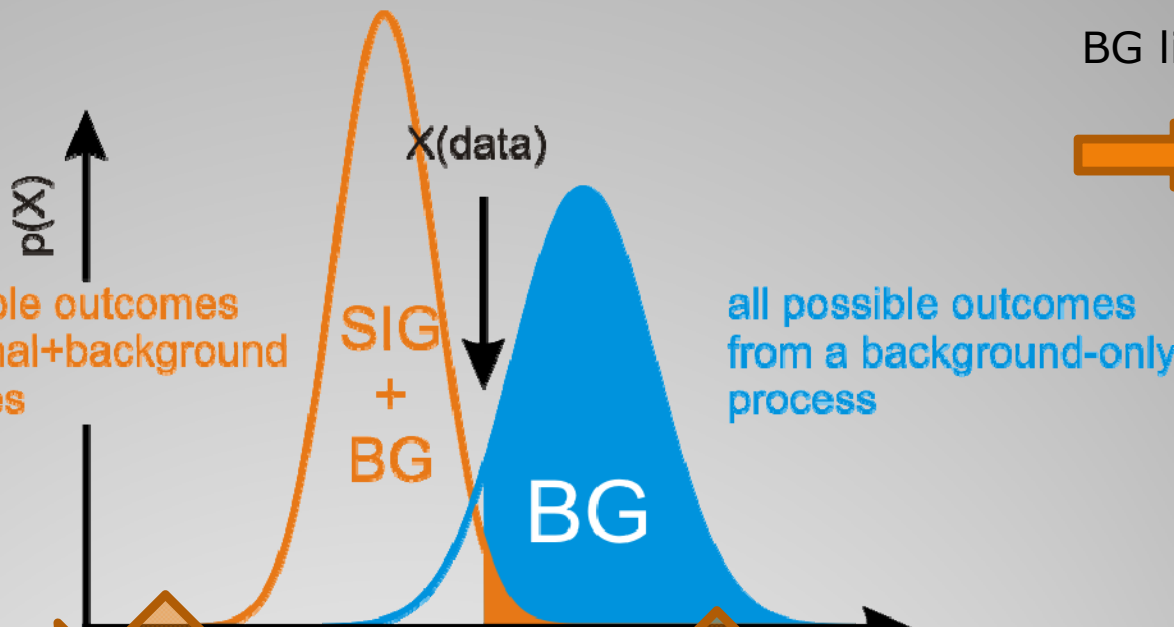
CL_{s+b} can be used for exclusion of a signal (e.g. when $CL_{s+b} < 5\%$)

Confidence levels

SIG+BG
like



BG like



Type I and type II errors

- In practice, we're faced with four possible situations:

<div style="text-align: center;">Actual situation</div>	<div style="text-align: center;">Background only present in the data</div>	<div style="text-align: center;">Signal + background present in the data</div>
<div style="text-align: center;">Our conclusion</div>	<div style="text-align: center;">(True) exclusion of signal</div>	<div style="text-align: center;">False exclusion of signal / missed discovery</div> <div style="text-align: center;">CL_{s+b}</div>
<div style="text-align: center;">Claim background only (accept H_0)</div>	<div style="text-align: center;">False discovery</div> <div style="text-align: center;">$1-CL_b$</div>	<div style="text-align: center;">(True) discovery</div>

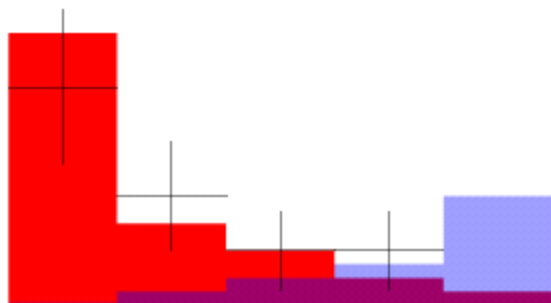
Frequentist calculation of confidence levels

- How do we get the background and signal+background distributions of $-2 \ln Q$?
- Remember Frequentist's definition of probability:

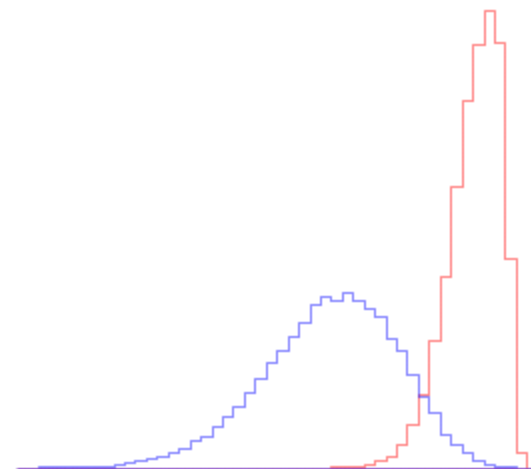
Probability = relative frequency (of an outcome) in a large number of trials.

- With sufficient amount of CPU power:
 - simulate a large number of experimental outcomes (throw Poisson numbers for each bin of background and signal+background expectation)
 - Include **systematic uncertainties** by varying e.g. the expected background **before generating each trial**
 - arbitrarily complex **correlations** can be done
 - treat each trial exactly the **same way the data is treated**
 - E.g. **fit background** from data in some channels etc.
 - **Count** the number of simulated outcomes which satisfy the criterion for which one wants to calculate the probability

CL calculation illustration



- next s+b trial
- next b trial
- next 1000 s+b trials
- next 1000 b trials
- remaining s+b trials
- remaining b trials
- Reset



Candidate weights

- One can rewrite the log likelihood ratio in the following way:

$$X = -2 \ln Q = -2 \sum_{i \in \text{bins}} \left(-s_i + d_i \cdot \ln \left(1 + \frac{s_i}{b_i} \right) \right)$$

Signal to background
ratio in bin i

Number of data
events bin i

- Scan/visualize **most significant events** (highest local s/b) and show them at conferences !
- WARNING:** People get attached to these events and will remember them after you have modified your detector simulation/physics simulation/analysis !!

Limits on parameters

- So far, we had two **point hypotheses**:

H_0 : only background is present in the data

H_1 : signal+background is present in the data

- If we want to test for Higgs production, the signal efficiency and the distributions used to test the hypotheses strongly depend on the Higgs mass

→ Straightforward approach:

For each Higgs mass in question, **repeat** the hypothesis test

Can become **computationally intensive** if more than one parameter (e.g. m_h and $\tan \beta$) need to be scanned

The CL_s method

- Example (counting experiment):
 - Expected **background**: 100 events
 - Expected **signal**: 0.5 events
 - Could one discover a signal ? (Is one sensitive to it ?)
 - Clearly **not** (the signal is much smaller than the statistical uncertainty on the background alone).
 - Assume one **observes**: 80 events in data
 - Can one exclude the signal at 95% CL ?
 - **Yes**, $CL_{s+b} = 2\%$ despite no sensitivity to the signal !
→ this is clearly **not a desirable feature**

The CL_s method

A.Read, [J.Phys.G28:2693,2002](#)

- 'Underfluctuations' can e.g. come from:
 - Underestimation of the detector efficiency (detector simulation too optimistic)
 - Choosing cuts (a posteriori) in order to remove 'unwanted events'
- To protect against this, do the following:

for counting experiments (n events observed):

Count fraction of signal+background experiments with less than n but **consider only those** signal+background experiments where the **contribution from background is less than n** .

more generally: use $CL_s = \frac{CL_{s+b}}{CL_b} < 5\%$ instead of $CL_{s+b} < 5\%$

(in the previous slide: $CL_{s+b}=2\%$, $CL_s=89\%$)

- Note: This gives **more conservative** limits !

Tevatron Higgs searches

- Combined CDF and DØ Upper Limits on Standard-Model Higgs-Boson Production (April 10, 2008 / "Winter 2008 Combination prepared for hep-ex.", with $L=1.0-2.4 \text{ fb}^{-1}$), [arXiv:0804.3423](https://arxiv.org/abs/0804.3423):
 - Uses two methods (giving the same results within 10%):
 - Modified CL_s method to include e.g. fitting the background from data
 - Bayesian method: Integrate

$$L(\mathbf{R}, \vec{s}, \vec{b} \mid \vec{n}, \vec{\theta}) \cdot \pi(\vec{\theta}) = \prod_{i \in \text{channels}} \prod_{j \in \text{bins}} \frac{\mu_{ij}^{n_{ij}} e^{-\mu_{ij}}}{n_{ij}!} \cdot \prod_{k \in \text{np}} e^{-\theta_k^2 / 2}$$

Tevatron Higgs searches

'nuisance' (unknown) parameters, typically with an uncertainty

Probability densities for nuisance parameters

Signal scaling factor

$$\mu_{ij}^{n_{ij}} = R \cdot s_{ij}(\vec{\theta}) + b_{ij}(\vec{\theta})$$

$$L(\mathbf{R}, \vec{s}, \vec{b} | \vec{n}, \vec{\theta}) \cdot \pi(\vec{\theta}) = \prod_{i \in \text{channels}} \prod_{j \in \text{bins}} \frac{\mu_{ij}^{n_{ij}} e^{-\mu_{ij}}}{n_{ij}!} \cdot \prod_{k \in \text{np}} e^{-\theta_k^2/2}$$

Observed data

Expected background distributions

Poisson Probability in bin j of channel i

Expected signal distributions

Tevatron Higgs searches

- Assume a flat prior for the total number of selected Higgs events
- Integrate over all parameters except the relative signal rate R and normalize to obtain the probability $p(R|\text{observed data})$
- Set the limit R_{95} on the relative signal rate R by requiring:

$$\int_0^{R_{95}} p(R | \text{observed data}) \cdot dR = 0.95$$

Limit calculations in ROOT

- TLimit

- Frequentist “with Bayesian treatment of uncertainties in nuisance parameters”
 - ☺ supports inclusion of (correlated) systematic uncertainties
 - ☺ arbitrary number of bins → combination of different channels, experiments etc. without losing sensitivity straightforward
 - ☹ might need a large number of MC trials, especially when it comes to high significances
 - ☹ Conclusion of whether a limit is derived or a signal is claimed is left to the user (does not come naturally out of the method)
- This is essentially what has been used at LEP for Higgs searches

Summary

- The method used at LEP to calculate limits on the mass of the standard model Higgs boson (and for constrained MSSM models) was presented
- At Tevatron and LHC, the situation is somewhat different due to the fact that the uncertainties on the background are much more important than at LEP

Lets hope for (large) signals
at the LHC and ILC so we
don't need to set limits !

Backup slides

Cross section limits

- For cross section limits, introduce a **cross section scaling factor** as another parameter
- Repeat confidence level calculations for several scaling factors until the exclusion condition (e.g. $CL_{s+b} = 5\%$) is reached

The profile likelihood method

- Rolke et. al. studied the profile likelihood method for limit and two sided intervals ('errors') for an experiment with:
 - a signal dominated bin to which background contributes
 - A background only bin
 - Uncertainty on the signal efficiency
- Idea: Given

Fit the nuisance parameters θ , leave model parameters π free

$$\lambda(\pi_0 | \text{data}) = \frac{\sup_{\theta} \{\text{Likelihood}(\pi_0, \theta | \text{data})\}}{\sup_{\theta, \pi} \{\text{Likelihood}(\pi, \theta | \text{data})\}}$$

Overall best fit of nuisance parameters θ , and model parameters π

The profile likelihood method

- $-2 \log \lambda$ has approximately a χ^2 distribution
- Therefore we can look for π_0^{\min} which gives the minimal

$$L_{\min} = -2 \log \lambda(\pi_0^{\min})$$

and set the interval boundaries / limits where $L = L_{\min} + c$

- For certain forms of signal, background and efficiency, one can get analytical results
- Some care is needed for special cases (e.g. number of observed events in signal bin is less than the number of expected background events)
- In their paper, they have found good coverage for the method

Limit calculations in ROOT

- TRolke

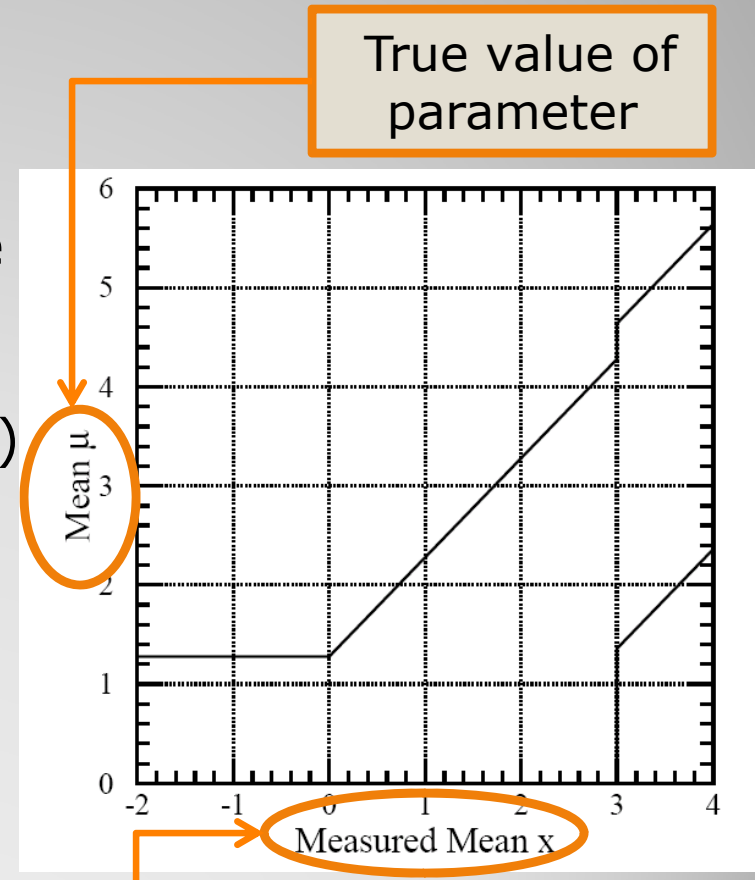
- Based on profile likelihood method, fully frequentist
 - ☺ includes uncertainties in nuisance parameters
 - ☺ seems **not** to generate MC trials (→ fast) despite frequentist method
 - ☹ Only for a (limited) scenario of a signal and a background (counting) region

- TFeldmanCousins

- Fully frequentist construction
 - ☺ solves the problem of undercoverage due to flip-flops between exclusion and measurement
 - ☹ does not handle uncertainties in nuisance parameters (e.g. background rate)

The problem of flip-flopping

- When does one go from limit setting (one-sided intervals on parameters) to two sided intervals (measurement of a parameter) ?
- Example policy:
 - If the result x is less than 3, I will state an upper limit from the standard tables. If the result is greater than 3, I will state a central confidence interval (90%) from the standard tables."



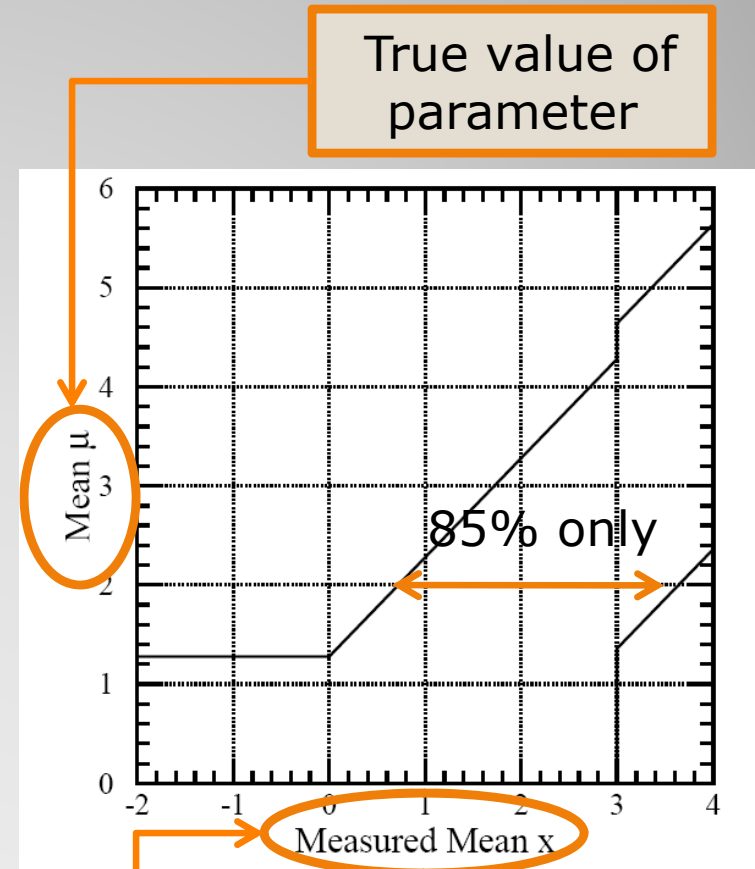
The problem of flip-flopping

- Problem:

If true value of the parameter is $\mu=2$, then for only in 85% of experiments $\mu=2$ is in the quoted interval

i.e. $P(x |$

- This should be 90% by construction !
- This effect is called **undercoverage**



The problem of flip-flopping

- Solution (Feldman & Cousins '97):

use of an ordering principle of the possible measurements (event counts n) according to:

$$R(n) = \frac{P(n | \mu)}{P(n | \mu_{\text{best}})}$$

include n (in decreasing order of R) into interval until the sum of $P(n|\mu)$ is $\geq 90\%$

- This removes undercoverage by construction and gives a natural way when to switch from one-sided (limits) to two-sided intervals (measurement)

Links

- Some words in this presentation are linked to Wikipedia
- Confidence limits workshop at CERN:
http://preprints.cern.ch/cgi-bin/setlink?base=cernrep&categ=Yellow_Report&id=2000-005

Literature

- LEP Higgs WG: 'Searches for Higgs bosons: Preliminary combined results using LEP data collected at energies up to 202-GeV', [CERN-EP-2000-055](#), Appendix A:
 - Describes the confidence level calculation used by the LEP Higgs WG in less than three pages
- Kyle Cranmer: 'Statistical Challenges for Searches for New Physics at the LHC' (Proceedings of PhyStat2005), [arXiv:physics/0511028](#):
 - Review of methods used in the past and for the LHC

Glossary

- Type I error (α):
 - rejecting a null hypothesis (e.g. only background present in the data) when it is actually true
 - `False positive', `False discovery'
- Type II error (β):
 - failing to reject a null hypothesis (e.g. only background present in the data) when the alternative hypothesis is true
 - `false negative', `False exclusion of new physics'
- Power of a hypothesis test:
 - probability that a test will reject a false null hypothesis
 - $1 - \text{probability of type II error} = 1 - \beta$
 - is 100% in the ideal case (i.e. a type II error can not happen)
- Coverage of a hypothesis test:
 - probability that a test will accept a true alternative hypothesis (????)
 - $1 - \text{probability of type I error} = 1 - \alpha$
- Undercoverage:
 - Coverage is less than the method claims it to be, i.e. type I error (missed discovery rate) larger than claimed (????). Example: The limit at 95% confidence limit is in fact only a 93% confidence limit.

Glossary

- Nuisance parameter:
 - A parameter which is not of immediate interest but must nevertheless be accounted for
 - e.g. amount of background estimated from the data
- Test statistic:
 - A function which summarizes the outcome of an experiment (typically in a single real number).
Typically used to order outcomes of (monte carlo) experiments
Example: A function which gives
 - very negative values if data signal+background like,
 - very positive values if data is background like
- P-value:
 - probability of obtaining a value of the test statistic at least as extreme as the one that was actually observed, given that the null hypothesis (background only) is true
(‘how often would I get a deviation from the expected background larger than the one I saw in my data ?’)
 - Examples: CL_{s+b} , CL_b , χ^2 upper tail probability

Glossary

- Marginalization of a parameter:
 - Integrate (a conditional probability) over this parameter (i.e. consider all possible choices for this parameter and their respective probability)