Statististical methods

Oldřich Kepka
Institute of Physics, Prague

April 19, 2017
Výjezdní seminář MFF, Malá Skála

Outline

- Hypothesis testing framework
- Application to discovery, limits, confidence intervals
- How to read Higgs search plots

Likelihood function

- Suppose the result of the experiment given by $\vec{x}$ of numbers
- Joint pdf for the result given

$$f(\vec{x}, \theta)$$

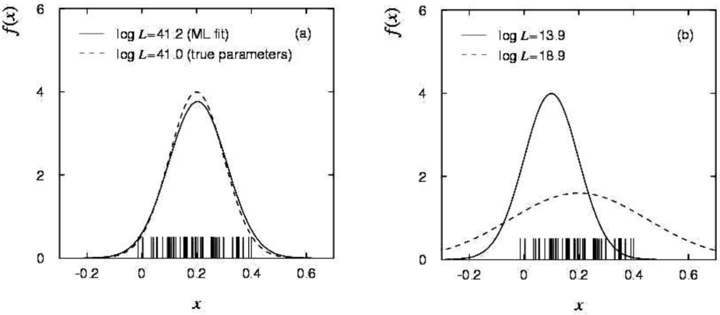
- Consider this expression as a function θ for fixed measured values $\vec{x}$.
- This is the Likelihood function

$$L(\vec{x}, \theta) = f(\vec{x}, \theta)$$

- n independent observation $\rightarrow$ product of probabilities

$$L(\vec{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta) \qquad x_i \text{ constant}$$

Maximum likelihood estimator



- Observation: Likelihood function large if θ close to the true value
- $\rightarrow$ An estimator $\hat{\theta}$ of the true parameter θ is obtained by finding $\hat{\theta}$ which maximizes the likelihood

$$\left.\frac{\partial L}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

Hypothesis tests

- We want to make a decision given observed data
- Compatibility of assumed model (discovery), compatibility of a model parameter (limits or confidence intervals) with data
- Place a cut event-by-event to distinguish signal/background

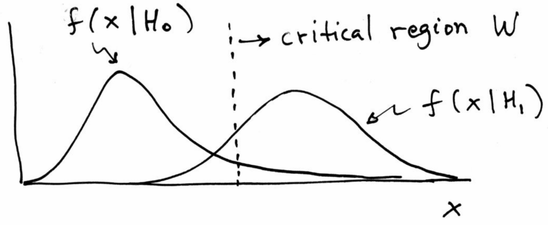
- Hypothesis H defines probability to observe data $\vec{x}$, $f(\vec{x}|H)$ (likelihood)
  - $\vec{x}$ e.g. single particle, single event, entire experiment
  - All possible values of $\vec{x}$ define sample space Ω
- Simple (point) hypothesis - $f(\vec{x}|H)$ completely specified
- Composite hypothesis - $f(\vec{x}|H)$ contains one or more unspecified parameters

Hypothesis test (frequentistic approach)

- Suppose null hypothesis H_0 and alternative hypothesis H_1
- The test is defined by a specific choice of a crictical region K = part of the data space where events fall with a probability α if H_0 is valid

$$P(x \in K|H_0) \leq \alpha$$

- α called size of test, typically a small number
- Define the critical region before taking data
- Carry experiment
  - If $\vec{x}$ falls in K $\rightarrow$ rejects hypothesis H_0



Errors and Power of the Test

Type I Error

- Probability to reject the hypothesis, H_0, even if true (no larger than size of the test)

$$P(x \in K | H_0) \leq \alpha$$

Type II Error

- Probability to accept the hypothesis H_0, when alternative H_1 is true

$$P(x \notin K | H_1) = \beta$$

Test Power

- Power of the test with respect to the alternative H_1

$$\mathrm{Power} = 1 - \beta$$

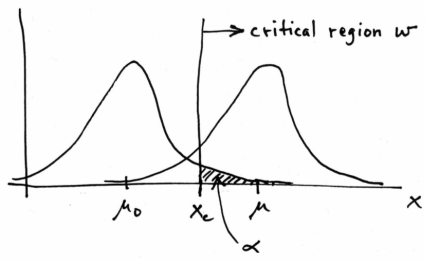
How to choose a critical region?

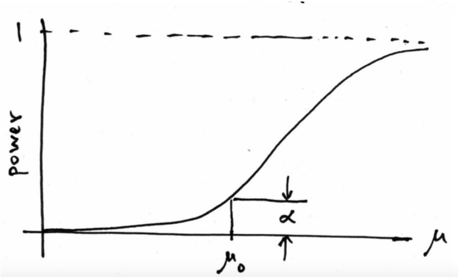
- The choice of critical region will depend on the relevant alternatives H_1
- Want to maximize power with respect to $H_1 \rightarrow$ reject H_0 if H_1 is true

- Often such a test has high power not only to one specific simple hypothesis, but also wrt. to a class of hypothesis

- Example: a measurement of $x \sim N(\mu, \sigma)$ (σ known)
  - H_0: $\mu = \mu_0$
  - versus H_1: $\mu > \mu_0$ (composite hyp.)
  - The highest power with respect to $\mu > \mu_0$ is obtained by defining the critical region as $x > x_c$. The exact value is determined by the size of the test $\alpha = P(x \geq x_c | \mu_0)$

Test of $\mu = \mu_0$ vs. $\mu > \mu_0$ with $x \sim N(\mu, \sigma)$





$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$

$$x_c = \mu_0 + \sigma\Phi^{-1}(1-\alpha)$$

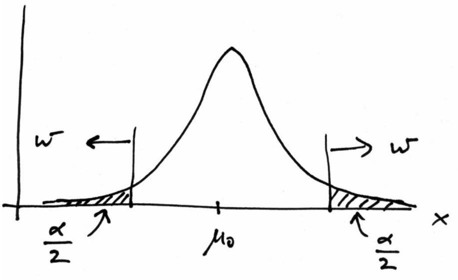
Φ - Gaussian cumulative distribution
Φ^{-1} - Gaussian quantile

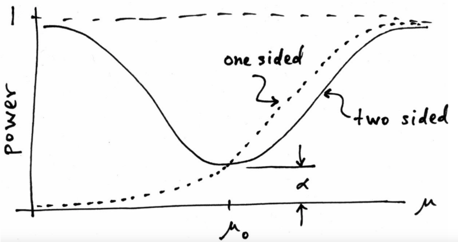
$$\text{Power} = 1 - \beta = P(x > x_c) =$$

$$1 - \Phi\left(\frac{x_c - \mu}{\sigma} + \Phi^{-1}(1-\alpha)\right)$$

- large power for large values of μ parameter

Critical region for two sided test





- We may want to construct the test to be sensitive to both $\mu > \mu_0$ and $\mu < \mu_0$
- Case for confidence interval construction

- Significant improvement for $\mu < \mu_0$
- Smaller power for $\mu > \mu_0$ than one-sided test

Illustration that generally there does not exist a test which would be most powerful with respect to any hypothesis

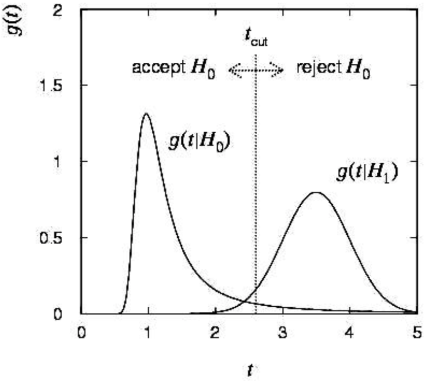
Test statistics

- The optimal critical region is often complicated selection in n-dimensional space
- Boundary of the critical region can be defined by

$$t(x_1, \ldots, x_n) = t_c$$

where the scalar function $t(x_1, \ldots, x_n)$ is called test statistic.

- Once we find out the distributions of t under null $g(t|H_0)$ and alternative $g(t|H_1)$ hypotheses, t can be used to define the test
- Reduction of the n-dimensional to 1-dimensional problem



Constructing a test statistics

Neyman-Pearson lemma:

- Allows to choose the critical region in an optimal way
- To obtain the highest power of a test of simple hypothesis H_0 wrt. to simple alternative hypothesis H_1, the critical region should be chosen such that likelihood ratio is

$$\frac{f(x|H_1)}{f(x|H_0)} > k$$

  for all $x \in K$, and less than k for x outside K. The value of k is chosen such that the test has size α.

Equivalent formulation:

- The test statistic giving highest power of the test is

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

Simple example

- Each event characterized by two variables, $\vec{x} = \{x_1, x_2\}$
- Background hypothesis (H_0)

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

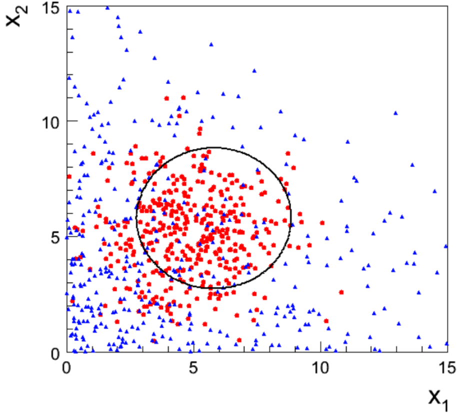
- Signal hypothesis (alternative H_1)

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

with $x_i > 0$ and normalization C.

Test statistic

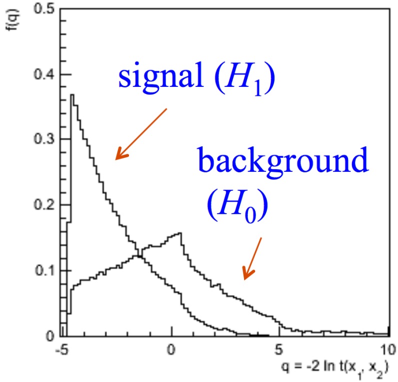
- We know the pdfs of $f(\vec{x}|H_0)$ and $f(\vec{x}|H_1)$ $\rightarrow$ can evaluate $t = \frac{f(\vec{x}|H_1)}{f(\vec{x}|H_0)}$
- In general this is not the case $\rightarrow$ multivariate techniques used to approximate LLR to get best out of data
- Contour of constant likelihood ratio defines the critical region



Event selection using LLR

- Use MC experiments to determine the distribution of t or equivalently of q

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



Generate events

- according to $H_1 \to f(q|H_1)$
- according to $H_0 \to f(q|H_0)$

- N-P Lemma implies that by placing a cut, we select the signal with highest efficiency (test power) for a given background contamination (size of a test)

Search for a signal

- Suppose that signal does not exist $\rightarrow$ search
- Hypotheses are
  - H_0: events are only background (b events)
  - H_1: events are mixture signal + background ($s + b$ events)
- Discovery: reject H_0 with large significance
- Likelihood function given H_0

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

- Likelihood function given H_1

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

- Test statistic

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

p-value

- Used to describe the level of compatibility of data with hypothesis H
- Probability, using hypothesis H, to observe data with equal or worse agreement than what we actually have seen in data

$$p_H = \int_{t(obs)}^{\infty} f(t'|H)\mathrm{d}t'$$

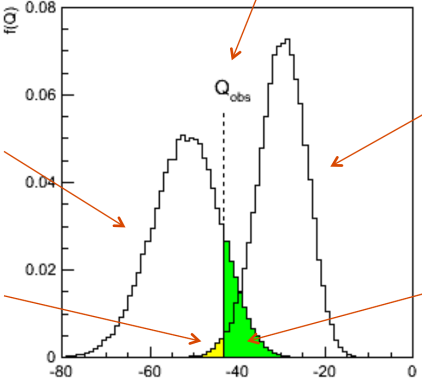
- Significance Z - defined as number of standard deviations a Gaussian variable would fluctuate in one direction to give the same p-value.

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

- Discovery $Z = 5$, p-value$= 2.9 \times 10^{-7}$

Distribution of Q

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



- $Q = -2\ln\frac{L_{s+b}}{L_b}$
- If $p_{s+b} < \alpha$ at confidence level $1-\alpha$
- If $p_b < 2.9 \times 10^{-7}$, reject background only $(Z = 5)$

Test Choices

- Discovery
  - H_0: background only hypothesis
  - H_1: events are mixture signal + background
  - Reject H_0 (typically $Z = 5$ significance)
- Limit
  - H_0: events are mixture signal + background
  - H_1: background only hypothesis
  - Upper limit: reject s which gives too high prediction for the signal yield ($s + b > b$ e.g. Higgs mass search)
  - Lower limit: reject s which gives too low prediction for the signal yield ($s + b < b$ e.g. neutrino disappearance)
- Confidence intervals
  - H_0: events are mixture signal + background
  - H_1: background only hypothesis
  - Reject H_0, parameters s which give both too high and too low predictions for the signal yield
  - Parameters s not rejected $\rightarrow$ CL intervals for s at confidence level (1-α)
- CL $= 1 - \alpha$ typically 95% for limit and 68% for confidence interval

Prototype analysis for profile likelihood ratio

- Suppose that a search analysis for signal is carried using some variable x leading to histogram (e.g. mass distribution $m_{\gamma\gamma}$)

$$\boldsymbol{n} = \{n_1, \ldots, n_N\}$$

- Assume that n_i are Poisson distributed

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

with signal strength μ and
signal and background predictions.

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

Control region for prototype analysis

- Often control regions are defined in the analysis to constrain some of the unknown parameters (e.g. with different selection cuts)
- Suppose that we have M auxiliary measurement

$$\boldsymbol{m} = \{m_1, \ldots, m_M\}$$

each Poisson distributed with the expectation value

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

- $\boldsymbol{\theta} = (b_{tot}, \boldsymbol{\theta}_s, \boldsymbol{\theta}_b)$ called nuisance parameters
Likelihood function of the problem

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_k}}{n_j!} e^{-(\mu s_i + b_i)} \prod_{j=1}^{M} \frac{u_j^{m_j}}{m_j!} e^{-u_j}$$

- Note the implicit dependence on many nuisance parameters
- Only one parameter of interest μ

Profile likelihood ratio

- Test based on profiled likelihood ratio test statistic

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

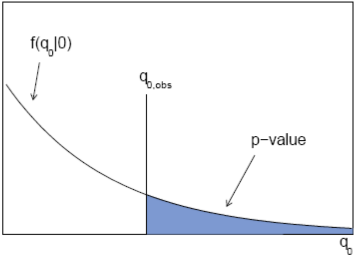
- $L(\mu, \hat{\hat{\boldsymbol{\theta}}})$ - maximize L for given μ; parameters $\hat{\hat{\boldsymbol{\theta}}}$ estimated from data
- $L(\hat{\mu}, \hat{\boldsymbol{\theta}})$ - find global maximum of L to determine $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ estimates
- $0 < \lambda(\mu) < 1$
  - $\lambda = 1$ - good agreement with data, $\hat{\mu}$ comes out close to μ
  - $\lambda = 0$ - model does not agree with data
- Test based on profile likelihood ratio gives near optimum performance

Test statistic for discovery

- Aim to reject background-only ($\mu = 0$) hypothesis with

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

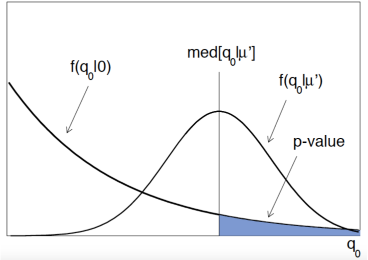
- Only positive values of $\hat{\mu}$ regarded as evidence against H_0
- Note in 'neutrino disappearance' experiment, the interesting region is $\hat{\mu} \leq 0$



- Large values of q_0, increasing discrepancy
- $p_0 = \int_{q_0,\text{obs}}^{\infty} f(q_0|H_0)\mathrm{d}q_0$
- if $p_0 < 2.9 \times 10^{-7}$ $\rightarrow$ discovery
- Note that H_1 not explicitly present. However alternative hypothesis determines the test to look for excess of events.

Expected sensitivity

- In the planing phase of the experiment, we want to know the expected sensitivity to reject background hypothesis given some alternative H_1
- Generate pseudo-experiments using alternative hypothesis H_1: $\mu = \mu'$
- Take the median as the expected $q_{0,\mathrm{obs}}$ and calculate p-value



- For p-value, we need $f(q_0|0)$, for sensitivity $f(q_0|\mu)$

Asymptotic formulae

- Profile likelihood ratio for large n $\rightarrow$ exponential form
- Simple asymptotic formulae for $f(q_0|0)$, $f(q_0|\mu)$
- p-value of $\mu = 0$ hypothesis

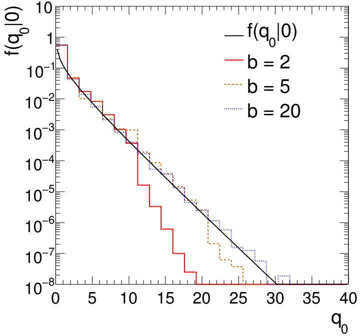
$$p_0 = 1 - \Phi(\sqrt{q_0})$$

- Significance of observed signal Z

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

- $n \sim \mathrm{Poiss}(\mu s + b)$
- $m \sim \mathrm{Poiss}(b)$

- Asymptotic formulae are good approximation for discovery ($q_0 = 25$) for $b > 25$



Test for upper limits

- Aim is to reject large values of μ which are incompatible with data $\hat{\mu}$

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

- Only small values of $\hat{\mu}$ regarded as evidence to reject H_0: $\mu \neq 0$ with alternative H_1: $\mu = 0$
- $p_\mu = \int\limits_{q_\mu,\mathrm{obs}}^{\infty} f(q_\mu|\mu)\mathrm{d}q$
- 95% CL is the highest value of μ for which the p-value is not less than size of the test $\alpha = 0.05$.

Unified (Feldman-Cousin) intervals

- Aim is to reject large and low values of μ which are incompatible with data $\hat{\mu}$

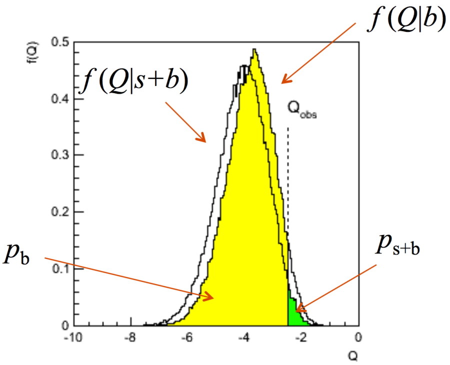
$$q_\mu = -2 \ln \lambda(\mu)$$

- Essentially the statistic used for Feldman-Cousin intervals
  G. Feldman and R.D. Cousins, Phys. Rev. D 57 (1998) 3873
- Here also including treatment of nuisance parameters

- Asymptotic formulae for discovery, limits, intervals
  Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554
- `RooStat` framework
  - Implements profile likelihood statistic
  - Allows to formulate the statistical model and perform MC pseudo-experiments for test inversion

Limits in experiments with low sensitivity

If model predicts very small signal (μ), we can run into problem excluding a parameter to which we have small or no sensitivity



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

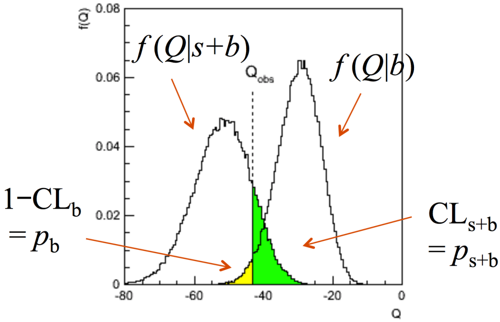
- Reject $s + b$ $(\mu > 0)$ hypothesis if $p_{s+b} < \alpha$
- For $\mu \sim 0$, parameter μ rejected with a probability $\sim \alpha$ size of the test
- Typically $\alpha = 0.05$, on average every 20^{th} limit measurement would give spurious exclusion

CL_s method

- Instead of the usual p-value (CL_{s+b}), define the test using ratio of the CL_b which equals to $1 - p_b$
  Alex Read, J. Phys., G28, 2002, 2693-2704.

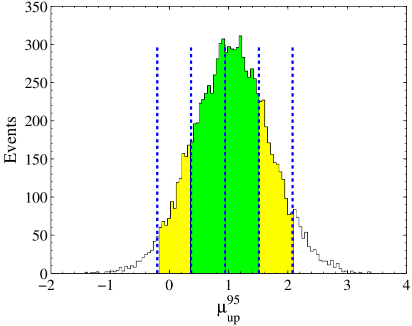
$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{p_{s+b}}{1 - p_b}$$

- Reject the null $s + b$ hypothesis if $CL_s < \alpha$
- $1/(1 - p_b)$ large when Q distribution close $\rightarrow$ prevent exclusion for low sensitivity
- In a way reduce Type II Error (accept H_0 when H_1 true)



Setting limits on $\mu = \sigma/\sigma_{SM}$

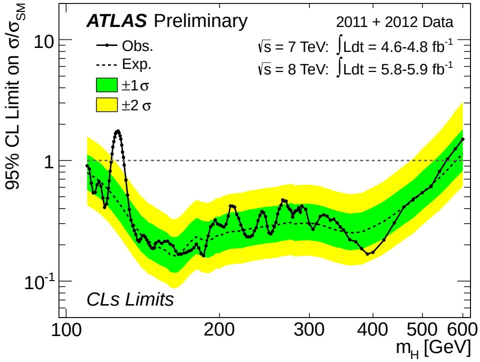
- The CL_s limit procedure results in a upper limit on the production μ_{up}
- Can be repeated as a function of some variable (m_H, $m_{ll}, \dots$)
- Pseudo-experiments used to sample what is the distribution of μ_{up} under background only hypotheses



- Dashed: asymptotic formulae
- Green (yellow): 1σ (2σ) expected limit from toy MC

Limit example: Higgs search

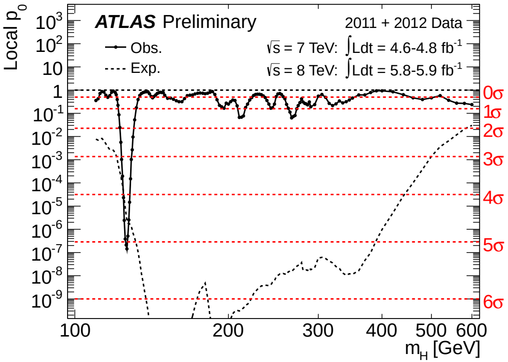
- The model has only one unknown parameter, Higgs mass
- A scaling factor, μ, on the Higgs cross-section used as a second parameter ATLAS-CONF-2011-163



- Solid: observed limit
- Dashed: expected limit (generate pseudo-experiments with bkg. only hypothesis)
- Green (yellow): 1σ (2σ) expected limit
- Expecting to exclude SM higgs from 110 GeV - 580 GeV

Local p-value

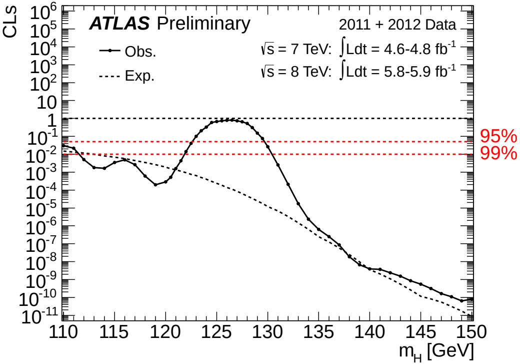
- Local p-value shows compatibility with background only hypothesis



- Dashed: expected p-value for SM higgs.
- Easier to discover if the Higgs was heavier

CL_s values for Higgs search

- If $CL_s(x_{obs};\mu) < \alpha$ we reject the parameter μ
- Here the level of confidence for SM Higgs with data: $CL_s(x_{obs};\mu = 1)$

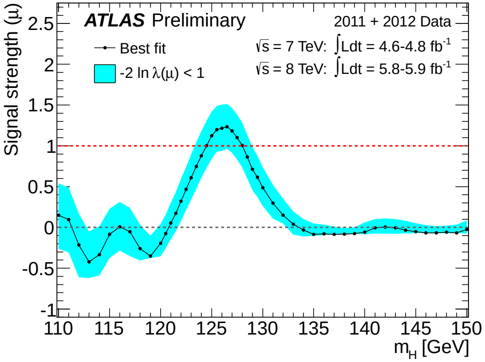


Signal strength

- The band defined by

$$-2\ln\lambda(\mu) = -2\ln L(\mu)/L(\hat{\mu}) < 1 \rightarrow \ln L(\mu) > \log L(\hat{\mu}) - 1/2$$

- Approximately the typical 1σ 68% CL confidence interval



Summary

- Hypothesis test based on likelihood ratio or profile likelihood ratio (in case of unknown parameters in the model) are most optimal
- Discovery - want to reject the background only hypothesis
- Limits or confidence intervals - inversion of hypothesis tests, reject the signal + background hypothesis

Backups

Maximum likelihood fit

$$\log L(\theta) = \underbrace{\log L(\hat{\theta})}_{\log L_{\max}} + \underbrace{\left[\frac{\partial \log L(\theta)}{\partial \theta_k} \right]}_{=0}{}_{\theta=\hat{\theta}} (\theta_k - \hat{\theta}_k) + \frac{1}{2}(\theta_i - \hat{\theta}_i) \underbrace{\left[\frac{\partial^2 \log L(\theta)}{\partial \theta_i \partial \theta_j} \right]}_{-\hat{V}_{ij}^{-1}[\hat{\theta}]}{}_{\theta=\hat{\theta}} (\theta_j - \hat{\theta}_j) + \ldots$$

- For sufficiently large n, the likelihood function is a paraboloid
- Several methods exploiting the shape to determine fit uncertainties
  - RFC method - inverse of covariance matrix defined by second derivatives of the likelihood (HESSE)
  - graphical method MINOS - graphical method (MINOS)

$$\log L(\theta) \approx \log L_{\max} - \frac{1}{2} \frac{(\theta - \hat{\theta})^2}{\hat{\sigma}^2_{\hat{\theta}}}$$

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{\max} - \frac{1}{2}$$

- MC pseudo-experiments for the case of small number of events

Distribution of q_μ

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

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

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

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

Independent of nuisance parameters.