Some aspects of statistics at LHC

N.V.Krasnikov

INR, Moscow

October 2013

Outline

1. Introduction
2. Parameters estimation
3. Confidence intervals
4. Systematics
5. Upper limits at LHC
6. Conclusion

1. Introduction

The statistical model of an analysis provides the complete description of that analysis

The main problem – from the known probability density $f(\vec{x},\vec{\theta})$ and x = x_{obs} to extract some information on θ parameter

Two approaches

1.Frequentist method

2. Bayesian method

Also very important – the notion of the likelihood

Likelihood - the probability density evaluated at the observed value $x=x_{obs}$

$$L(\vec{\theta}|\vec{x}_{obs,i}) = \prod_{i=1}^{l} f(\vec{x}_{obs,i}|\vec{\theta}) \,,$$

Frequents statistics – general philosophy

In frequentist statistics, probabilities are associated only with data, i.e. outcomes of repeatable observations. The preferred models are those for which our observations have non small probabilities

Quick review of probablility

Frequentist (A = outcome of repeatable observation):

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

Subjective (A = hypothesis):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

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

Bayes' theorem:

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

Bayesian statistics – general philosophy

In Bayesian statistics , interpretation of probability is extended to degree of belief(subjective probability). Bayesian methods can provide more natural treatment of non repeatable phenomena :

systematic uncertainties

Parameters estimation

Maximum likelihood principle

$$\frac{\partial}{\partial\vec{\theta}}L(\vec{\theta}|\vec{x}_{obs,i}) = 0\,.$$

$$lim_{l\to\infty}\vec{\theta}_0(\vec{x}_1,...\vec{x}_l) = \vec{\theta}\,.$$

Normal distribution

$$N(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}.$$

Bayesian method

- In Bayes approach

$$P(\vec{\theta}|\vec{x}_{obs,i}) = \frac{\pi(\vec{\theta})L(\vec{\theta}|\vec{x}_{obs,i})}{\int \pi(\vec{\theta})L(\vec{\theta}|\vec{x}_{obs,i})d\vec{\theta}}\,.$$

$$\frac{\partial P(\vec{\theta}|\vec{x}_{obs,i})}{\partial\vec{\theta}}|_{\vec{\theta}=\vec{\theta}_0} = 0\,.$$

For flat prior π(θ) = const

Bayes and likelihood coincide

Confidence intervals

Suppose we measure $x = x_{obs}$

- What are possible values of θ parameter?
- Frequentist answer:

Neyman belt construction

Alternative:

Bayes credible interval

Neyman belt construction

$$P(x_1 < X < x_2|\theta) = 1 - \alpha = \int_{x_1}^{x_2} f(x|\theta)dx.$$

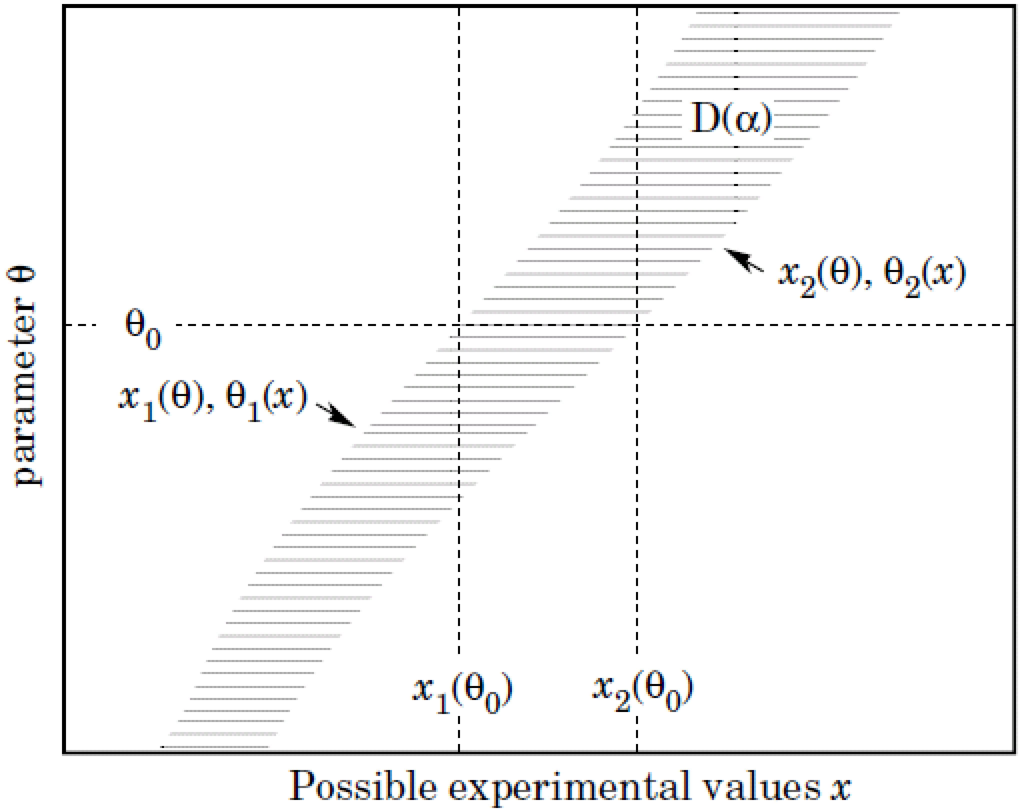
(1-α) – confidence level. The choice of x_1 and x_2 is not unique

$$1 - \alpha = P(x_1(\theta) < X < x_2(\theta)) = P(\theta_2(X) < \theta < \theta_1(X)),$$

$$\theta_2(X) = max_\theta x_1(\theta, x_{obs}),$$

$$\theta_1(X) = min_\theta x_2(\theta, x_{obs}).$$

Neyman belt construction



D(α)
$x_2(\theta), \theta_2(x)$
θ_0
$x_1(\theta), \theta_1(x)$
$x_1(\theta_0)$
$x_2(\theta_0)$
parameter θ
Possible experimental values x

Neyman belt construction

$$\int_{x_{obs}}^{\infty} f(x'|\theta_1)dx' = \beta' \,,$$

$$\int_{-\infty}^{x_{obs}} f(x'|\theta_2)dx' = \alpha' \,,$$

$$\alpha' + \beta' = \alpha \,.$$

Neyman belt construction

- For normal distribution Neyman belt equations for lower limit lead to

$$1-\alpha = P(-\infty < X < x_{obs}) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x_{obs}} e^{-\frac{(x-\mu_{low})^2}{2\sigma^2}}dx =$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x_{obs}-\mu_{low}}{\sigma}} e^{-\frac{y^2}{2}}dy = 1 - \frac{1}{\sqrt{2\pi}}\int_{\frac{x_{obs}-\mu_{low}}{\sigma}}^{\infty} e^{-\frac{y^2}{2}}dy.$$

Neyman belt equations

$$\mu \geq \mu_{low} = x_{obs} - \sigma s(\alpha)$$

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_s^\infty e^{-\frac{y^2}{2}} dy \, .$$

Maximal likelihood

Approximate estimate

$$2[lnL(\vec{\theta}_{max}) - lnL(\vec{\theta})] \leq s^2,$$

For normal distribution

$$\frac{(x_{obs} - \mu)^2}{\sigma^2} \leq s^2.$$

Bayes approach

Bayes theorem

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

$P(A|B)$ – conditional probability

Bayes approach

- Due to Bayes formula

$$P(\mu|x_{obs},\sigma) = \frac{\pi(\mu)N(x_{obs}|\mu,\sigma)}{\int_{-\infty}^{\infty}\pi(\mu')N(x_{obs}|\mu',\sigma)d\mu'}\,,$$

the statistics problem is reduced to the probability problem

$$\int_{\mu_{low}}^{\mu_{up}} P(\mu,\sigma|x_{obs})d\mu = 1-\alpha'-\beta'\,.$$

$$\int_{\mu_{up}}^{\infty} P(\mu|x_{obs},\sigma)d\mu = \alpha'\,, \qquad \int_{-\infty}^{\mu_{low}} P(\mu|x_{obs},\sigma)d\mu = \beta'\,.$$

Bayes approach

- The main problem – prior function π(θ) is not known
- For what prior frequentist and Bayes approaches coincide?

$$1-\alpha'-\beta' = \int_{-\infty}^{x_{obs}} [f(x,\theta_1) - f(x,\theta_2)]dx = \int_{\theta_1}^{\theta_2} P_f(\theta|x_{obs})d\theta\,.$$

$$P_f(\theta|x_{obs}) = -\int_{-\infty}^{x_{obs}} \frac{\partial}{\partial\theta} f(x,\theta)dx = \int_{x_{obs}}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx$$

$$\pi_f(\theta|x_{obs}) = \frac{P_f(\theta,x_{obs})}{f(x_{obs},\theta)}\,.$$

The relation between Bayes and frequentist approaches

- Two examples

1. Example A

$$f(x,\theta) = \Phi(x-\theta)$$

$$P_f(\theta|x_{obs}) = \Phi(x_{obs}-\theta)\,,$$

$$\pi_f(\theta|x_{obs}) = 1\,.$$

The relation between Bayes and frequentist approaches

2.Example B

$$F(x,\theta) = \frac{1}{\theta}\Phi(\frac{x}{\theta})$$

$$P_f(\theta|x_{obs}) = \frac{x_{obs}}{\theta^2}\Phi(\frac{x_{obs}}{\theta})\,,$$

$$\pi_f(\theta|x_{obs}) = \frac{x_{obs}}{\theta}\,.$$

Parameter determination with additional constraint

- Consider the case of normal distribution

$$N(x|\mu, \sigma^2 = 1),$$

with additional constraint $\mu \geq 0$

Maximum likelihood method gives

$$\mu_{best} = \begin{cases} x_0, & x \geq 0 \\ 0, & x < 0 \end{cases} = max(0, x_0).$$

Parameter determination with additional constraint

- How to construct the confidence interval for the μ parameter?

  Cousins, Feldman method

- Maximum of

$$R(\mu|x) = \frac{N(x|\mu,\sigma=1)}{N(x|\mu_{best},\sigma=1)} = \begin{cases} e^{-\frac{(x-\mu)^2}{2}}, & x \geq 0 \\ e^{x\mu-\frac{\mu^2}{2}}, & x < 0 \end{cases}.$$

Neyman belt construction

- The ordering principle on $R(\mu|x)$
- As a consequence we find

$$R(\mu|x_1) = R(\mu|x_2),$$

$$\int_{x_1}^{x_2} N(x|\mu, \sigma = 1)dx = 1 - \alpha.$$

Likelihood method

- For $x_0 < 0$

$$(x_0 - \mu)^2 - x_0^2 \leq s^2$$

- or

$$0 \leq \mu \leq x_0 + \sqrt{s^2 + x_0^2}.$$

Likelihood principle

- For $x_0 > 0$

$$(\mu - x_0)^2 \leq s^2.$$

- or

$$max(0, -s + x_0) \leq \mu \leq x_0 + s.$$

Bayes approach

- We choose π(μ) = θ(μ)

So prior function is zero for negative μ automatically

- The equation for the credible interval determination

$$\frac{\int_{\mu_1}^{\mu_2} e^{-\frac{1}{2}(\mu - x_0)^2} d\mu}{\int_{-x_0}^{\infty} e^{-\frac{1}{2}y^2} dy} = 1 - \alpha.$$

Confidence intervals for Poisson distribution

- The generalization of Neyman belt construction is

$$\sum_{n=n_1}^{n_2} P(n|\lambda) \geq 1-\alpha.$$

$$P(r|\mu) = \frac{\mu^r e^{-\mu}}{r!},$$

Klopper-Pearson interval

$$\lambda_{low} \leq \lambda \leq \lambda_{up}$$

$$\sum_{n=n_{obs}}^{\infty} P(n|\lambda_{low}) = \beta',$$

$$\sum_{n=0}^{n_{obs}} P(n|\lambda_{up}) = \alpha'.$$

$$\alpha' + \beta' = \alpha$$

Poisson distribution

- The Kloper-Pearson interval is conservative and it does not have the coverage property. Coverage is the probability that interval covers true value with the probability $1-\alpha.$ Besides for $\lambda_{up} = \lambda_{low}$

$$P(n_{obs}|\lambda_{up} = \lambda_{low}) = \alpha - 1$$

So we have negative probability - contradiction

Stevens interval

- To overcome these problems Stevens (1952) suggested to introduce new random variable U. Modified equations are

$$\sum_{n=n_{obs}+1}^{\infty} P(n|\lambda_{low}) + (1-U)\cdot P(n_{obs}|\lambda_{low}) = \beta' \,,$$

$$\sum_{n=0}^{n_{obs}-1} P(n|\lambda_{up}) + U\cdot P(n_{obs}|\lambda_{up}) = \alpha' \,.$$

Stevens equations

- One can derive Stevens equations using the regularization of discrete Poisson distribution(S.Bityukov,N.V.K). Namely let us introduce Poisson generalization

$$P_0(x,\lambda) = \sum_{n=0}^{\infty} \delta(x-n) P(n|\lambda)\,.$$

- The integral

$$\int_n^{\infty} \delta(x-n)dx$$

- is not well defined

Stevens interval

Let us introduce the regularization

$$\delta(x-n) \to \delta_{reg}(x,n|\delta_n,\gamma_n)\,,$$

$$\delta_{reg}(x,n|\delta_n,\gamma_n) = \frac{1}{\delta_n+\gamma_n}\cdot\left(-\theta(x-n-\delta_n)+\theta(x-n+\gamma_n)\right).$$

Stevens interval

- We can use Neyman belt construction for regularized distribution and we find

$$\sum_{n=n_{obs}+1}^{\infty} P(n|\lambda_{low}) + (1 - U(\gamma_{n_{obs}}, \delta_{n_{low}}))P(n_{obs}|\lambda_{low}) = \beta' ,$$

$$\sum_{n=0}^{n_{obs}-1} P(n|\lambda_{up}) + U(\gamma_{n_{obs}}, \delta_{n_{obs}})P(n_{obs}|\lambda_{up}) = \alpha' ,$$

$$U(\gamma_n, \delta_n) = \frac{\gamma_n}{\delta_n + \gamma_n} .$$

Stevens interval

In the limit of the regularization removement we find

$$\sum_{n=n_{obs}+1}^{\infty} P(n|\lambda_{low}) + (1 - U(n_{obs},\lambda_{low}))P(n_{obs}|\lambda_{low}) = \beta' \,,$$

$$\sum_{n=0}^{n_{obs}-1} P(n|\lambda_{up}) + U(n_{obs},\lambda_{up})P(n_{obs}|\lambda_{up}) = \alpha' \,,$$

$$U(n,\lambda)) = lim_{(\gamma_n,\delta_n)\to 0}\frac{\gamma_n}{\delta_n + \gamma_n} \leq 1 \,.$$

Likelihood method

- The use of likelihood method gives

$$\frac{d}{d\lambda}L(\lambda|n_{obs})|_{\lambda=\lambda_{max}} = 0$$

$$L(\lambda|n_{obs}) = \frac{1}{n_{obs}!}(\lambda)^{n_{obs}}e^{-\lambda}.$$

- The solution is

$$\lambda_{max} = n_{obs}.$$

Likelihood method

$$2lnL(\lambda_{max} = n_{obs}|n_{obs}) - 2lnL(\lambda|n_{obs}) \leq s^2$$

$$2[(\lambda - n_{obs}) + n_{obs}(ln\ n_{obs} - ln\lambda)] \leq s^2.$$

$$\lambda_{low}(n_{obs}, s) \leq \lambda \leq \lambda_{up}(n_{obs}, s).$$

Bayes approach

- The basic equations are

$$P(\lambda|n_{obs}) = \frac{\pi(\lambda)P(n_{obs}|\lambda)}{\int_0^\infty \pi(\lambda')P(n_{obs}|\lambda')d\lambda'}$$

$$\int_{\lambda_{low}}^{\lambda_{up}} P(\lambda|n_{obs})d\lambda = 1 - \alpha\,.$$

- Due to identities

$$\int_{\lambda_{up}}^{\infty} P(n_{obs}|\lambda)d\lambda = \sum_{n=0}^{n_{obs}} P(n|\lambda_{up}),$$

$$\int_0^{\lambda_{low}} P(n_{obs}|\lambda)d\lambda = \sum_{n=n_{obs}+1}^{\infty} P(n|\lambda_{low}).$$

Bayes approach

Upper Klopper-Pearson limit coincides

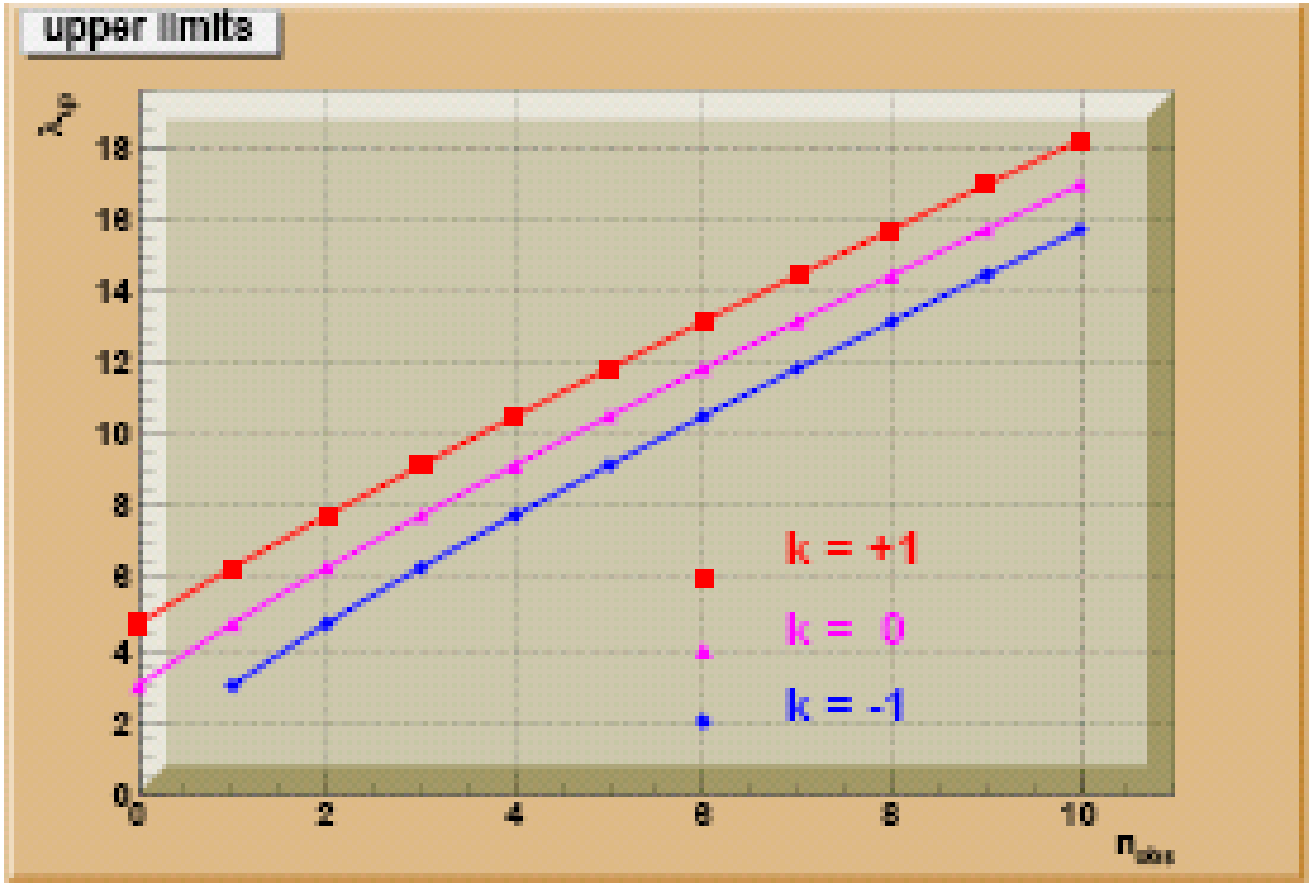
with Bayesian limit for flat prior and lower limit corresponds to prior $\pi(\lambda) = \frac{const}{\lambda}$

The Stevens equations for $U(n, \lambda)$ non dependent on λ

are equivalent to Bayes approach with prior function

$$\pi(\lambda|n_{obs}, U) = [U + (1 - U)\frac{n_{obs}}{\lambda}] \, .$$

Uncertainties in extraction of an upper limit



Modified frequentist definition

- We require(S.Bityukov,N.V.K.,2012) that

$$1 - \beta^{'} \geq P_{-}(n_{obs}|\lambda; c_k) \geq \alpha^{'} \, ,$$

$$P_{-}(n_{obs}|\lambda; c_k) \equiv \sum_k c_k^2 P_{-}(n_{obs} + k|\lambda),$$

$$P_{-}(n_{obs}|\lambda) \equiv \sum_{n=0}^{n_{obs}} P(n_{obs}|\lambda),$$

- Our definition is equivalent to Bayes
- approach with prior function $\pi(\lambda) = \sum_k c_k^2 l_k \lambda^k,$

$$l_k = \frac{n!}{(n+k)!}.$$

Signal extraction for nonzero background

- Consider the case

$$\lambda = \lambda_b + \lambda_s,$$

$$\lambda_b = L\epsilon_b\sigma_b$$

$$\lambda_s = L\sigma_s\epsilon_s$$

- Cousins-Feldman method

$$\sum_{n=n_-}^{n_+} P(n|\lambda_b + \lambda_s)$$

Nonzero background

- Likelihood ordering

$$R = \frac{P(n|\lambda_b + \lambda_s)}{P(n|\lambda_b + \lambda_{s,best})},$$

Plus Neyman construction

$$\sum_{n=n_1}^{n_2} P(n|\lambda_b + \lambda_s) \geq 1 - \alpha.$$

CL$_S$ method(T.Junk,A.Read)

- Upper bound

$$P(n \leq n_{obs}|\lambda_b + \lambda_s) = \sum_{n=0}^{n_{obs}} P(n|\lambda_b + \lambda_s) \geq \alpha.$$

- CL$_S$ method

$$\frac{P(n \leq n_{obs}|\lambda_b + \lambda_s)}{P(n \leq n_{obs}|\lambda_b)} \geq \alpha,$$

- In Bayes approach it corresponds to the replacement

$$\theta(\lambda) \rightarrow \quad const \cdot \theta(\lambda - \lambda_b).$$

Bayes method

$$P(\lambda_s|n_{obs},\lambda_b) = P(n_{obs}|\lambda_b+\lambda_s)\pi(\lambda_b,\lambda_s)\cdot\frac{1}{\int_0^\infty P(n_{obs}|\lambda_b+\lambda'_s)\pi(\lambda_b,\lambda'_s)d\lambda'_s}.$$

- For flat prior

$$P(\lambda_s|n_{obs},\lambda_b) = P(n_{obs}|\lambda_b+\lambda_s)\cdot\frac{1}{\int_{\lambda_b}^\infty P(n_{obs}|\lambda)d\lambda}.$$

- We can interprete this formula in terms of conditional probability

Bayes method

- Namely the probability that parameter λ lies in the interval [λ, λ+dλ] provided $\lambda \geq \lambda_b$ is determined by the formula

$$P(\lambda|n_{obs}, \lambda \geq \lambda_b)d\lambda = \frac{P(\lambda|n_{obs})d\lambda}{P(\lambda \geq \lambda_b)} = \frac{P(\lambda|n_{obs})d\lambda}{\int_{\lambda_b}^{\infty} P(n_{obs}|\lambda')d\lambda'}$$

that coincides with the previous Bayes formula

Systematics

- 1. Systematics that can be eliminated by the measurement of some variables in other kinematic region
- 2. Uncertainties related with nonexact accuracy in determination of particle momenta, misidentification...
- 3. Uncertainties related with nonexact knowledge of theoretical cross sections

Systematics

- 3 methods to deal with systematics(at least)

1. Suppose we measure some events in two kinematic regions with distribution functions $N(x|\mu_B + \mu_S, \sigma^2_{B+S})$, $N(y|\mu_B, \sigma^2_B)$.

The random variable Z = X-Y obeys normal distribution $N(z|\mu_S, \sigma^2_{B+S} + \sigma^2_B)$

As a consequence we find

$$|x_{obs} - y_{obs} - \mu_S| \leq k \cdot \sqrt{\sigma^2_{B+S} + \sigma^2_B} \,.$$

Systematics

- For Poisson distributions

$P(n, \lambda_b + \lambda_s).$ and $P(m, \tau\lambda_b)$

due to identity

$$P(n|\lambda_1)P(m|\lambda_2) = P(n+m|\lambda_1+\lambda_2) \cdot Bi(n|n+m,\rho)\,,$$

$$Bi(n|m,\rho) = \frac{m!}{(m-n)!}\rho^n(1-\rho)^{m-n}\,,$$

$$\rho = \frac{\lambda_1}{\lambda_2} = \frac{\tau}{1+\frac{\lambda_s}{\lambda_b}}.$$

Systematics

- The problem is reduced to the determination of the ρ parameter from experimental data

Systematics

2.Bayesian treatment or Cousins-Highland

method is based on integration over nonessential variables

$$P_{av}(x|\theta) = \frac{\int d\theta' \pi(\theta') P(x|\theta, \theta')}{\int d\theta' dx \pi(\theta') P(x|\theta, \theta')}.$$

For normal distributions and flat prior we find

$$G(x|\mu_0, \sigma^2) = \int_{-\infty}^{\infty} d\mu N(x|\mu, \sigma^2) N(\mu|\mu_0, \sigma_\mu^2) = N(x|\mu_0, \sigma^2 + \sigma_\mu^2).$$

Systematics

- In other words the main effect is the replacement

$$\sigma^2 \to \sigma^2 + \sigma_\mu^2.$$

and the significance is

$$s = \frac{|x_0 - \mu_0|}{\sqrt{\sigma^2 + \sigma_\mu^2}}.$$

So for normal distribution this method coincides with the first method

Systematics

- Profile likelihood method

Suppose likelihood function $L(\lambda, \vec{\theta})$

depends on nonessential variables θ

and essential variables λ

Profile likelihood

$$\frac{\partial L(\lambda, \vec{\theta})}{\partial \vec{\theta}} |_{\vec{\theta}=\vec{\theta}_0} = 0.$$

$$\bar{L}(\lambda) = L(\lambda, \vec{\theta}_0(\lambda)) \, ,$$

Profile likelihood

New variable(statistics) $t_\lambda = -2\ \ln \frac{L(\lambda, \vec{\theta}_0(\lambda))}{L(\hat{\lambda}, \vec{\theta}_0(\lambda))},$

$$\frac{\partial L}{\partial \lambda}|_{\hat{\theta}=\hat{\theta}_0, \lambda=\hat{\lambda}_0} = \frac{\partial L}{\partial \hat{\theta}}|_{\hat{\theta}=\hat{\theta}_0, \lambda=\hat{\lambda}_0} = 0.$$

- Per construction $t_\lambda \geq 0.$
- For new statistics t_λ defines probability density $f(t_\lambda|\lambda, \vec{\theta})$

Profile likelihood

- For normal distributions profile likelihood method coincides with the Cousins-Highland method
- Very often p-value is used
- By definition

$$p_\lambda = \int_{t_{\lambda,obs}}^{\infty} f(t_\lambda, \vec{\theta}) dt_\lambda .$$

p-value determines the agreement of data with a model

- Small p-value($p < 5.9*10^{-7}$) - the model is excluded by experimental data

P-value

- For Poisson distribution p-value definition is

$$P_+(n_{obs}|\lambda_b) \equiv P(n \geq n_{obs}|\lambda_b) = \sum_{n=n_{obs}}^{\infty} P(n|\lambda_b).$$

Limits on new physics at LHC

For the Higgs boson search CMS and ATLAS introduce the extended model $\sigma_H \to \mu\sigma_H$

with additional μ parameter and the replacement cross section the same. The case μ =1 corresponds to SM. The case μ=0 corresponds to the absence of the SM Higgs boson.

The likelihood of the general model can be written in the form

Likelihood of the model

$$L(data|\mu,\theta) = Poisson(data|\mu \cdot s(\theta) + b(\theta)) \cdot p(\tilde{\theta}|\theta)\,,$$

$$Poisson(data|\mu \cdot s + b) = \prod_{i=1}^{k} P(n_{obs,i}|\mu \cdot s_i + b_i)\,.$$

Here $p(\tilde{\theta}|\theta)$ is the probability density of nonessential parameters . Usually $p(\tilde{\theta}|\theta)$

Is taken as normal or lognormal distribution

Bayes approach

- In Bayes approach the use of the formula

$$P(\mu) = \frac{1}{C}\int_{\theta} L(data|\mu,\theta)\rho_{\theta}(\theta)\pi_{\mu}(\mu))d\theta\,.$$

- allows to determine the probability density for μ parameter. Upper limit μ_{up} is detemined as

$$\int_0^{\mu_{up}} P(\mu)d\mu = 1-\alpha.$$

Usually α= 0.05

Frequentist approach

- CMS and ATLAS use statistics

$$q_\mu = -2\ \ln \frac{L(data|\mu s + b)}{L(data|\hat{\mu} s + b)} ,$$

Often modifications are used with additional conditions as

(a) $\hat{\mu} \geq 0;$

(b) $\hat{\mu} \leq \mu;$

(c) $0 \leq \hat{\mu} \leq \mu.$

Frequentist approach

Very often the hypothesis μ=0 is tested against μ>0. For such case it is convenient to use

$$q_0 = \begin{cases} -2\ \ln \dfrac{L(data|b(\hat{\theta}_0))}{L(data|b(\hat{\theta}) + \hat{\mu}s(\hat{\theta}))}, & \hat{\mu} \geq 0, \\ 0, \quad \hat{\mu} < 0. \end{cases}$$

For single Poisson

$$q_0 = \begin{cases} -2\ [n \ln b - n \ln n + n - b], & n \geq b, \\ 0, \quad n < b. \end{cases}$$

Single Poisson

- By construction $q_0 \geq 0$ and

$$p_0 = \int_{q_{0,obs}}^{\infty} f(q_0|0) dq_0 = \sum_{n=n_{obs}}^{\infty} P(n|b)$$

In the limit $n_{obs} \gg 1$ the probability density is

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

Upper limits

- To derive upper limits the statistics

$$q_\mu = \begin{cases} -2\ \ln \dfrac{L(data|\mu s + b)}{L(data|\hat{\mu} s + b)}, & \hat{\mu} \leq \mu, \\ 0, \quad \hat{\mu} > \mu. \end{cases}$$

is used. For single Poisson

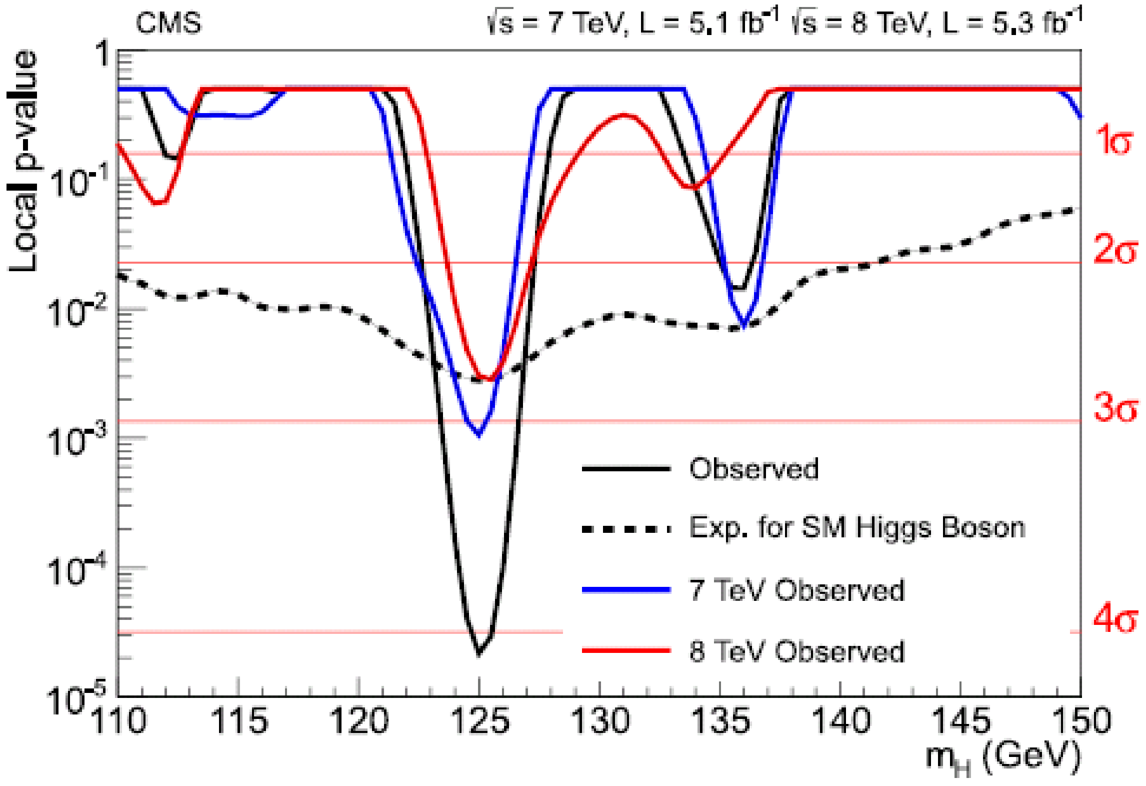
$$q_\mu = \begin{cases} -2\ [-n \ln(\mu s + b) + n \ln n - (\mu s + b) + n], & n \leq \mu s + b, \\ 0, \quad n > \mu s + b. \end{cases}$$

$$p_\mu = \int_{q_{\mu,obs}}^{\infty} f(q_\mu|\mu) dq_\mu = \sum_{n=0}^{n_{obs}} P(n|\mu s + b).$$

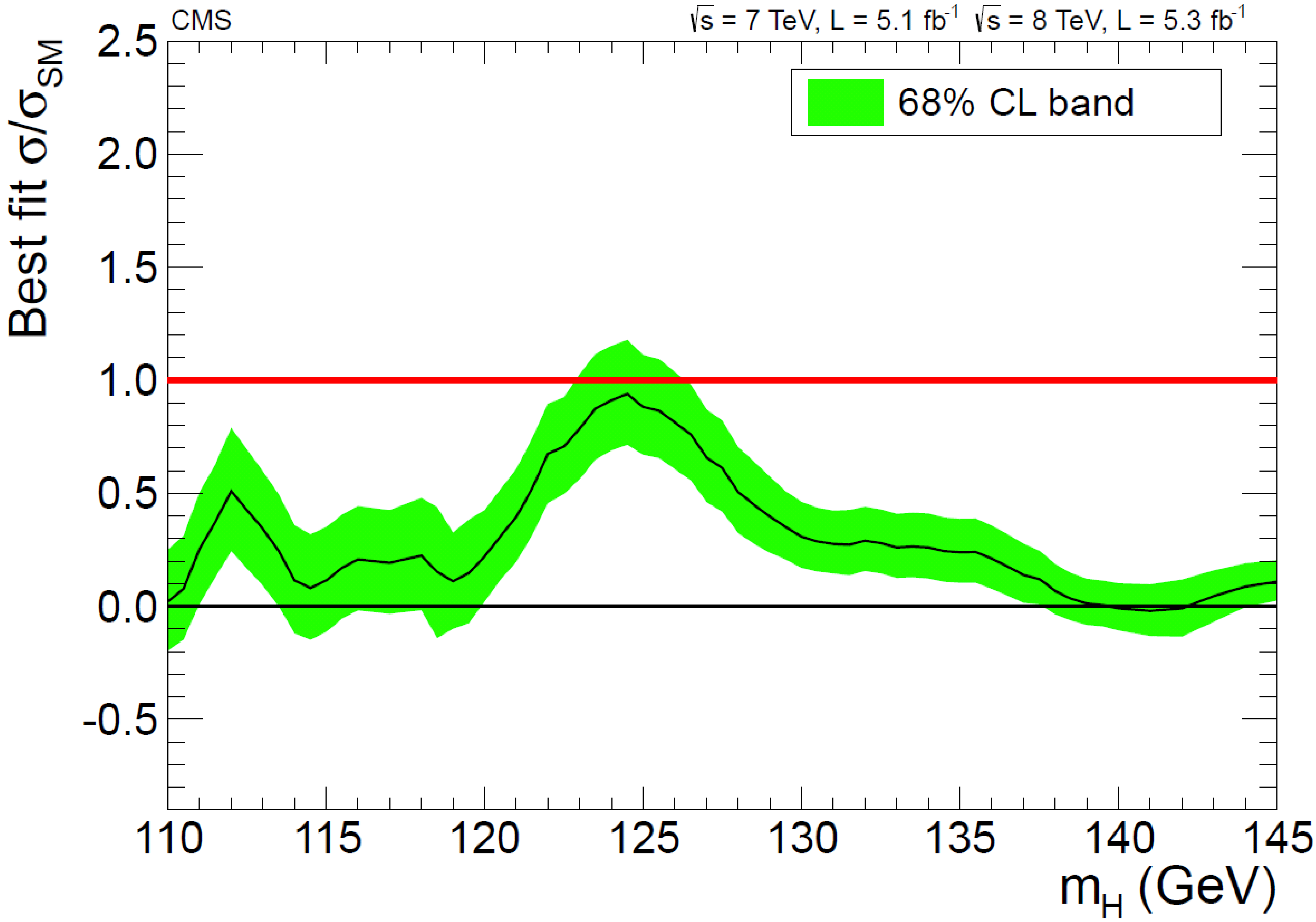
Higgs boson search at CMS

As an illustration consider the Higgs boson search at CMS detector

P-value for Higgs boson search

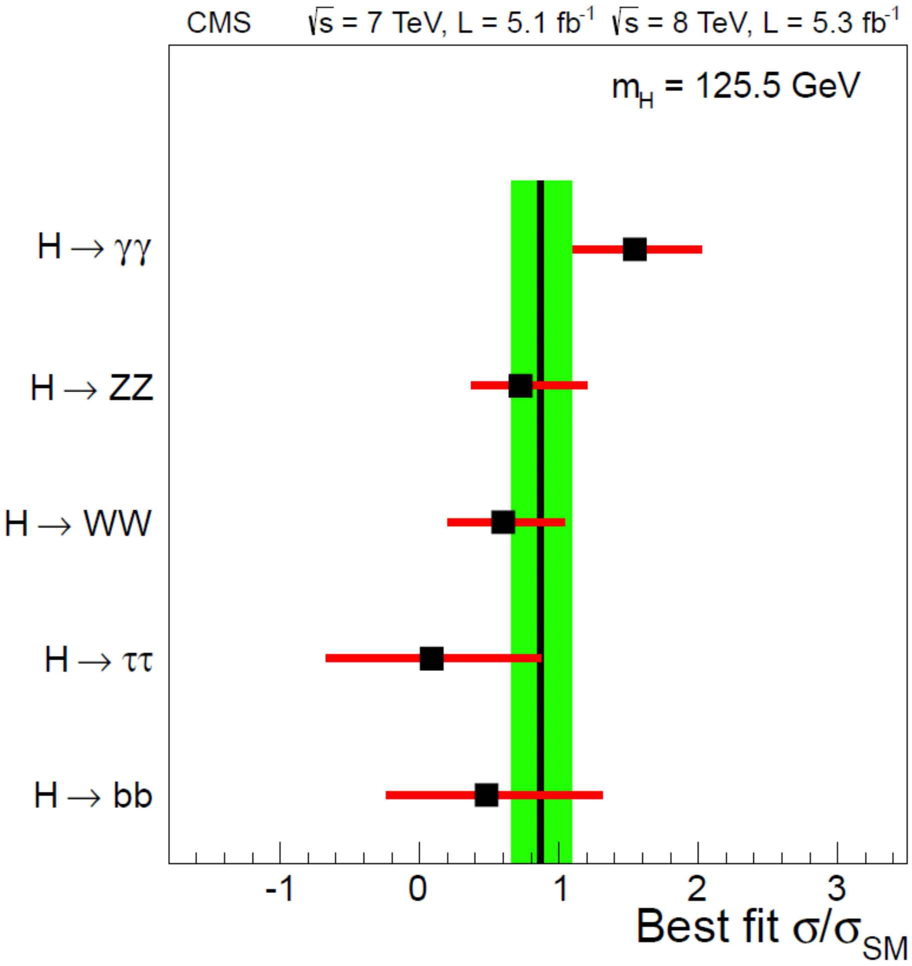


October 2013

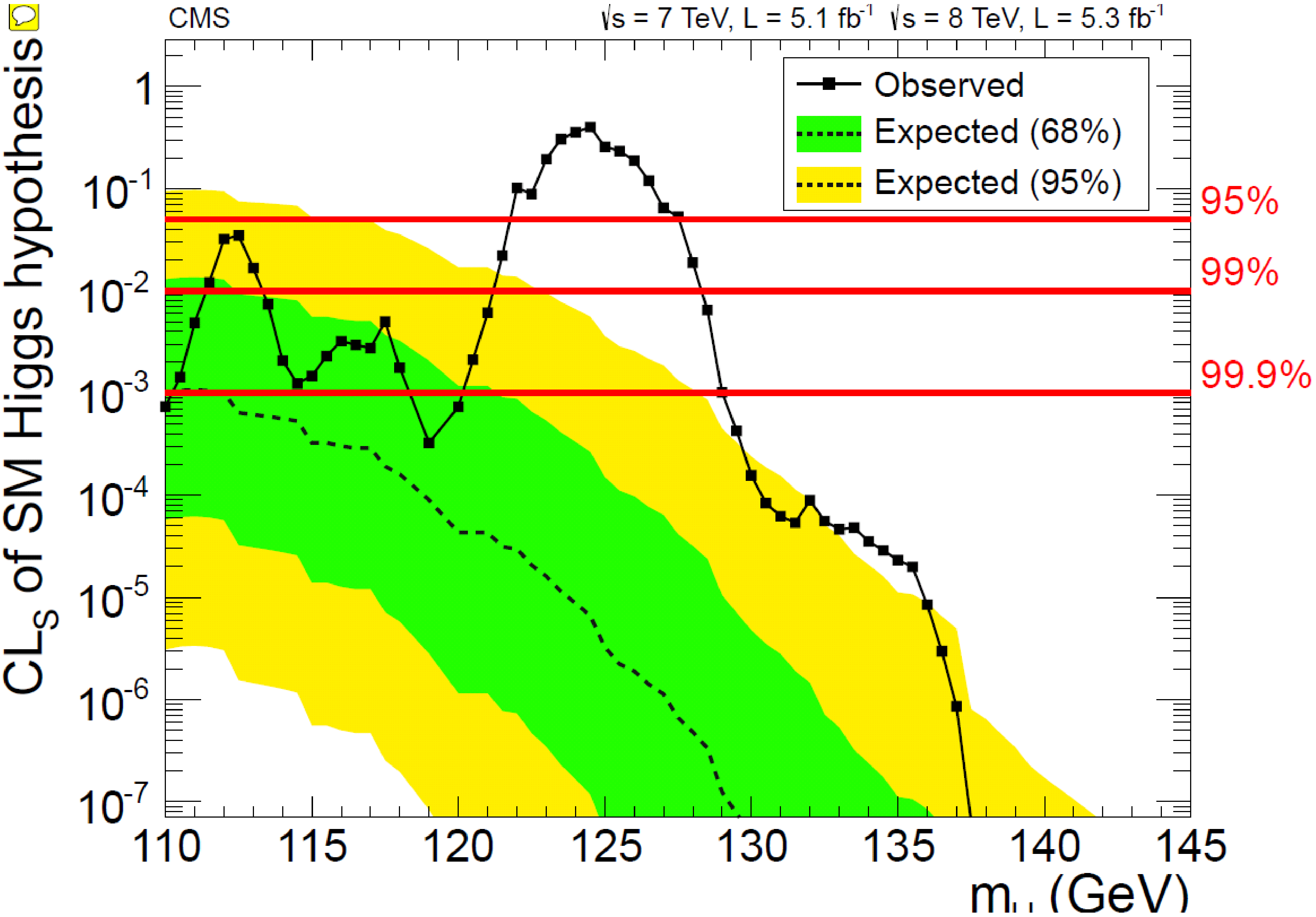


October 2013

Summary of Higgs boson measurements



October 2013



October 2013

Conclusions

Experiments CMS and ATLAS use both frequentist and Bayesian methods to extract the parameters of Higgs boson and limits on new physics. As a rule they give numerically similar results

BACKUP

Hypotheses testing

Simple vs. Compound Hypotheses

A quick review of frequentist statistical tests

Consider a hypothesis H_0 and alternative H_1.

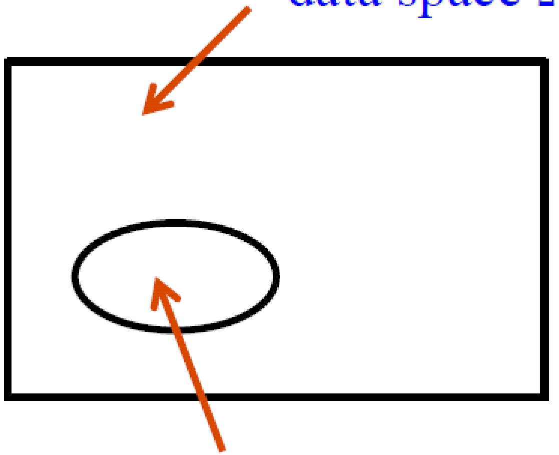
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0.



data space Ω

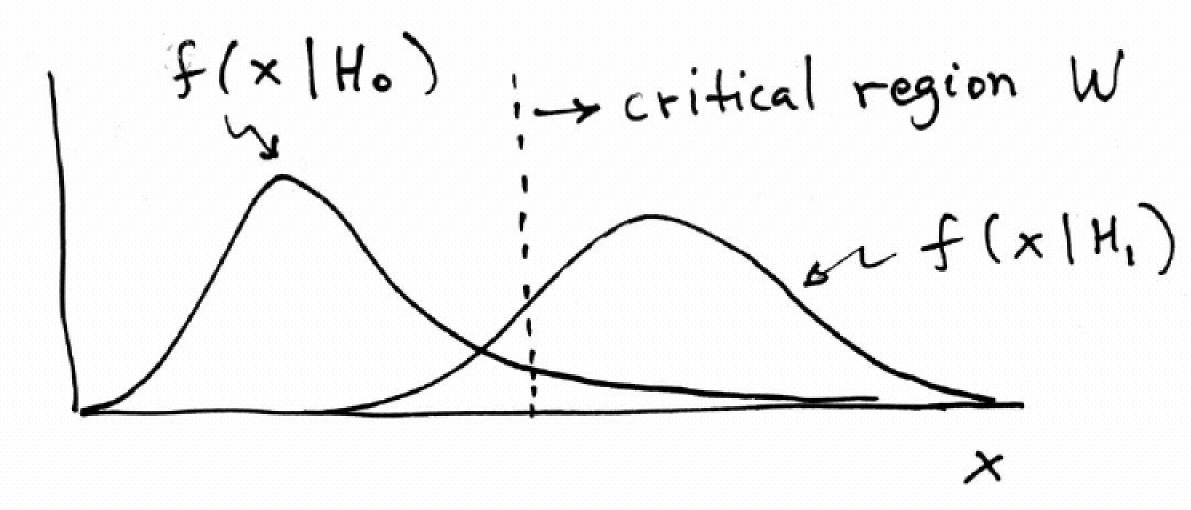
critical region w

Definition of a test

But in general there are an infinite number of possible critical regions that give the same significance level α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in S - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

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

Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of H_0, (background) versus H_1, (signal) the critical region should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

inside the region, and $\leq c$ outside, where c is a constant which determines the power.

Equivalently, optimal scalar test statistic is $\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$

N.B. any monotonic function of this is leads to the same test.

p-values

Suppose hypothesis *H* predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\mathrm{obs}}$

What can we say about the validity of *H* in light of the data?

Express level of compatibility by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

Using a p-value to define test of H_0

One can show the distribution of the p-value of H, under assumption of H, is uniform in [0,1].

So the probability to find the p-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We can define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

> Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.
>
> If data observed in the critical region, reject the value θ .

Now invert the test to define a confidence interval as:

> set of θ values that would not be rejected in a test of size α (confidence level is $1 - \alpha$).

The interval will cover the true value of θ with probability $\geq 1 - \alpha$.

Equivalently, the parameter values in the confidence interval have p-values of at least α.

Ingredients for a frequentist test

In general to carry out a test we need to know the distribution of the test statistic $t(x)$, and this means we need the full model $P(x|H)$.

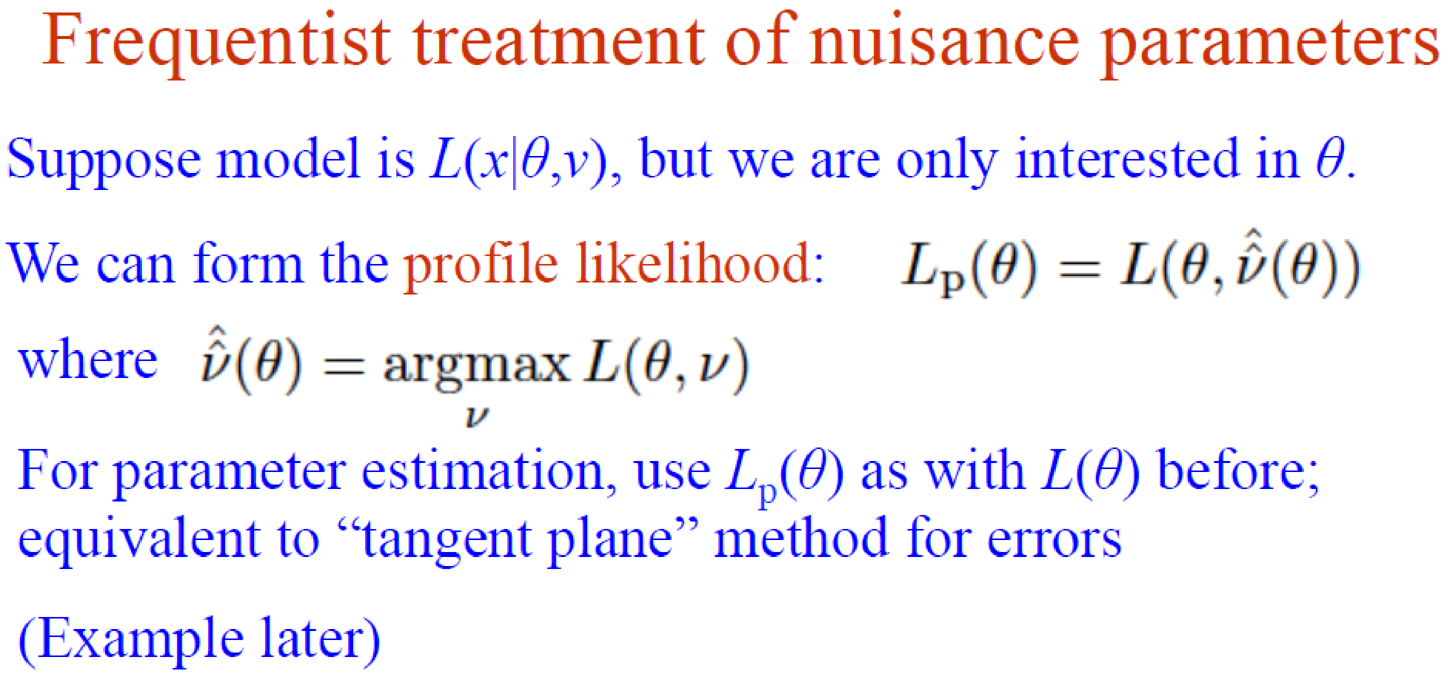
Often one can construct a test statistic whose distribution approaches a well-defined form (almost) independent of the distribution of the data, e.g., likelihood ratio to test a value of θ:

$$t_\theta = -2 \ln \frac{L(\theta)}{L(\hat{\theta})}$$

In the large sample limit t_θ follows a chi-square distribution with number of degrees of freedom = number of components in θ (Wilks' theorem).

So here one doesn't need the full model $P(x|\theta)$, only the observed value of t_θ.

Nuisance parameters(systematics)



Frequentist treatment of nuisance parameters

Suppose model is $L(x|\theta,\nu)$, but we are only interested in θ.

We can form the profile likelihood: $L_{\mathrm{p}}(\theta) = L(\theta, \hat{\hat{\nu}}(\theta))$

where $\hat{\hat{\nu}}(\theta) = \underset{\nu}{\operatorname{argmax}}\, L(\theta, \nu)$

For parameter estimation, use $L_{\mathrm{p}}(\theta)$ as with $L(\theta)$ before; equivalent to "tangent plane" method for errors

(Example later)

Frequentist treatment of nuisance parameters in a test

Suppose we test a value of θ with the profile likelihood ratio:

$$t_\theta = -2 \ln \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

We want a *p*-value of θ:

$$p_\theta = \int_{t_{\theta,\mathrm{obs}}}^{\infty} f(t_\theta|\theta, \nu)\, dt_\theta$$

Wilks' theorem says in the large sample limit (and under some additional conditions) $f(t_\theta|\theta,\nu)$ is a chi-square distribution with number of degrees of freedom equal to number of parameters of interest (number of components in θ).

Simple recipe for *p*-value; holds regardless of the values of the nuisance parameters!

Frequentist treatment of nuisance parameters in a test (2)

But for a finite data sample, $f(t_\theta|\theta,\nu)$ still depends on ν.

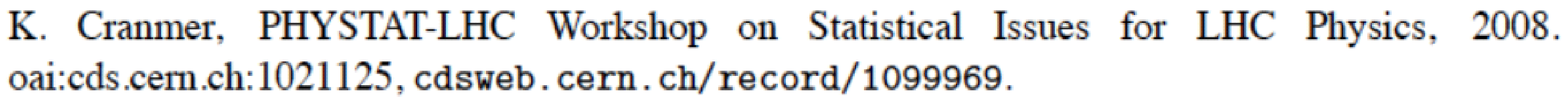
So what is the rule for saying whether we reject θ?

Exact approach is to reject θ only if $p_\theta < \alpha$ (5%) for all possible ν.

This can make it very hard to reject some values of θ; they might not be excluded for value of ν known to be highly disfavoured.

Resulting confidence level too large (“over-coverage”).

Profile construction (“hybrid resampling”)



K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008. oai:cds.cern.ch:1021125, cdsweb.cern.ch/record/1099969.

Compromise procedure is to reject θ if $p_\theta \leq \alpha$ where the p-value is computed assuming the value of the nuisance parameter that best fits the data for the specified θ (the profiled values):

$$\hat{\hat{\nu}}(\theta) = \underset{\nu}{\operatorname{argmax}} \, L(\theta, \nu)$$

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\hat{\nu}}(\theta))$

Elsewhere it may under- or over-cover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

Bayesian treatment of nuisance parameters

Conceptually straightforward: all parameters have a prior: $\pi(\theta, \nu)$

Often $\pi(\theta, \nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

Often $\pi_\theta(\theta)$ "non-informative" (broad compared to likelihood).

Usually $\pi_\nu(\nu)$ "informative", reflects best available info. on *ν*.

Use with likelihood in Bayes' theorem:

$$p(\theta, \nu | x) \propto L(x | \theta, \nu)\pi(\theta, \nu)$$

To find $p(\theta|x)$, marginalize (integrate) over nuisance param.:

$$p(\theta|x) = \int p(\theta, \nu | x)\, d\nu$$

The marginal (integrated) likelihood

If the prior factorizes: $\pi(\theta, \nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

then one can compute the marginal likelihood as:

$$L_{\mathrm{m}}(x|\theta) = \int L(x|\theta, \nu)\,\pi_\nu(\nu)\,d\nu$$

This represents an average of models with respect to $\pi_\nu(\nu)$ (also called "prior predictive" distribution).

Does not represent a realistic model for the data;
ν would not vary upon repetition of the experiment.

Leads to same posterior for θ as before:

$$p(\theta|x) = \int p(\theta, \nu|x)\,d\nu \propto \int L(x|\theta, \nu)\pi_\nu(\nu)\pi_\theta(\theta)\,d\nu = L_{\mathrm{m}}(x|\theta)\pi_\theta(\theta)$$

The (pure) frequentist equivalent

In a purely frequentist analysis, one would regard both *x* and *y* as part of the data, and write down the full likelihood:

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

"Repetition of the experiment" here means generating both *x* and *y* according to the distribution above.

So we could either say that $\pi_v(v)$ encapsulates all of our prior knowledge about *v* and forget that it came from a measurement,

$$p(\theta, \nu|x) \propto L(x|\theta, \nu)\pi_\theta(\theta)\pi_\nu(\nu)$$

or regard both *x* and *y* as measurements,

$$p(\theta, \nu|x, y) \propto L(x|\theta, \nu)L(y|\nu)\pi_\theta(\theta)\pi_0(\nu)$$

In the Bayesian approach both give the same result.

Frequentist use of Bayesian ingredients

For subjective Bayesian, end result is the posterior $p(\theta|x)$.

Use this, e.g., to compute an upper limit at 95% "credibility level":

$$P(\theta < \theta_{\mathrm{up}}|x) = \int_{-\infty}^{\theta_{\mathrm{up}}} p(\theta|x)\,d\theta = 95\%$$

→ Degree of belief that $\theta < \theta_{\mathrm{up}}$ is 95%.

But θ_{up} is $\theta_{\mathrm{up}}(x)$, a function of the data. So we can also ask

$$P(\theta < \theta_{\mathrm{up}}(x)|\theta) = ? \qquad \text{(a frequentist question)}$$

Here we are using a Bayesian result in a frequentist construct by studying the coverage probability, which may be greater or less than the nominal credibility level of 95%.

More Bayesian ingredients in frequentist tests

Another way to use Bayesian ingredients to obtain a frequentist result is to construct a test based on a ratio of marginal likelihoods:

$$t_{\rm m}(x) = \frac{L_{\rm m}(x|s)}{L_{\rm m}(x|b)} = \frac{\int L(x|\nu, s)\pi_\nu(\nu)\,d\nu}{\int L(x|\nu, b)\pi_\nu(\nu)\,d\nu}$$

Except in simple cases this will be difficult to compute; often use instead ratio of profile likelihoods,

$$t_{\rm p}(x) = \frac{L_{\rm p}(x|s)}{L_{\rm p}(x|b)} = \frac{L(x|\hat{\hat{\nu}}(s), s)}{L(x|\hat{\hat{\nu}}(b), b)}$$

or in some cases one may just use the ratio of likelihoods for some chosen values of the nuisance parameters.

Here the choice of statistic influences the optimality of the test, not its "correctness".

Prior predictive distribution for statistical test

The more important use of a Bayesian ingredient is in computing the distribution of the statistic. One can take this to be the Bayesian averaged model (prior predictive distribution), i.e.,

Generate $x \sim L_{\mathrm{m}}(x|\mathrm{s})$ to determine $f(t(x)|\mathrm{s})$,

Generate $x \sim L_{\mathrm{m}}\,(x|\mathrm{b})$ to determine $f(t(x)|\mathrm{b})$.

Use of the marginal likelihood results in a broadening of the distributions of $t(x)$ and effectively builds in the systematic uncertainty on the nuisance parameter into the test.

Prior predictive distribution for statistical test

Note the important difference between two approaches:

1) Pure frequentist: find "correct" model (enough nuisance parameters) and construct a test statistic whose distribution is almost independent of the nuisance parameters (and/or use profile construction).

2) Hybrid frequentist/Bayesian: construct an averaged model by integrating over a prior for the nuisance parameters; use this to find sampling distribution of test statistic (which itself may be based on a ratio of marginal or profile likelihoods).

Search for a signal process

Suppose a signal process is not known to exist and we want to search for it.

We observe n events and for each measure a set of numbers $\boldsymbol{x}$. The relevant hypotheses are:

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

Rejecting H_0 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:

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

Likelihoods for full experiment

We observe *n* events, and thus measure *n* instances of ***x***.

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.

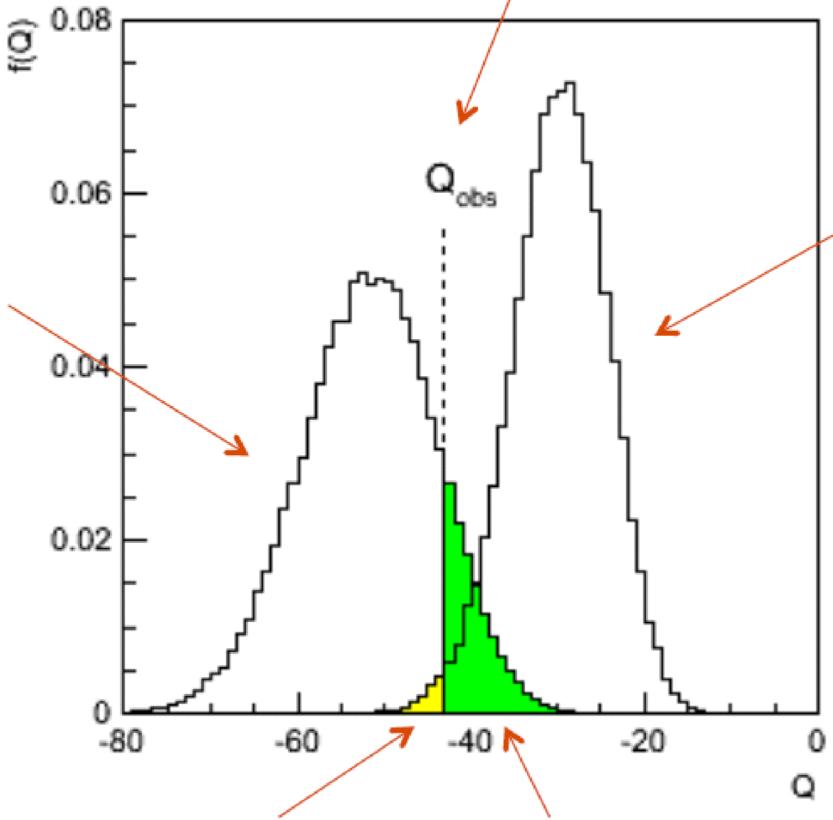
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

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.



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

$f\,(Q|\mathrm{b})$

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:

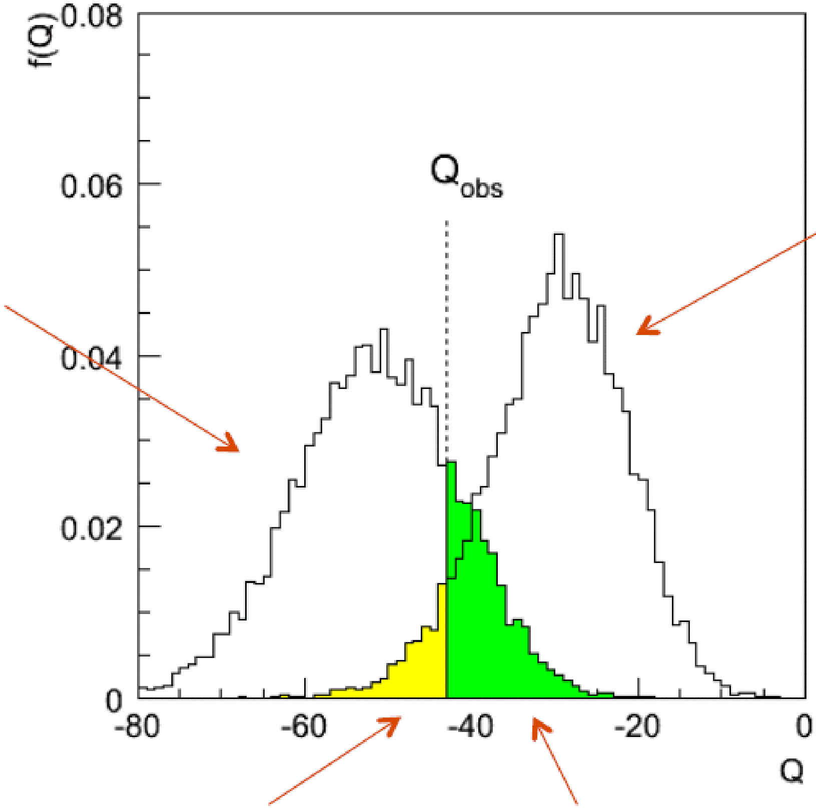
$$f(Q) = \int f(Q|b)\pi(b)\,db$$

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



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

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

p-value of b only

p-value of s+b

Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right] .$$

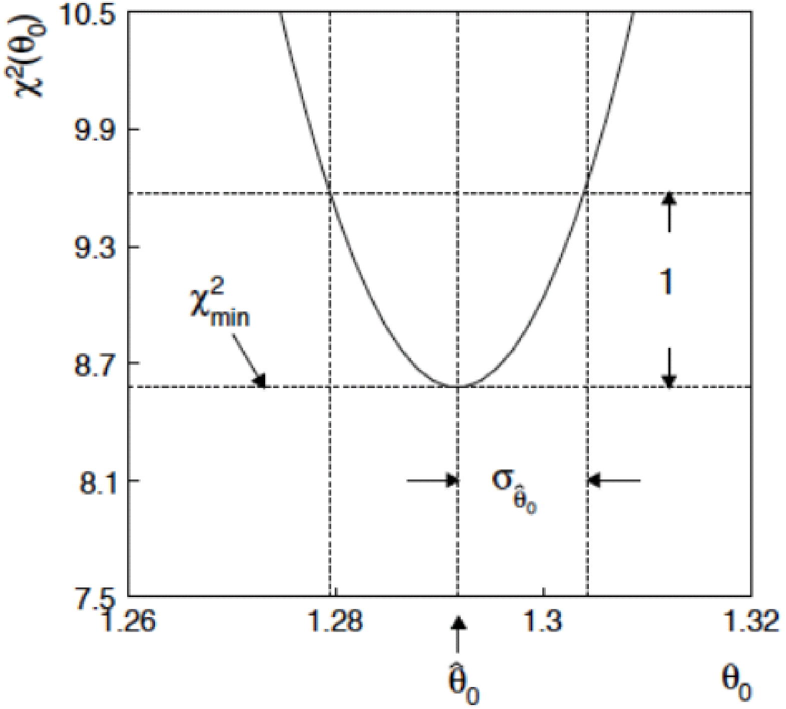
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ^2_{min}

to find $\sigma_{\hat{\theta}_0}$.



ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0,\theta_1) = -2\ln L(\theta_0,\theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} \; .$$

Standard deviations from

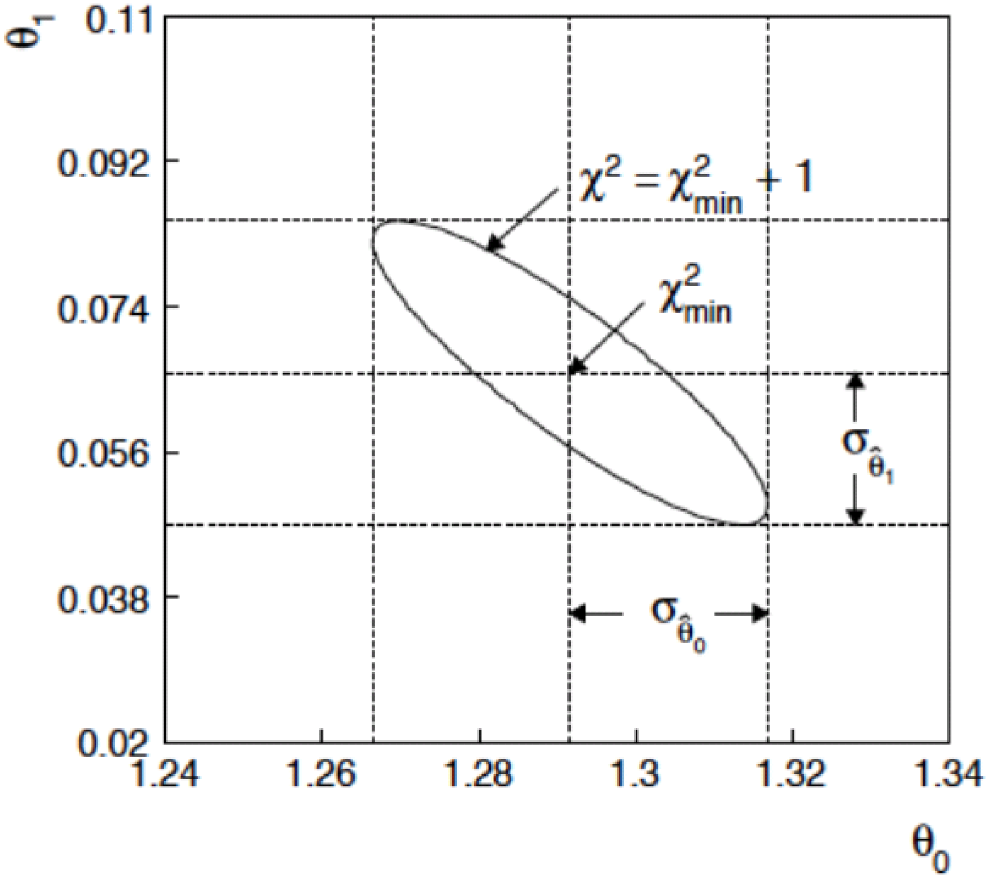
tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 \; .$$

Correlation between

$\hat{\theta}_0, \; \hat{\theta}_1$ causes errors

to increase.

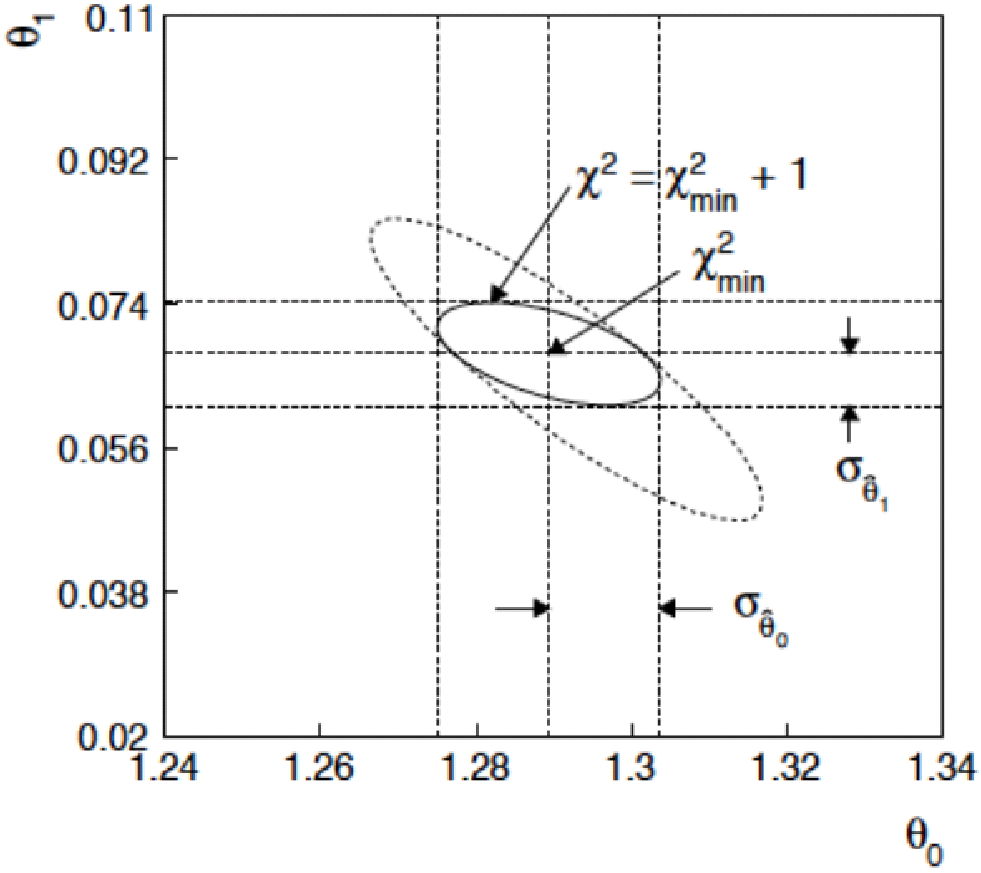


If we have a measurement $t_1 \sim$ Gauss (θ_1, σ_{t_1})

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2} \,.$$

The information on θ_1

improves accuracy of $\hat{\theta}_0$.



October 2013