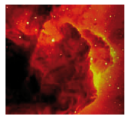
Frequentist Hypothesis Testing

with ATLAS at the LHC

Daniel Rauch

July 9th, 2018 – HEP Students' Seminar

Particles, Strings, and the Early Universe
Collaborative Research Center SFB 676

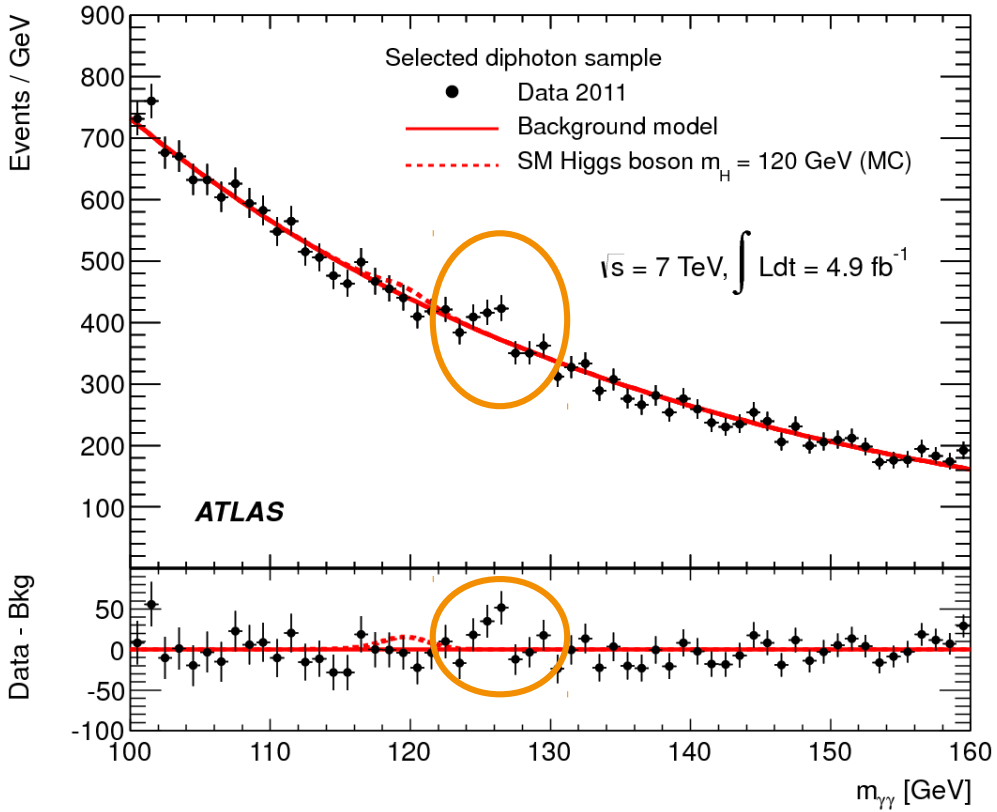


Agenda.

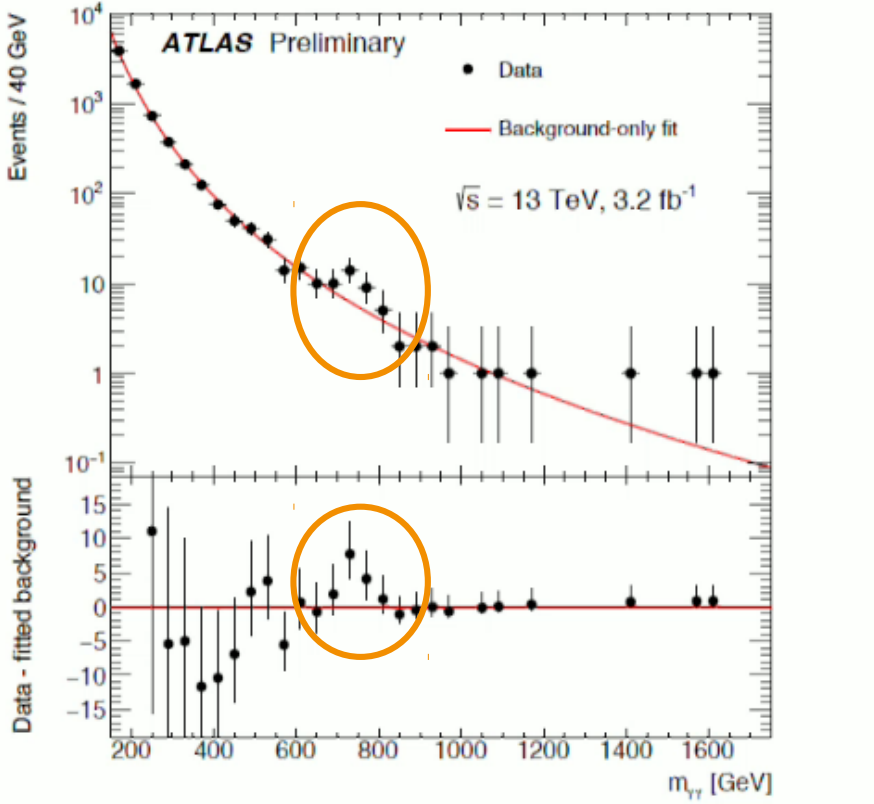
- **Motivation: Why all that?**
- **Theory Intro: Statistics basics**
- **Stats by Guts: Statistics Techniques for Searches** at the LHC
- **Retrospective:** The **Higgs Discovery** in LHC Run 1
- **The Gory Details:** A few **test statistics**, **CLs**, **Feldman-Cousins** and all that
- **And finally: The Look-Elsewhere Effect**

New Physics? Are You Sure? Really?

➔ Or just old physics and ‘bad luck’ conspiring to fake new physics? → **Just how sure are we?**



[arXiv:1202.1414 [hep-ex]]



[ATLAS-CONF-2015-081]

Bayes Theorem and the Likelihood.

- **Conditional probability** $P(A|B)$
- **Bayes' theorem**

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

and therefore also

[wikipedia]

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

likelihood

prior

posterior probability / density

normalisation

- **Keep in mind for later:** $\text{Likelihood} = L(\text{theory}) = P(\text{data}|\text{theory}) \neq P(\text{theory}|\text{data})$
  - The likelihood is a function of the parameters of the theory

Bayesian Approach to Hypothesis Testing in a Nutshell.

- There are two schools of thought: **Frequentism** and **Bayesianism**
  - Frequentist definition of probability: Probability is the relative frequency after an infinite number of trials
  - Bayesian definition of probability: How much money are you willing to bet on the outcome?
- Bayesian hypothesis testing

$$P(\vec{\theta}|\{\vec{x}_i\}) = \frac{P(\{\vec{x}_i\}|\vec{\theta}) \cdot P(\vec{\theta})}{P(\{\vec{x}_i\})} = \frac{P(\{\vec{x}_i\}|\vec{\theta}) \cdot P(\vec{\theta})}{\int P(\{\vec{x}_i\}|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}}$$

theory data

  - Choose priors, use Bayes' theorem and calculate $P(\text{theory}|\text{data}) = P(\vec{\theta}|\{\vec{x}_i\})$
    - **Common criticism:** choice of priors is subjective
  - Find the interval / region where $P(\text{theory}|\text{data})$ is max → **credible intervals**
    - Smallest, central, symmetric intervals / regions possible

Very nice lectures on Bayesian Data Analysis by Christian Graf @ iCSC 2018 → http://cern.ch/go/HC9w

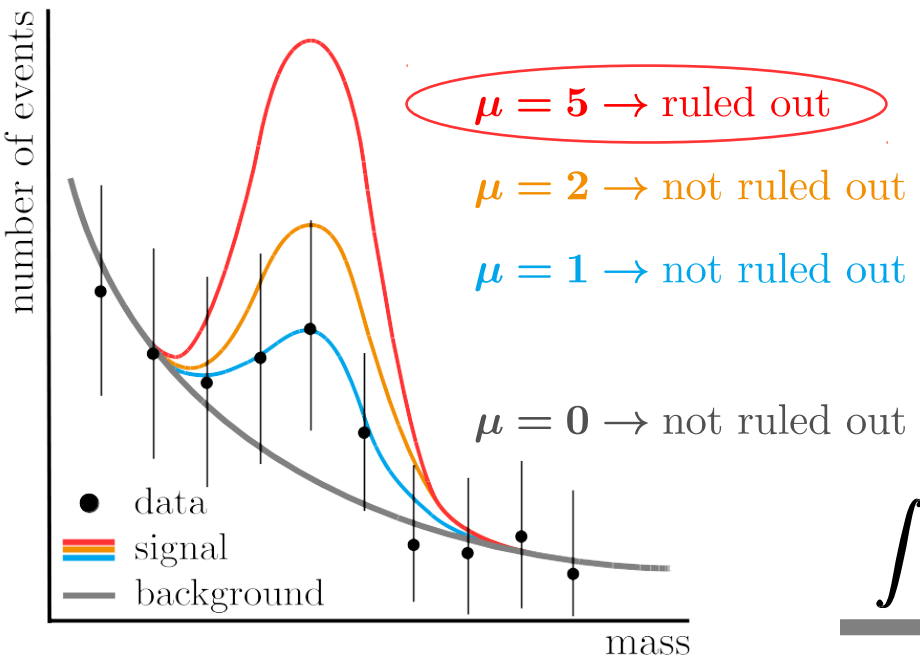
- This procedure, however, is not possible in the frequentist approach! → More complicated conceptually!

Development of a Physics Signal with Increasing Amount of Data.

➔ **Signal strength**

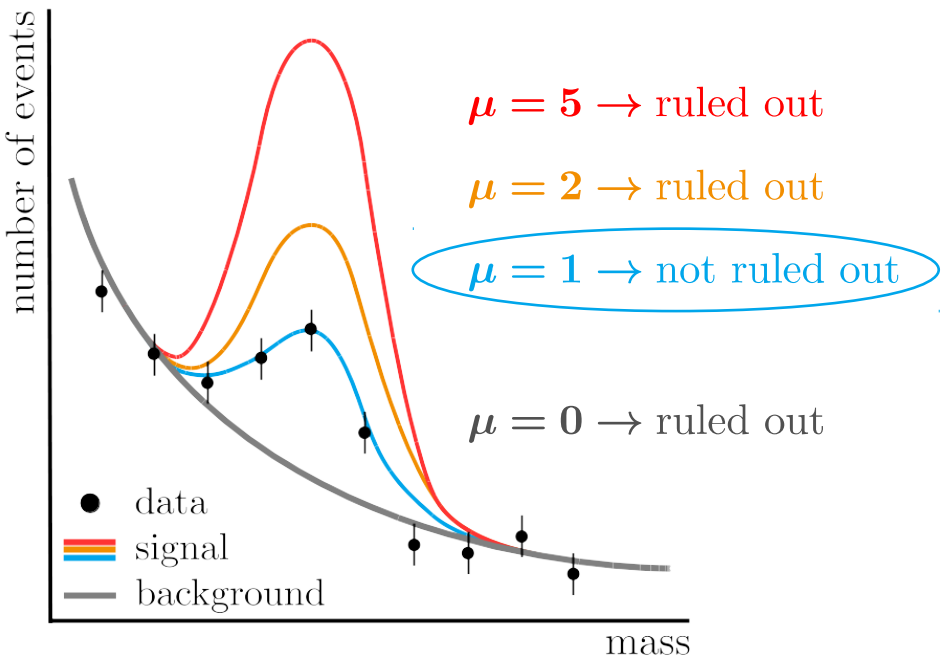
- Global factor that multiplies the signal cross section
- $\mu = 0$: no signal at all
- $\mu = 1$: signal as expected from the theory

$$\mu = \sigma^{(\text{signal})} / \sigma^{(\text{signal})}_{\text{theory}}$$



$\int \mathcal{L} dt$

more data



Discovery vs. Limit Setting vs. Measurement.

Is there a new signal? Or just a back-ground fluctuation?

DISCOVERY

Check compatibility of the data with the background-only hypothesis.

$1\sigma \rightarrow 68.3\%$ $4\sigma \rightarrow 1 - 6.3E\text{-}5$
$2\sigma \rightarrow 95.5\%$ $5\sigma \rightarrow 1 - 5.7E\text{-}7$
$3\sigma \rightarrow 99.7\%$ (in 1 dimension)

Significance: How likely is the background to fake the signal?

no discrepancy

discrepancy

$< 5\sigma$

$\geq 5\sigma$

How strong could the signal be right now while still hiding?

What exactly are we looking at?

LIMIT SETTING

- upper limits: $\mu \leq \mu_{95\%}$
- two-sided limits: $\mu_{95\%}^{(\text{low})} \leq \mu \leq \mu_{95\%}^{(\text{up})}$

$\int \mathcal{L} dt$

more data

MEASUREMENT

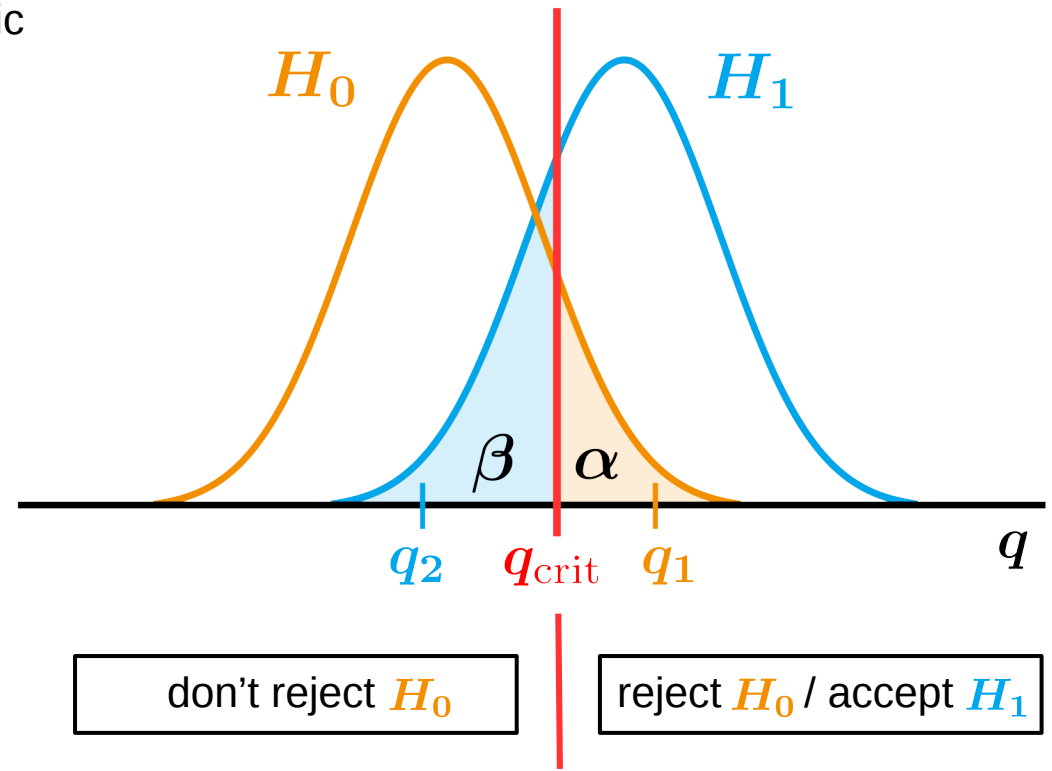
$$\mu = \mu_{\text{best fit}} \; {}^{+\text{stat}}_{-\text{unc}} \; {}^{+\text{sys}}_{-\text{unc}}$$

Basics of Hypothesis Testing.

- H_0: null hypothesis → e.g. standard model H_1: alternative hypothesis → e.g. new physics
  - Can the experimental data be explained with known physics or are we forced to believe in new physics?

- Cut on the **critical value** q_{crit} of a test statistic
  - **Type-1 error** α
    - Wrongly **reject** the null hypothesis
    - $1 - \alpha$ is the **confidence level**
  - **Type-2 error** β
    - Wrongly **accept** the null hypothesis
    - $1 - \beta$ is called **power**

- **Neyman-Pearson lemma**
  - Likelihood ratio is the most powerful test for a desired confidence level $1 - \alpha$



Basics of Hypothesis Testing.

- H_0: null hypothesis → e.g. standard model H_1: alternative hypothesis → e.g. new physics
  - Can the experimental data be explained with known physics or are we forced to believe in new physics?
- **Likelihood function**
  - Fit the model to the data
  - The likelihood function is a measure for how good the model describes the data!

$$L(\mu, \vec{\theta}) = P(\{\vec{x}_i\}|\mu, \vec{\theta})$$

signal strength

data

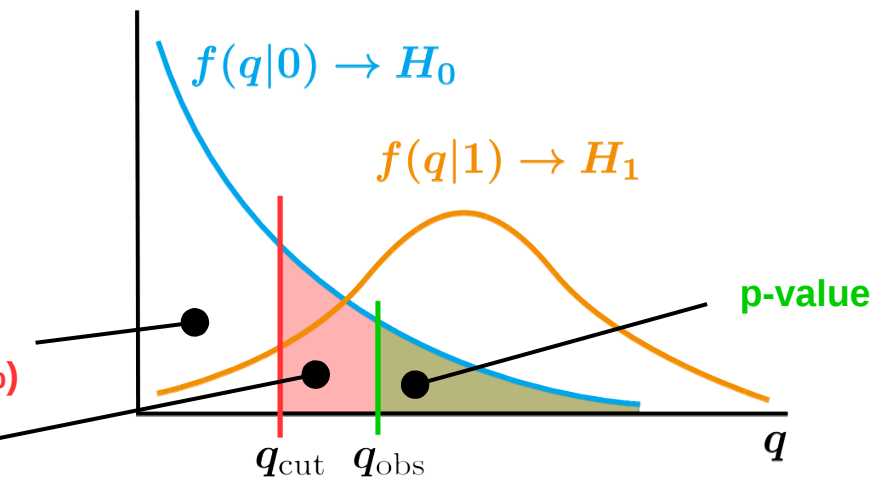
theory & model parameters

- Need **one single number** to compress all information and decide between the two hypotheses → **test statistic**

e.g.

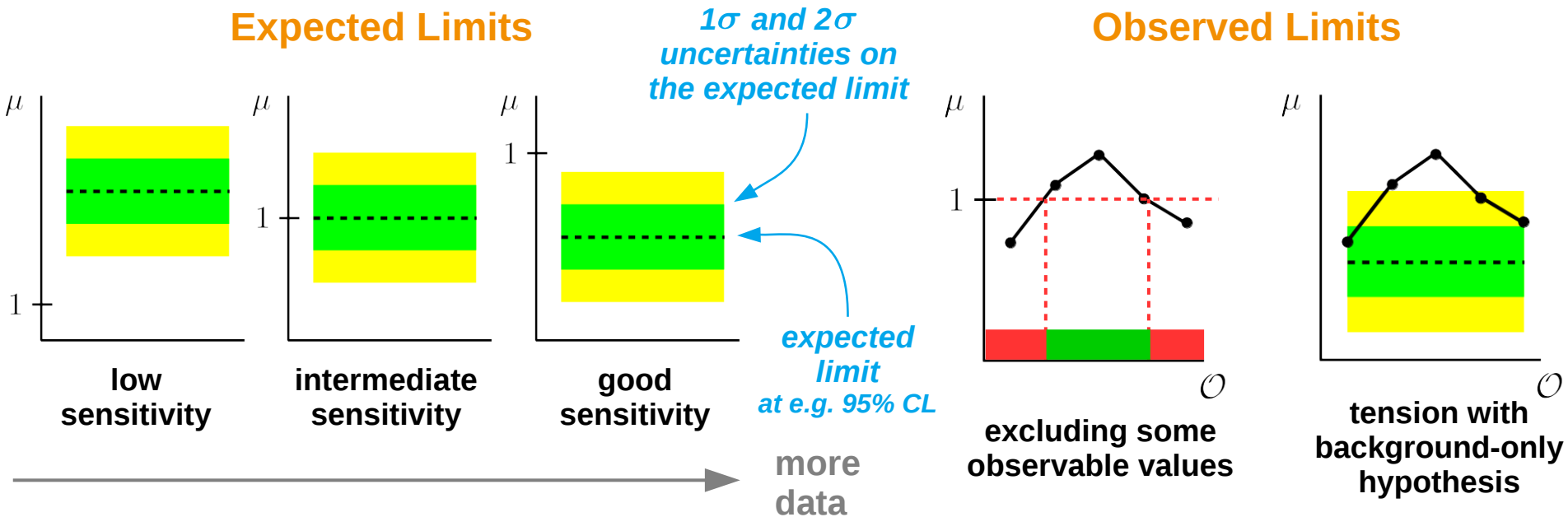
$$q = -2 \ln \frac{L(\{\vec{x}_i\}|0, \hat{\theta}(0))}{L(\{\vec{x}_i\}|\hat{\mu}, \hat{\theta}(\hat{\mu}))}$$

- Good agreement → $q \approx 0$
- Bad agreement → $q \gg 0$



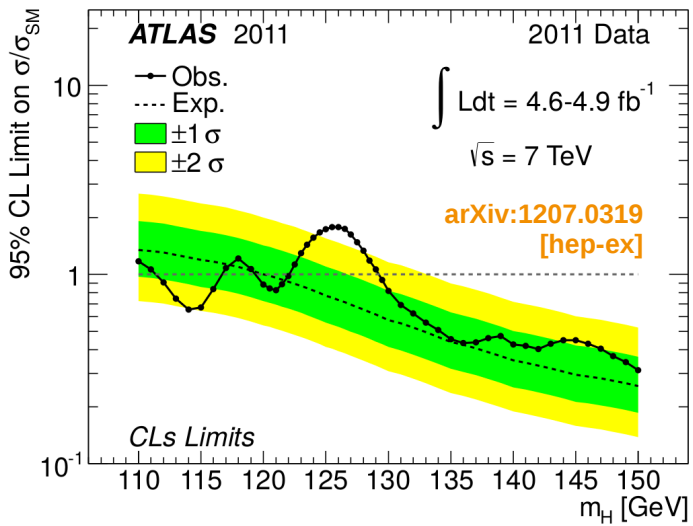
Brazil Plots - From Speculation to Certainty.

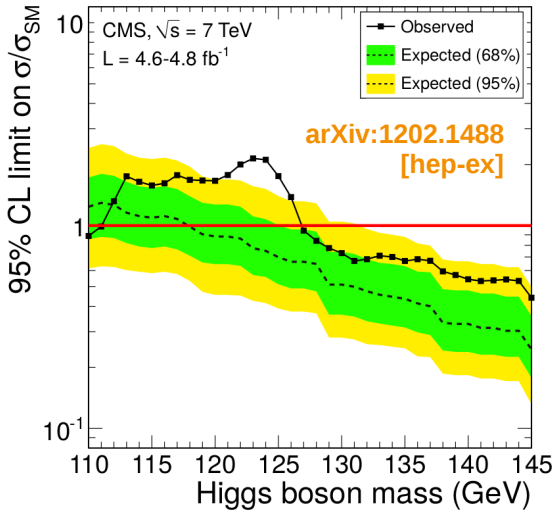
- **Confidence level:** Amount of confidence / trust in the statement that is made
- **p-value:** Chance of an observation at least as extreme as the one that was made to come from a background fluctuation faking the signal
- **Signal strength:** Global factor that multiplies the signal cross section
- **Limits:** Parameter values that mark the transition between the ranges that are allowed / excluded by data

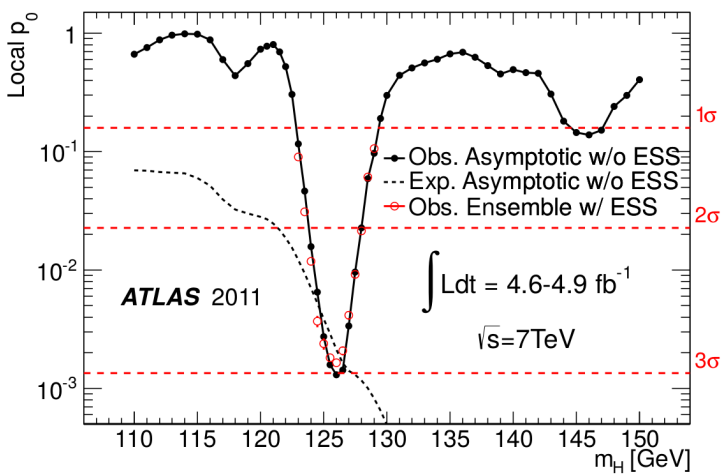


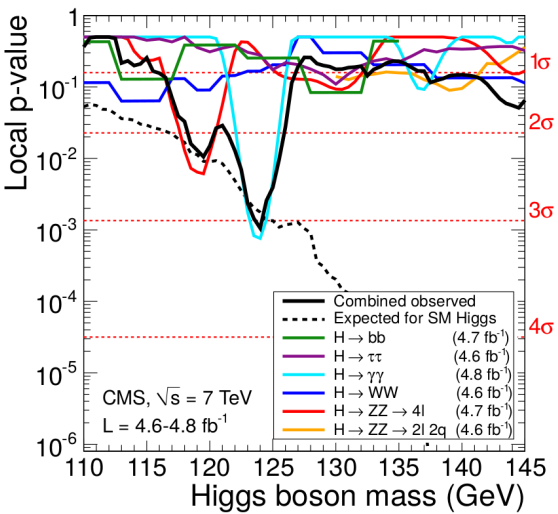
Retrospective: Higgs Searches at the LHC - Full 2011 / 7 TeV Dataset.

- Scan across different values of the mass parameter
- Combination of many different channels
- Dataset
  - Full 2011 / 7 TeV
- Local significances observed (expected)
  - ATLAS: 2.9σ (2.9σ)
  - CMS: 3.1σ $(\approx 2.9\sigma?)$



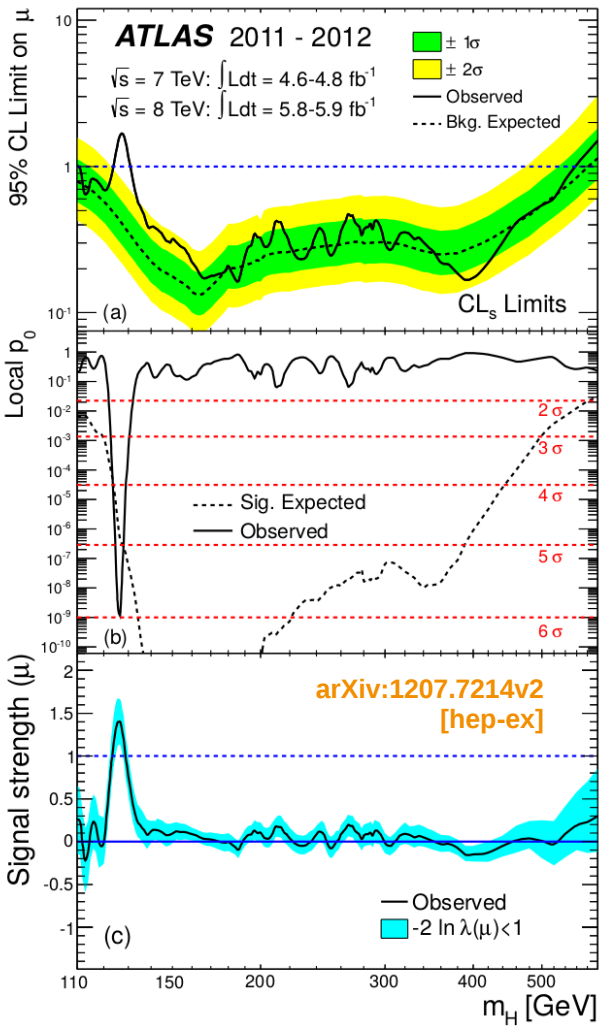


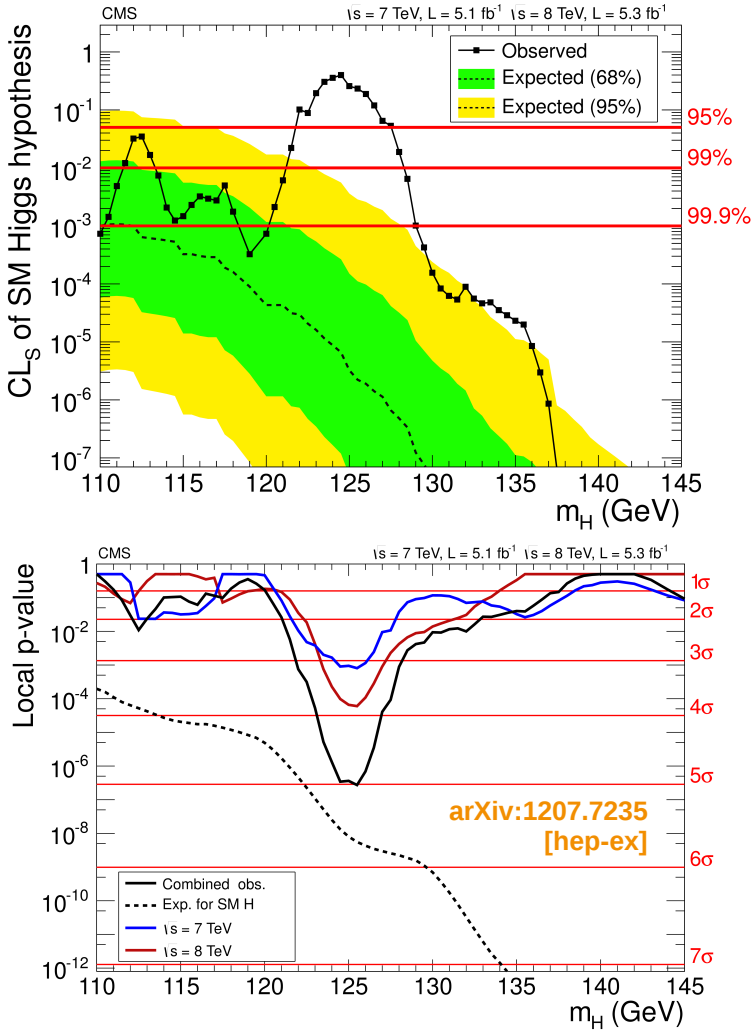




Retrospective: Higgs Searches at the LHC – Discovery in Summer 2012.

- Scan across different values of the mass parameter
- Combination of many different channels
- Dataset
  - Full 7 TeV / 2011
  - First 6/fb of 8 TeV / 2012
- Local significances observed (expected)
  - ATLAS: 5.9σ (4.9σ)
  - CMS: 5.0σ (5.8σ)

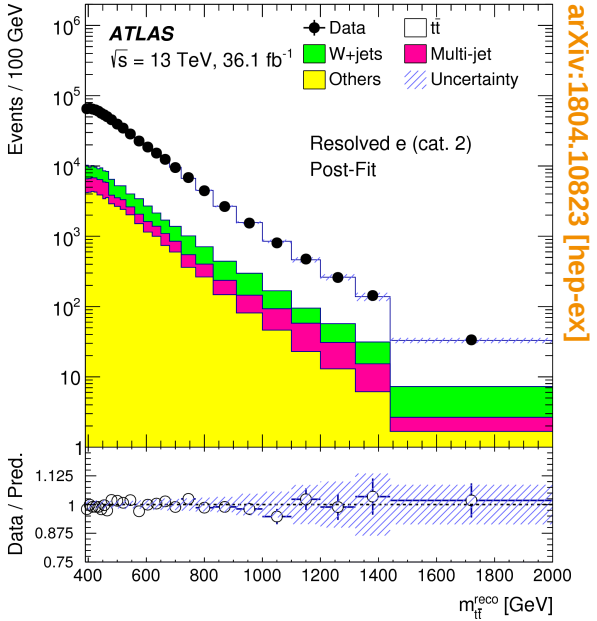




More Vocabulary.

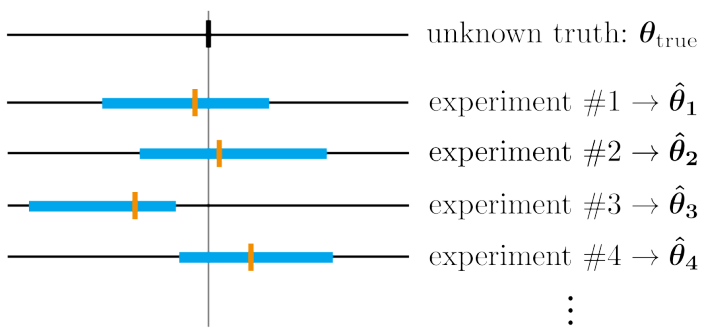
➔ Nuisance parameters

- All parameters other than the parameter of interest (POI), auxiliary parameters that are needed to adjust the model
- Their precise values are not of primary interest
- Make the model more flexible and adjustable
- Reflect our imperfect knowledge / ignorance about many parameters
- e.g. related to luminosity or contributions from different background processes



➔ Coverage

- How often does the measurement contain the true value?
- Property of the statistics method / procedure, not the individual measurement(s)!



More Vocabulary.

➔ **Pull** *(→ "pull distribution")*

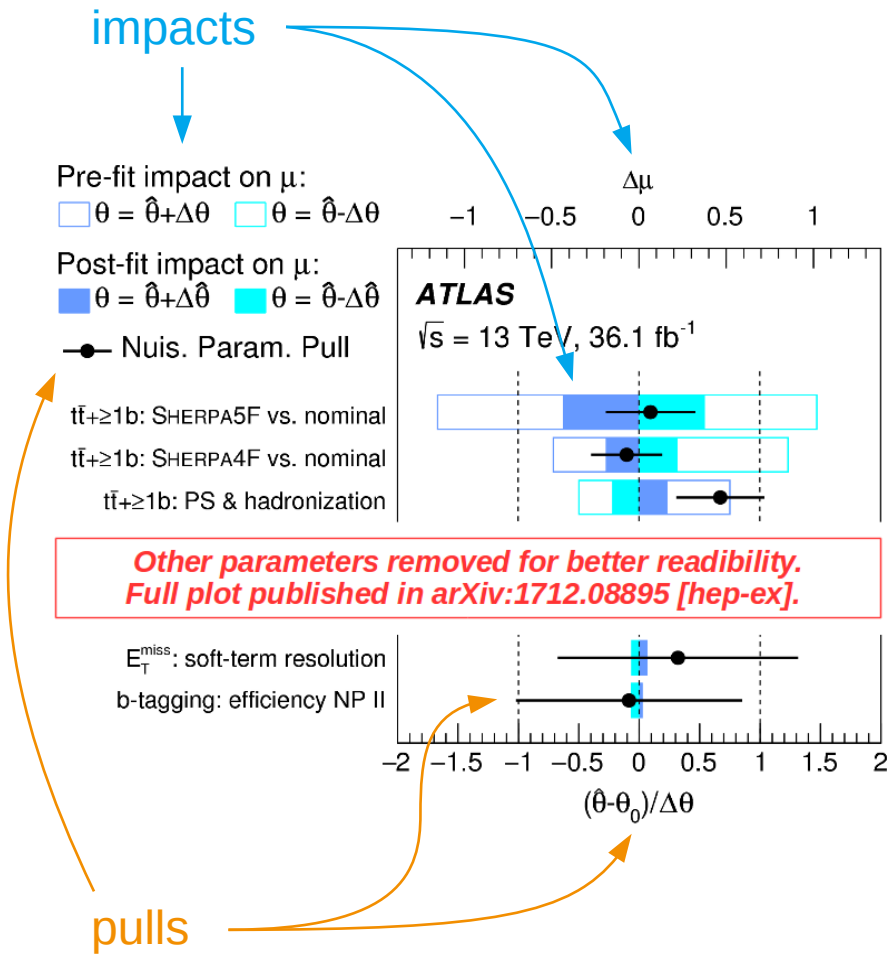
- How were a nuisance parameter and its uncertainty changed by the fit?

➔ **Impact**

- **Pre-fit:** Fix a single parameter at $1\sigma_{\text{prefit}}$ above / below its best-fit value
- **Post-fit:** Fix a single parameter at $1\sigma_{\text{postfit}}$ above / below its best-fit value
- Then fit all remaining parameters and see how POI changes

θ_0

θ_0	Pre-fit parameter value
$\hat{\theta}$	Post-fit parameter value
$\Delta\theta$	Pre-fit parameter uncertainty
$\Delta\hat{\theta}$	Post-fit parameter uncertainty



LHC-Era Test Statistics.

arXiv:1007.1727 [physics.data-an]

- **Idea:** Profiling of nuisance parameters → asymptotic formulae for distributions of test statistics
- Different test statistics for the different use cases

symbol	purpose	rejection region	signal
t_0	discovery	two-sided	+ / -
q_0	discovery	one-sided	+
t_μ	limit setting	two-sided (→ confidence intervals)	+ / -
$\tilde{t}_\mu$	limit setting	two-sided (→ confidence intervals)	+
q_μ	limit setting	one-sided (→ upper limits)	+ / -
$\tilde{q}_\mu$	limit setting	one-sided (→ upper limits)	+

$$t_0 = -2 \ln \frac{L(0, \hat{\theta}(0))}{L(\hat{\mu}, \hat{\theta}(\hat{\mu}))}$$

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

LHC-Era Test Statistics – Example: Discovery.

arXiv:1007.1727 [physics.data-an]

- One-sided / capped test statistic

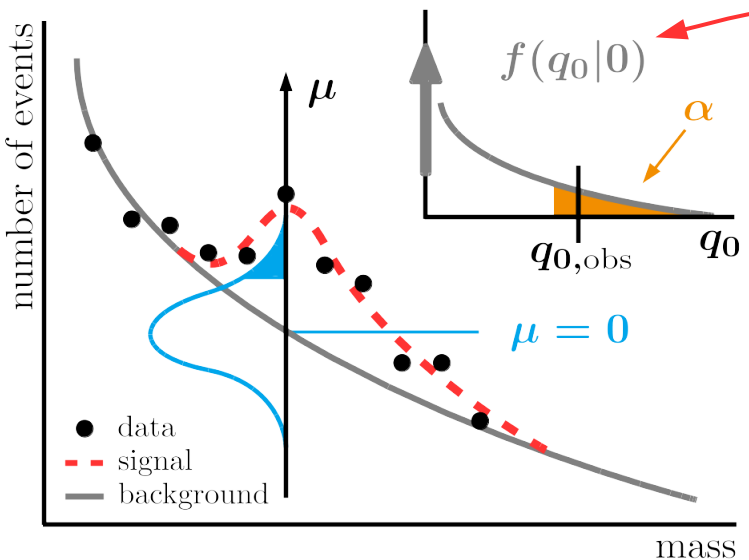
$$q_0 = \begin{cases} -2\ln\frac{L(0,\hat{\theta}(0))}{L(\hat{\mu},\hat{\theta}(\hat{\mu}))} & \text{if } \hat{\mu} \geq 0 \\ 0 & \text{if } \hat{\mu} < 0 \end{cases}$$

  - Only upwards deviations can lead to rejection of the null hypothesis

- Two-sided / uncapped test statistic

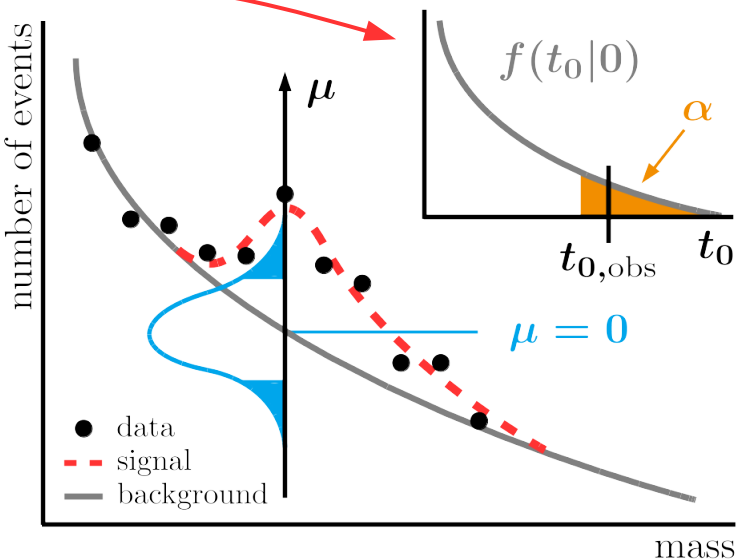
$$t_0 = -2\ln\frac{L(0,\hat{\theta}(0))}{L(\hat{\mu},\hat{\theta}(\hat{\mu}))}$$

  - Both upwards and downwards deviations can lead to rejection of the null hypothesis



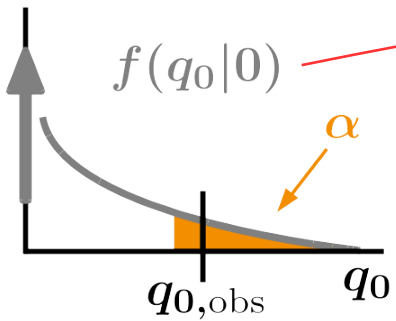
analytically known!*

* in the large sample limit



LHC-Era Test Statistics – Example: Discovery.

- **p-Value:** What is the probability to see an excess of events at least as large as the one we observe in the absence of the signal, i.e. just from a background fluctuation?



integration variable

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

observed value of the test statistic

$$q_{0,\mathrm{obs}} = \begin{cases} -2 \ln \frac{L(0,\hat{\theta}(0))}{L(\hat{\mu},\hat{\theta}(\hat{\mu}))} & \text{if } \hat{\mu} \geq 0 \\ 0 & \text{if } \hat{\mu} < 0 \end{cases}$$

- if $p_0 \geq \alpha$ → don't reject H_0
- if $p_0 < \alpha$ → reject H_0/ accept H_1

LHC-Era Test Statistics – Example: Upper Limits.

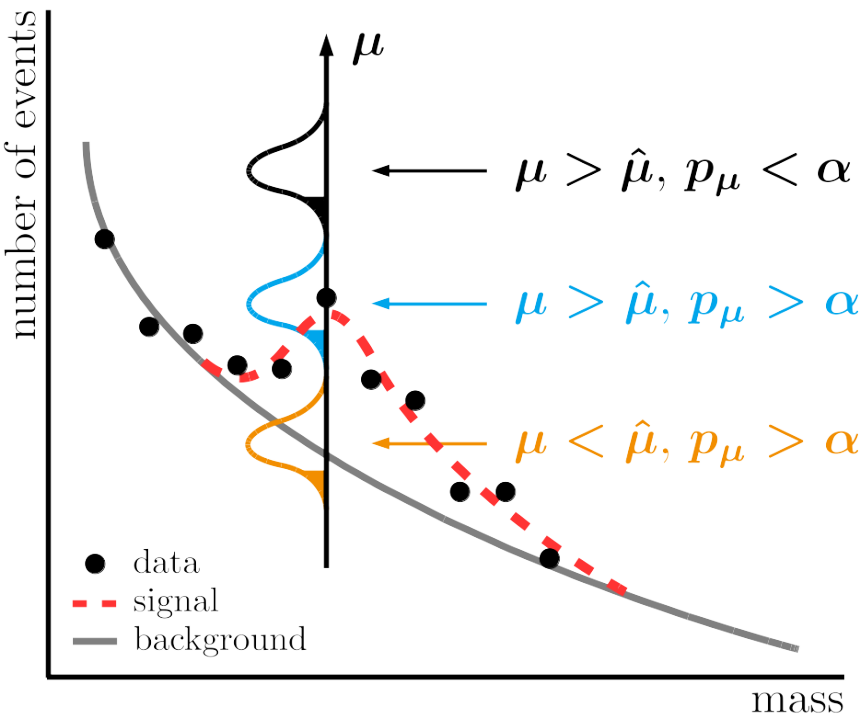
arXiv:1007.1727 [physics.data-an]

- One-sided rejection region

$$q_\mu = \begin{cases} -2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta}(\hat{\mu}))} & \text{if } \mu \geq \hat{\mu} \\ 0 & \text{if } \mu < \hat{\mu} \end{cases}$$

  - Reject null hypothesis if hypothesised μ is significantly larger than best-fit $\hat{\mu}$
  - Sublety
    - Here the signal may be both positive or negative
    - This may lead to the exclusion of the null hypothesis if downwards fluctuation of the background

- Try different hypothetical signal strengths μ until the transition between rejecting and accepting the null hypothesis is found
- This then is the measured upper limit



Feldman-Cousins and CLs.

- Motivation / problem
  - When the expected sensitivity is very low, a downward fluctuation of the background may result in rejection of the null hypothesis → false signal claim
  - This is precisely the sublety mentioned on the previous slide!

- The method by Feldman and Cousins
  - Restrict the fitted signal strength to $\hat{\mu} \geq 0$ when evaluating the likelihood ratio

$$\tilde{q}_\mu = \begin{cases} -2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta}(\hat{\mu}))} & \text{if } \mu \geq \hat{\mu} \text{ and } \hat{\mu} \geq 0 \\ \boldsymbol{-2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(0, \hat{\theta}(0))}} & \text{if } \mu \geq \hat{\mu} \text{ and } \boldsymbol{\hat{\mu} < 0} \\ 0 & \text{if } \mu < \hat{\mu} \end{cases}$$

- The CLs method
  - Increase the p-value to make rejection of the null hypothesis less likely
  - Reject the null hypothesis if

$$\mathrm{CL_s} = \frac{\mathrm{CL_{s+b}}}{\mathrm{CL_b}} < \alpha$$

with

$$\mathrm{CL_{s+b}} = p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q'_\mu|\mu)\, dq'_\mu$$

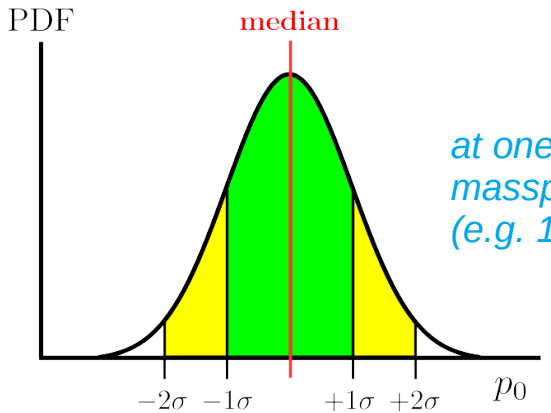
$$\mathrm{CL_b} = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q'_\mu|0)\, dq'_\mu$$

Expected Limits.

- **So far:** Only **observed** p-values and limits
  - Only need $f(q_\mu|\mu)$, but not $f(q_\mu|\mu' \neq \mu)$

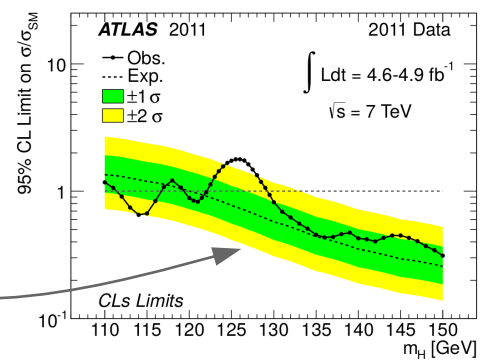
- **Idea:** Generate pseudo data / toys, find median p-values as well as 68% and 95% percentiles
  - Expected sensitivity for discovery
    - In the absence of signal: $p_0 = \int_{q_{0,\mathrm{toy}}}^{\infty} f(q_0|\mathbf{0})dq_0$
    - In the presence of signal: $p_0 = \int_{q_{0,\mathrm{toy}}}^{\infty} f(q_0|\mathbf{1})dq_0$
  - Expected upper limits:
    - In the absence of signal: $p_\mu = \int_{q_{\mu,\mathrm{toy}}}^{\infty} f(q_\mu|\mathbf{0})dq_\mu$
    - In the presence of signal: $p_\mu = \int_{q_{\mu,\mathrm{toy}}}^{\infty} f(q_\mu|\mathbf{1})dq_\mu$

- **Now: Expected** p-values and limits
  - Now also need $f(q_\mu|\mu' \neq \mu)$



at one fixed masspoint (e.g. 125 GeV)

scan across the entire search range



arXiv:1207.0319 [hep-ex]

Recap / At One Glance.

Discovery

- Observed
  - Decide on confidence level CL, e.g. 5σ
  - Calculate observed $q_{0,\mathrm{obs}}$ from data
  - Get $f(q_0|0)$ from asymptotic formulae or toy experiments
  - Calculate $p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0'|0)dq_0'$
  - If $p_0 \geq \alpha \rightarrow$ don't reject H_0,
    if $p_0 < \alpha \rightarrow$ reject H_0 and accept H_1

Limits (example: upper limits)

- Observed
  - Decide on confidence level CL, e.g. 95%
  - Hypothesise a signal strength μ
    - Calculate observed $q_{\mu,\mathrm{obs}}$ from data
    - Get $f(q_\mu|\mu)$ from asymptotic formulae or toy experiments
    - Calculate $p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu'|\mu)dq_\mu'$
    - If $p_\mu \geq \alpha = 1 - \mathrm{CL} \rightarrow$ don't reject H_0 ,
      if $p_\mu < \alpha \rightarrow$ reject H_0 and accept H_1
  - Repeat for different values of μ
  - Find the transition between rejection and non-rejection of $H_0 \rightarrow$ this is the upper limit

Possibly repeat for different parameters of the theory (e.g. Higgs masses)

Recap / At One Glance.

Discovery

- Expected
  - Decide on confidence level CL, e.g. 5σ
  - Get $f(q_0|\mu)$ from asymptotic formulae or toy experiments, usually for $\mu = 0$ or 1
  - Generate toy data set for the same μ
    - Calculate observed $q_{0,\text{toy}}$ from data
    - Calculate $p_0 = \int_{q_{0,\text{toy}}}^{\infty} f(q_0'|\mu) dq_0'$
    - Enter p_0 into histogram
  - Generate more toy experiments and repeat
  - Get median p_0 (expected sensitivity) as well as the 68% and 95% percentiles (statistical uncertainty bands)

Limits (example: upper limits)

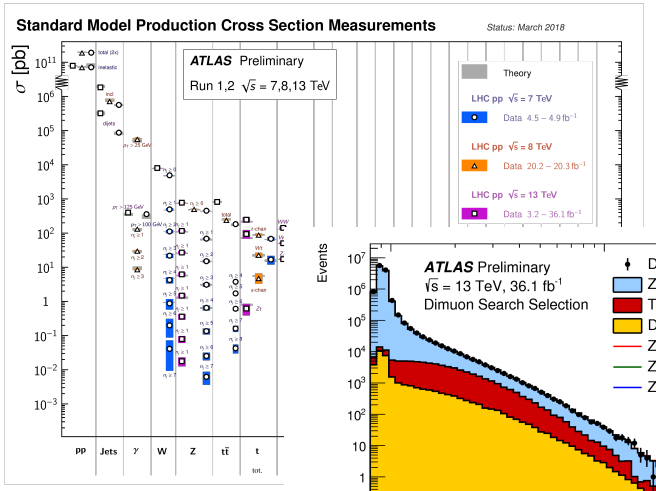
- Expected
  - Decide on confidence level CL, e.g. 95%
  - Get $f(q_\mu|\mu')$ from asymptotic formulae or toy experiments, usually for $\mu' = 0$ or 1
  - Generate toy data set for the same μ'
    - Find the upper limit $\mu_{95\%}$ for this toy data set as explained on the previous slide
    - Enter $\mu_{95\%}$ into histogram
  - Generate more toy experiments and repeat
  - Get median $\mu_{95\%}$ (expected upper limit) as well as the 68% and 95% percentiles (statistical uncertainty bands)

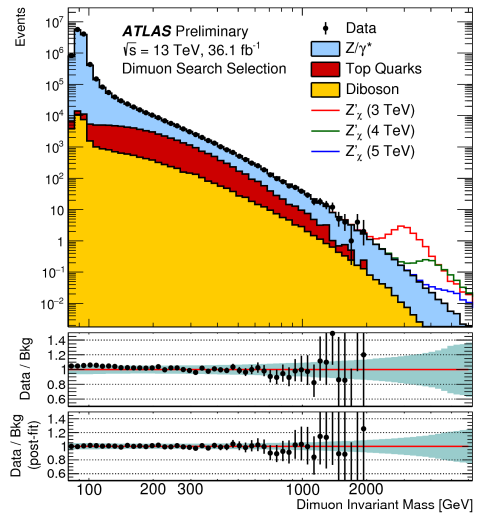
Possibly repeat for different parameters of the theory (e.g. Higgs masses)

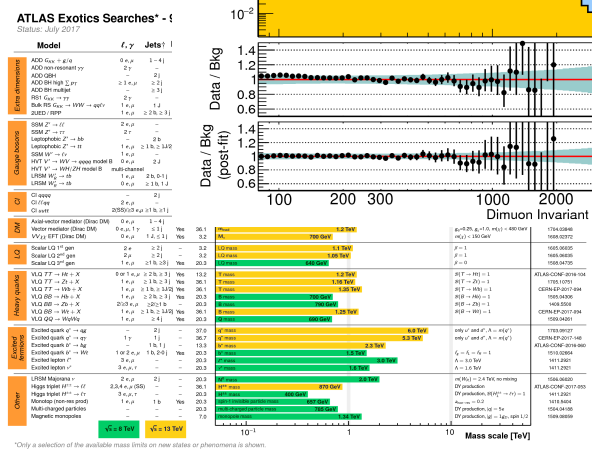
The Look-Elsewhere Effect.

- Plethora of measurements, searches and bins at the LHC
  - We are bound to see “significant” excesses somewhere!
    - If 100 bins, 5 bins should deviate by 2 standard deviations!
  - What do we really consider to be “significant”?
- **Idea:** Calculate a reduced global significance
- Dunn-Šidák correction
  - Assume all bins to be uncorrelated
  - Recall: $\alpha = P(\text{type-1 error})$
  - Define: $\alpha_{\text{global}} = P(\text{at least one type-1 error somewhere})$

    $\rightarrow\ \alpha_{\text{global}} = 1 - (1-\alpha)^n$ for n bins
  - Not commonly done in practice



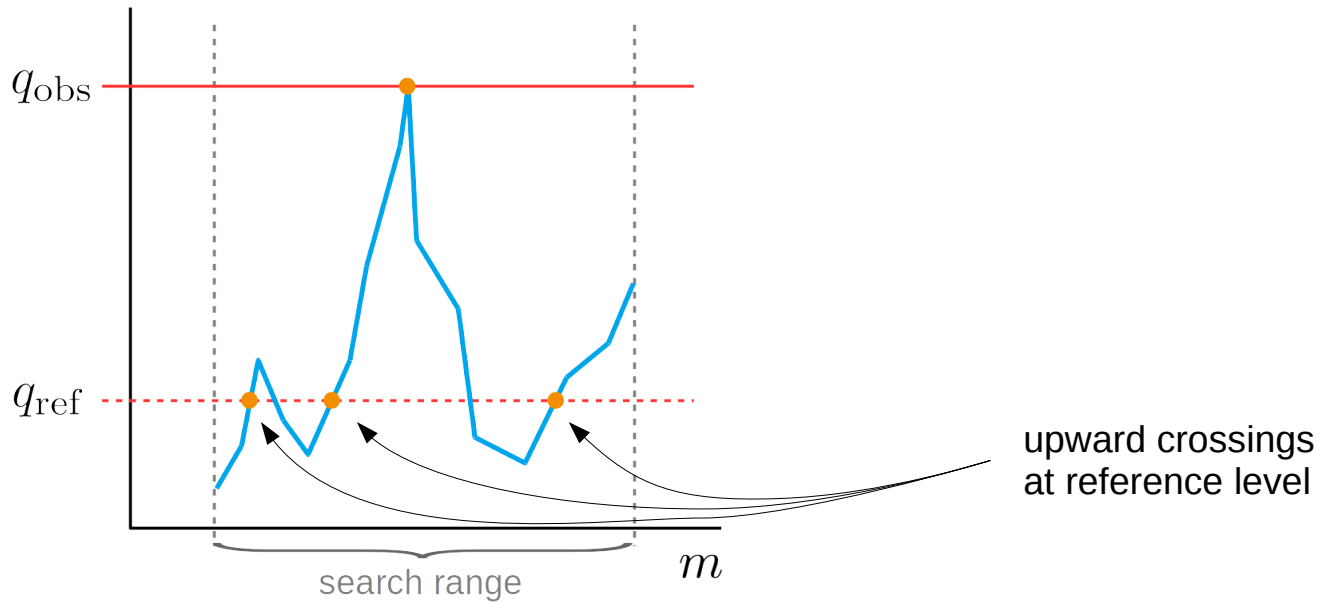




The Look-Elsewhere Effect.

arXiv:1005.1891 [physics.data-an]

- **Brute force approach:** Generate random experiments / pseudo data
  - Very inefficient if done “naively” as one is precisely interested in very rare cases
- **Idea:** Get number of upcrossings at lower reference level and scale it according to the analytically known asymptotic behaviour of $f(q)$
  - Much less random experiments needed at lower reference level



Summary.

- LHC test statistics
  - Different test statistics for discovery, upper and two-sided limits
  - Based on likelihood ratios and profiling of the nuisance parameters
  - Known closed-form asymptotic behaviour in the large sample limit
- Feldman-Cousins and CLs methods
  - Modifications to protect against wrongly rejecting the null hypothesis in case of a downward fluctuation of the background at low sensitivity
- Look-Elsewhere Effect
  - The more bins, the more likely it is to find a seemingly “significant” excess somewhere
  - Derive a reduced global significance that takes into account the number of bins / search range

Many thanks for your attention!