

Statistics for LHC searches

23.02.2017 - IMPRS PPSMC
Philipp Gadow | Max-Planck-Institut für Physik, München

Statistics for LHC searches

It is often said that the language of science is mathematics. It could be well said that the language of experimental science is statistics.

— Kyle Cranmer

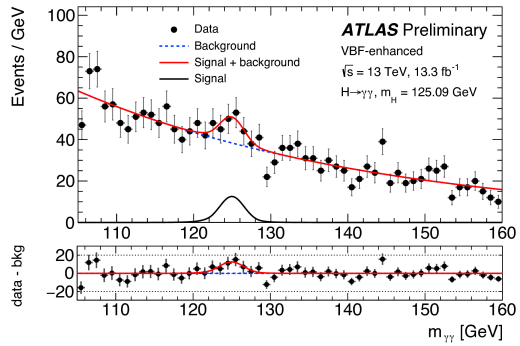
Structure of this talk:

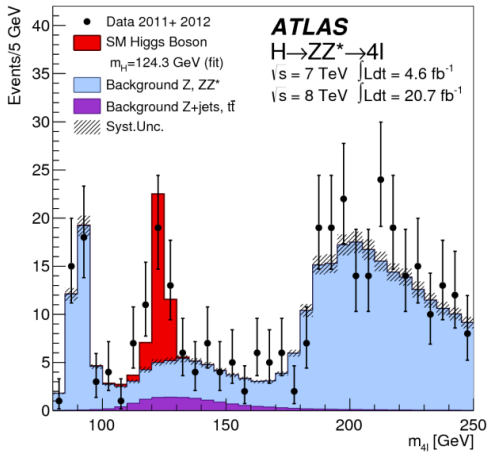
- Scientific narrative $\leftrightarrow$ model building
- Discovery $\leftrightarrow$ hypothesis tests
- Exclusion limits $\leftrightarrow$ confidence intervals

The scientific narrative

How would you explain your HEP analysis over lunch? You tell a story about signal and background:

- What do you want to find out?
- Analysis strategy (event counting, discriminating variable, ...)
- Estimation of backgrounds (MC simulations, side-band, data-driven technique, ...)
- Dominant uncertainties in the rate and shape of signal and background





A statistical model is the mathematical representation of the scientific narrative behind an analysis.

Model building

- Start model building with experiment and physical theory.
- **Model parameters α represent parameters of a physical theory or unknown properties of a detector's response.**
- Distinguish between parameters of interest μ and nuisance parameters θ, meaning that $\alpha = \{\mu, \theta\}$

Consider parametric family of probability density functions:
probability model $f(x|\alpha)$

parameter α fixed:	random variable x fixed:
probability density function	likelihood function

Model building

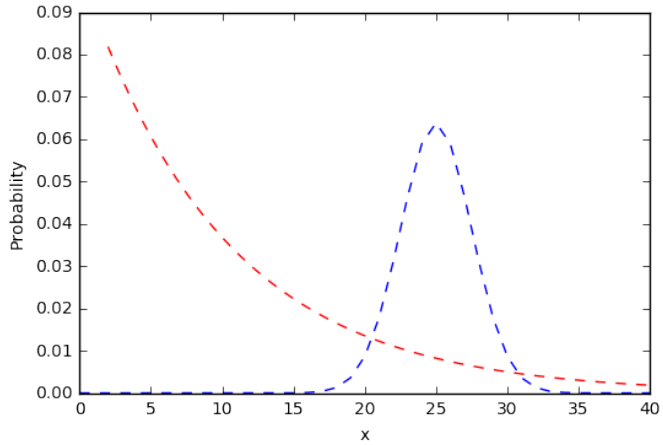
Example: Look for an excess in invariant mass spectrum (signal) that is not explained by Standard Model physics (background)

- Signal model $s := \{x_0, \sigma\}$:

$$f(x|s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x - x_0)^2/2\sigma^2\right)$$

- Background model $b := \{\Gamma\}$:

$$f(x|b) = \frac{\exp\left(-x/\Gamma\right)}{\Gamma}$$

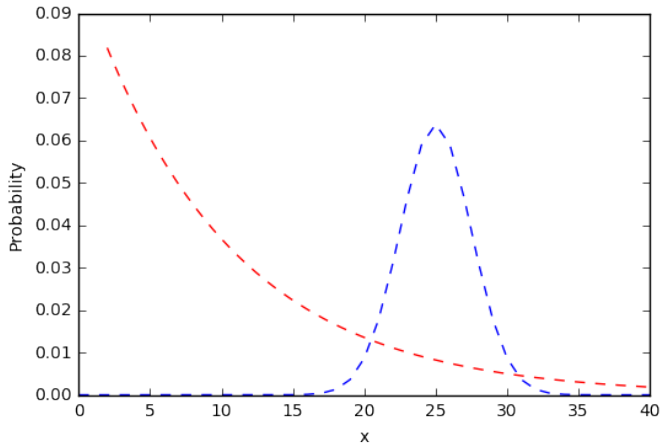


In reality: $f(x|b)$ and $f(x|s)$ are seldom analytic functions but determined by complex MC simulation.

Model building

Probability model: mixture of signal and background

- Signal strength parameterised by μ
- Expected events: $\nu(\mu) = \mu\nu_s + \nu_b$
  (ν_s, ν_b fixed by model)



$$f(x|\mu) = \frac{\mu\nu_s}{\mu\nu_s + \nu_b} f(x|s) + \frac{\nu_b}{\mu\nu_s + \nu_s} f(x|b)$$

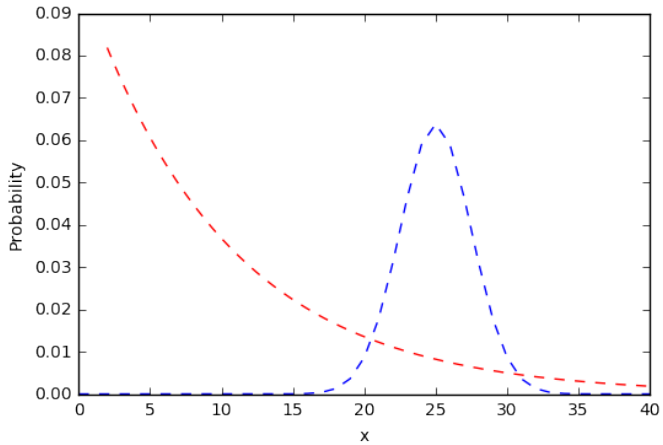
Probability distribution for single event:

$$f(x|\mu)$$

Model building

Probability model: mixture of signal and background

- Signal strength parameterised by μ
- Expected events: $\nu(\mu) = \mu\nu_s + \nu_b$
  (ν_s, ν_b fixed by model)



$$f(x|\mu) = \frac{\mu\nu_s}{\mu\nu_s + \nu_b} f(x|s) + \frac{\nu_b}{\mu\nu_s + \nu_s} f(x|b)$$

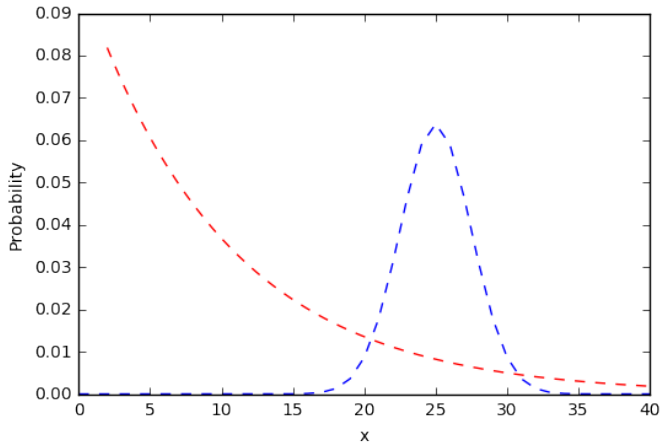
Data set $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ of n events from same distribution:

$$f(\mathcal{D}|\mu) = \prod_{e=1}^{n} f(x_e|\mu)$$

Model building

Probability model: mixture of signal and background

- Signal strength parameterised by μ
- Expected events: $\nu(\mu) = \mu\nu_s + \nu_b$
  (ν_s, ν_b fixed by model)



$$f(x|\mu) = \frac{\mu\nu_s}{\mu\nu_s + \nu_b} f(x|s) + \frac{\nu_b}{\mu\nu_s + \nu_s} f(x|b)$$

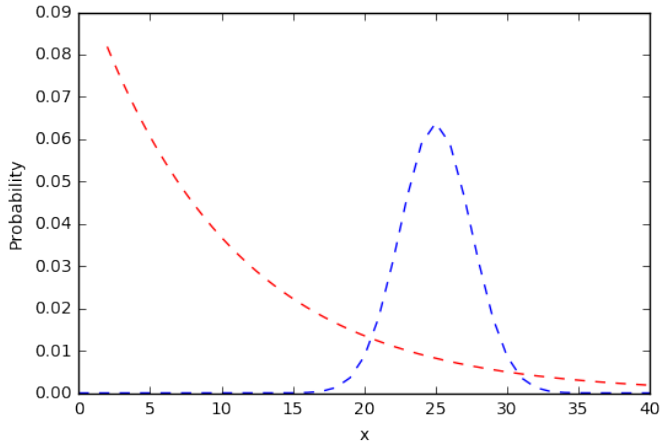
Expected number of events $\nu(\mu) = \mu\nu_s + \nu_b$ is specified by theory:

$$f(\mathcal{D}|\mu) = \text{Pois}(n|\nu(\mu)) \prod_{e=1}^{n} f(x_e|\mu)$$

Model building

Probability model: mixture of signal and background

- Signal strength parameterised by μ
- Expected events: $\nu(\mu) = \mu\nu_s + \nu_b$
  (ν_s, ν_b fixed by model)



$$f(x|\mu) = \frac{\mu\nu_s}{\mu\nu_s + \nu_b} f(x|s) + \frac{\nu_b}{\mu\nu_s + \nu_s} f(x|b)$$

Expected number of events $\nu(\mu) = \mu\nu_s + \nu_b$ is specified by theory:

$$f(\mathcal{D}|\mu) = \text{Pois}(n|\nu(\mu)) \prod_{e=1}^{n} f(x_e|\mu)$$

Marked Poisson model

Model building

Dealing with systematic uncertainty

Problem: background modeling could have systematic uncertainty on number of background events ν_b, model as nuisance parameter

$$f(\mathcal{D}|\mu) = \text{Pois}(n|\nu(\mu)) \prod_{e=1}^{n} f(x_e|\mu)$$

Model building

Dealing with systematic uncertainty

Problem: background modeling could have systematic uncertainty on number of background events ν_b, model as nuisance parameter

$$f(\mathcal{D}|\mu,\nu_b) = \text{Pois}(n|\nu(\mu,\nu_b)) \prod_{e=1}^{n} f(x_e|\mu,\nu_b)$$

Model building

Dealing with systematic uncertainty

Problem: background modeling could have systematic uncertainty on number of background events ν_b, model as nuisance parameter

$$f(\mathcal{D}|\mu,\nu_b) = \text{Pois}(n|\nu(\mu,\nu_b)) \prod_{e=1}^{n} f(x_e|\mu,\nu_b)$$

Solution: auxiliary measurement to constrain ν_b

- Define some control sample with n_{CR} events.
- Assume that control region has no signal contamination and is populated by background process

Model building

Dealing with systematic uncertainty

Problem: background modeling could have systematic uncertainty on number of background events ν_b, model as nuisance parameter

$$f(\mathcal{D}|\mu,\nu_b) = \text{Pois}(n|\nu(\mu,\nu_b)) \prod_{e=1}^{n} f(x_e|\mu,\nu_b) \cdot \text{Pois}(n_{\text{CR}}|\nu_b)$$

Solution: auxiliary measurement to constrain ν_b

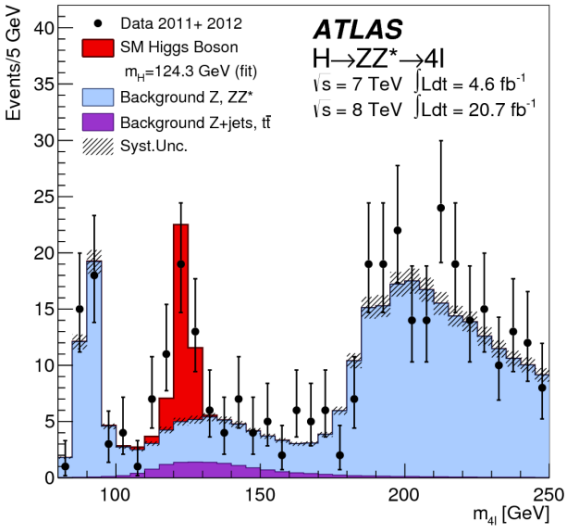
- Define some control sample with n_{CR} events.
- Assume that control region has no signal contamination and is populated by background process
- Add factor $\text{Pois}(n_{\text{CR}}|\nu_b)$ to probability model

Model building

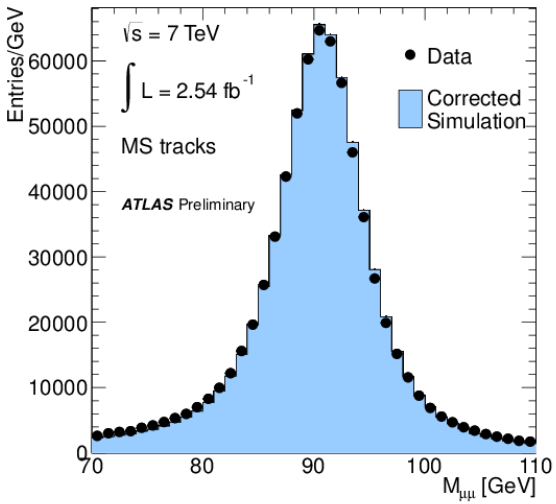
Dealing with systematic uncertainty

$$f(\mathcal{D}|\mu,\nu_b) = \text{Pois}(n|\nu(\mu,\nu_b)) \prod_{e=1}^{n} f(x_e|\mu,\nu_b) \cdot \text{Pois}(n_{\text{CR}}|\nu_b)$$

Signal + Background



Control Region



Model building

Dealing with systematic uncertainty

$$f_{\text{sim}}(\mathcal{D}_{\text{sim}}|\mu,\theta) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu(\mu,\theta)) \prod_{e=1}^{n_c} f(x_{ce}|\mu,\theta) \right]$$

- In practice control regions are modeled as additional channels to constrain nuisance parameters θ

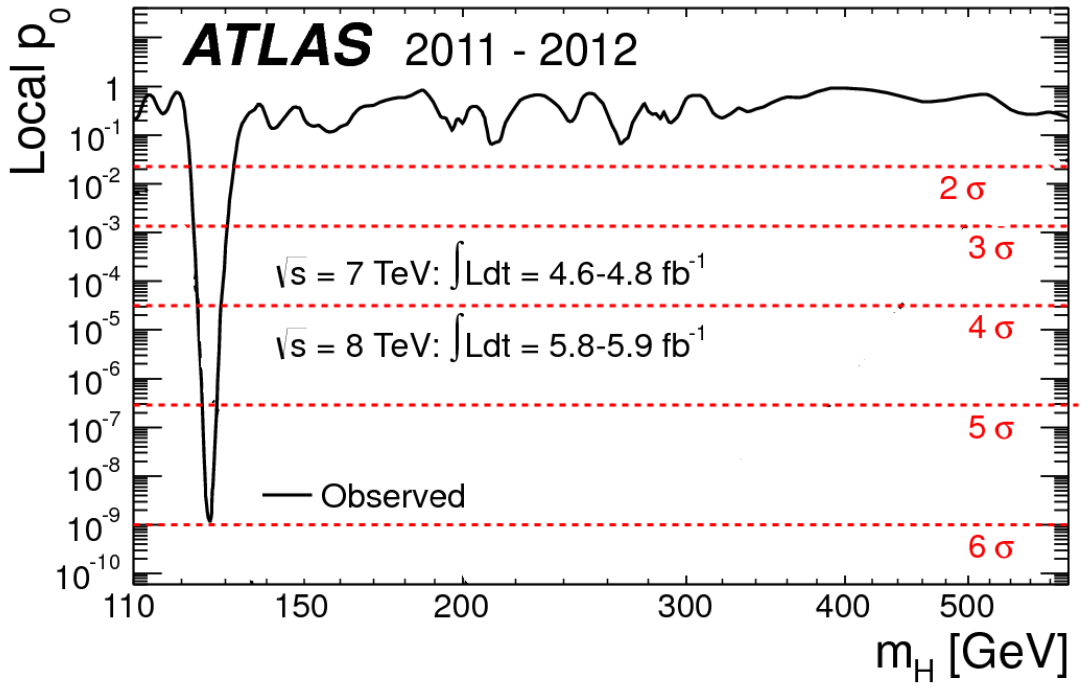
Model building

Dealing with systematic uncertainty

$$f_{\text{tot}}(\mathcal{D}_{\text{sim}}|\mu,\theta) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu(\mu,\theta)) \prod_{e=1}^{n_c} f(x_{ce}|\mu,\theta) \right] \cdot \prod_{p\in\mathbb{S}} f_p(a_p|\theta_p)$$

- In practice control regions are modeled as additional channels to constrain nuisance parameters θ
- When auxiliary measurements can't be included in model, use idealized constraint terms with central value a_p and uncertainty.
- They are included as $f(a_p|\theta_p)$ in the model for a set of parameters with constraint terms $\mathbb{S}$

Discovery as hypothesis test



They claimed that by combining two data sets, they had attained a confidence level just at the "five-sigma"point - about a one-in-3.5 million chance that the signal they see would appear if there were no Higgs particle.
— Paul Rincoln, BBC

Discovery as hypothesis test

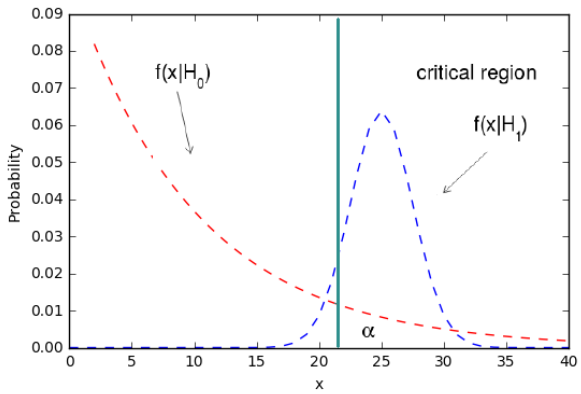
Typical search: look for signal on top of background
This leads to two competing hypotheses:

symbol	statistical name	physics name	probability model
H_0	null hypothesis	background-only	$\mathrm{Pois}(n_{SR}\|\nu_\mathrm{B})$
H_1	alternate hypothesis	signal-plus-background	$\mathrm{Pois}(n_{SR}\|\nu_\mathrm{S}+\nu_\mathrm{B})$

***p*-value**: probability that background-only hypothesis would produce at least n_0 events in critical region

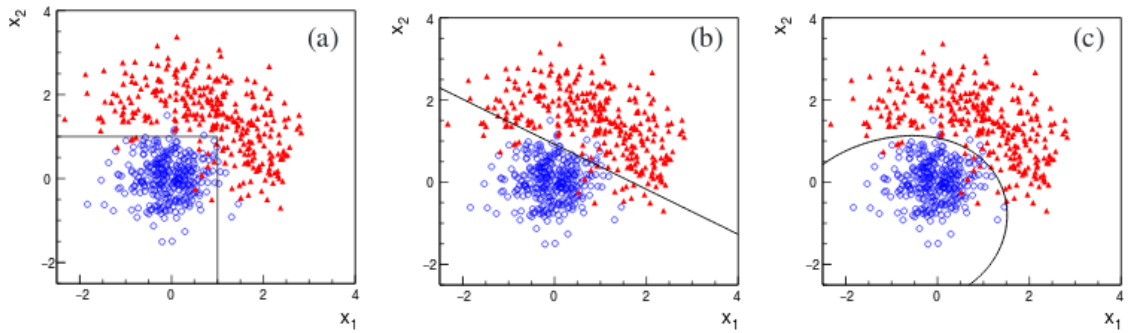
$$p = \sum_{n=n_0}^{\infty} \mathrm{Pois}(n|\nu_\mathrm{B})$$

If p-value is smaller than a certain threshold α (e.g. 2.9×10^{-7}) one rejects the null hypothesis.

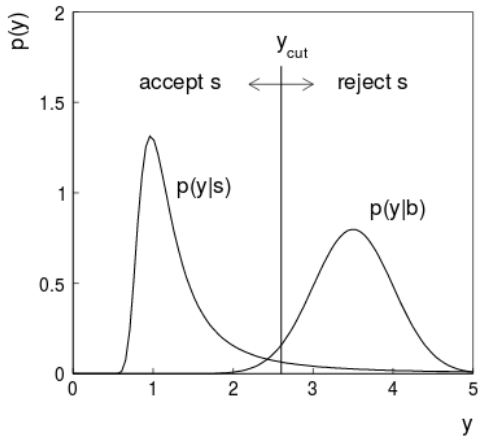


Discovery as hypothesis test

Problem: in practice more than one variable: how to define critical region?



Solution: Test statistic $y(x) : \mathcal{D} \to \mathbb{R}$ allows scalar definition of critical region



Discovery as hypothesis test

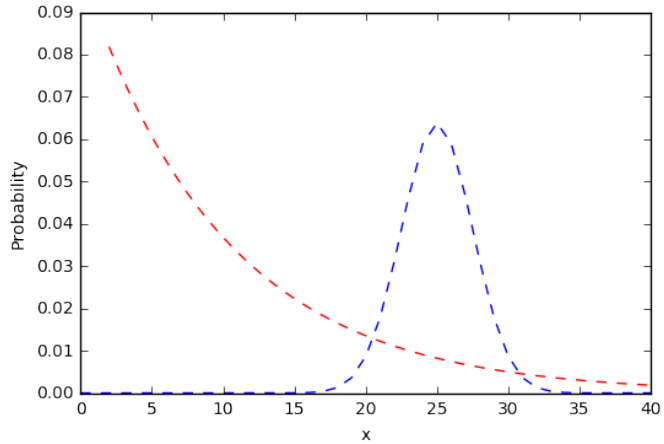
Most powerful test statistic: likelihood ratio of probability density functions for signal and background hypothesis

$$y(x) = \frac{f(x|s)}{f(x|b)}$$

For the example from the beginning:

$$q := -2 \ln y(x) = \left(\frac{x - x_0}{\sigma} \right)^2 - \frac{2x}{\Gamma} + \text{const}$$

Require $q < q_{cut}$ to select a sample enhanced in signal events and then compute p-value.



Discovery as hypothesis test

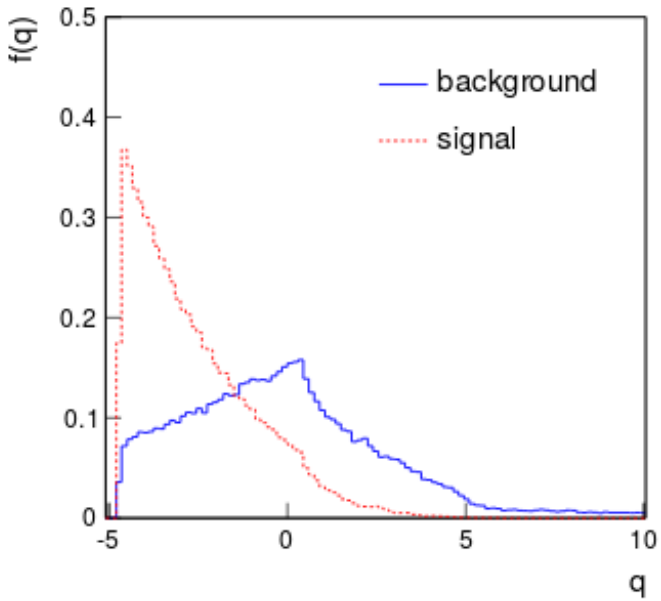
Most powerful test statistic: likelihood ratio of probability density functions for signal and background hypothesis

$$y(x) = \frac{f(x|s)}{f(x|b)}$$

For the example from the beginning:

$$q := -2 \ln y(x) = \left(\frac{x - x_0}{\sigma} \right)^2 - \frac{2x}{\Gamma} + \text{const}$$

Require $q < q_{cut}$ to select a sample enhanced in signal events and then compute p-value.



Discovery as hypothesis test

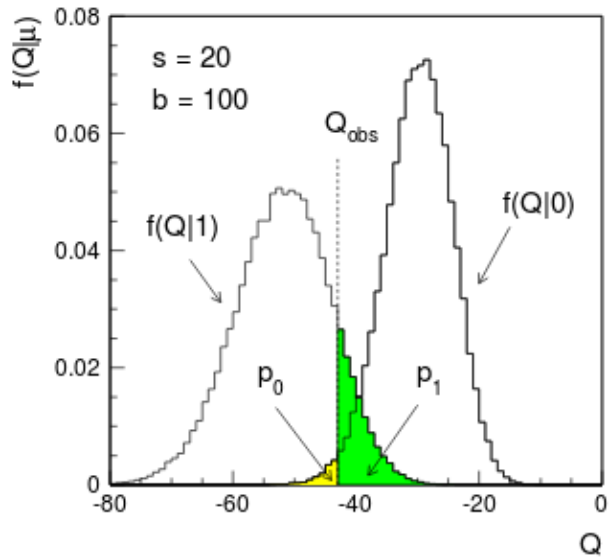
For searches we don't know if the signal exists, either it is in all measured samples or there is only background.
Use extended likelihood function for test:

$$L(\mu) = f(\mathcal{D}_{obs}|\mu) = \text{Pois}(n|\nu(\mu)) \prod_{e=1}^{n} f(x_e|\mu)$$

Construct test statistic as likelihood ratio:

$$Q = -2 \ln \frac{L(1)}{L(0)}$$

For discovery:
require $p_0 < 2.9 \times 10^{-7}$.



Discovery as hypothesis test

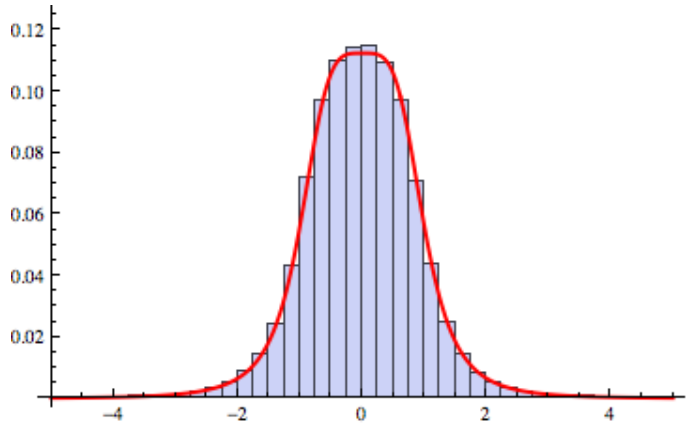
For searches we don't know if the signal exists, either it is in all measured samples or there is only background.
Use extended likelihood function for test:

$$L(\mu) = f(\mathcal{D}_{obs}|\mu) = \text{Pois}(n|\nu(\mu)) \prod_{e=1}^{n} f(x_e|\mu)$$

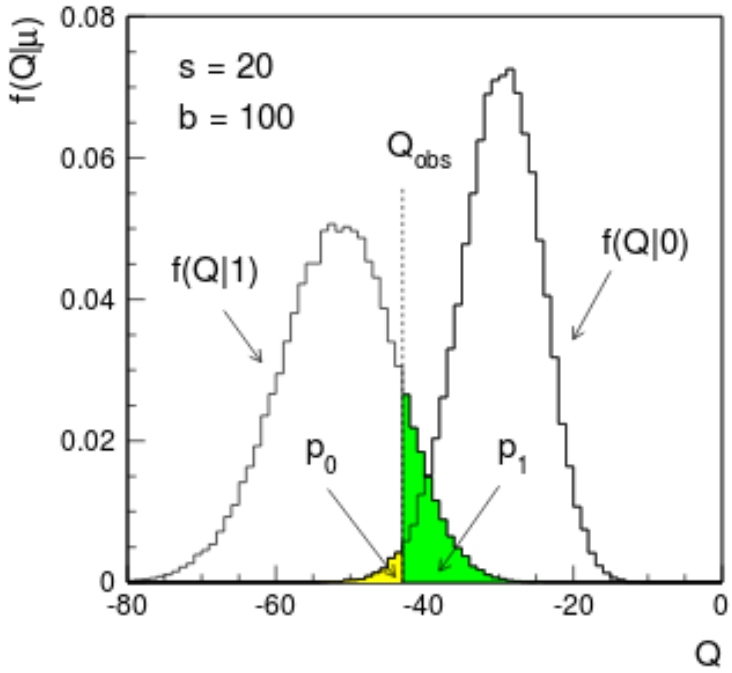
Construct test statistic as likelihood ratio:

$$Q = -2 \ln \frac{L(1)}{L(0)}$$

For discovery:
require $p_0 < 2.9 \times 10^{-7}$ (5 σ).

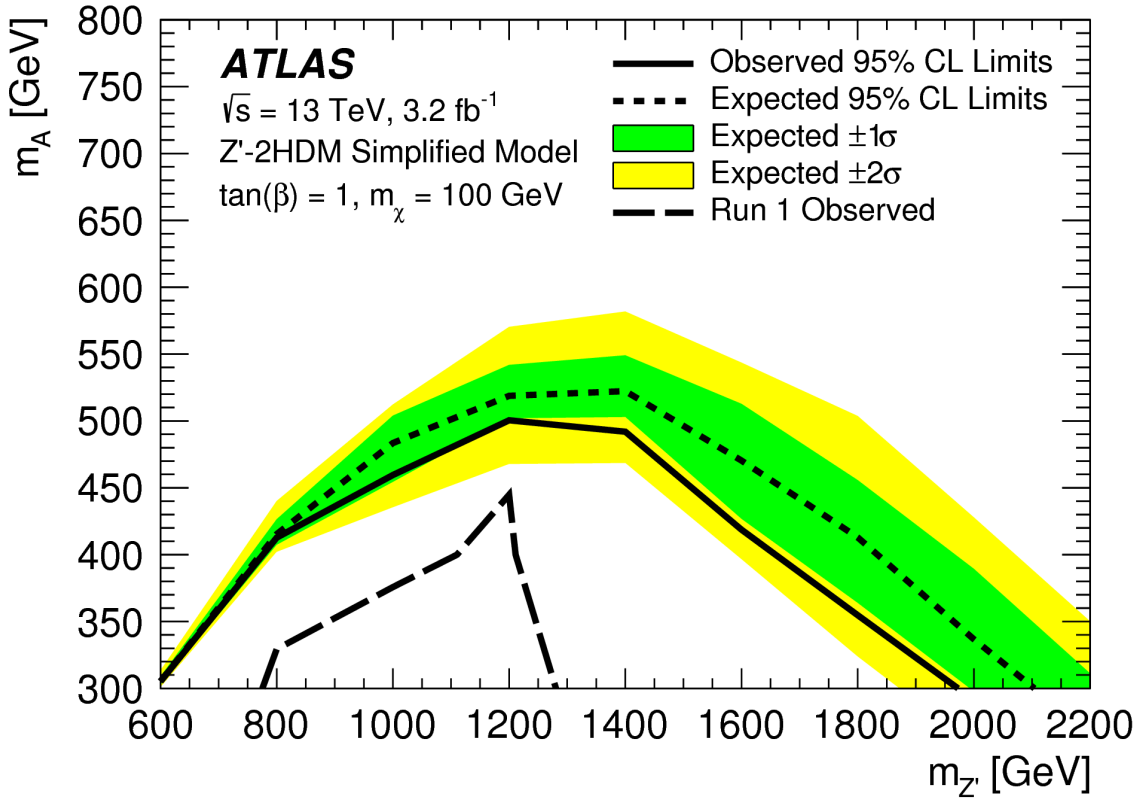


Excluded regions as confidence intervals



- If we find $p_\mu \leq \alpha$, we reject this value of μ. The highest level of μ at confidence $1 - \alpha$ that is not rejected is called upper limit of μ.
- For exclusion at $1 - \alpha = 95\%$ confidence level we require that the critical region starts at 1.64 standard deviations below the value of μ being tested.

Excluded regions as confidence intervals



Example: exclusion limits for ATLAS Dark Matter search $\chi + H \to bb$ with 2015 data

Summary

- Statistical analysis and tools are necessary to interpret LHC experiments
- Parameterise physics analysis in explicit statistical model
- Likelihood-based analysis (frequentist) to estimate parameters, claim discovery or set up limits
- Further reading: References in back-up


K. Cranmer, doi:10.5170/CERN-2015-001.247,10.5170/CERN-2014-003.267 arXiv:1503.07622 [physics.data-an].


G. Cowan, doi:10.1007/978-3-319-05362-2_9 arXiv:1307.2487 [hep-ex].