Statistics & Machine Learning for HEP 2

Harrison B. Prosper
Florida State University

11th CERN | 15 – 28 MARCH 2023
LATIN-AMERICAN SCHOOL OF HIGH-ENERGY PHYSICS
San Esteban, Chile
https://indico.cern.ch/e/clashep2023

Topics

- **Lecture 1**
  - Frequentist Analysis (1)
- **Lecture 2**
  - Frequentist Analysis (2)
  - Bayesian Analysis
- **Lectures 3, 4**
  - Introduction to Machine Learning

FREQUENTIST ANALYSIS (2)

BY EXAMPLE

Parameter(s) of Interest

Last time, we succeeded in creating a *confidence set* for the 2 parameters of the likelihood

$$p(D|s, b) = \text{Poisson}(N, s + b)\ \text{Poisson}(M, kb)$$

In practice, however, we usually make inferences about a subset of the parameters, i.e., the *parameters of interest* (POI). Here there is only one: the mean signal ***s***. The mean background ***b*** is an example of a *nuisance parameter*.

If we wish to make inferences about the signal, we must rid our likelihood of *all* nuisance parameters; in particular, we must get rid of ***b***.

Profile Likelihood: $H \to ZZ \to 4l$

The standard practice is to *replace* all nuisance parameters in the likelihood function by their *conditional* MLEs, that is, their MLE for given values of the parameters of interest.

In this example, this means solving,

$$\frac{\partial \ln p(D|s,b)}{\partial b} = 0$$

for a *fixed* s to find $\hat{b} = f(s)$.

Exercise 4: Show that

$$f(s) = \frac{g + \sqrt{g^2 + 4(1+k)Ms}}{2(1+k)}$$

$$g = N + M - (1+k)s$$

The resulting function $L_p(s) = p(D|s, f(s))$ is called the *profile likelihood*.

Profile Likelihood: $H \to ZZ \to 4l$

Obviously, when we replace the parameter ***b*** by an estimate of it

$$\hat{b} = f(s)$$

we are making an *approximation*.

Therefore, we cannot expect the *frequentist principle* to be satisfied exactly: there could be subsets of the parameter space where the *coverage probability* dips below the specified *confidence level*. If the dips are no bigger than about 10% most physicists would be happy about this.

Furthermore, *profiling* has a sound justification…

Profile Likelihood: $H \to ZZ \to 4l$

The profiling procedure rests, again, upon Wilks' theorem: given the *profile likelihood ratio*

$$\lambda_s(D) = \frac{L_p(s)}{L_p(\hat{\boldsymbol{s}})}$$

where $\hat{\boldsymbol{s}}$ is the MLE of $\boldsymbol{s}$, the distribution of the *statistic*

$$t_s(D) = -2 \ln \lambda_s(D)$$

approximates a χ^2 *density*, this time, of **1** degree of freedom.

Profile Likelihood: $H \to ZZ \to 4l$

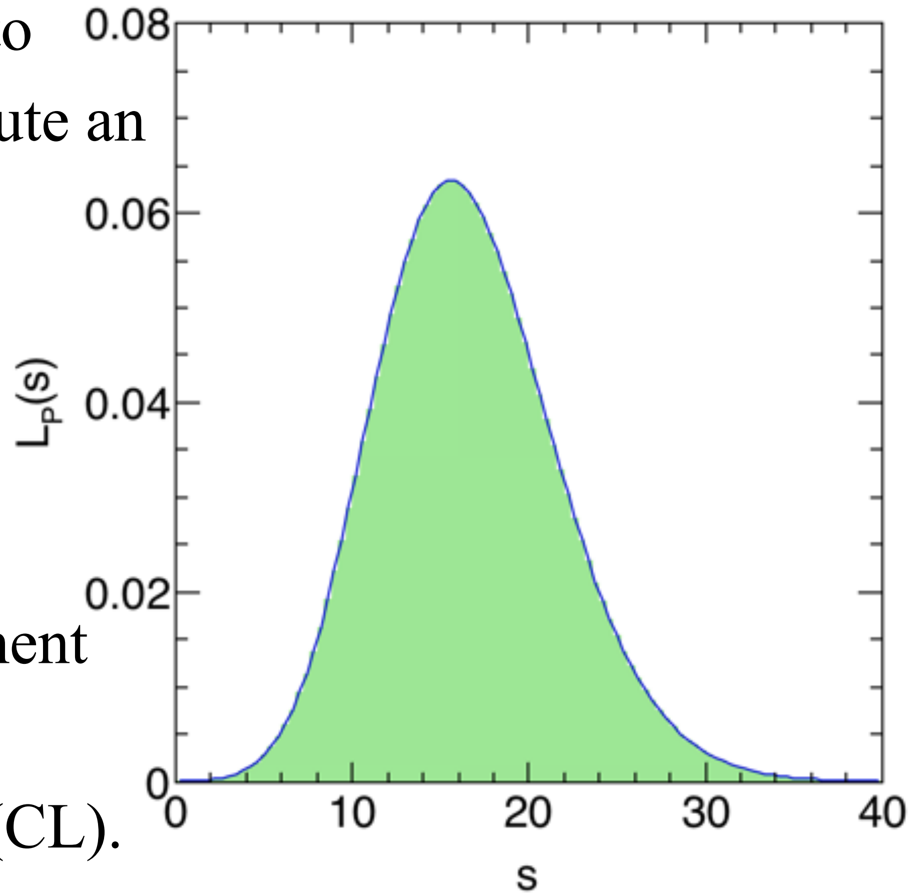
Since $t_s(D) \approx \chi_1^2$, according to Wilks' theorem, we can compute an *approximate* 68% *confidence interval* by solving

$$t_s(D) = 1$$

for ***s***. This results in the statement

$$s \in [10.9, 21]$$

@ ~ 68% *confidence level* (CL).



Exercise 5: Show this by solving $t_s(D) = 1$ numerically

HYPOTHESIS TESTS
BY EXAMPLE

Is The Signal Real?

In experimental physics, it is rare that we can we make definitive statements about signals.

What we do instead is make *probabilistic statements* about whether, or not, a putative signal is real.

In high-energy physics, the consensus is that we declare a signal real, that is, we announce a *discovery*, if the background-only hypothesis is judged to be extremely unlikely.

But, to be quantitative, we need a way to test *hypotheses*.

Hypothesis Tests (1)

1. Decide which hypothesis is to be *rejected* and call it the *null* hypothesis, denoted by H_0. At the LHC, this is usually the *background-only* hypothesis.

2. Construct a function of the data called a *test statistic* such that large values of it would cast doubt on the null hypothesis.

3. Choose a test statistic threshold above which we agree to *reject* the null. Do the experiment, compute the test statistic, and reject the null if the threshold is exceeded.

Hypothesis Tests (2)

There are two variations on this general procedure:

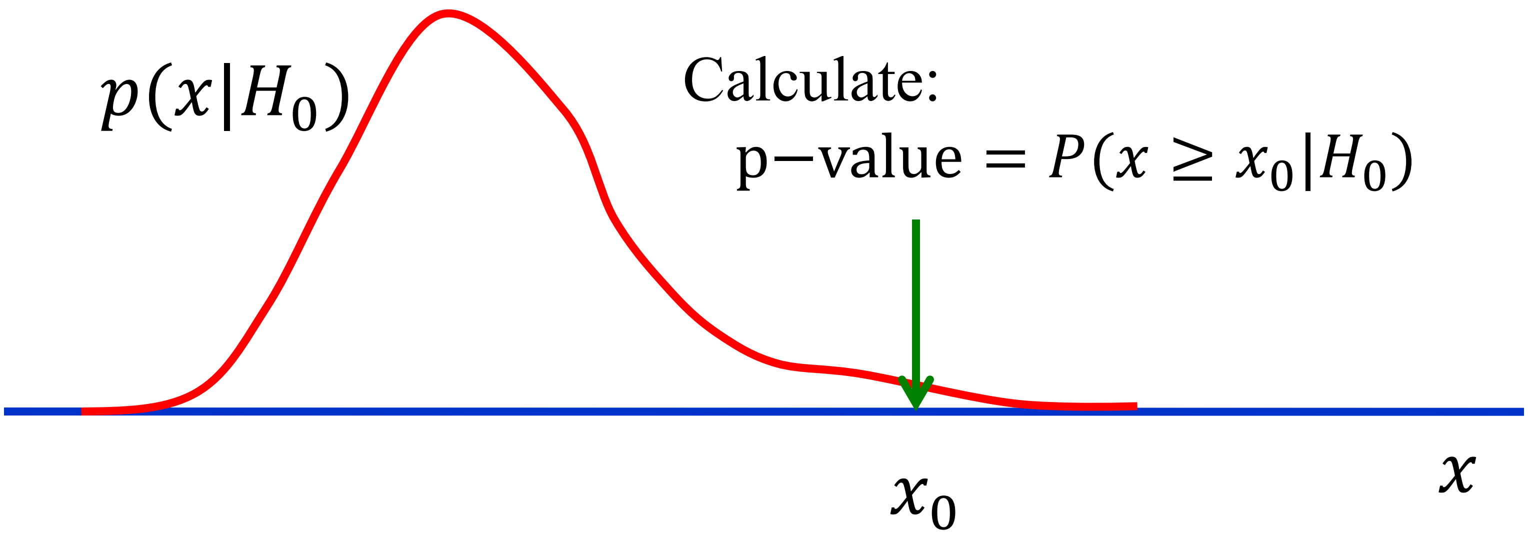
1. Fisher: reject the null if the test statistic is large enough.

2. Neyman: compare the null to an *alternative hypothesis* using a test statistic that depends on *both* hypotheses. Reject the null if the alternative is preferred.

In high-energy physics, we do both!

Hypothesis Tests (3)

Fisher's Approach: *Null* hypothesis (H_0), e.g., background-only



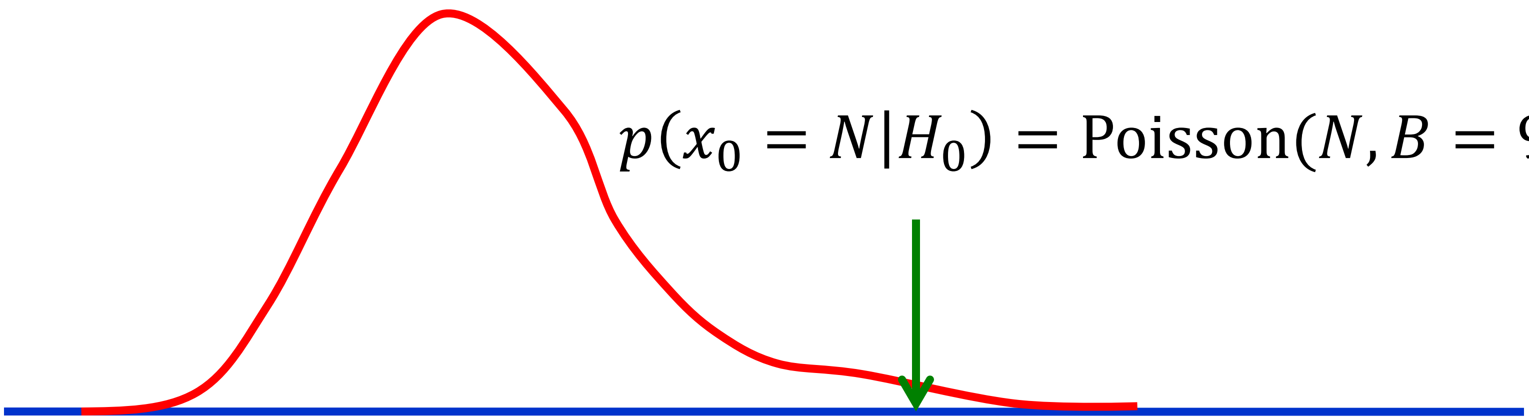
x_0 is the *observed* value of the test statistic x.

The null hypothesis is *rejected* if the p-value is judged to be small enough, i.e., if x_0 is large enough.

Example: $H \to ZZ \to 4l$

Background, $B = 9.4$ events (ignoring uncertainty in background)

$$p(x_0 = N|H_0) = \text{Poisson}(N, B = 9.4)$$



$N = 25$ observed count

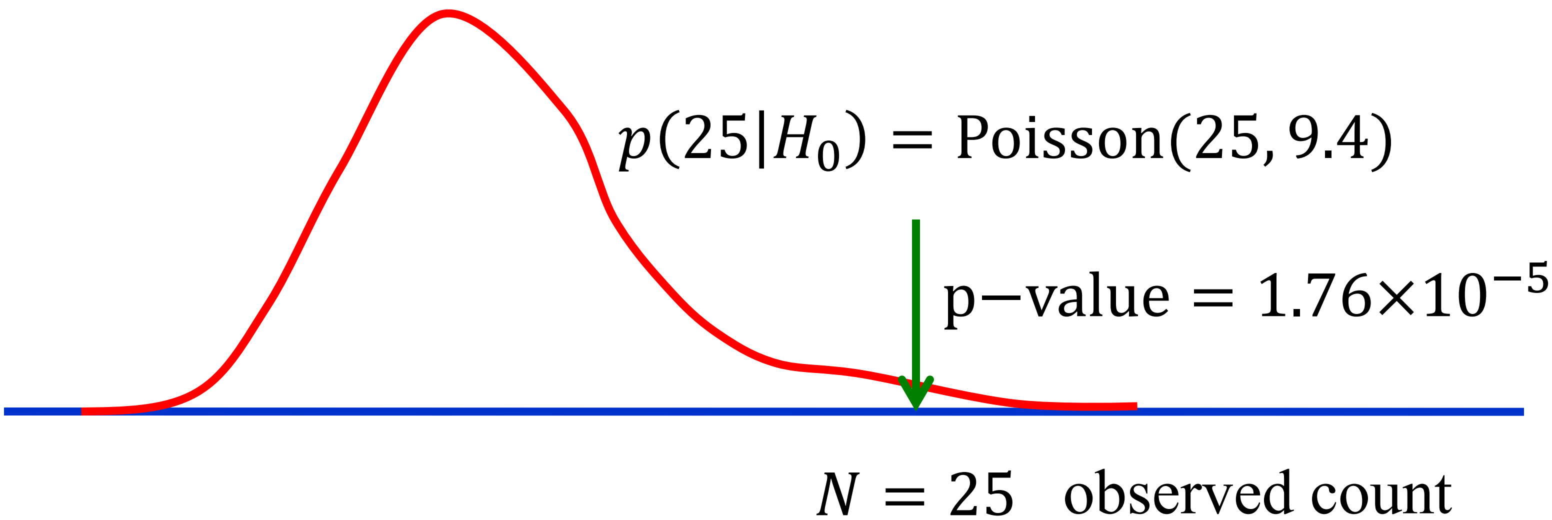
$$\text{p-value} = \sum_{k=N}^{\infty} \text{Poisson}(k, 9.4) = 1.76\times10^{-5}$$

$$\sum_{k=N}^{\infty} \text{Poisson}(k, a) = \int_0^a t^{N-1} e^{-t} dt / \Gamma(N)$$

scipy.special.gammainc(N, a)

Example: $H \rightarrow ZZ \rightarrow 4l$

Background, $B = 9.4$ events (ignoring uncertainty)



We usually map a p-value to a *Z*-value, that is, to the number of standard deviations *away from the null* if the distribution were a Gaussian. This yields $Z = 4.14$.

We say we have a 4.14σ signal.

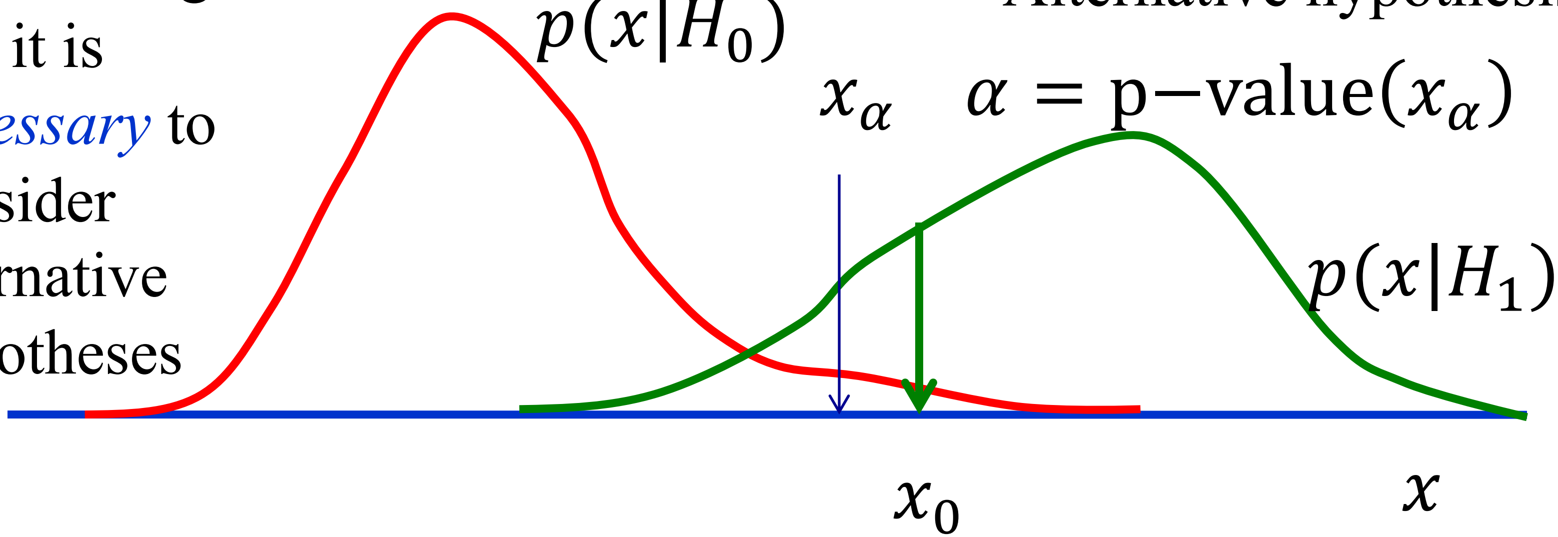
Hypothesis Tests (4)

Neyman's Approach: *Null* hypothesis (H_0) + alternative (H_1)

Neyman argued that it is *necessary* to consider alternative hypotheses

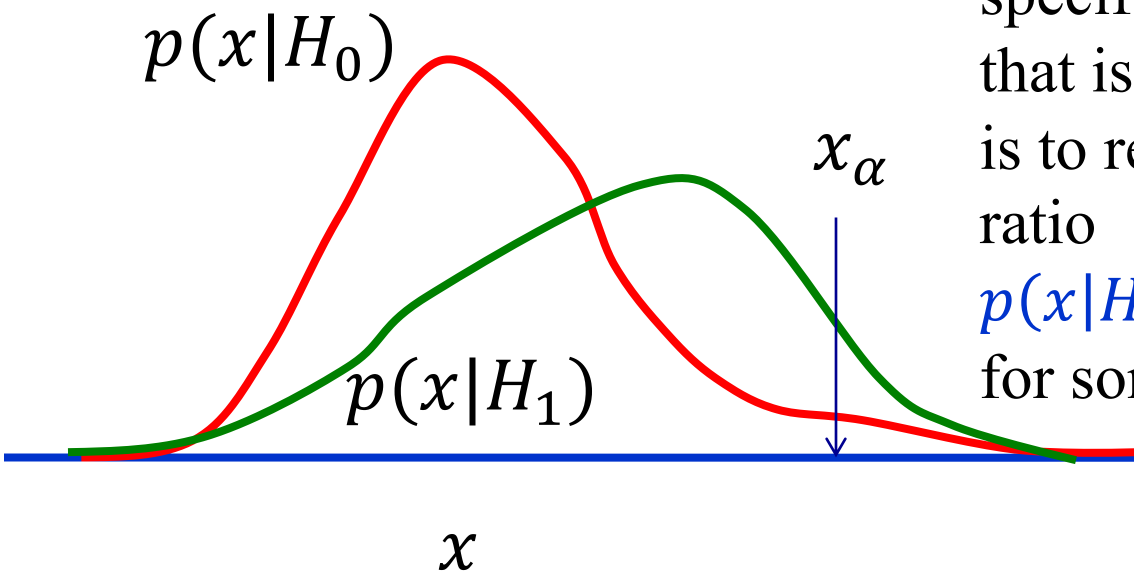
H_1

Alternative hypothesis



Choose a *fixed* value of α *before* data are analyzed. Reject the null in favor of the alternative if the p-value < α. Statisticians call α the *significance* (or size) of the test, while particle physicists call the Z-value the significance!

The Neyman-Pearson Test



The optimal test for fully specified hypotheses, that is, *simple hypotheses*, is to reject the null if the ratio

$p(x|H_1)/\, p(x|H_0) > \lambda$

for some threshold λ.

$$\alpha = \int_{x_\alpha}^{\infty} p(x|H_0)\,dx$$

significance of test

$$p = \int_{x_\alpha}^{\infty} p(x|H_1)\,dx$$

power of test

Hypothesis Tests (5)

All realistic analyses contain *nuisance parameters* that we must get rid of to perform an hypothesis test on the parameters of interest

There two primary ways:

Profiling: Use the profile likelihood.

Marginalizing: Use the *marginal* likelihood, i.e., a likelihood integrated over the nuisance parameters.

Example: $H \to ZZ \to 4l$ (Profiling)

We need to compute

p-value = $P[\boldsymbol{x} > x_0]$

given the observed value $x_0 = t_{s_0}(D)$ of $x = t_{s_0}$.

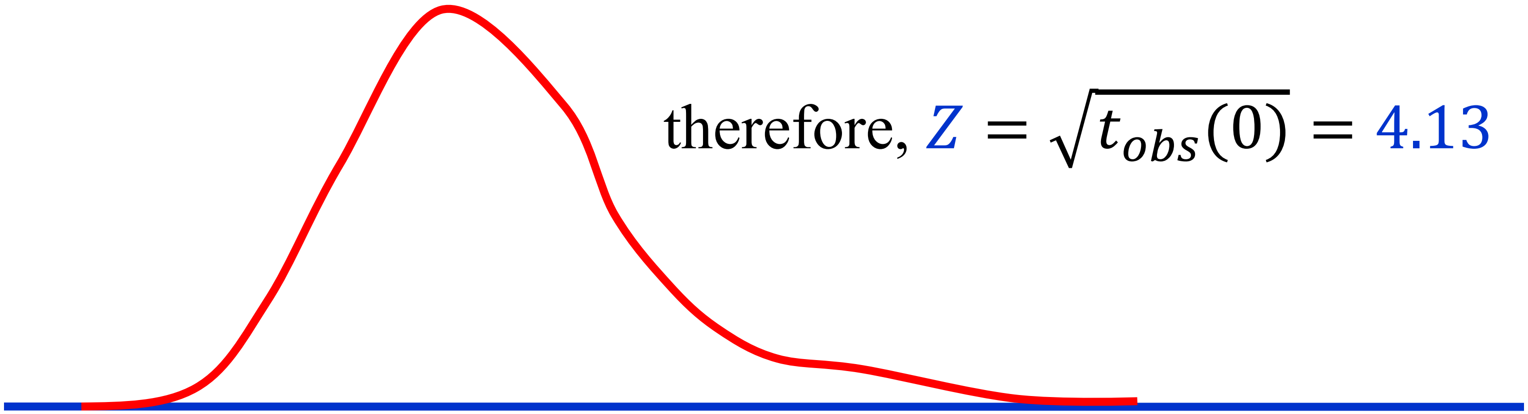
If the p-value < $\boldsymbol{\alpha}$ we agree to reject the $\boldsymbol{s} = \boldsymbol{s_0}$ hypothesis and we also report the p-value.

But, since $Z = \sqrt{t_{s_0}(D)}$, we can avoid the calculation of the p-value and just report Z!

Example: $H \rightarrow ZZ \rightarrow 4l$ (Profiling)

Background, $B = 9.4 \pm \mathbf{0.5}$ events. For this example, $\boldsymbol{s_0} = \mathbf{0}$.

$t_{\text{obs}}(0) = 17.05$



therefore, $Z = \sqrt{t_{obs}(0)} = 4.13$

$$L_p(s) = L(s, \boldsymbol{f}(\boldsymbol{s}))$$

$$\hat{b} = f(s) = \frac{g + \sqrt{g^2 + 4(1+k)Ms}}{2(1+k)}$$

$$g = N + M - (1+k)s$$

$$t_s(D) = -2\ln[L_p(s)/L_p(\hat{s})]$$

Exercise 6: Verify this calculation

BAYESIAN ANALYSIS
BY EXAMPLE

Bayesian Inference (1)

Bayesian methods are

1. based on the *degree of belief* interpretation of probability
2. and use Bayes' theorem

$$p(\theta_H, H | D) = \frac{p(D \mid \theta_H, H)\pi(\theta_H, H)}{p(D)}$$

for *all* inferences, where

D	observed data
θ_{H}	parameters pertaining to hypothesis H (parameters of interest and nuisance parameters)
H	hypothesis
π	*prior density*

BAYESIAN ANALYSIS
BY EXAMPLE

Example: Bayesian Analysis $H \to 4l$

Step 1: Construct a probability model for the observations

$$p(D|s, b) = \frac{(s+b)^N e^{-(s+b)}}{N!} \frac{(kb)^M e^{-kb}}{\Gamma(M+1)}$$

knowns:

$N = 25$	observed event count
$M = 353$	effective background event count
$k = 37.6$	effective background scale factor

unknowns:

b	mean background count
s	mean signal count

Example: Bayesian Analysis $H \rightarrow 4l$

Step 2: Write down Bayes' theorem:

$$p(s,b|D) = \frac{p(D \mid s,b)\, \pi(s,b)}{p(D)}$$

and specify the prior:

$$\pi(s,b) = \pi(b|s)\, \pi(s)$$

Sometimes it is convenient to compute the *marginal likelihood* of the parameters of interest by integrating over the nuisance parameters, here b (as we did earlier),

$$p(D|s) = \int_0^\infty p(D \mid s,b)\, \pi(b|s) db$$

Example: Bayesian Analysis $H \to 4l$

The Prior:

What does

$$\pi(s, b) = \pi(b|s)\,\pi(s)$$

represent?

The prior encodes what we *know*, or *assume*, about the mean background and signal in the absence of *new* observations.

We shall *assume* that s and b are non-negative.

Unfortunately, there is no *unique* way to encode such vague information.

Example: Bayesian Analysis $H \to 4l$

For simplicity, we shall take $\pi(b \mid s) = 1$, though one can do better*.

The marginal likelihood and can be computed exactly:

$$p(D \mid s) = \frac{(1-x)^2}{M} \sum_{r=0}^{N} \text{beta}(x, r+1, M)\, \text{Poisson}(N-r, s)$$

where, $x = \frac{1}{1+k}$.

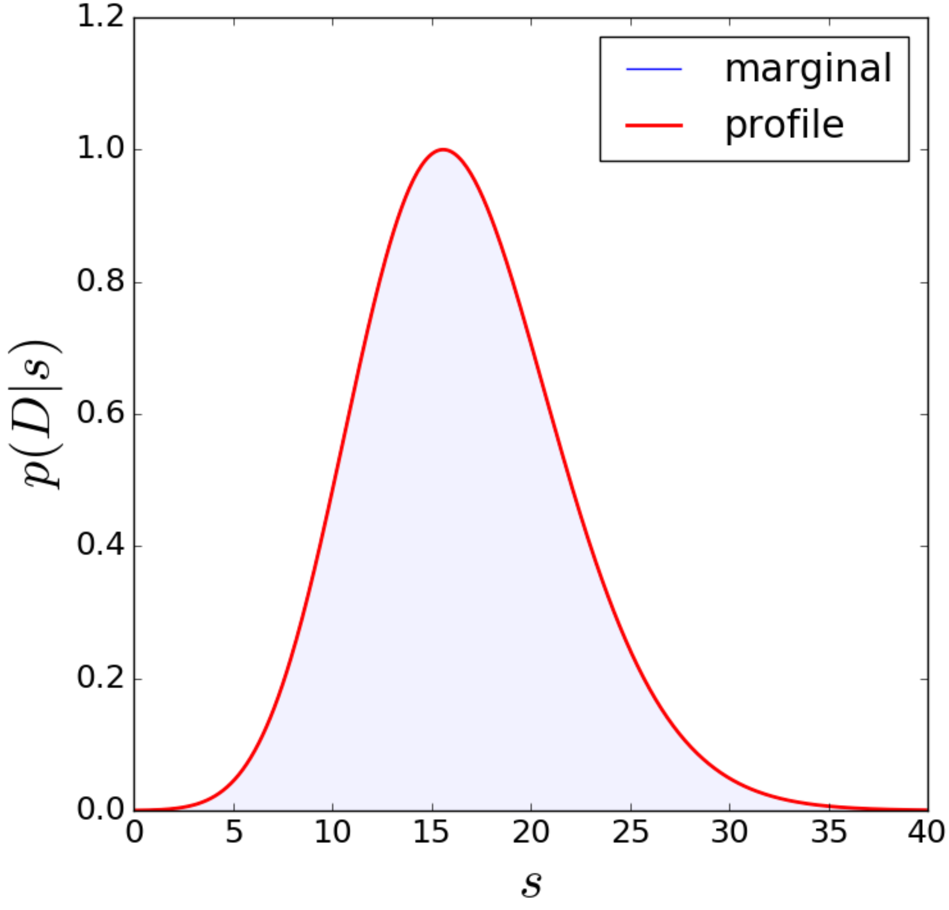
*Luc Demortier, Supriya Jain, HBP, Reference priors for high energy physics, Phys.Rev.D82:034002 (2010)

Example: Bayesian Analysis $H \rightarrow 4l$

$L(s) = P(25 \mid s)$ is the marginal likelihood for the expected signal s.

Here, we compare the marginal and profile likelihoods. For this problem they are almost identical.

But, this does not always happen!



Example: Bayesian Analysis $H \to 4l$

Given $p(D \mid s)$ we can compute the posterior density of the signal

$$p(s \mid D) = \frac{p(D \mid s)\pi(s)}{p(D)}$$

Again, for simplicity, let's assume $\pi(s) = 1$, then

$$p(s \mid D) = \frac{\sum_{r=0}^{N} \text{beta}(x, r+1, M)\text{Poisson}(N - r, s)}{\sum_{r=0}^{N} \text{beta}(x, r+1, M)}$$

Exercise 7: Derive an expression for $p(s \mid D)$ assuming a gamma prior Gamma$(qs, U+1)$ for $\pi(s)$

Example: Bayesian Analysis $H \to 4l$

Computing Central Credible Intervals

Solve

$$\int_0^{l(N)} p(s\,|D)\,ds = (1 - \text{CL})/2$$

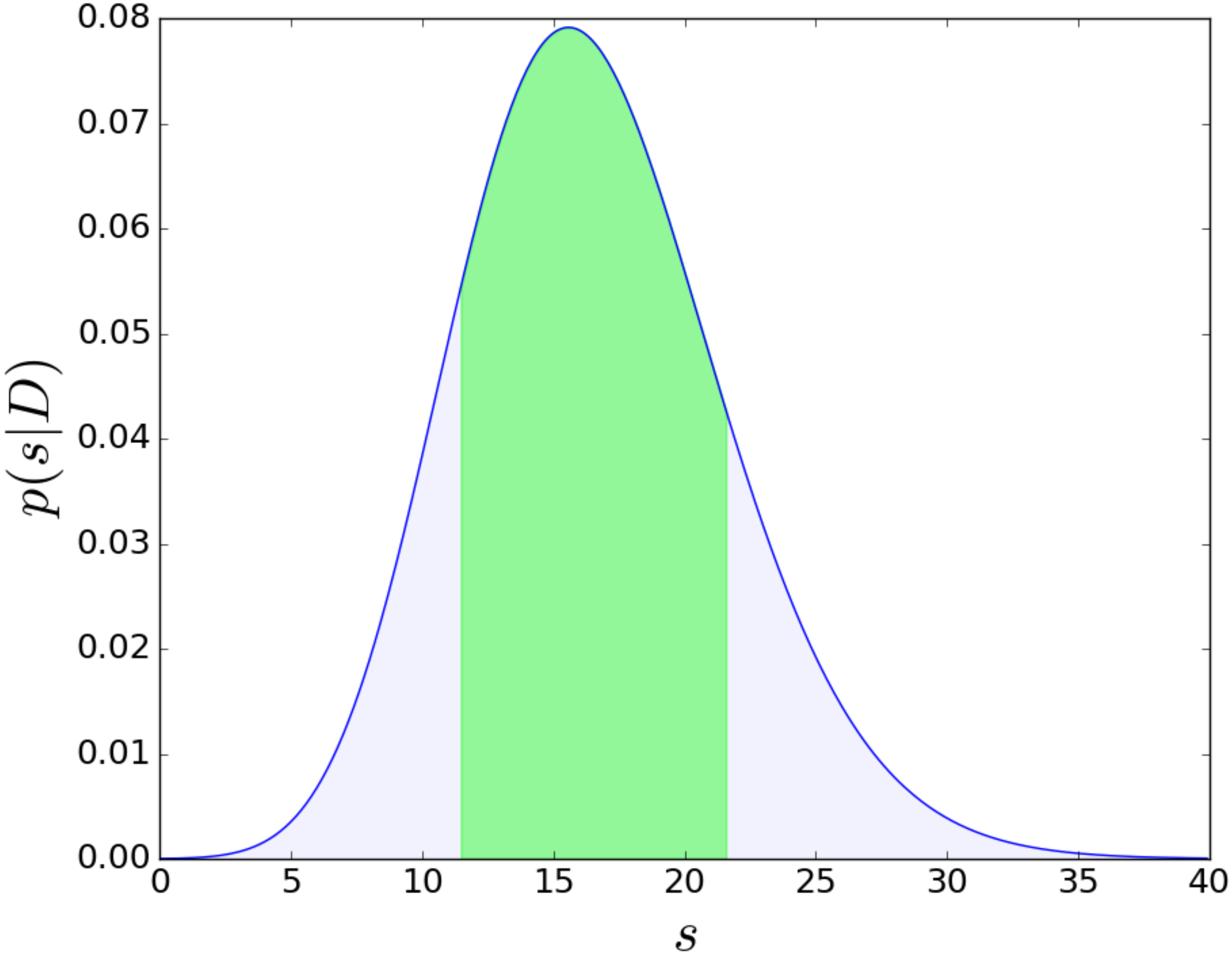
$$\int_0^{u(N)} p(s\,|D)\,ds = (1 + \text{CL})/2$$

with CL = 0.683, we obtain $s \in [11.5, 21.7]$ at 68% credible level (CL).

Since this is a Bayesian calculation, this statement means:

the probability that s lies in [11.5, 21.7] *is* 0.68.

Example: Bayesian Analysis $H \rightarrow 4l$



Example: Bayesian Analysis $H \to 4l$

Finally, we can test different hypotheses H about the signal $\boldsymbol{s}$ by marginalizing over the parameters of each hypothesis. In our case, the parameters are $\theta_{H_0} = b$ and $\theta_{H_1} = b, s$ for hypotheses H_0 and H_1, respectively.

Since we have already marginalized over b, we just need to compute

$$p(D\,|H_1) = \int_0^\infty p(D|s, H_1\,)\pi(s|H_1)ds$$

The simplest choice for the prior is $\pi\,(s\mid H_1) = \delta(s - 15.6)$, which yields

$$p(D|H_1) \equiv p(D|\,\boldsymbol{s = 15.6}) = 7.91{\times}10^{-2}.$$

Note also that

$$p(D|H_0)\ \equiv p(D|\,\boldsymbol{s = 0}) = 1.59{\times}10^{-5}$$

Example: Bayesian Analysis $H \to 4l$

From

$$p(D \mid H_1) = 7.91\times10^{-2} \text{ and}$$
$$p(D \mid H_0) = 1.59\times10^{-5}$$

we conclude that the CMS results increase the probability of hypothesis H_1 relative to H_0 by ~5000.

The increased odds can be converted to a *Z*-value (S. Sekmen, HBP) roughly equivalent to the frequentist measure using

$$Z = \text{sign}(\ln B_{10})\sqrt{2[\ln B_{10}]}$$

This yields Z = 4.13.

Exercise 8: Verify this number

Summary (1)

Probability

Interpretations: degree of belief, relative frequency

Likelihood Function

Probability model into which data have been inserted.

Frequentist Principle

Construct statements such that a fraction $f \geq$ CL of them will be true over a population of statements.

Frequentist Analysis

1. Eliminate nuisance parameters by profiling likelihood.
2. Tests: decide on a fixed threshold α and *reject* null hypothesis if the p-value $< \alpha$; report the p-value.

Summary (2)

Frequentist Analysis (2)

- Profile Likelihood
  - Standard way to eliminate nuisance parameters. But strict adherence to frequentist principle not guaranteed.
- Hypothesis Tests
  - Decide on a fixed threshold α and *reject* null hypothesis if the p-value < α *and* report the p-value.

Bayesian Analysis

- Uses Bayes' theorem for all inferences.
- Needs both a likelihood and a *prior*.
- Must compare at least *two* hypotheses.