Introduction to Statistics in Particle Physics

Harrison B. Prosper

Florida State University

ENHEP 19, Ain Shams University, Egypt

29 January, 2019

Topics

- Introduction
- Frequentist Analysis: An Example
- Hypothesis Tests: An Example
- Summary

- Bibliography
- Tutorials

INTRODUCTION

Introduction: Statistical Inference

The main goal of statistical inference is to use a *sample*, which is necessarily finite, to infer something about its associated *population*, which, by definition, is infinite.

Note:

- The great thing about physics is that there is a *single* judge of its correctness (namely, *Nature*). The bad thing about statistics is that there too many judges!
- Consequently, there is no such thing as “the right approach”; rather, there are different approaches, with different assumptions, and different opinions about them.

Introduction: Statistical Inference

Everyone at least agrees that the key concept is *probability*.
But, it is interpreted in at least two ways:

1. **Degree of belief** in, or assigned to, a proposition e.g.:

   proposition: It will rain in Cairo tonight
   probability: $p = 5 \times 10^{-2}$

2. **Relative frequency** of outcomes in an *infinite* sequence of trials, e.g.:

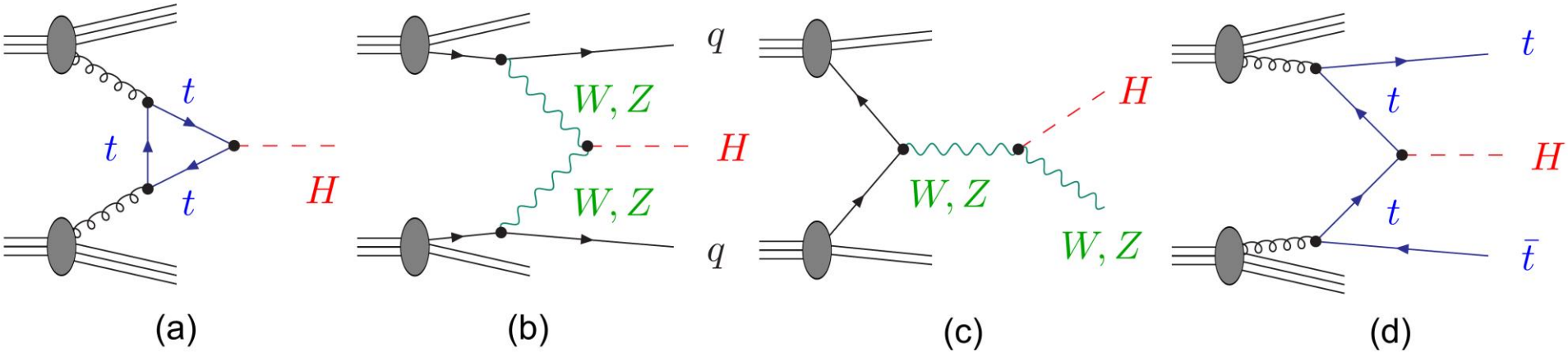
   trial: a proton-proton collision at the LHC

   outcome: creation of a Higgs boson

   probability: $p = 5 \times 10^{-10}$

FREQUENTIST ANALYSIS
AN EXAMPLE

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

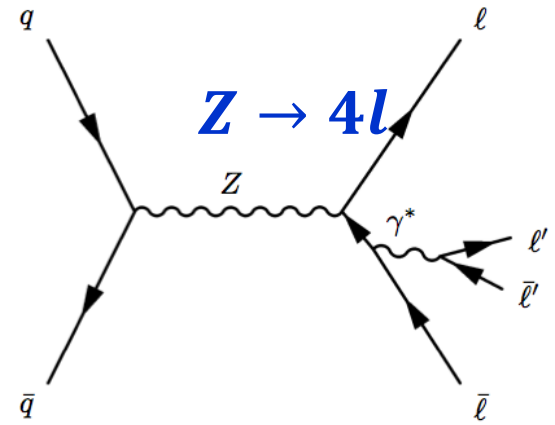


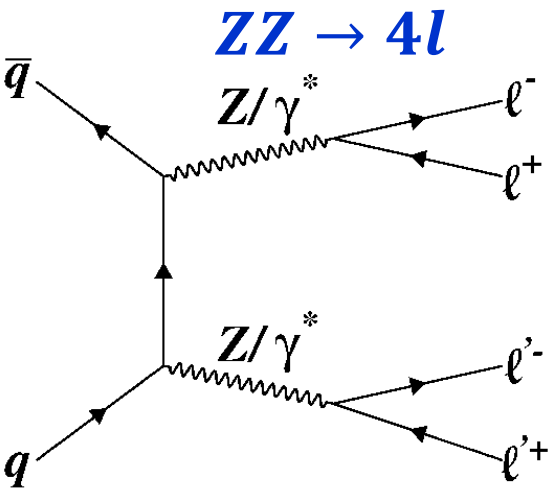
Process		$\sigma \times BR$ (fb)
(a) Gluon gluon fusion	(ggF)	12.18
(b) Vector boson fusion	(VBF)	1.044
(c) Associated production	(VH)	1.047
(d) Top anti-top fusion	(ttH)	0.393

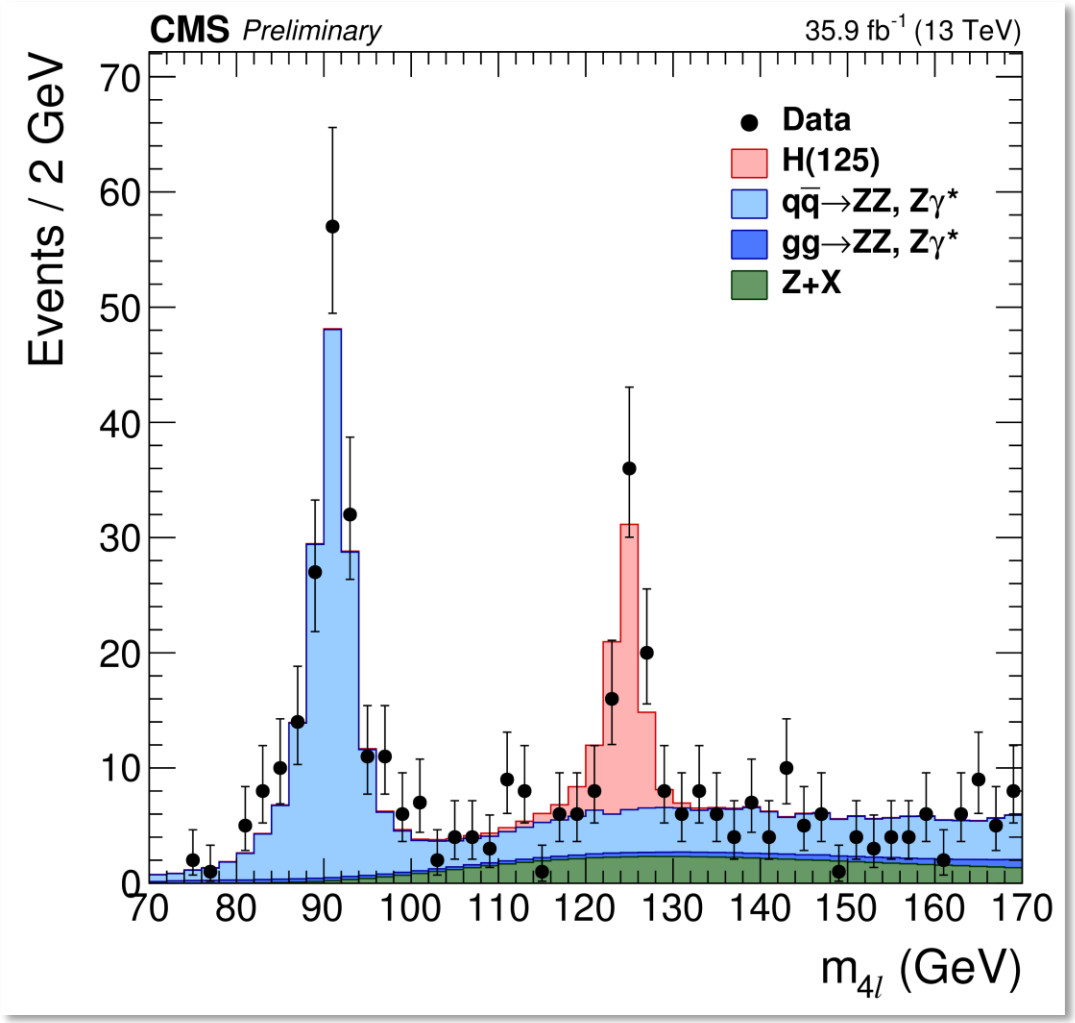
Before event selection, the background ~ **1700** times larger!

CMS (2018) $pp \rightarrow H \rightarrow ZZ \rightarrow 4l$

The main backgrounds:







Knowns and Unknowns: $H \rightarrow ZZ \rightarrow 4l$

In 2014, the CMS Collaboration published a summary of its work on $pp \rightarrow H \rightarrow ZZ \rightarrow 4l$ (Phys. Rev. **D89**, 092007 (2014))

knowns:

$N = 25$	observed event count
$B \pm \delta B = 9.4 \pm 0.5$	background event count
$S \pm \delta S = 17.3 \pm 1.3$	predicted signal count @ $m_H = 125$ GeV

unknowns:

b	mean background count
s	mean signal count
$d = s + b$	mean event count

Probability Model: $H \to ZZ \to 4l$

Goals:

1. Estimate (i.e., measure) the mean signal, ***s***.
2. Quantify the accuracy of the estimate.
3. Quantify the significance of the signal.

In order to do the above, we need first to construct a *probability model* of the *data generation mechanism*.

Let's start from scratch…

Bernoulli Trial (1): $H \rightarrow ZZ \rightarrow 4l$

A Bernoulli trial has two outcomes:

S = success or ***F*** = failure.

Example: Each collision between protons at the LHC is a Bernoulli trial in which either something interesting happens (***S***) or does not happen (***F***).

What is the probability of this sequence of events?

Without assumptions, there is ***no*** answer!

Bernoulli Trial (2) : $H \rightarrow ZZ \rightarrow 4l$

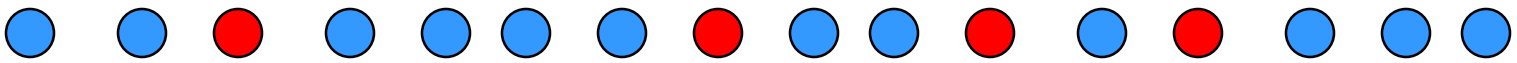
Assumption 1: Let ***p*** be the probability of a success.

Assumption 2: Let ***p*** be the same for every collision (trial).

Assumption 3: Let ***S*** and ***F*** be *exhaustive* and *mutually exclusive*. Therefore, the probability of a failure is **1 – *p***.

Consequently, for a sequence ***Q*** of ***n*** trials, the probability $P(\boldsymbol{k}|Q, p, n)$ of *exactly* ***k*** successes and *exactly* ***n – k*** failures is

$$P(k|\,Q, p, n) = p^k(1-p)^{n-k}$$



Binomial Distribution: $H \to ZZ \to 4l$

Note: the sequence $\boldsymbol{Q}$ of successes at the LHC is unknown!

According to the rules of probability theory, we can eliminate the unknown (discrete) parameter $\boldsymbol{Q}$ from the problem by *summing* over all sequences that are possible *a priori*:

$$P(k|p,n) = \sum_{\boldsymbol{Q}} P(k|\boldsymbol{S},p,n) = \sum_{Q} p^k (1-p)^{n-k}$$

Each sequence has the same probability and there are $\binom{n}{k}$ of them in the sum. Therefore,

$$P(k|\,p,n) = \binom{n}{k} p^k (1-p)^{n-k},$$

which is the *binomial distribution*, Binomial(k,n,p).

Poisson Distribution: $H \to ZZ \to 4l$

The mean number of successes $\boldsymbol{a}$ is

$\boldsymbol{a = pn}$. **Exercise 1**: Show this

For the Higgs boson outcomes, $\boldsymbol{p \sim 10^{-10}}$ and $\boldsymbol{n >> 10^{12}}$.

So, let's consider $\boldsymbol{p} \to 0$ and $\boldsymbol{n} \to \infty$, with ***a*** *constant*,

Binomial$(k, n, p) \to$ **Poisson**$(k, a) = a^k \exp(-a)\,/k!$

Exercise 2: Show that Binomial$(k, n, p) \to$ Poisson(k, a)

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

Probability Function:

The probability to observe a count n is, therefore,

$$p(n|s,\ b) = \text{Poisson}(n, s+b) = \frac{(s+b)^n e^{-(s+b)}}{n!}$$

Likelihood Function:

$p(N|s,\ b)$, N =25

The *likelihood function* is simply the *probability function* evaluated at the observed data.

What about $B \pm \delta B = 9.4 \pm 0.5$?

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

One way to proceed is to suppose that

$$B \pm \delta B = 9.4 \pm 0.5$$

is the result of *scaling* down a count M by some factor k

$$B = M / k, \quad \delta B = \sqrt{M} / k,$$

and that the count M is sampled from a Poisson distribution with variance $\approx M$ (perhaps from a Monte Carlo simulation). We can then solve for M and k to get $M = 353.4$, $k = 37.6$.

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

Therefore, the likelihood for the count M is

$$\text{Poisson}(M, kb) = (kb)^M e^{-kb} / \Gamma(M+1),$$

where we have continued the function to non-integer M.

The full likelihood for the data $D = (N, M, k)$ is, therefore,

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

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

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

Given $p(D|s,\ b)$, we can answer:

1. How does one estimate (measure) the mean signal, s?
2. How does one quantify the accuracy of the estimate?
3. How does one decide if a signal has been found?

A common way to estimate a parameter is to choose the value that maximizes the likelihood.

Estimates obtained this way are called *maximum likelihood estimates* (MLE).

Nuisance Parameters are a Nuisance!

But, there is a problem! The likelihood

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

contains two parameters ***s*** and ***b*** only one of which, ***s***, is of interest to us, that is, only one is the *parameter of interest.*

The parameter ***b*** is an example of a *nuisance parameter*.

If we wish to make inferences about the parameter of interest, we must rid our probability model of *all* nuisance parameters.

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

This means, we must transform the 2-parameter function

$$\frac{(s+b)^N e^{-(s+b)}}{N!} \frac{(kb)^M e^{-kb}}{\Gamma(M+1)}$$

into one involving ***s*** only.

Moreover, in principle, this should be done while respecting the foundational principle of *frequentist statistics*, namely, the *frequentist principle*.

The Frequentist Principle

The Frequentist Principle (FP) (Jerzy Neyman, 1937)

Given a probability ***p***, statements should be constructed so that we can guarantee that a fraction $f \geq p$ of them are true over an *ensemble* of statements, even if we do not know which ones are true and which are false.

The fraction ***f*** is called the *coverage probability* (or *coverage* for short) and ***p*** is called the *confidence level* (C.L.).

Statements that obey the frequentist principle are said *to cover*.

The Frequentist Principle

Example

Consider an ensemble of statements, each associated with a pair of numbers, a mean count θ randomly sampled from **uniform**(0, 3) and a count N randomly sampled from a **Poisson** distribution with mean θ.

In a simulation, both numbers are known. Therefore, we can compute the coverage probability ***f*** of statements of the form $N + \sqrt{N} > \theta$.

Exercise 3:
What is the coverage probability of these statements? Repeat using **uniform**(0, 3000).

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

As noted, in order to make an inference about the parameter of interest, we must get rid of the nuisance parameters.

The common practice, is to replace the nuisance parameters in the likelihood function by their conditional MLEs, that is, their MLE for given values of the parameter (or parameters) of interest.

The resulting function is called the *profile likelihood*, L_p.

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

In our example, this means replacing the unknown parameter ***b*** by its MLE

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

for a given value of s: $L_p(s) = p(D|s, f(s))$.

Replacing the true value of a parameter by an estimate of it is an *approximation*.

Consequently, if we use the profile likelihood as if it were a likelihood then the *frequentist principle* is not guaranteed to be satisfied exactly.

Nevertheless, *profiling* has a sound justification…

Wald Approximation: $H \to ZZ \to 4l$

Consider the *profile likelihood ratio*

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

where $\hat{\boldsymbol{s}}$ is the MLE of $\boldsymbol{s}$. Taylor expand the quantity

$$t(s) = -2 \ln \lambda(s)$$

about $\hat{\boldsymbol{s}}$:

$$t(\hat{\boldsymbol{s}} + \ s - \hat{\boldsymbol{s}}) = t(\hat{\boldsymbol{s}}) + t'(\hat{\boldsymbol{s}})(s - \hat{\boldsymbol{s}}) + \frac{t''(\hat{\boldsymbol{s}})(s - \hat{\boldsymbol{s}})^2}{2} + \cdots$$

$$\approx (s - \hat{\boldsymbol{s}})^2 / \ \sigma^2 + \mathrm{O}(1/\sqrt{N})$$

where $\sigma^2 \approx 2/t''(\hat{\boldsymbol{s}})$.

This approximation is called the *Wald approximation* (1943).

Wilks' Theorem: $H \rightarrow ZZ \rightarrow 4l$

If $\hat{\boldsymbol{s}}$ does not occur on the boundary of the parameter space, and the data sample is large enough (typically, when the density of $\hat{\boldsymbol{s}}$ is approximately Gaussian$(\hat{\boldsymbol{s}}, \boldsymbol{s}, \sigma)$), and $\boldsymbol{s}$ is the *true value* of the mean signal, then

$$t(s) = -2 \ln \lambda(s)$$

has a χ^2 density of one degree of freedom.

This result is called Wilks' Theorem (1938)*.

Exercise 4: Verify this theorem through simulation

Note, this theorem implies that the probability density of $\boldsymbol{t}(\boldsymbol{s})$ is independent of *all* the parameters of the problem!

(*Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells "Asymptotic formulae for likelihood-based tests of new physics." Eur.Phys.J.C71: 1554, 2011)

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

The MLE of *b*, given *s*, is

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

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

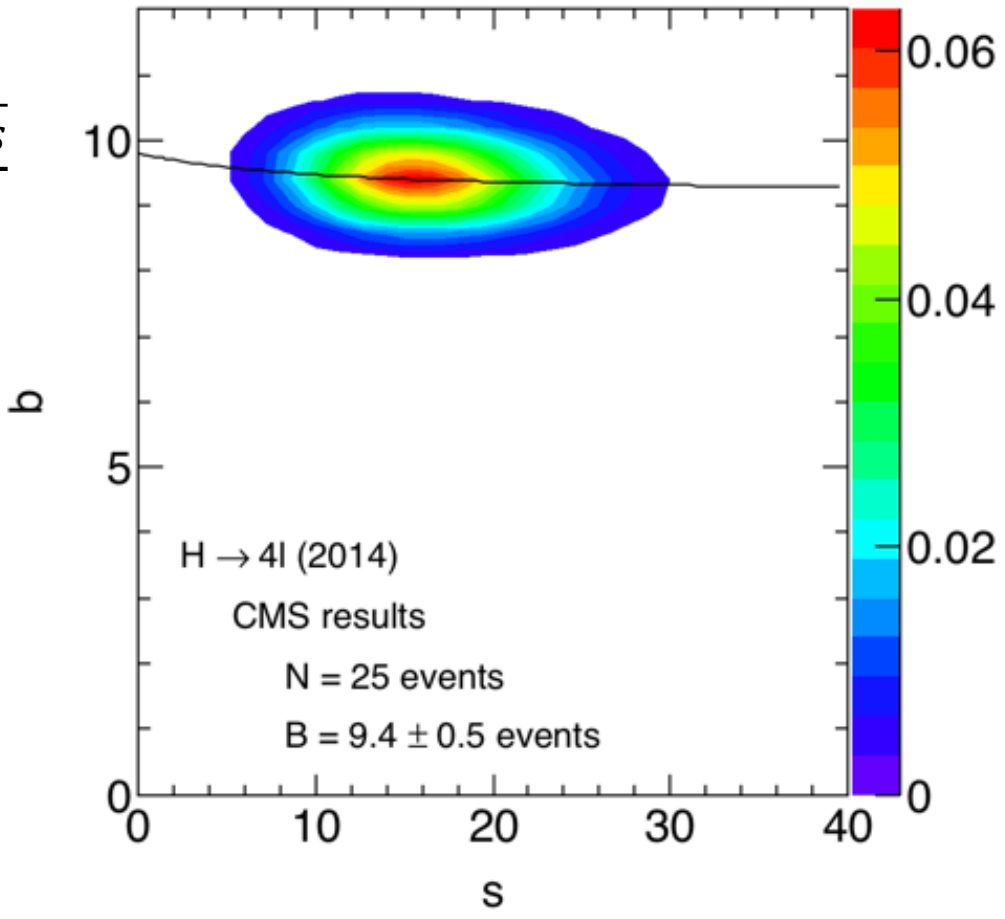
Note,

$s = N - B = 15.6$ events

$b = B$

is the *mode* (location of the peak) of the likelihood function.

Exercise 5: Show this



Confidence Interval: $H \to ZZ \to 4l$

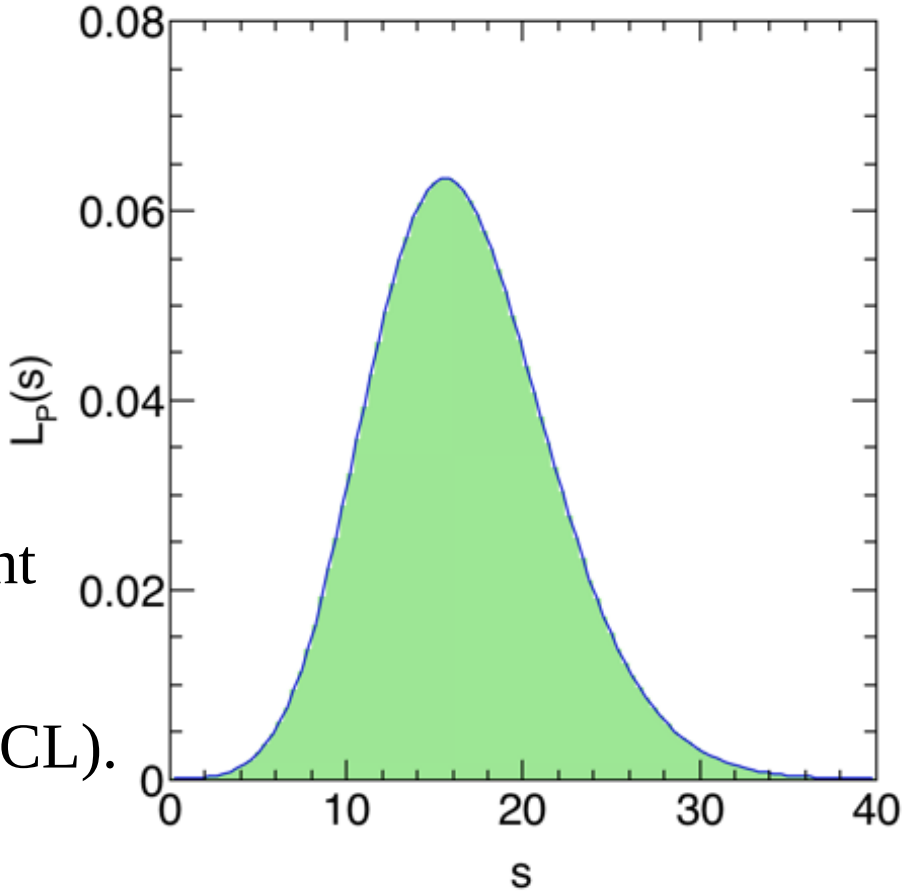
Since $t(s) \approx \chi^2$, we can compute an *approximate* 68% confidence interval by solving

$$t(s) = -2 \ln \lambda(s) = 1$$

for ***s***. The result is the statement

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

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



Exercise 6: Show this by solving $t(s) = 1$ numerically

Confidence Interval: $H \rightarrow ZZ \rightarrow 4l$

But what *exactly* does the statement

$s \in [10.9, 21]$ @ 68% CL

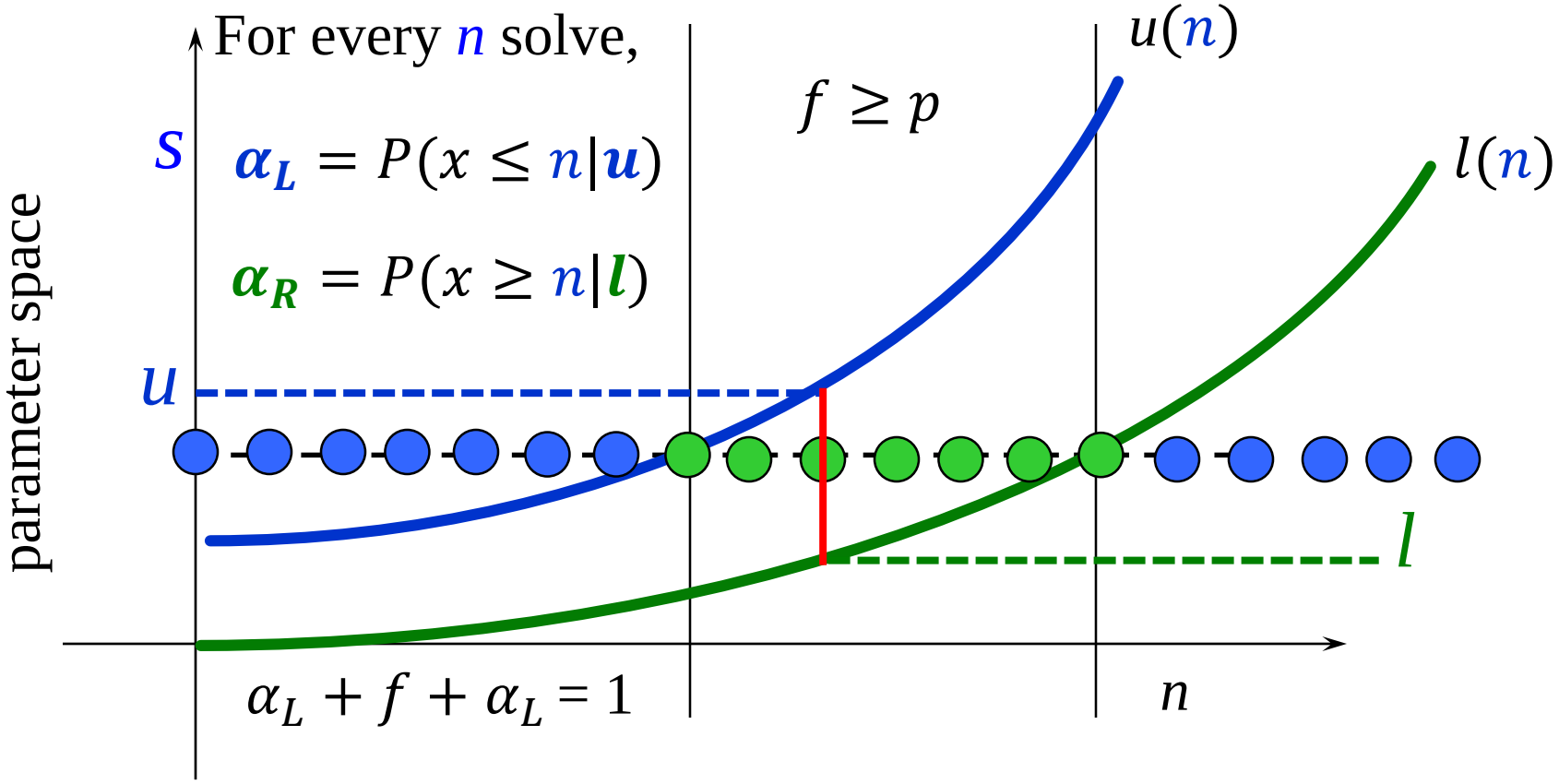
mean? First note that this statement is either true or false.

While we do not know the truth value of this particular statement, we can assert that this statement is a member of an ensemble of statements for which it is *guaranteed* that (approximately) 68% are true.

If the frequentist principle were satisfied exactly, we could remove the word “approximately”.

Confidence Interval: $H \to ZZ \to 4l$

The confidence interval algorithm just described approximates the method devised by Neyman (1937):



If $\alpha_L = \alpha_R$ the resulting intervals are called *central intervals*.

But Is The Signal Real?

In the real world, we can never know for sure.

We can, however, make a *probabilistic statement* about whether the signal is real or the result of a fluctuation of the background.

In particle physics, the broad consensus is that we declare a signal real if the background-only hypothesis is extremely unlikely.

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

HYPOTHESIS TESTS
AN EXAMPLE

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 truth of the null hypothesis.

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

Hypothesis Tests – 2

There are two related variations on this general procedure:

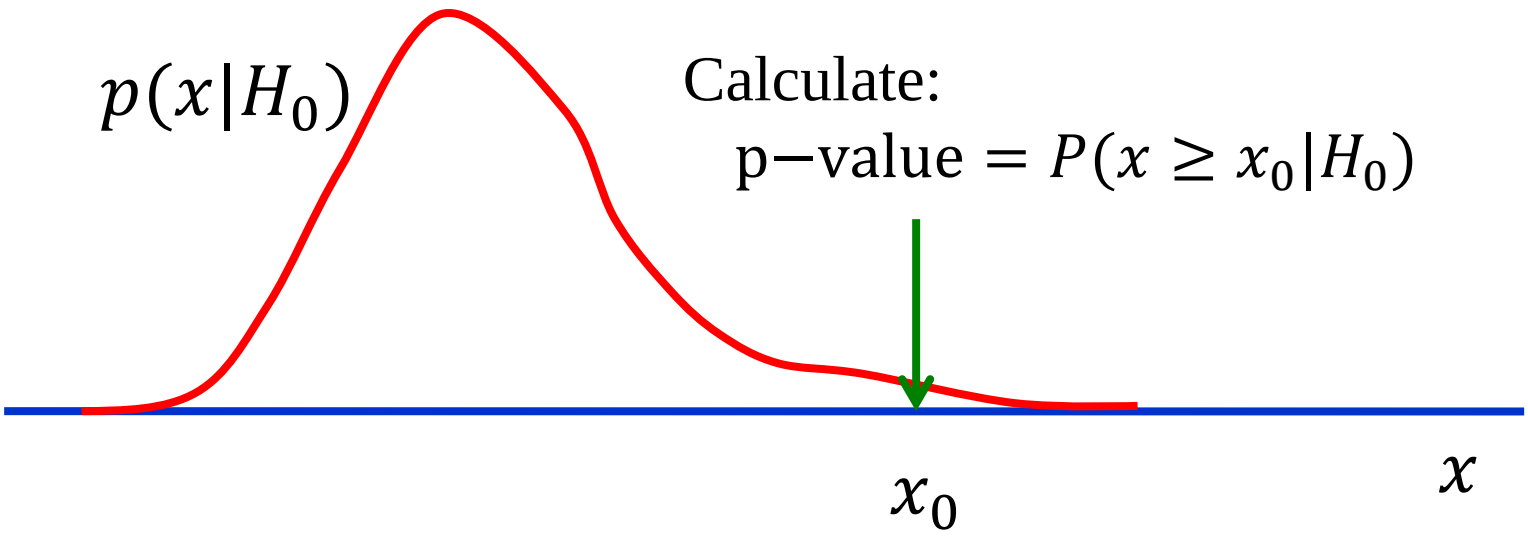
1. Fisher: reject the null if the test statistic is large enough.

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

In particle physics, we do a mixture of both!

Hypothesis Tests – 3

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

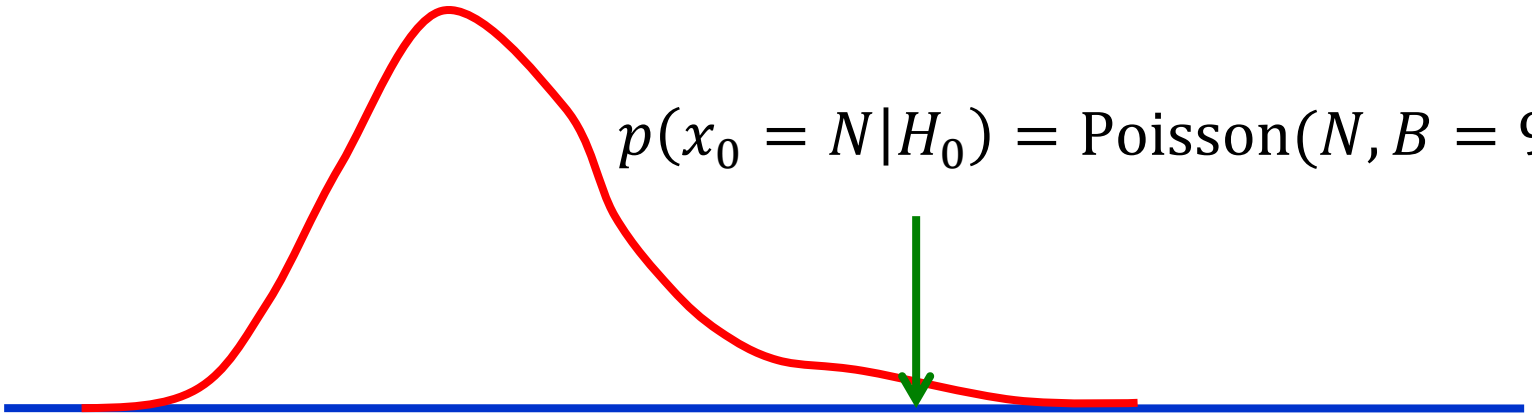


x_0 is the *observed* value of the test statistic x.

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)



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

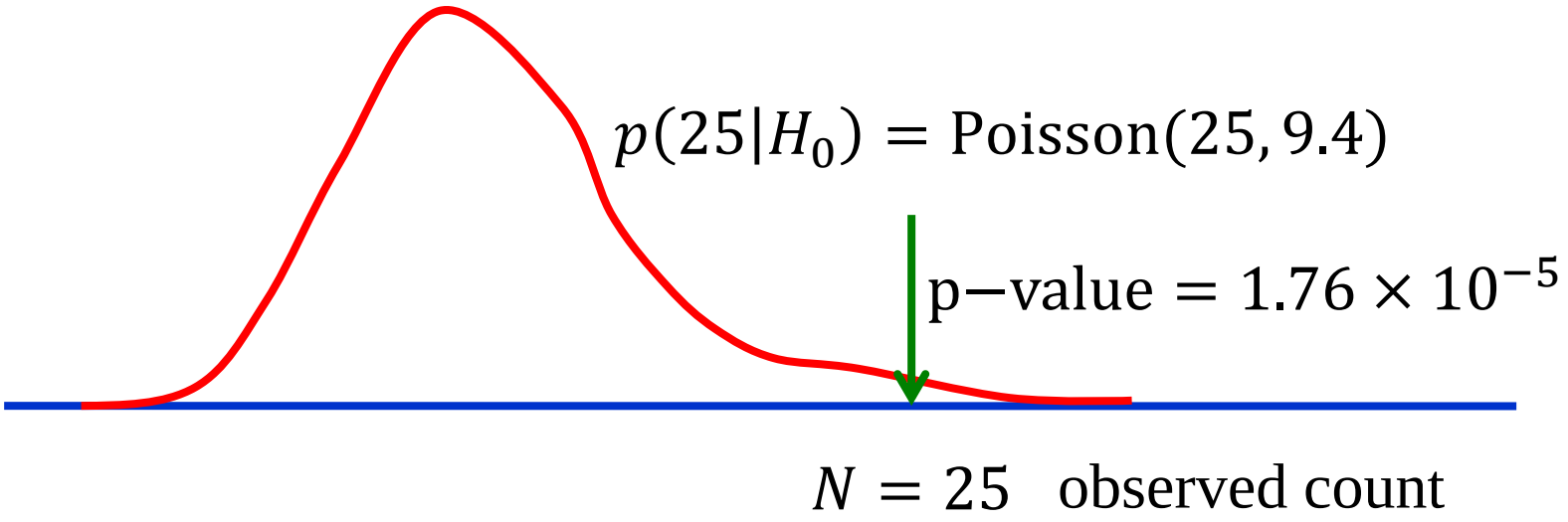
$N = 25$ observed count

$$\text{p-value} = \sum_{k=N}^{\infty} \text{Poisson}(k, 9.4) = 1.76 \times 10^{-5}$$

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

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

Background, B = 9.4 events (ignoring uncertainty)



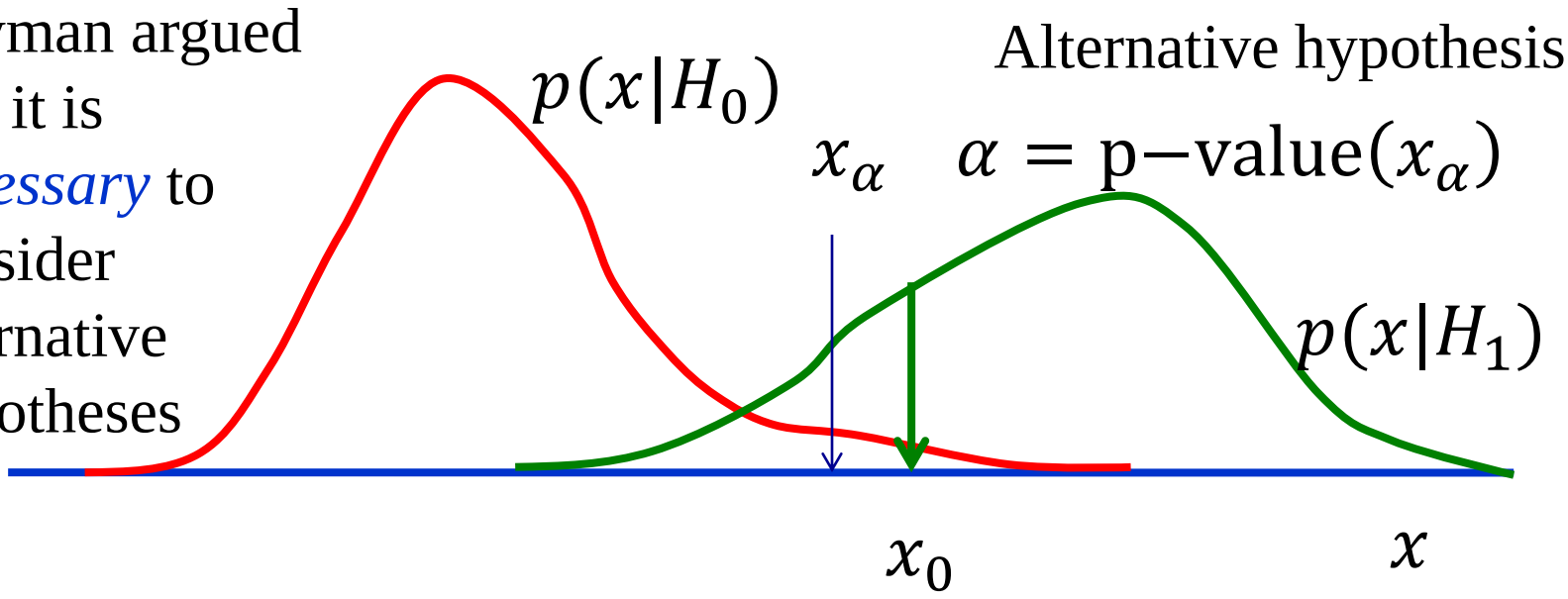
We often map a p-value to a *Z*-value, that is, to the number of standard deviations *away from the null* if the distribution were a Gaussian. This yields $Z = 4.14$.

We say we have a 4.14σ signal.

Hypothesis Tests – 4

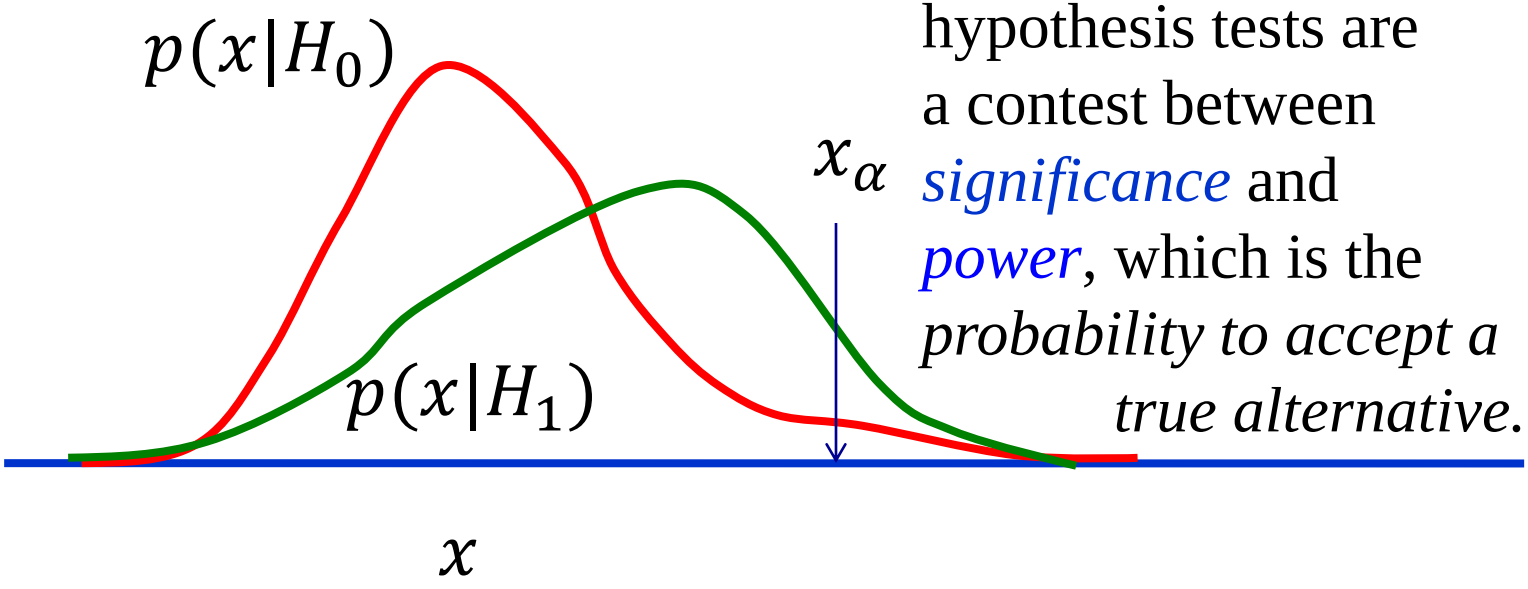
Neyman's Approach: *Null* hypothesis (H_0) + alternative (H_1)

Neyman argued that it is *necessary* to consider alternative hypotheses H_1



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



In Neyman's approach, hypothesis tests are a contest between *significance* and *power*, which is the *probability to accept a true alternative.*

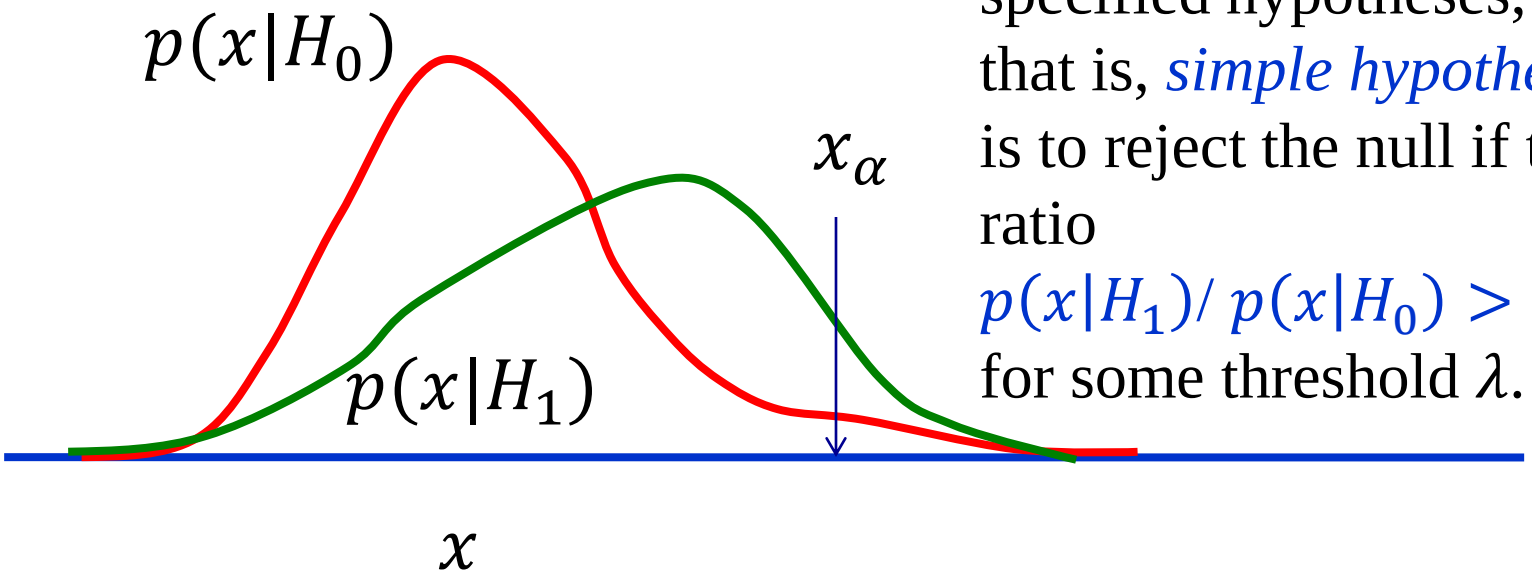
$$\alpha = \int_{x_\alpha}^{\infty} p(x|H_0)dx$$

significance of test

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

power of test

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 in order to perform an hypothesis test.

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)

In the 2-parameter likelihood $L(s, \boldsymbol{b}) \equiv p(D \mid s, \boldsymbol{b})$, we replace $\boldsymbol{b}$ by $\hat{\boldsymbol{b}} = \boldsymbol{f}(\boldsymbol{s})$

to get the profile likelihood $L_p(s) = L(s, \boldsymbol{f}(\boldsymbol{s}))$.

We can use the quantity

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

as a *test statistic* to test the null hypothesis $\boldsymbol{s} = \boldsymbol{s_0}$.

Note that the tests are *approximate* because we used an approximation to arrive at the profile likelihood.

For HEP applications, however, the tests work very well.

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

For this example, Wilks' theorem is equivalent to the statement that for large samples the density of the signal estimate $\hat{\boldsymbol{s}}$ is approximately Gaussian.

If, furthermore, $\boldsymbol{s_0}$ is the true value of $\boldsymbol{s}$, then the distribution of $\boldsymbol{x} \equiv \boldsymbol{t}(\boldsymbol{s_0})$ will be approximately a χ^2 density of one degree of freedom.

Therefore, for any value of $\boldsymbol{x}$, the associated p-value($\boldsymbol{x}$) can be computed using the χ^2 density.

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

So, we need to compute

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

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

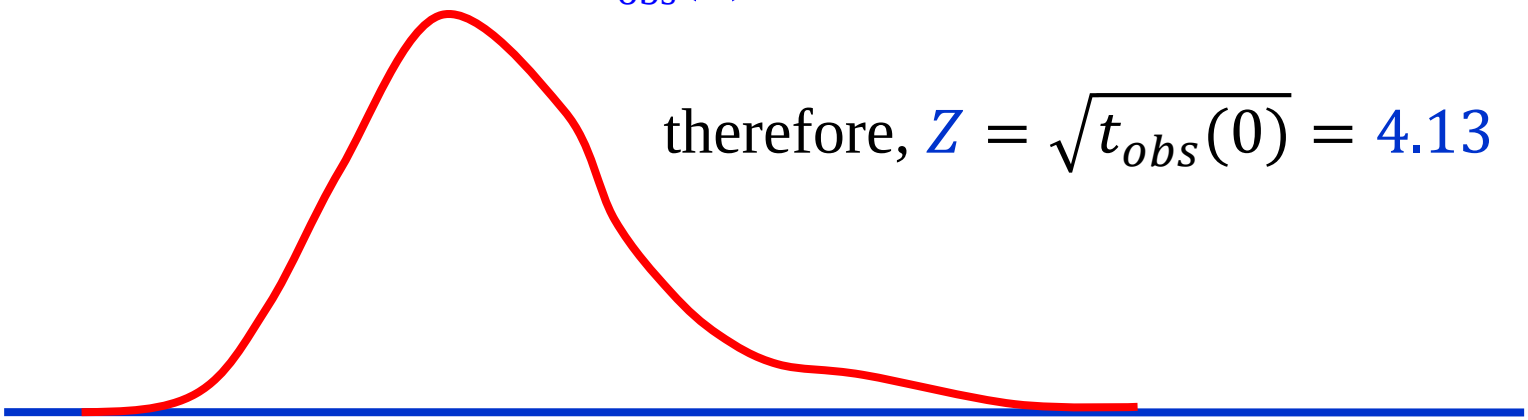
Then, if the p-value < $\boldsymbol{\alpha}$ we reject the $\boldsymbol{s = s_0}$ hypothesis. In addition, the p-value is reported.

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

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

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

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



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

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

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

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

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

Exercise 7: Verify this calculation

Summary

Probability

Interpretations: degree of belief, relative frequency

Likelihood Function

Probability function into which data have been inserted.

Frequentist Principle

Construct statements such that a fraction $f \geq$ C.L. of them are true over an *infinite* ensemble 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 < α; report the p-value.

Bibliography

1. G. Cowan, *Statistical Data Analysis*, Oxford University Press, Oxford (1998).
2. L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, Cambridge (1989).
3. R. J. Barlow, *Statistics: A Guide To The Use Of Statistical Methods In The Physical Sciences*, The Manchester Physics Series, John Wiley and Sons, New York (1989).
4. F. James, *Statistical Methods in Experimental Physics*, 2nd Edition, World Scientific, Singapore (2006).
5. H. B. Prosper, *Practical Statistics for Particle Physicists*, CERN-2015-004 (2016).

Tutorials

Download the tutorials from github using

git clone https://github.com/hbprosper/ENHEP

or update your local copy using git pull

Requirements

1. ROOT (from https://root.cern.ch) 6.14/06 or greater.
2. jupyter, which is needed to run the Python notebooks, with Python 2.7.X (X > 9) or Python 3 provided you replace statements such as **print** "hello" by **print**("hello") (thanks for testing this Ulrich!).
3. Python packages: pandas, numpy, matplotlib, scipy, and scikit-learn; sympy is also very handy.

Tutorials

Statistics notebooks

1. roofit — basics of Python, PyROOT, and RooFit
2. rootN — coverage of statements: $N + \sqrt{N} > \theta$
3. Poisson — Poisson confidence intervals
4. Wilks — demonstration of Wilks' theorem
5. hzz4l — analysis of data $N = 25, B = 9.4 \pm 0.5$
6. Type1afit — fitting ΛCDM model to Type1a Sne data.

Machine learning notebooks

1. higgs_vbf_ggf_bdt — VBF vs. ggF Higgs production
2. higgs_vbf_ggf_dnn — as above, but using a DNN

BACKUP SLIDES

The Poisson Process

The Poisson distribution is also the outcome of a *stochastic process*. Suppose that at time $t + \Delta t$ there are n successes.

Assume the following:

1. The probability of **1** success in the interval $[t, t + \Delta t]$ is $q\Delta t$.
2. The probability of ≥ 2 successes is negligible.

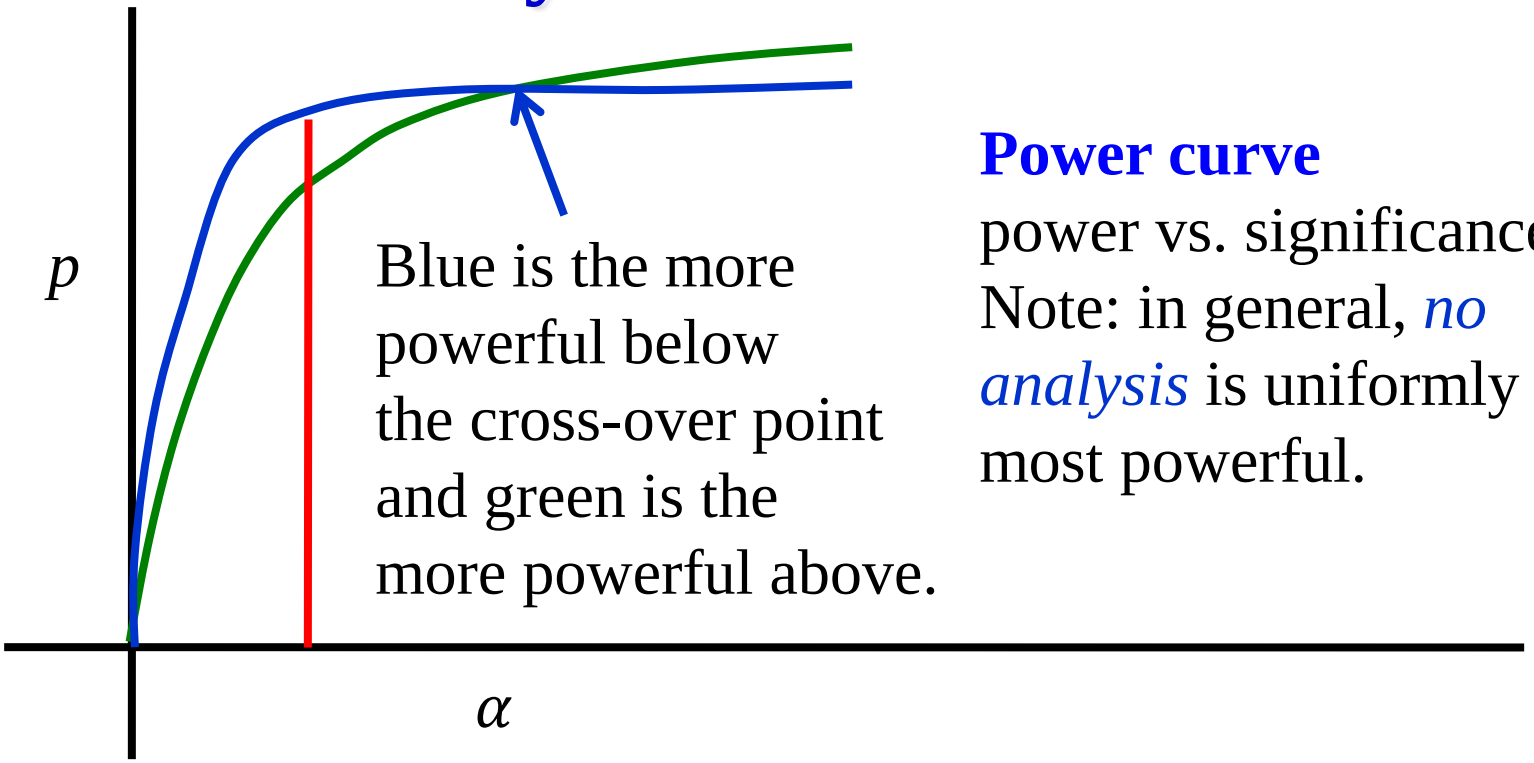
The possible *state transitions*, and *transition probabilities*, are

1. $n - 1 \to n$ $\quad\quad$ $p_{n-1}(t)q\Delta t$
2. $n \to n$ $\quad\quad$ $p_n(t)(1 - q\Delta t)$

Therefore, $p_n(t + \Delta t) = p_n(t)(1 - q\Delta t) + p_{n-1}(t)q\Delta t$, where $p_n(t)$ is the probability to have n successes by time t.

Solving the 1st order ODE yields Poisson(qt) for constant q.

The Neyman-Pearson Test



Blue is the more powerful below the cross-over point and green is the more powerful above.

Power curve
power vs. significance.
Note: in general, *no analysis* is uniformly the most powerful.

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

significance of test

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

power of test

The Bayesian Approach in a Nutshell!

Bayesian methods are

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

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

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*

Nuisance parameters are removed by marginalization.

Example: Bayesian Analysis $H \rightarrow 4l$

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
$M = 353.4$	effective background event count
$k = 37.6$	effective background scale factor

unknowns:

b	expected background count
s	expected signal count
$d = s + b$	expected event count

Example: Bayesian Analysis $H \to 4l$

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 $H \rightarrow 4l$

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*.

We have already calculated the marginal likelihood and found

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

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

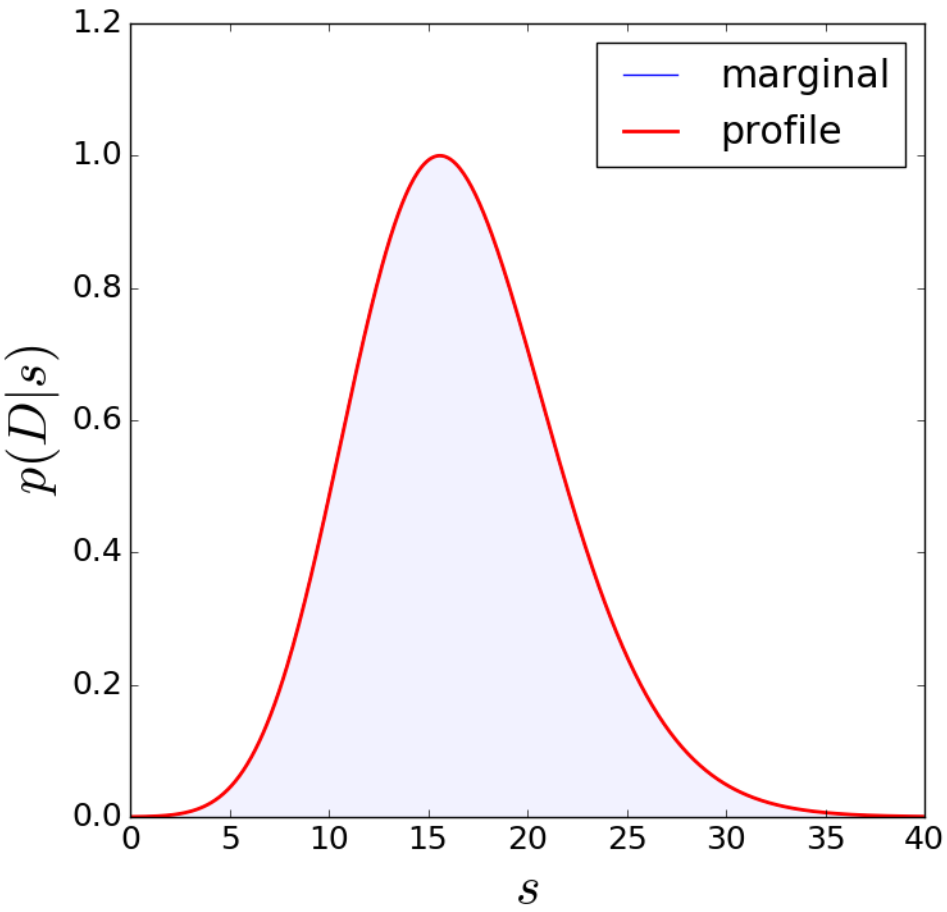
*Luc Demortier, Supriya Jain, Harrison B. Prosper,
Reference priors for high energy physics, Phys.Rev.D82:034002 (2010)

Example: Bayesian Analysis $H \to 4l$

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

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 9: 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

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

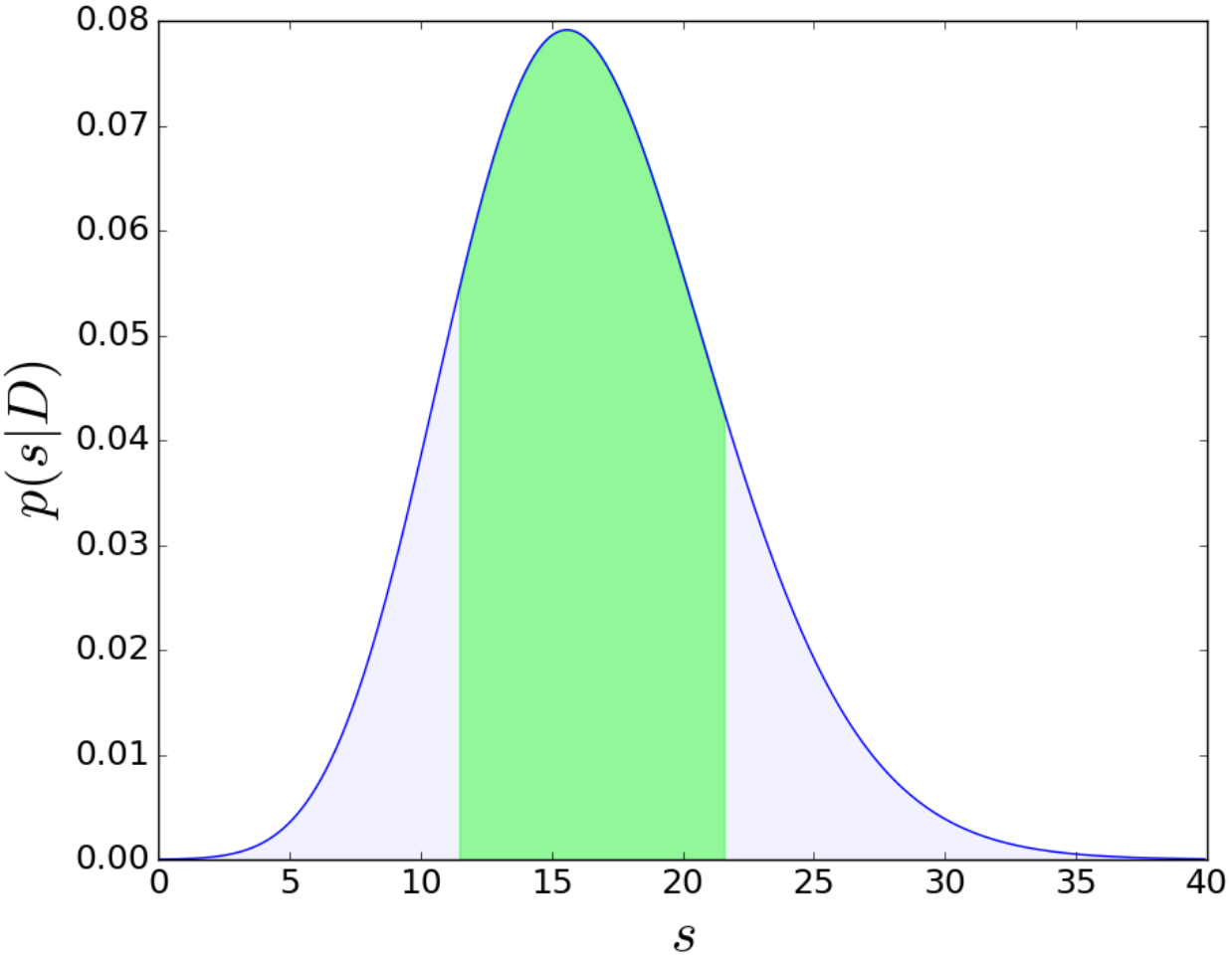
$$\int_0^{u(N)} p(s\,|D)\,ds = (1 + CL)/2$$

with CL = 0.683, we obtain $s \in [11.5, 21.7]$ at 68% 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 $H \rightarrow 4l$

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 $\pi\,(s\,|\,H_1) = \delta(s - 15.6)$, 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 $H \rightarrow 4l$

From

$p(D \mid H_1) = 7.91 \times 10^{-2}$ and

$p(D \mid H_0) = 1.59 \times 10^{-5}$

we conclude that the 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

$$Z = \text{sign}(\ln B_{10})\sqrt{2\lceil \ln B_{10} \rceil}$$

This yields Z = 4.13.

Exercise 10: Verify this number

Generalization to Multiple Bins

The generalization to so-called "shape" analyses, that is, to multiple bins introduces no new concepts.

Here is a model for M independent bins, each with N sources:

1. Mean count in ith bin: $d_i = \sum_{j=1}^{N} p_j a_{ji}$, where each bin contains N sources with mean counts a_{ji}. The p_j are parameters such as the signal strength μ.
2. Likelihood for ith bin: $p(D_i|d_i) = \text{Poisson}(D_i, d_i)$.
3. Likelihood for ith bin of jth source: $p(A_{ji}|r_{ji}a_{ji}) = \text{Poisson}(A_{ji}, r_{ji}a_{ji})$, where r_{ji} are known scale factors.

Generalization to Multiple Bins

The overall probability model is

$$p(D|a) = \prod_{i=1}^{M} p(D_i|d_i) \prod_{j=1}^{N} p\left(A_{ji}|r_{ji}a_{ji}\right)$$

which can be marginalized with respect to $\boldsymbol{a_{ji}}$ *exactly**:

$$p\left(D\middle|p_j, r_{ji}\right) = \prod_{i=1}^{M} \sum_{k_1 \ldots, k_N = 0}^{D_i} \prod_{j=1}^{N} \binom{A_{ji}+k_j}{k_j} \frac{{p_j}^{k_j} {r_{ji}}^{A_{ji}+k_j}}{\left(p_j + r_{ji}\right)^{k_j}}$$

with $k_1 + \cdots + k_N = D_i$.

If the scale factors r_{ji} are not known precisely, the above can be extended to incorporate the appropriate uncertainties.

*Former FSU undergraduate Robert Orlando.