STATISTICS AND MACHINE LEARNING 1

Harrison B. Prosper

Florida State University

Topics

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

Jupyter Notebooks

The only way to learn something well is by "doing"!

Therefore, I encourage you to try the jupyter notebooks at

https://github.com/hbprosper/AEPSHEP

Do

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

to download them. The GitHub repo will be ready later today.

INTRODUCTION

Introduction: Sample

Fundamental Assumption of Statistics: data are *randomly sampleed*.

A *statistic* is a function of a data *sample*, $\boldsymbol{x} = x_1, x_2, \ldots x_n$.

Here are some well-known statistics:

sample average $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

sample variance $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

sample moments $m_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$

Introduction: Population

An *infinitely* large sample is called a *population*.

A population is an *abstraction*.

But, like many abstractions, we can study populations mathematically and we can study them *approximately* by simulating large samples.

Introduction: Population

A few characteristics of populations

Expected Value	$E[x]$
Mean	μ
Error	$\epsilon = x - \mu$
Mean Square Error	$MSE = E[\epsilon^2]$
Bias	$b = E[x] - \mu$
Variance	$V[x] = E[(x - E[x])^2]$

These characteristics are also abstractions!

Introduction: Statistical Inference

The main goal of *statistical inference* in high-energy physics is to use a data *sample* to infer interesting attributes of the associated *population*. These attributes are typically physical parameters such as particle masses.

Important point to note:

- In statistics, there is no such thing as "the right answer".
- Rather, there are many answers based on different assumptions and different opinions about which assumptions are reasonable.

Introduction: Statistical Inference

However, everyone agrees that the key concept in statistics is *probability*, which is interpreted in at least two ways:

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

   statement: it will rain tomorrow.
   probability: $p = 2\times10^{-3}$

This interpretation of probability is the basis of the *Bayesian approach* to statistical inference.

Introduction: Statistical Inference

2. **Relative frequency** of a given outcome in a long sequence of trials, e.g.:

   trial: a proton-proton collision at the LHC

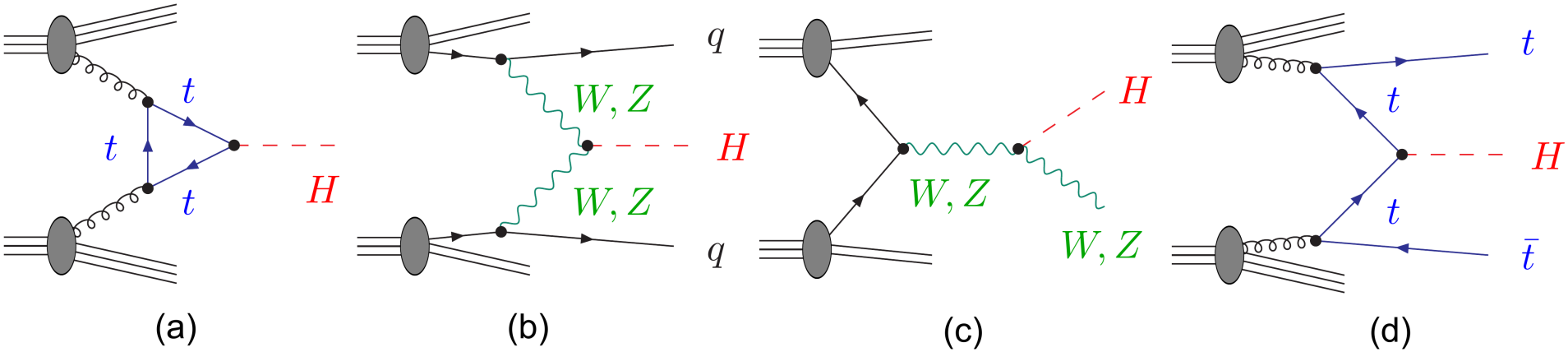
   outcome: creation of a Higgs boson

   probability: $p = 5\times10^{-10}$

This interpretation of probability is the basis of the *frequentist approach* to statistical inference.

FREQUENTIST ANALYSIS (1)
EXAMPLE 1

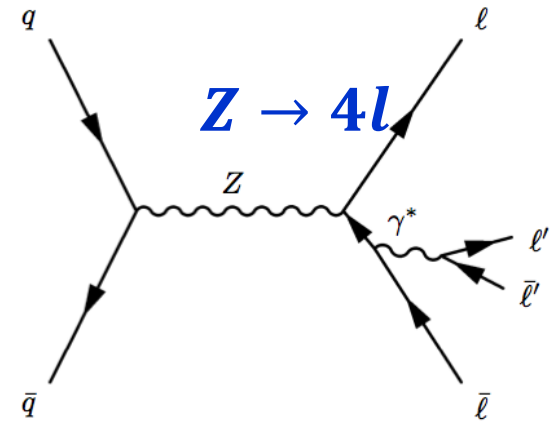
LHC: $pp \rightarrow H \rightarrow ZZ \rightarrow 4l$

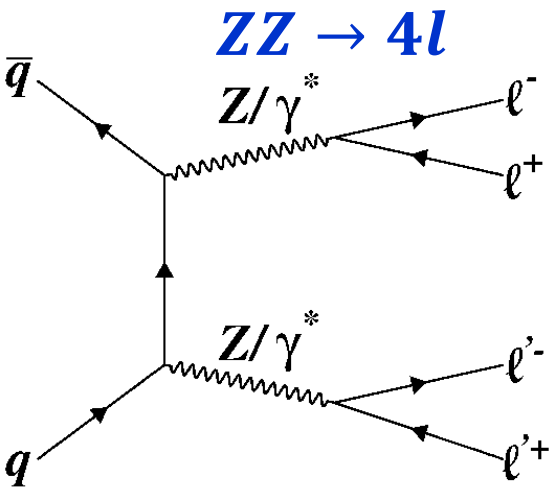


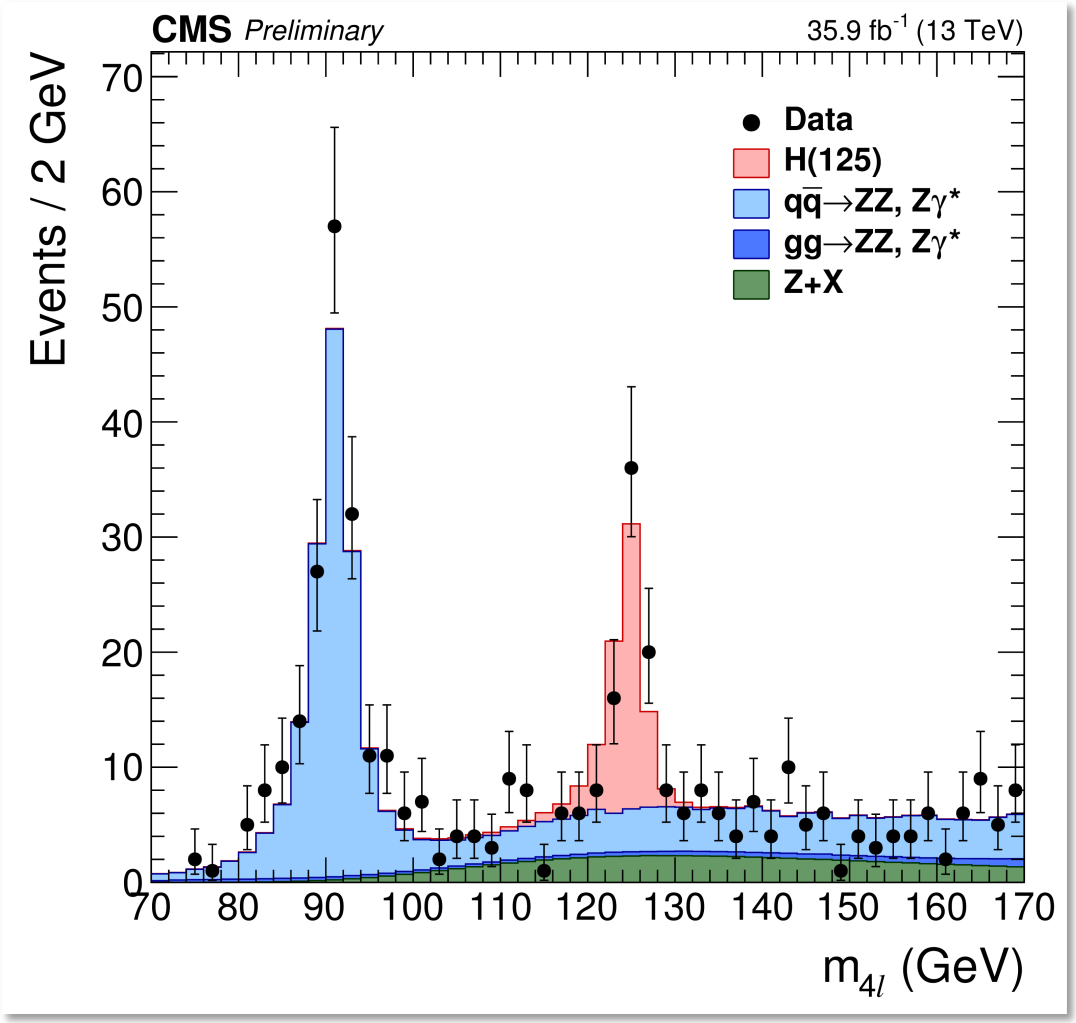
Process		$\sigma \times BR$ (fb)
(a) Gluon-gluon fusion	(ggF)	12.18
(b) Vector-boson fusion	(VBF)	1.044
(c) Associated production	(VH)	1.047
(d) Top anti-top fusion	(ttH)	0.393

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

The main backgrounds:







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

Let's consider some results published by the CMS Collaboration in 2014 ($pp \rightarrow H \rightarrow ZZ \rightarrow 4l$ (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

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

Our goals:

1. Estimate (i.e., measure) the mean signal, ***s***.
2. Quantify the accuracy of the estimate.
3. Quantify how confident we are that the signal is real.

In order to do the above, we need to construct a *probability (or statistical) model* of the mechanism that generated the data.

Let's start from the very beginning…

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

If we assume:

1. The probability ***p*** of a success is the same for every proton-proton collision (trial).
2. A success ***S*** and a failure ***F*** are *exhaustive* and *mutually exclusive*.
3. Every sequence of collisions (trials) is equally probable.

Then the probability of k successes in n trails is

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

that is, the *binomial distribution*, $\text{Binomial}(k, n, p)$.

Bernoulli Trial (3) : $H \rightarrow ZZ \rightarrow 4l$

The mean number of successes a is

$a = pn$. **Exercise 1**: Show this

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

Let's, therefore, consider $p \rightarrow 0$ and $n \rightarrow \infty$, with *a constant*,

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

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

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

Probability Model:

The probability to observe n events is, therefore,

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

where s and b are the *mean* signal and background counts, respectively.

Likelihood Function:

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

The *likelihood function* is simply the *probability model* into which data have been entered.

But what about $B \pm \delta B = 9.4 \pm 0.5$?

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

We need more assumptions! (Or we need to study in detail how CMS arrived at $B \pm \delta B = 9.4 \pm 0.5$.)

Let's assume that

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

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

M could be the result of a Monte Carlo (MC) simulation of the background or the event count in a background-dominated sample. Let's also assume that the probability model for M is a Poisson with mean $\approx M$ and standard deviation $\approx \sqrt{M}$.

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

Solving for M and k, we get $M = 353$, $k = 37.6$.

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

Given the last assumption, the likelihood for the count M is

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

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

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

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

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

Now that we have our statistical model, $p(D|s,\ b)$, we can answer the questions:

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 is real?

Maximum Likelihood

1. How does one estimate (measure) the mean signal, *s*?

The standard answer to this question is to choose as estimates of the parameters the values that *maximize the likelihood*:

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

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

For this example, we find the unsurprising results:

$$\hat{s}(D) = N - B, \qquad \hat{b}(D) = B$$

The Frequentist Principle

2. How does one quantify the accuracy of the estimate?

A general answer to this question was proposed by Jerzy Neyman in 1937:

Statistical statements should be constructed with the guarantee that a fraction $f \geq p$ of them are true over a *population* of statements where p is chosen *a priori*.

This is called the *frequentist principle* (FP). The fraction f is called the *coverage probability* (or *coverage* for short) and p is called the *confidence level* (CL).

The Frequentist Principle

Example 1

Consider statements of the form $\theta < N + \sqrt{N}$, 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. *Note: each statement is either true or false.*

In a real experiment, we do not know which statements are true and which are false, but we do in a simulation. So we can compute the coverage ***f*** and determine ***p***.

Exercise 3: Estimate by simulation the coverage probability of these statements. Repeat using uniform(0, 10). Then repeat for *fixed* values of θ in steps of 0.2 from 0.1 to 9.9 and plot the coverage versus θ. What is ***p***?

The Frequentist Principle

Example 2

Consider $x = D$ sampled from a Gaussian statistical model

$$p(x|\mu,\sigma) = \exp\left(-\frac{\chi^2}{2}\right)/(\sigma\sqrt{2\pi}), \qquad \chi^2 = \frac{(x-\mu)^2}{\sigma^2}$$

with *known* standard deviation σ but *unknown* mean μ.

The MLE of μ is $\hat{\mu}(D) = D$. According to Neyman, we should quantify the accuracy of the estimate with a statement of the form

$$\mu \in \left[\underline{\mu}(x), \overline{\mu}(x)\right]$$

with a specified *confidence level*.

The Frequentist Principle

Example 2

For a Gaussian, the standard method for constructing such a statement is to solve the equation

$$\chi^2 = -2 \ln p(x|\mu, \sigma) = \mathbf{1}$$

The solutions are $\underline{\mu}(x) = x - \sigma$ and $\overline{\mu}(x) = x + \sigma$.

A statement of the form $\mu \in \left[\underline{\mu}(x), \overline{\mu}(x)\right]$ is either true or false. Consider a large number of experiments, each yielding an *confidence interval* $\left[\underline{\mu}(D), \overline{\mu}(D)\right]$, which varies *randomly* from one experiment to another.

The Frequentist Principle

Example 2 (contd.)

For the Gaussian, the coverage probability ***f*** of statements of the form

$$\mu \in \left[\underline{\mu}(x), \overline{\mu}(x)\right] \text{ is } \mathbf{0.683}.$$

Moreover, for any given point μ, the coverage probability never to falls below 0.683.

Therefore, the *confidence level* (CL) associated with the above statements is $100\boldsymbol{f}\,\% = 68\%$.

Ideally, we would like to arrive at similar statements in our Higgs boson example.

The Frequentist Principle

Example 3

Our (greatly simplified) Higgs boson likelihood

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

contains two parameters ***s*** and ***b.***

Suppose we want to make statements about *both* parameters such as the following: $s, b \in R(D)$ except that this time $R(D)$ is not a confidence interval but rather a *confidence set*.

How do we construct such a set?

Wilks' Theorem (1)

If certain conditions are met (e.g., we have enough data, estimates do not occur on the boundary of the parameter space) then the quantity $t(x) = -2 \ln \lambda(x)$

where,

$$\lambda(x) = \frac{p(x|s,\ b)}{p\left(x|\hat{s},\hat{b}\right)}$$

has a distribution that approximates a χ^2 *density* of **2** degrees of freedom (because there are **2** free parameters).

This is a special case of Wilks' Theorem (1938)*.

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

Wilks' Theorem (2)

If we want to create confidence sets with a confidence level of 68%, Wilks' theorem suggests that we construct the set by finding all points (s, b) that satisfy the inequality

$$t(D) \approx \chi^2_2 \leq \mathbf{2.296}$$

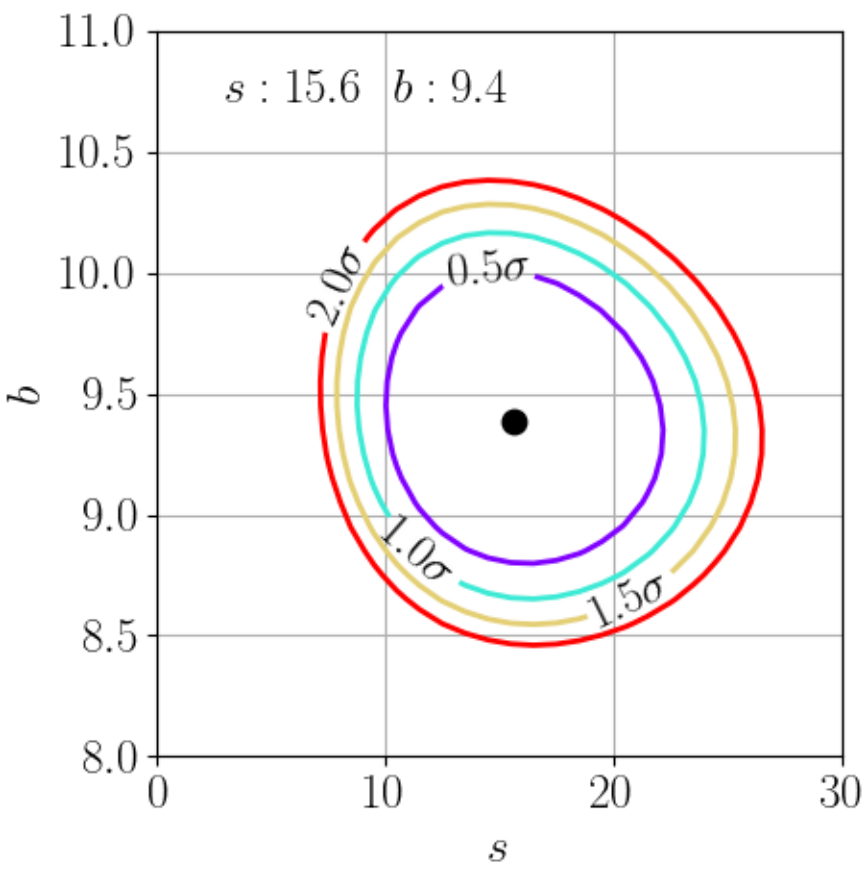
for observed data $x = D$, or, equivalently, from the inequality

$$C_2\big(t(D)\big) \leq \mathbf{0.683}$$

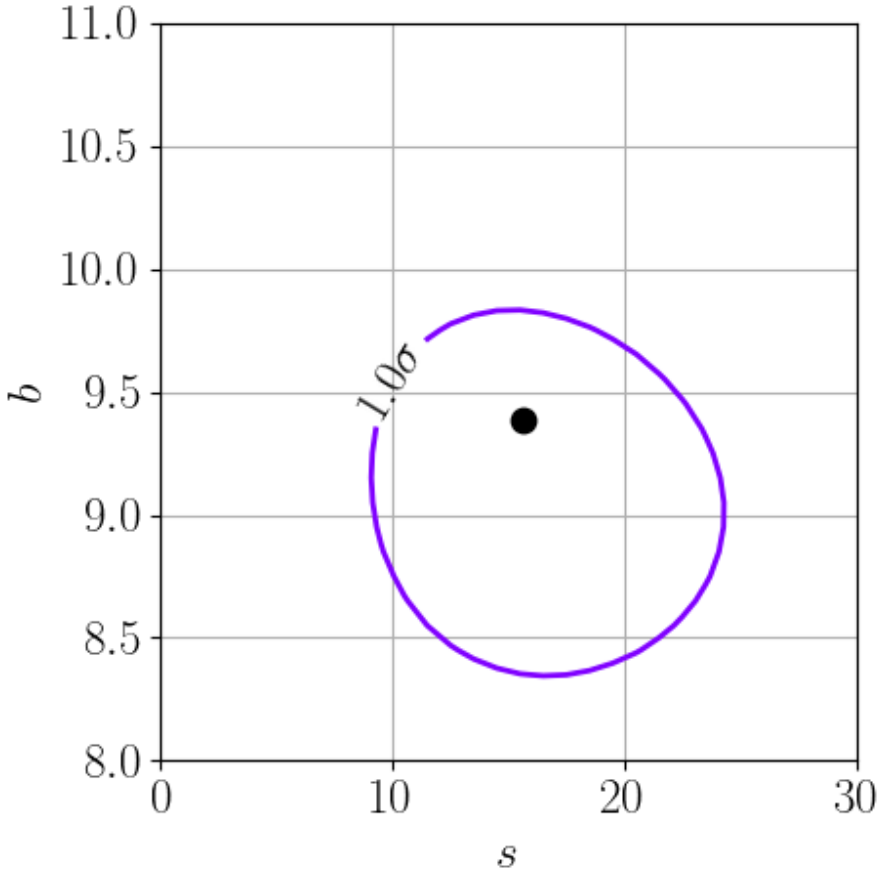
where $C_2\big(t(D)\big) = \int_0^{t(D)} p_2(z)dz$ is the *cumulative distribution function* of the χ^2_2 density.

Wilks' Theorem (3)

Confidence sets



Checking coverage

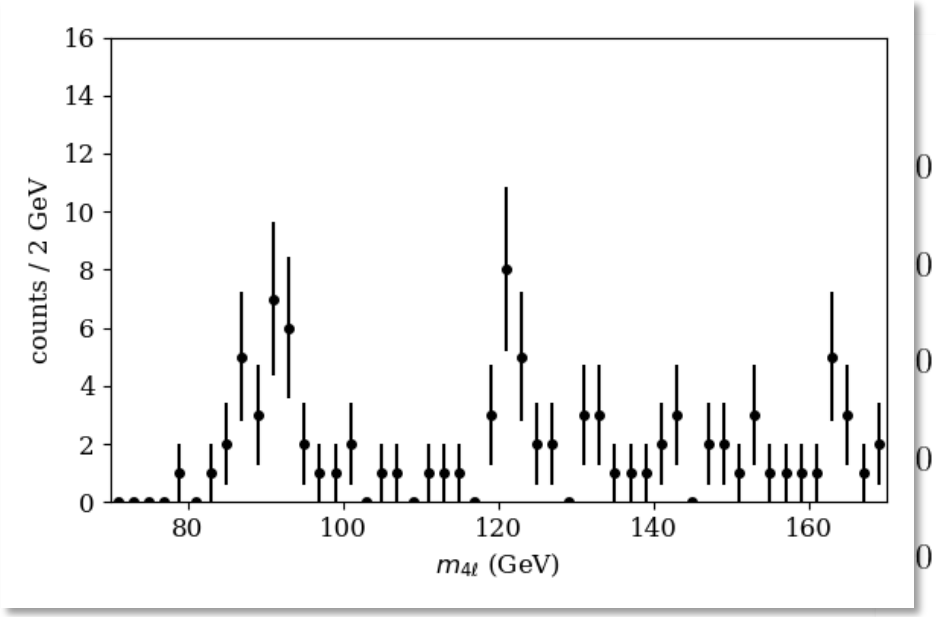


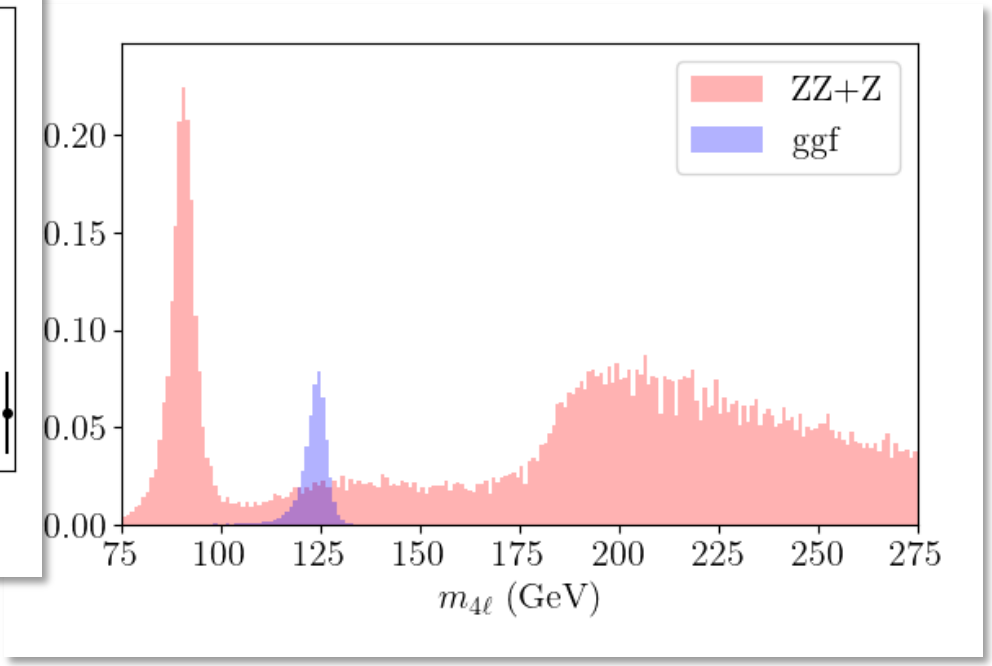
FREQUENTIST ANALYSIS (1)
EXAMPLE 2

Multi-bin Analysis: $pp \rightarrow 4l$

The analysis of a binned spectrum (a "shape" analysis) is a direct generalization of the double-Poisson model:

$$p(D|s,b) = \prod_{i=1}^{K} \frac{(\mu s_i + b_i)^{N_i} e^{-(\mu s_i + b_i)}}{N_i!} \frac{(k_i b_i)^{M_i} e^{-k_i b_i}}{M_i!} \frac{(l_i s_i)^{Q_i} e^{-l_i s_i}}{Q_i!}$$





https://atlas-opendata.web.cern.ch/atlas-opendata/samples/2020/4lep/

Un-binned Analysis: $pp \rightarrow \gamma\gamma$

If one let's the bin size go to zero, then one arrives at an un-binned model, e.g.:

$$p(D|\theta) = \exp[-(b+s)] \prod_{i=1}^{K} (1-\epsilon) f_b(x_i, \theta_b) + \epsilon f_s(x_i, \theta_s)$$

which can be used to fit, for example, the CMS Higgs boson discovery data. See example on GitHub.

Summary

Probability

Interpretations: degree of belief, relative frequency

Likelihood Function

Statistical model into which data have been inserted.

Frequentist Principle

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