Practical Statistics for Particle Physicists

Lecture 1

Harrison B. Prosper

Florida State University

ESHEP 18, Maratea, Italy

26 June, 2018

Tutorials on github

Download:

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

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

Dependencies:

- Python 2.7.x, ROOT (6.12.x), jupyter
- numpy, pandas, matplotlib, scikit-learn

Tutorials on github

stats

00.1_roofit.ipynb
00.2_roofit.ipynb

07.1_rootn.ipynb
07.2_poisson.ipynb
07.3_poisson.ipynb

09_wilks.ipynb
11_16_hzz4l.ipynb

ml

00_prepare_hzz_data.ipynb

01_bdt_wine.ipynb
01_dnn_wine.ipynb

02_dnn_hzz_vbf_ggf.ipynb

03_hmc.ipynb, hmc.ipynb

Outline

- Lecture 1
  - Introduction
  - The Frequentist Principle
  - Confidence Intervals
  - The Profile Likelihood
- Lecture 2
  - Hypothesis Tests
  - Introduction to Bayesian Inference
- Lecture 3
  - Introduction to Machine Learning

Outline

- Lecture 1
  - Introduction
  - The Frequentist Principle
  - Confidence Intervals
  - The Profile Likelihood
- Lecture 2
  - Hypothesis Tests
  - Introduction to Bayesian Inference
- Lecture 3
  - Introduction to Machine Learning

Introduction: Samples

Definition: A **statistic** is any function of the data sample, $\boldsymbol{x} = x_1, x_2, \ldots x_n$. Here are some simple examples:

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

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

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

These quantities are computable because the sample is *known*.

Introduction: Populations – 1

Now consider an *infinitely* large sample. This is referred to as a population.

A population is clearly an *abstraction*, and exists only in the sense that the set of real numbers exist.

But, like many abstractions, we can study this one both mathematically and *approximately* through simulation.

Introduction: Populations – 2

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

Introduction: Populations – 3

$$MSE = E[\epsilon^2] = V + b^2$$

Exercise 1: Show this

The **MSE** is the most widely used measure of how close an **ensemble** of statistics $\{\mathbf{x}\}$ is to the population mean $\boldsymbol{\mu}$.

The **root mean square** (RMS) is $RMS = \sqrt{MSE}$

Introduction: Populations – 4

Consider the expected value (or ensemble average) of the *sample* **variance**

$$E[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}x\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} x^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(E\left[x_i^2\right] - E[x^2]\right)$$

$$= E[x^2] - E[x^2]$$

Introduction: Populations – 5

The expected value of the sample variance (as we have defined it) is *biased*

$$E[S^2] = E[x^2] - E[x^2]$$
$$= \mathrm{V}[x] + b$$

where $b = -\frac{1}{n} V[x]$

Exercise 2: Show this.
Hint:
write $E[x^2]$ in terms of $E[x^2]$ and $E[x]$

Introduction: Statistical Inference

The main goal of statistical inference is to use a sample to infer something about its associated population.

But, note:

- Statistics is not physics. The great thing about physics is that there is a single supreme judge of its correctness, namely, Nature. Statistics has no such judge!
- Consequently, in statistics, there is no such thing as "the right approach"; rather, there are different approaches, with different assumptions, and different opinions about them.
- The only assumption that is universally accepted is that the foundation of statistics is **probability**.

Introduction: Statistical Inference

Probability is an abstraction, with several *interpretations* of which the two most important are:

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

   proposition: It will rain in Maratea tomorrow
   probability: $p = 5\times10^{-2}$

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

   trial: a proton-proton collision at the LHC

   outcome: creation of a Higgs boson

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

https://plato.stanford.edu/entries/probability-interpret/

Example $H \rightarrow ZZ \rightarrow 4l$

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

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

unknowns:

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

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

In order to infer something about the population associated with the 4-lepton events, which in this simple example is characterized by two parameters ***s*** and ***b***, we need first to construct a probability model of the data generation mechanism.

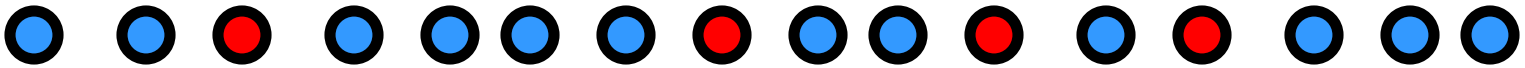
Let's start at the very beginning…

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

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 of interest happens (***S***) or does not happen (***F***).



What is the probability of this sequence of events? There is ***no*** answer unless we are prepared to make assumptions.

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

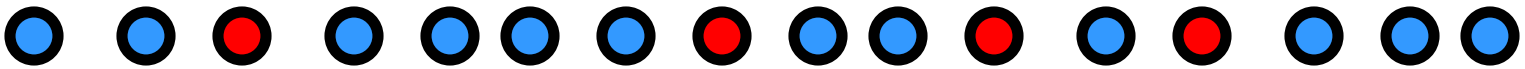
Assumption 1: Let ***p*** be the probability of a success.

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

Assumption 3: ***S*** and ***F*** are *exhaustive* and *mutually exclusive*. Therefore, the probability of a failure is $\mathbf{1 - p}$.

Therefore, for a given sequence ***S*** of ***n*** trials, the probability $P(\mathbf{k}\,|S, p, n)$ of *exactly* ***k*** successes and *exactly* $\mathbf{n - k}$ failures is

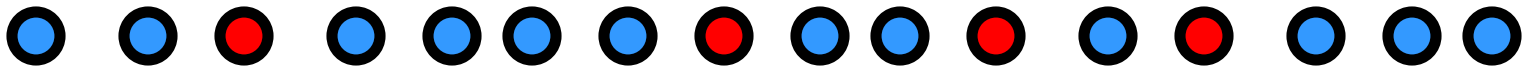
$$P(k|\,S, p, n) = p^k (1-p)^{n-k}$$



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

Assumption 4: The specific sequence ***S*** at the LHC is irrelevant. Therefore, according to probability theory, we can eliminate the (discrete) parameter ***S*** from the probability model by ***summing*** over all potential sequences:

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



Note, there are $\binom{n}{k}$ sequences with the same outcomes.

Therefore,

$P(k|\,p,n) = \binom{n}{k} p^k (1-p)^{n-k}$, that is, we arrive at the **binomial distribution**, Binomial(k,n,p).

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

The mean number of successes ***a*** is

$a = p\,n$. **Exercise 4**: Prove it

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

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

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

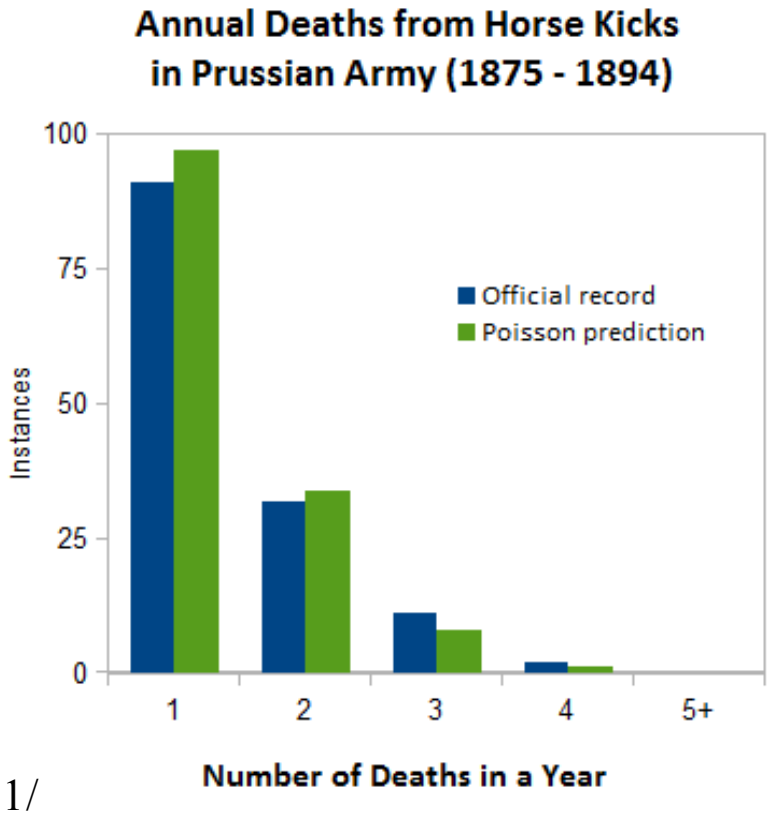
Exercise 5.1: Show that Binomial$(k, n, p) \rightarrow$ Poisson(k, a)

An Aside: The Poisson Distribution

The Poisson distribution is a good model for the probability of rare events.

In 1898, von Bortkiewicz published data on the number of Prussian soldiers kicked to death per year by horses.

He noticed that the observed counts could be modeled by the distribution first described in 1837 by Siméon Poisson (1781 – 1840).



https://mindyourdecisions.com/blog/2013/06/21/what-do-deaths-from-horse-kicks-have-to-do-with-statistics/

An Aside: The Poisson Distribution

The Poisson distribution is the most widely used distribution in particle physics and astronomy because we love to ***count*** rare things.

Astronomers count photons, while we count, for example, Higgs bosons.

The mean of Poisson(k, $\boldsymbol{a}$) is $\boldsymbol{a}$ and its variance is also $\boldsymbol{a}$.

To see this, first compute

$$M(x) = \sum_{k=0}^{\infty} e^{kx}\,\text{Poisson}(k, a) = \exp(-a + a\exp(x))$$

that is, its moment generating function.

Exercise 5.2: Show this

An Aside: The Poisson Distribution

Given $M(x)$, the moments are given by

$$m_r = \sum_{k=0}^{\infty} k^r \,\text{Poisson}(k, a) = \frac{d^r}{dx^r} M(x)_{x=0}$$

You can verify the results using the Python package sympy:

```
import sympy as sp
x, a  = sp.symbols("x, a")          # define variables
M     = sp.exp(-a + a*sp.exp(x))    # define function
m1    = sp.diff(M, x, 1)            # take first derivative
m2    = sp.diff(M, x, 2)            # take second derivative
sp.limit(m1, x, 0)                  # a
sp.limit(m2, x, 0)                  # a**2 + a
```

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

Probability Function:

The probability to observe a count n is

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

Likelihood Function:

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

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

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

What likelihood should we write down for these data?

Introduction: 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 scale factor ***k***

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

and that the count M is sampled from a Poisson distribution with variance $\approx M$. We can then solve for M and k to get

$$\boldsymbol{M} = 353.4, \ \boldsymbol{k} = 37.6.$$

Therefore, the likelihood function for the count M is

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

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

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

The full likelihood is then

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

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

where the data are $D = (N, M, k)$.

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

Given the likelihood function, we can answer several questions including:

1. How do I estimate (measure) a parameter?
2. How do I quantify its accuracy?
3. How do I test an hypothesis?
4. How do I quantify the significance of a result?

Recap: Writing down the likelihood function requires:

1. Identifying all that is *known*, e.g., the observations
2. Identifying all that is *unknown*, e.g., the parameters
3. Constructing a probability model *for both* and plugging data into it.

Outline

- Lecture 1
  - Introduction
  - The Frequentist Principle
  - Confidence Intervals
  - The Profile Likelihood
- Lecture 2
  - Hypothesis Tests
  - Bayesian Inference
- Lecture 3
  - Introduction to Machine Learning

The Frequentist Principle

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

Construct statements such that a fraction $f \geq p$ of them are *guaranteed* to be true over an *ensemble* of statements.

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

An ensemble of statements that obey the frequentist principle is said to cover.

The Frequentist Principle

Points to Note:

1. The FP applies to *real* ensembles, not just the *virtual* ones we simulate on a computer. Moreover, the ensembles can contain statements about different quantities.

   Example: all published measurements x, since 1897, of the form $l(x) \leq \theta \leq u(x)$, where θ is a parameter of interest.

2. Coverage is an *objective* characteristic of ensembles of statements. However, in order to *verify* whether an ensemble of statements covers, we need to *know* which statements are true and which ones are false. Alas, this information is generally not available in the real world.

The Frequentist Principle

Example

Consider an ensemble of *different* experiments, each with a *different* mean count θ, and each yielding a count N. Each experiment makes a single statement of the form

$$N + \sqrt{N} > \theta,$$

which is either True or False.

Obviously, some fraction of these statements are true.

But if we don't know which ones, we have no *operational* way to compute the coverage, that is, the fraction of true statements within the ensemble.

The Frequentist Principle

Example continued

Suppose each mean count θ is randomly sampled from uniform(0, 3), and suppose we *know* these numbers.

Since we know the numbers, we can compute the coverage probability ***f***.

Exercise 7.1:
Show that the coverage of statements of the form $N + \sqrt{N} > \theta$ is ~ 0.62

Outline

- Lecture 1
  - Introduction
  - The Frequentist Principle
  - Confidence Intervals
  - The Profile Likelihood
- Lecture 2
  - Hypothesis Tests
  - Bayesian Inference
- Lecture 3
  - Introduction to Machine Learning

Confidence Intervals – 1

Consider an experiment that observes ***N*** events with expected (i.e., mean) count ***s***.

In 1937, the statistician Jerzy Neyman devised a way to make statements of the form

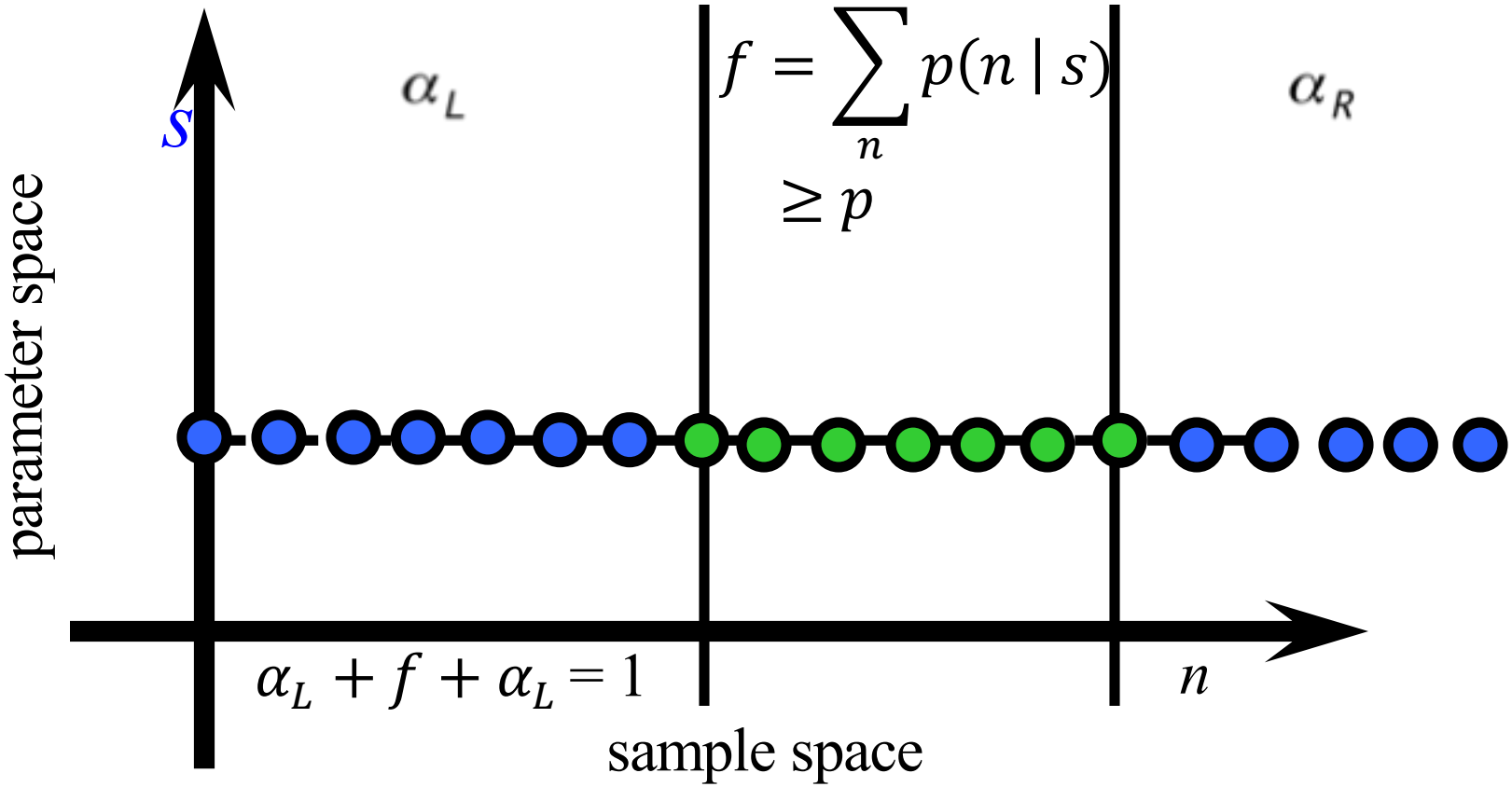
$$s \in [l(N), u(N)]$$

such that a fraction $f \geq p$ of them are *guaranteed* to be true whatever the true mean count ***s***.

Neyman's invention is called a Neyman construction.

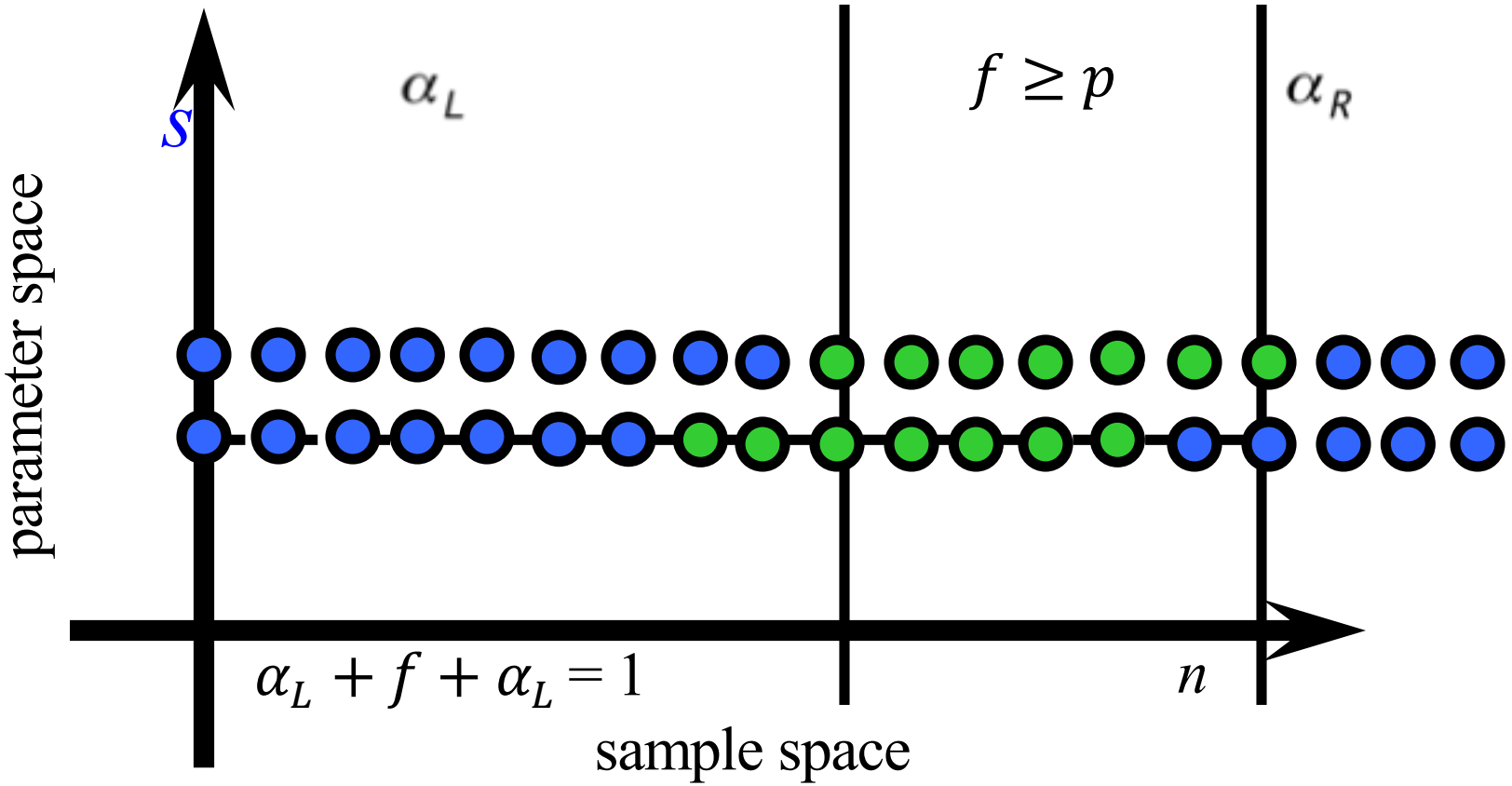
Confidence Intervals – 2

Suppose we knew ***s***. We could then find a region in the *sample space* with probability $f \geq \boldsymbol{p}$ = *confidence level* (C.L.)



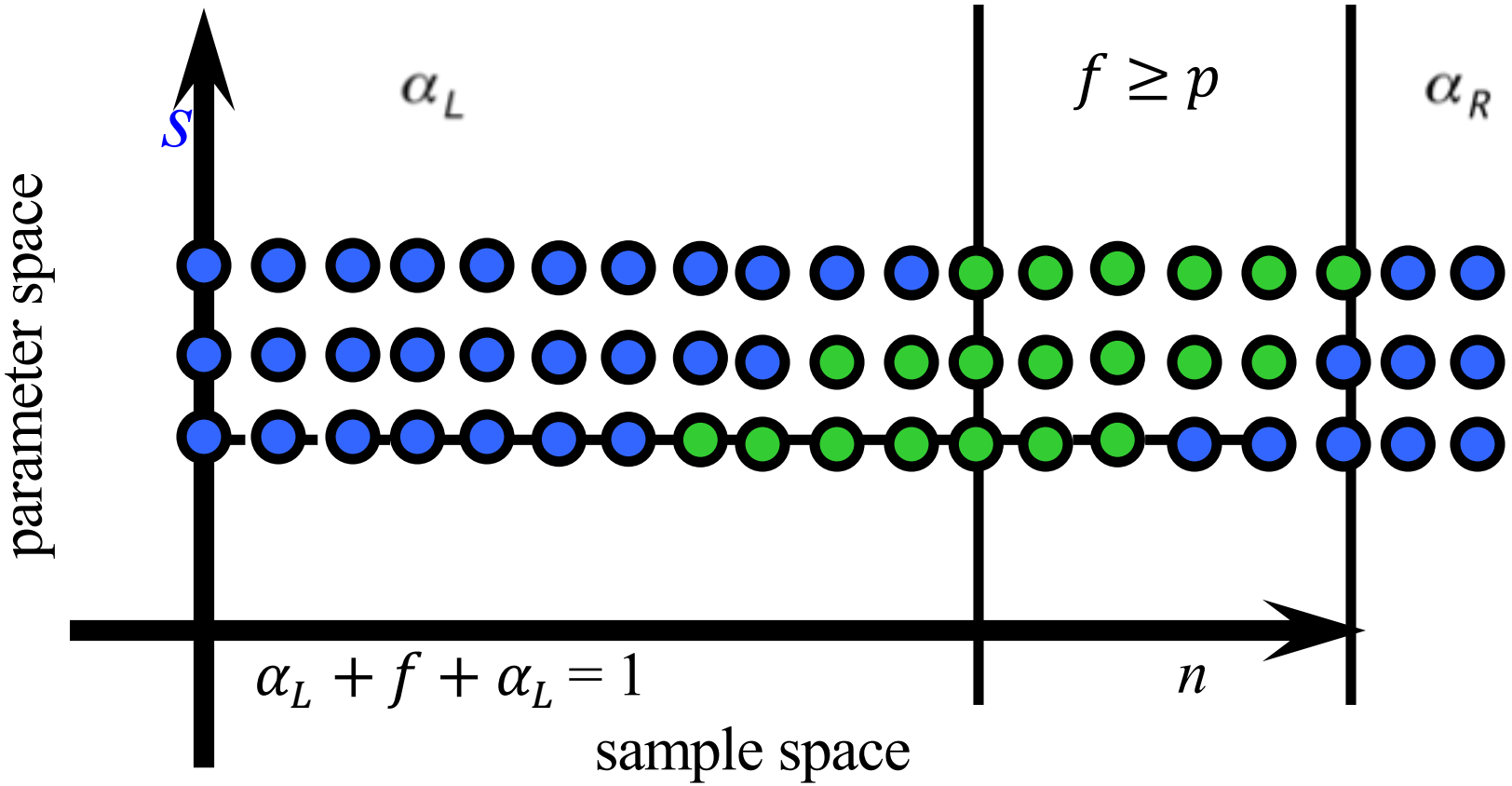
Confidence Intervals – 3

But, in reality, we do not know ***s***! So, we repeat this procedure for every ***s*** that is possible, *a priori*.



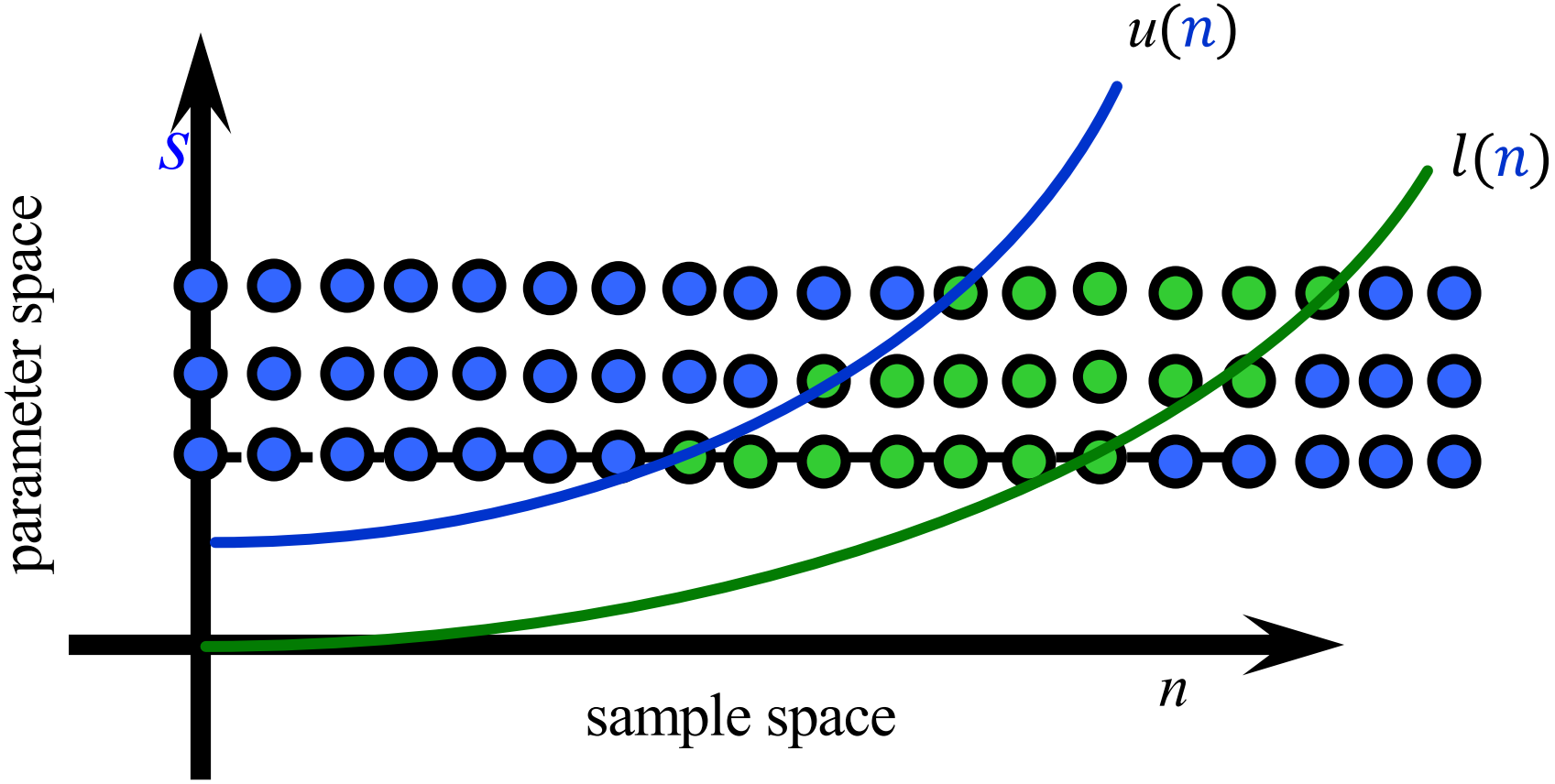
Confidence Intervals – 4

But, in reality. we do not know ***s***! So, we repeat this procedure for every ***s*** that is possible, *a priori*.



Confidence Intervals – 5

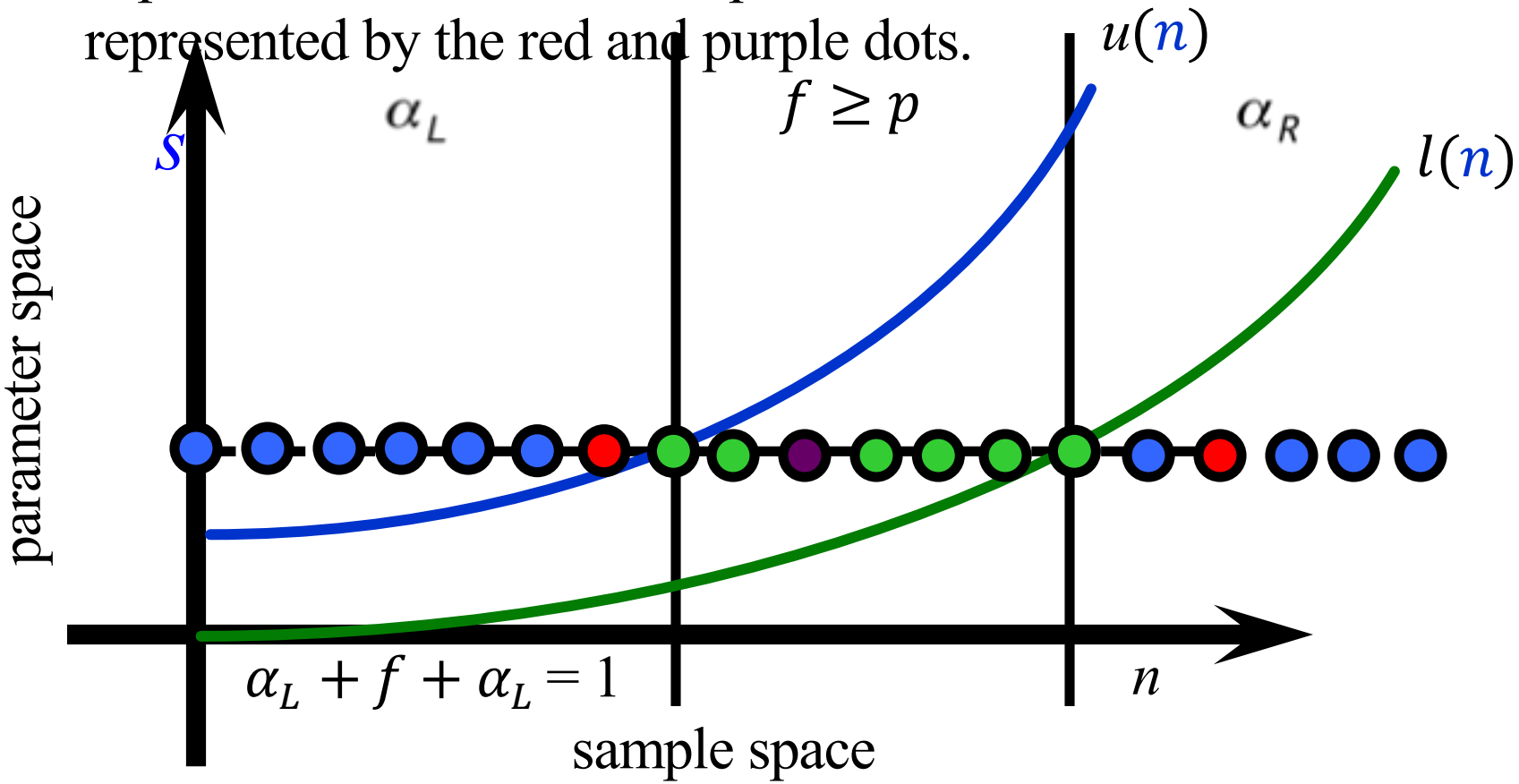
Through this step-by-step procedure, we build two curves $l(n)$ and $u(n)$* which define the confidence intervals $[l(n), u(n)]$.



* Since we've assumed the sample space is discrete, these "curves" are discontinuous.

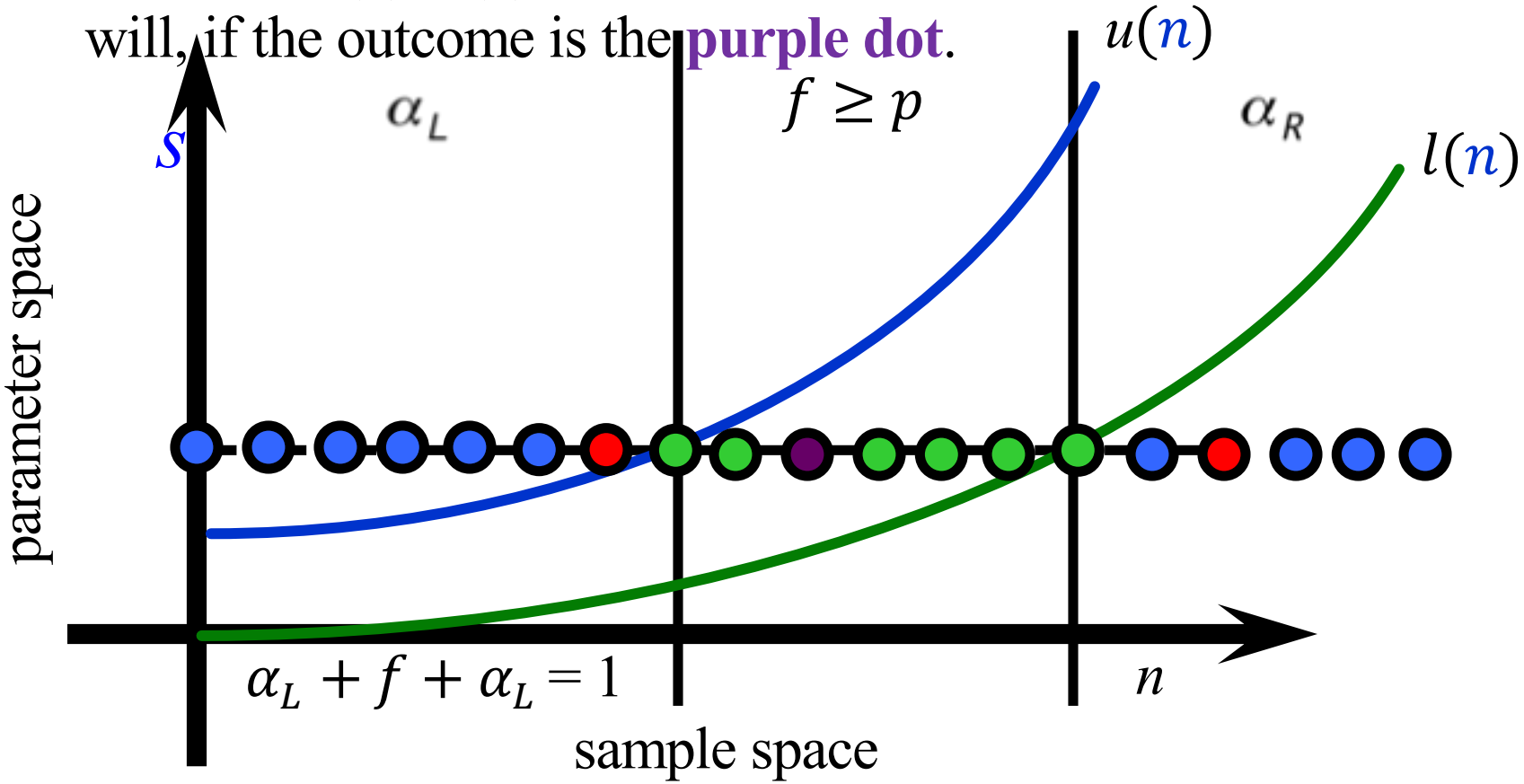
Confidence Intervals – 6.1

Suppose, the ***s*** shown is the true (unknown) value for a given experiment. There are three possible classes of outcome represented by the red and purple dots.



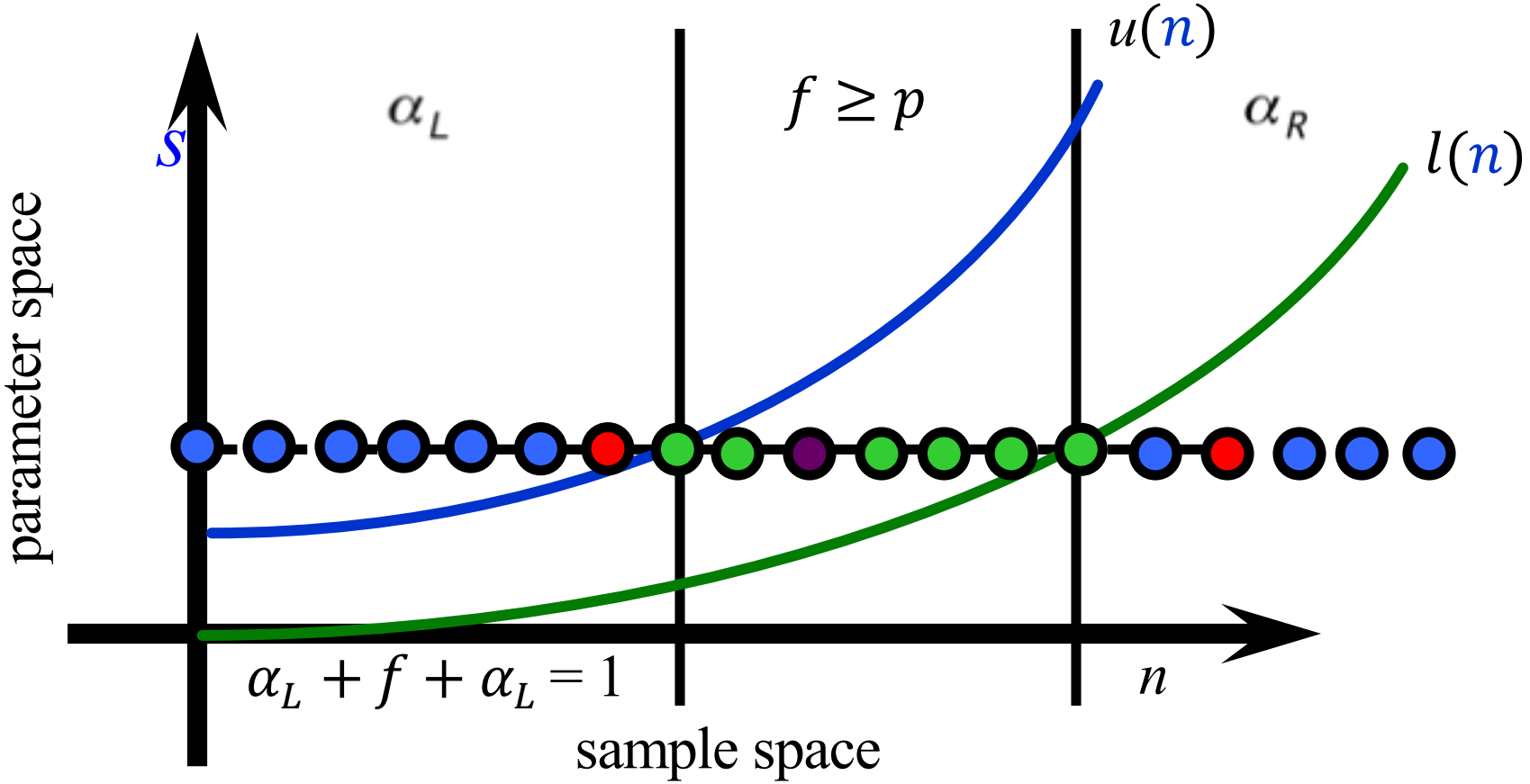
Confidence Intervals – 6.2

If the experiment yields a count corresponding to either **red dot**, then $[l(n), u(n)]$ will *not* contain the true value of s. But, it will, if the outcome is the **purple dot**.



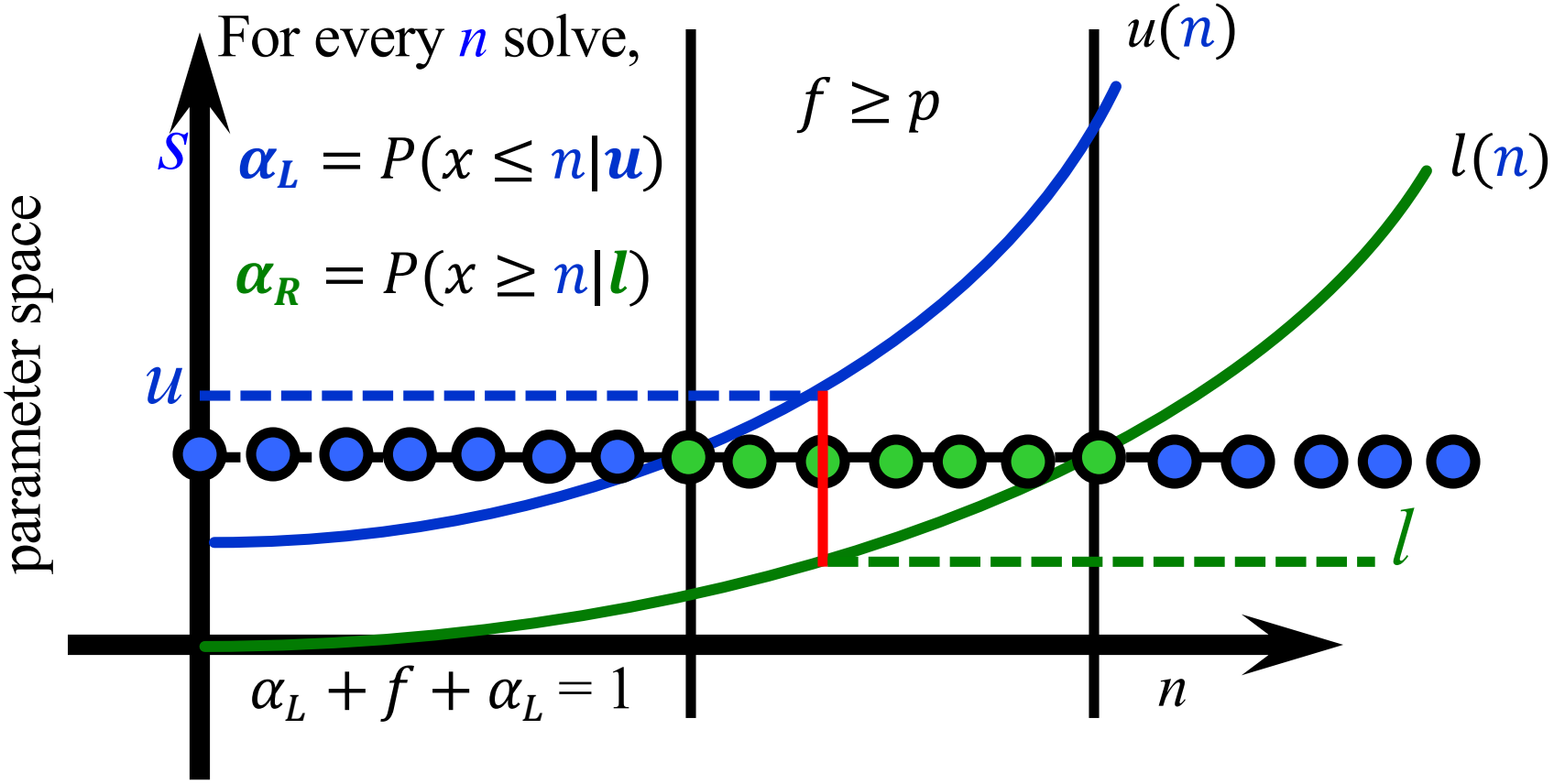
Confidence Intervals – 6.3

The probability to get an interval [$l(n)$, $u(n)$] that brackets ***s*** is, by construction, $f \geq p$, where ***p*** is the desired confidence level.



Confidence Intervals – 7

There are many ways to create a region in the sample space with probability f. Here is Neyman's (1937) method:



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

Confidence Intervals – 8

Here are two other ways to construct *sample space* intervals that yield *parameter space* intervals that satisfy the FP.

1. **Feldman & Cousins** (1997)

   Find intervals with the largest values of the ratio $\lambda(s) = P(n|\, s) \,/\, P(n|\, s^*)$, where s^* is an estimate of s.

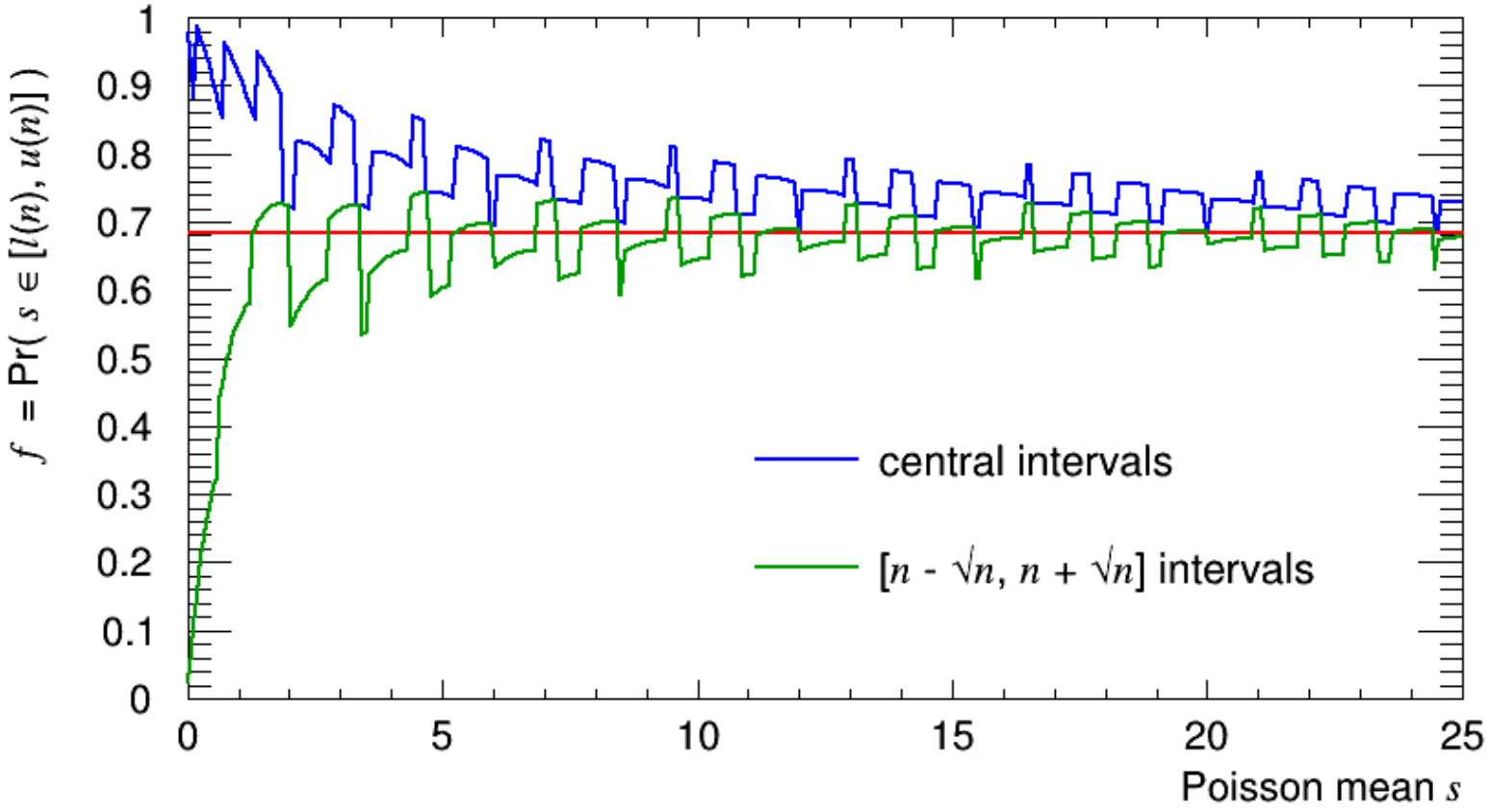
2. **Mode Centered** (HBP, late 20th century!)

   Find intervals with the largest value of $P(n \mid s)$.

3. $[n - \sqrt{n},\, n + \sqrt{n}]$. However, these do *not* satisfy the frequentist principle (FP).

Exercise 7.2: Compute Poisson central intervals and their coverage

Confidence Intervals – 9



Coverage probability ***f*** as a function of the Poisson mean ***s***
The **red line** is $p = 0.683$ the desired **confidence level**.

Outline

- Lecture 1
  - Introduction
  - The Frequentist Principle
  - Confidence Intervals
  - The Profile Likelihood
- Lecture 2
  - Hypothesis Tests
  - Bayesian Inference
- Lecture 3
  - Introduction to Machine Learning

Nuisance Parameters

Consider again the likelihood

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

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

for the data

$D = (N, M, k) = (25,\ 353.4,\ 37.6)$ of our example.

This function contains two parameters ***s*** and ***b***. Suppose we are interested in the mean signal ***s***, but not the mean background ***b***. In this case, ***b*** is an example of a nuisance parameter.

Nuisance Parameters are a Nuisance!

One way or another, we must rid our probability models of *all* nuisance parameters if we wish to make interferences about the *parameters of interest*, such as the mean signal ***s***.

We'll show how this works using our example.

Example: Higgs to 4-Leptons

Recall the knowns and unknowns:

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

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

unknowns:

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

Example: Higgs to 4-Leptons

Given the likelihood, $L(s,b) \equiv p(D\,|s,b)$, we can estimate its parameters by maximizing the likelihood:

$$\frac{\partial \ln L(s,b)}{\partial s} = \frac{\partial \ln L(s,b)}{\partial s} = 0$$

This yields the obvious results

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

with $N = 25$ observed events

$B = 9.4 \pm 0.5$ background events

Estimates found this way (first done by Karl Frederick Gauss) are called maximum likelihood estimates (MLE).

Maximum Likelihood – An Aside

The **Good**

- Maximum likelihood estimates are *consistent*: the RMS goes to zero as more data are acquired.
- If an *unbiased* estimate for a parameter exists, the maximum likelihood procedure will find it.
- Given the MLE for $\boldsymbol{s}$, the MLE for $y = g(\boldsymbol{s})$ is just

$$\hat{y} = g(\hat{\boldsymbol{s}})$$

The **Bad** (according to some!)

- In general, MLEs are *biased*

The **Ugly** (according to some!)

- Correcting for bias can sometimes waste data.

Exercise 8: Prove this.
Hint: write
$g(\hat{s}) = g(s + \hat{s} - s)$,
Taylor expand about $\boldsymbol{s}$ and consider the ensemble average of $\hat{y} = g(\hat{\boldsymbol{s}})$.

Example: Higgs to 4-Leptons

In order to make an inference about the signal, ***s***, the 2-parameter problem,

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

must be reduced to one involving ***s*** *only* by getting rid of the nuisance parameter ***b***.

In principle, this must be done while respecting the frequentist principle:

coverage probability $\geq$ *confidence level*.

In general, *this is difficult to do exactly*.

Example: Higgs to 4-Leptons

In practice, one replaces *all* nuisance parameters by their conditional maximum likelihood estimates (CMLE), which yields a function called the **profile likelihood**, $L_P(s)$.

For the Higgs boson example, we find an estimate of ***b*** as a function of ***s***

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

Then, in the likelihood $L(s, b)$, b is replaced with its *estimate*.

Since this is an approximation, the frequentist principle is not guaranteed to be satisfied exactly.

But this procedure has a sound justification…

Example: Higgs to 4-Leptons

Consider the **profile likelihood ratio**

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

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

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

about $\hat{s}$:

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

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

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

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

Example: Higgs to 4-Leptons

If $\hat{s}$ does not occur on the boundary of the parameter space, and the data sample is large enough (typically, when the density of $\hat{s}$ is approximately Gaussian($\hat{s}$, **s**, σ)), and ***s*** is the true value of the 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 9: Verify this theorem through simulation

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

Example: Higgs to 4-Leptons

The CMLE of b is

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

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

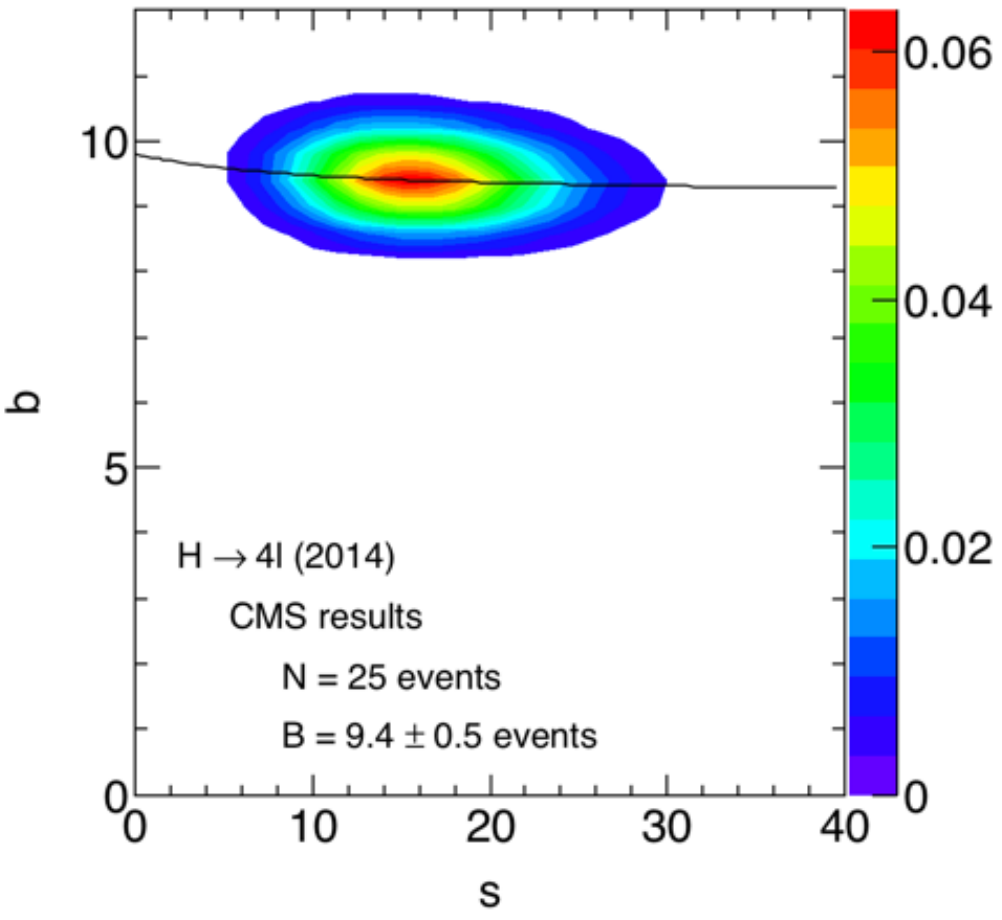
Note,

$$s = N - B = 15.6 \text{ events}$$

$$b = B$$

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

Exercise 10: Show this



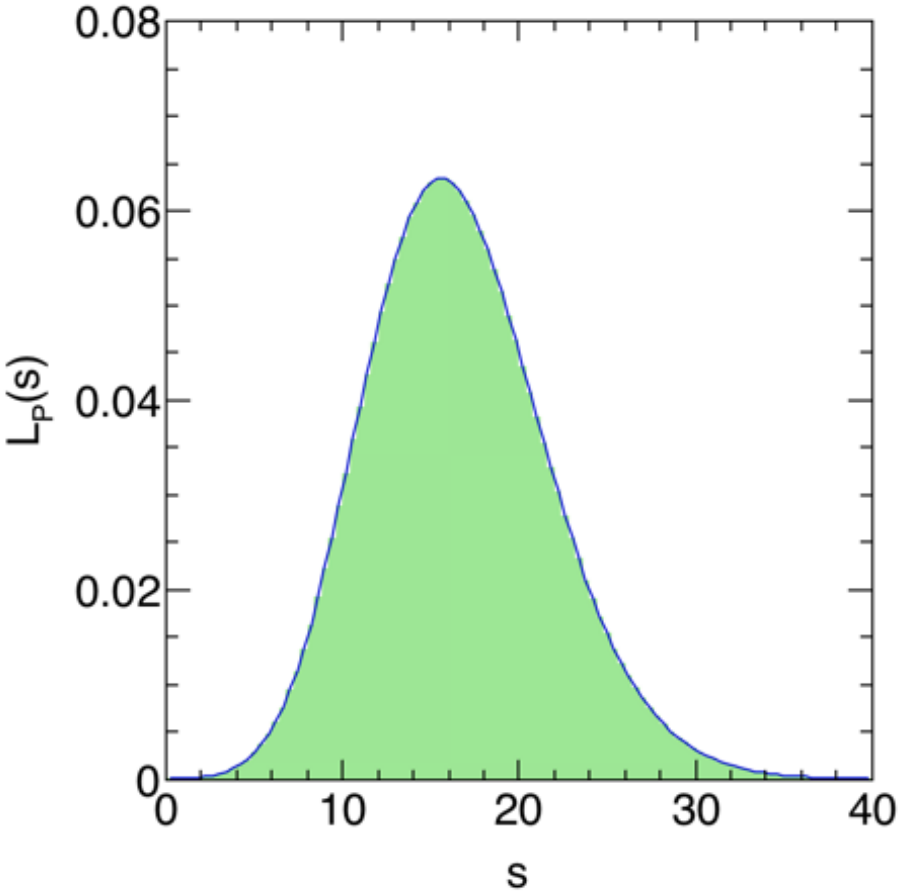
Example: Higgs to 4-Leptons

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

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

for s, which yields the statement

$s \in [10.9, 21]$ @ ~ 68% C.L



Exercise 11: Show this by solving $t(s) = 1$ numerically

Summary

Frequentist Approach

1) Uses the relative frequency interpretation of probability.

2) Ideally, respects the frequentist principle.

3) Uses the likelihood.

4) In general, nuisance parameters are removed through the *approximate* procedure of profiling.

Introduction: Samples – 2

If one orders data $\boldsymbol{x} = x_1, x_2, \ldots x_n$ so that

$x_{(1)} < x_{(2)} < \ldots < x_{(n)}$

$x_{(k)}$ is called the k^{th} order statistic

$x_{(k)}$ is also the α-quantile if $\alpha = k / n$

When $\alpha = 0.5$, $x_{(k)}$ is called the median.

Common Probability Functions

Uniform(x, a)	$1/a$
Normal(x, μ, σ)	$\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)/(\sigma\sqrt{2\pi})$
LogNormal(x, μ, σ)	$\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)/(x\,\sigma\sqrt{2\pi})$
Chisq(x, n)	$x^{\frac{n}{2}-1}\exp(-x/2)/\left[2^{\frac{n}{2}}\Gamma(n/2)\right]$
Gamma(x, a, b)	$a(ax)^{b-1}\exp(-ax)/\Gamma(b)$
Exp(x, a)	Gamma$(x, a, 1)$
Binomial(k, n, p)	$\binom{n}{k} p^k (1-p)^{n-k}$
Poisson(k, a)	$a^k \exp(-a)\,/k!$
Multinomi(k, n, p)	$n! \prod_{i=1}^{n} \frac{p_i^{n_i}}{k_i!}$, $\sum_{i=1}^{n} p_i = 1$, $\sum_{i=1}^{n} n_i = n$