Practical Statistics for LHC Physicists

Frequentist Inference

Harrison B. Prosper

Florida State University

CERN Academic Training Lectures

8 April, 2015

Outline

- The Frequentist Principle
- Confidence Intervals
- The Profile Likelihood
- Hypothesis Tests

The Frequentist Principle

The Frequentist Principle

The Frequentist Principle (FP) (Neyman, 1937)

Construct statements such that a fraction $f \geq p$ of them are 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 FP is said to cover.

The Frequentist Principle

Points to Note:

1. The frequentist principle applies to *real* ensembles, not just the ones we simulate on a computer. Moreover, the statements need not all be about the same quantity.

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

2. Coverage ***f*** 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, in the real world, we are typically not privy to this knowledge.

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.

The Frequentist Principle

Example continued

Suppose each mean count θ is randomly sampled from Uniform(θ, 5), and suppose we *know* these numbers.

Now, of course, we can compute the coverage probability ***f*** , i.e., the fraction of true statements.

Exercise 7:
Show, that the coverage f is 0.62

Confidence Intervals

Confidence Intervals – 1

Consider an experiment that observes N events with expected count s.

Neyman (1937) devised a way to make statements of the form

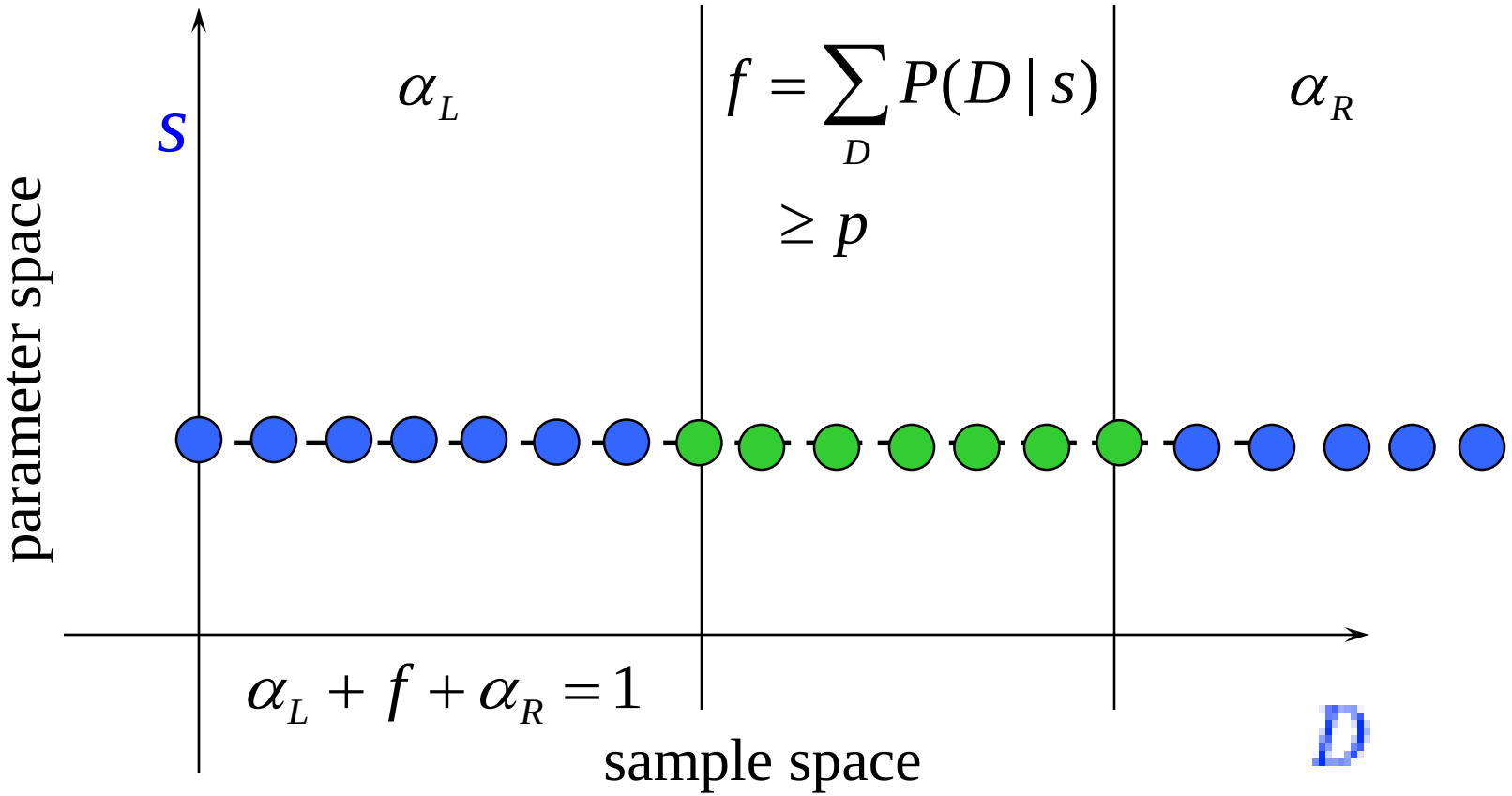
$$s \in [l(N), u(N)]$$

such that a fraction $f \geq p$ of them are true. Note, again, that the expected count s may not be, and extremely unlikely to be, exactly the same for every experiment.

Neyman's brilliant invention is called a Neyman construction.

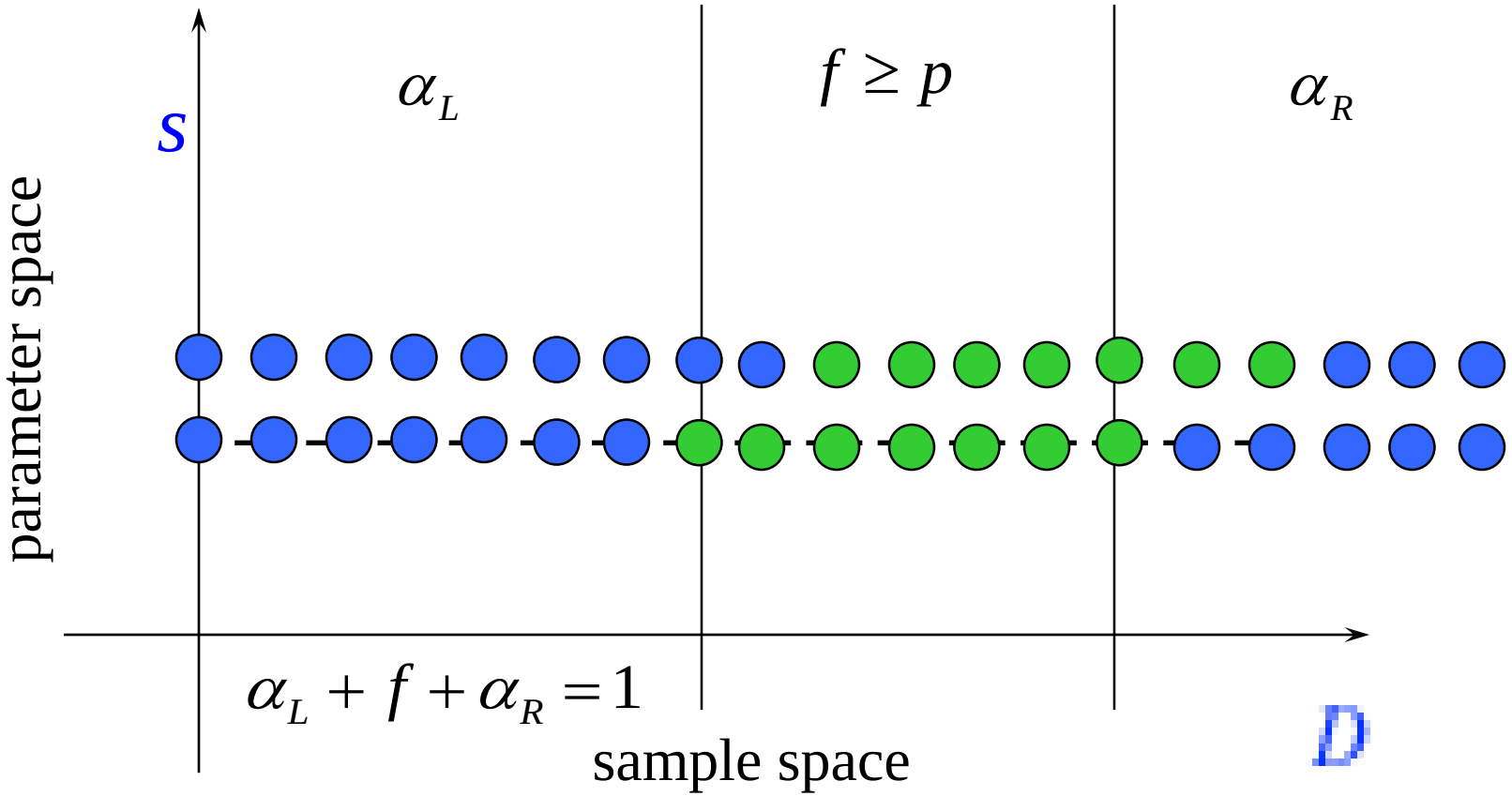
Confidence Intervals – 2

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



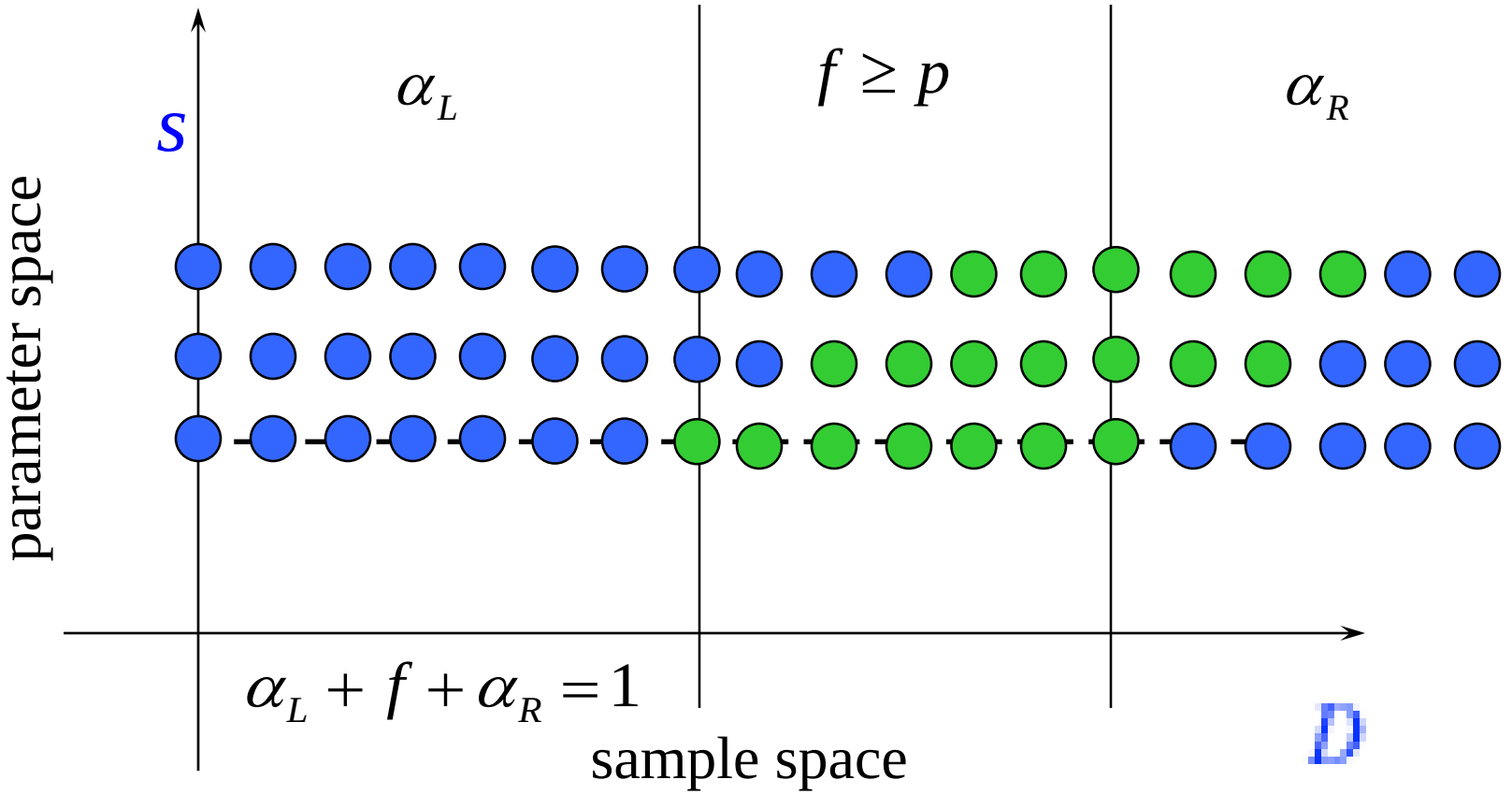
Confidence Intervals – 3

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



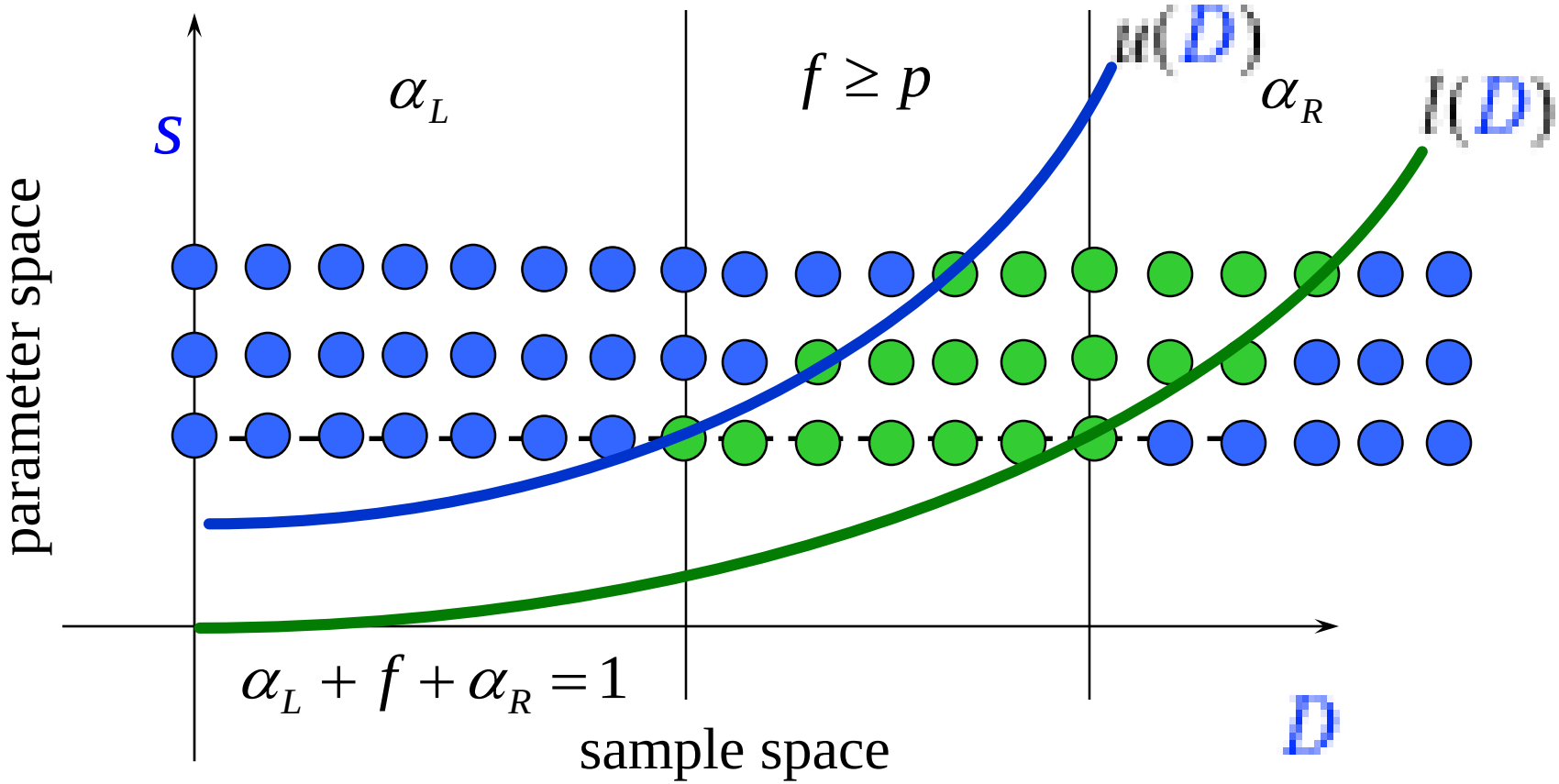
Confidence Intervals – 4

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



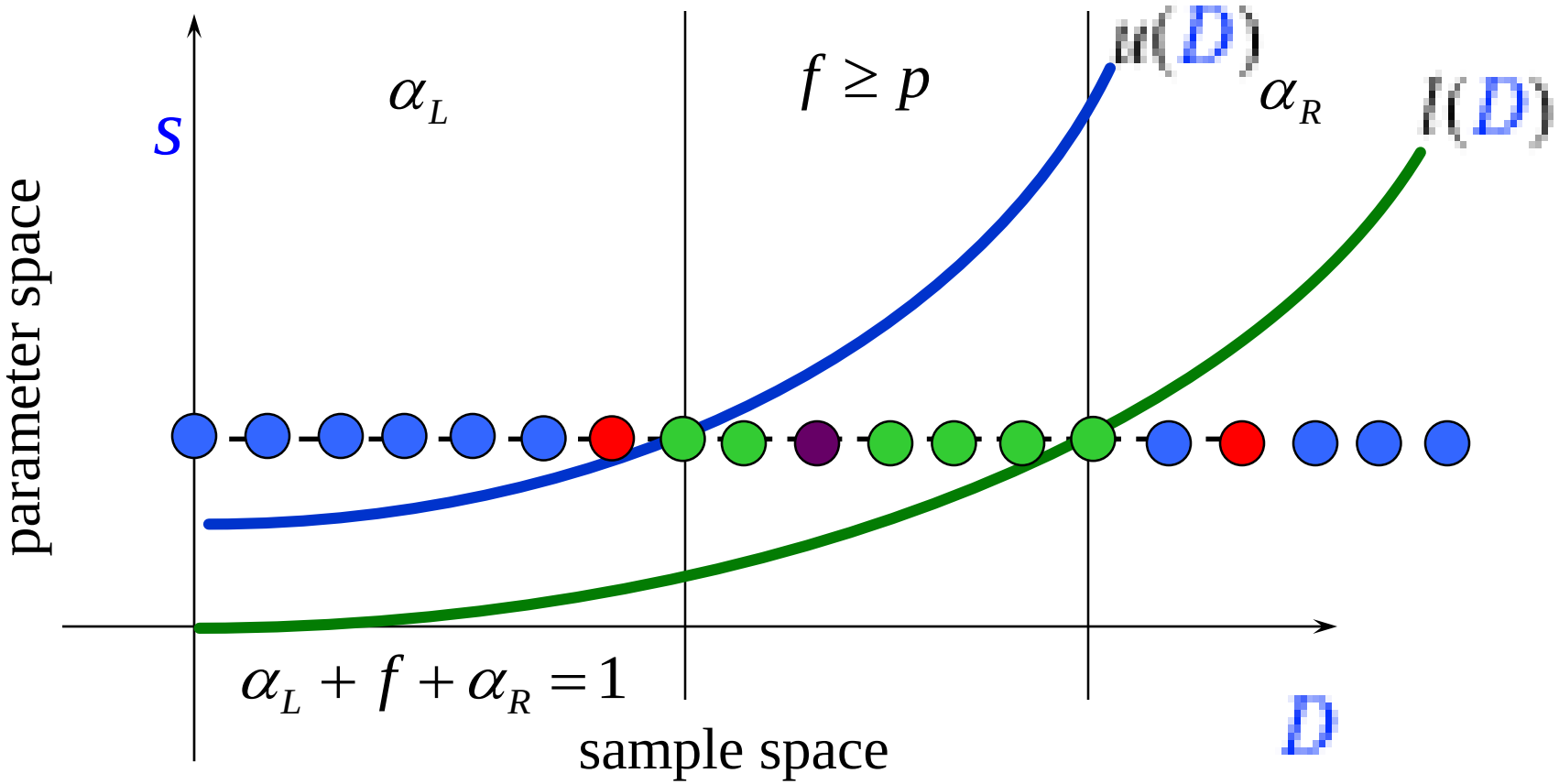
Confidence Intervals – 5

Through this procedure we build two curves $l(D)$ and $u(D)$ that define lower and upper limits, respectively.



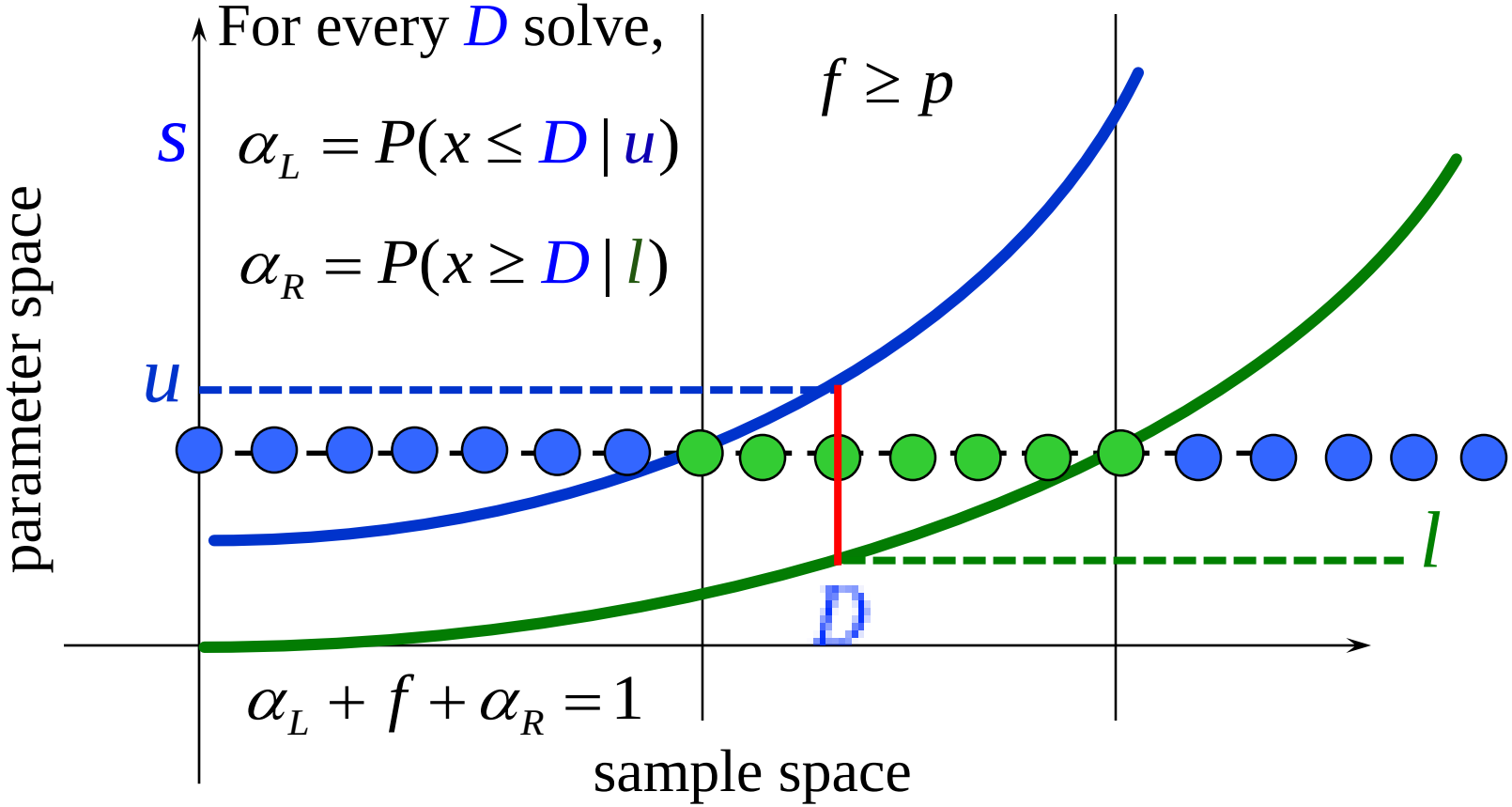
Confidence Intervals – 6

Suppose, the s shown is the true value for one experiment. The probability to get an interval $[l(D), u(D)]$ that includes s is $\geq p$.



Confidence Intervals – 7

There are many ways to create a region, in the sample space, with probability f. Here is *the* classic way (Neyman, 1937):



Confidence Intervals – 8

Here are a few ways to construct *sample space* intervals

1. Central Intervals (Neyman, 1937)

   Solve $\alpha_R = P(x \leq D|\, u)$ and $\alpha_R = P(x \geq D|\, l)$
   with $\alpha_R = \alpha_L = (1 - CL)/2$

2. Feldman & Cousins (1997)

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

1. Mode Centered (HBP, some time in the late 20th century)

   Find intervals with the largest value of $P(D|\, s)$.

By construction, all these yield intervals satisfy the FP.

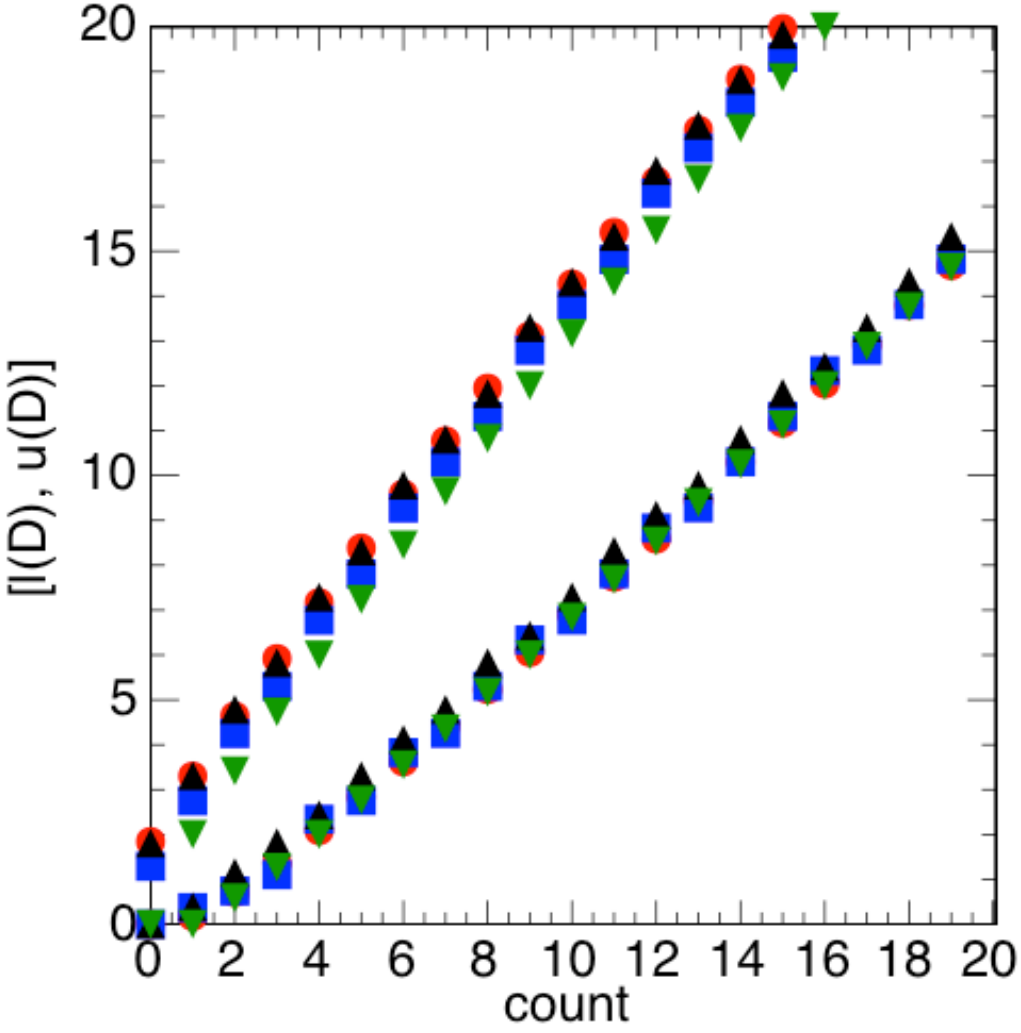
Confidence Intervals – 9

Central

Feldman & Cousins

Mode Centered

$[D - \sqrt{D}, D + \sqrt{D}]$



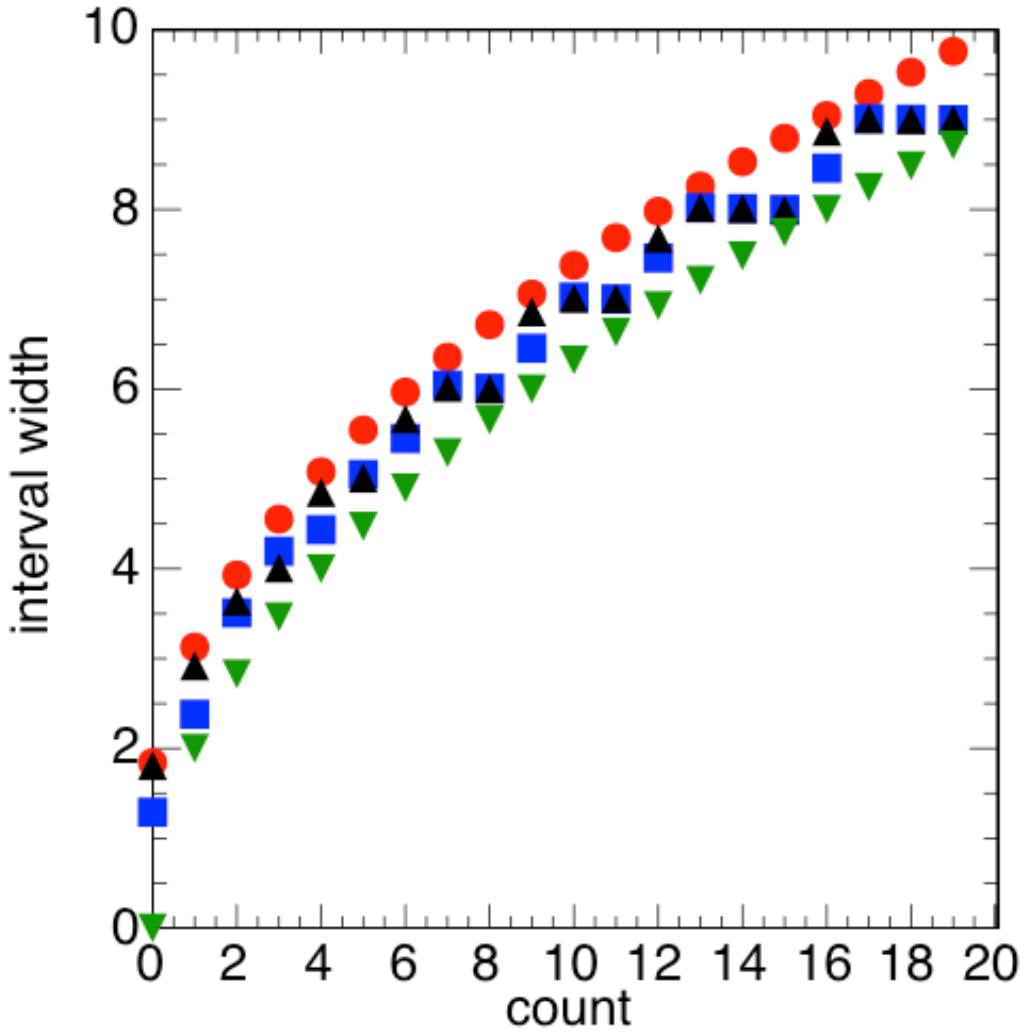
Confidence Intervals – 10

Central

Feldman & Cousins

Mode Centered

$[D - \sqrt{D}, D + \sqrt{D}]$



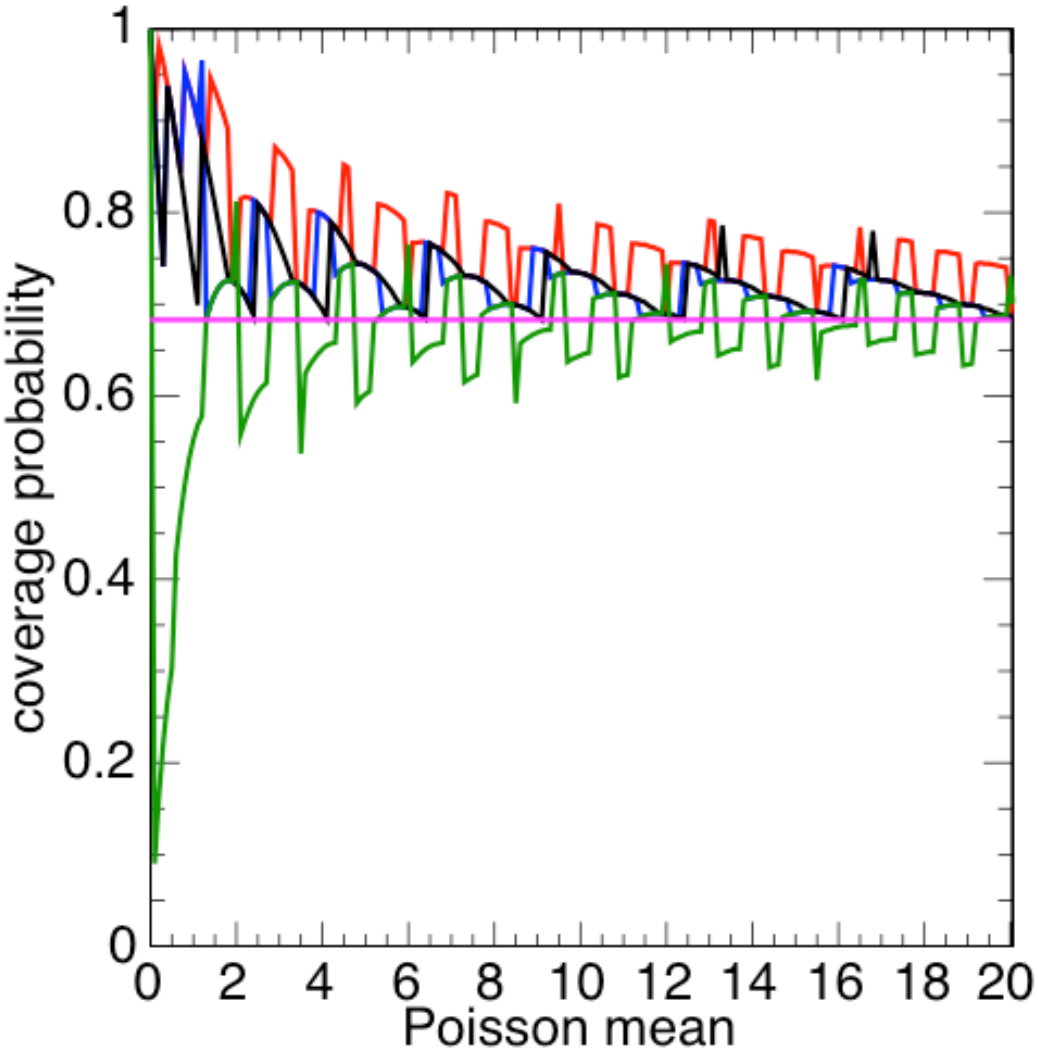
Confidence Intervals – 11

Central

Feldman & Cousins

Mode Centered

$[D - \sqrt{D}, D + \sqrt{D}]$



The Profile Likelihood

Nuisance Parameters are a Nuisance!

All models are "wrong"! But,…

…to the degree that the probability models are accurate models of the data generation mechanisms, the Neyman construction, *by construction*, satisfies the FP *exactly*.

However, to achieve this happy state, we must construct confidence regions for all the parameters *simultaneously*.

But, what if we are *not* interested in the expected background, b, but only in the expected signal s? The expected background count is an example of a nuisance parameter…for once, here's jargon that says it all!

Nuisance Parameters are a Nuisance!

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

We'll show how this works in practice, using

Example 1:

Evidence for electroweak production of $W^{\pm}W^{\pm}jj$ (ATLAS, 2014)

PRL 113, 141803 (2014)

Example 1: $W^{\pm}W^{\pm}jj$ Production (ATLAS)

First, let's be clear about knowns and (*known*) unknowns:

knowns:

$D = 12$	observed events	($\mu^{\pm}\mu^{\pm}$ mode)
$B = 3.0 \pm 0.6$	background events	

unknowns:

b	expected background count
s	expected signal count

Next, we construct a probability model.

Example 1: $W^{\pm}W^{\pm}jj$ Production (ATLAS)

Probability:

$$P(D \mid s, b) = \text{Poisson}(D, s+b)\ \text{Poisson}(Q, bq)$$

$$= \frac{(s+b)^D e^{-(s+b)}}{D!} \frac{(bk)^Q e^{-bq}}{\Gamma(Q+1)}$$

Likelihood:

$$L(s, b) \equiv P(12 \mid s, b)$$

We model the background estimate as an *effective* count:

$$B = Q / q$$

$$\delta B = \sqrt{Q} / q$$

$$Q = (B / \delta B)^2 = (3.0 / 0.6)^2 = 25.0$$

$$q = B / \delta B^2 = 3.0 / 0.6^2 = 8.33$$

Example 1: $W^{\pm}W^{\pm}jj$ Production (ATLAS)

Now that we have a likelihood, we can estimate its parameters, for example, by maximizing the likelihood:

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

$$\hat{s} = D - B, \quad \hat{b} = B$$

with $D = 12$ observed events ($\mu^{\pm}\mu^{\pm}$ mode)

$B = 3.0 \pm 0.6$ 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 and more data are acquired.
- If an *unbiased* estimate for a parameter exists, the maximum likelihood procedure will find it.
- Given the MLE for s, the MLE for $y = g(s)$ is just $\hat{y} = g(\hat{s})$

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

- In general, MLEs are biased

Exercise 8: Show this
Hint: perform a Taylor expansion about the MLE and consider its ensemble average.

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

- Correcting for bias, however, can waste data and sometimes yield absurdities. (See Seriously Ugly)

The Profile Likelihood – 1

In order to make an inference about the $W^{\pm}W^{\pm}jj$ signal, s, the
2-parameter problem,

$$p(D \mid s, b) = \frac{(s+b)^D e^{-(s+b)}}{D!} \frac{(bq)^Q e^{-bq}}{\Gamma(Q+1)}$$

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 prob.* $\geq$ *confidence level*.

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

The Profile Likelihood – 2

In practice, what we do is replace *all* nuisance parameters by their conditional maximum likelihood estimates (CMLE), which yields a function called the profile likelihood, $L_P(s)$.

For the $W^{\pm}W^{\pm}jj$ evidence 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, as we shall now see…

The Profile Likelihood – 3

Consider the profile likelihood ratio

$$\lambda(s) = L_P(s) / L_p(\hat{s})$$

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

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

about $\hat{s}$:

$$t(s) = t(\hat{s}) + t'(\hat{s})(s - \hat{s}) + t''(\hat{s})(s - \hat{s})^2 / 2 + \ldots$$

$$= (s - \hat{s})^2 / [2 / t''(\hat{s})] + \mathrm{d}\,(1/\sqrt{N})$$

The result is called the Wald approximation (1943).

The Profile Likelihood – 4

If $\hat{s}$ does not occur on the boundary of the parameter space, and if the data sample is large enough so that the density of $\hat{s}$ is approximately,

$$\text{Gaussian}(\hat{s}, s, \sigma)$$

then

$$t(s) \approx (s - \hat{s})^2 / \sigma^2$$

has a χ^2 density of one degree of freedom, where $\sigma^2 = 2 / t''(\hat{s})$

This result, Wilks' Theorem (1938) and its generalization, is the basis of formulae popular in ATLAS and CMS.

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

The Profile Likelihood – 5

The CMLE of b is

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

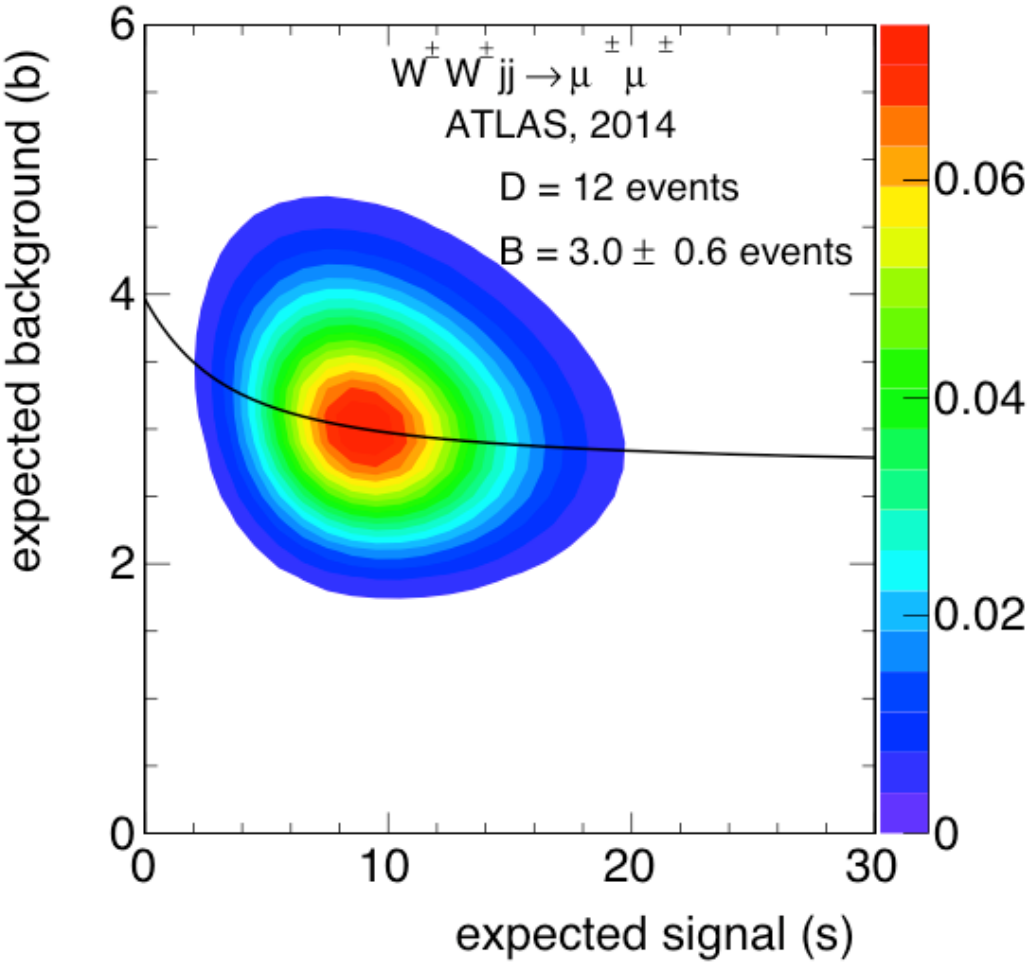
$$g = D + Q - (1+q)s$$

with

$$s = D - B$$

$$b = B$$

the **mode** (peak) of the likelihood



The Profile Likelihood – 6

By solving

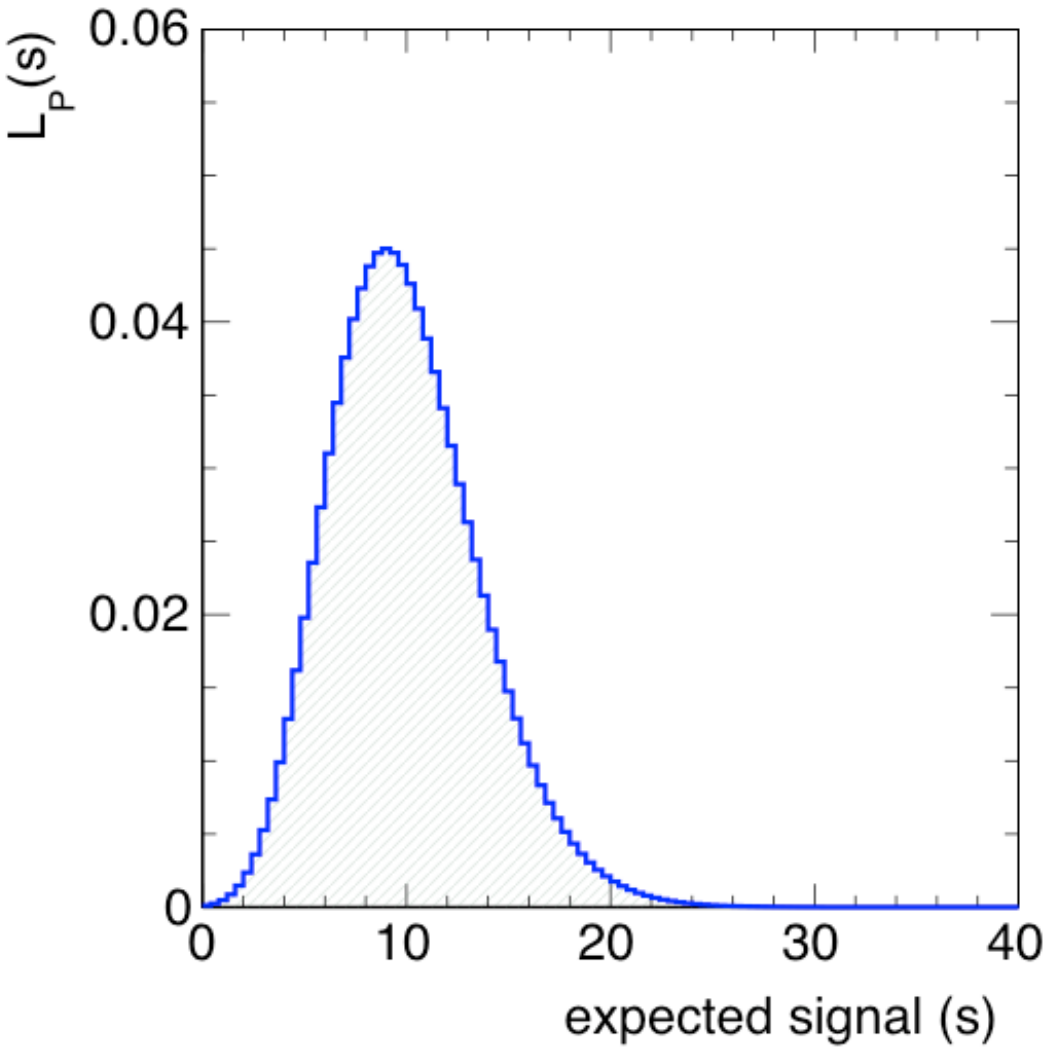
$$t(s) = -2\ln\lambda(s) = 1$$

for s, we can make the statement

$$s \in [5.8, 12.9]$$

@ ~ 68% C.L.

Exercise 9: Show this



The Hypothesis Tests

Hypothesis Tests

The basic idea is simple:

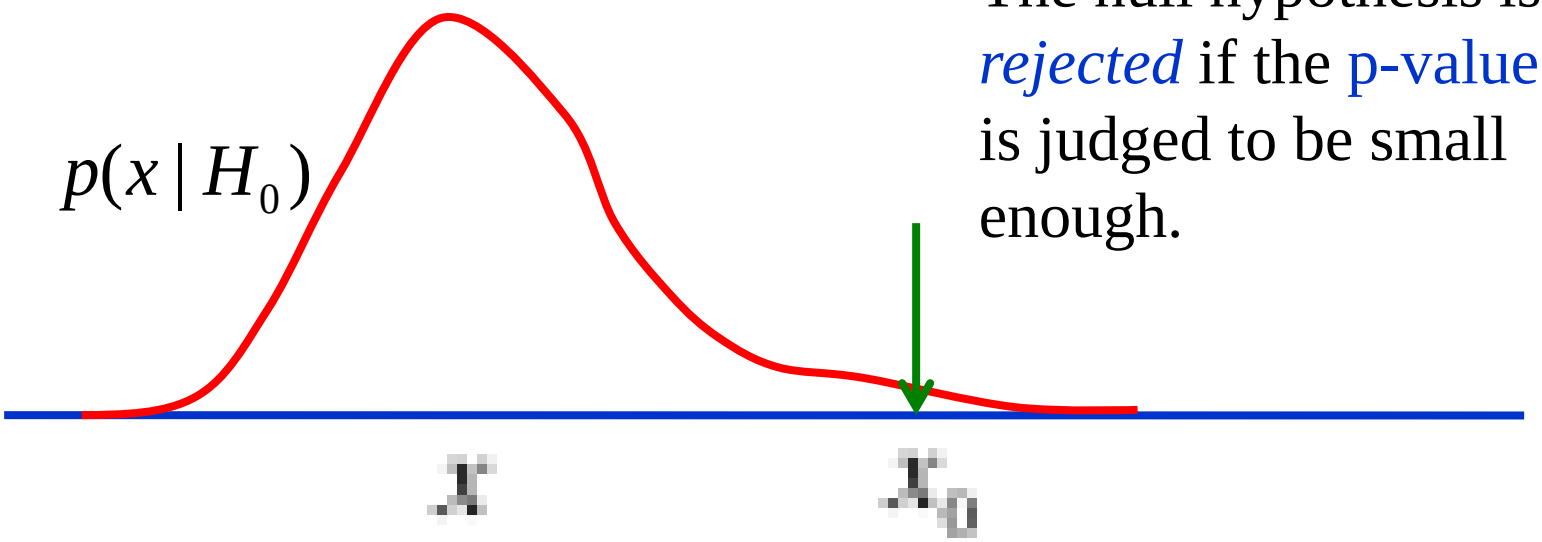
1. Decide which hypothesis you may end up *rejecting*. This is called the null hypothesis. At the LHC, this is typically the background-only hypothesis.

2. Construct a number, called a test statistic that depends on data, such that large values of the test statistic would cast doubt on the veracity of the null hypothesis.

3. Decide on a threshold above which you are prepared to reject the null hypothesis. Do the experiment, compute the statistic, compare it to the agreed upon rejection threshold and reject the null if the threshold is breached.

Hypothesis Tests

Fisher's Approach: *Null* hypothesis (H_0), say, background-only



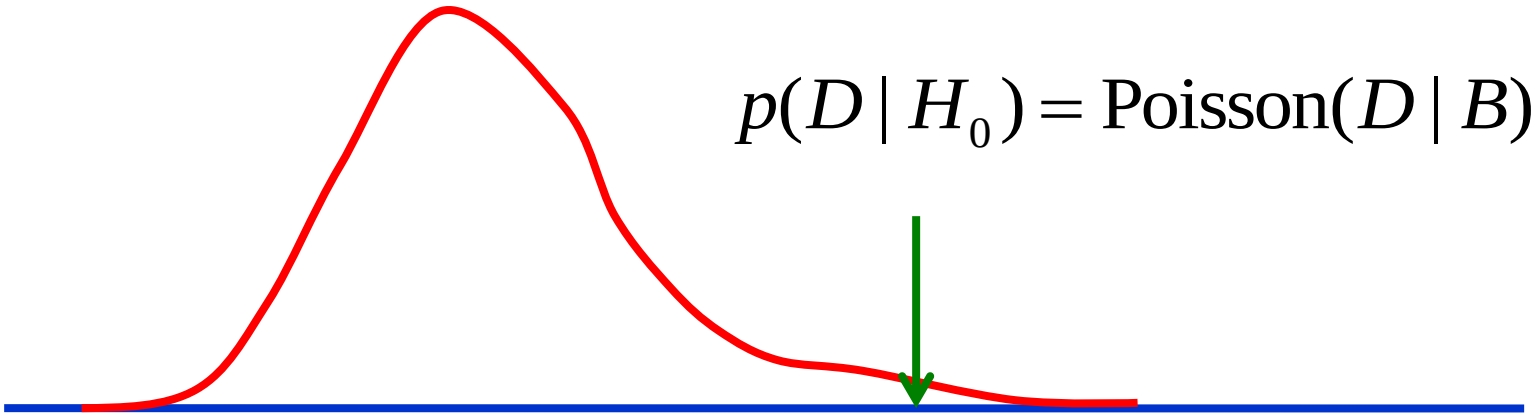
The null hypothesis is *rejected* if the p-value is judged to be small enough.

$$\text{p-value} = P(x \geq x_0 \mid H_0)$$

x_0 is the observed value of the test statistic.

Example 1: $W^{\pm}W^{\pm}jj$ Production (ATLAS)

Background, B = 3.0 events (ignoring uncertainty)



$$p(D \mid H_0) = \text{Poisson}(D \mid B)$$

$D = 12$ observed count

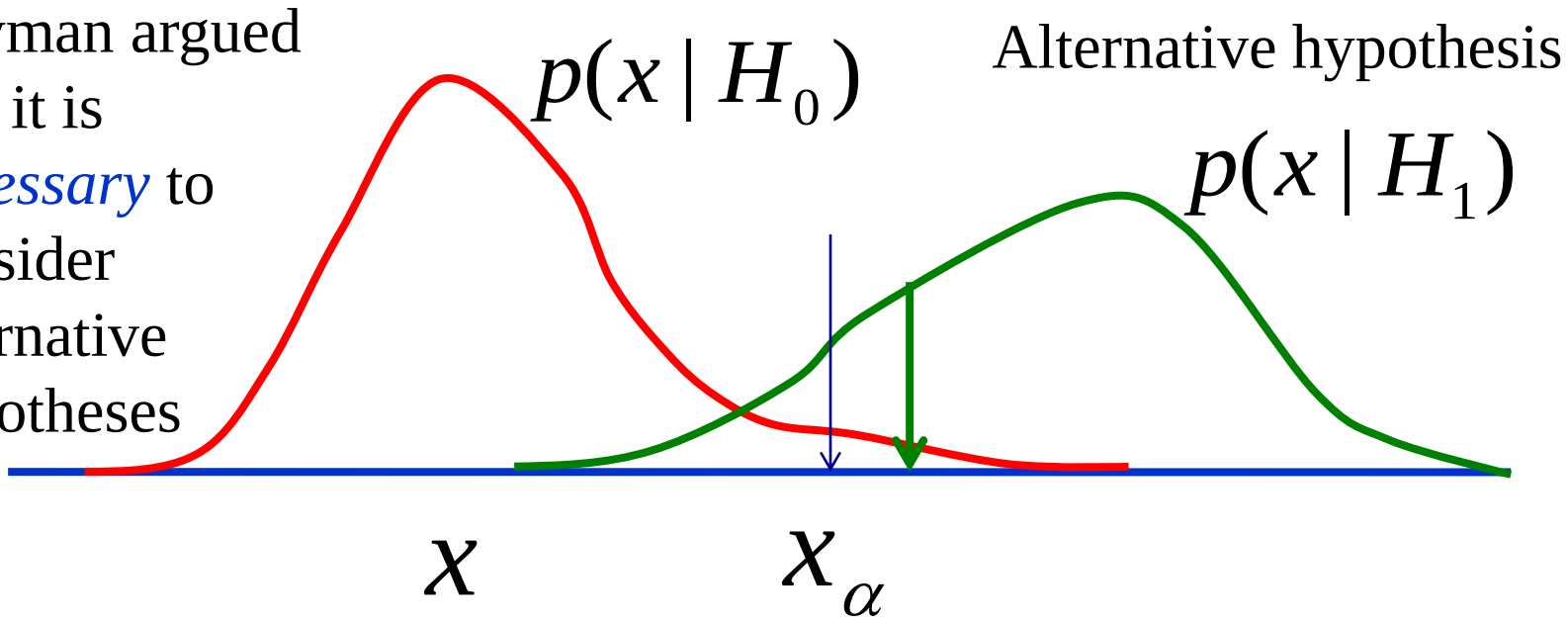
$$\text{p-value} = \sum_{D=12}^{\infty} \text{Poisson}(D \mid 3.0) = 7.1 \times 10^{-5}$$

This is equivalent to a 3.8σ excess above background if the density were a Gaussian.

Hypothesis Tests – 2

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

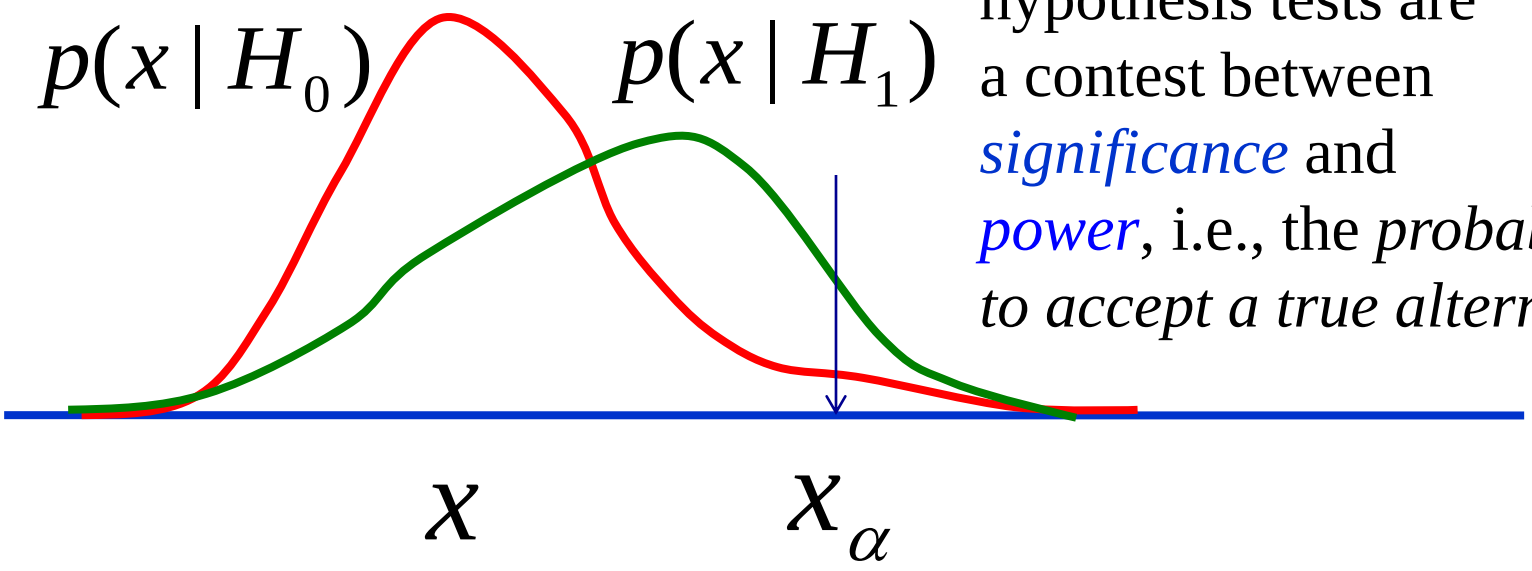
Neyman argued that it is *necessary* to consider alternative hypotheses H_1



$\alpha = \text{p-value}(x_\alpha)$ Choose a *fixed* value of α *before* data are analyzed.

α is called the significance (or size) of the test.

The Neyman-Pearson Test



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

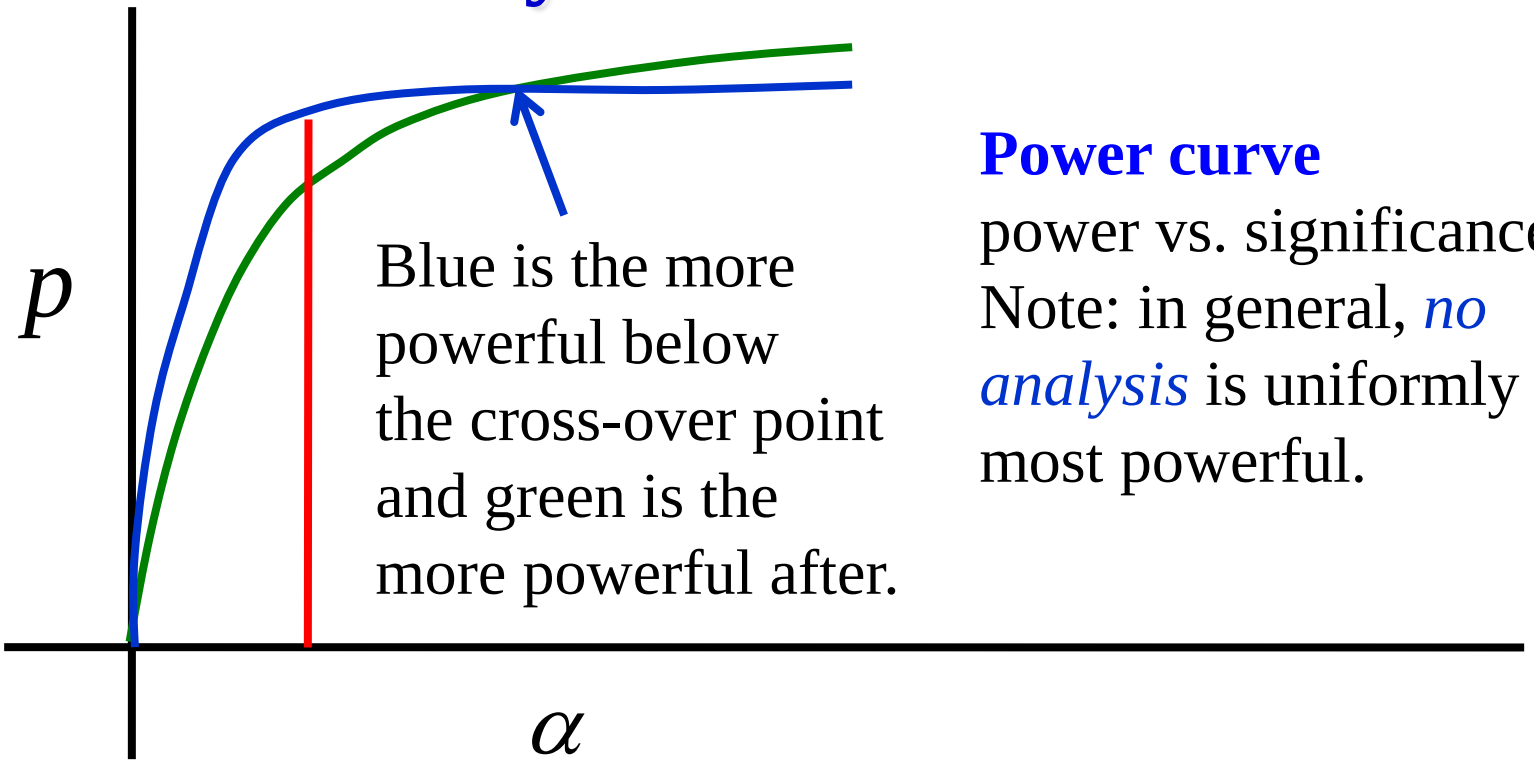
$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0)dx$$

significance of test

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

power

The Neyman-Pearson Test



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

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

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

significance of test

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

power

Hypothesis Tests

This is all well and good, but what do we do when we are bedeviled with nuisance parameters?

…well, we'll talk about that tomorrow and also talk about Bayesian inference.

Summary

Frequentist Inference

1) Uses the likelihood.

2) Ideally, respects the frequentist principle.

3) In practice, nuisance parameters are eliminated through the *approximate* procedure of profiling.

4) A hypothesis test reduces to a comparison between an observed p-value and a p-value threshold called the significance α. Should the p-value < α, the null hypothesis is rejected.

The Seriously Ugly

The **moment generating function** of a probability distribution $P(k)$ is the average:

$$G(x) \equiv \left\langle e^{xk} \right\rangle$$

For the binomial, this is

$$G(x) = \left(e^x p + 1 - p\right)^n$$

Exercise 8a: Show this

which is useful for calculating **moments**

$$M_r = \left.\frac{d^r G}{dx^r}\right|_{x=0} = \sum_{k=0}^{n} k^r \text{Binomial}(k, n, p)$$

e.g.,

$$M_2 = (np)^2 + np - \text{n}p^2$$

The Seriously Ugly

Given that k events out of n pass a set of cuts, the MLE of the event selection efficiency is

$$p = k / n$$

and the obvious estimate of p^2 is

$$k^2 / n^2$$

But

$$\left\langle k^2 / n^2 \right\rangle = p^2 + V / n$$

Exercise 8b: Show this

is a ***biased*** estimate of p^2. The best unbiased estimate of p^2 is

$$k\,(k-1) / [n\,(n-1)\,]$$

Exercise 8c: Show this

Note: for a single success in n trials, $p = 1/n$, but $\boldsymbol{p^2 = 0}$!