Practical Statistics for Particle Physicists
Lecture 2

Harrison B. Prosper

Florida State University

European School of High-Energy Physics
Parádfürdő, Hungary

5 – 18 June, 2013

Likelihood – Higgs to γγ (CMS)

Example 4: Higgs to γγ

background model

$$x = m_{\gamma\gamma}$$

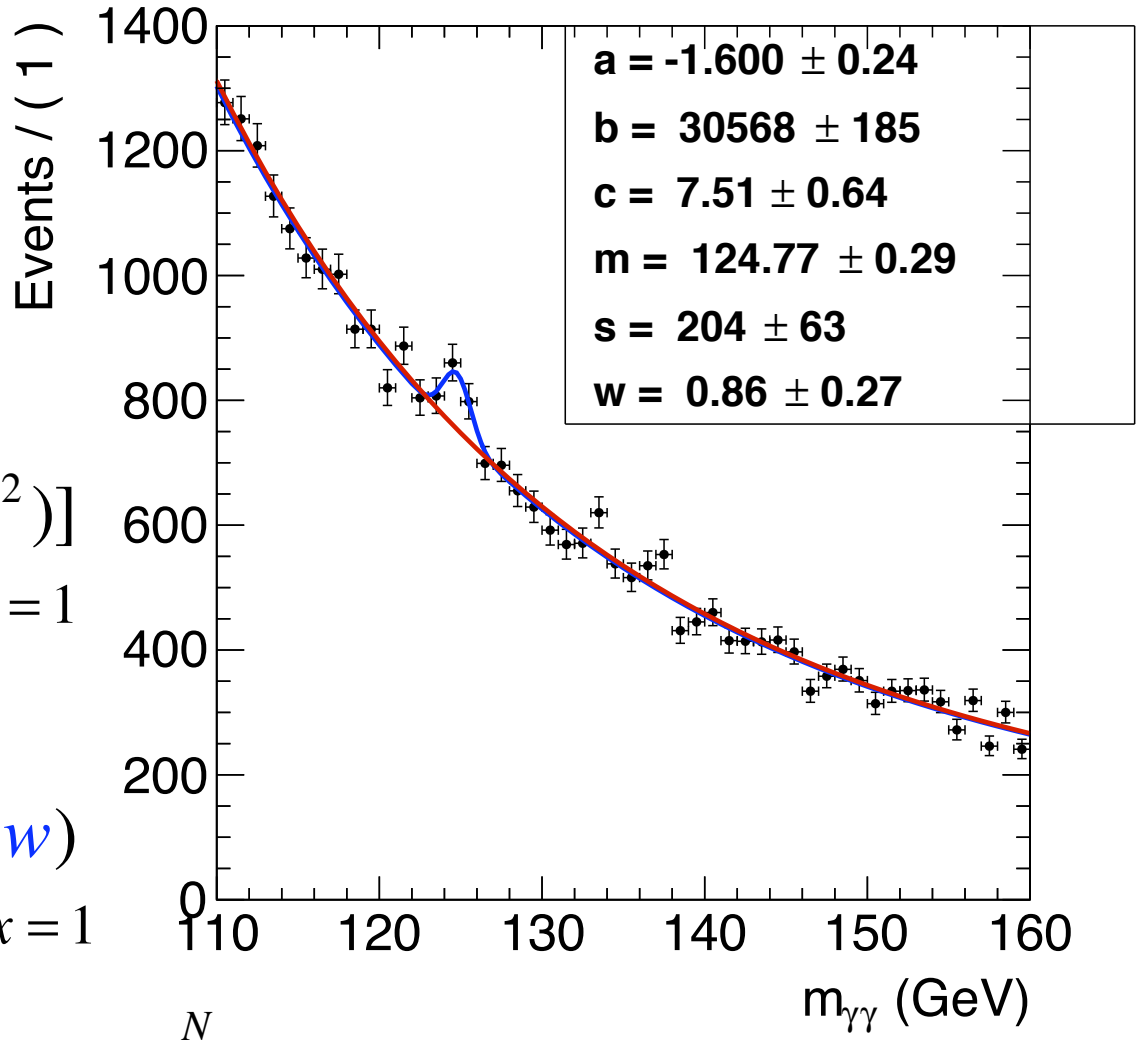
$$f_b(x \mid c, a) = A\exp[-(cx + ax^2)]$$

$$\int f_b(x \mid c, a)\,dx = 1$$

signal model

$$f_s(x \mid m, w) = \text{Gaussian}(x, m, w)$$

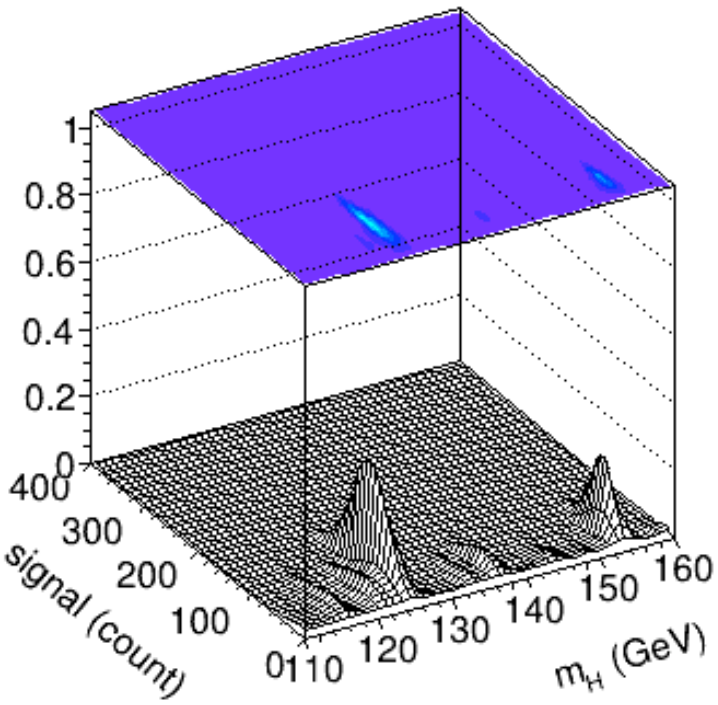
$$\int f_s(x \mid m, w)\,dx = 1$$



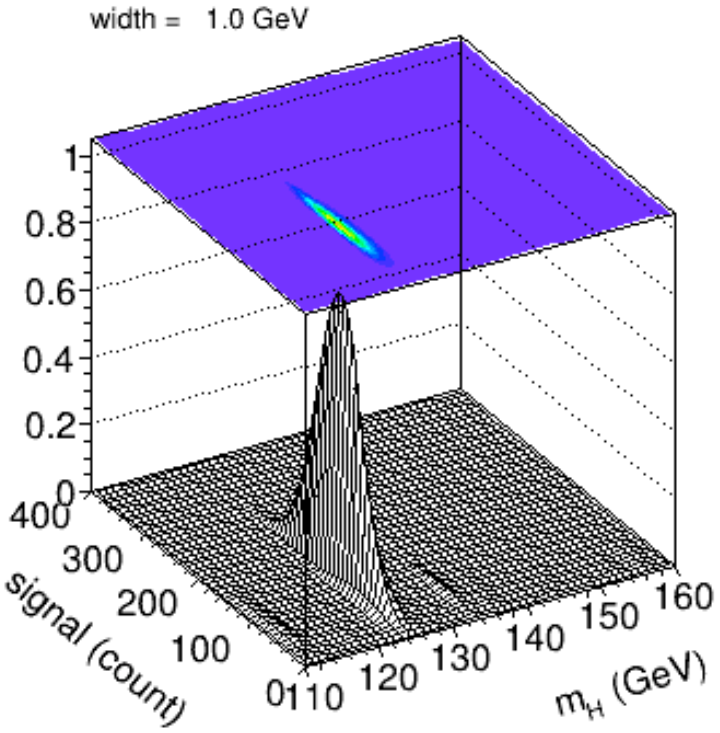
$$p(x \mid s, m, w, c, b, a) = \exp[-(s+b)]\prod_{i=1}^{N}\left[s f_s(x_i \mid m, w) + b f_b(x_i \mid c, a)\right]$$

Likelihood – Higgs to γγ (CMS)

7 TeV



8 TeV



Outline

- Lecture 1
  - Descriptive Statistics
  - Probability & Likelihood
- Lecture 2
  - The Frequentist Approach
  - The Bayesian Approach
- Lecture 3
  - Analysis Example

The Frequentist Approach
Confidence Intervals

The Frequentist Principle

The Frequentist Principle (Neyman, 1937)

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

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

Note: The confidence level is a property of the ***ensemble*** to which the statements belong. Consequently, the confidence level may change if the ensemble changes.

Confidence Intervals – 1

Consider an experiment that observes D events with expected signal s and no background.

Neyman devised a way to make statements of the form

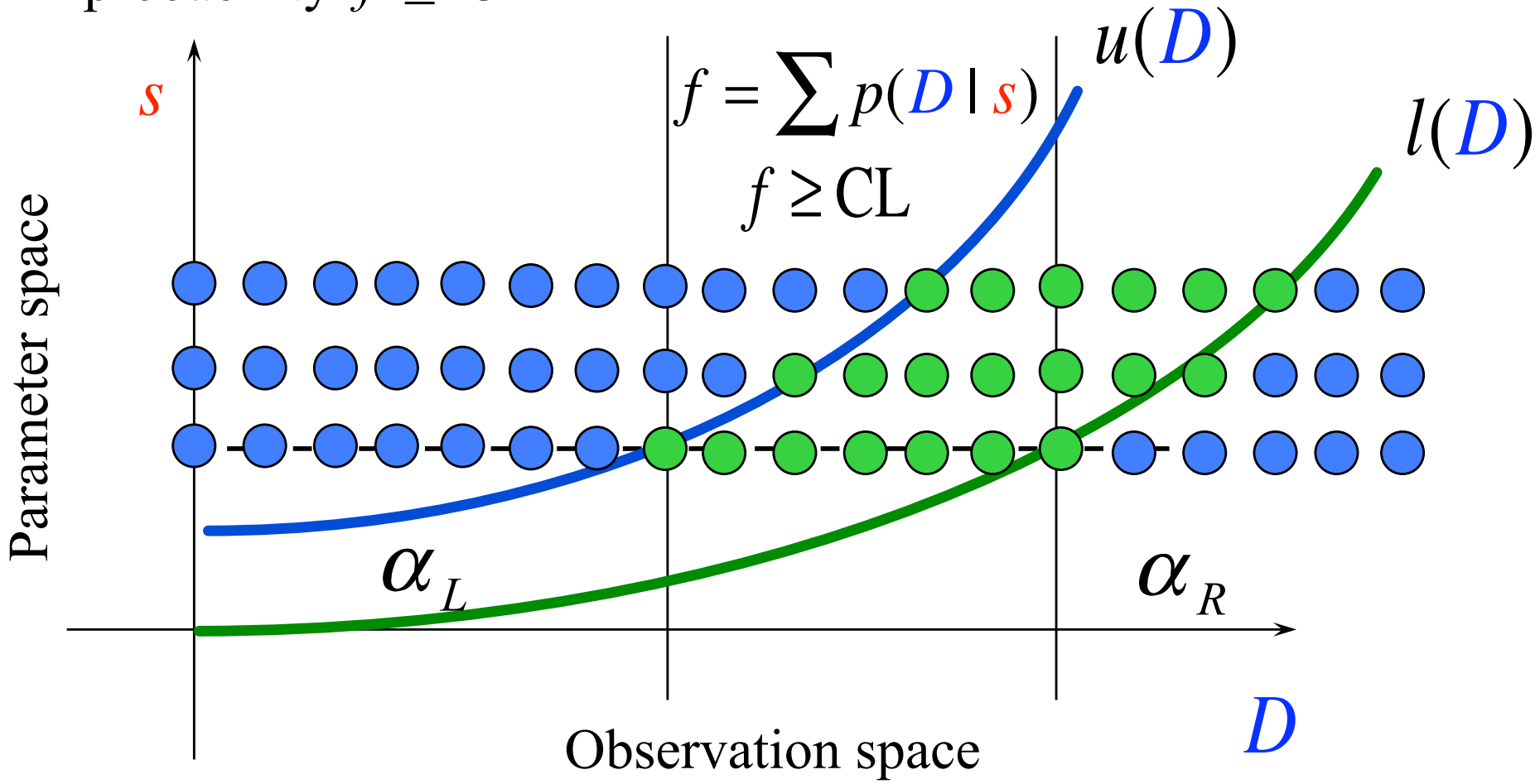
$$s \in [l(D),\ u(D)]$$

with the guarantee that at least a fraction p of them will be true.

But, since we don't know the true value of s, this property must hold ***whatever the true value*** of s.

Confidence Intervals – 2

For each value ***s*** find a region in the ***observation space*** with probability $f \geq \mathrm{CL}$



Confidence Intervals – 3

- **Central Intervals (Neyman)**

  Has equal probabilities on either side

- **Feldman – Cousins Intervals**

  Contains largest values of the ratios $p(D \mid s) / p(D \mid D)$

- **Mode – Centered Intervals**

  Contains largest probabilities $p(D \mid s)$

By construction, all these intervals satisfy the frequentist principle: ***coverage probability*** $\geq$ ***confidence level***

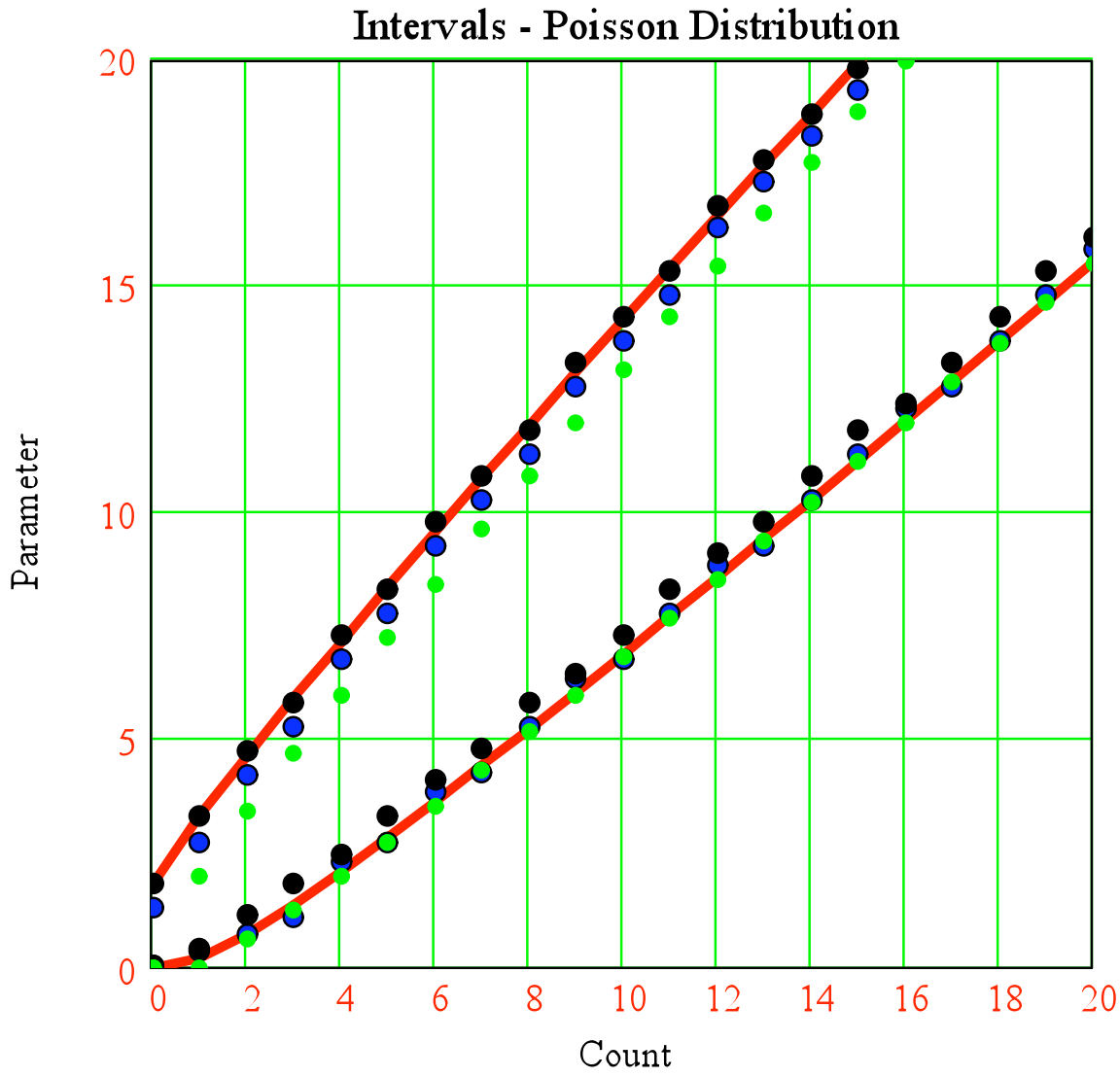
Confidence Intervals – 4

Central

Feldman-Cousins

Mode-Centered

$D \pm \sqrt{D}$



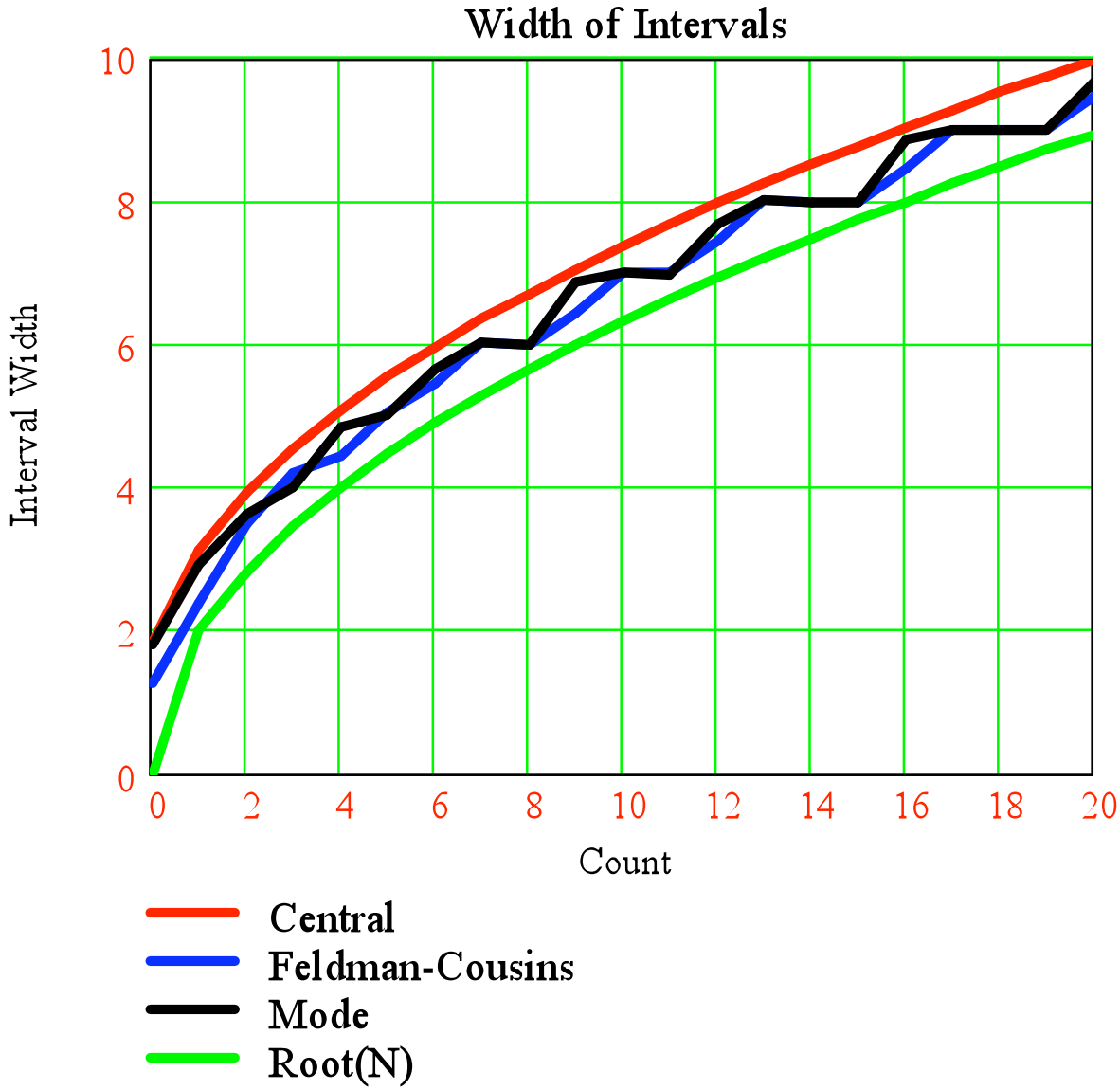
Confidence Intervals – 5

Central

Feldman-Cousins

Mode-Centered

$D \pm \sqrt{D}$



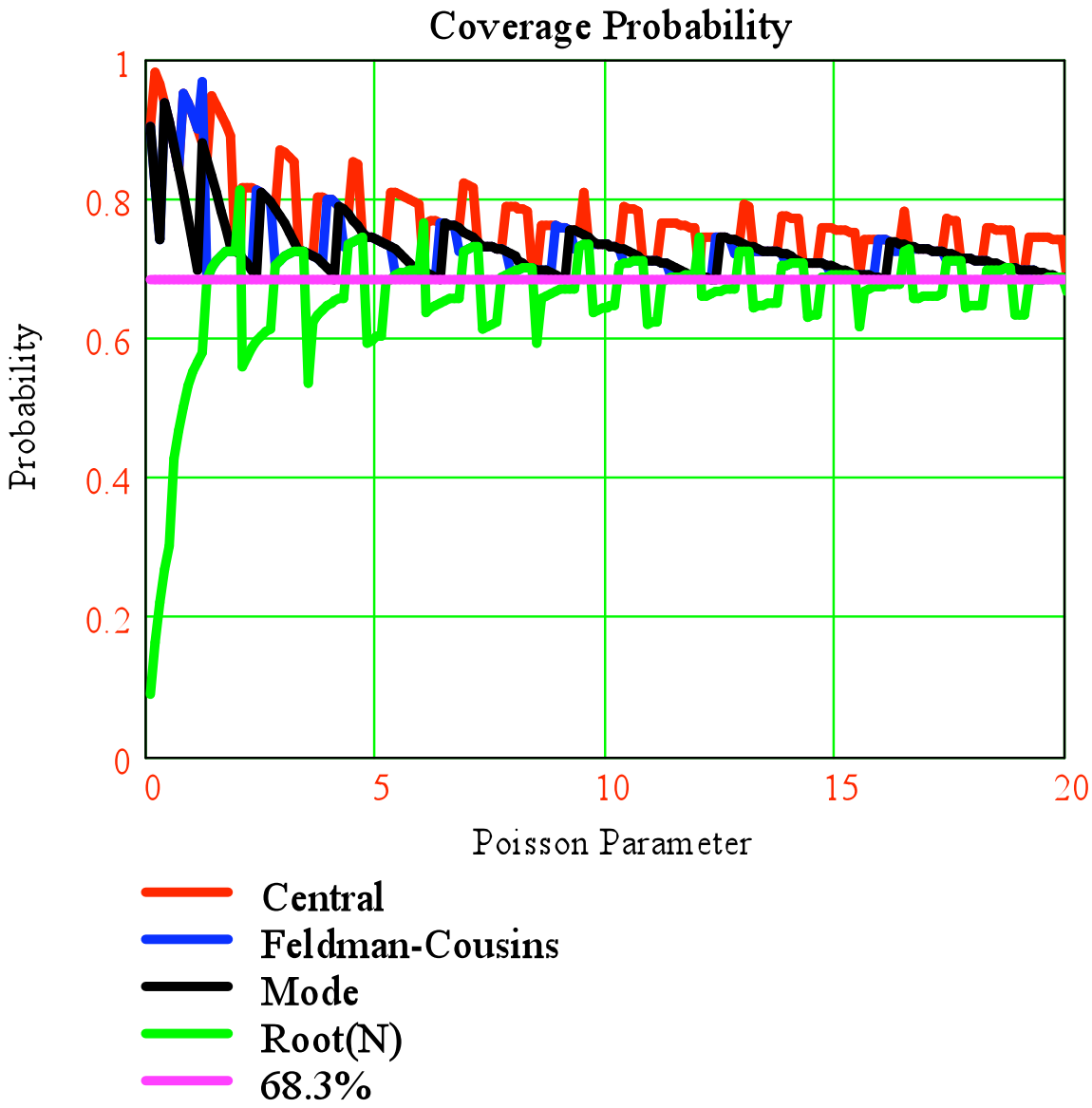
Confidence Intervals – 6

Central

Feldman-Cousins

Mode-Centered

$D \pm \sqrt{D}$



The Frequentist Approach
The Profile Likelihood

Maximum Likelihood – 1

Example: Top Quark Discovery (1995), D0 Results

D = 17 events

B = 3.8 ± 0.6 events

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

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

where

$B = Q / k$ $\qquad Q = (B / \delta B)^2 = (3.8 / 0.6)^2 = 41.11$

$\delta B = \sqrt{\text{Q}} / \text{k}$ $\qquad k = B / \delta B^2 = 3.8 / 0.6^2 = 10.56$

Maximum Likelihood – 2

knowns:

$D = 17$ events

$B = 3.8 \pm 0.6$ background events

unknowns:

b	expected background count
s	expected signal count

Find maximum likelihood estimates (MLE):

$$\frac{\partial \ln p(17 \mid s,b)}{\partial s} = \frac{\partial \ln p(17 \mid s,b)}{\partial b} = 0 \quad \Rightarrow \quad \hat{s},\ \hat{b}$$

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

Maximum Likelihood – 3

The **Good**

- Maximum likelihood estimates (MLE) are **consistent**: 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 7: Show this
Hint: consider a Taylor expansion about the MLE

The **Ugly**

- Correcting for bias, however, can waste data and sometimes yield absurdities

Example: The Seriously Ugly

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

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

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.,

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

Example: 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 $\boldsymbol{p^2 = 0}$!

The Profile Likelihood – 1

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

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

must be reduced to one involving s *only* by getting rid of all ***nuisance parameters***, such as ***b***.

In principle, this must be done while respecting the frequentist principle: ***coverage prob. ≥ confidence level***.

This is very difficult to do exactly.

The Profile Likelihood – 2

In practice, we replace all nuisance parameters by their conditional maximum likelihood estimates (CMLE), which yields a function called the ***profile likelihood***, $p_{PL}(D \mid s)$.

In the top quark discovery example, we find an estimate of b as a function of s

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

Then, in the likelihood $p(D \mid s, b)$, b is replaced with its ***estimate***.

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

The Profile Likelihood – 3

Wilks' Theorem (1938)

If certain conditions are met, and $p_{\max}$ is the value of the likelihood $p(D \mid s, b)$ at its maximum, the quantity

$$y(s) = -2 \ln \frac{p_{PL}(D \mid s)}{p_{\max}}$$

has a density that is asymptotically χ^2. Therefore, by setting $y(s) = 1$ and solving for s, we can compute *approximate* 68% confidence intervals.

This is what Minuit (now TMinuit) has been doing for 40 years!

The Profile Likelihood – 4

The CMLE of b is

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

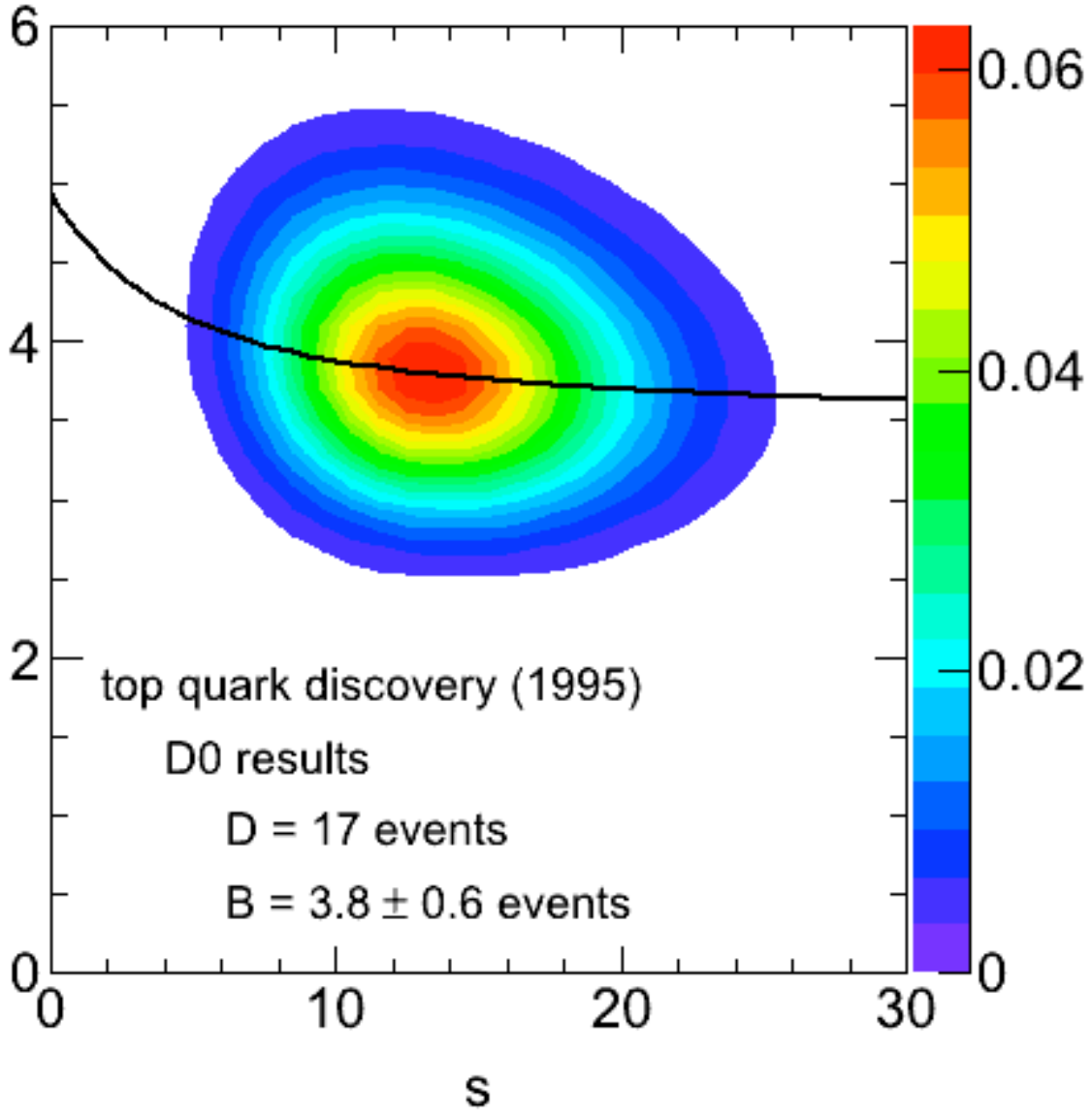
$$g = D + Q - (1+k)s$$

with

$$s = D - B$$

$$b = B$$

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



The Profile Likelihood – 5

By solving

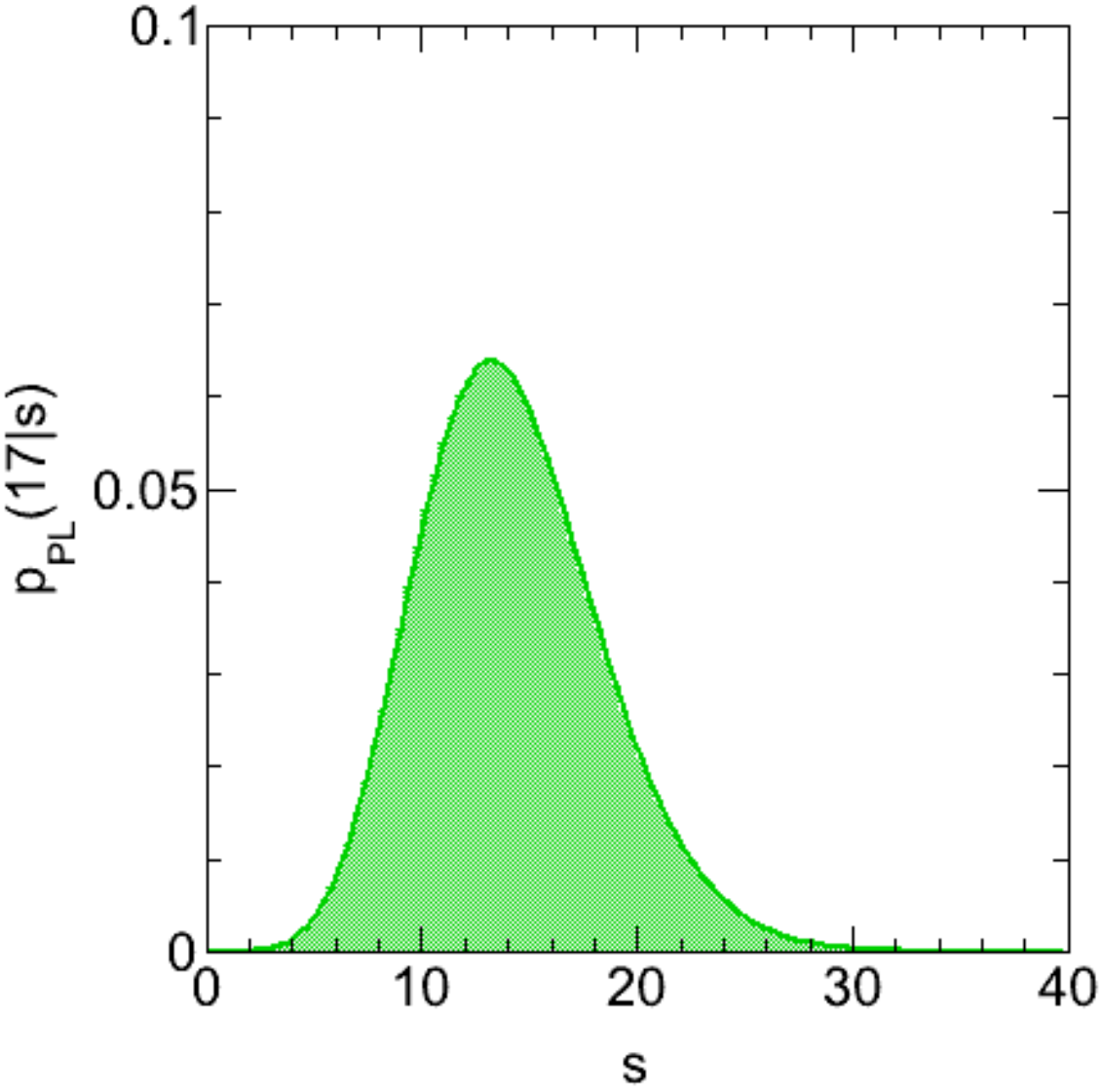
$$-2\ln\frac{p_{PL}(17\,|\,s)}{p_{\max}} = 1$$

for s, we can make the statement

$$s \in [9.4, 17.7]$$

@ 68% C.L.

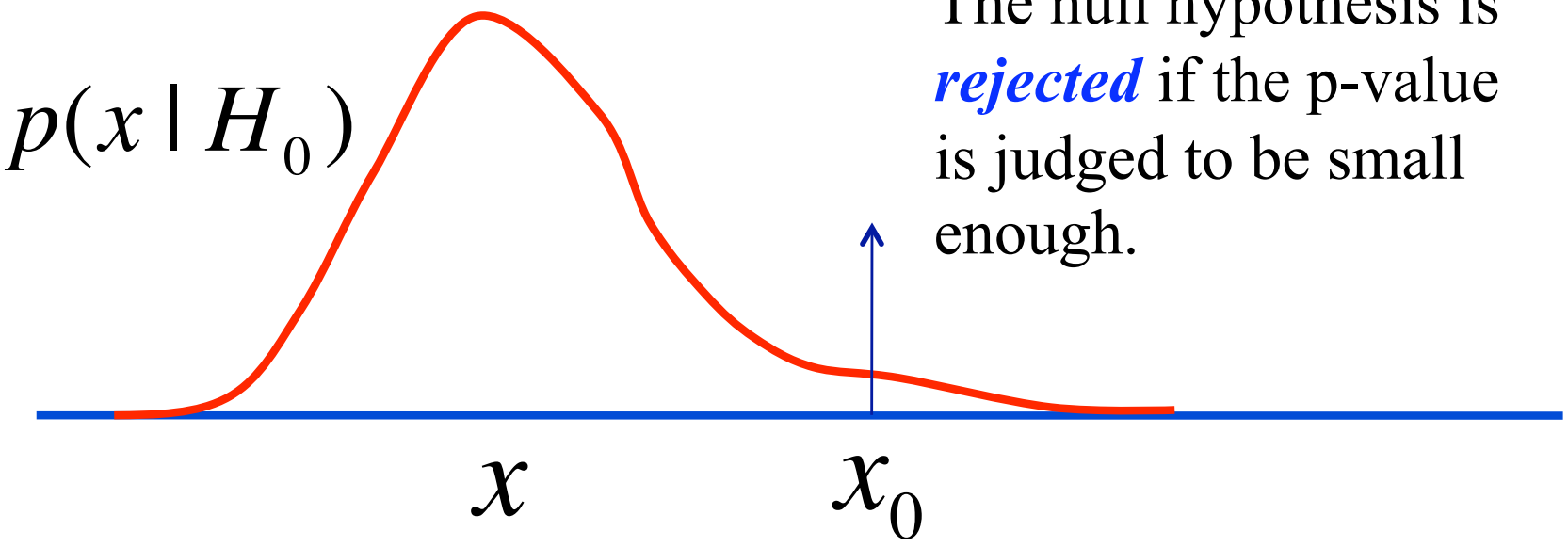
Exercise 9: Show this



The Frequentist Approach
Hypothesis Tests

Hypothesis Tests

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



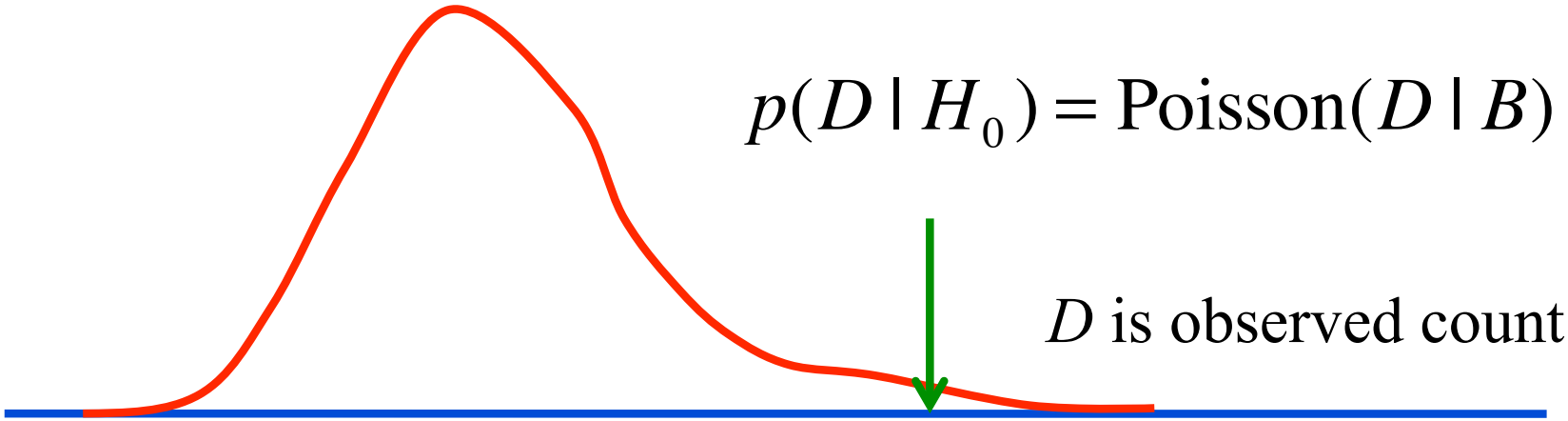
The null hypothesis is ***rejected*** if the p-value is judged to be small enough.

$$\text{p-value} \equiv \int_{x_0}^{\infty} p(x \mid H_0)\,dx$$

Note: this can be calculated only if $p(x \mid H_0)$ is known

Example – Top Quark Discovery

Background, $B = 3.8$ events (*ignoring uncertainty*)



$$D = 17$$

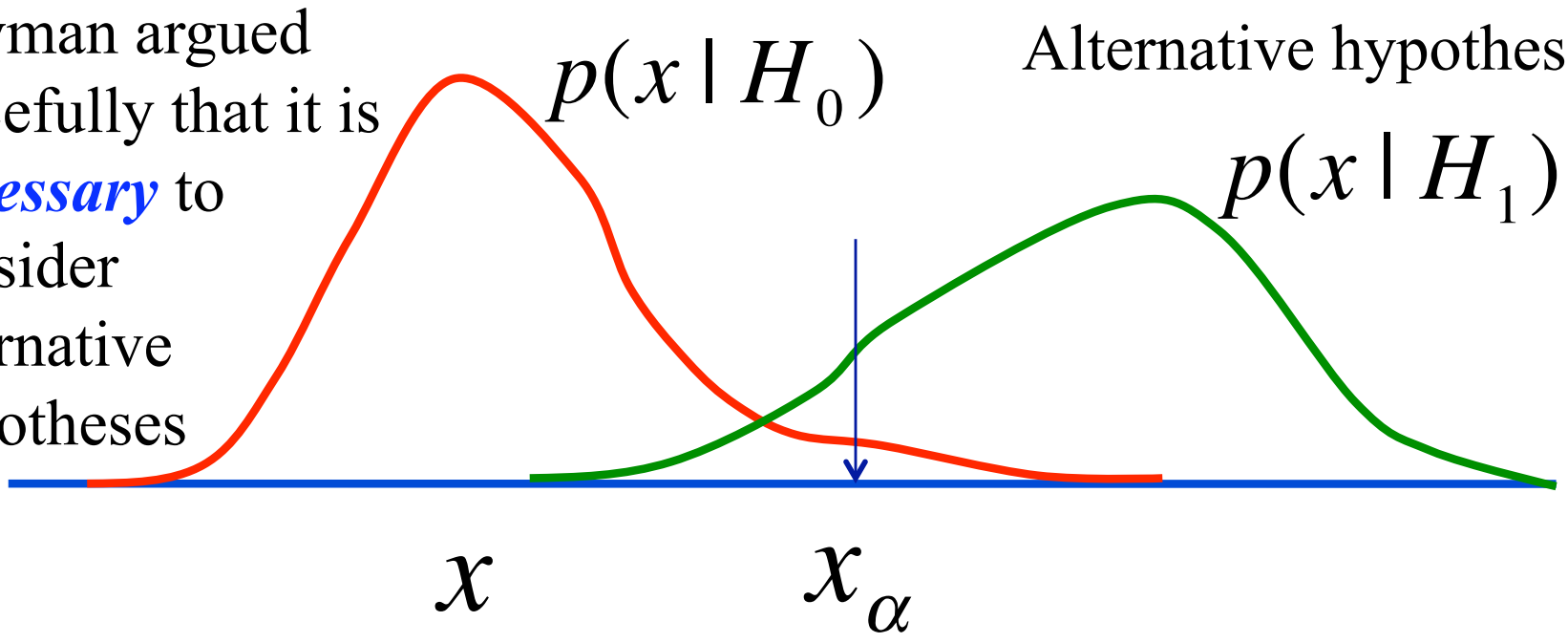
$$\text{p-value} = \sum_{D=17}^{\infty} \text{Poisson}(D \mid 3.8) = 5.7 \times 10^{-7}$$

This is equivalent to **4.9 σ**

Hypothesis Tests – 2

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

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

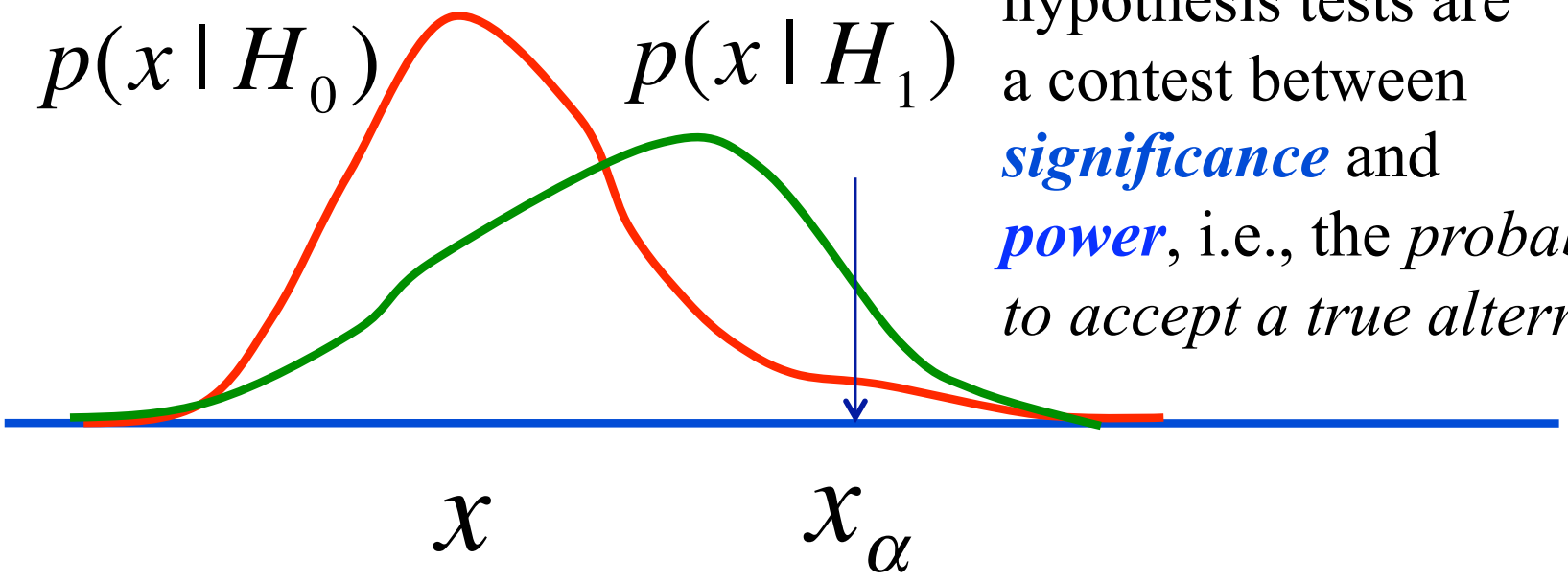


$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0)\,dx$$

A ***fixed*** significance α is chosen ***before*** data are analyzed.

significance (or size) of 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.*

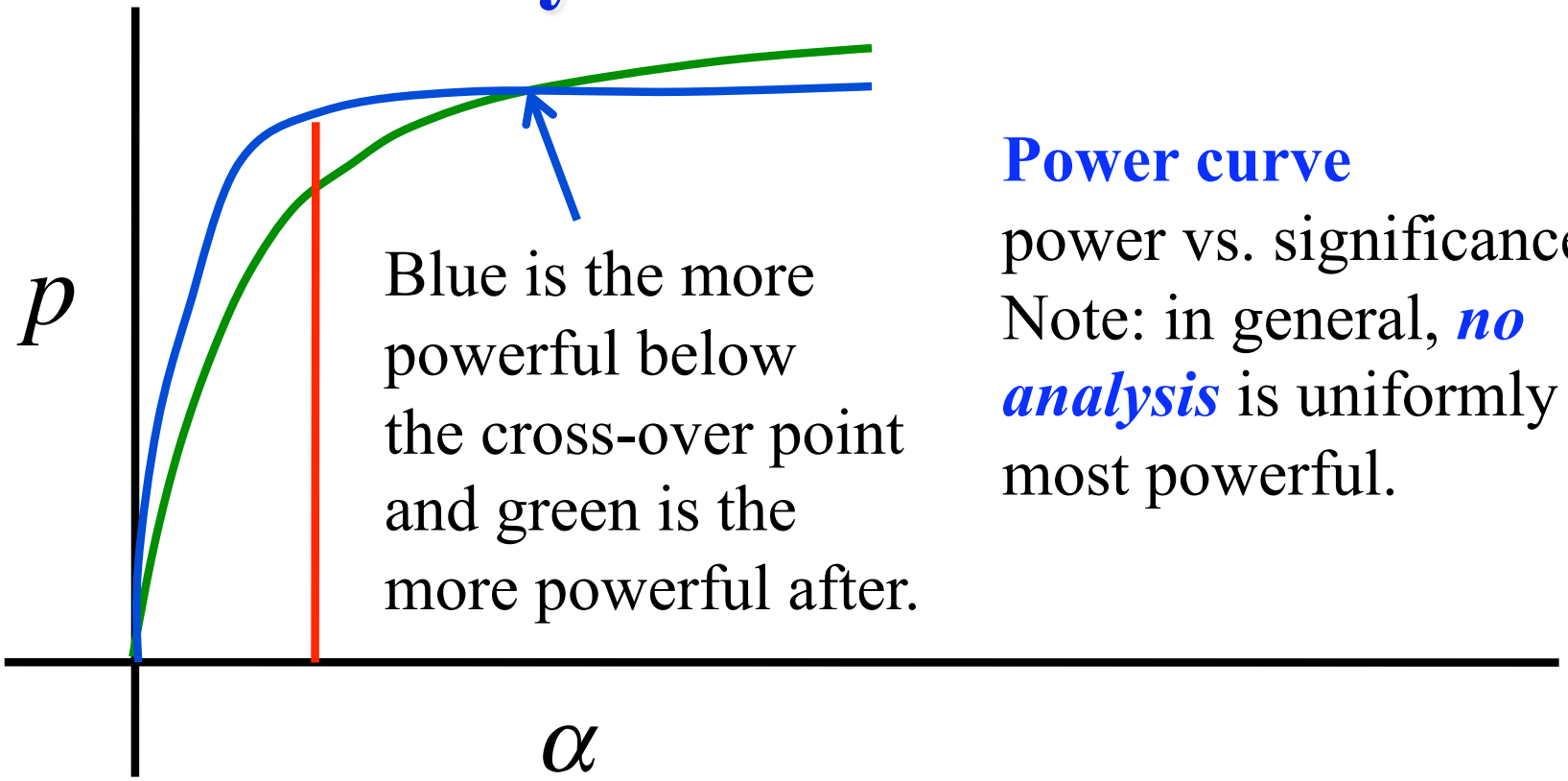
$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0)dx$$

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

The Bayesian Approach

The Bayesian Approach

Definition:

A method is Bayesian if

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

$$p(\theta,\omega \mid D) = \frac{p(D \mid \theta,\omega)\pi(\theta,\omega)}{p(D)}$$

for ***all*** inferences.

D	observed data
θ	parameter of interest
ω	nuisance parameters
π	***prior density***

The Bayesian Approach – 2

Bayesian analysis is just applied probability theory.

Therefore, the method for eliminating nuisance parameters is "simply" to integrate them out:

$$p(\theta \mid D) = \int p(\theta, \omega \mid D)\, d\omega$$

$$\propto \int p(D \mid \theta, \omega)\pi(\theta, \omega)\, d\omega$$

a procedure called ***marginalization***.

The integral is a weighted average of the likelihood.

The Bayesian Approach

An Example

Example – Top Quark Discovery – 1

D0 1995 Top Discovery Data

D = 17 events

B = 3.8 ± 0.6 events

Calculations

1. Compute the ***posterior density*** $p(s \mid D)$

2. Compute the ***Bayes factor*** $B_{10} = p(D \mid H_1) / p(D \mid H_0)$.

3. Compute the ***p-value***, including the effect of background uncertainty.

Example – Top Quark Discovery – 2

Step 1: Construct a probability model for the observations

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

then put in the data

$D = 17$ events

$B = 3.8 \pm 0.6$ background events

$Q = (B / \delta B)^2 = 40.1$ $\qquad B = Q / k$

$k = B / \delta B^2 = 10.6$ $\qquad \delta B = \sqrt{Q} / k$

to arrive at the likelihood.

Example – Top Quark Discovery – 3

Step 2: Write down Bayes' theorem:

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

and specify the prior:

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

It is useful to compute the following ***marginal likelihood***:

$$p(D \mid s) = \int p(D \mid s, b)\, \pi(b \mid s)\, db$$

sometimes referred to as the ***evidence*** for s.

Example – Top Quark Discovery – 4

The Prior: What do

$$\pi(b \mid s)$$

and

$$\pi(s)$$

represent?

They encode 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.

After a century of argument, the consensus today is that there is no *unique* way to represent such vague information.

Example – Top Quark Discovery – 5

For simplicity, we take $\pi(b \mid s) = 1$.

We may now eliminate b from the problem:

$$p(D \mid s, H_1) = \int_0^\infty p(D \mid s, b)\, \pi(b \mid s)\, d(kb)$$

$$= \frac{1}{Q}(1-x)^2 \sum_{r=0}^{D} \text{Beta}(x, r+1, Q)\, \text{Poisson}(D-r \mid s)$$

Exercise 10: Show this

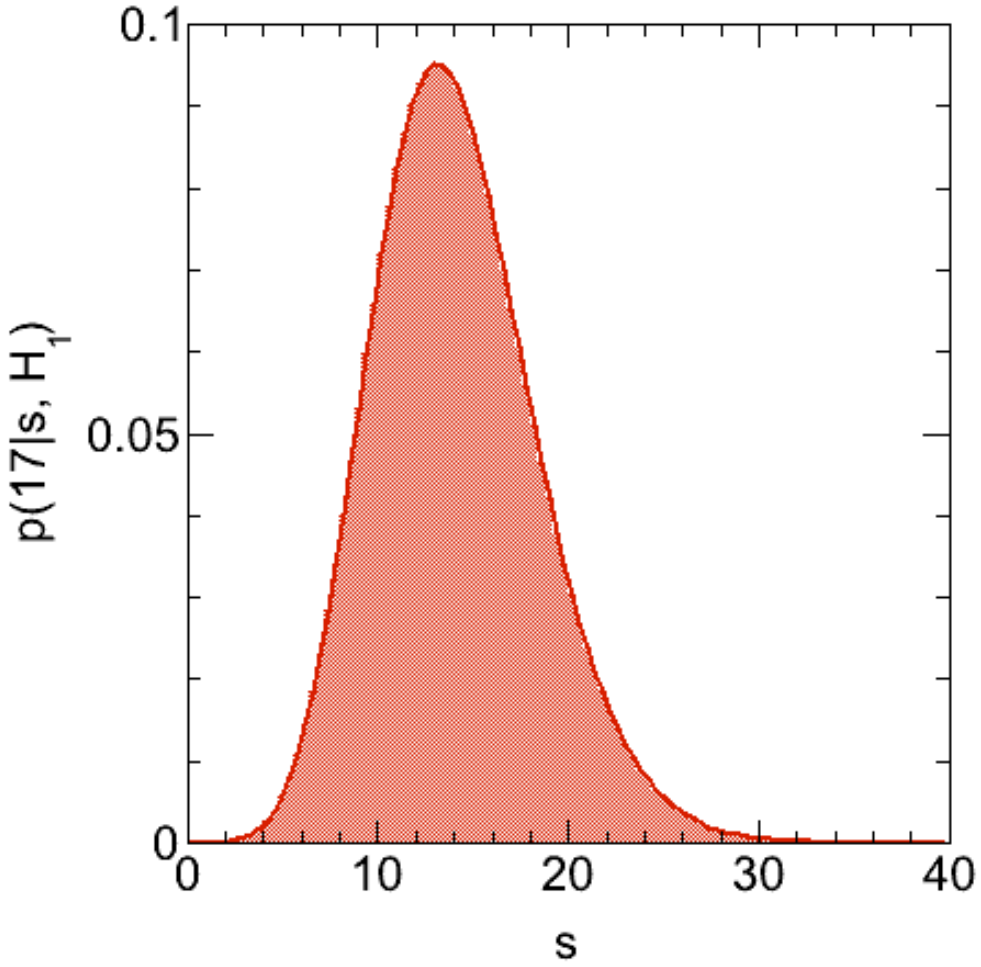
where,

$$x = \frac{1}{1+k}, \quad \text{Beta}(x, n, m) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} x^{n-1}(1-x)^{m-1}$$

and where we have introduced the symbol H_1 to denote the background + signal hypothesis.

Example – Top Quark Discovery – 6

$p(17|s, H_1)$ as a function of the expected signal s.



Example – Top Quark Discovery – 7

The posterior density

Given the marginal likelihood

$$p(D \mid s, H_1)$$

we can compute

$$p(s \mid D, H_1) = \frac{p(D \mid s, H_1)\,\pi(s \mid H_1)}{p(D \mid H_1)}$$

where

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

Example – Top Quark Discovery – 8

Assuming a ***flat prior*** for the signal $\pi(s \mid H_1) = 1$, the posterior density is given by

$$p(s \mid D, H_1) = \frac{\sum_{r=0}^{D} \text{Beta}(x, r+1, Q)\,\text{Poisson}(D-r \mid s)}{\sum_{r=0}^{D} \text{Beta}(x, r+1, Q)}$$

from which we can compute the *central interval*

$$s \in [9.9, 18.4]$$

@ 68% C.L.

Exercise 11: Derive an expression for $p(s \mid D, H_1)$ assuming a gamma prior Gamma(qs, $U+1$) for $\pi(s \mid H_1)$

The Bayesian Approach
Hypothesis Testing

Bayesian Hypothesis Testing – 1

Conceptually, Bayesian hypothesis testing proceeds in exactly the same way as any other Bayesian calculation: one computes the posterior density

posterior **likelihood** **prior**

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

and marginalize it with respect to all parameters except those label the hypotheses

$$p(H \mid D) = \iint p(\theta,\phi,H \mid D)\,d\theta\,d\phi$$

and of course get your ***pHD***!

Bayesian Hypothesis Testing – 2

However, just like your PhD, it is usually more convenient, and instructive, to arrive at the ***p(H | D)*** in stages.

1. Factorize the priors: $\pi(\theta, \varphi, H) = \pi(\theta, \varphi \mid H)\ \pi(H)$

2. Then, for each hypothesis, H, compute the function

$$p(D \mid H) = \iint p(D \mid \theta, \phi, H)\, \pi(\theta, \phi \mid H)\, d\theta\, d\phi$$

3. Then, compute the **probability of each hypothesis, *H***

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

Bayesian Hypothesis Testing – 3

It is clear, however, that to compute ***p(H | D)***, it is necessary to specify the priors ***π(H)***.

Unfortunately, consensus on these numbers is highly unlikely!

Instead of asking for the probability of an hypothesis, ***p(H | D)***, we could compare probabilities:

$$\frac{p(H_1 \mid D)}{p(H_0 \mid D)} = \left[\frac{p(D \mid H_1)}{p(D \mid H_0)}\right]\left[\frac{\pi(H_1)}{\pi(H_0)}\right]$$

The ratio in the first bracket is called the **Bayes factor**, B_{10}.

Bayesian Hypothesis Testing – 4

In principle, in order to compute the number

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

twe need to specify a ***proper*** prior for the signal, that is, a prior that integrates to one.

For simplicity, let's assume $\pi(s \mid H_1) = \delta(s - 14)$.

This yields:

$p(D \mid H_0) = 3.86 \times 10^{-6}$

$p(D \mid H_1) = p(D \mid 14, H_1) = 9.28 \times 10^{-2}$

Bayesian Hypothesis Testing – 5

Since,

$p(D \mid H_0)$ = **3.86 x 10^{-6}**

$p(D \mid H_1)$ = **9.28 x 10^{-2}**

we conclude that the hypothesis with ***s*** = 14 events is favored over that with ***s*** = 0 by **24,000** to **1**.

The Bayes factor can be mapped to a measure akin to "n-sigma"

$$Z = \sqrt{2 \ln B_{10}} = 4.5$$

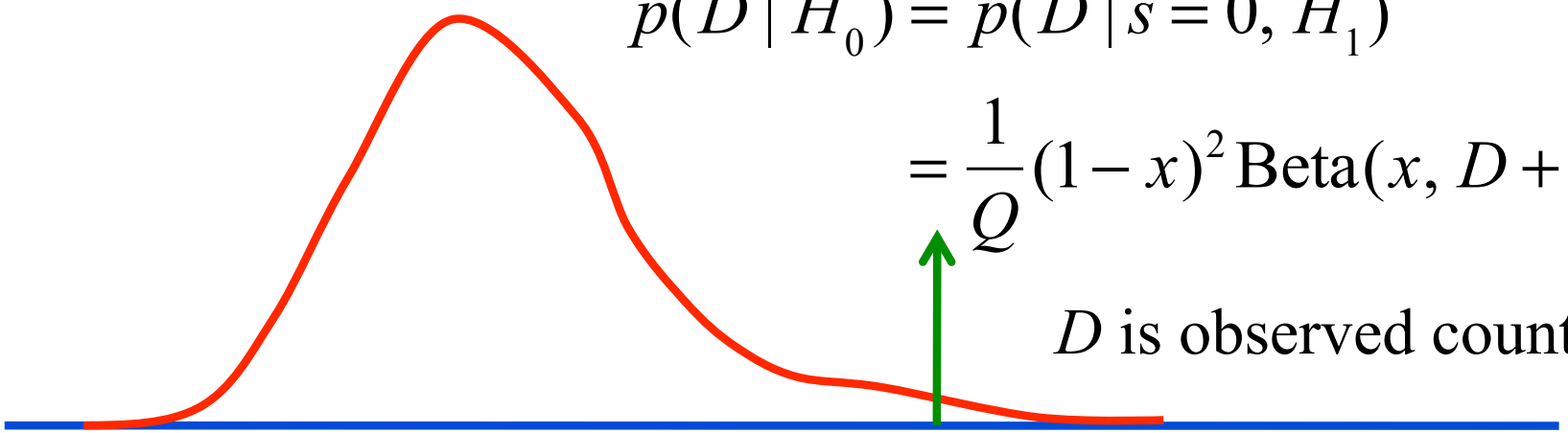
Exercise 12: Compute Z for the D0 results

Hypothesis Testing – A Hybrid Approach

Background, $B = 3.8 \pm \mathbf{0.6}$ events

$$p(D \mid H_0) = p(D \mid s = 0, H_1)$$

$$= \frac{1}{Q}(1-x)^2 \,\mathrm{Beta}(x, D+1, Q)$$



$$D = 17$$

$$\text{p-value} = \sum_{D=17}^{\infty} p(D \mid H_0) = 5.4 \times 10^{-6}$$

This is equivalent to **4.4 σ** which may be compared with the 4.5 σ obtained with B_{10}

Exercise 13: Verify this calculation

Summary

Frequentist Approach

1) Models statistical uncertainties with probabilities
2) Uses the likelihood
3) Ideally, respects the frequentist principle
4) In practice, eliminates nuisance parameters through the *approximate* procedure of profiling

Bayesian Approach

1) Models ***all*** uncertainties with probabilities
2) Uses likelihood and prior probabilities
3) Eliminate nuisance parameters through marginalization.