Practical Statistics for Particle Physicists
Lecture 2

Harrison B. Prosper

Florida State University

European School of High-Energy Physics

Anjou, France

6 – 19 June, 2012

Outline

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

Practicum

Toy data and code at

http://www.hep.fsu.edu/~harry/ESHEP12

topdiscovery.tar (already there)

contactinteractions.tar

just download and unpack

Recap – The Frequentist Approach

The Frequentist Principle (Neyman, 1937)

Construct statements such that a fraction $f \geq p$ of them will be true over an (infinite) ensemble of statements.

f is called the ***coverage probability***

p is called the ***confidence level*** (CL).

The Frequentist Approach – 2
The Profile Likelihood

The Profile Likelihood – 1

Example: Top Quark Discovery (1995), D0 Results

$D \quad = 17$ events

$B \quad = 3.8 \pm 0.6$ events

$$p(D \mid s, b) = \text{Poisson}(D, s+b)\ \text{Gamma}(k, b, Q+1)$$

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

where

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

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

The Profile Likelihood – 2

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 a problem 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***.

In general, this is very difficult to do exactly.

The Profile Likelihood – 3

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 $p(D|s, b)$, b is replaced with its estimate.

This is an approximation because the likelihood principle is not guaranteed to be satisfied exactly

The Profile Likelihood – 4

Wilks' Theorem (1938)

If certain conditions are met, and $p_{\max}$ is the value of the likelihood $p(D|s, b)$ 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 be doing for 40 years!

The Profile Likelihood – 5

The CMLE of b is

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

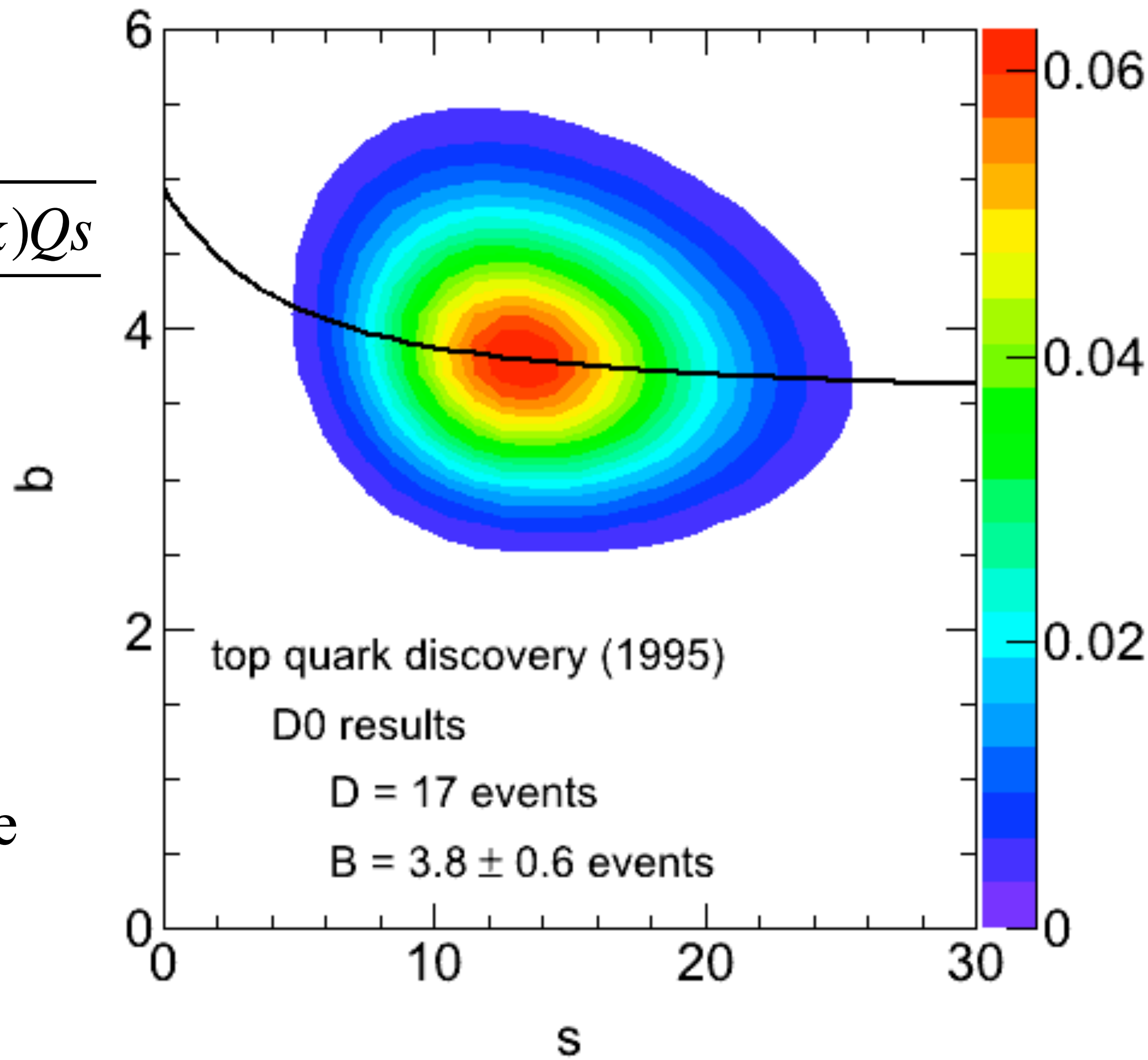
$$g = D + Q - (1+k)s$$

with

$$s = D - B$$

$$b = B$$

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



The Profile Likelihood – 5

By solving

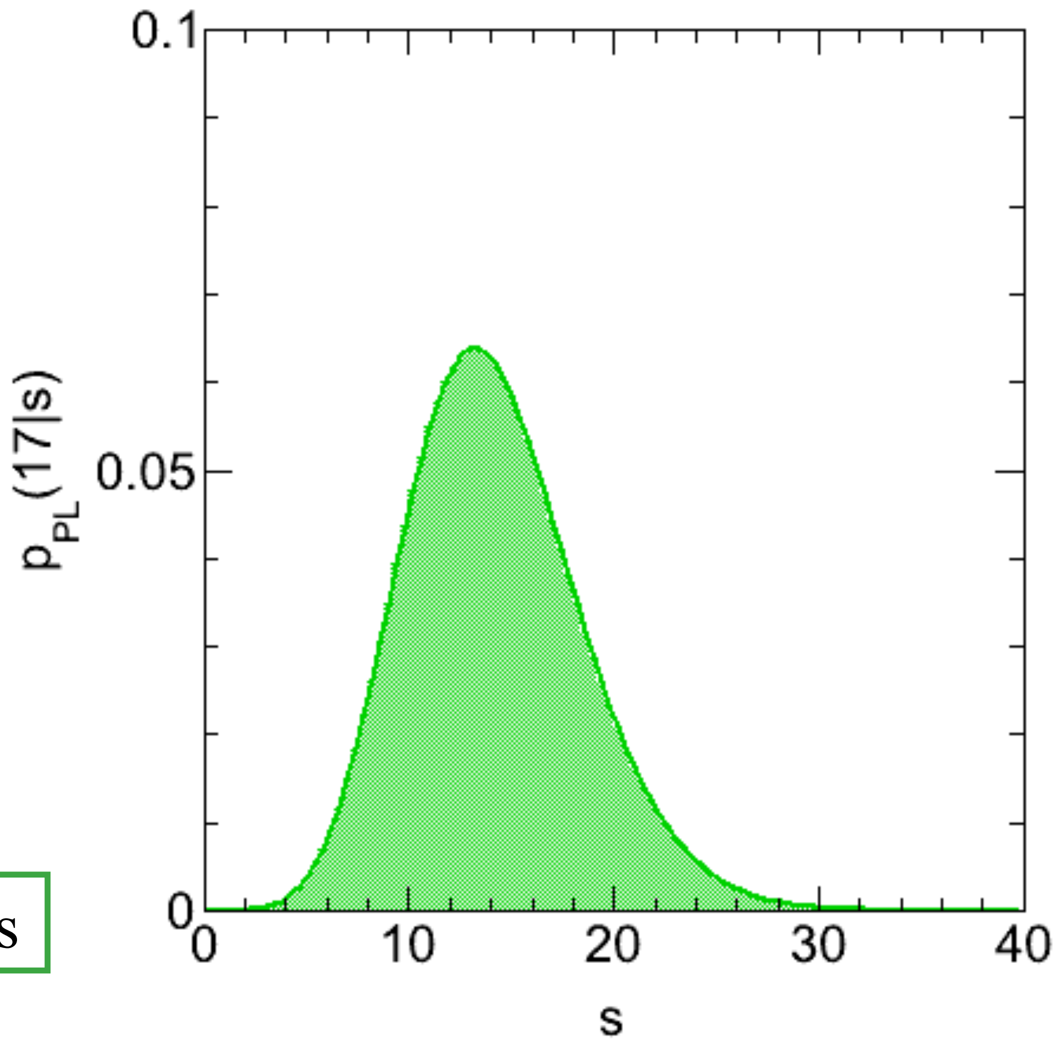
$$-2\ln\frac{p_{PL}(17\,|\,s)}{p_{\max}} = 1$$

for *s*, we can make the statement

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

@ 68% CL

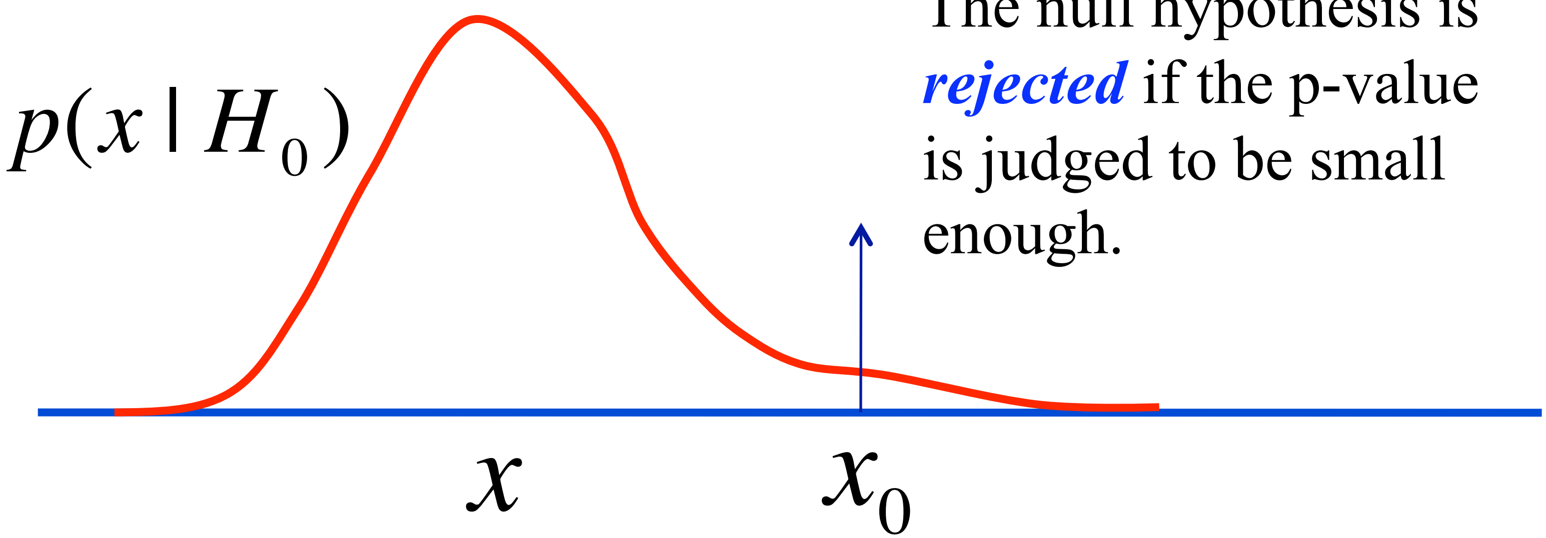
Exercise 6: Show this



The Frequentist Approach
Hypothesis Tests

Hypothesis Tests

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



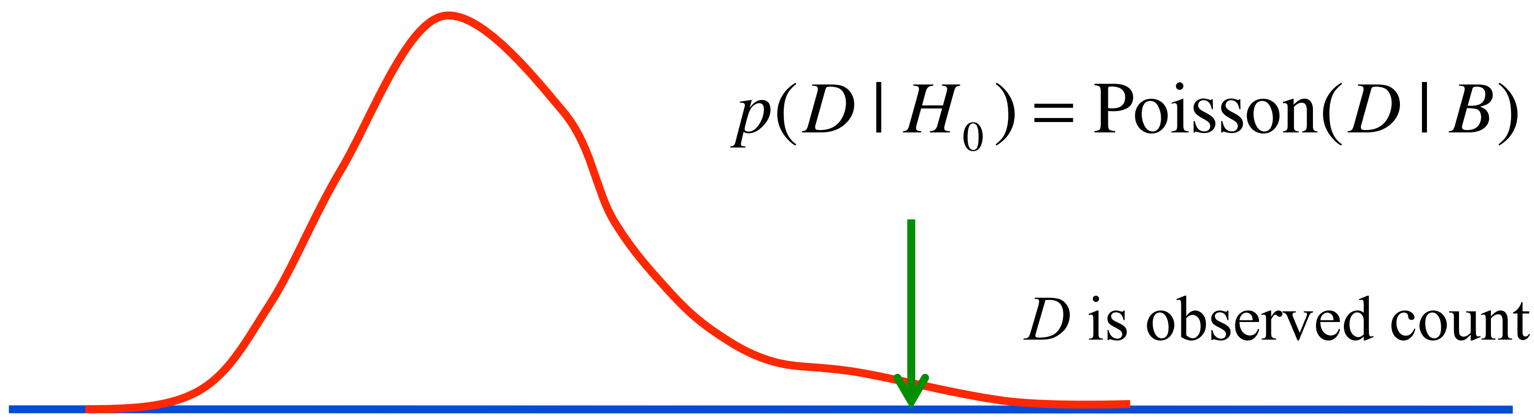
The null hypothesis is ***rejected*** if the p-value is judged to be small enough.

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

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

Example – Top Quark Discovery

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



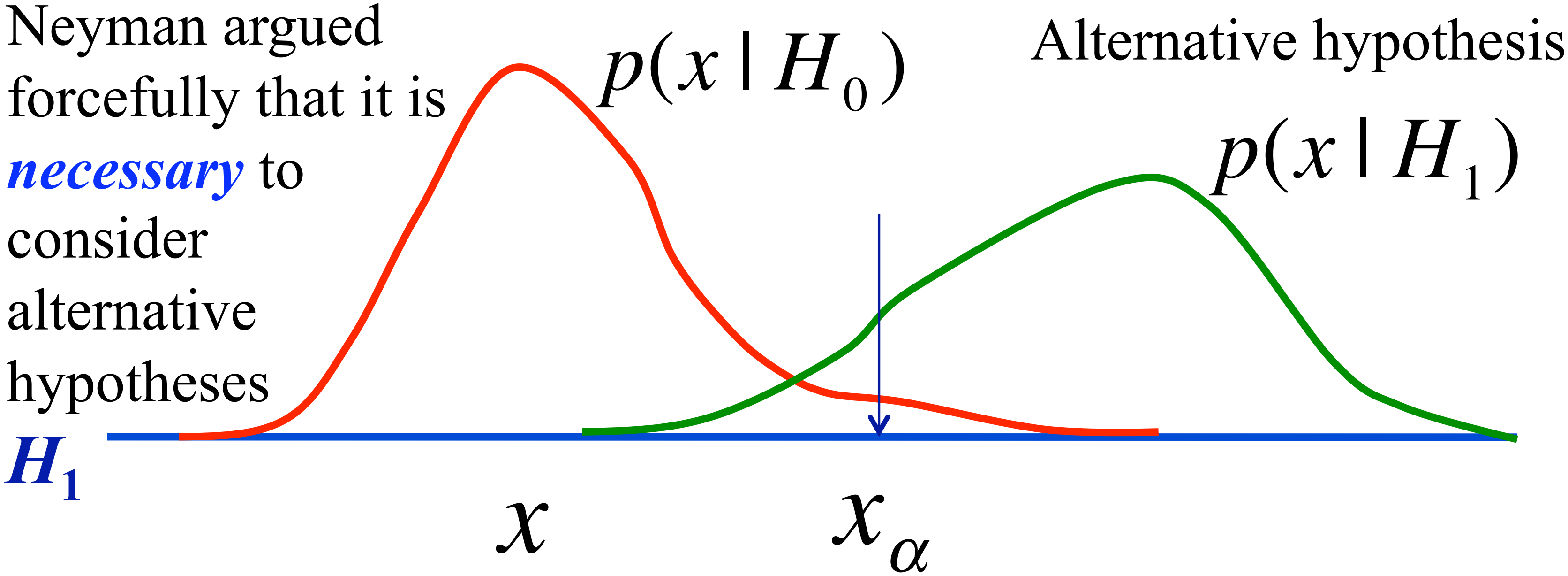
$$\text{p-value} = \sum_{D=17}^{\infty} \text{Poisson}(D \mid 3.8) = 5.7 \times 10^{-7}$$

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

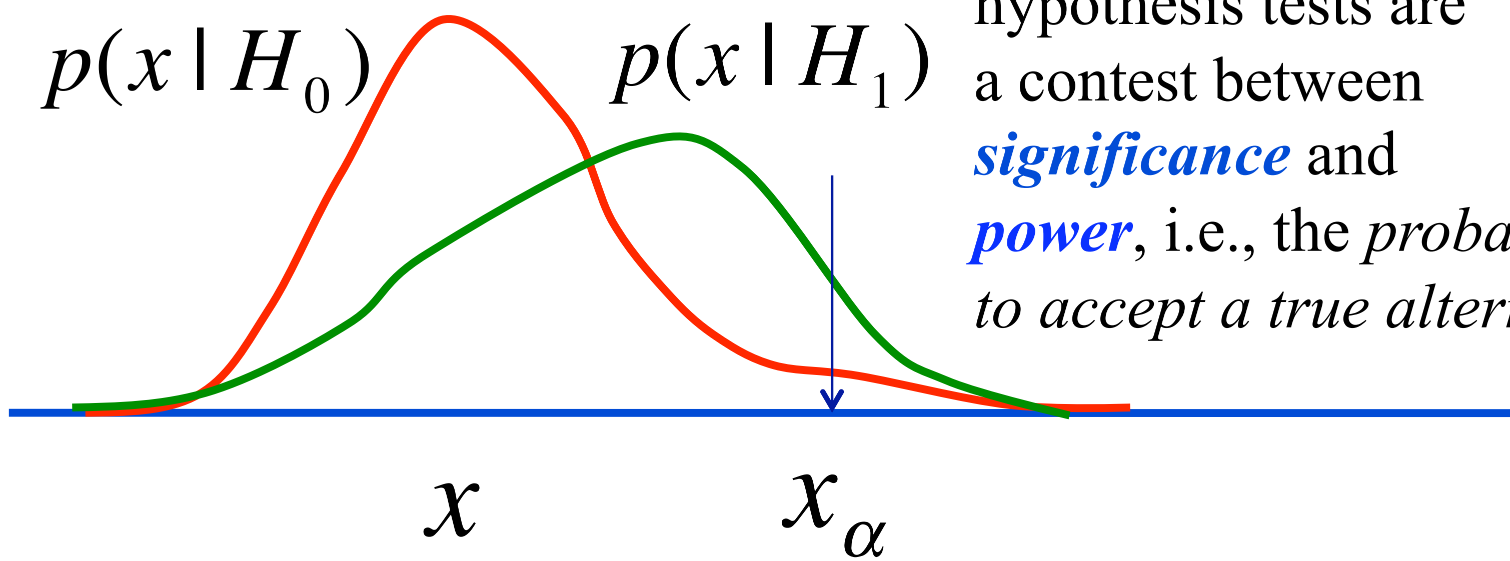


$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0)\,dx$$

significance of test

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

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.*

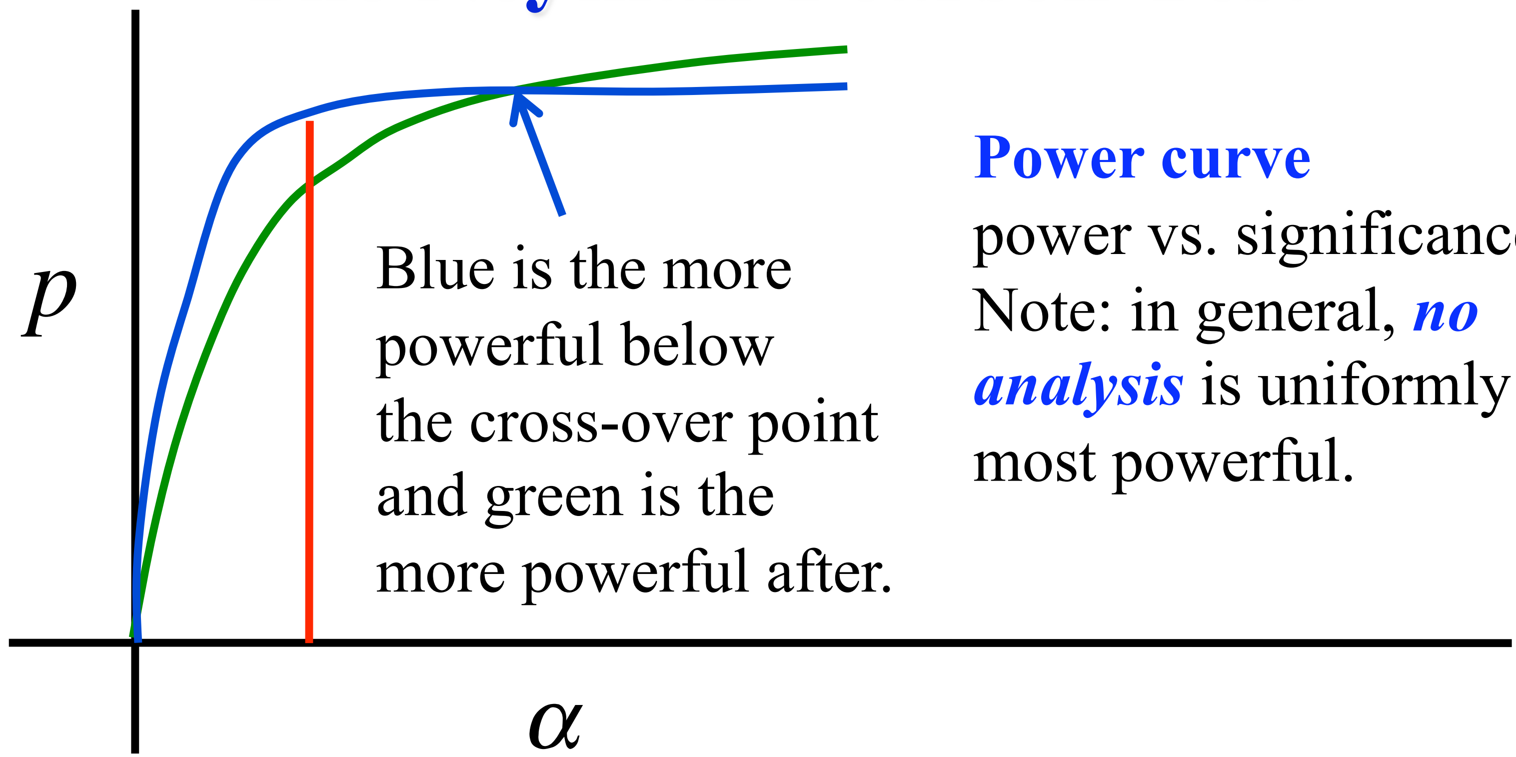
$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0)\,dx$$

significance of test

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

power

The Neyman-Pearson Test



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 Frequentist Approach
A Word About Bias

Bias is Not a Four-Letter Word!

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 7: Show this

which is useful for calculating **moments**

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

e.g.,

$$M_2 = (np)^2 + np\,(1 - p)$$

Bias is Not a Four-Letter Word!

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

$$p = k / n$$

and the obvious estimate of p^2 is

$$k^2 / n^2$$

But

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

Exercise 8: Show this

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

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

Exercise 9: Show this

but it is crazy: for a single success, $p = 1/n$, but $p^2 = 0$!

The Bayesian Approach

The Bayesian Approach

Definition:

A method is Bayesian if

1. it is based on the ***subjective*** 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.

Consequently, 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***.

This is simply a weighted average of the likelihood.

The Bayesian Approach
An Example

Example – The Top Quark Discovery

D0 Data (1995)

$D = 17$ events

$B = 3.8 \pm 0.6$ estimated background events

Parameters

b = expected (i.e., mean) background count

s = expected (i.e., mean) signal count

$d = b + s$

Analysis goals:

1. Estimate (i.e., *measure*) the expected signal s
2. Quantify its significance
3. Drink champagne, if warranted!

Example – The Top 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 plug in the data

$D = 17$ events

$B = 3.8 \pm 0.6$ background events

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

$k = B / \delta B^2 = 10.6$

$B = Q / k$

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

to arrive at the likelihood.

Example – The Top 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 – The Top 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 – The Top Discovery – 5

For now, we take $\pi(b|s) = 1$.

We may now eliminate b from the problem:

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

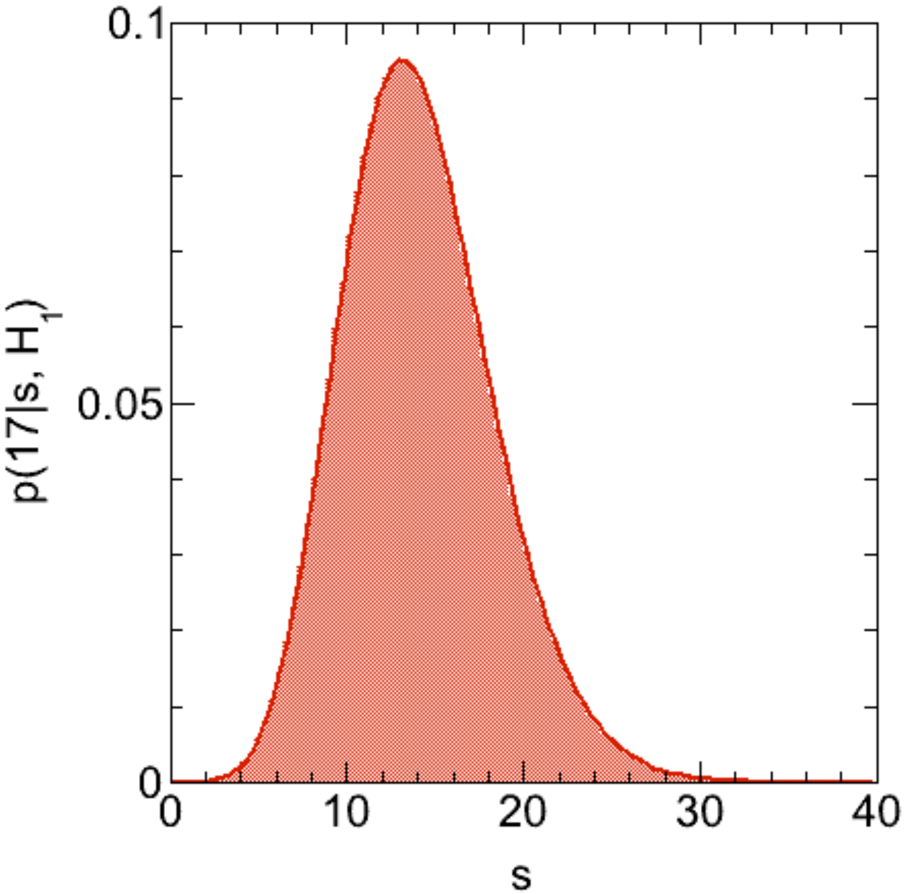
$$= \left(\frac{k}{1+k}\right)^{Q+1} \sum_{r=0}^{D} \frac{1}{(1+k)^r} \frac{\Gamma(Q+1+r)}{\Gamma(Q+1)r!} \text{Poisson}(D-r\,|\,s)$$

The symbol H_1 denotes the background + signal hypothesis.

Note: $p(D|s = 0, H_0) = p(D|H_0)$ is the evidence for the background-only hypothesis, H_0.

Example – The Top Discovery – 6

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

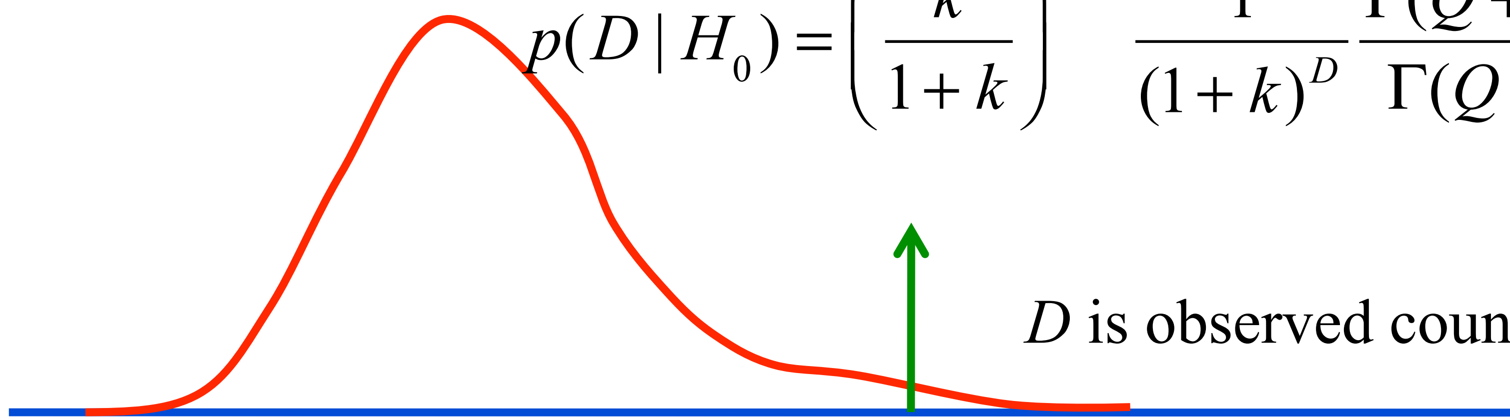


The Bayesian Approach
Hypothesis Testing

Hypothesis Testing - A Hybrid Approach

Background, $B = 3.8$ **± 0.6** events

$$p(D \mid H_0) = \left(\frac{k}{1+k}\right)^{Q+1} \frac{1}{(1+k)^D} \frac{\Gamma(Q+1+D)}{\Gamma(Q+1)D!}$$



$$D = 17$$

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

This is equivalent to **4.4 σ**

Hypothesis Testing – 1

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

posterior **likelihood** **prior**

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

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

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

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

Hypothesis Testing – 2

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

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

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

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

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

$$p(H | D) = \frac{p(D | H)\pi(H)}{\sum_H p(D | H)\pi(H)}$$

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 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}.

Exercise 10: Compute B_{10} for the D0 results

Summary

Frequentist Hypothesis Testing

Two approaches:

1) reject null if p-value is judged to be too small
2) decide on a fixed threshold for rejection and reject null if threshold has been breached.

Bayesian Approach

1) Model all uncertainty using probabilities and use Bayes' theorem to make inferences.
2) Eliminate nuisance parameters through marginalization.

Summary

- **Hypothesis Tests**
  - It is necessary to specify priors for each of hypothesis.
  - In particular, for our simple counting experiment, we need to specify the prior $p(s)$ for the signal since it is part of the specification of the background+signal hypothesis.
  - Unfortunately, doing so ***sensibly*** is hard!

More tomorrow!