Statistics in Physics Analysis
Lecture 2

Harrison B. Prosper
Florida State University

Workshop on LHC Physics
TIFR, Mumbai

21 – 27 October, 2009

Outline

- Lecture 2 – **Foundations & Applications**
  - The Bayesian Approach
  - Decisions & Loss
  - Hypothesis Tests
  - Summary

The Bayesian Approach

The Bayesian Approach

A Bayesian calculation requires the following ingredients:

$p(D \mid \theta, \varphi)$	the **probability model** that represents the mechanism that gave rise to the observed data D, given some *unknown* values of the parameters θ, φ.
$p(\theta, \varphi)$	the **prior probability density** over the parameter space of the probability model

The Bayesian Approach

Then one calculates the posterior density as follows:

posterior marginal likelihood prior

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\int p(D \mid \theta)\, p(\theta)\, d\theta}$$

where the **marginal (or integrated) likelihood** is given by

$$p(D \mid \theta) = \int_{\Phi} p(D \mid \theta, \phi)\, p(\phi \mid \theta)\, d\phi$$

and $p(\theta, \phi) = p(\phi \mid \theta)\, p(\theta)$ is the full prior density.

Example – D0 Top Discovery Data

D0 1995 Top Discovery Data

$n \qquad = 17$ events

$b_0 \qquad = 3.8 \pm 0.6$ events

Example – D0 Top Discovery Data

Likelihood Functions

$p(n|b+s, H_1)$ = Poisson$(n|b+s)$ = $\exp[-(b+s)]\,(b+s)^n / n!$

$p(n|b, H_0)$ = Poisson$(n|b)$ = $\exp[-b]\,b^n / n!$

Prior Density

$p(\mathrm{b}, \mathrm{s})$ = $p(\mathrm{b}|\mathrm{s})\,p(\mathrm{s})$

$p(b|s)$ = Gamma$(kb|B+1)$ = $k \exp(-kb)\,(kb)^B / \Gamma(B+1)$

where the effective scale factor k and count B are

$b_0 = B / k$ $\qquad B = (b_0 / \delta b)^2 = (3.8 / 0.6)^2 = 41.11$

$\delta b = \sqrt{\mathrm{B}} / \mathrm{k}$ $\qquad k = b_0 / \delta b^2 = 3.8 / 0.6^2 = 10.56$

Example – Integrated Likelihoods

The **integrated likelihoods** are

$$p(n \mid s, H_1) = \int_0^\infty \text{Poisson}(n \mid b+s)\text{Gamma}(kb \mid B+1)db$$

$$= \left(\frac{k}{1+k}\right)^{B+1} \sum_{r=0}^{n} \frac{1}{(1+k)^r} \frac{\Gamma(B+1+r)}{\Gamma(B+1)r!} \text{Poisson}(n-r \mid s)$$

and

$$p(n \mid H_0) = p(n \mid s=0, H_1) = \left(\frac{k}{1+k}\right)^{B+1} \frac{1}{(1+k)^n} \frac{\Gamma(B+1+n)}{\Gamma(B+1)n!}$$

Exercise 1: Compute these integrated likelihoods

Example – Posterior Density

Given the integrated likelihood

$$p(n \mid s, H_1) = \left(\frac{k}{1+k}\right)^{B+1} \sum_{r=0}^{n} c_r(k,B)\,\text{Poisson}(n-r \mid s)$$

where

$$c_r(k,B) \equiv \frac{1}{(1+k)^r}\,\frac{\Gamma(B+1+r)}{\Gamma(B+1)r!}$$

we can compute

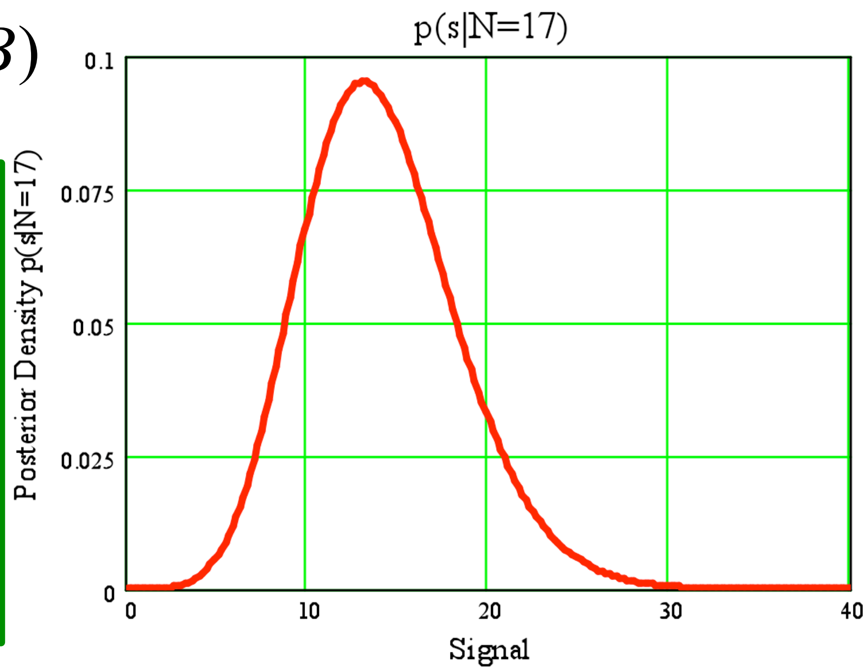
$$p(s \mid n, H_1) = \frac{p(n \mid s, H_1)\,p(s \mid H_1)}{\int_0^\infty p(n \mid s, H_1)\,p(s \mid H_1)\,ds}$$

Example – Posterior Density

Assuming a ***flat prior*** for the signal $p(s|H_1) = constant$, the posterior density is given by

$$p(s \mid n, H_1) = \frac{\sum_{r=0}^{n} c_r(k,B)\,\text{Poisson}(n-r \mid s)}{\sum_{r=0}^{n} c_r(k,B)}$$

> Exercise 2: Compute $p(s|n, H_1)$.
>
> Repeat assuming the prior density
> $p(s|H_1) = \text{Gamma}(qs|S+1)$,
> where $S = (s_0 / \delta s)^2$ and
> $q = s_0 / \delta s^2$



Decisions & Loss

Decisions and Loss

The posterior density ***p(θ| D)*** is the ***complete*** answer to an inference about the parameter ***θ***.

However, it is often of interest to summarize this answer with a **point estimate *θ*** * (a measurement) and, or, an **interval estimate** [θ_L, θ_U].

Or, we wish to decide which of two or more competing models is preferred by the data.

Decision theory provides a general way to model such problems.

Decisions and Loss

One way to render a decision about the value of $\boldsymbol{\theta}$ is to implement the decision as a function $\boldsymbol{d}$ that returns an **estimate** $\boldsymbol{\theta}$* of $\boldsymbol{\theta}$. A function $\boldsymbol{d}$ that returns estimates is called an **estimator**.

In principle, we also need to specify a **loss function** $\boldsymbol{L(d, \theta)}$ that quantifies what we lose should the estimate turn out to have been a bad one.

Decisions and Loss

In practice, since our knowledge of the parameter θ is encoded in the posterior density $p(\theta \mid D)$, our decisions will be more ***robust*** if we average (E[*]) the loss $L(d, \theta)$ with respect to $p(\theta \mid D)$

$$\begin{aligned} R(d) &= \mathrm{E}[L(d, \theta)] \\ &= \int L(d, \theta)\, p(\theta \mid D)\, \mathrm{d}\theta \end{aligned}$$

The quantity $R(d)$ is called the **risk function**.

By definition, the **optimal estimate** of θ is the one that minimizes the risk

$$\theta^* = \arg\min_{\mathrm{d}} R(d)$$

Comments

In general, different loss functions will yield different estimates.

Therefore, even with ***exactly the same data*** one should not be surprised to obtain different results.

Reasonable people can disagree about the results simply because they disagree about what properties of the results are thought to be most useful.

For example, many insist that a result should always be ***unbiased***, while others do not!

Comments

Consider a loss function $L(d, m)$ to extract a value for the Higgs mass, m, from a posterior density $p(m|D)$.

Suppose $L(d, m)$ is invariant in the following sense: it yields an estimate m^* of m which, when inserted into the prediction $\sigma = g(m)$ for the Higgs cross section, yields an estimate of the cross section $\sigma^* = g(m^*)$ that is *identical* to the one obtained using the loss function $L(d, \sigma)$.

$L(d, \sigma)$ is the loss function $L(d, m)$ with m replaced by σ.

In general, either m^* or σ^* (or both) will be **biased**.

Comments

To see this, expand $\sigma^* = g(m^*)$ about the ***true*** Higgs mass m

$$\sigma^* \approx g(m) + (m^* - m)\, g' + \tfrac{1}{2}\,(m^* - m)^2\, g''$$

and average both sides over an **ensemble** of estimates. This gives

$$E[\sigma^*] \approx \sigma + \textbf{bias}\; g' + \tfrac{1}{2}\textbf{mse}\; g'',$$

$$E[\sigma^*] \approx \sigma + \textbf{bias}\; g' + \tfrac{1}{2}[\textbf{bias}^2 + \textbf{variance}]\; g'',$$

where **bias** $= E[m^*] - m$ and **variance** $= E[m^{*2}] - E[m^*]^2$.

mse: mean square error (note: **rms** $= \sqrt{\textbf{mse}}$)

Decisions and Loss

Point Estimation

quadratic loss

$L(d, \theta) = (d - \theta)^2$

Average with respect to $p(\theta | D)$

risk $R(d) = E[(d - \theta)^2]$

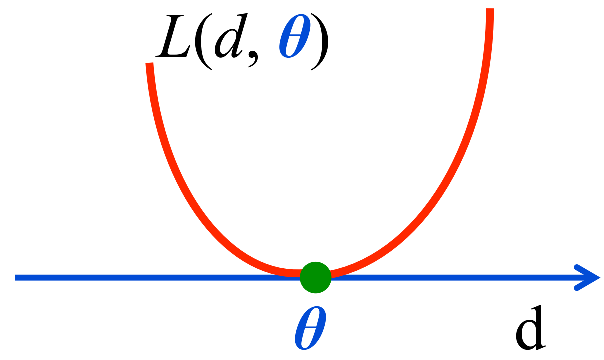
$= E[d^2] - 2E[\theta d] + E[\theta^2]$

$= d^2 - 2E[\theta]d + E[\theta^2]$

minimize with respect to d

$dR/dd = 2d - 2E[\theta] = 0$

obtaining, $\theta^* = \mathbf{E}[\theta]$

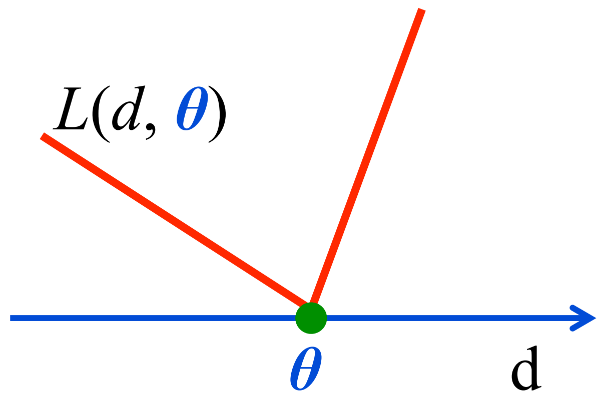


Note: quadratic loss is ***not invariant***. If $a = g(\theta)$, then $L(d, a) = (d - a)^2$ gives $a^* = E[a] \neq g(\theta^*)$

Decisions and Loss

Point Estimation

bilinear loss



$$L(d,\theta) = \begin{cases} a(\theta - d), & d < \theta \\ b(d - \theta), & d \geq \theta \end{cases}$$

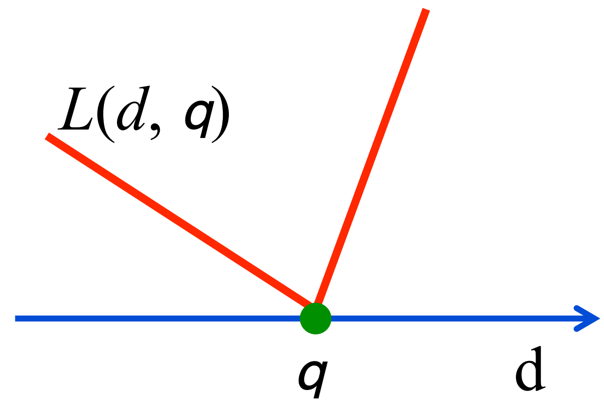
risk
$$R(d) = a\int H(\theta - d)(\theta - d)p(\theta \mid D)d\theta + b\int H(d - \theta)(d - \theta)p(\theta \mid D)d\theta$$

$$\boxed{H(x) = 1 \text{ if } x > 0 \text{ else } 0}$$

Decisions and Loss

Point Estimation

bilinear loss



The optimal estimate is

$$\boldsymbol{\theta}^* = \arg\min_d R(d)$$

where $\boldsymbol{\theta}^*$ is the $a/(a+b)$ **quantile**

$$\int^{\theta^*} p(\theta \mid D) d\theta = a/(a+b)$$

of $p(\boldsymbol{\theta} \mid D)$. If we set $a = b$, $\boldsymbol{\theta}^*$ = **median** of $p(\boldsymbol{\theta} \mid D)$

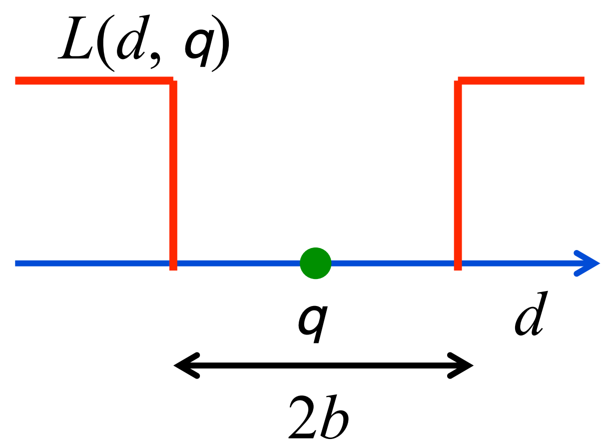
Note: estimates based on quantiles are ***invariant***.

Decisions and Loss

Point Estimation

zero-one loss

$$L(d,\theta)=\begin{cases}0, & |d-\theta|\le b\\ 1, & |d-\theta|>b\end{cases}$$



Its risk function is

$$R(d)=\int[H(\theta-b-d)+H(d-\theta-b)]p(\theta\,|\,D)d\theta$$

and the optimal estimate $\boldsymbol{\theta}^* = \min_d R(d)$ is the solution of

$$p(\boldsymbol{\theta}^* + b|D) = p(\boldsymbol{\theta}^* - b|D).$$

In the limit $b \to 0$, one obtains $\boldsymbol{\theta}^*$ = **mode** of $p(\boldsymbol{\theta}\,|\,D)$. The mode is ***not invariant***.

Example – Posterior Mean

Compute the moments of $p(s|n, H_1)$ about zero

$$M_m = \int_0^\infty s^m p(s \mid n, H_1) ds$$

$$= \sum_{r=0}^{n} c_r(k,B)(n-r+m)!/\,(n-r)!\,/\sum_{r=0}^{n} c_r(k,B)$$

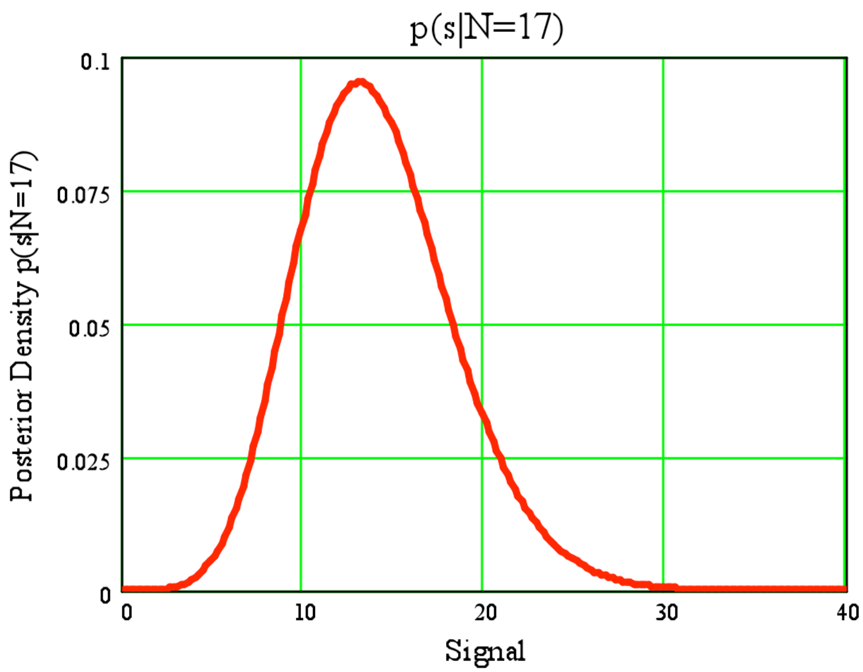
For the D0 top quark discovery data we find:

mean

M_1 = 14.0 events

standard deviation

$\sqrt{(M^2 - M_1^{\,2})}$ = 4.3 events

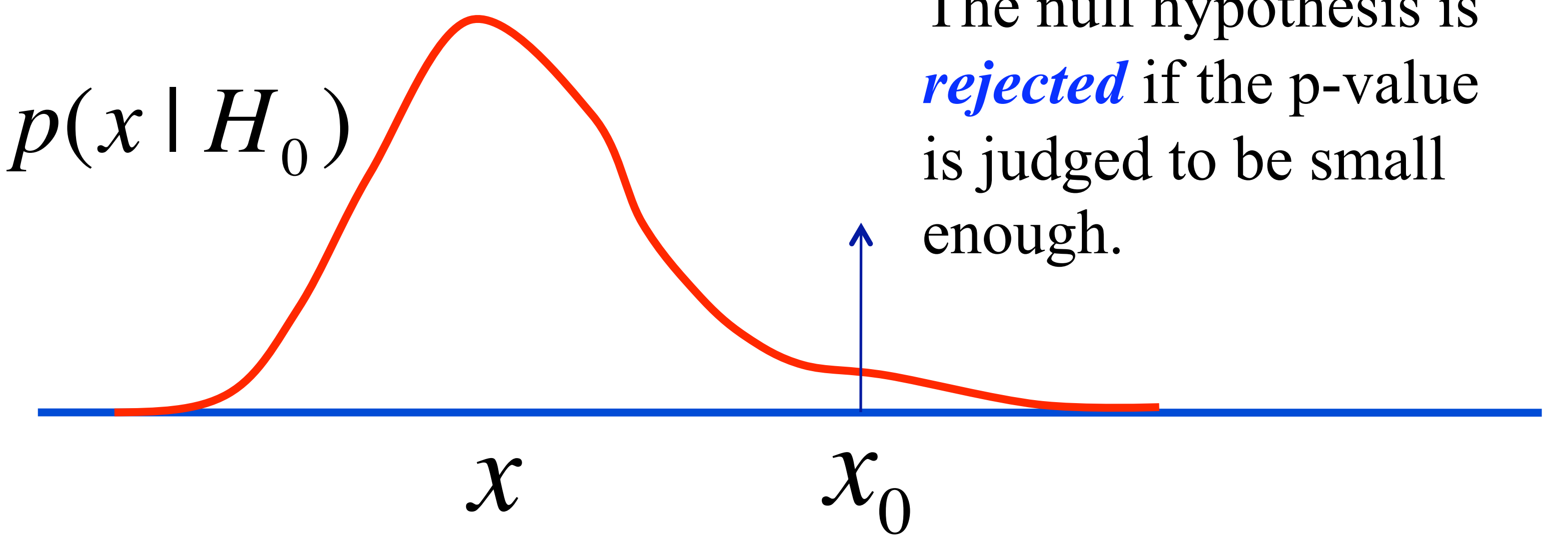


Exercise 3: Compute M_m.

Hypothesis Testing

P-Values

Null hypothesis (***H_0***): **background-only**

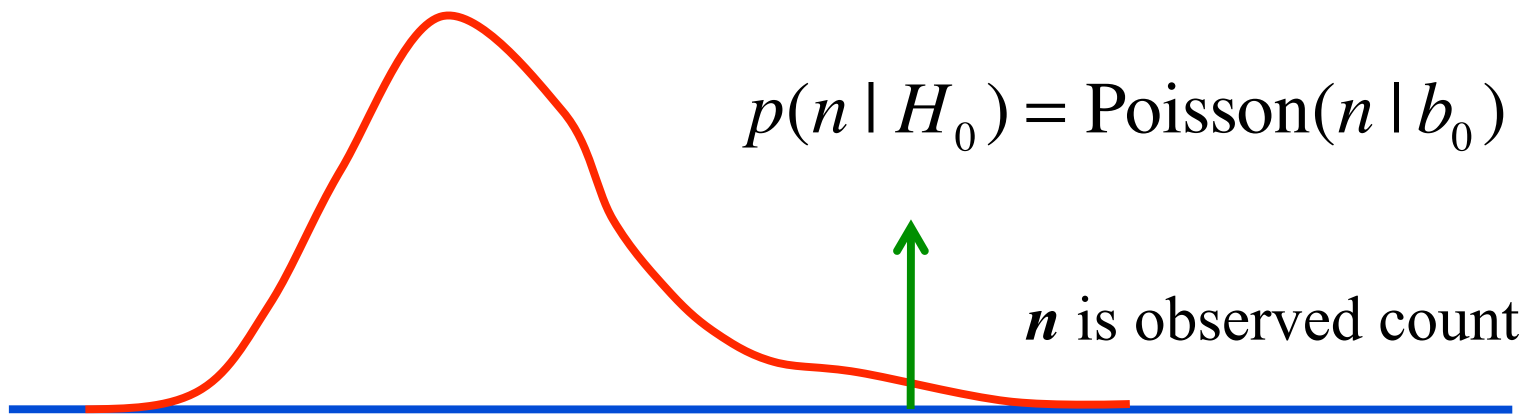


The null hypothesis is ***rejected*** if the p-value is judged to be small enough.

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

Example – Top Discovery p-value (a)

Background, $b_0 = 3.8$ events (*ignoring uncertainty*)



$$n = 17$$

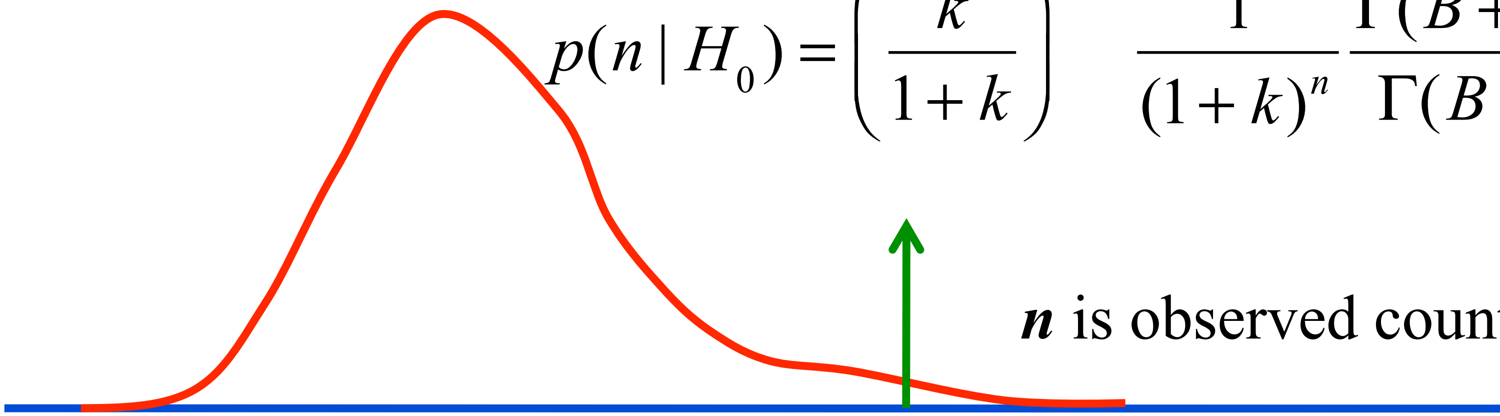
$$\text{p-value} = \sum_{n=17}^{\infty} \text{Poisson}(n \mid 3.8) = 5.7 \times 10^{-7}$$

This is equivalent to **4.9 σ**

Example – Top Discovery p-value (b)

Background, $b_0 = 3.8$ **± 0.6** events

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



n is observed count

$$n = 17$$

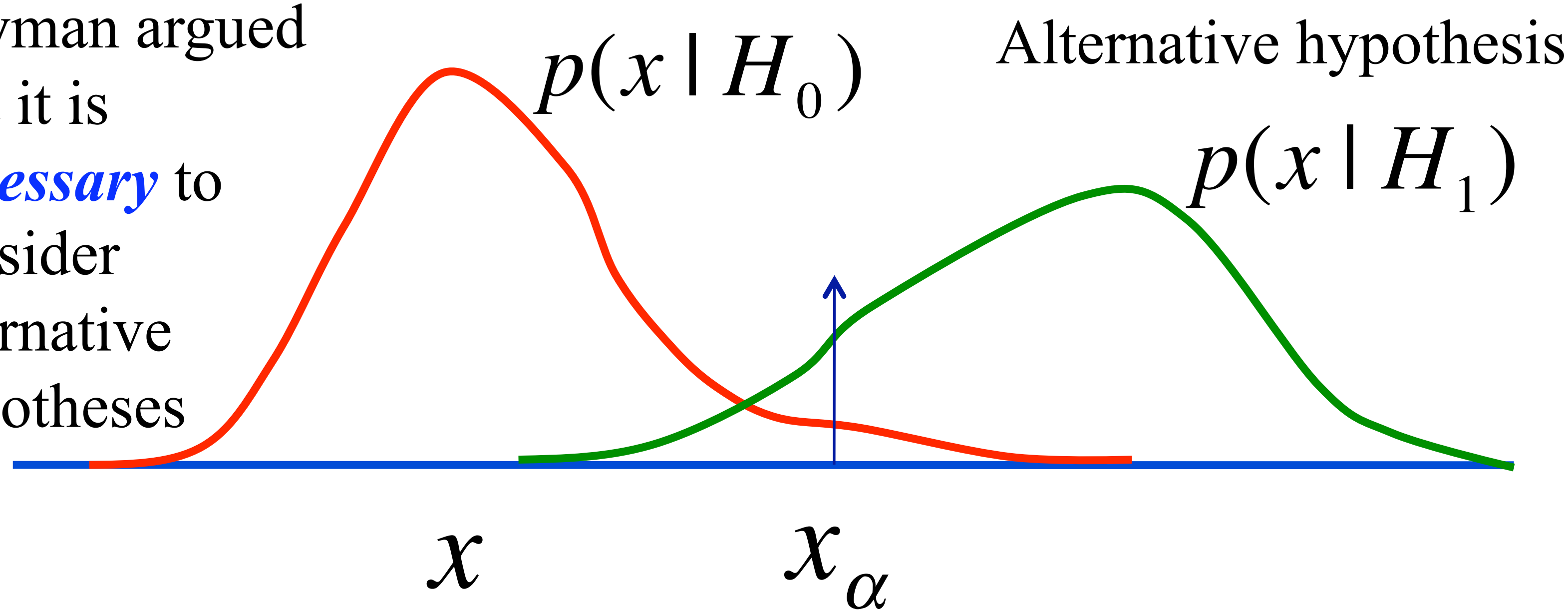
$$\text{p-value} = \sum_{n=17}^{\infty} p(n \mid H_0) = 5.4 \times 10^{-6}$$

This is equivalent to **4.4 σ**

The Neyman-Pearson Test

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

H_1

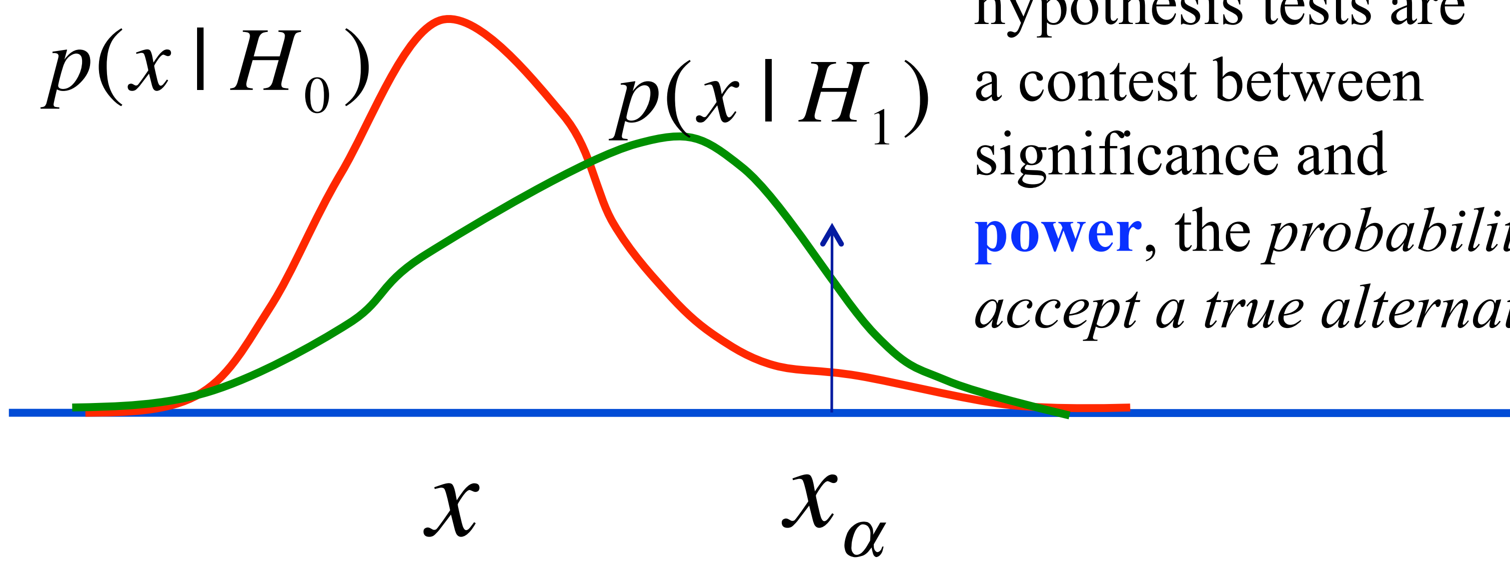


$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0)\,dx$$

significance of test

A ***fixed*** significance, the *probability to reject a true null*, is chosen before data are analyzed.

The Neyman-Pearson Test



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

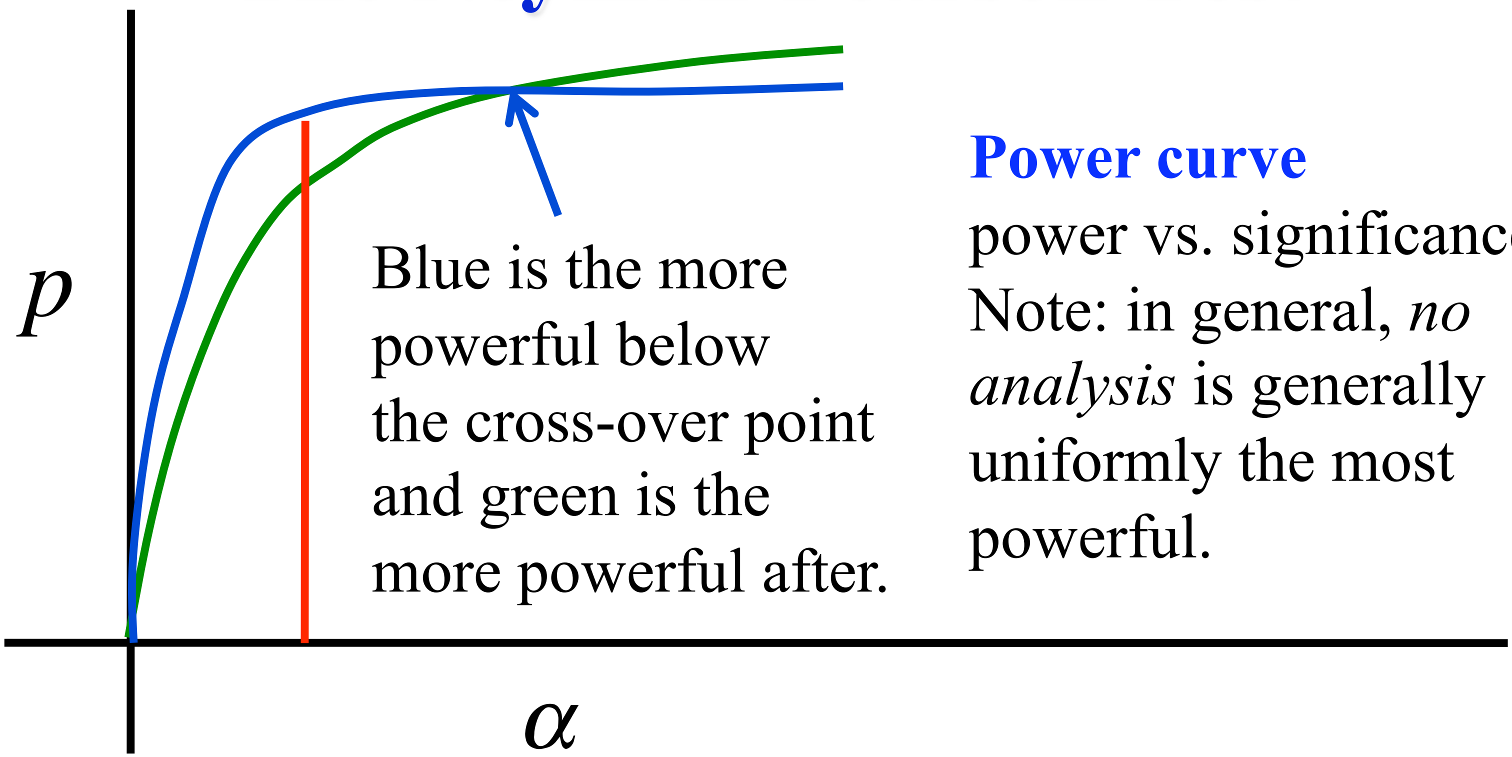
$$\alpha = \int_{x_\alpha}^{\infty} p(x \mid H_0) dx$$

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 generally uniformly the most powerful.

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

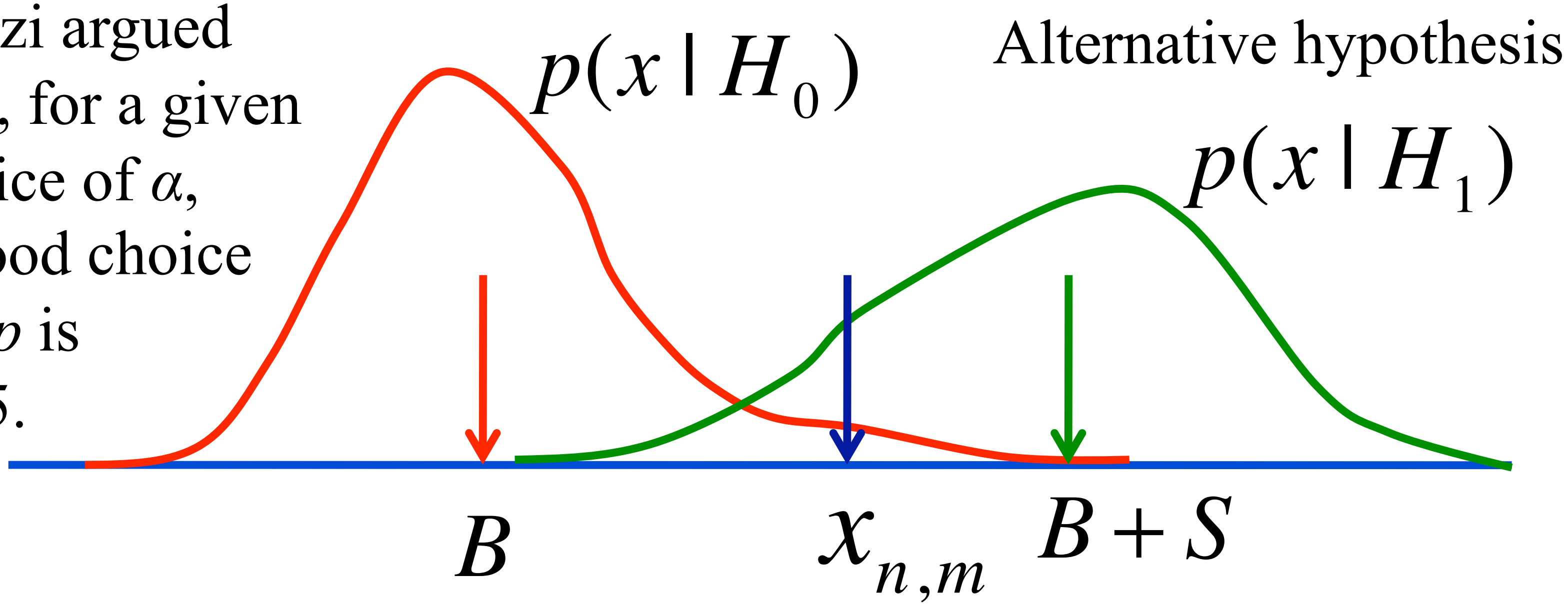
significance of test

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

power

Punzi's Test

Punzi argued that, for a given choice of α, a good choice for p is 0.95.



$$S = n\sqrt{B} + m\sqrt{B + S}$$

Exercise 4: Write as Q = S/√(B+a) and find a

Summary

- **Decision Theory**
  - The basic insight is that optimal decision making entails combining a ***loss function*** with a ***posterior density***. Since loss functions can differ, it is unsurprising that results can differ even when using the same data.

- **Hypothesis Tests**
  - The standard non-Bayesian approach is that of Neyman and Pearson, plus the calculation of p-values. Neyman argued (in agreement with Bayesians) that it is necessary to consider pairs of hypotheses.