Is there evidence for a peak in this data?
Is there evidence for a peak in this data?

“Observation of an Exotic S=+1 Baryon in Exclusive Photoproduction from the Deuteron”

“The statistical significance of the peak is 5.2 ± 0.6 σ”
Is there evidence for a peak in this data?

“Observation of an Exotic S=+1 Baryon in Exclusive Photoproduction from the Deuteron”
“The statistical significance of the peak is 5.2 ± 0.6 σ”

“A Bayesian analysis of pentaquark signals from CLAS data”
“The ln(RE) value for g2a (-0.408) indicates weak evidence in favour of the data model without a peak in the spectrum.”

p-values and Discovery

Louis Lyons
IC and Oxford
l.lyons@physics.ox.ac.uk

SLAC,
May 2011
Progress on Statistical Issues in Searches

http://www-conf.slac.stanford.edu/statisticalissues2012/
Discoveries

H0 or H0 v H1

p-values: For Gaussian, Poisson and multi-variate data

Goodness of Fit tests

Why 5σ?

Blind analyses

What is p good for?

Errors of 1st and 2nd kind

What a p-value is not

P(theory|data) ≠ P(data|theory)

THE paradox

Optimising for discovery and exclusion

Incorporating nuisance parameters
DISCOVERIES

“Recent” history:

Charm      SLAC, BNL      1974
Tau lepton SLAC          1977
Bottom     FNAL          1977
W,Z        CERN          1983
Top        FNAL          1995
{Pentaquarks ~Everywhere 2002 }
?           CERN LHC      2012?

? = Higgs, SUSY, q and l substructure, extra dimensions,
    free q/monopoles, technicolour, 4\textsuperscript{th} generation, black holes,…..

QUESTION: How to distinguish discoveries from fluctuations?
Penta-quarks?

Hypothesis testing: New particle or statistical fluctuation?
H0  or  H0 versus H1 ?

H0 = null hypothesis
e.g. Standard Model, with nothing new
H1 = specific New Physics  e.g. Higgs with $M_H = 120$ GeV

H0: “Goodness of Fit” e.g. $\chi^2$, p-values
H0 v H1: “Hypothesis Testing” e.g. $L$-ratio
Measures how much data favours one hypothesis wrt other

H0 v H1 likely to be more sensitive

or
Testing H0: Do we have an alternative in mind?

1) Data is number (of observed events)
   “H1” usually gives larger number
   (smaller number of events if looking for oscillations)

2) Data = distribution. Calculate $\chi^2$.
   Agreement between data and theory gives $\chi^2 \sim ndf$
   Any deviations give large $\chi^2$
   So test is independent of alternative?
   Counter-example: Cheating undergraduate

3) Data = number or distribution
   Use $\mathcal{L}$-ratio as test statistic for calculating p-value

4) H0 = Standard Model
p-values

Concept of pdf
Example: Gaussian

\( y = \text{probability density for measurement } x \)

\[ y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -0.5 \frac{(x-\mu)^2}{\sigma^2} \right\} \]

p-value: probability that \( x \geq x_0 \)

Gives probability of “extreme” values of data (in interesting direction)

<table>
<thead>
<tr>
<th>((x_0-\mu)/\sigma)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>16%</td>
<td>2.3%</td>
<td>0.13%</td>
<td>0.003%</td>
<td>0.3*10^{-6}</td>
</tr>
</tbody>
</table>

i.e. Small \( p = \text{unexpected} \)
p-values, contd

Assumes:
  Gaussian pdf (no long tails)
  Data is unbiased
  \( \sigma \) is correct
If so, Gaussian \( x \) \( \rightarrow \) uniform \( p \)-distribution

(Events at large \( x \) give small \( p \))
p-values for non-Gaussian distributions

e.g. Poisson counting experiment, bgd = b

\[ P(n) = e^{-b} \cdot \frac{b^n}{n!} \]

\{P = probability, not prob density\}

For \( n=7 \), \( p = \text{Prob( at least 7 events)} = P(7) + P(8) + P(9) + \ldots \ldots = 0.03 \)
Poisson p-values

n = integer, so p has discrete values
So p distribution cannot be uniform
Replace \( \text{Prob}\{p \leq p_0\} = p_0 \), for continuous p
by \( \text{Prob}\{p \leq p_0\} \leq p_0 \), for discrete p
(equality for possible \( p_0 \))

p-values often converted into equivalent Gaussian \( \sigma \)
e.g. \( 3 \times 10^{-7} \) is “5\( \sigma \)” (one-sided Gaussian tail)
Does NOT imply that pdf = Gaussian
Significance

Significance = \( S / \sqrt{B} \) ?

Potential Problems:

• Uncertainty in B
• Non-Gaussian behaviour of Poisson, especially in tail
• Number of bins in histogram, no. of other histograms [FDR]
• Choice of cuts (Blind analyses)
• Choice of bins (………………..)

For future experiments:

• Optimising \( S / \sqrt{B} \) could give S =0.1, B = 10^{-4}
Goodness of Fit Tests

Data = individual points, histogram, multi-dimensional, multi-channel

$\chi^2$ and number of degrees of freedom
$\Delta \chi^2$ (or $ln\mathcal{L}$-ratio): Looking for a peak
Unbinned $\mathcal{L}_{\text{max}}$?
Kolmogorov-Smirnov
Zech energy test
Combining p-values

Lots of different methods. Software available from:
http://www.ge.infn.it/statisticaltoolkit
$\chi^2$ with $\nu$ degrees of freedom?

1) $\nu = \text{data} - \text{free parameters}?$

Why asymptotic (apart from Poisson $\rightarrow$ Gaussian)?

a) Fit flatish histogram with

$$y = N \{ 1 + 10^{-6} \exp\{-0.5(x-x_0)^2\} \}$$

$x_0 = \text{free param}$

b) Neutrino oscillations: almost degenerate parameters

$$y \sim 1 - A \sin^2(1.27 \Delta m^2 L/E)$$

2 parameters

$$\longrightarrow 1 - A (1.27 \Delta m^2 L/E)^2$$

1 parameter

Small $\Delta m^2$
χ² with ν degrees of freedom?

2) Is difference in χ² distributed as χ²?

H0 is true.
Also fit with H1 with k extra params
e. g. Look for Gaussian peak on top of smooth background

\[ y = C(x) + A \exp\{-0.5 \frac{(x-x_0)/\sigma)^2}{ }\} \]

Is \( \chi^2_{H0} - \chi^2_{H1} \) distributed as \( \chi^2 \) with \( \nu = k = 3 \) ?

Relevant for assessing whether enhancement in data is just a statistical fluctuation, or something more interesting

N.B. Under H0 (\( y = C(x) \)) : \( A=0 \) (boundary of physical region)
\( x_0 \) and \( \sigma \) undefined
Is difference in $\chi^2$ distributed as $\chi^2$?

Demortier:
H0 = quadratic bgd
H1 = \ldots +
Gaussian of fixed width, variable location & ampl

Protassov, van Dyk, Connors, \ldots
H0 = continuum
(a) H1 = narrow emission line
(b) H1 = wider emission line
(c) H1 = absorption line

Nominal significance level = 5%
Is difference in $\chi^2$ distributed as $\chi^2$ ?, contd.

So need to determine the $\Delta\chi^2$ distribution by Monte Carlo

N.B.

1) Determining $\Delta\chi^2$ for hypothesis H1 when data is generated according to H0 is not trivial, because there will be lots of local minima

2) If we are interested in $5\sigma$ significance level, needs lots of MC simulations (or intelligent MC generation)
Unbinned $\mathcal{L}_{\text{max}}$ and Goodness of Fit?

Find params by maximising $\mathcal{L}$
So larger $\mathcal{L}$ better than smaller $\mathcal{L}$
So $\mathcal{L}_{\text{max}}$ gives Goodness of Fit ??

Monte Carlo distribution of unbinned $\mathcal{L}_{\text{max}}$

---

$\mathcal{L}_{\text{max}}$ vs Frequency

- Bad
- Good?
- Great?

$\mathcal{L}_{\text{max}} \rightarrow 22$
Not necessarily: $\mathcal{L}(\text{data}, \text{params})$

Contrast $\text{pdf}(\text{data}, \text{params})$

\[\begin{align*}
\uparrow & \quad \uparrow \\
\text{fixed} & \quad \text{vary}
\end{align*}\]

\[\begin{align*}
\uparrow & \quad \uparrow \\
\text{vary} & \quad \text{fixed}
\end{align*}\]

e.g. $p(t, \lambda) = \lambda \cdot \exp(-\lambda t)$

Max at $t = 0$

Max at $\lambda = 1/t$
Example 1: Exponential distribution

Fit exponential $\lambda$ to times $t_1, t_2, t_3 \ldots \ldots$ [Joel Heinrich, CDF 5639]

$$\mathcal{L} = \prod \lambda e^{-\lambda t}$$

$$\ln \mathcal{L}_{\text{max}} = -N(1 + \ln t_{\text{av}})$$

i.e. $\ln \mathcal{L}_{\text{max}}$ depends only on AVERAGE $t$, but is INDEPENDENT OF DISTRIBUTION OF $t$ (except for…….)

(Average $t$ is a sufficient statistic)

Variation of $\mathcal{L}_{\text{max}}$ in Monte Carlo is due to variations in samples’ average $t$, but NOT TO BETTER OR WORSE FIT

Same average $t \rightarrow$ same $\mathcal{L}_{\text{max}}$
Example 2

\[
\frac{dN}{d \cos \theta} = \frac{1 + \alpha \cos^2 \theta}{1 + \alpha / 3}
\]

\[
\mathcal{L} = \prod_{i} \frac{1 + \alpha \cos^2 \theta_j}{1 + \alpha / 3}
\]

pdf (and likelihood) depends only on \(\cos^2 \theta_i\)
Insensitive to sign of \(\cos \theta_i\)
So data can be in very bad agreement with expected distribution
  e.g. all data with \(\cos \theta < 0\), but \(\mathcal{L}_{\text{max}}\) does not know about it.

Example of general principle
Example 3

Fit to Gaussian with variable $\mu$, fixed $\sigma$

$$pdf = \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\}$$

$$\ln \mathcal{L}_{\text{max}} = N(-0.5 \ln 2\pi - \ln \sigma) - 0.5 \Sigma (x_i - x_{\text{av}})^2 / \sigma^2$$

constant $\uparrow$ ~variance(x) $\uparrow$

i.e. $\mathcal{L}_{\text{max}}$ depends only on variance(x),

which is not relevant for fitting $\mu$ \quad ($\mu_{\text{est}} = x_{\text{av}}$)

Smaller than expected variance(x) results in larger $\mathcal{L}_{\text{max}}$

Worse fit, larger $\mathcal{L}_{\text{max}}$  
Better fit, lower $\mathcal{L}_{\text{max}}$
Conclusion:

\( \mathcal{L} \) has sensible properties with respect to parameters

NOT with respect to data

\( \mathcal{L}_{\text{max}} \) within Monte Carlo peak is NECESSARY

not SUFFICIENT

(‘Necessary’ doesn’t mean that you have to do it!)
Goodness of Fit: 
Kolmogorov-Smirnov

Compares data and model cumulative plots
Uses largest discrepancy between dists.
Model can be analytic or MC sample

Uses individual data points
Not so sensitive to deviations in tails
(so variants of K-S exist)
Not readily extendible to more dimensions
Distribution-free conversion to p; depends on n
(but not when free parameters involved – needs MC)
Goodness of fit: ‘Energy’ test

Assign +ve charge to data ➡️ ; -ve charge to M.C. ⬅️

Calculate ‘electrostatic energy E’ of charges

If distributions agree, E ~ 0

If distributions don’t overlap, E is positive

Assess significance of magnitude of E by MC

N.B.
1) Works in many dimensions
2) Needs metric for each variable (make variances similar?)
3) \[ E \sim \sum q_i q_j f(\Delta r = |r_i - r_j|) , \quad f = 1/(\Delta r + \varepsilon) \text{ or } -\ln(\Delta r + \varepsilon) \]
   Performance insensitive to choice of small \( \varepsilon \)

See Aslan and Zech’s paper at:
http://www.ippp.dur.ac.uk/Workshops/02/statistics/program.shtml
Combining different p-values

Several results quote p-values for same effect: $p_1$, $p_2$, $p_3$……
e.g. 0.9, 0.001, 0.3 ……..
What is combined significance? Not just $p_1 \times p_2 \times p_3$…….
If 10 expts each have $p \sim 0.5$, product $\sim 0.001$ and is clearly NOT correct combined $p$

$$S = z^{*} \sum_{j=0}^{n-1} (-\ln z)^j / j!$$
= $z = p_1 p_2 p_3$…….
(e.g. For 2 measurements, $S = z^* (1 - \ln z) \geq z$)

Slight problem: Formula is not associative

Combining $\{p_1$ and $p_2\}$, and then $p_3$ gives different answer
from $\{p_3$ and $p_2\}$, and then $p_1 \}$, or all together
Due to different options for “more extreme than $x_1, x_2, x_3$".
Combining different p-values

Conventional:
Are set of p-values consistent with H0?

SLEUTH:
How significant is smallest p?

\[ 1-S = (1-p_{\text{smallest}})^n \]

\[ \begin{align*}
  p_1 &= 0.01 & p_1 &= 10^{-4} \\
  p_2 &= 0.01 & p_2 &= 10^{-4} & p_2 &= 1 \\
\end{align*} \]

Combined S

<table>
<thead>
<tr>
<th></th>
<th>Conventiona</th>
<th>SLEUTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0 10^{-3}</td>
<td>5.6 10^{-2}</td>
</tr>
<tr>
<td></td>
<td>2.0 10^{-2}</td>
<td>2.0 10^{-2}</td>
</tr>
</tbody>
</table>
What is $p$ good for?

Used to test whether data is consistent with $H_0$

Reject $H_0$ if $p$ is small: $p \leq \alpha$ (How small?)

Sometimes make wrong decision:

Reject $H_0$ when $H_0$ is true: Error of 1\textsuperscript{st} kind

Should happen at rate $\alpha$

OR

Fail to reject $H_0$ when something else ($H_1,H_2,…$) is true: Error of 2\textsuperscript{nd} kind

Rate at which this happens depends on...........
Errors of 2\textsuperscript{nd} kind: How often?

e.g.1. Does data line on straight line?  
Calculate $\chi^2$  
Reject if $\chi^2 \geq 20$

Error of 1\textsuperscript{st} kind: $\chi^2 \geq 20$  Reject H0 when true

Error of 2\textsuperscript{nd} kind: $\chi^2 \leq 20$  Accept H0 when in fact quadratic or...

How often depends on:

Size of quadratic term  
Magnitude of errors on data, spread in x-values,.......  
How frequently quadratic term is present
Errors of 2\textsuperscript{nd} kind: How often?

e.g. 2. Particle identification (TOF, dE/dx, Čerenkov,…….)
Particles are $\pi$ or $\mu$
Extract p-value for $H_0 = \pi$ from PID information

\begin{align*}
\text{p} & \rightarrow \\
0 & \quad 1
\end{align*}

$\pi$ and $\mu$ have similar masses

Of particles that have $p \sim 1\%$ (‘reject $H_0$’), fraction that are $\pi$ is
\begin{enumerate}
\item[a)] $\sim$ half, for equal mixture of $\pi$ and $\mu$
\item[b)] almost all, for “pure” $\pi$ beam
\item[c)] very few, for “pure” $\mu$ beam
\end{enumerate}
What is P good for?

Selecting sample of wanted events
e.g. kinematic fit to select $t \bar{t}$ events

$t \rightarrow bW, \ b \rightarrow jj, \ W \rightarrow \mu \nu \quad t \rightarrow bW, \ b \rightarrow jj, \ W \rightarrow jj$

Convert $\chi^2$ from kinematic fit to p-value

Choose cut on $\chi^2$ to select $t \bar{t}$ events

Error of 1$^{st}$ kind: Loss of efficiency for $t \bar{t}$ events

Error of 2$^{nd}$ kind: Background from other processes

Loose cut (large $\chi^2_{\text{max}}$, small $p_{\text{min}}$): Good efficiency, larger bgd

Tight cut (small $\chi^2_{\text{max}}$, larger $p_{\text{min}}$): Lower efficiency, small bgd

Choose cut to optimise analysis:

More signal events: Reduced statistical error

More background: Larger systematic error
p-value is not ........

Does **NOT** measure \( \text{Prob(H0 is true)} \)
i.e. It is **NOT** \( P(H0|\text{data}) \)
It is \( P(\text{data}|H0) \)

N.B. \( P(H0|\text{data}) \neq P(\text{data}|H0) \)
\[ P(\text{theory}|\text{data}) \neq P(\text{data}|\text{theory}) \]
\[ P(\text{rainy ! Dec 25th}) \neq P(\text{Dec 25th ! rainy}) \]

“Of all results with \( p \leq 5\% \), half will turn out to be wrong”

N.B. Nothing wrong with this statement

e.g. 1000 tests of energy conservation

~50 should have \( p \leq 5\% \), and so reject \( H0 = \text{energy conservation} \)

Of these 50 results, **all are likely to be “wrong”**
Why $5\sigma$?

- Past experience with $3\sigma$, $4\sigma$, … signals
- Look elsewhere effect:
  - Different cuts to produce data
  - Different bins (and binning) of this histogram
  - Different distributions Collaboration did/could look at
  - Defined in SLEUTH

- Bayesian priors:
  \[
  \frac{P(H_0|\text{data})}{P(H_1|\text{data})} = \frac{P(\text{data}|H_0) \cdot P(H_0)}{P(\text{data}|H_1) \cdot P(H_1)}
  \]

Bayes posteriors \quad Likelihoods \quad Priors

Prior for \{H_0 = \text{S.M.}\} \gggg Prior for \{H_1 = \text{New Physics}\}
Why $5\sigma$?

BEWARE of tails,
especially for nuisance parameters

Same criterion for all searches?
  Single top production
  Higgs
  Highly speculative particle
  Energy non-conservation
BLIND ANALYSES

Why blind analysis?
Selections, corrections, method

Methods of blinding

Add random number to result *
Study procedure with simulation only
Look at only first fraction of data
Keep the signal box closed
Keep MC parameters hidden
Keep unknown fraction visible for each bin

After analysis is unblinded, ………

* Luis Alvarez suggestion re “discovery” of free quarks
PARADOX

Histogram with 100 bins
Fit 1 parameter

$S_{\text{min}}$ : $\chi^2$ with NDF = 99  (Expected $\chi^2 = 99 \pm 14$)

For our data, $S_{\text{min}}(p_0) = 90$
Is $p_1$ acceptable if $S(p_1) = 115$?

1) YES.    Very acceptable $\chi^2$ probability
2) NO.    $\sigma_p$ from $S(p_0 + \sigma_p) = S_{\text{min}} + 1 = 91$
But $S(p_1) - S(p_0) = 25$
So $p_1$ is $5\sigma$ away from best value
Is this value of $\beta$ acceptable?

$NDF = 99$
ANOTHER EXAMPLE

\[ X = -1 \rightarrow +1 \]

\[ H_1: 1 + \alpha x \quad \alpha = 0.05 \]

\[ H_2: 1 + b \cos(\pi x) \quad b = 0.05 \]

Generate events according to \( H_1 \) (+ star function)

Try fitting according to \( H_1 \) or \( H_2 \)

\( \chi^2 \), \( \chi_{\alpha}^2 \)

Look at dist of \( \chi^2_1 \) As expected for \( NDF=100 \)

\( \chi^2_2 \) But bigger Many

\( \chi^2_2 - \chi^2_1 \) Decision based in \( \Delta \chi^2 \)

has much better power

Repeat for events generated according to \( H_2 \)

Look at dist of \( \chi^2 \)

\( \chi^2_2 - \chi^2_1 \)

\[ \# \text{ of 691 have } \chi^2_2 < 130 \]
Distinguishing 2 hypotheses on basis of $\Delta \chi^2$  

(See simulations)  

```
H_2 = 1 + 0.05 \cos(\pi x)  
H_1 = 1 + 0.05 x  
```

Comparing data with different hypotheses


- Supernova Cosmology Project
- High-Z Supernova Search
- Calan/Tololo Supernova Survey

Accelerating Universe
Decelerating Universe

With vacuum energy
Without vacuum energy
Mass density
Empty
Choosing between 2 hypotheses

Possible methods:

\( \Delta \chi^2 \)
\( p \)-value of statistic \( \rightarrow \)
\( \ln \mathcal{L} \)-ratio

Bayesian:

Posterior odds
Bayes factor
Bayes information criterion (BIC)
Akaike \( \ldots \ldots \) (AIC)

Minimise “cost”
1) No sensitivity

2) Maybe

3) Easy separation

**Procedure:** Choose $\alpha$ (e.g. 95%, 3$\sigma$, 5$\sigma$?) and CL for $\beta$ (e.g. 95%)  

Given $b$, $\alpha$ determines $n_{\text{crit}}$  

$s$ defines $\beta$. For $s > s_{\text{min}}$, separation of curves $\rightarrow$ discovery or excln  

$s_{\text{min}}$ = Punzi measure of sensitivity  

For $s \geq s_{\text{min}}$, 95% chance of 5$\sigma$ discovery  

Optimise cuts for smallest $s_{\text{min}}$

Now data:  

- If $n_{\text{obs}} \geq n_{\text{crit}}$, discovery at level $\alpha$  
- If $n_{\text{obs}} < n_{\text{crit}}$, no discovery. If $\beta_{\text{obs}} < 1 - \text{CL}$, exclude H1
p-values or \( \text{Likelihood ratio?} \)

\[ \mathcal{L} = \text{height of curve} \]
\[ p = \text{tail area} \]

Different for distributions that

a) have dip in middle

b) are flat over range

Likelihood ratio favoured by Neyman-Pearson lemma (for simple H0, H1)

Use \( \mathcal{L} \)-ratio as statistic, and use p-values for its distributions for H0 and H1

Think of this as either

i) p-value method, with \( \mathcal{L} \)-ratio as statistic; or

ii) \( \mathcal{L} \)-ratio method, with p-values as method to assess value of \( \mathcal{L} \)-ratio
Bayes’ methods for H0 versus H1

Bayes’ Th: \[ P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \]

\[
\begin{array}{c}
\frac{P(H0|data)}{P(H1|data)} \quad = \quad \frac{P(data|H0) \times \text{Prior}(H0)}{P(data|H1) \times \text{Prior}(H1)}
\end{array}
\]

Posterior odds ratio \quad \longleftarrow \quad \text{Likelihood ratio} \quad \longleftarrow \quad \text{Priors}

N.B. Frequentists object to this
(and some Bayesians object to p-values)
Bayes’ methods for H0 versus H1

\[
\frac{P(H0|\text{data})}{P(H1|\text{data})} = \frac{P(\text{data}|H0) \times \text{Prior}(H0)}{P(\text{data}|H1) \times \text{Prior}(H1)}
\]

Posterior odds = Likelihood ratio \times Priors

e.g. data is mass histogram
   H0 = smooth background
   H1 = .......................... + peak

1) Profile likelihood ratio also used but not quite Bayesian
   (Profile = \textit{maximise} wrt parameters.
   Contrast Bayes which \textit{integrates} wrt parameters)

2) Posterior odds

3) Bayes factor = Posterior odds/Prior ratio
   (= Likelihood ratio in simple case)

4) In presence of parameters, need to integrate them out, using priors.
   e.g. peak’s mass, width, amplitude
   Result becomes dependent on prior, and more so than in parameter determination.

5) Bayes information criterion (BIC) tries to avoid priors by
   \[
   \text{BIC} = -2 \times \ln{\text{L ratio}} + k \times \ln{n} \\
   \text{where } k = \text{ free params; } n = \text{no. of obs}
   \]

6) Akaike information criterion (AIC) tries to avoid priors by
   \[
   \text{AIC} = -2 \times \ln{\text{L ratio}} + 2k
   \]

etc etc etc
Why $p \neq$ Bayes factor

Measure different things:
$p_0$ refers just to $H_0$; $B_{01}$ compares $H_0$ and $H_1$

Depends on amount of data:
e.g. Poisson counting expt little data:
   For $H_0$, $\mu_0 = 1.0$. For $H_1$, $\mu_1 = 10.0$
   Observe $n = 10$ $p_0 \sim 10^{-7}$ $B_{01} \sim 10^{-5}$
Now with 100 times as much data, $\mu_0 = 100.0$ $\mu_1 = 1000.0$
   Observe $n = 160$ $p_0 \sim 10^{-7}$ $B_{01} \sim 10^{+14}$
$CL_S = \frac{p_1}{1-p_0}$
$p_0$ versus $p_1$ plots
Optimisation for Discovery and Exclusion

Giovanni Punzi, PHYSTAT2003:
“Sensitivity for searches for new signals and its optimisation”

Simplest situation: Poisson counting experiment,
\[ \text{Bgd} = b, \text{Possible signal} = s, \ n_{\text{obs}} \text{ counts} \]
(More complex: Multivariate data, \( \ln \mathcal{L} \)-ratio)

Traditional sensitivity:
  
  Median limit when \( s=0 \)
  
  Median \( \sigma \) when \( s \neq 0 \) (averaged over \( s \)?)

Punzi criticism: Not most useful criteria
  
  Separate optimisations
Procedure: Choose $\alpha$ (e.g. 95%, $3\sigma$, $5\sigma$ ?) and CL for $\beta$ (e.g. 95%)

Given $b$, $\alpha$ determines $n_{\text{crit}}$

s defines $\beta$. For $s > s_{\text{min}}$, separation of curves $\rightarrow$ discovery or excln

$s_{\text{min}} =$ Punzi measure of sensitivity  For $s \geq s_{\text{min}}$, 95% chance of $5\sigma$ discovery

Optimise cuts for smallest $s_{\text{min}}$

Now data:  

If $n_{\text{obs}} \geq n_{\text{crit}}$, discovery at level $\alpha$

If $n_{\text{obs}} < n_{\text{crit}}$, no discovery. If $\beta_{\text{obs}} < 1 - \text{CL}$, exclude $H1$
1) No sensitivity
Data almost always falls in peak
$\beta$ as large as 5%, so 5% chance of H1 exclusion even when no sensitivity. ($CL_s$)

2) Maybe
If data fall above $n_{crit}$, discovery
Otherwise, and $n_{obs} \rightarrow \beta_{obs}$ small, exclude H1

(95% exclusion is easier than $5\sigma$ discovery)

But these may not happen $\rightarrow$ no decision

3) Easy separation
Always gives discovery or exclusion (or both!)

<table>
<thead>
<tr>
<th>Disc</th>
<th>Excl</th>
<th>1)</th>
<th>2)</th>
<th>3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>![ ]</td>
<td>![ ]</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>![ ]</td>
<td>![ ]</td>
<td>![ ]</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>![ ]</td>
<td>![ ]</td>
<td>![ ]</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>![ ]</td>
<td>![ ]</td>
<td>![ ]</td>
</tr>
</tbody>
</table>
Incorporating systematics in p-values

Simplest version:

Observe $n$ events

Poisson expectation for background only is $b \pm \sigma_b$

$\sigma_b$ may come from:

- acceptance problems
- jet energy scale
- detector alignment
- limited MC or data statistics for backgrounds
- theoretical uncertainties
Luc Demortier, “p-values: What they are and how we use them”, CDF memo June 2006
http://www-cdfd.fnal.gov/~luc/statistics/cdf0000.ps

Includes discussion of several ways of incorporating nuisance parameters

Desiderata:

- Uniformity of p-value (averaged over $\nu$, or for each $\nu$?)
- p-value increases as $\sigma_\nu$ increases

Generality

Maintains power for discovery
Ways to incorporate nuisance params in p-values

- Supremum: Maximise p over all \( \nu \). Very conservative
- Conditioning: Good, if applicable
- Prior Predictive Box: Most common in HEP
  \[ p = \int p(\nu) \pi(\nu) \, d\nu \]
- Posterior predictive: Averages p over posterior
- Plug-in: Uses best estimate of \( \nu \), without error
- \( L \)-ratio
- Confidence interval: Berger and Boos.
  \[ p = \text{Sup}\{p(\nu)\} + \beta, \text{ where } 1-\beta \text{ Conf Int for } \nu \]
- Generalised frequentist: Generalised test statistic

Performances compared by Demortier
Summary

• $P(H0|\text{data}) \neq P(\text{data}|H0)$
• p-value is NOT probability of hypothesis, given data
• Many different Goodness of Fit tests
  Most need MC for statistic $\rightarrow$ p-value
• For comparing hypotheses, $\Delta\chi^2$ is better than $\chi^2_1$ and $\chi^2_2$
• Blind analysis avoids personal choice issues
• Different definitions of sensitivity
• Worry about systematics

PHYSTAT-LHC Workshop at CERN, June 2007
“Statistical issues for LHC Physics Analyses”