Statistical Issues in Searches for New Physics

Louis Lyons
Imperial College, London
and
Oxford

LHCP2014, June 2014
Columbia, NY

Version of 30/5/14

Theme:

Using data to make judgements about H1 (New Physics) versus
H0 (S.M. with nothing new)

Why?

HEP is expensive and time-consuming
so
Worth investing effort in statistical analysis
→ better information from data

Topics:

- Why 5σ for discovery?
- P(A|B) ≠ P(B|A)
- Meaning of p-values
- Wilks' Theorem
- LEE = Look Elsewhere Effect
- Background Systematics
- Coverage
- p_0 v p_1 plots

(N.B. Several of these topics have no unique solutions from Statisticians)

Conclusions

Why 5σ for Discovery?

Statisticians ridicule our belief in extreme tails (esp. for systematics)

Our reasons:

1) Past history (Many 3σ and 4σ effects have gone away)

2) LEE (see later)

3) Worries about underestimated systematics

4) Subconscious Bayes calculation

$$\frac{p(H_1|x)}{p(H_0|x)} = \frac{p(x|H_1)}{p(x|H_0)} * \frac{\pi(H_1)}{\pi(H_0)}$$

Posterior prob — Likelihood ratio — Priors

"Extraordinary claims require extraordinary evidence"

N.B. Points 2), 3) and 4) are experiment-dependent

Alternative suggestion:

L.L. "Discovering the significance of 5σ" http://arxiv.org/abs/1310.1284

How many σ's for discovery?

SEARCH	SURPRISE	IMPACT	LEE	SYSTEMATICS	No. σ
Higgs search	Medium	Very high	M	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/Low	Medium	Δm	No	4
Neutrino osc	Medium	High	$\sin^2 2\vartheta$, Δm^2	No	4
$B_s \rightarrow \mu\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/V. high	M, decay mode	Medium	7
$(g\text{-}2)_\mu$ anom	Yes	High	No	Yes	4
H spin ≠ 0	Yes	High	No	Medium	5
4^{th} gen q, l, ν	Yes	High	M, mode	No	6
Dark energy	Yes	Very high	Strength	Yes	5
Grav Waves	No	High	Enormous	Yes	8

Suggestions to provoke discussion, rather than `delivered on Mt. Sinai'

Bob Cousins: "2 independent expts each with 3.5σ better than one expt with 5σ"

$P(A|B) \neq P(B|A)$

Remind Lab or University media contact person that:

Prob[data, given H0] is very small

does not imply that

Prob[H0, given data] is also very small.

e.g. Prob{data | speed of $\nu \leq c$}= very small

does not imply

Prob{speed of $\nu \leq c$ | data} = very small

or Prob{speed of $\nu > c$ | data} ~ 1

Everyday example p(pregnant|female) ~ 3%

$P(A|B) \neq P(B|A)$

Remind Lab or University media contact person that:

Prob[data, given H0] is very small

does not imply that

Prob[H0, given data] is also very small.

e.g. Prob{data | speed of $\nu \leq c$}= very small

does not imply

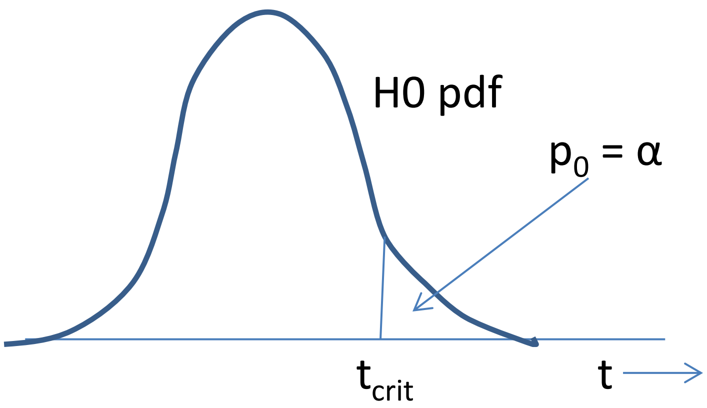
Prob{speed of $\nu \leq c$ | data} = very small

or Prob{speed of $\nu > c$ | data} ~ 1

Everyday example p(pregnant|female) ~ 3%

p(female|pregnant) >> 3%

What p-values are (and are not)



Reject H0 if $t > t_{crit}$ ($p < \alpha$)

p-value = prob that $t \geq t_{obs}$

Small p → data and theory have poor compatibility

Small p-value does **NOT** automatically imply that theory is unlikely

Bayes prob(Theory|data) related to prob(data|Theory) = Likelihood

by Bayes Th, including Bayesian prior

p-values are misunderstood. e.g. Anti-HEP jibe:

"Particle Physicists don't know what they are doing, because half their $p < 0.05$ exclusions turn out to be wrong"

Demonstrates lack of understanding of p-values

[All results rejecting energy conservation with $p < \alpha = .05$ cut will turn out to be 'wrong']

Wilks' Theorem

Data = some distribution e.g. mass histogram

For H0 and H1, calculate best fit weighted sum of squares S_0 and S_1

Examples: 1) H0 = polynomial of degree 3

H1 = polynomial of degree 5

2) H0 = background only

H1 = bgd + peak with free M_0 and cross-section

3) H0 = normal neutrino hierarchy

H1 = inverted hierarchy

If H0 true, S_0 distributed as χ^2 with ndf = ν_0

If H1 true, S_1 distributed as χ^2 with ndf = ν_1

If H0 true, what is distribution of $\Delta S = S_0 - S_1$? Is it χ^2?

Wilks' Theorem: ΔS distributed as χ^2 with ndf = $\nu_1 - \nu_0$ provided:

a) H0 is true
b) H0 and H1 are nested
c) Params for H1 → H0 are well defined, and not on boundary
d) Data is asymptotic

Wilks' Theorem, contd

Examples: Does Wilks' Th apply?

1) H0 = polynomial of degree 3
   H1 = polynomial of degree 5

YES: ΔS distributed as χ^2 with ndf = (d-4) – (d-6) = 2

2) H0 = background only
   H1 = bgd + peak with free M_0 and cross-section

NO: H0 and H1 nested, but M_0 undefined when H1→ H0. $\Delta S \neq \chi^2$

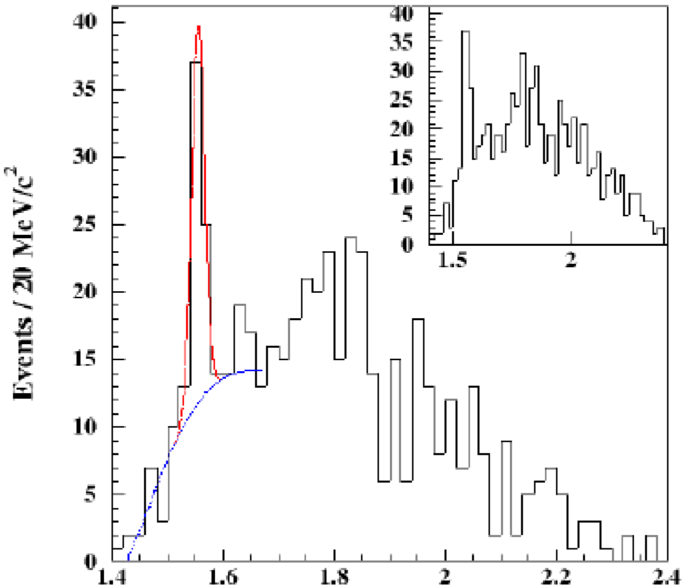
3) H0 = normal neutrino hierarchy
   H1 = inverted hierarchy

NO: Not nested. $\Delta S \neq \chi^2$

N.B. 1: Even when W. Th. does not apply, it does not mean that ΔS is irrelevant, but you cannot use W. Th. for its expected distribution.

N.B. 2: For large ndf, better to use ΔS, rather than S_1 and S_0 separately

Look Elsewhere Effect (LEE)



Prob of bgd fluctuation at that place = local p-value
Prob of bgd fluctuation 'anywhere' = global p-value
Global p > Local p
Where is `anywhere'?

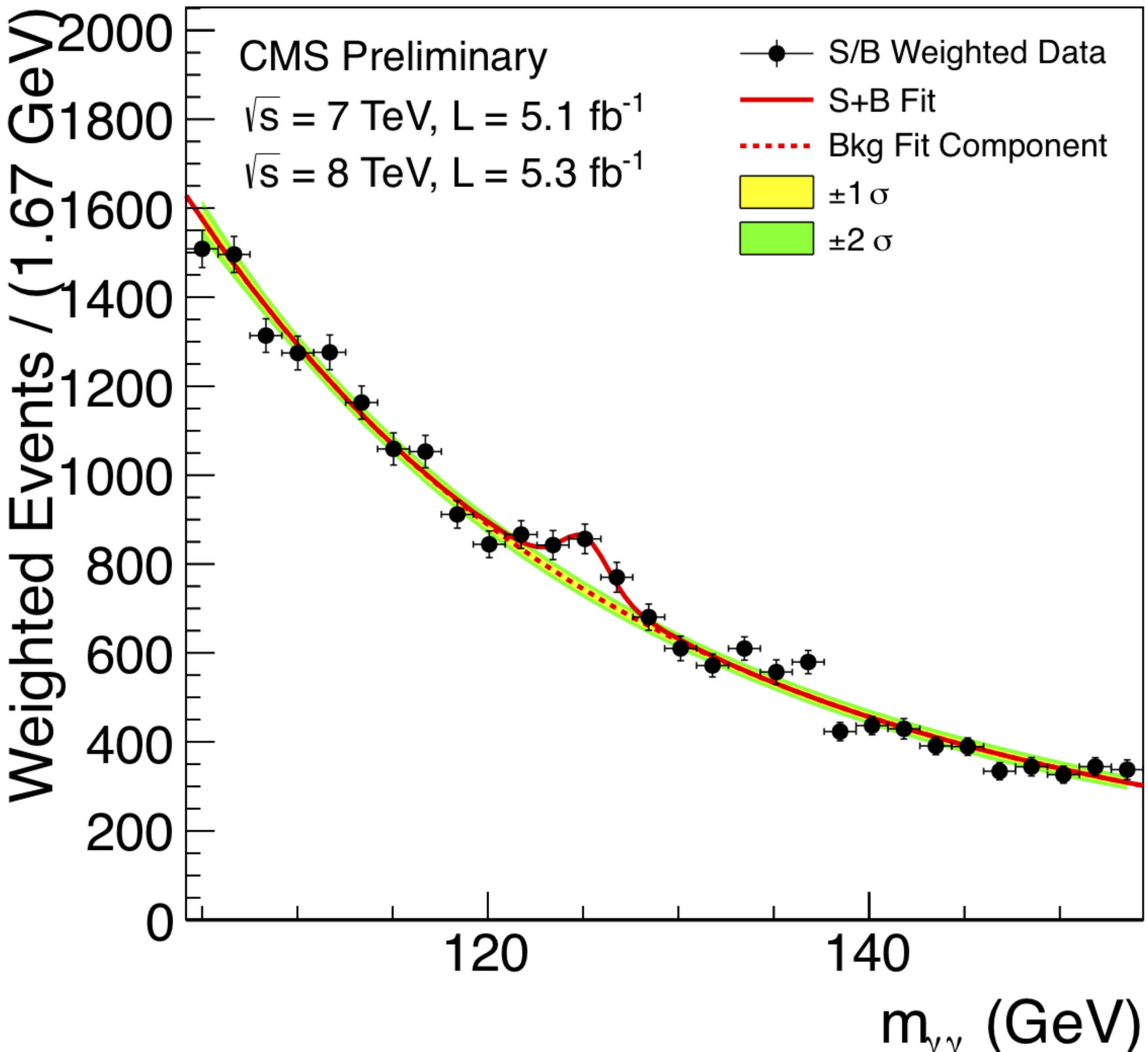
a) Any location in this histogram in sensible range
b) Any location in this histogram
c) Also in histogram produced with different cuts, binning, etc.
d) Also in other plausible histograms for this analysis
e) Also in other searches in this PHYSICS group (e.g. SUSY at CMS)
f) In any search in this experiment (e.g. CMS)
g) In all CERN expts (e.g. LHC expts + NA62 + OPERA + ASACUSA + ….)
h) In all HEP expts

etc.

d) relevant for graduate student doing analysis
f) relevant for experiment's Spokesperson

INFORMAL CONSENSUS:
Quote local p, and global p according to a) above.
Explain which global p

Background systematics



Background systematics, contd

Signif from comparing χ^2's for H0 (bgd only) and for H1 (bgd + signal)

Typically, bgd = functional form f_a with free params

e.g. 4^{th} order polynomial

Uncertainties in params included in signif calculation

But what if functional form is different ? e.g. f_b

Typical approach:

If f_b best fit is bad, not relevant for systematics

If f_b best fit is ~comparable to f_a fit, include contribution to systematics

But what is '~comparable'?

Other approaches:

Profile likelihood over different bgd parametric forms

Background subtraction

sPlots

Non-parametric background

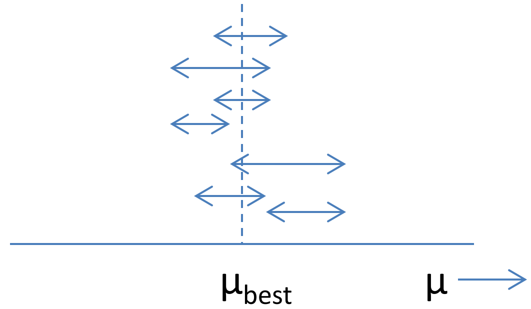
Bayes

etc

No common consensus yet among experiments on best approach

{Spectra with multiple peaks are more difficult}

Coverage



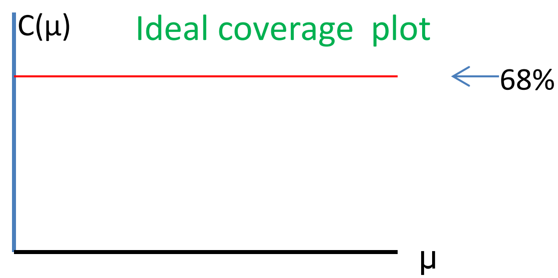
* What it is:

For given statistical method applied to many sets of data to extract confidence intervals for param μ, coverage C is fraction of ranges that contain true value of param.   Can vary with μ

* Does not apply to **your** data:

It is a property of the **statistical method** used

It is **NOT** a probability statement about whether μ_{true} lies in your confidence range for μ



* Coverage plot for Poisson counting expt

Observe n counts

Estimate μ_{best} from maximum of likelihood

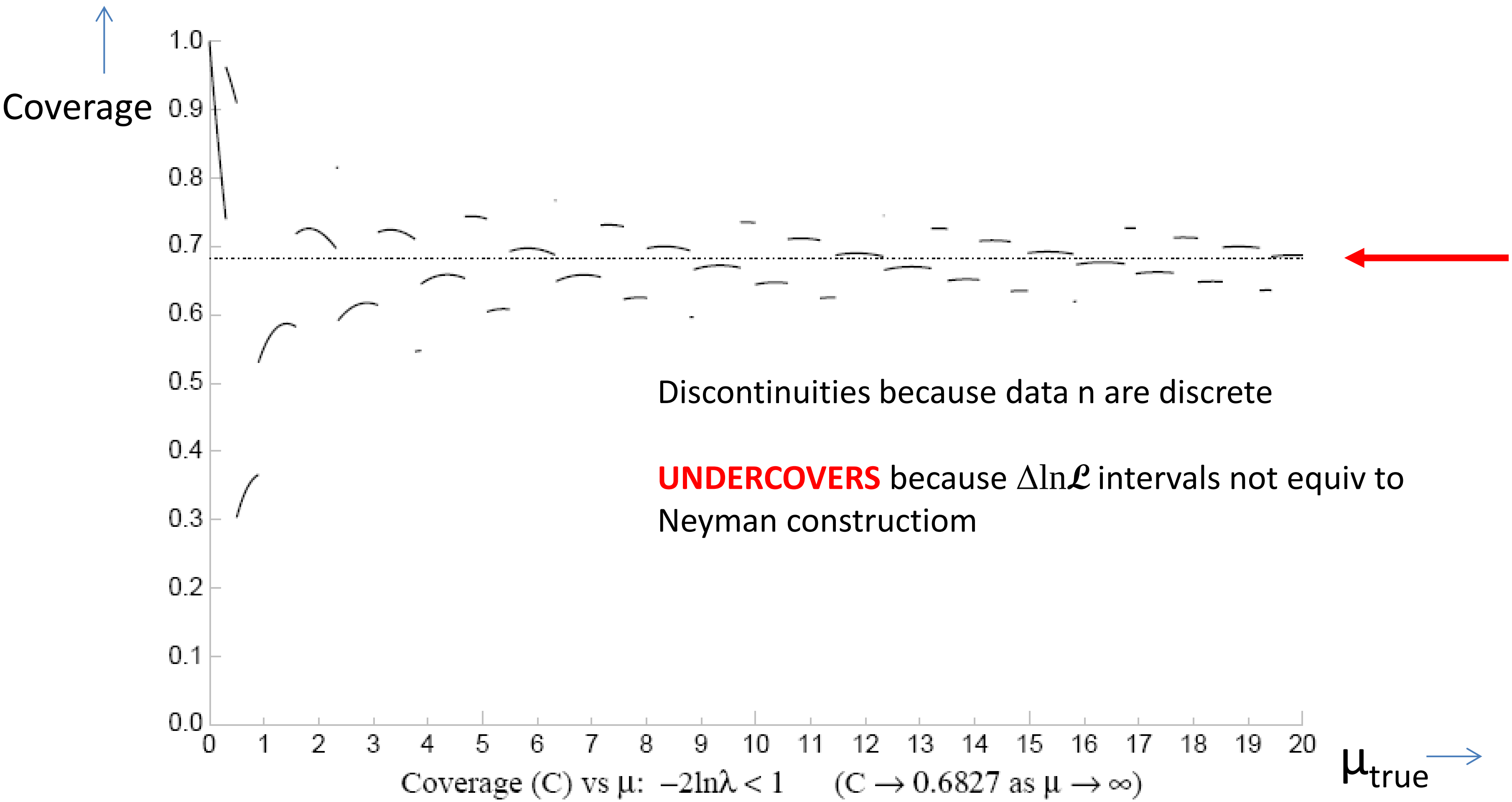
$L(\mu) = e^{-\mu} \mu^n/n!$ and range of μ from $\ln\{L(\mu_{best})/L(\mu)\} < 0.5$

For each μ_{true} calculate coverage $C(\mu_{true})$, and compare with nominal 68%

Coverage : $\Delta\ln\mathcal{L}$ intervals for μ

$P(n,\mu) = e^{-\mu}\mu^n/n!$ (Joel Heinrich CDF note 6438)

$-2\ln\lambda < 1$ $\lambda = p(n,\mu)/p(n,\mu_{best})$



Discontinuities because data n are discrete

UNDERCOVERS because $\Delta\ln\mathcal{L}$ intervals not equiv to Neyman constructiom

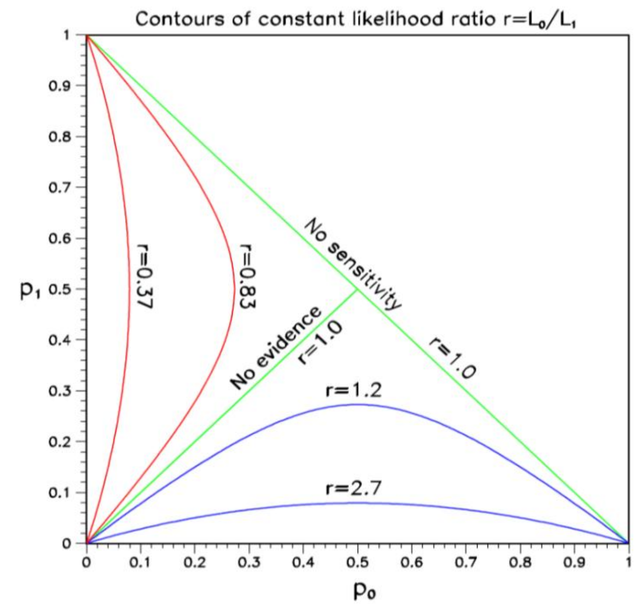
p_0 v p_1 plots

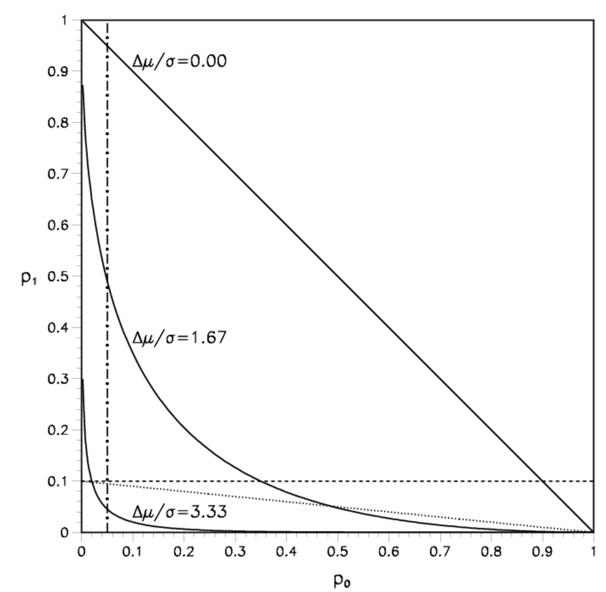
Preprint by Luc Demortier and LL, "Testing Hypotheses in Particle Physics: Plots of p_0 versus p_1"
.

For hypotheses H0 and H1, p_0 and p_1 are the tail probabilities for data statistic t

Provide insights on:

- CLs for exclusion
- Punzi definition of sensitivity
- Relation of p-values and Likelihoods
- Probability of misleading evidence
- Sampling to foregone conclusion
- Jeffries-Lindley paradox





Conclusions

Resources:

Software exists: e.g. RooStats

Books exist: Barlow, Cowan, James, Lyons, Roe,…..

New: `Data Analysis in HEP: A Practical Guide to Statistical Methods' , Behnke et al.

PDG sections on Prob, Statistics, Monte Carlo

CMS and ATLAS have Statistics Committees (and BaBar and CDF earlier) – see their websites

Before re-inventing the wheel, try to see if Statisticians have already found a solution to your statistics analysis problem.

Don't use a square wheel if a circular one already exists.

"Good luck"