Hard Scattering in Hadron-Hadron Collisions: Physics and Anatomy

Section 8: Data Analysis Challenges

1. Some Tools to Extract Knowledge
2. Incorporation of Systematic Uncertainties
3. Significance (or not)
4. Perils of Running Blind
5. Resources

Introduction: Some Tools

- Our understanding of high energy hadron collisions has limits
  - It's why we are studying them in the first place
  - But some of the limitations in knowledge "get in the way"
  - Progress is made by being able to control or minimize the uncertainties that issues not relevant to your analysis
- Generally, particle physicists have become pretty good at doing basic statistics
  - But we do get into trouble
  - Discuss a number of tools (and pitfalls) in common use

- Treatment of systematic uncertainties
  - Essential, but often riddled with assumptions and approximations
- Significance – how do we make statements about belief from data?
  - But we do get into trouble
- Blind Analyses
  - All about avoiding unconscious or conscious bias
  - But there are challenges
- Resources Available
  - No re-invention of wheels please

Literature Summary

- **Some classic statistics resources**
  - F. Solmitz, "Analysis of Experiments in Particle Physics", Annu. Rev. Nucl. Sci. 1964:14, 375-402.
  - J. Orear, "Notes on Statistics for Physicists", CLNS 82/511 (1982), http://pages.physics.cornell.edu/p510/w/images/p510b/6/62/Notes_on_Statistics_for_Physicists.pdf
- **Systematic Uncertainty References**
  - P. Sinervo, "Definition and Treatment of Systematic Uncertainties", http://www.slac.stanford.edu/econf/C030908/papers/TUAT004.pdf

1. Systematic Uncertainties

- **Systematic uncertainties play key role in physics measurements**
  - **Few formal definitions exist, much "oral tradition"**
  - **"Know" they are different from statistical uncertainties**

Random Uncertainties	**Systematic Uncertainties**
Arise from stochastic fluctuations	Due to uncertainties in the apparatus or model
Uncorrelated with previous measurements	Usually correlated with previous measurements
Well-developed theory	Limited theoretical framework
Examples: **measurement resolution**; **finite statistics**; **random variations in system**	Examples: **calibrations uncertainties**; **detector acceptance**; **poorly-known theoretical parameters**

Literature Summary

- **Increasing literature on the topic of "systematics"**

  A representative list:

  - R.D.Cousins & V.L. Highland, NIM **A320**, 331 (1992).
  - C. Guinti, Phys. Rev. D **59** (1999), 113009.
  - G. Feldman, "Multiple measurements and parameters in the unified approach," presented at the FNAL workshop on Confidence Limits (Mar 2000).
  - R. J. Barlow, "Systematic Errors, Fact and Fiction," hep-ex/0207026 (Jun 2002), and several other presentations in the Durham conference.
  - G. Zech, "Frequentist and Bayesian Confidence Limits," Eur. Phys. J, **C4:12** (2002).
  - R. J. Barlow, "Asymmetric Systematic Errors," hep-ph/0306138 (June 2003).
  - A. G. Kim et al., "Effects of Systematic Uncertainties on the Determination of Cosmological Parameters," astro-ph/0304509 (April 2003).
  - J. Conrad et al., "Including Systematic Uncertainties in Confidence Interval Construction for Poisson Statistics," Phys. Rev. D **67** (2003), 012002
  - G.C.Hill, "Comment on "Including Systematic Uncertainties in Confidence Interval Construction for Poisson Statistics"," Phys. Rev. D **67** (2003), 118101.
  - G. Punzi, "Including Systematic Uncertainties in Confidence Limits", CDF Note in preparation.

I. Case Study #1: W Boson Cross Section

- **Rate of W boson production**
  - **Count candidates N_s+N_b**
  - **Estimate background N_b & signal efficiency e**

$$\sigma = (N_c - N_b)/(\varepsilon L)$$

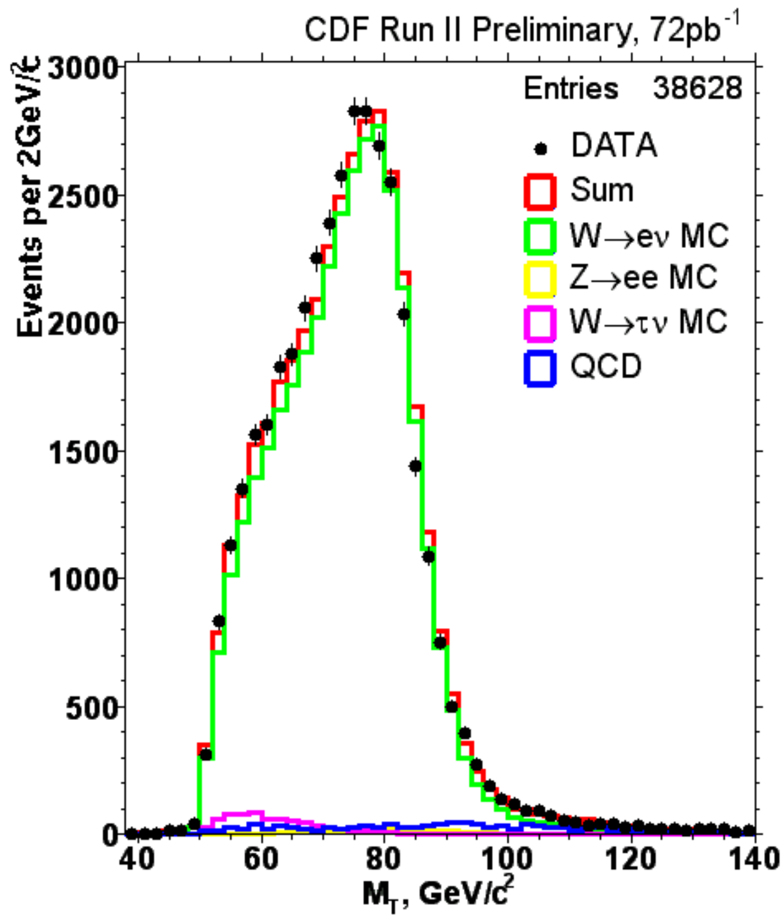
  - **Measurement reported as**

$$\sigma = 2.64 \pm 0.01 \text{ (stat)} \pm 0.18 \text{ (syst) nb}$$

  - **Uncertainties are**

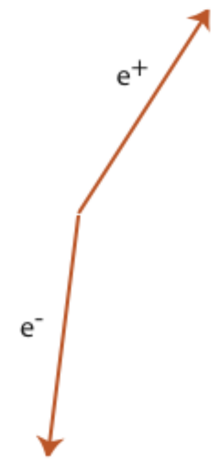
$$\sigma_{stat} \cong \sigma_0^{stat} \sqrt{1/N_c}$$

$$\sigma_{syst} \cong \sigma_0^{syst} \sqrt{(\delta N_b/N_b)^2 + (\delta\varepsilon/\varepsilon)^2 + (\delta L/L)^2}$$



Definitions are Relative

- **Efficiency uncertainty estimated using Z boson decays**
  - **Count up number of Z candidates N_Z^{cand}**
    - Can identify using charged tracks
    - Count up number reconstructed N_Z^{recon}



$$\varepsilon = \frac{N_Z^{recon}}{N_Z^{cand}} \Rightarrow \delta\varepsilon \cong \sqrt{\frac{N_Z^{recon}\left(N_Z^{cand} - N_Z^{recon}\right)}{N_Z^{cand}}}$$

  - **Redefine uncertainties**

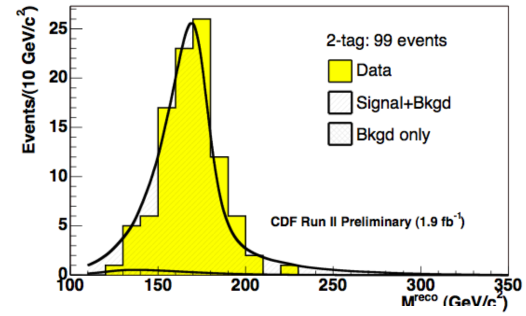
$$\sigma_{stat} \cong \sigma_0 \sqrt{1/N_c + \left(\delta\varepsilon/\varepsilon\right)^2}$$

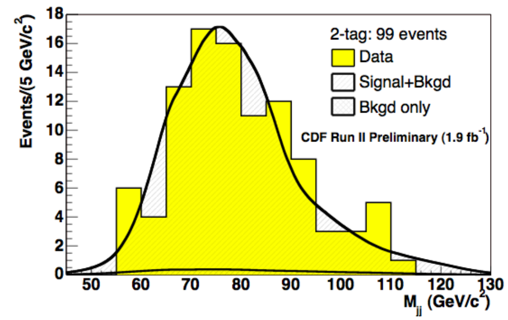
$$\sigma_{syst} \cong \sigma_0 \sqrt{\left(\delta N_b / N_b\right)^2 + \left(\delta L/L\right)^2}$$

Lessons:
- Some systematic uncertainties are really "random"
- Good to know this
  - Uncorrelated
  - Know how they scale
- May wish to redefine
- Call these "CLASS 1" Systematics

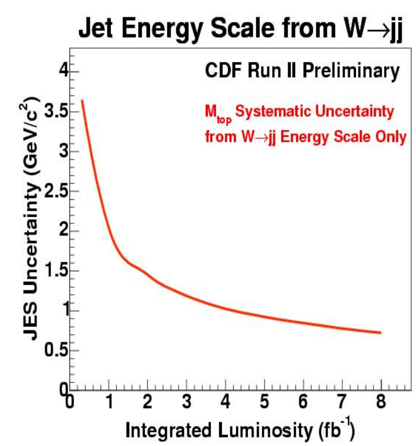
Top Mass Good Example

- Top mass uncertainty in template analysis
  - Statistical uncertainty from shape of reconstructed mass distribution and statistics of sample
  - Systematic uncertainty coming from jet energy scale (JES)
    - Determined by calibration studies, dominated by modelling uncertainties
    - 5% systematic uncertainty





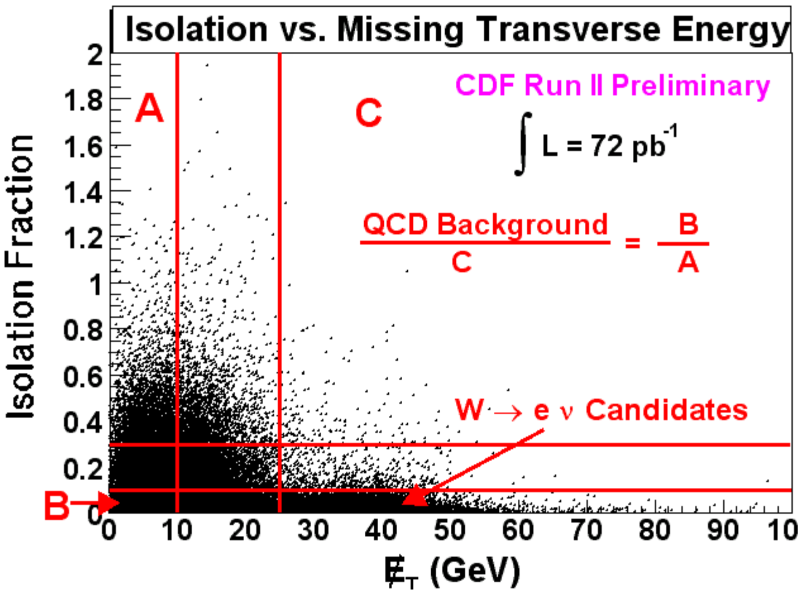
- Latest techniques determine JES uncertainty from dijet mass peak (W->jj)
  - Turn JES uncertainty into a largely statistical one
  - Introduce other smaller systematics



$$M_{top} = 171.8 \pm 1.9\,(\text{stat} + \text{JES}) \pm 1.0\ (\text{syst})\ \text{GeV/c}^2$$
$$= 171.9 \pm 2.1\ \text{GeV/c}^2$$

Case Study #2: Background Uncertainty

- **Look at same W cross section analysis**
  - **Estimate of N_b dominated by QCD backgrounds**
    - Candidate event
      - **Have non-isolated leptons**
      - **Less missing energy**
    - Assume that isolation and MET uncorrelated
    - Have to estimate the uncertainty on N_b^{QCD}



  - **No direct measurement has been made to verify the model**
  - **Estimates using Monte Carlo modelling have large uncertainties**

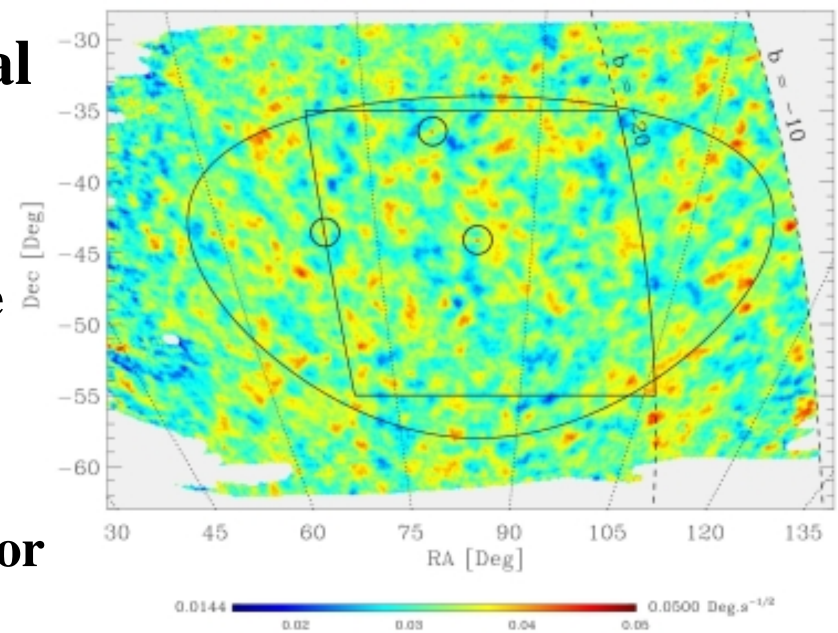
Estimation of Uncertainty

- **Fundamentally different class of uncertainty**
  - **Assumed a model for data interpretation**
  - **Uncertainty in N_b^{QCD} depends on accuracy of model**
  - **Use "informed judgment" to place bounds on one's ignorance**
    - Vary the model assumption to estimate robustness
    - Compare with other methods of estimation
- **Difficult to quantify in consistent manner**
  - **Largest possible variation?**
    - Asymmetric?
  - **Estimate a "1 s" interval?**
  - **Take** $\sigma \approx \frac{\Delta}{\sqrt{12}}$?

Lessons:
- Some systematic uncertainties reflect ignorance of one's data
- Cannot be constrained by observations
- Call these "CLASS 2" Systematics

Case Study #3: Boomerang CMB Analysis

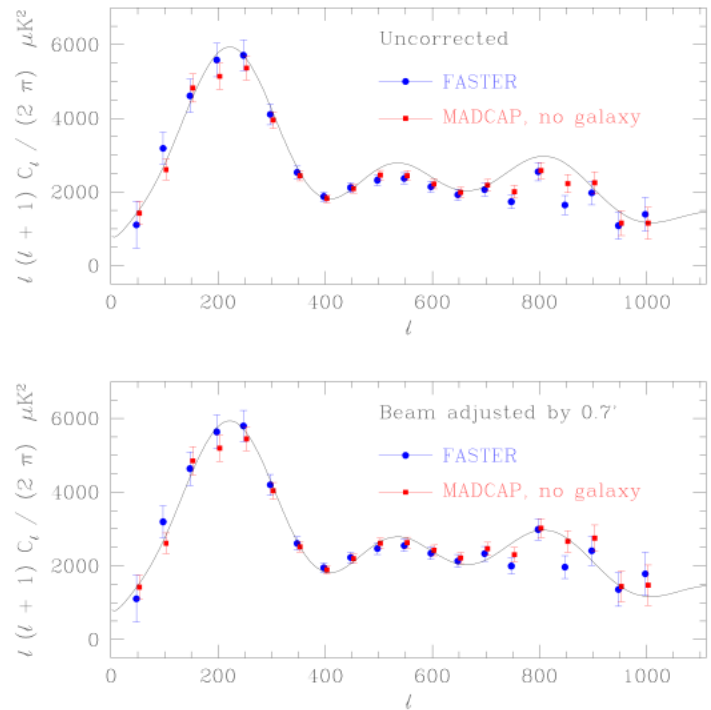
- **Boomerang is one of several CMB probes**
  - **Mapped CMB anisoptropy**
  - **Data constrain models of the early universe**



- **Analysis chain:**
  - **Produce a power spectrum for the CMB spatial anisotropy**
    - > Remove instrumental effects through a complex signal processing algorithm
  - **Interpret data in context of many models with unknown parameters**

Incorporation of Model Uncertainties

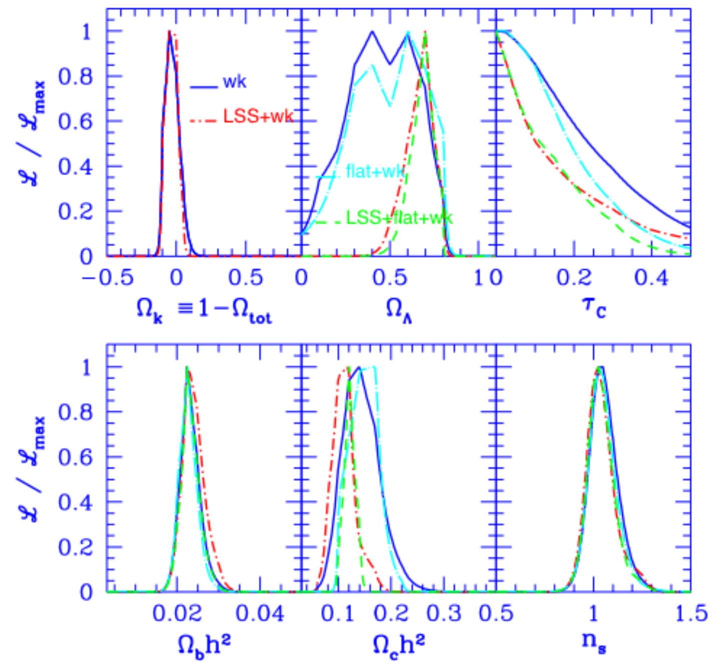
- **Power spectrum extraction includes all instrumental effects**
  - **Effective size of beam**
  - **Variations in data-taking procedures**



- **Use these data to extract 7 cosmological parameters**
  - **Take Bayesian approach**
    - Family of theoretical models defined by 7 parameters
    - Define a 6-D grid (6.4M points), and calculate likelihood function for each

Marginalize Posterior Probabilities

- **Perform a Bayesian "averaging" over a grid of parameter values**
  - **Marginalize w.r.t. the other parameters**
    - NB: instrumental uncertainies included in approximate manner
  - **Chose various priors in the parameters**



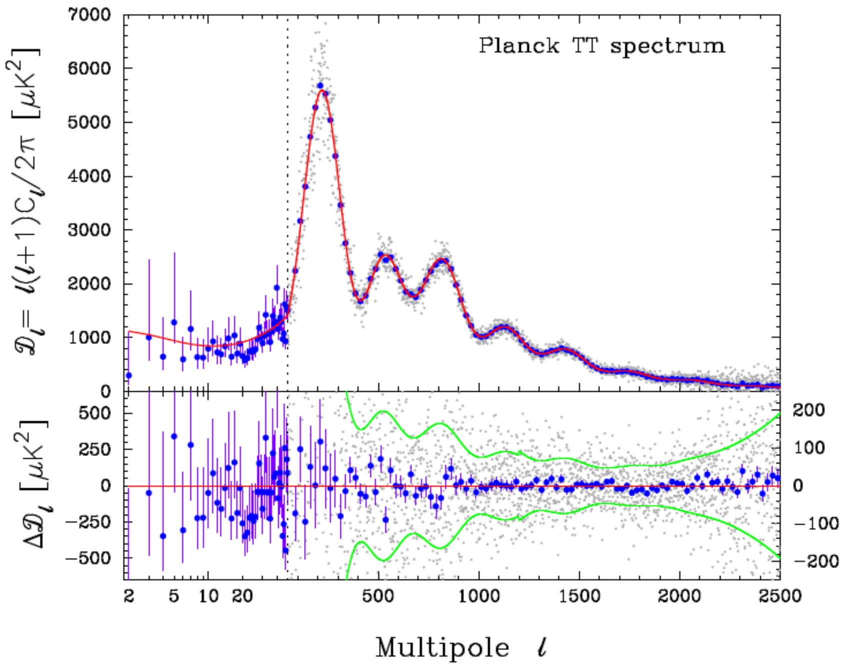
- **Comments:**
  - **Purely Bayesian analysis with no frequentist analogue**
  - **Provides path for inclusion of additional data (eg. WMAP)**

Lessons:
- Some systematic uncertainties reflect paradigm uncertainties
- No relevant concept of a frequentist ensemble
- Call these "CLASS 3" Systematics

Latest Planck Results

- **The prior uncertainties dominate**

	Planck	
Parameter	Best fit	68% limits
$\Omega_b h^2$	0.022068	0.02207 ± 0.00033
$\Omega_c h^2$	0.12029	0.1196 ± 0.0031
$100\theta_{MC}$	1.04122	1.04132 ± 0.00068
τ	0.0925	0.097 ± 0.038
n_s	0.9624	0.9616 ± 0.0094
$\ln(10^{10} A_s)$	3.098	3.103 ± 0.072
Ω_Λ	0.6825	0.686 ± 0.020
Ω_m	0.3175	0.314 ± 0.020
σ_8	0.8344	0.834 ± 0.027
z_{re}	11.35	$11.4^{+4.0}_{-2.8}$
H_0	67.11	67.4 ± 1.4
$10^9 A_s$	2.215	2.23 ± 0.16
$\Omega_m h^2$	0.14300	0.1423 ± 0.0029
$\Omega_m h^3$	0.09597	0.09590 ± 0.00059
Y_P	0.247710	0.24771 ± 0.00014
Age/Gyr	13.819	13.813 ± 0.058



Planck Collaboration, 1303.5076v3 (2014)

Proposed Taxonomy for Systematic Uncertainties

- **Three "classes" of systematic uncertainties**
  - **Uncertainties that can be constrained by ancillary measurements**
  - **Uncertainties arising from model assumptions or problems with the data that are poorly understood**
  - **Uncertainties in the underlying models**
- **Estimation of Class 1 uncertainties straightforward**
  - **Class 2 and 3 uncertainties present unique challenges**
  - **In many cases, have nothing to do with statistical uncertainties**
    - Driven by our desire to make inferences from the data using specific models

II. Estimation Techniques

- **No formal guidance on how to define a systematic uncertainty**
  - **Can identify a possible source of uncertainty**
  - **Many different approaches to estimate their magnitude**
    - Determine maximum effect D

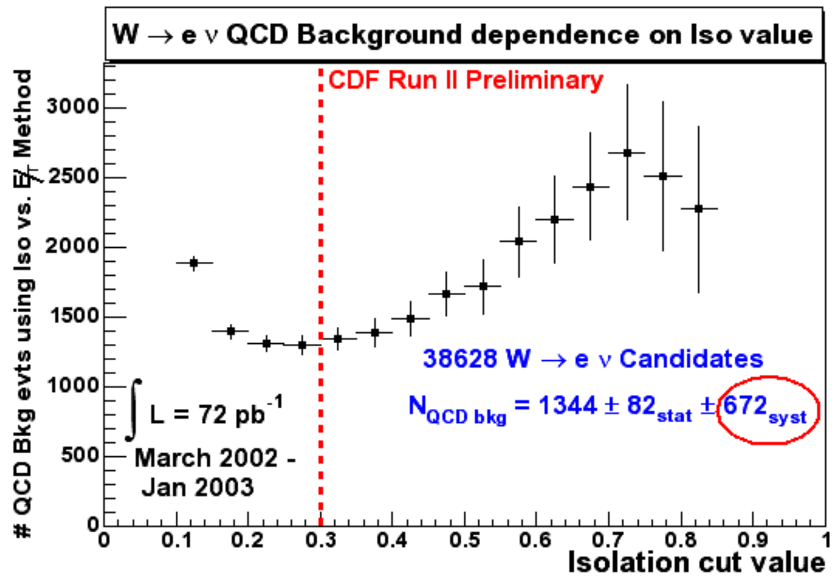
$$\sigma = \frac{\Delta}{2}?$$

$$\sigma = \frac{\Delta}{\sqrt{12}}?$$

- **General rule:**
  - **Maintain consistency with definition of statistical intervals**
  - **Field is pretty glued to 68% confidence intervals**
  - **Recommend attempting to reflect that in magnitudes of systematic uncertainties**
  - **Avoid tendency to be "conservative"**

Estimate of Background Uncertainty in Case Study #2

- Look at correlation of Isolation and MET
  - Background estimate increases as isolation "cut" is raised
  - Difficult to measure or accurately model
    - Background comes primarily from very rare jet events with unusual properties
    - Very model-dependent



- Assume a systematic uncertainty representing the observed variation
  - Authors argue this is a "conservative" choice

Cross-Checks Vs Systematics

- **R. Barlow makes the point in Durham(PhysStat02)**
  - **A cross-check for robustness is not an invitation to introduce a systematic uncertainty**
    - Most cross-checks confirm that interval or limit is robust,
      - **They are usually not designed to measure a systematic uncertainty**
- **More generally, a systematic uncertainty should**
  - **Be based on a hypothesis or model with clearly stated assumptions**
  - **Be estimated using a well-defined methodology**
  - **Be introduced *a posteriori* only when all else has failed**

III. Statistics of Systematic Uncertainties

- **Goal has been to incorporate systematic uncertainties into measurements in coherent manner**
  - **Increasing awareness of need for consistent practice**
    - Frequentists: interval estimation increasingly sophisticated
      - **Neyman construction, ordering strategies, coverage properties**
    - Bayesians: understanding of priors and use of posteriors
      - **Objective vs subjective approaches, marginalization/conditioning**
  - **Systematic uncertainties threaten to dominate as precision and sensitivity of experiments increase**
- **There are a number of approaches widely used**
  - **Summarize and give a few examples**
  - **Place it in context of traditional statistical concepts**

Formal Statement of the Problem

- **Have a set of observations x_i, $i=1,n$**
  - **Associated probability distribution function (pdf) and likelihood function**

$$\mathrm{p}(x_i \mid \theta) \Rightarrow \mathcal{L}(\theta) = \prod_i \mathrm{p}(x_i \mid \theta)$$

  - Depends on unknown random parameter *q*
  - Have some additional uncertainty in pdf
    - **Introduce a second unknown parameter *l***

$$\mathcal{L}(\theta,\lambda) = \prod_i \mathrm{p}(x_i \mid \theta,\lambda)$$

- **In some cases, one can identify statistic y_j that provides information about *l***

$$\mathcal{L}(\theta,\lambda) = \prod_{i,j} \mathrm{p}(x_i, y_j \mid \theta,\lambda)$$

  - **Can treat *l* as a "nuisance parameter"**

Bayesian Approach

- **Identify a prior *p(l)* for the "nuisance parameter" *l***
  - **Typically, parametrize as either a Gaussian pdf or a flat distribution within a range ("tophat")**
  - **Can then define Bayesian posterior**

$$\mathcal{L}(\theta,\lambda)\ \pi(\lambda)\ d\theta\ d\lambda$$

  - **Can marginalize over possible values of *l***
    - Use marginalized posterior to set Bayesian credibility intervals, estimate parameters, etc.

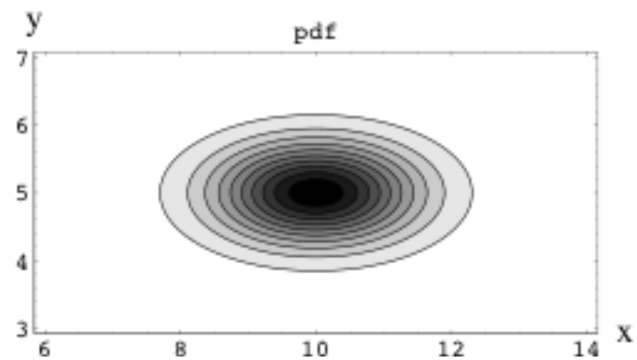
- **Theoretically straightforward ….**
  - **Issues come down to choice of priors for both *q, l***
    - No widely-adopted single choice
    - Results have to be reported and compared carefully to ensure consistent treatment

Frequentist Approach

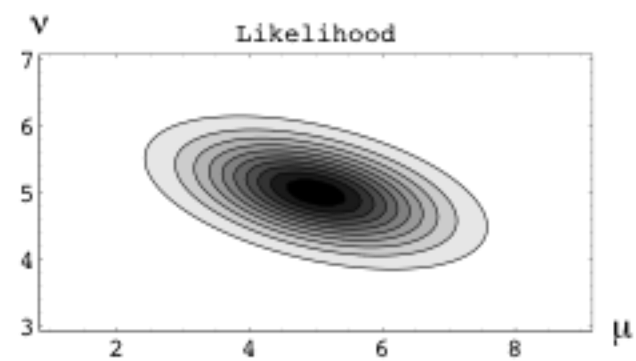
- **Start with a pdf for data** $p(x_i, y_j \mid \theta, \lambda)$
  - **In principle, this would describe frequency distributions of data in multi-dimensional space**
  - **Challenge is take account of nuisance parameter**
  - **Consider a toy model**

$$p(x, y \mid \mu, \nu) = G(x - (\mu + \nu), 1)\, G(y - \nu, s)$$

    - Parameter s is Gaussian width for *n*



- **Likelihood function (*x=10, y=5*)**
  - **Shows the correlation**
  - **Effect of unknown n**



Formal Methods to Eliminate Nuisance Parameters

- **Number of formal methods exist to eliminate nuisance parameters**
  - **Of limited applicability given the restrictions**
  - **Our "toy example" is one such case**
    - Replace *x* with *t=x-y* and parameter *n* with

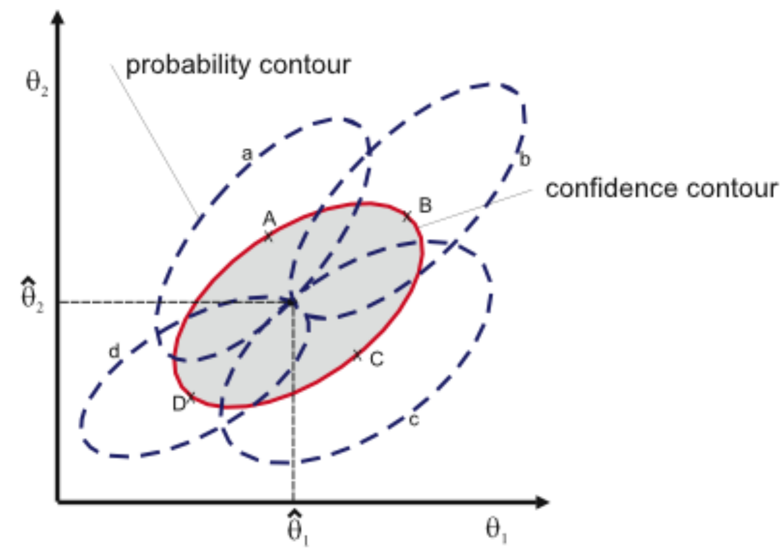
$$\nu' \equiv \nu + {\mu s^2}/{1+s^2}$$

$$\Rightarrow \mathrm{p}(t, y \mid \mu, \nu') = \mathrm{G}\left(t-\mu, \sqrt{1+s^2}\right) \mathrm{G}\left(y - \nu' + {ts^2}/{1+s^2}, {s}/{\sqrt{1+s^2}}\right)$$

    - Factorized pdf and can now integrate over *n'*
    - Note that pdf for *m* has larger width, as expected
  - **In practice, one often loses information using this technique**

Alternative Techniques for Treating Nuisance Parameters

- **Project Neyman volumes onto parameter of interest**
  - **"Conservative interval"**
  - **Typically over-covers, possibly badly**



From G. Zech

- **Choose best estimate of nuisance parameter**
  - **Known as "profile method"**
  - **Coverage properties require definition of ensemble**
  - **Can possible under-cover when parameters strongly correlated**
    - Feldman-Cousins intervals tend to over-cover slightly (private communication)

Example: Solar Neutrino Global Analysis

- **Many experiments have measured solar neutrino flux**
  - **Gallex, SuperKamiokande, SNO, Homestake, SAGE, etc.**
  - **Standard Solar Model (SSM) describes n spectrum**
  - **Numerous "global analyses" that synthesize these**
- **Fogli et al. have detailed one such analysis**
  - **81 observables from these experiments**
  - **Characterize systematic uncertainties through 31 parameters**
    - 12 describing SSM spectrum
    - 11 (SK) and 7 (SNO) systematic uncertainties
- **Perform a χ^2 analysis**
  - **Look at χ^2 to set limits on parameters**

Hep-ph/0206162, 18 Jun 2002

Formulation of χ^2

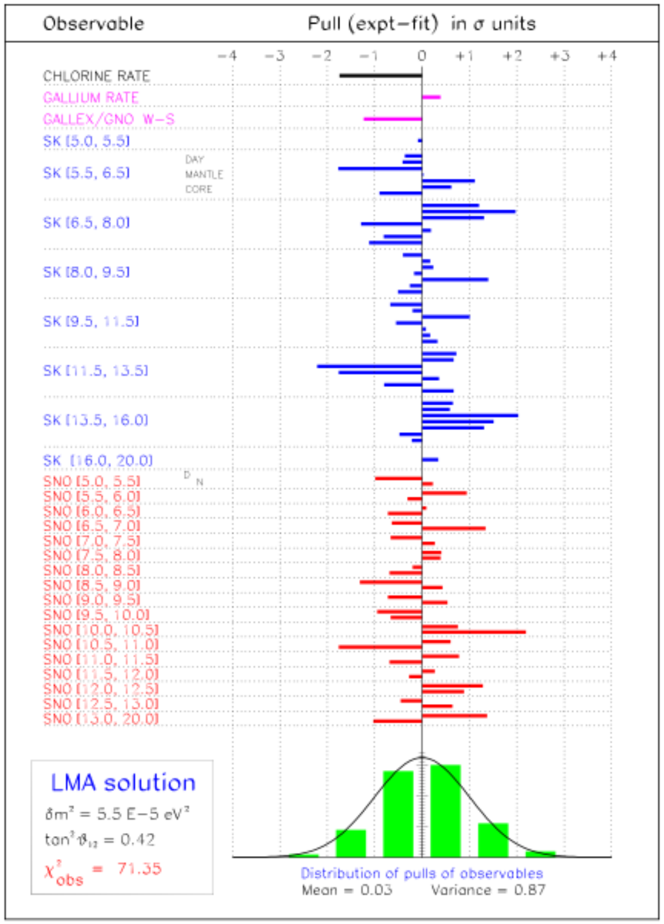
- **In formulating χ^2, linearize effects of the systematic uncertainties on data and theory comparison**

$$\chi^2_{pull} \equiv \min_{\{\xi\}} \left[\sum_{n=1}^{N} \left(\frac{R_n^{\exp t} - R_n^{theor} - \sum (c_n^k \xi_k)}{u_n} \right)^2 + \sum_{k=1}^{K} \xi_k^2 \right]$$

  - > Uncertainties u_n for each observable
  - **Introduce "random" pull x_k for each systematic**
    - > Coefficients $c_k{}^n$ to parameterize effect on *nth* observable
    - > Minimize χ^2 with respect to x_k
    - > Look at contours of equal $\Delta\chi^2$

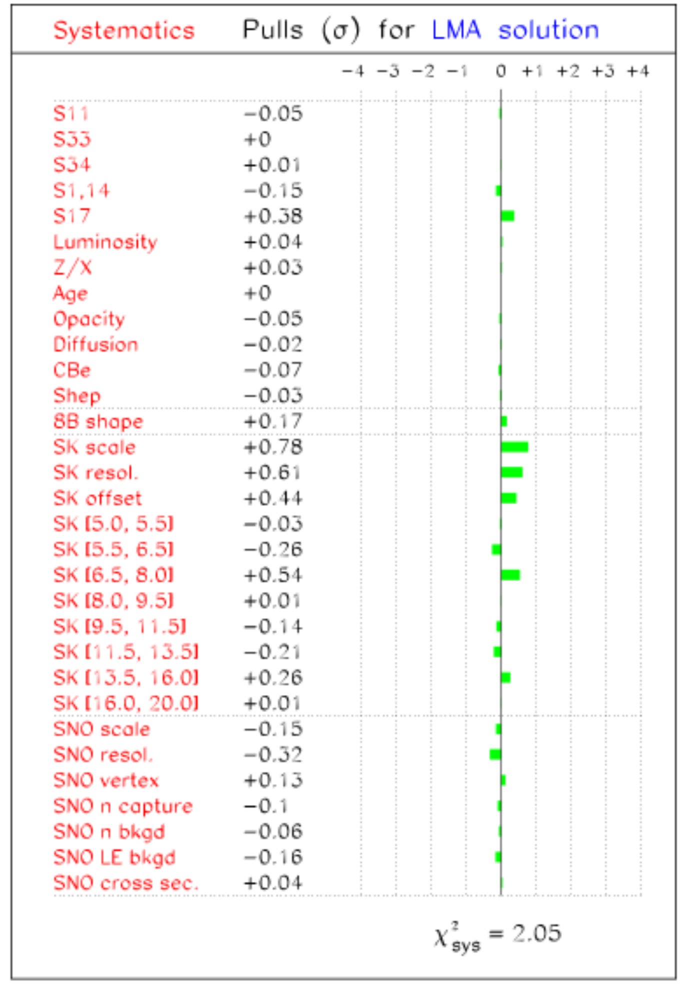
Solar Neutrino Results

- **Can look at "pulls" at χ^2 minimum**
  - **Have reasonable distribution**
  - **Demonstrates consistency of model with the various measurements**
  - **Can also separate**
    - Agreement with experiments
    - Agreement with systematic uncertainties



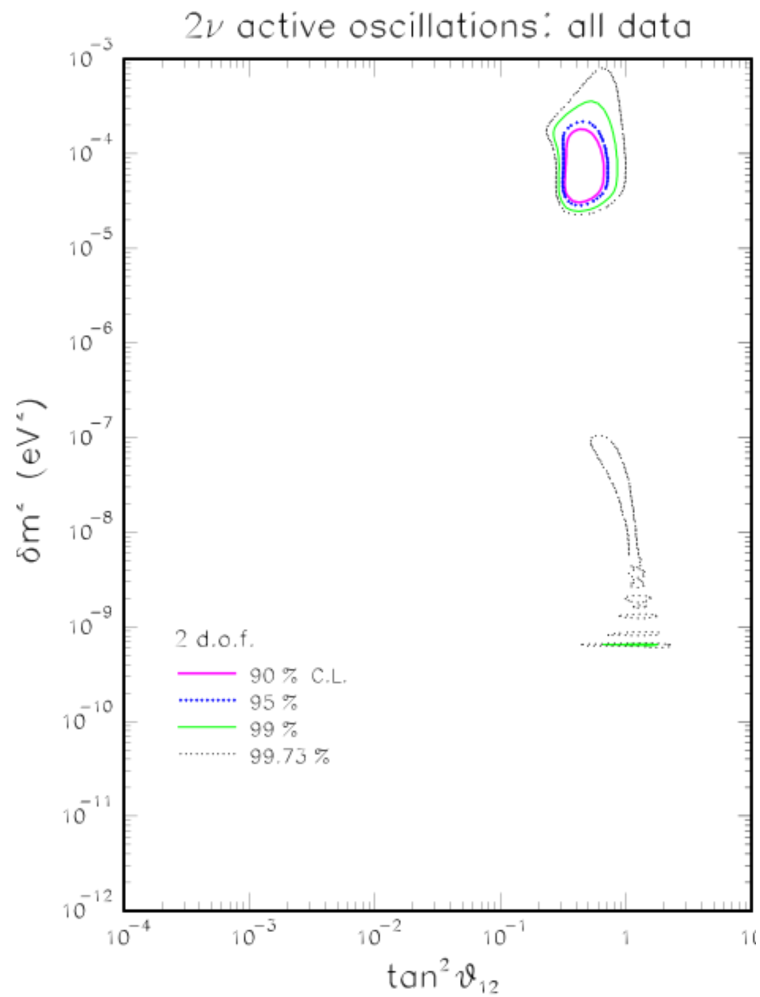
Pull Distributions for Systematics

- **Pull distributions for x_k also informative**
  - **Unreasonably small variations**
  - **Estimates are globally too conservative?**
  - **Choice of central values affected by data**
    - Note this is NOT a blind analysis
- **But it gives us some confidence that intervals are realistic**



Typical Solar Neutrino Contours

- Can look at probability contours
  - Assume standard χ^2 form
  - Probably very small probability contours have relatively large uncertainties



Hybrid Techniques

- **A popular technique (Cousins-Highland) does an "averaging" of the pdf**
  - **Assume a pdf for nuisance parameter g(/)**
  - **"Average" the pdf for data *x***

$$p_{CH}(x \mid \theta) \equiv \int p(x \mid \theta,\lambda)\, g(\lambda)\; d\lambda$$

  - **Argue this approximates an ensemble where**
    - Each measurement uses an apparatus that differs in parameter /
      - **The pdf g(/) describes the frequency distribution**
    - Resulting distribution for *x* reflects variations in /

- **Intuitively appealing**
  - **But fundamentally a Bayesian approach**
  - **Coverage is not well-defined**

See, for example, J. Conrad et al.

Computationally Challenging

- **In many measurements**
  - **Can have several dozen sources of systematic uncertainty**
  - **Creating a tractable ensemble is not possible**
  - **Even the definition of the ensemble is controversial**

- **Current state of the art is to perform a Bayesian-like "marginalization"**
  - **Treat the new probability function in the same way as before**
  - **But**
    - Not clear how to evaluate coverage
    - Not strongly grounded in theory

2. What is Significance?

- **Typical HEP approach**
  - **Have a set of observations**
  - **We say the data are "statistically significant" when**
    - We can use data to support a specific hypothesis, eg.
      - **"We see a phenomenom not predicted by the Standard Model"**
      - **"We report the discovery of X"**
    - The interpretation eliminates a number of competing hypotheses
    - The conclusion will not likely be altered with larger statistics or further analysis
- **Want a statistical framework that**
  - **Measures "degree of belief"**
  - **Ensures robust conclusions**

Some “Obvious” Discoveries

- **Observation of $B^0\overline{B}^0$ Mixing**
  - 24.8 ± 7.6 ± 3.8 like-sign events vs 25.2 ± 5.0 ± 3.8 opposite sign
  - “3σ” discovery

Albrecht et al., PLB 192, 245 (1987)

- **W Boson**
  - 6 eν events, no background!

Arnison et al., PLB 122, 103 (1983)

- **Upsilon**
  - 770 events on 350 background
  - Described as “significant” but no measure of it

Herb et al., PRL 39, 252 (1977)

- **B mesons**
  - 18 events on 4-7 background
  - No measure of significance

Behrends et al., PRL 50, 881 (1983)

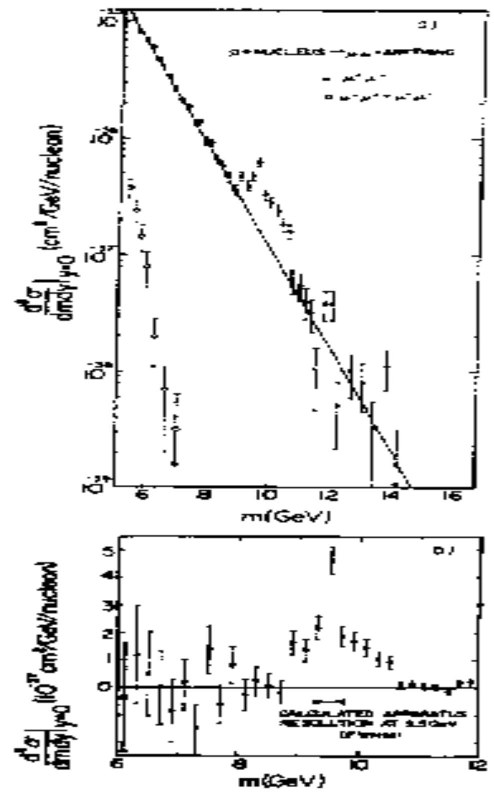
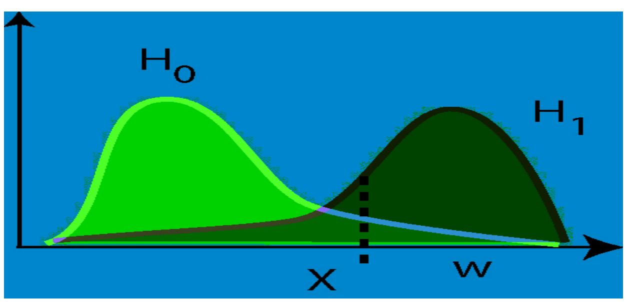


FIG. 3. (a) Measured dimuon production cross sections as a function of the invariant mass of the muon pair. The solid line is the continuum fit outlined in the text. The equal-sign—dimuon cross section is also shown. (b) The same cross sections as in (a) with the smooth exponential continuum fit subtracted in order to reveal the 9—10-GeV region in more detail.

A Frequentist Definition

- **Significance defined in context of "hypothesis testing"**
  - **Have two hypotheses, H_0 and H_1, and possible set of observations X**
    - Choose a "critical region", w, in the space of observations X
    - Define **significance**, α, as probability of $X \in w$ when H_0 is true
    - Define the **power**, $1-\beta$, as probability of $X \in w$ when H_1 is true



Typically, H0 is "null" hypothesis

- **In this language, an observation is "significant" when**
  - **Significance α is small & β is small**
    - Typically $a < \text{few } 10^{-5}$

Some Comments on Formal Definition

- **Definition depends on**
  - **Choice of statistic X**
    - Left up to the experimenter as part of design
    - More on that later
  - **Choice of "critical region" w**
    - Depends on hypotheses
    - Often chosen to minimize systematic uncertainties?
    - Not necessarily defined in advance!
  - **Definition of "probability"**
    - A frequentist definition
    - Raises issue of how systematic uncertainties are managed
  - **Choice of α and β**
    - Matter of "taste" and precedent
    - A small α is safe, but comes with less "discovery reach"

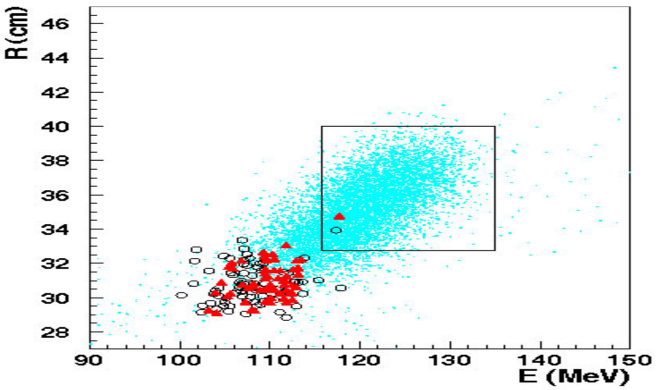
- **More fundamentally:**
  - **Is this an adequate definition of "significance?"**

The Choice of Statistic & Critical Region

- **Choice of statistic motivated by specific experimental design**
  - **Informed by the measurement to be made**
  - **Critical region is chosen at the same time**
  - **Good example: E787/E949 search**

$$K^+ \to \pi^+ \nu\nu$$

    - Look for $\pi^+ \to \mu^+ \nu$ decay
    - Define a "box" a priori
      - **Expected 0.15±0.05 event bkgd**



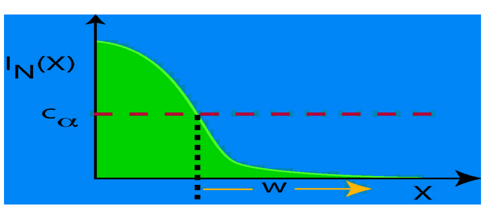
Only two events Observed

Significance 0.02%

Have used the "box" Since 1988

Optimal Tests: Neyman-Pearson

- **In some cases, possible to identify the "most powerful" test**
  - **Must involve only "simple" hypotheses (no free parameters)**
    - PDF's given by $f_i(X)$
    - Must have two hypotheses
  - **For given α, can identify region to minimize β for alternative H_1**
    - Order observations by $l_N(X) \equiv f_0(X) / f_1(X)$
    - Can minimize β by choosing critical region as all X s.t. $l_N(X) \geq c_a$
      - **Chose c_a so that** $\int_w f_0(X)\,dX = \alpha$



Caveats to Neyman-Pearson

- **Neyman-Pearson limited**
  - **Only true for simple hypotheses**
    - Not for composite hypotheses (where unknown parameter)
  - **Compares two hypotheses**
    - Depends on alternative hypothesis
    - Makes results model-dependent
- **But does give some insight**
  - **The ratio $I_N(X)$ is proportional to ratio of likelihoods**

$$f_0(X) / f_1(X) \cong \mathcal{L}_0(X) / \mathcal{L}_1(X)$$

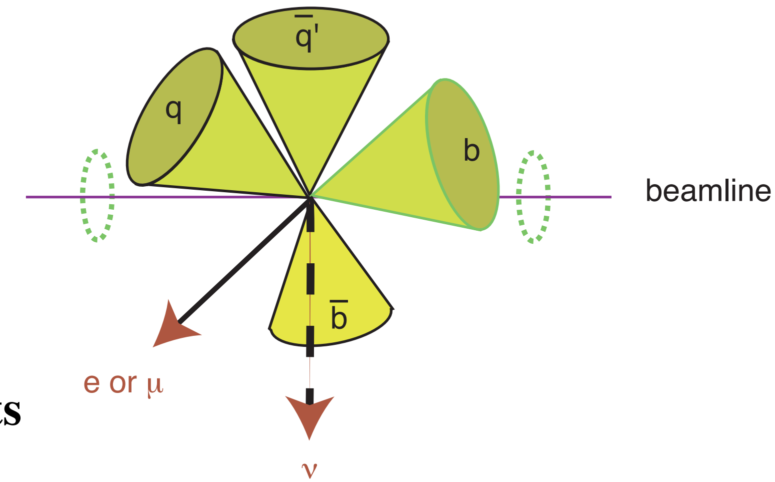
  - **Provides guidance for definition of effective tests**

Definition of Critical Region

- **Challenge is not to bias choice of critical region with data**
  - **However, observer required to understand data**
    - Identify instrumental pathologies
    - Identify unexpected backgrounds
    - Estimate systematic uncertainties
    - Verify stable run conditions
  - **Studies may lead to unconscious bias (see, eg. RPP plots!)**
- **"Blind" analyses are popular**
    - Study data complementary to signal
    - However, implementation varies
      - **SNO's pure D_2O results set aside about 40% of data**
      - **Not clear that this really helps!**
    - Even E787/E949 reserve right to examine background rejection

Significance in Counting Experiments

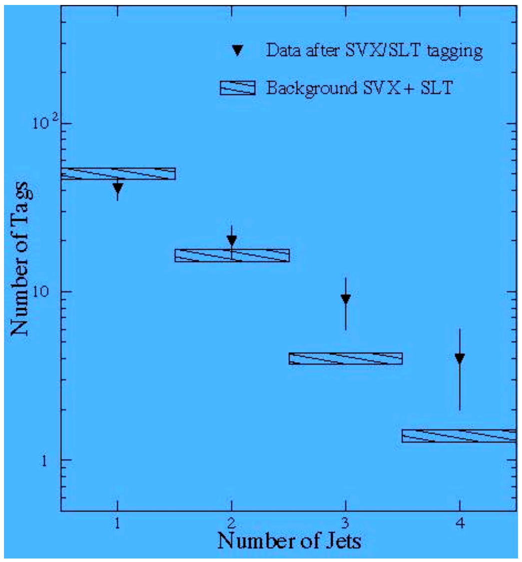
- **Top quark search is textbook example**
  - **By 1991, CDF had ruled out top quark with mass < 91 GeV/c^2**
  - **Searching for top quark pair production and decay into**
    - Lepton + n + jets (20%)
    - Dilepton + n + jets (8%)



- **In a sample of 20 pb^{-1}, expected handful of events**
  - **Large background from W + jets**
  - **"Fake" b-quark tags**

Definition of the Measurement

- **Defined clear strategy in 1990**
  - **Identify lepton+jets and dilepton candidates**
  - **Count "b" tags in lepton+jet events**
    - Use two b-tagging algorithms
      - **Use events with 1-2 jets as control**
      - **Signal sample events with ≥3 jets**
      - **Expected 3.5 evts (M_{top}=160 GeV/c^2)**



Expect **5.4±0.4** tags from background

Observed **13** tagged "b jets" in 10 evts

7 SVX tags
6 lepton tags

  - **For dileptons:**
    - Require 2 or more jets
    - Expected 1.3 evts (M_{top}=160 GeV/c^2)
    - Observed 2 evts, bkd of 0.6±0.3 evts

Significance Calculation

- **Calculated probability of background hypothesis**
  - **Dilepton significance $\alpha_{dil} = 0.12$**
  - **Used MC calculation**
    - Treated background uncertainty as a normally distributed uncertainty on acceptance
  - **For lepton+jets, MC gives**
    - SVX b tags: $\alpha_{SVX} = 0.032$
    - SLT b tags: $\alpha_{SLT} = 0.038$
- **To combine, take into account correlations**
  - **Gives $\alpha_{tot} = 0.0026$**
  - **If assume independent, then**

    $$\alpha_{tot} = \alpha_{dil}\,\alpha_{ljets}\,[1 - \ln(\alpha_{dil}\,\alpha_{ljets})]$$

    - Gives $\alpha_{tot} = 0.0088$
  - **Collaboration reported only "evidence for top quark…."**
    - Factor 2 more data -- α_{tot} = few 10^{-5}

Power of the Top Quark Statistic

- **Choice of statistic driven by need to reduce background**
  - **Note ε_{ljets} = 0.074 before b-tagging**
    - Predict 12 events signal and 60 events background
    - Tagging efficiency 0.40
      - **Background "efficiency" 0.09**
  - **Definition of "power" problematic**
    - Arbitrary
      - **Power of lepton+jets selection? b-tagging?**
      - ***A posteriori* choice of X = $N_{tags} + N_{dil}$**
    - Experimenter chooses "critical region" based on hypothesis
      - **Lepton+jets Higgs search useδ different selection**

$$W\,H \rightarrow l\,\nu\,b\,b$$

  - **Usually characterized by sensitivity**
    - Size of expected signal

Significance using Data Distributions

- **Measurements often involve continuous observables**
  - **Can assess agreement with "null" hypothesis**
    - Generally "goodness-of-fit" tests

- **Number of tests in common use**
    - χ^2 Test
      - **Depends on choice of binning**
      - **Limited to "large" statistics samples**
        - **Bin contents > 5-10 (?)**
    - Smirnov-Cramer-Von Mises
      - **Define statistic based on cumulative distributions $S_N(x)$**

$$W^2 \equiv \int [S_N(X) - F(X)]^2 f(X)\, dX$$

      - **Probability distribution for W^2 independent of distribution**
        - **$E[W^2] = (6N)^{-1}$ and $V[W^2] = (4N-3)/180N^3$**
    - Kolmogorov-Smirnov
      - **Popular form of test based on $S_N(x)$**
      - **Distribution for D_N proportional to χ^2**

$$D_N \equiv \max |S_N(X) - F(X)|$$

Multivariate Significance

- **Often difficult to reduce data to 1-dimensional statistic**
  - **Typical case has several variables**
    - Different correlations between signal and "null" hypothesis
    - Any straightforward transformation causes loss of information
  - **Several techniques used**
    - Characterize significance of each component and then combine into a single measure of significance
    - More sophisticated, e.g.
      - **Combine information using any one of the techniques discussed by Prosper, Towers, etc.**

- **In practice, two approaches:**

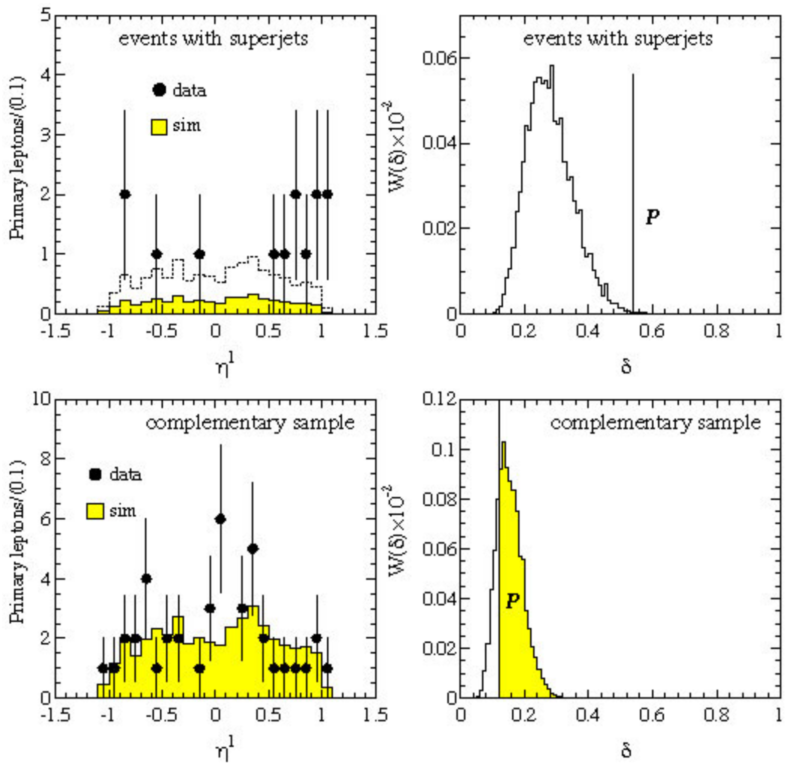
1. **Assume independent statistics**
   - Check for any correlations
2. **Model correlations using MC approaches or "bootstrapping"**
   - Computationally expensive
   - Relies on understanding correlations

An Infamous Example: "Superjets"

- **CDF Run I data contained**
  - **Unusual lepton + ν + 2,3 jet events**
    - 13 events with jets that are both SLT and SVX tagged
      - **Expect 4.4±0.6 events from background sources**
      - **Significance is 0.001!**
  - **Led to examination of 9 kinematical distributions**
      - **P_T & η for leptons & jets, and azimuthal angle between lepton, jet**
      - **P_T and η for lepton+jet system**
    - Perform independent K-S tests
      - **Use control sample defined by events without a "supertag"**
      - **Combined significance of 1.6×10^{-6}**
    - Also defined a new statistic
      - **Sum of K-S distances**
      - **MC gives significance of 3.3×10^{-6}**

K-S Tests on Superjet Data

- **Lepton η distribution**



- **Some approximations:**
  - Control sample events w/o superjet
  - Randomly pick 13 of 42 events

Comments on Superjet Study

- **Choice of statistic (number of superjets) problematic**
  - **Made *a posteriori* after anomaly noted**
    - Significance difficult to assess
  - **Ignored lepton + 1 jet data (where one observes a deficit of events)**
    - Why?

- **Choice of distributions also problematic**
  - **Justified *a posteriori***
  - **Correlations difficult to assess**

- **Aside:**
  - **Interpretation of excess requires unusual physics process**
    - Not a problem in itself
    - But small statistics allow for many hypotheses

Some Practical Proxies for Significance

- **HEP suffers Gaussian tyranny**
  - **Many people will quote numbers of "σ" as measures of significance**
    - Belief that this can be more readily interpreted by lay person
      - **Shorthand for the significance of an ns measurement**
    - 5σ seems to have become conventional "discovery threshold"
      - $\alpha = 2.8\times10^{-7}$
      - **Used for LHC discovery reach**
- **In situations where expected signal S and background B**
  - **Various figures of merit**
    - S/N -- signal versus noise
      - **Doesn't scale with N**
    - More natural definition is

$$\frac{S}{\sqrt{B}}$$

      - **Just normal Gaussian estimate of # of s.d.**
      - **Does scale with N**

See papers by Bityukov & Krasnikov for more discussion

The "Flip-Flopping" Physicist

- **Feldman & Cousins highlighted the problem of "flip-flopping"**
  - **A physicist who uses**
    - One set of criteria to set a limit in the absence of a signal
    - Different criteria to claim a significant signal
  - **Results in confidence intervals with ill-defined frequentist coverage**

- **This should be anticipated in any experiment that wishes to be sensitive to small signals**
  - **F-C propose their "unified approach"**

What About Reverend Bayes?

- **Bayesian approach to classifying hypotheses is**

$$\frac{P(H_1 \mid X)}{P(H_0 \mid X)} = \frac{P(X \mid H_1)}{P(X \mid H_0)} \bullet \frac{\pi(H_1)}{\pi(H_0)}$$

  - **Few comments:**
    - $P(X|H_i)$ is typically likelihood
    - Only meaningful in comparison of two hypotheses
    - Can handle composite hypotheses readily
      - **Just integrate over any "nuisance" variables**

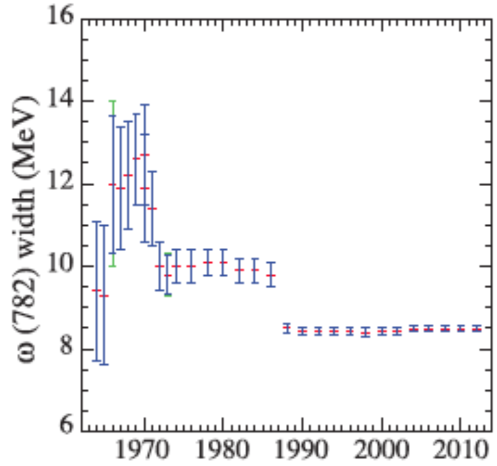
- **Is it used? Not often…**
  - **Only relative "degree of belief"**
    - Requires at least two hypotheses
  - **"Prior" avoidance**
  - **Challenges where single points in parameter space are important**
    - Is sin2b = 0?

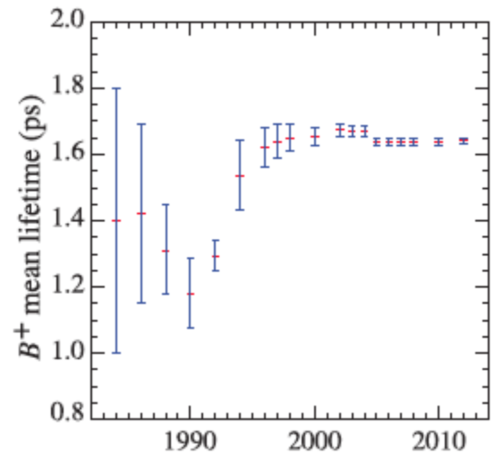
Some Recommendations

- **Define strategy in advance of data analysis**
  - **Otherwise, significance estimates could and will be biased**
  - **"Blind" analyses can play a role**
    - However, this should not limit the ability to "explore" the data
- **Take consistent approach to CL setting & signal measurement**
  - **Avoid "flip-flopping" -- F-C offers one approach to this problem**
- **Describe clearly how you are determining "significance"**
  - **Things to remember:**
    - Definition of probability
    - Definition of critical region
    - What decisions were taken *a posteriori*?

3. Blind Analyses

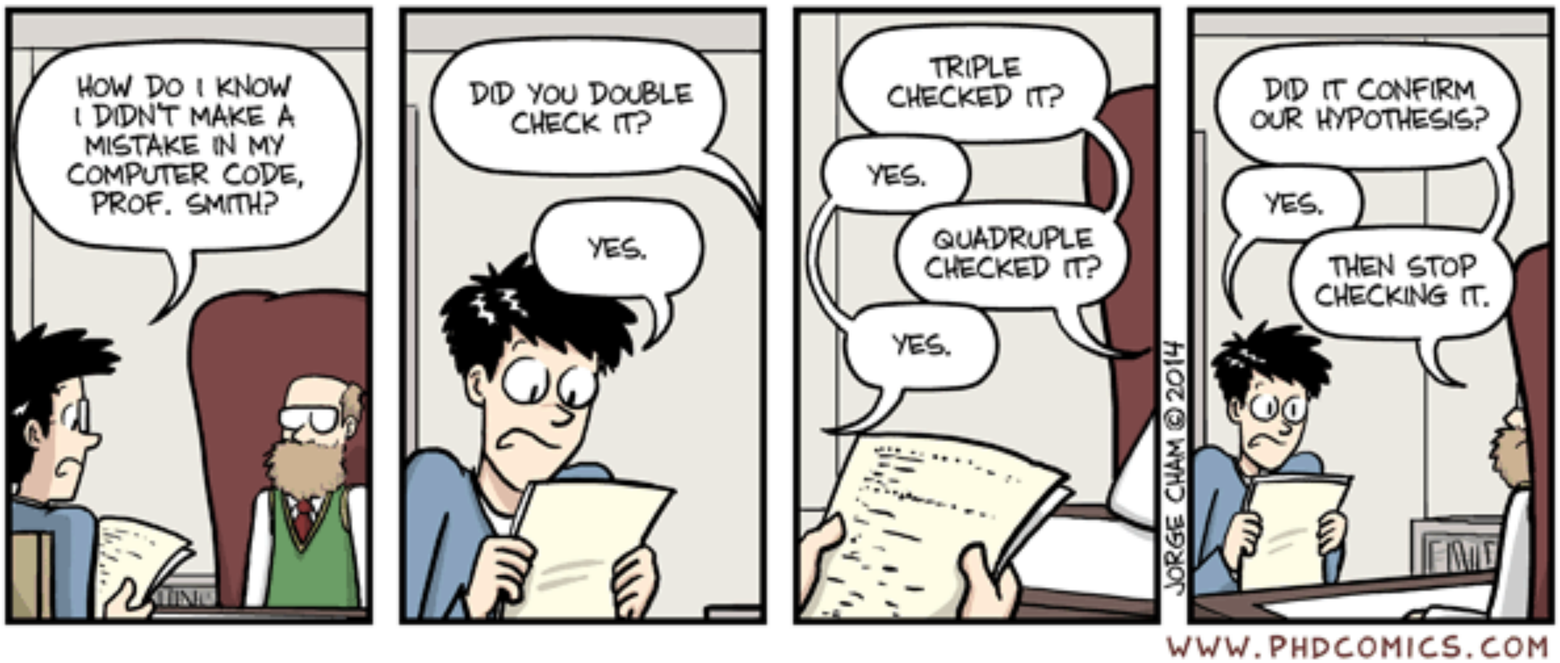
- **To make inferences, we have to assume:**
  - Random events free from correlation
  - More data results in greater precision
  - Procedures used are free of bias
- **Are these reasonable assumptions?**
- **PDG has a set of "history" plots**
  - Reveal that some measurements are just **wrong**
  - Post mortems have indicated that some bias had crept into analysis
    - Looking for the right answer?
    - Selection biased by data itself?





Piled Higher and Deeper

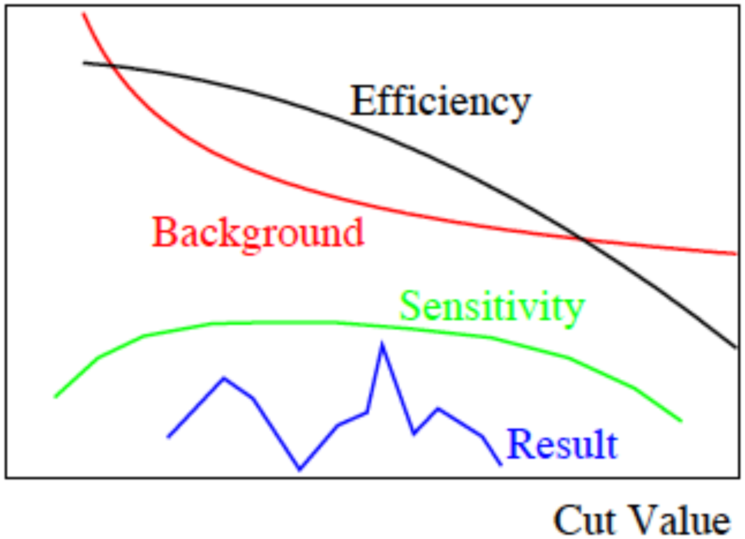
Piled Higher and Deeper ***by Jorge Cham*** **www.phdcomics.com**



title: "Check it" - originally published 3/31/2014

How Can This Happen?

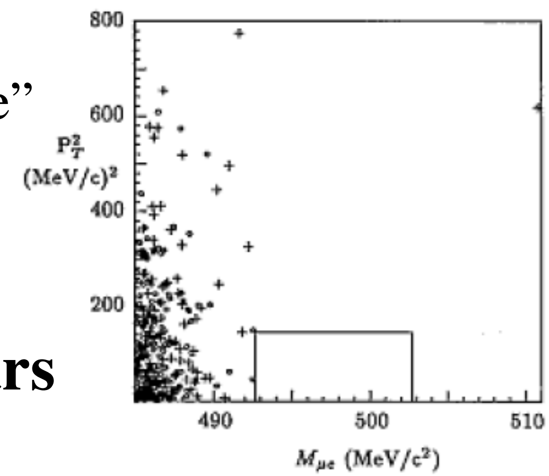
- **Simple carton illustrates a typical situation**
  - One is “exploring” the data
  - Finds a “cut” that miraculously reduces the background with high efficiency
  - But what is the right value of the cut?
- In some cases, it is not so clear
  - Experimenter can make an arbitrary choice
  - But behavioural psychologists claim there is no such thing!



A. Roodman,
ArXiv:0312102v1 (2003)

Avoiding Experimenter's Bias

- **A standard solution has been to formally "blind" the analysis**
  - Define *a priori* "signal region" or "measurable" that will not be looked at during analysis
  - Define a procedure for "opening the box"



- **Now been used in HEP for about 20 years**
  - Popularized by the BaBar collaboration
    - committed to using "blind techniques"
  - Goes back to 1662 by John Baptista von Helmont
    - Adopted in the biomedical community as the "gold standard" – double-blind studies – as far back as 1948

Too Good to be True?

- **Actually, works pretty well in practice**
  - Generally accepted as one strategy for reducing the bias
- **Some pitfalls/challenges:**
  - "Blinding" obscures an unanticipated instrumental or theoretical problem
    - Discover that half the data was missed (true example)!
  - After "opening the box", procedure changes because of ancillary studies or measurements
    - Current example in ATLAS is where
      - Box opened and 5 signal events
      - New "jet cleaning tool being implemented" – kills 1 event
      - 17% of background events also reduced, though 9 events in "sideband" all survived
      - Do you use the new "jet cleaning tool"?