Practical Statistics for Particle Physicists
Lecture 3

Harrison B. Prosper

Florida State University

European School of High-Energy Physics

Parádfürdő, Hungary

5 – 18 June, 2013

Outline

- Lecture 1
  - Descriptive Statistics
  - Probability & Likelihood
- Lecture 2
  - The Frequentist Approach
  - The Bayesian Approach
- Lecture 3
  - The Bayesian Approach
  - Analysis Examples

The Bayesian Approach

The Bayesian Approach – 1

Definition:

A method is Bayesian if

1. it is based on the ***degree of belief*** interpretation of probability and if
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

Nuisance parameters are removed by marginalization:

$$p(\theta \mid D) = \int p(\theta, \omega \mid D)\, d\omega$$
$$= \int p(D \mid \theta, \omega) \pi(\theta, \omega)\, d\omega \,/\, p(D)$$

in contrast to profiling, which can be viewed as marginalization with the (*data*-dependent) prior $\pi(\theta, \omega) = \delta[\omega - \hat{\omega}(\theta, D)]$

$$p(\theta \mid D) = \int p(D \mid \theta, \omega) \pi(\theta, \omega)\, d\omega \,/\, p(D)$$
$$= \int p(D \mid \theta, \omega) \delta(\omega - \hat{\omega})\, d\omega \,/\, p(D)$$
$$\simeq p(D \mid \theta, \hat{\omega}) \,/\, p(D)$$

The Bayesian Approach – 3

Bayes' theorem can be used to compute the probability of a model. First compute the posterior density:

$$p(\theta_H, \omega, H \mid D) = \frac{p(D \mid \theta_H, \omega, H)\,\pi(\theta_H, \omega, H)}{p(D)}$$

D	observed data
θ_H	parameters of model, or hypothesis, H
H	model or hypothesis
ω	nuisance parameters
π	prior density

The Bayesian Approach – 4

1. Factorize the priors: $\pi(\theta_H, \omega, H) = \pi(\theta_H, \omega \mid H)\, \pi(H)$

2. Then, for each model, H, compute the function

$$p(D \mid H) = \iint p(D \mid \theta_H, \omega, H)\, \pi(\theta_H, \omega \mid H)\, d\theta_H\, d\omega$$

3. Then, compute the probability of each model, H

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

The Bayesian Approach – 5

In order to compute ***p*(*H* |*D*)**, however, two things are needed:

1. Proper priors over the parameter spaces

$$\iint \pi(\theta_H, \omega \mid H)\, d\theta_H\, d\omega = 1$$

2. The priors ***π* (*H*)**.

In practice, we compute the Bayes factor:

$$\frac{p(H_1 \mid D)}{p(H_0 \mid D)} = \left[\frac{p(D \mid H_1)}{p(D \mid H_0)}\right]\left[\frac{\pi(H_1)}{\pi(H_0)}\right]$$

which is the ratio in the first bracket, B_{10}.

Examples

1. Top Quark Discovery
2. Search for Contact Interactions

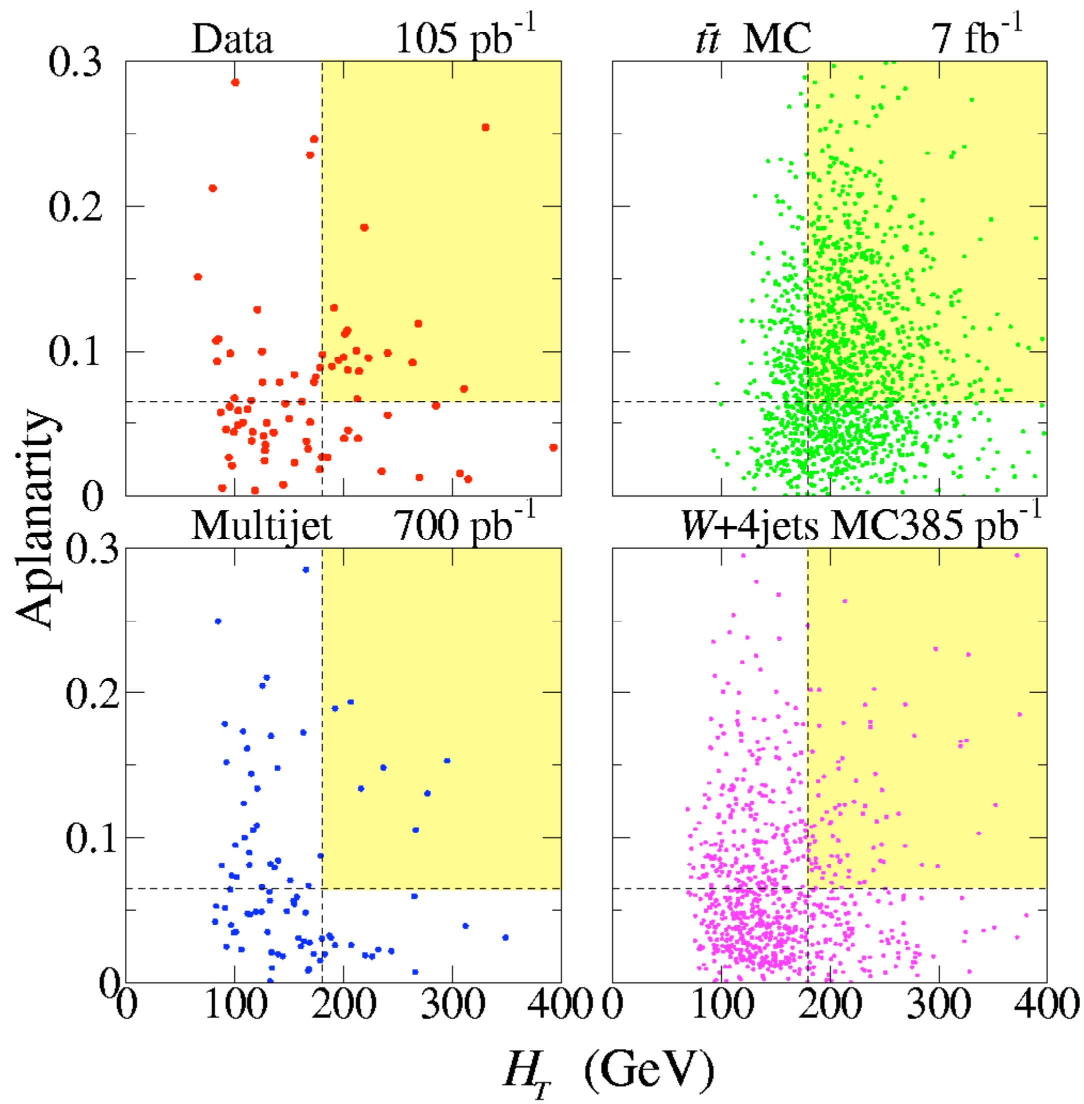
Example – Top Quark Discovery – 1

D0 1995 Top Discovery

Data

$D \quad = 17$ events

$B \quad = 3.8 \pm 0.6$ events



Example – Top Quark 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 put 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 – Top Quark 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 – Top Quark 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 – Top Quark Discovery – 5

For simplicity, we take $\pi(b \mid s) = 1$.

We may now eliminate b from the problem:

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

$$= \frac{1}{Q}(1-x)^2 \sum_{r=0}^{D} \mathrm{Beta}(x, r+1, Q)\, \mathrm{Poisson}(D-r \mid s)$$

Exercise 10: Show this

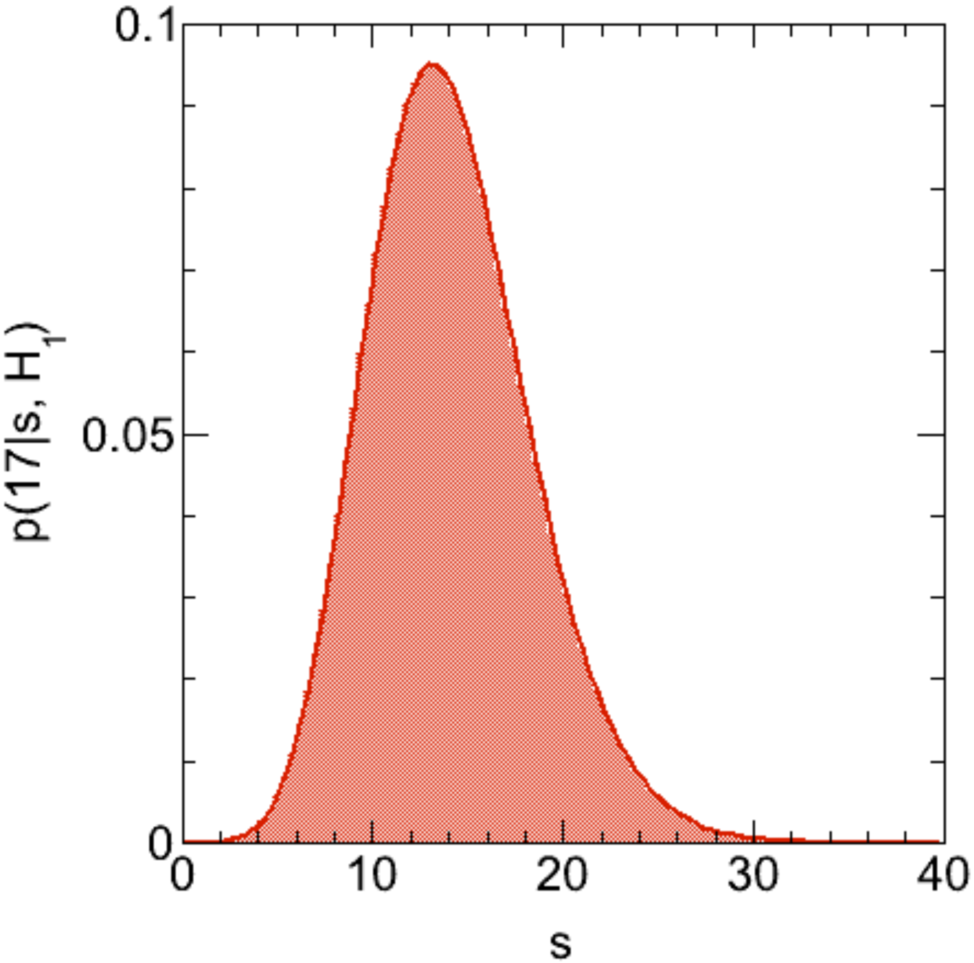
where,

$$x = \frac{1}{1+k}, \quad \mathrm{Beta}(x, n, m) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} x^{n-1}(1-x)^{m-1}$$

and where we have introduced the symbol $\boldsymbol{H_1}$ to denote the background + signal hypothesis.

Example – Top Quark Discovery – 6

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



Example – Top Quark Discovery – 7

Given the marginal likelihood

$$p(D \mid s, H_1)$$

we can compute the **the posterior density**

$$p(s \mid D, H_1) = \frac{p(D \mid s, H_1)\,\pi(s \mid H_1)}{p(D \mid H_1)}$$

and the **evidence** for hypothesis H_1

$$p(D \mid H_1) = \int_0^\infty p(D \mid s, H_1)\,\pi(s \mid H_1)\,ds$$

Example – Top Quark Discovery – 8

Assuming a ***flat prior*** for the signal $\pi(s \mid H_1) = 1$, the posterior density is given by

$$p(s \mid D, H_1) = \frac{\sum_{r=0}^{D} \text{Beta}(x, r+1, Q)\,\text{Poisson}(D-r \mid s)}{\sum_{r=0}^{D} \text{Beta}(x, r+1, Q)}$$

The posterior density of the parameter (or parameters) of interest is the ***complete*** answer to the inference problem and should be made available. Better still, publish the likelihood and the prior

Exercise 11: Derive an expression for $p(s \mid D, H_1)$ assuming a gamma prior Gamma$(qs, U+1)$ for $\pi(s \mid H_1)$

Example – $p(s \mid 17, H_1)$

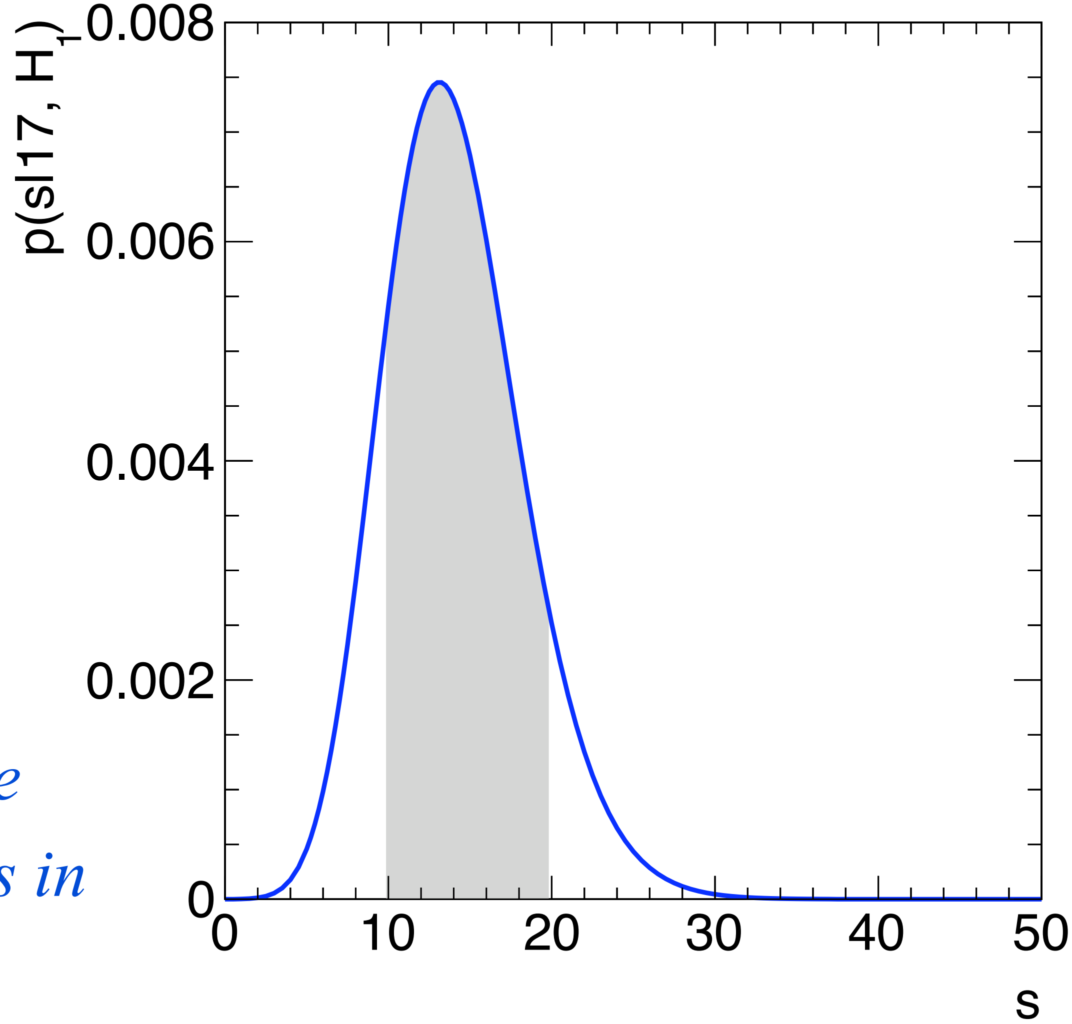
The current practice is to report summaries of the posterior density, such as

$$s \in [9.9, 19.8]$$

$$@ \; 95\% \text{ C.L.}$$

Note, since this is a Bayesian calculation, this statement means:

the probability (that is, the degree of belief) that s lies in [9.9, 19.8] *is* 0.95



Example – Top Quark Discovery – 9

As noted, the number

$$p(D \mid H_1) = \int_0^\infty p(D \mid s, H_1)\,\pi(s \mid H_1)\,ds$$

can be used to perform a hypothesis test. But, to do so, we need to specify a ***proper*** prior for the signal, that is, a prior $\pi(s| H_1)$ that integrates to one.

The simplest such prior is a δ-function, e.g.:

$\pi\,(s \mid H_1) = \delta(s - 14)$, which yields

$$p(D \mid H_1) = p(D \mid 14, H_1) \quad = \mathbf{9.28 \times 10^{-2}}$$

Example – Top Quark Discovery – 10

Since,

$$p(D \mid H_1) \quad = \mathbf{9.28 \times 10^{-2}} \text{ and}$$

$$p(D \mid H_0) \quad = \mathbf{3.86 \times 10^{-6}}$$

we conclude that the hypothesis s = 14 events is favored over the hypothesis s = 0 by **24,000** to 1.

To avoid big numbers, the Bayes factor can be mapped to a (signed) measure akin to "n-sigma" (Sezen Sekmen, HBP)

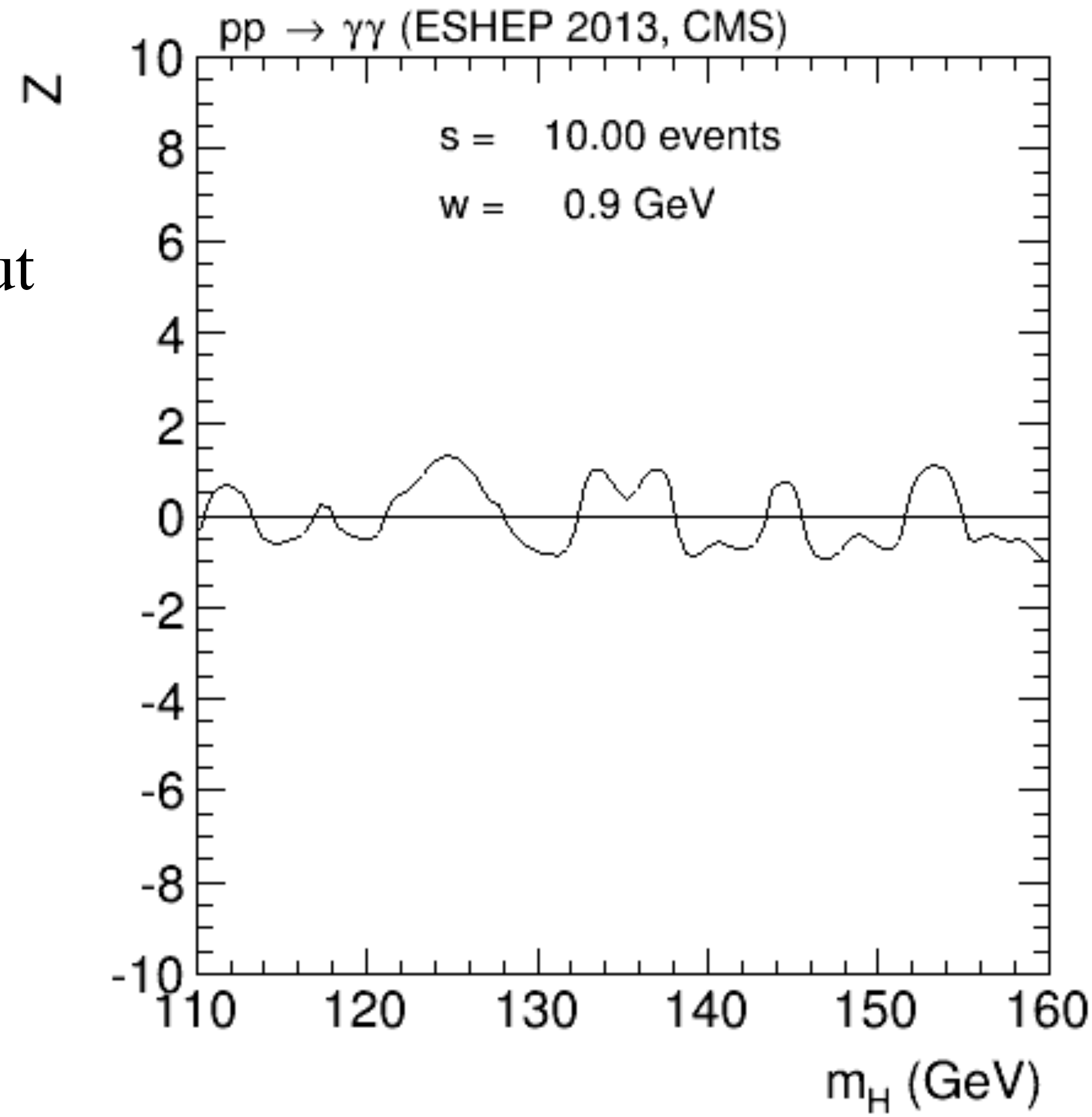
$$Z = \text{sign}(\ln B_{10})\sqrt{2\,|\ln B_{10}\,|} = 4.5, \quad B_{10} = p(D \mid H_1) \,/\, p(D \mid H_0)$$

$\boxed{\textbf{Exercise 12}\text{: Compute } Z \text{ for the D0 results}}$

Example – Z vs m_H for pp to $\gamma\gamma$ events

Here is a plot of $Z(m_H)$ as we scan through different hypotheses about the expected signal *s*.

The signal width and background parameters have been fixed to their maximum likelihood estimates

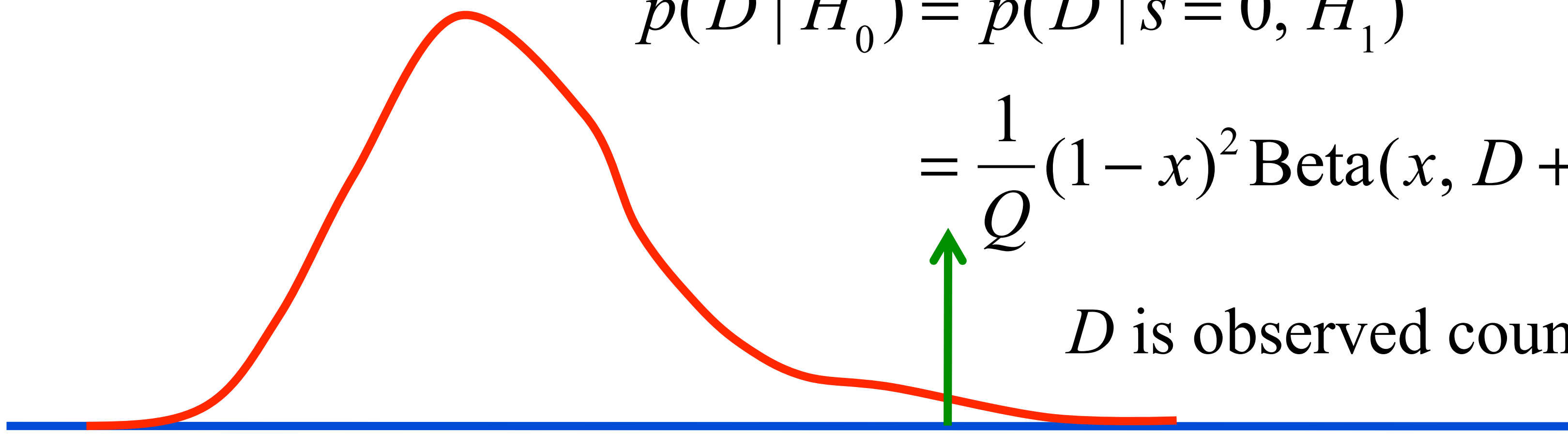


Hypothesis Testing – A Hybrid Approach

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

$$p(D \mid H_0) = p(D \mid s = 0, H_1)$$

$$= \frac{1}{Q}(1 - x)^2 \, \text{Beta}(x, D + 1, Q)$$



$$D = 17$$

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

This is equivalent to **4.4 σ** which may be compared with the **4.5 σ** obtained with B_{10}

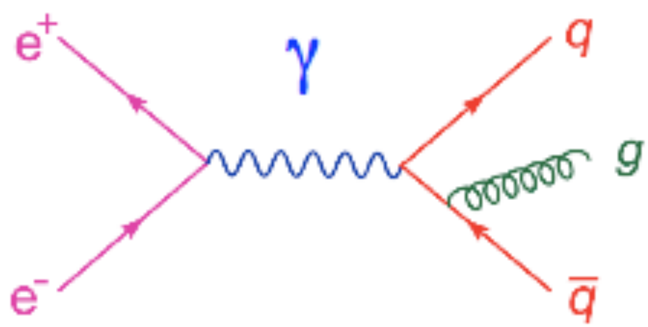
Exercise 13: Verify this calculation

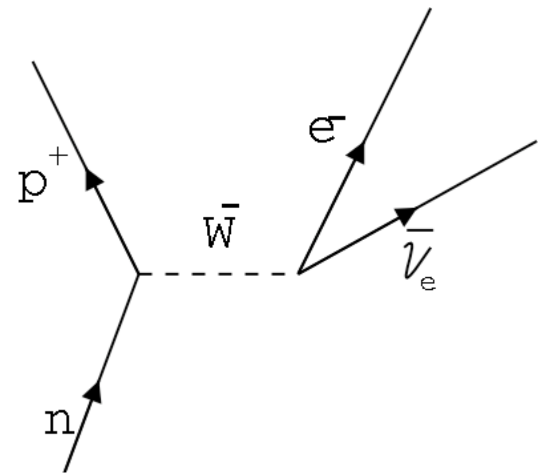
Example 2
CMS Search for Contact Interactions using Inclusive Jet Events

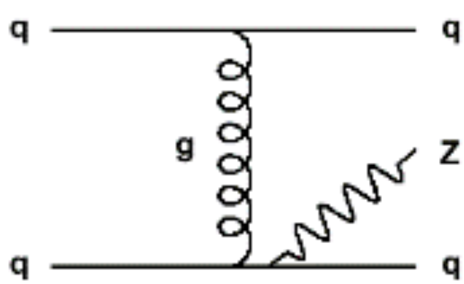
CMS Exotica/QCD Group
PhD work of Jeff Haas (FSU, PhD, 2013)

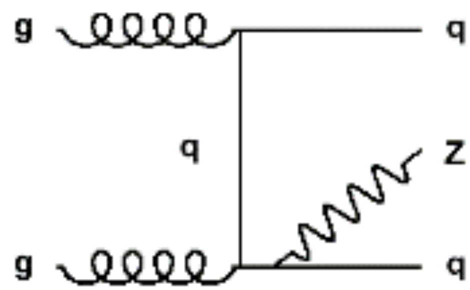
Contact Interactions – 1

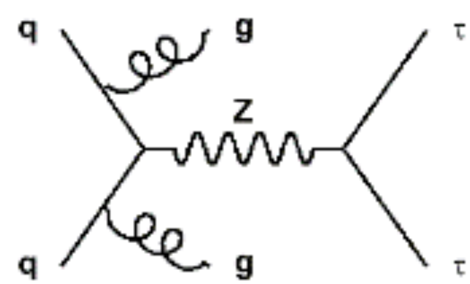
In our current theories, all interactions are said to arise through the exchange of bosons:









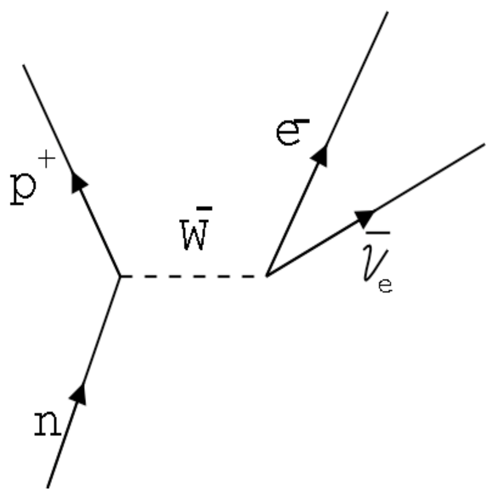


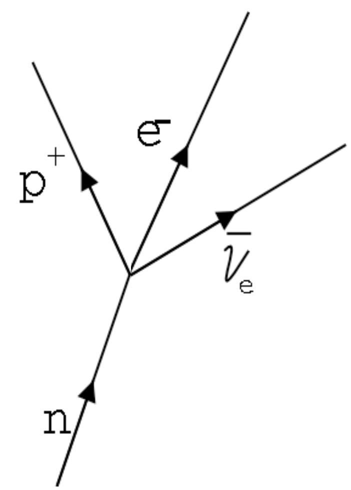
But,…

Contact Interactions – 2

… when the experimentally available energies are << than the mass of the exchanged particles, the interactions can be modeled as ***contact interactions*** (CI).

Here is the most famous example:





Contact Interactions – 3

The modern view of the Standard Model (SM) is that it is an *effective theory*: the low-energy limit of a more general (unknown) theory.

For the strong interactions, we assume that the Lagrangian of the unknown theory can be approximated as follows

$$L_{NEW} = L_{QCD} + 2\pi \lambda \sum_{i=1}^{6} \beta_i O_i + \cdots$$

where the O_i are a set of dim-6 operators, $\lambda = 1/\Lambda^2$ defines the scale of the new physics, and β_i are coefficients defined by the new theory.

Contact Interactions – 4

The CMS contact interaction analysis, using *inclusive* jet events, that is, events of the form

$$pp \to jet + X$$

where X can be any collection of particles, was a search for deviations from the prediction of QCD, calculated at next-to-leading order (NLO) accuracy

We searched for new QCD-like physics that can be modeled with a set of dim-6 operators of the form*

$$O_1 = (\bar{q}_L \gamma^\mu q_L)(\bar{q}_L \gamma_\mu q_L)$$

*Eichten, Hinchliffe, Lane, Quigg, Rev. Mod. Phys. **56**, 579 (1984)

Contact Interactions – 5

At NLO* the cross section per jet p_T bin can be written as

$$\sigma = c + \lambda[b - b'(\ln \mu_0 + \ln\sqrt{\lambda})] + \lambda^2[a - a'(\ln \mu_0 + \ln\sqrt{\lambda})]$$

where, c, b, a, b', a' are calculable and μ_0 is p_T–dependent scale. At leading order (LO) the primed terms vanish.

The 7 TeV CMS jet data, however, were analyzed using the model

$$\sigma = c + CI(\Lambda), \text{ where } CI(\Lambda) = b\lambda + a\lambda^2, \quad \lambda \equiv 1/\Lambda^2$$

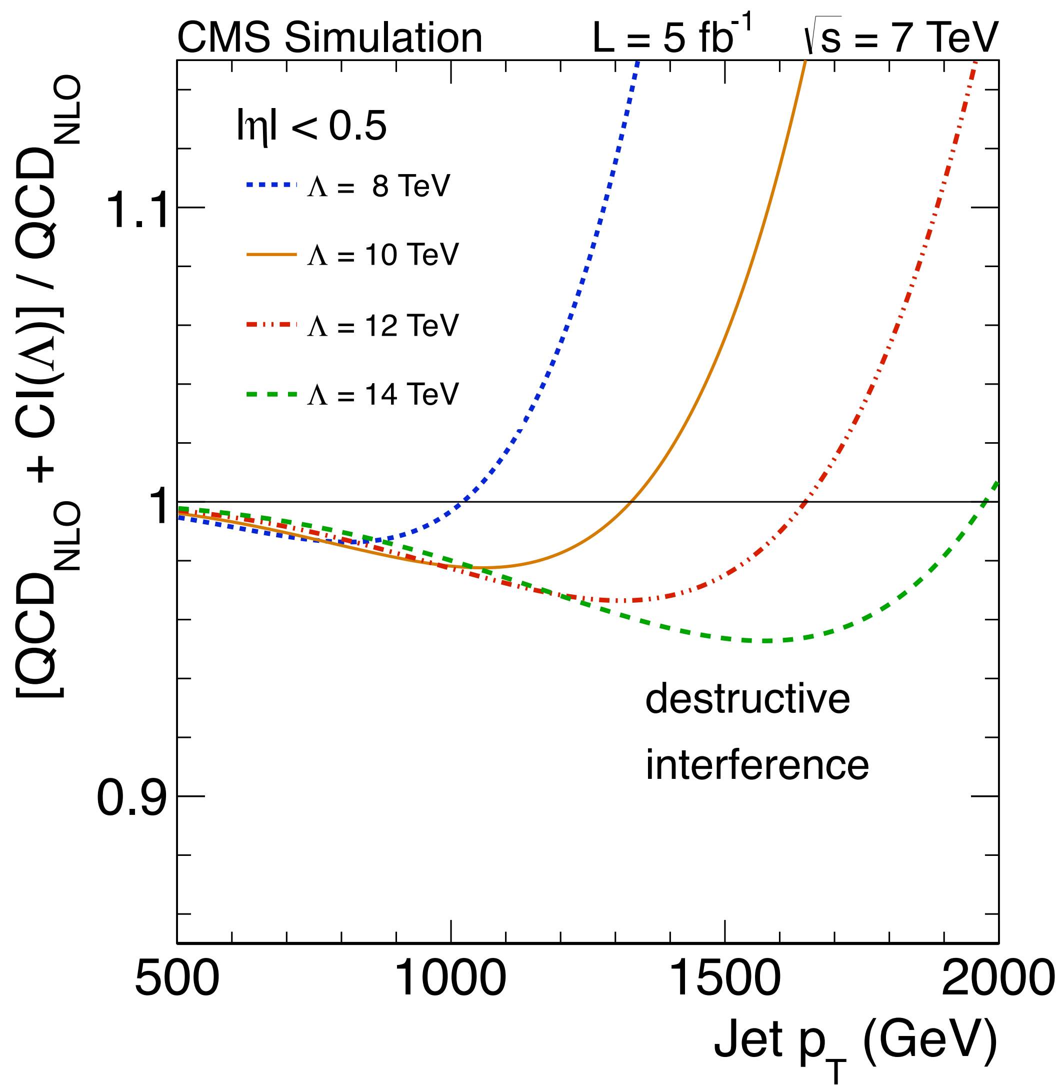
with c and $CI(\Lambda)$ computed at NLO at LO accuracy, respectively

*J. Gao, CIJET, arXiv:1301.7263

Contact Interactions – 6

The CI spectra were calculated with PYTHIA 6.422 and the QCD spectrum with fastNLO 2.1.0-1062.

This is an instructive example of physics in which the signal can be both positive and *negative*



Analysis

Inclusive Jet Data

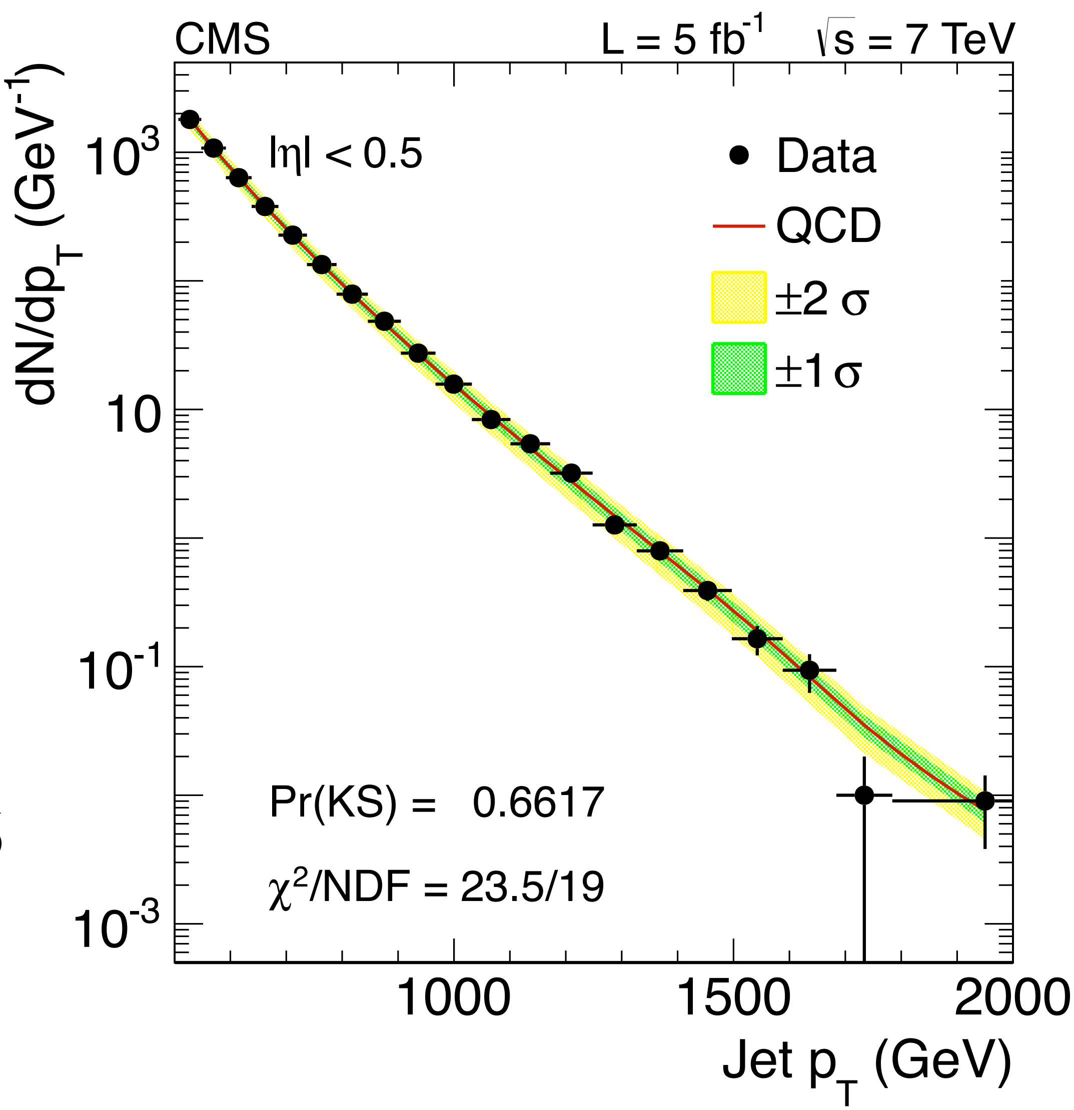
Data

$M = 20$ bins

$507 \leq p_T \leq 2116$ GeV

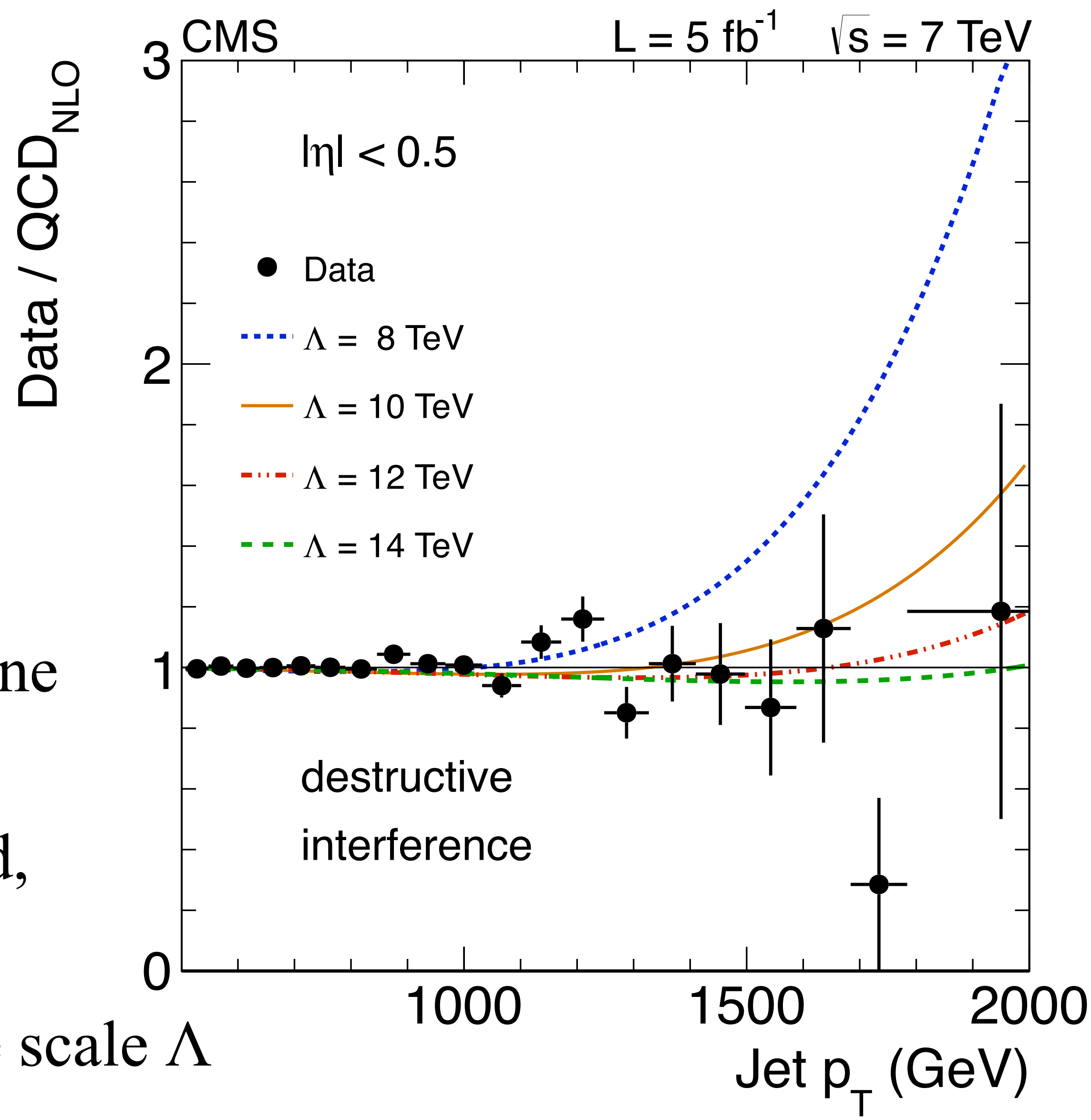
$D = 73{,}792$ to 3

The plot compares the observed dN/dp_T spectrum with the NLO QCD prediction (using CTEQ6.6 PDFs) convolved with the CMS jet response function



Analysis Goal

Data/QCD spectrum compared with (QCD+CI)/QCD spectra for several values of the scale Λ



Analysis Goal: Determine if there is a significant deviation from QCD and, if so, measure it; if not, set a lower bound on the scale Λ

Analysis – 1

First Attempt

Assume the following probability model for the observations

$$p(D \mid \lambda, \alpha, \nu) = \prod_{i=1}^{K} \text{Poisson}(N_i \mid \alpha \, \sigma_i)$$

where

$$\sigma_i = c_i + b_i \lambda + a_i \lambda^2$$

$$D = N_1, \cdots, N_K, \quad K = 20$$

$$\nu = c_1, b_1, a_1, \cdots, c_K, b_K, a_K$$

$$\alpha = \text{total count / total cross section}$$

Analysis Issues

1. Counts range from ~70,000 to 3! This causes the limits on Λ to be *very* sensitive to the normalization α. For example, increasing α by 1% decreases the limit by 25%!

2. Spectrum sensitive to the jet energy scale (JES)

3. And to the parton distribution functions (PDF)

4. Simulated CI models (using PYTHIA) were available for only 4 values of Λ, namely, Λ = 3, 5, 8, and 12 TeV for destructive interference models only

5. Insane deadlines and the need, occasionally, to sleep!

Analysis Issues

Solution: (Channel the Reverend Thomas Bayes!)

1. Integrate the likelihood over the scale factor α

2. Integrate the likelihood over the JES

3. Integrate the likelihood over the PDF parameters

4. Interpolate over the 4 PYTHIA CI models

5. Ignore insane deadlines and sleep as needed!☺

Analysis – 2

Step 1: Re-write

$$p(D\,|\,\lambda,\alpha,\nu) = \prod_{i=1}^{K} \text{Poisson}(N_i\,|\,\alpha\,\sigma_i)$$

as

$$p(D\,|\,\lambda,\alpha,\nu) = \text{Poisson}(N\,|\,\alpha\,\sigma)$$
$$\times\,\text{Multinomial}(N_1,\ldots,N_K\,|\,\theta_1,\ldots,\theta_K)$$

where

$$\theta_i = \sigma_i\,/\,\sigma, \quad \sigma = \sum \sigma_i, \;\; N = \sum N_i$$

Exercise 14: Show this

Analysis – 3

Step 2: Now eliminate α by integrating

$$p(D \mid \lambda, \alpha, \nu) = \text{Poisson}(N \mid \alpha \sigma) \times \text{Multinomial}(N_1, \ldots, N_K \mid \theta_1, \ldots, \theta_K)$$

with respect to α.

To do so, we need a prior density for α. In the absence of reliable information about this parameter, we use

$$\pi(\alpha \mid \lambda, \nu) = \sqrt{\sigma / \alpha}$$

which is an example of a **reference prior***.

*L. Demortier, S. Jain, HBP, arxiv:1002.1111 (2010)

Analysis – 4

Step 3: The integration with respect to α yields

$$p(D \mid \lambda, \nu) \propto \text{Multinomial}(N_1, \ldots, N_K \mid \theta_1, \ldots, \theta_K)$$

But, after more thought, we realized that almost all the information about the models is contained in the *shapes* of their jet p_T spectra, especially given that the total jet count is large (~200,000). This causes the multinomial to be particularly sensitive to the spectral shapes

Therefore, we could simply start with the multinomial and sidestep the normalization problem

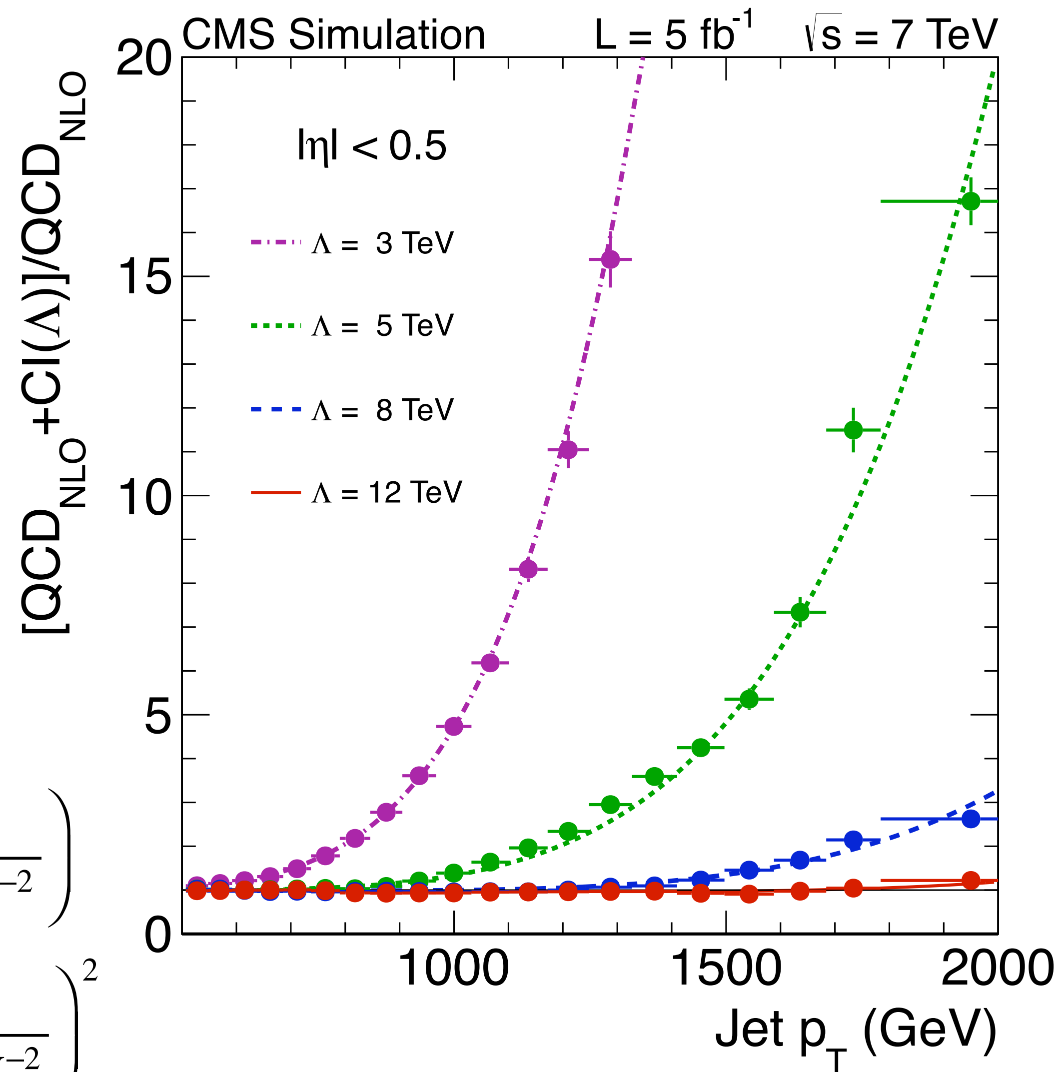
Analysis – 5

Step 4: Fit a 4-parameter interpolation function $\boldsymbol{f}$ to the four spectral ratios $(\mathrm{QCD}_{\mathrm{NLO}}+\mathrm{CI}_{\mathrm{LO}})/\mathrm{QCD}_{\mathrm{NLO}}$ simultaneously. The cross section (per p_{T} bin) is then modeled with

$\sigma = \boldsymbol{f}(\lambda, p_1, \ldots, p_4)\, \sigma_{\mathrm{QCD}}$

where

$$f = 1 + p_1 \left(\frac{p_T}{100}\right)^{p_2} \left(\frac{\lambda}{1\ \mathrm{TeV}^{-2}}\right) + p_3 \left(\frac{p_T}{100}\right)^{p_4} \left(\frac{\lambda}{1\ \mathrm{TeV}^{-2}}\right)^2$$



Analysis – 6

List of nuisance parameters ("systematics" in HEP jargon):

1. the jet energy scale (JES),
2. jet energy resolution (JER),
3. the PDF parameters (PDF),
4. the factorization an renormalization scales (μ_F, μ_R)
5. the parameters $\boldsymbol{\omega} = p_1 \ldots p_4$ of the function $\boldsymbol{f}(\lambda, \boldsymbol{\omega})$

Analysis – 7

Step 5: We use Bayes' theorem to calculate the posterior density of the parameter of interest λ,

$$
\begin{aligned}
p(\lambda \mid D) &= \int p(\lambda, \omega \mid D)\, d\omega \\
&= \int p(D \mid \lambda, \omega)\, \pi(\lambda, \omega)\, d\omega \,/\, p(D) \\
&= \pi(\lambda) \left[\int p(D \mid \lambda, \omega)\, \pi(\omega \mid \lambda)\, d\omega \right] / \, p(D)
\end{aligned}
$$

VIP (**V**ery **I**mportant **P**oint): *whatever the nature or provenance of nuisance parameters, whatever words we use to describe them, statistical, systematic, best guess, gut feeling…, in a Bayesian calculation we "simply" integrate them out of the problem.*

Analysis – 8

Bayesian Hierarchical Modeling

The parameters $\omega = p_1 \ldots p_4$ that appear in the likelihood depend on φ = JES, JER, PDFs, μ_F, and μ_R.

This fact can be modeled *hierarchically* as follows

$$p(\lambda \mid D) = p(D \mid \lambda)\pi(\lambda) / p(D)$$

where

$$p(D \mid \lambda) = \int p(D \mid \lambda, \omega)\, \pi(\omega \mid \lambda)\, d\omega \quad \text{and}$$

$$\pi(\omega \mid \lambda) = \int \pi(\omega \mid \lambda, \varphi)\, \pi(\varphi)\, d\varphi$$

and the density $\pi(\omega \mid \lambda, \varphi)$ models how ω depends on φ

Analysis – 9

$$p(D \mid \lambda) = \int p(D \mid \lambda, \omega)\, \pi(\omega \mid \lambda)\, d\omega$$

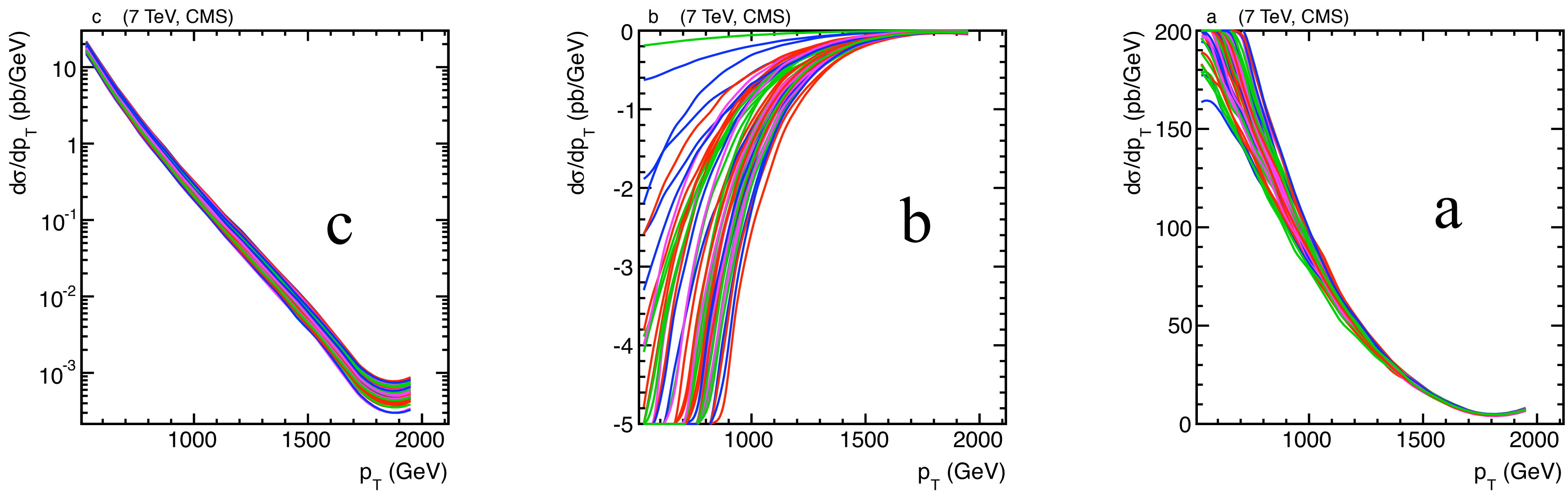
$$\approx \frac{1}{T} \sum_{i=1}^{T} p(D \mid \lambda, \omega_i)\,, \quad T \sim 500$$

Step 6: *Simultaneously* sample:

1. the jet energy scale,
2. the jet energy resolution,
3. the (CTEQ6.6) PDF parameters,
4. the factorization and renormalization scales

and, for each set of parameters, fit the parameters $\omega = p_1 \ldots p_4$ thereby creating points $\{\boldsymbol{\omega}_\mathrm{i}\}$ that constitute the prior $\pi(\omega \mid \lambda)$

Analysis – Ensemble of Coefficients



Ensemble of coefficients c, b, a, as a function of jet p_{T}, created by *simultaneous* sampling of all "systematics"

$$\sigma = c + b\lambda + a\lambda^2$$

Analysis – 10

Step 7: Finally, we compute a 95% Bayesian interval by solving

$$\int_0^{\lambda^{UP}} p(\lambda \mid D)\, d\lambda = 0.95$$

for λ^{UP}, from which we compute $\Lambda = 1/\sqrt{\lambda^{UP}}$.

The published limits were calculated for $\pi(\lambda) = 1$ and for $\pi(\lambda)$ = a reference prior* (and annoyingly, using CL_s):

$\Lambda > 10.1$ TeV or $\Lambda > 14.1$ TeV @ 95% C.L. for models with destructive or constructive interference, respectively

*L. Demortier, S. Jain, HBP, arxiv:1002.1111 (2010)

Summary – 1

Probability

Two main interpretations:

1. Degree of belief
2. Relative frequency

Likelihood Function

Main ingredient in any non-trivial statistical analysis

Frequentist Principle

Construct statements such that a fraction $p \geq \text{CL}$ of them will be true over a specified ensemble of statements.

Summary – 2

Frequentist Approach

1. Use likelihood function only
2. Eliminate nuisance parameters by profiling
3. Fisher: Reject null if p-value is judged to be small enough
4. Neyman: Decide on a fixed threshold α for rejection and reject null if p-value $< \alpha$, but do so only if the probability of the alternative is judged to be high enough

Bayesian Approach

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

The End

"Have the course to use your ***own*** understanding!"

Immanuel Kant