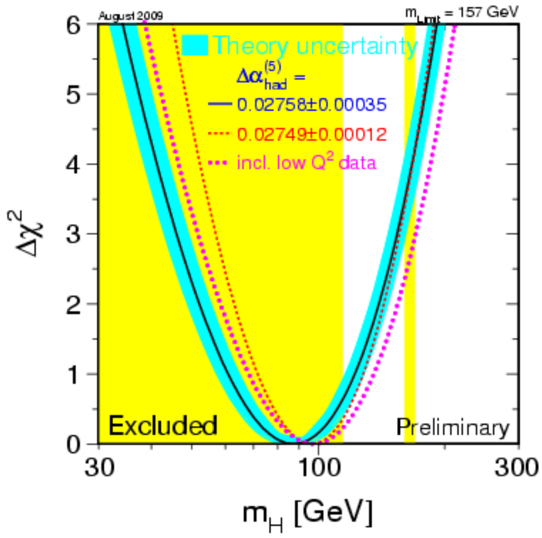
Statistics

Lent Term 2015
Prof. Mark Thomson



Lecture 4 : The Dark Arts

Course Synopsis

Lecture 1: The basics
Introduction, Probability distribution functions, Binomial distributions, Poisson distribution

Lecture 2: Treatment of Gaussian Errors
The central limit theorem, Gaussian errors, Error propagation, Combination of measurements, Multi-dimensional Gaussian errors, Error Matrix

Lecture 3: Fitting and Hypothesis Testing
The χ^2 test, Likelihood functions, Fitting, Binned maximum likelihood, Unbinned maximum likelihood

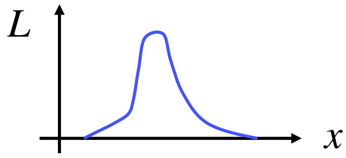
Lecture 4: The Dark Arts
Bayesian Inference, Credible Intervals
The Frequentist approach, Confidence Intervals
Systematic Uncertainties

Parameter Estimation Revisited

★ Let's consider more carefully the maximum likelihood method
for simplicity consider a single parameter x

★ Construct the likelihood that our data are consistent with the model, i.e.
the probability that the model would give the observed data

$$L = P(\mathrm{data}; x)$$



★ We have then (very reasonably) taken the value of x which maximises the likelihood as our best estimate of the parameter

★ With less justification we then took our error *estimate* from

$$-\ln L \rightarrow -\ln L + \tfrac{1}{2}$$

★ Does this really make sense ?

★ What we really want to calculate is the posterior PDF for the parameter given the data, i.e. $P(x; \mathrm{data})$

"assumed" $P(x; \mathrm{data}) = P(\mathrm{data}; x)$

Can not justify this – in general it is not the case

Conditional Probabilities and Bayes' Theory

★ A nice example of conditional probability (from L. Lyons)

- In the general population, the probability of a randomly selected woman being pregnant is 2%

$$P(\mathrm{pregnant}; \mathrm{woman}) = 0.02$$

- But $$P(\mathrm{woman}; \mathrm{pregnant}) \gg 0.02$$

★ Correct treatment of conditional probabilities requires Bayes' theorem

- Probability of A and B can be expressed in terms of conditional probabilities

$$P(AB) = P(A;B)P(B) = P(B;A)P(A)$$

$$\boxed{P(A;B) = \frac{P(B;A)P(A)}{P(B)}}$$

★ Here the prior probability of selecting a woman is

$$P(\mathrm{woman}) = 0.50$$ i.e. half population are women

and the prior probability of selecting a pregnant person is

$$P(\mathrm{pregnant}) = 0.01$$ i.e. 1 % of population are pregnant

$$P(\mathrm{woman}; \mathrm{pregnant}) = \frac{P(\mathrm{pregnant}; \mathrm{woman})P(\mathrm{woman})}{P(\mathrm{pregnant})} = \frac{0.02 \times 0.5}{0.01} = 1$$

Sanity restored…

★ Apply Bayes' theory to our the measurement of a parameter x

- We determine $P(\text{data}; x)$, i.e. the likelihood function
- We want $P(x; \text{data})$, i.e. the PDF for x in the light of the data
- Bayes' theory gives:

$$\boxed{P(x; \text{data}) = \frac{P(\text{data}; x)P(x)}{P(\text{data})}}$$

$P(\text{data}; x)$ the likelihood function, i.e. what we measure

$P(x; \text{data})$ the posterior PDF for x, i.e. in the light of the data

$P(\text{data})$ { prior probability of the data. Since this doesn't depend on x it is essentially a normalisation constant

$\boxed{P(x)}$ { prior probability of x, i.e. encompassing our knowledge of x before the measurement

★ Bayes' theory tells us how to modify our knowledge of x in the light of new data

Bayes' theory is the formal basis of Statistical Inference

Applying Bayes' Theorem

★ Bayes' theory provides an unambiguous prescription for going from

$$P(\text{data}; x) \rightarrow P(x; \text{data})$$

★ But you need to provide the PRIOR PROBABILITY $P(x)$

★ This is fine if you have an objective prior, e.g. a previous measurement

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x - x_1)^2}{2\sigma_1^2}\right\}$$

- If we now make a new measurement, i.e. determine the likelihood function

$$P(\text{data}; x) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(x - x_2)^2}{2\sigma_2^2}\right\}$$

- Bayes' theory then gives

$$P(x; \text{data}) = \frac{P(\text{data}; x)P(x)}{P(\text{data})} = \frac{1}{P(\text{data})} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{(x - x_1)^2}{2\sigma_1^2} - \frac{(x - x_2)^2}{2\sigma_2^2}\right\}$$

$$P(x; \text{data}) = \frac{1}{P(\text{data})} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{(x - \overline{x})^2}{2\sigma^2}\right\}$$

Where $\overline{x}$ and σ are the usual mean and variance for combining two measurements

- For this to be a (normalised) PDF can infer (although it isn't of any interest):

$$P(\text{data}) = [2\pi(\sigma_1^2 + \sigma_2^2)]^{-\frac{1}{2}}$$

The Problem with Applying Bayes' Theorem

★ The problem arises when there is no objective prior

★ For example, in a hypothetical background free search for a Z', observe no events

- No problem in calculating the likelihood function (a conditional probability)

$$P(\text{data};x) = P(0;x) = e^{-x}$$

Poisson prob. for observing 0

x is the true number of expected events

- What is the best estimate of x and the **90 %** "confidence level upper limit" ?
- Depends on the choice of prior probability:

$$P(x;\text{data}) = P(x)e^{-x}$$

- What to do about the prior ?
- i.e. how do we express our knowledge (none) of x prior to the measurement

★ In general there is no objective answer, always putting in some extra information
- i.e. a subjective bias
- could argue that a flat prior, i.e. ***P(x)*** **= constant**, is objective
- but why not choose a prior that is flat in **ln** x ?
- for some limits/measurements (e.g. a mass) a flat prior in **ln** x is more natural
- the arbitrariness in the choice of prior is a problem for the Bayesian approach
- it can make a big difference…

Choice of Prior, example I

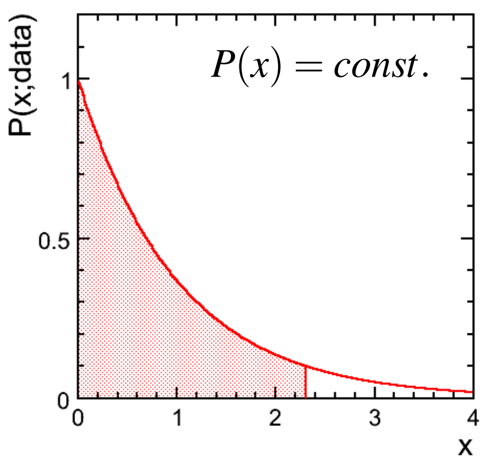
★ See no events…

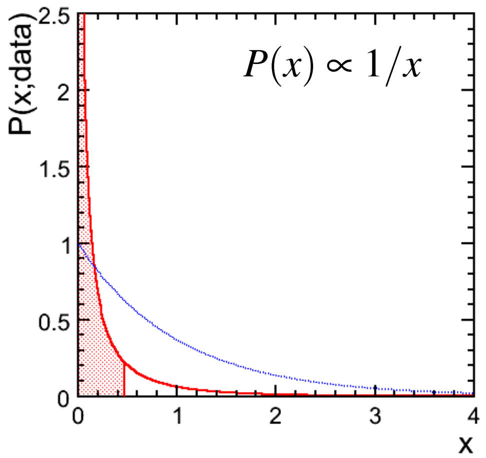
$$P(\text{data};x) = P(0;x) = e^{-x}$$

Poisson prob. for observing 0

Prior flat prior in x : $P(x) = const.$

Prior flat prior in lnx : $P(\ln x) = const.$





★ The Conclusions are very different. Compare regions containing **90 %** of probability

$x < 2.3$ $\qquad$ $x < 0.46$

- In this case, the choice of prior is important

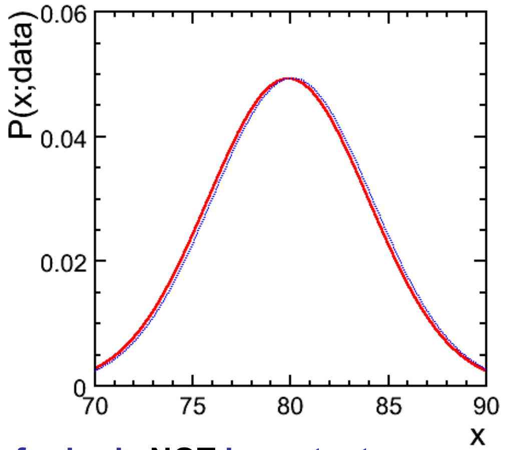
Choice of Prior, example II

★ **Suppose we measure the W-boson mass:** $80.1 \pm 4.1\,\mathrm{GeV}$

$$P(\mathrm{data};m) = G(80.1;m) \propto \exp\left\{-\frac{(80.1-m)^2}{2\times 4.1^2}\right\}$$

★ **We want** $P(m;\mathrm{data}) = P(m)P(\mathrm{data};x)$

- **Again consider two priors**



$P(x) = const.$

$P(\ln x) = const.$

- **Here the choice of prior is NOT important**
- **The data are "strong enough" to overcome our prior assumptions (subjective bias)**
- **Here, can interpret the measurement as a Gaussian PDF for *m***

Choice of Prior, example III

♦ **An example (apparently due to Newton), e.g. see CERN Yellow Report 2000-005**

★ **Suppose you are in the Tower of London facing execution.**

★ **The Queen arrives carrying a small bag and says**

"This bag contains 5 balls; the balls are either white or black. If you correctly guess the number of black balls, I will spare your life and set you free."

★ **The Queen is in a good mood and continues**

"To give you a better chance, you can take one of the balls from the bag."

It's BLACK

★ **The Queen points her pistol at you**

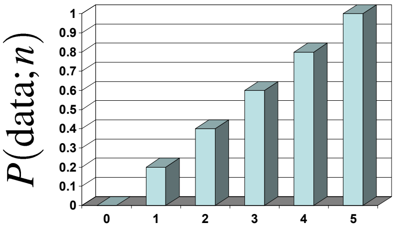
"Time to choose, sucker…"

★ **What do you guess to maximise your chance of survival ?**

★ **Use statistical inference to analyse the problem.**

- **Let *n* be the number of black balls in the bag.**
- **The data are "that you picked out a black ball"**
- **Can calculate** $P(data;n)$

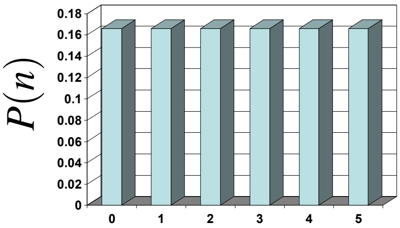
  e.g. if there were two black balls chance of picking out a black ball from the five in the bag was 2/5.

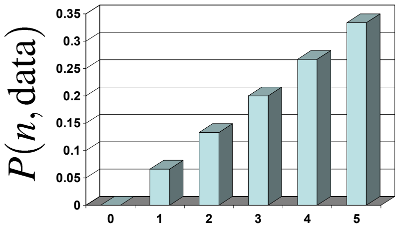


★ But we want $P(n;\text{data}) \propto P(n)P(\text{data};n)$

★ Answer depends on choice of Prior

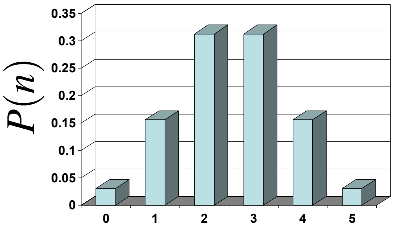
★ Could assume **flat Prior**

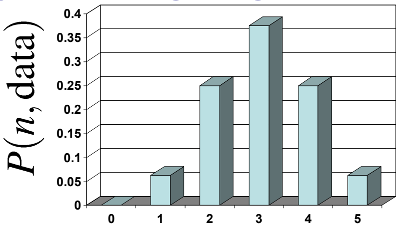




GUESS: 5

★ Could assume balls drawn randomly from a large bag containing equal nos. B & W





GUESS: 3

★ Oh dear… answer depends on Prior (unknown) assumptions

★ So what do we learn from this ?

(apart something about the role of the Monarchy in a modern democracy)

- Whilst we know how to apply Bayesian statistical inference, we have insufficient data, i.e. **we don't know the prior**
- Unless the data are "strong", i.e. override the information in the reasonable range of prior probabilities, we cannot expect to know

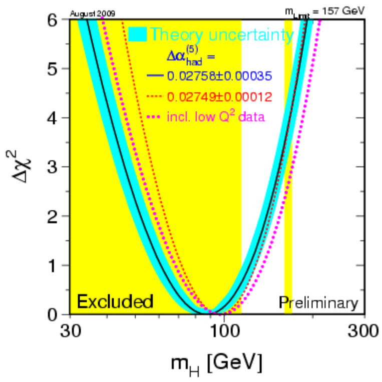
$$\boxed{P(x;\text{data})}$$

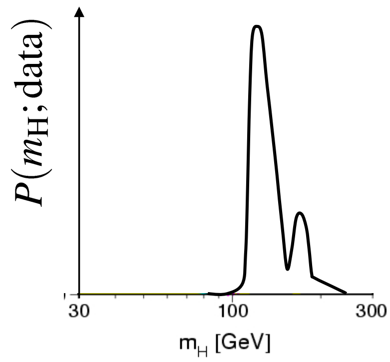
- Applies equally to our experiment where we saw zero events and wanted to arrive at a PDF for the expected mean number of events…

Don't have enough information to answer this question

Bayesian Credible Intervals

★ Ideally, (I) would like to work with probabilities, i.e. a PDF which encompasses all our knowledge of a particular parameter, e.g. $P(m_{\rm H};\mathrm{data})$





★ Could then integrate PDF to contain 95 % of probability. Can then define the "95 % Credible Interval*: $m_{\rm H}$ < 186 GeV"

★ To do this need to go from $P(\mathrm{data};m_{\rm H})$, i.e. from $\Delta\ln\mathscr{L}$, to $P(m_{\rm H};\mathrm{data})$
- requires "subjective" choice of prior probability

★ Hence Bayesian Credible Intervals necessarily include some additional input beyond the data alone…

*This is not what is done.

Bayesian Credible Intervals - example

★ Trying to estimate a selection efficiency using MC events. All **N** events pass cuts.
- what statement can we make about the efficiency?

★ Binomial distribution…

$$\boxed{P(\mathrm{data};x)} \rightarrow P(N;\varepsilon) \;=\; {}^{N}C_{N}\varepsilon^{N}(1-\varepsilon)^{0} = \varepsilon^{N}$$

★ Apply Bayes' theorem:

$$\boxed{P(x;\mathrm{data})} \rightarrow P(\varepsilon;N) = \frac{P(N;\varepsilon)P(\varepsilon)}{P(N)}$$

Prior

Constant

★ Choose prior, e.g. $P(\varepsilon)=1$

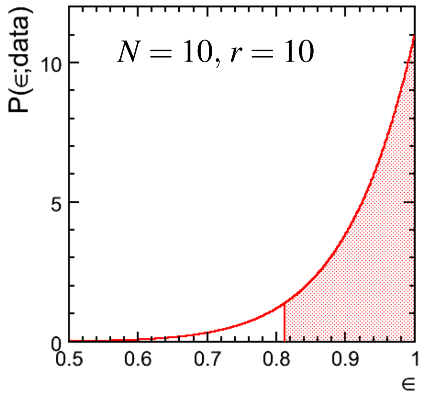
$$P(\varepsilon;N) = \kappa\varepsilon^{N}$$

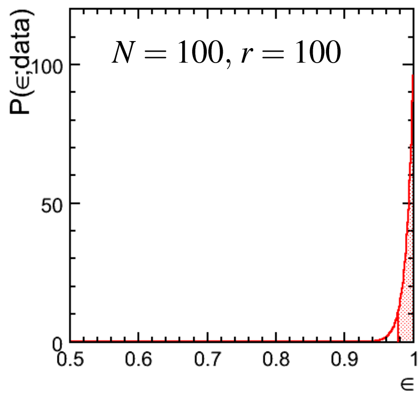
★ Normalise $\int_0^1 P(\varepsilon;N)d\varepsilon = 1 \;\Rightarrow\; \kappa = (N+1)$

$$\boxed{P(\varepsilon;N) = (N+1)\varepsilon^{N}}$$

★Integrate $\boxed{P(\varepsilon;N) = (N+1)\varepsilon^N}$ to find region containing 90% of probability

$$\Rightarrow \quad \varepsilon_{90\%} = (1-0.90)^{\frac{1}{N+1}}$$



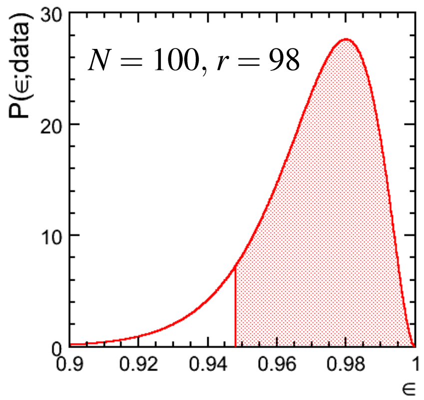


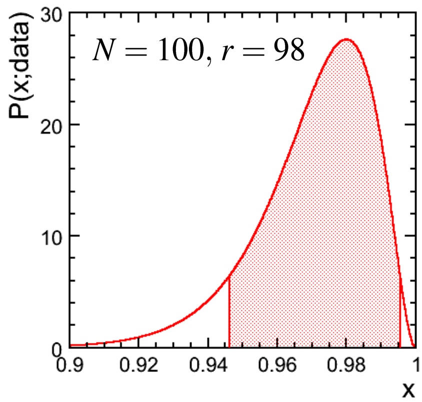
90 % Credible Interval: $\varepsilon > 0.81$ $\quad\quad$ $\varepsilon > 0.977$

(with a flat prior probability)

Likelihood Ordering

★ Note, 90 % credible interval is not uniquely defined
- more than one interval contains 90 % probability, e.g. $N=100,\, r=98$





90 % Credible Interval: $\varepsilon > 0.9482$ $\quad\quad$ $0.9456 < \varepsilon < 0.9956$

★ Natural, to choose the interval such that all points in the excluded region are lower in likelihood than those in the credible interval : likelihood ordering

★ **Credible intervals** provide an intuitive way of interpreting data, **but**:
- Rarely used in Particle Physics as a way of presenting data
- Because they represent the "data" and "prior" combined
- **NOTE:** all information from the experiment is in the likelihood $P(\mathrm{data};x)$

C.I. vs C.L.

★ From data obtain $P(\text{data};x)$

★ Bayes' theorem provides the mathematical framework for statistical inference

★ To go from $P(\text{data};x) \rightarrow P(x;\text{data})$ requires a (usually) subjective choice of Prior probability

★ For "weak" data, the choice of Prior can drive the interpretation of the data

★ Credible intervals are a useful way of interpreting data, but are generally not used in Particle Physics as a way of presenting the conclusions of an experiment.

★ Particle Physics to use Frequentist "Confidence limits" which are not $P(x;\text{data})$ [and do not form a mathematically consistent basis for statistical inference]

★ Finally, never forget that credible intervals (or confidence limits) are an interpretation of the data

The experimental result is the likelihood function $P(\text{data};x)$

A Few words on Systematic Uncertainties

★ Systematic Uncertainties are often associated with an internal unknown bias, e.g.
- How well do you know your calibration
- How well does MC model the data, e.g. jet fragmentation parameters

★ Parametric Uncertainties associated with uncertain parameters
- How does the uncertainty on the Higgs mass impact the interpretation of a a measurement

★ No over-riding principle – just some general guidelines
- Once a result is published, systematic errors will be treated as if they are Gaussian

$$x = a \pm b\ (stat.) \pm c\ (syst.)$$

- Some systematic errors are Gaussian: e.g. energy scale determined from data e.g. $Z \rightarrow e^+e^-$ to determine electron energy scale
- Others are not: e.g. impact of different jet hadronisation models, where one might compare PYTHIA with HERWIG – here one obtains a single estimate of the scale of the uncertainties
- Theoretical uncertainties: e.g. missing HO corrections. Again these are estimates – should not be treated as Gaussian (although they are)

★ Systematic dominated measurements
- Beware – if there is a single dominating systematic error and it is inherently non-Gaussian, this is a problem

Estimating Systematic Uncertainties

★ **No rules – just guidelines**

- **Remember syst. errors will be treated as Gaussian, so try to evaluate them on this basis, e.g. suppose use 3 alternative MC jet fragmentation models and result changes by $+\Delta_1$, $+\Delta_2$ and $-\Delta_3$ (where Δ_2 is the largest):**
  - **i) take largest shift as systematic error estimate: Δ_2 ?**
  - **ii) assume error distributed uniformly in "box" of width $2\Delta_2$ giving an rms of $2\Delta_2/\sqrt{12}$?**

- **Cut variation is evil (i.e. vary cuts and see how results change)**
  - **at best, introduces statistical noise**
  - **at worst, hides away lack of understanding of some data - MC discrepancy**
    - ⇨ **understand the origin of the discrepancy**

- **Wherever possible use data driven estimates, energy scales, control samples, etc.**

- **Remember that you are estimating the scale of a possible systematic bias**

Incorporating Systematics into Fits

★ **Two commonly used approaches**

- **Error matrix – with (correlated) systematic uncertainties**
- **Nuisance parameters**

★ **Nuisance parameter example:**

- **Suppose we are looking at WW decays and count numbers of events in three different decay channels** qqqq, qqlν and lνlν
- **Want to measure cross section and hadronic branching fractions accounting for common luminosity uncertainty**
  - **i) build physics model**

$$N^{\rm exp}_{\rm qqqq}(\sigma_{\rm WW}, B_{\rm qq}, \mathcal{L}) = \frac{\sigma_{\rm WW}}{\varepsilon\mathcal{L}} B^2_{\rm qq}$$

  - **ii) build likelihood function**

$$\chi^2(\sigma_{\rm WW}, B_{\rm qq}, \mathcal{L}) = -2\ln L = \frac{(N^{\rm exp}_{\rm qqqq} - N^{\rm obs}_{\rm qqqq})^2}{N^{\rm exp}_{\rm qqqq}} + \frac{(N^{\rm exp}_{\rm qqlv} - N^{\rm obs}_{\rm qqlv})^2}{N^{\rm exp}_{\rm qqlv}} + \frac{(N^{\rm exp}_{\rm lvlv} - N^{\rm obs}_{\rm lvlv})^2}{N^{\rm exp}_{\rm lvlv}}$$

  - **iii) add penalty term for nuisance parameters, here integrated lumi. Known to be L_0 with uncertainty σ_L**

$$\chi^2(\sigma_{\rm WW}, B_{\rm qq}, \mathcal{L}) = -2\ln L = \frac{(N^{\rm exp}_{\rm qqqq} - N^{\rm obs}_{\rm qqqq})^2}{N^{\rm exp}_{\rm qqqq}} + \frac{(N^{\rm exp}_{\rm qqlv} - N^{\rm obs}_{\rm qqlv})^2}{N^{\rm exp}_{\rm qqlv}} + \frac{(N^{\rm exp}_{\rm lvlv} - N^{\rm obs}_{\rm lvlv})^2}{N^{\rm exp}_{\rm lvlv}} + \boxed{\frac{(\mathcal{L}-\mathcal{L}_0)^2}{\sigma^2_{\mathcal{L}}}}$$

Incorporating Systematics into Fits

★ **Let's consider this more closely**

$$\chi^2(\sigma_{\mathrm{WW}}, B_{\mathrm{qq}}, \mathcal{L}) = -2\ln L = \frac{(N^{\mathrm{exp}}_{\mathrm{qqqq}} - N^{\mathrm{obs}}_{\mathrm{qqqq}})^2}{N^{\mathrm{exp}}_{\mathrm{qqqq}}} + \ldots \frac{(N^{\mathrm{exp}}_{\mathrm{qql\nu}} - N^{\mathrm{obs}}_{\mathrm{qql\nu}})^2}{N^{\mathrm{exp}}_{\mathrm{qql\nu}}} + \frac{(\mathcal{L} - \mathcal{L}_0)^2}{\sigma^2_{\mathcal{L}}}$$

- **We are now fitting 3 parameters**
  - the number of degrees of freedom has not changed, since we have added one parameter, but also one additional "data point"
- **Of the 3 parameters, we are "not interested" in the fitted value of the lumi.**
- **The penalty term constrains the luminosity to be consistent with the externally measured value**
- **The presence of the nuisance parameters will flatten the fitted likelihood surface – increasing the uncertainties on the fitted parameters**
- **Also have some measure of the tension in the fit**
  - if the data pull the nuisance parameter away from the expected value, could indicate a problem

That's All Folks