PREFIT-School

PRecision Effective Field Theory School

2-13 March 2020 at DESY, Hamburg

Fitting: Session 1

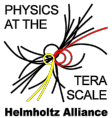
Andrea C. Marini, Nick Wardle

Assistants:

Lydia Brenner

Sebastian Wuchterl

Adinda de Wit



Introduction

There are many reasons for “fitting” in HEP

- Several test-statistics we use for discovery/limit setting involve one or more fits (eg. likelihood minimization)
- Training a ML algorithm requires minimising a loss function with some training data (eg fitting BDT weights)
- **Measuring** particle properties/model parameters (eg the parameters of an EFT)
  - → this is one we will focus on in these lectures!

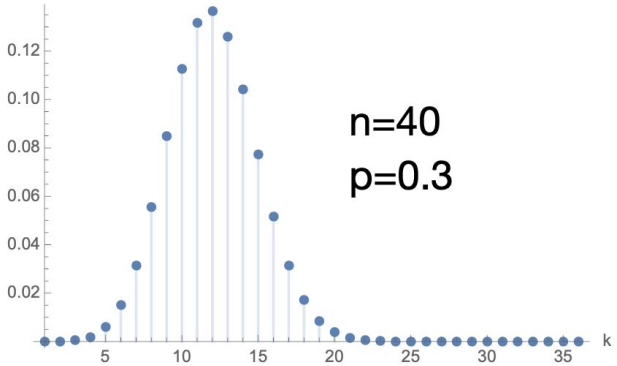
In this first part, we will introduce the concept of a measurement and some of the key statistical concepts before diving into the hands-on part.

You will recognize some of this from Kyle’s ML lectures (but it never hurts to refresh!)

Binomial distribution

$$\mathcal{P}(k|p,n) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$\mathbb{E}[k] = np$$
$$\mathbb{V}[k] = np(1-p)$$

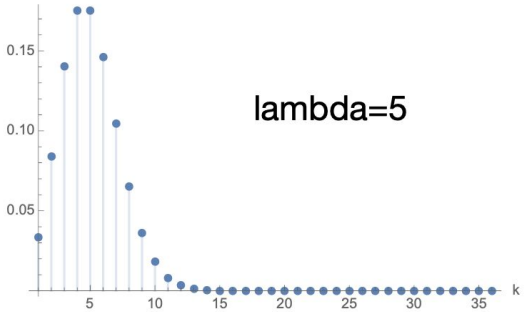


Poisson distribution

Binomial: $np = \lambda, n \gg 1$

$$\mathcal{P}(k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$\mathbb{E}[k] = \lambda$$
$$\mathbb{V}[k] = \lambda$$



lambda=5

The key to all HEP measurements!

Can you prove that the binomial → Poisson when n → infinity (during the hands on)?

- In HEP histograms usually follows Poisson distribution
  - n ~ number of collisions
  - p ~ cross-section (rare processes)

Likelihoods & Measurements

$$\mathcal{L}(\vec{\alpha}) \sim p(data|\vec{\alpha})$$

$\vec{\alpha}$ are the **parameters** of the likelihood

$p(data|\vec{\alpha})$ is the **probability to observe the data**, for a particular value of $\vec{\alpha}$

Likelihoods & Measurements

$$\mathcal{L}(\vec{\alpha}) \sim p(data|\vec{\alpha})$$

It's worth noting two things:

- The likelihood function is defined by **fixing the data** - that is, it takes different values for different outcomes
- The likelihood is **not a probability** - (e.g. there are various normalisation terms which we ignore for our purposes)

Likelihoods & Measurements

Often we want to separate parameters which are physics parameters of interest (POI = $\vec{\mu}$) vs uninteresting parameters (NP = $\vec{\theta}$)

$$\vec{\alpha} = \left(\vec{\mu}, \vec{\theta}\right)$$

Typically (though certainly not always!) the nuisance parameters are constrained by some external measurements (eg Jet energy scales) - we introduce ***constraint terms***

$$\pi(\vec{\theta}_0|\vec{\theta}) \sim p(\vec{\theta}|\vec{\theta}_0)$$

where again, we introduce some observed (or measured) values $\vec{\theta}_0$ to relate π to the probability to observe that outcome some value of the NPs

Likelihoods & Measurements

So then we have

$$\mathcal{L}(\vec{\mu}, \vec{\theta}) \sim p(data|\vec{\mu}, \vec{\theta}) \cdot \pi(\vec{\theta}_0|\vec{\theta})$$

Let's look at a very simple setup with 1 parameter of interest μ, and one nuisance parameter θ

Imagine we had an analysis which, after making some cuts, just counts the number of events left over in pp collisions (a ***cut-and-count*** analysis).

Our "*data*" in this case is just the observed number of events n, and the probability term is just a ***Poisson probability***

$$p(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

Likelihoods & Measurements

Often, we want to determine some total cross-section for production of these events which populate our selection. Typically we have some reference theory (eg the SM) in mind which predicts a value σ for this cross-section. Our POI then would be the cross-section, relative to that prediction - i.e

$$\lambda = \mu \cdot \sigma(pp \to X) \cdot \varepsilon \cdot A \cdot L$$

$\mathcal{E}$ - The efficiency of our selection

A - The acceptance of the detector

L - The integrated luminosity of our dataset

Likelihoods & Measurements

L - The integrated luminosity of our dataset

Any of these terms could have “uncertainties” associated to them. We can model this uncertainty by introducing nuisance parameters Eg.

Say, the luminosity is known to 10%*. We could think of this as saying the rate of events could increase by 10% (x1.1) or decrease by 10% (1/1.1)

$$L \rightarrow L(1+0.1)^{\theta}$$

When θ=0, we recover the nominal value
We identify $\theta = \pm 1$ as + or -1 sigma uncertainty so…

This is known as a **log-normal**** distributed nuisance parameter

$$\pi(\theta) = e^{-\frac{1}{2}\theta^2}$$

*Usually we usually know the luminosity better than this

**this is a very common, but not the only distribution we use

Likelihoods & Measurements

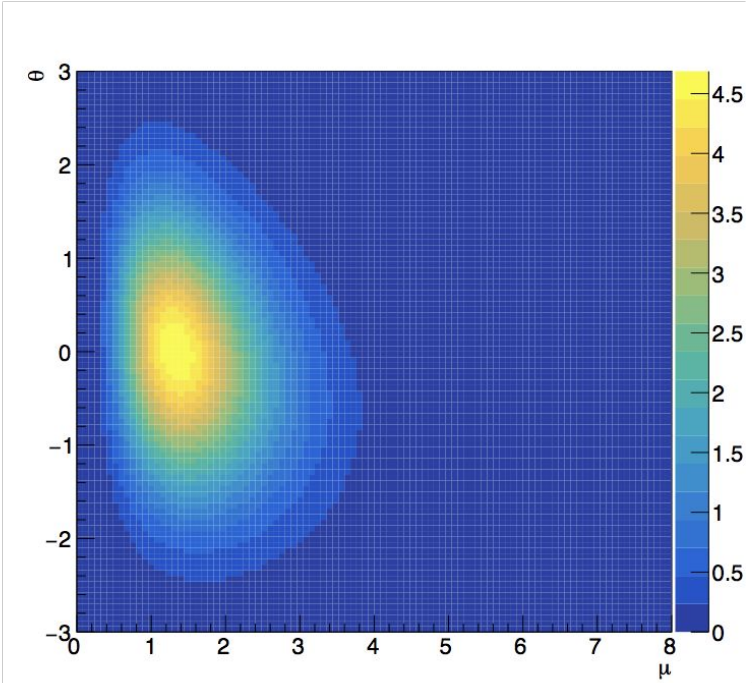
The likelihood is then $\mathcal{L}(\mu,\theta) = \lambda^n(\mu,\theta)e^{-\lambda(\mu,\theta)} \cdot e^{-\frac{1}{2}\theta^2}$

With $\lambda^n(\mu,\theta) = \mu\sigma(pp \to X)\varepsilon AL(1.1)^\theta$

Suppose in our experimental setup
A = 0.2, ε = 0.5, L = 30 fb^{-1} , σ = 1 fb

And we observe n = 4 events,
then our likelihood will look like …

You can see that the maximum of the likelihood is found at $\mu \sim 1.33$, θ = 0



Likelihoods & Measurements

Notice that the maximum value of the likelihood doesn't mean anything

- Its > 1 so it cannot be a "probability"
- It's somewhat arbitrary - I could have included a normalisation term in $\pi(\theta)$ without changing the values of μ or θ at which the maximum of the likelihood is found

Instead, the relative values of likelihoods are useful (we will see why soon)

Also in general, we often have many more than one observation (eg multiple bins in a histogram, or unbinned data). These are simple to deal with by using the product rule for probability …

Multiple bins: $$p \to \prod_i p_i = \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!}$$

Unbinned: $$p \to \prod_i \int_{x_i}^{x_i+\delta x_i} f(x_i)dx \xrightarrow[\delta x_i \to 0]{} \prod_i f(x_i)\delta x_i = \prod_i f(x_i)$$

Where $f(x)$ is the probability density function for *x* and we have used the fact that the data is fixed to ignore the δx_i

Marginalisation/Profiling

In general, we would also have many more than 1 nuisance parameter (usually, there is one per source of systematic uncertainty). In these cases, reporting the N-dim likelihood is not feasible and not interesting.

Instead, we tend to remove the nuisance parameters from the likelihood by one of two methods

Marginalisation or ***Profiling***

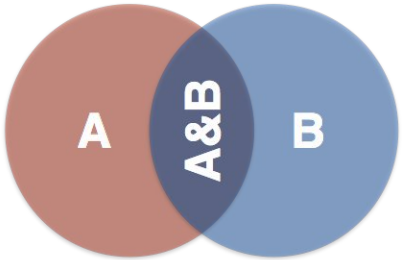
The two are often synonymous with **Bayesian** vs **Frequentist** methods

- We won't go into the long debate about which is better/worse/right/wrong etc, but rather just review the techniques we use in combine which are one or the other….

Marginalisation

Recall Bayes' theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$



For our purposes, we can write this as

$$p(\mu|data) = \int \frac{p(data|\mu,\theta)p(\theta)p(\mu)}{p(data)} d\theta$$

since $p(\mu|data) = \int p(data|\mu,\theta)p(\theta)d\theta$, by the probability sum rule. $p(\mu|data)$ is know as the posterior probability of μ

Notice how the **likelihood** has appeared in the definition of the posterior!

Marginalisation

We have also introduced $p(\mu)$, the "***prior***" on μ.

$$p(\mu|data) = \int \frac{p(data|\mu,\theta)p(\theta)p(\mu)}{p(data)} d\theta$$

This is something we need to choose* - for now, lets assume a "flat prior" such that

$$p(\mu) = \begin{cases} 1/20 & \text{if } 0 \leq \mu \leq 20 \\ 0 & \text{if else} \end{cases}$$

Finally $p(data) = \int p(\mu|data)d\mu$ is just a normalisation term.

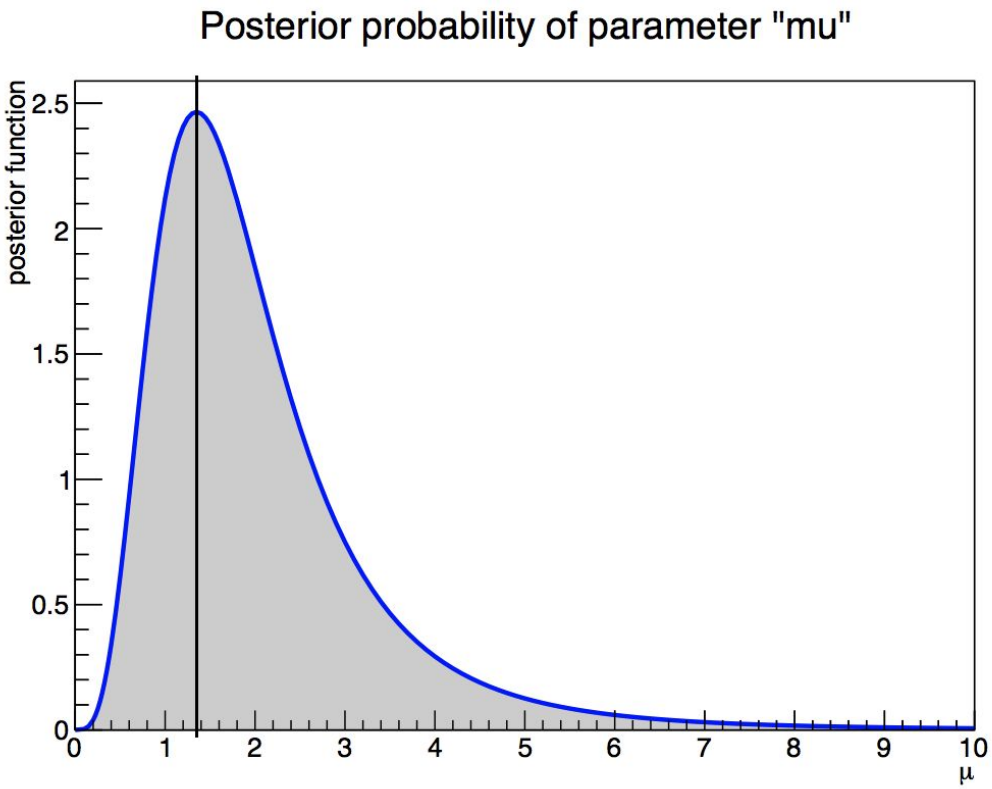
*The dependance of the results on this choice is something which must be checked in each analysis

Marginalisation

For our cut-and-count analysis, the posterior looks like …

The value of μ which is the most common is ~1.333!

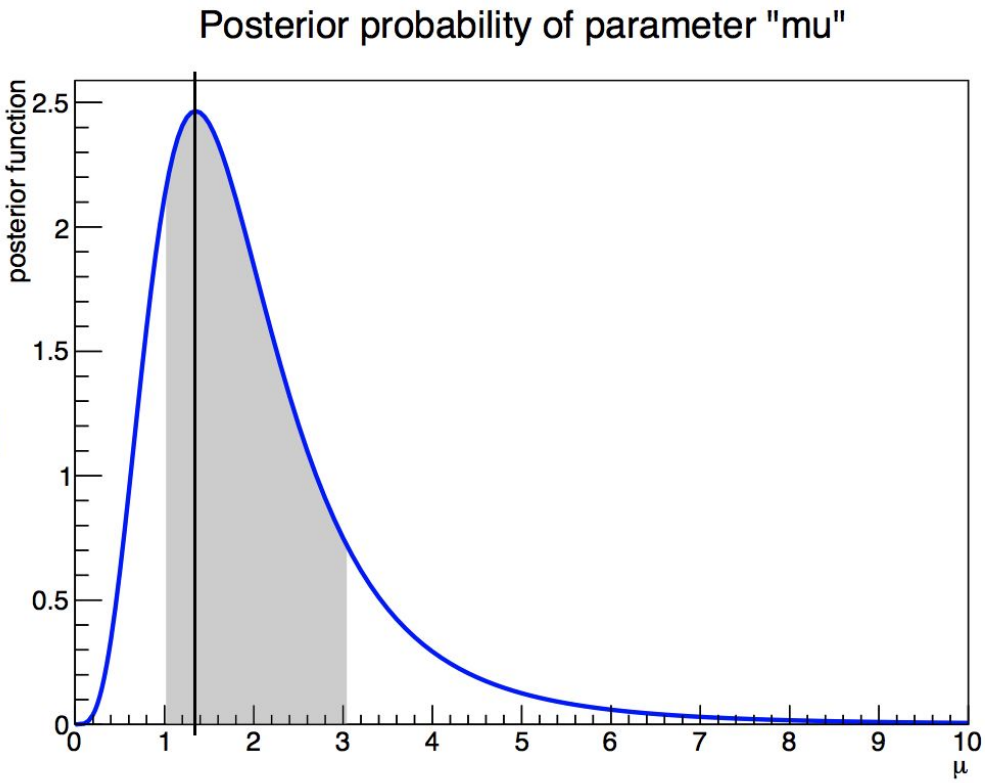
In this case, this is the same as the value which gave us the maximum likelihood value, when $\theta = 0$



Marginalisation

We could also ask which region contains 68% of the posterior distribution. This is known as a ***68% credible interval***

This is just one such example and another choice to be made is ***which*** such interval should be reported



Profiling

The other common method to remove nuisance parameters from the likelihood function is to find the value for θ which maximises the likelihood at each value of μ. This is known as ***profiling*** over the nuisance parameters

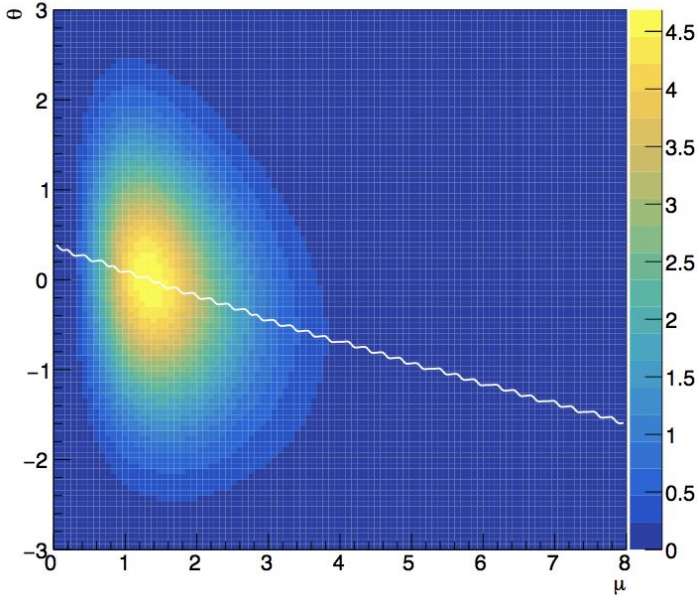
Taking our likelihood function, we can draw the line $\hat{\theta}(\mu)$, for which $\mathcal{L}(\mu, \theta)$ is maximum

The value of $\mathcal{L}(\mu, \theta)$, along this line is the "profiled likelihood"

$$\mathcal{L}(\mu, \theta) \rightarrow \mathcal{L}(\mu, \hat{\theta}(\mu)) := \max_{\theta} \mathcal{L}(\mu, \theta)$$

Or dropping, implicit dependencies,

$$\mathcal{L}(\mu, \theta) \rightarrow \mathcal{L}(\mu)$$



A quick word on gradient descent

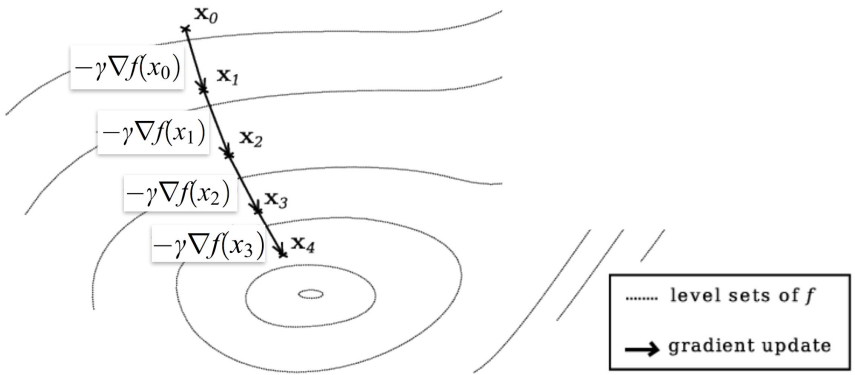
A very common class of minimizer algorithms (and the one we will be using) are based on Gradient Descent

If *f(x)* is defined and differentiable close to *x*, then *f(x)* decreases fastest in the direction $-\nabla f(x)$

Define a sequence

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

for some small γ

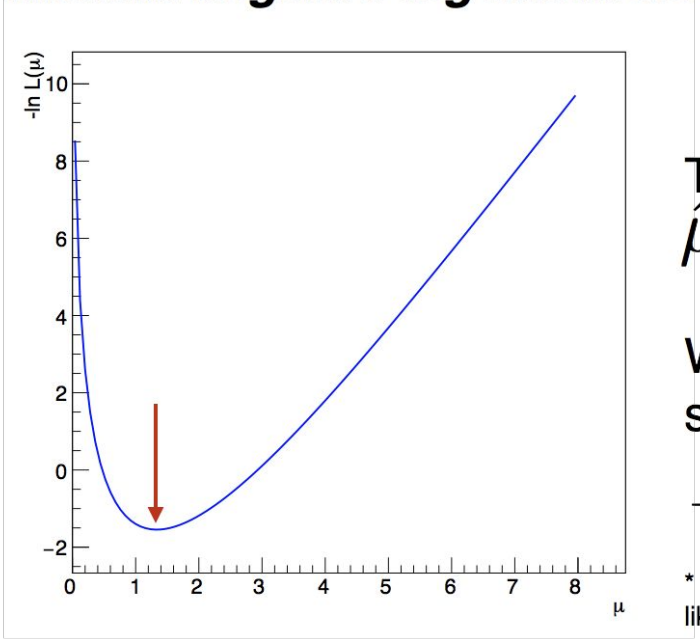


C.Mueller (2019) https://indico.cern.ch/event/839925/attachments/1945430/3234453/PHYSTAT2019.pdf

There are extensions - eg variable step-size (also known as “learning-rate”) or using Hessian to update. Depending on the problem, it can be faster to calculate the gradient numerically or analytically.

Profiling

Very often, to avoid dealing with small or large values of likelihoods (in this simple case it doesn't matter, but if we had many bins, the products can get quite small), we take negative logs of the likelihood —> maximum likelihood = ***minimum negative log likelihood***



$$\mathcal{L}(\mu) \rightarrow -\ln \mathcal{L}(\mu)$$

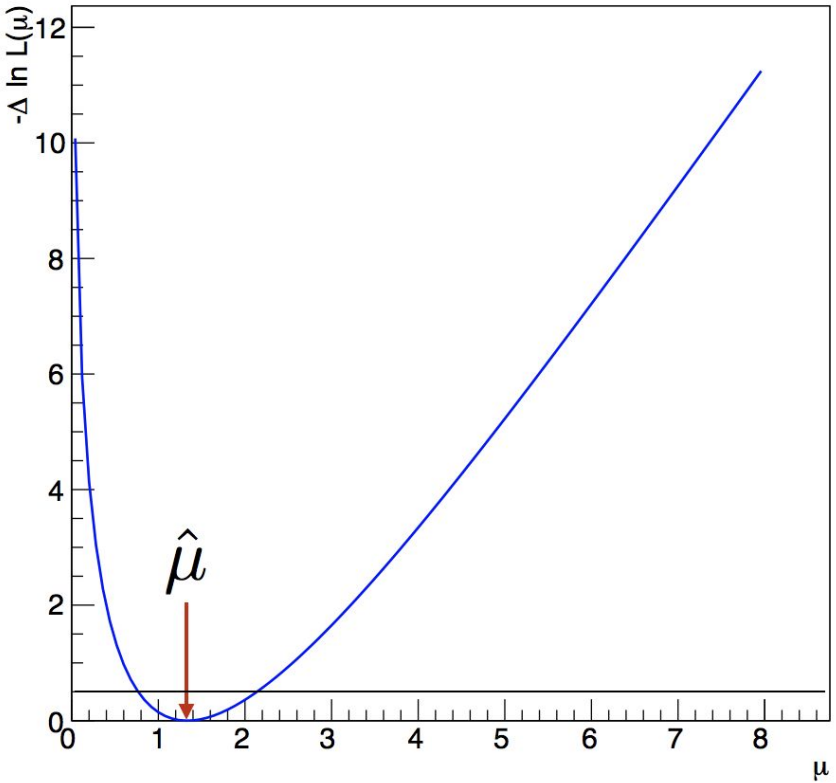
The minimum value of this curve is at $\hat{\mu}$ = 1.333!*

We often normalise this curve by subtracting the value at this minimum

$$-\ln \mathcal{L}(\mu) \rightarrow -\ln \mathcal{L}(\mu) - (-\mathcal{L}(\hat{\mu})) = -\Delta \ln \mathcal{L}(\mu)$$

* again the ^ notation indicates the value at which the likelihood (-ve log likelihood) is maximised (minimised)

Profiling



Wilkes' theorem* tells us that we can obtain a 68% confidence interval from the region for which

$$-\Delta \ln \mathcal{L}(\mu) < 0.5$$

Which we will call the "**minos**" method

What does that mean exactly?

*I won't go through why but think about taylor expanding the log-likelihood around the minimum

Frequentist intervals

So far, we have seen two methods for determining the value of the POI and ascribing a 68% confidence or credible interval

- For Bayesian thinkers, the 68% credible interval represents the probability that the parameter μ is in a certain region, given that we observed n=4 events.

- For frequentist thinkers, the interval we constructed was just one of many possibilities depending on the observation n. There is a little more work in defining what this interval means.

We can introduce the concept of ***coverage .***

"*When a 68% confidence interval has **good coverage** it means that 68% of intervals constructed in such a way will contain the **true** value of the POI.*"

This should hold regardless of the true value of the POI*. It is ***not*** a statement about the specific interval we constructed but rather a statement about the ***ensemble*** of intervals!

*And should also be true for any reasonable values of the nuisance parameters, but this is a subtlety which should be checked to hold

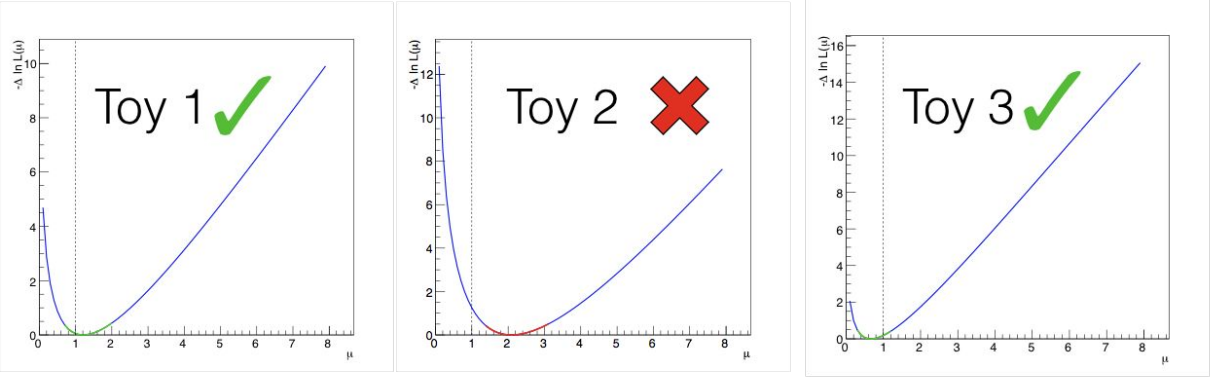
Frequentist intervals

Let's suppose our true value of $\mu = 1$.

We can generate outcomes for n and for θ_0

- Remember that $\theta_0 = 0$ was the value we observed in our 1 measurement. We could imagine measuring other values, generated from $\pi(\theta_0|\theta)$
- We need to pick a value for θ to generate from however. The most obvious choice* would be $\hat{\theta}(\mu = 1) = 0.09$

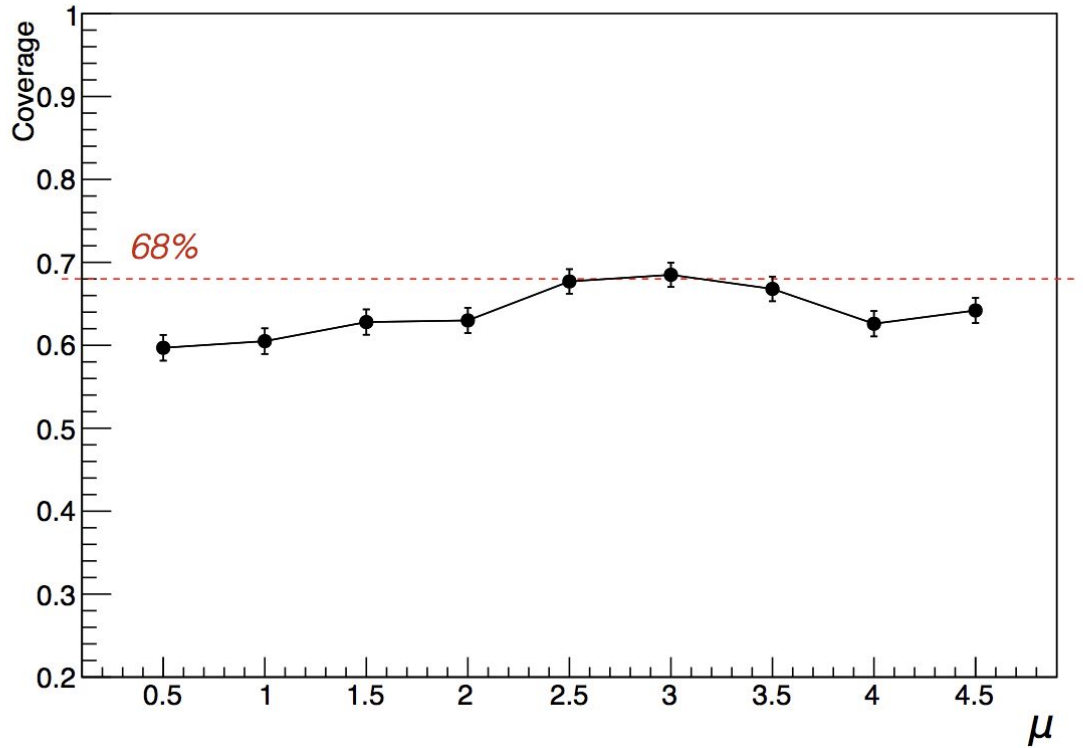
For each generation (toy) we determine the 68% confidence interval and determine whether or not 1 is contained in the interval…



*Again, this is common practise, but we should check that the coverage property doesn't depend on this choice!

Frequentist intervals

What is the coverage as a function of μ ?



The intervals undercover for smaller values of μ .

This is not so surprising since Wilkes' theorem assumes the likelihood follows a Gaussian distribution.

This becomes more appropriate for large values of n but is less accurate when n is small

Frequentist intervals

We can do better using a Neyman construction…

1. Pick a true value of $\mu = \mu_T$, (set θ_0 to the value which maximises the original likelihood defined at $\mu = \mu_T$)
2. Generate toy values for n and θ_0 (for generating, again set θ to the value which maximises the original likelihood defined at $\mu = \mu_T$)
3. Evaluate $q = -\Delta \ln \mathcal{L}(\mu_T)$ (this is called a test-statistic) for each toy and enter into a histogram* -> $f(q)$

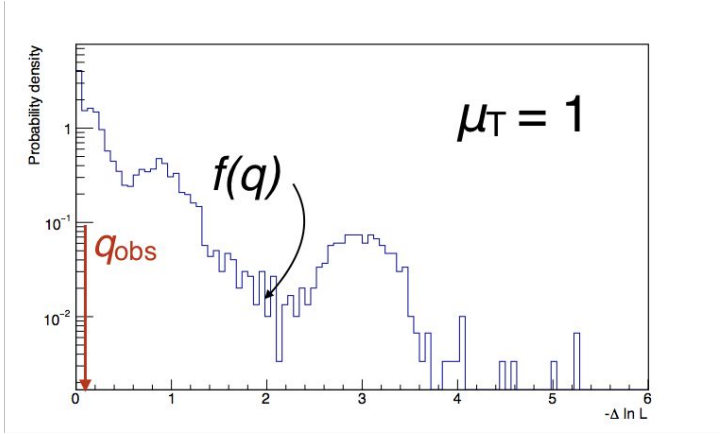
4. Calculate $p = \int_{q_{\rm obs}}^{+\infty} f(q) dq$ where $q_{\rm obs}$ is the value of q for the ***observed data***

5. If $p < 1 - 0.68$, then μ_T is in the 68% confidence interval, otherwise, its not
6. Start at 1. and repeat for another value of μ_T

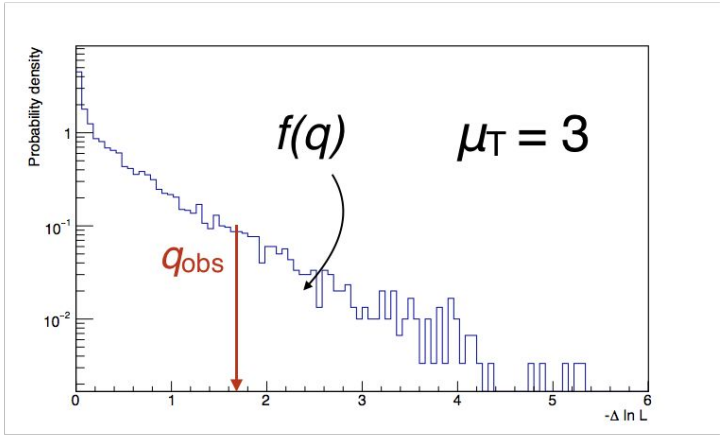
*note that in general, you need to choose an ordering principle for the outcomes, here we have used the Feldman-Cousins approach

Frequentist intervals

For a small value $\mu_T = 1$ (which means λ~3), we see the discrete nature of the Poisson probability appear in the distribution of q

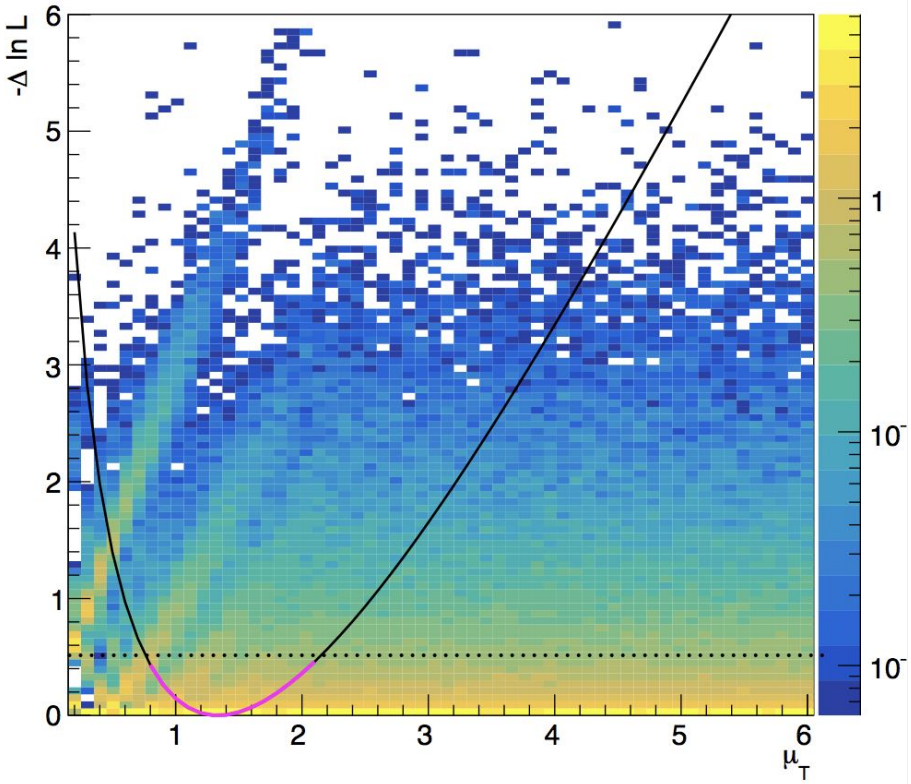


For large values $\mu_T = 3$ (which means λ~9) we see that the Poisson smooths out and the distribution of q looks like a χ^2 function with 1 degree of freedom. This is Wilkes' theorem kicking in !



Notice that the value of q_{obs} changes depending on μ. Of course, this is expected for our choice of procedure (test-statistic)

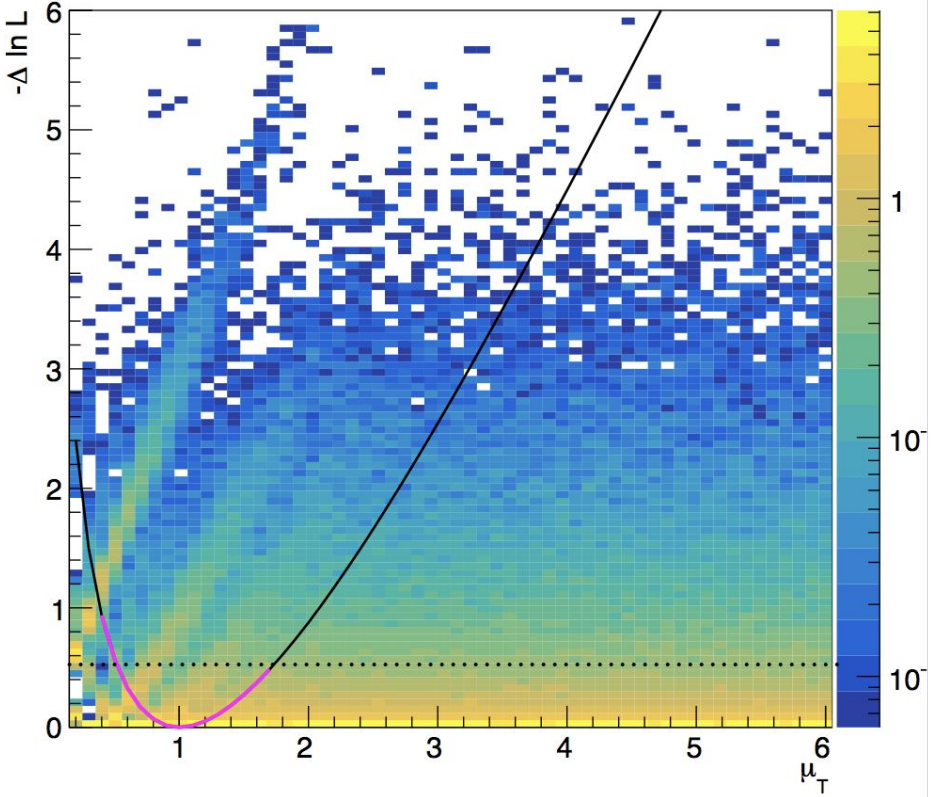
Frequentist intervals



The magenta curve shows the points for which $p < 1 - 0.68$, i.e the 68% confidence interval given our observation.

You can see its actually not so far off from what we got using the minos method which would just pick points for which q < 0.5

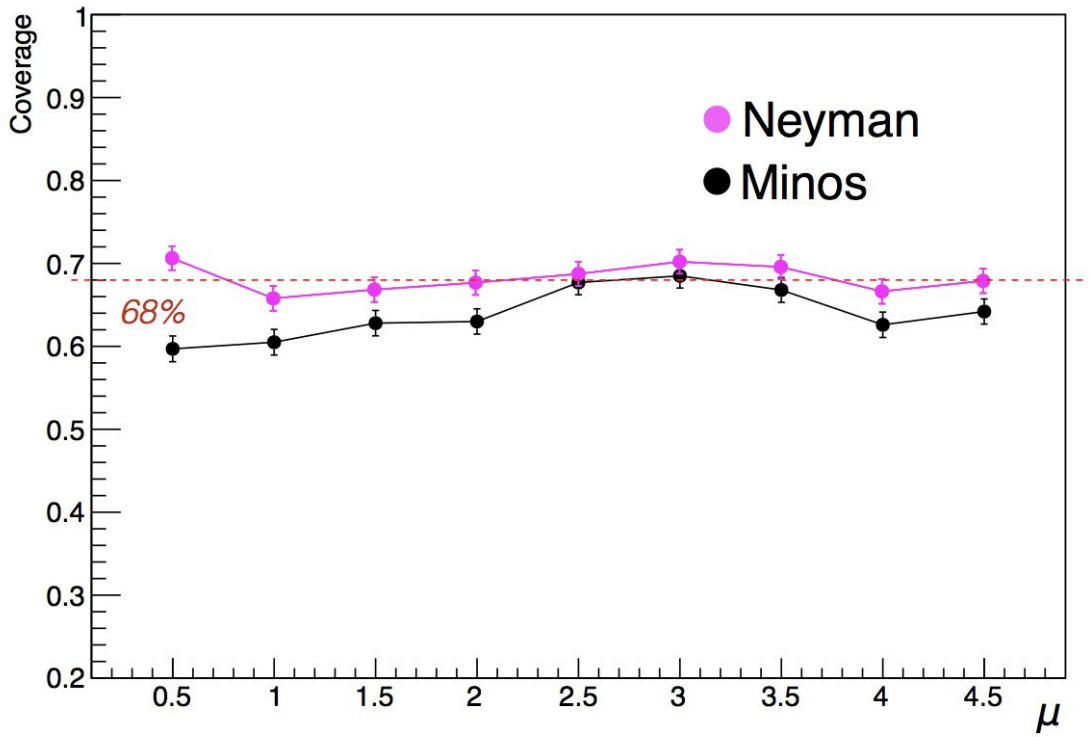
Frequentist intervals



This is not always the case however. In this example the interval is wider than what we would have gotten from the minos method!

Frequentist intervals

What is the coverage as a function of μ ?



The Neyman construction indeed does better in terms of coverage (by design) but of course it is more computationally intensive to run.

PREFIT-School

PRecision Effective Field Theory School

2-13 March 2020 at DESY, Hamburg

Hands On

Session 1

You need to checkout the GitHub repository (if you didn't already do so)

https://github.com/amarini/Prefit2020

Within a terminal run

```
git clone https://github.com/amarini/Prefit2020
```

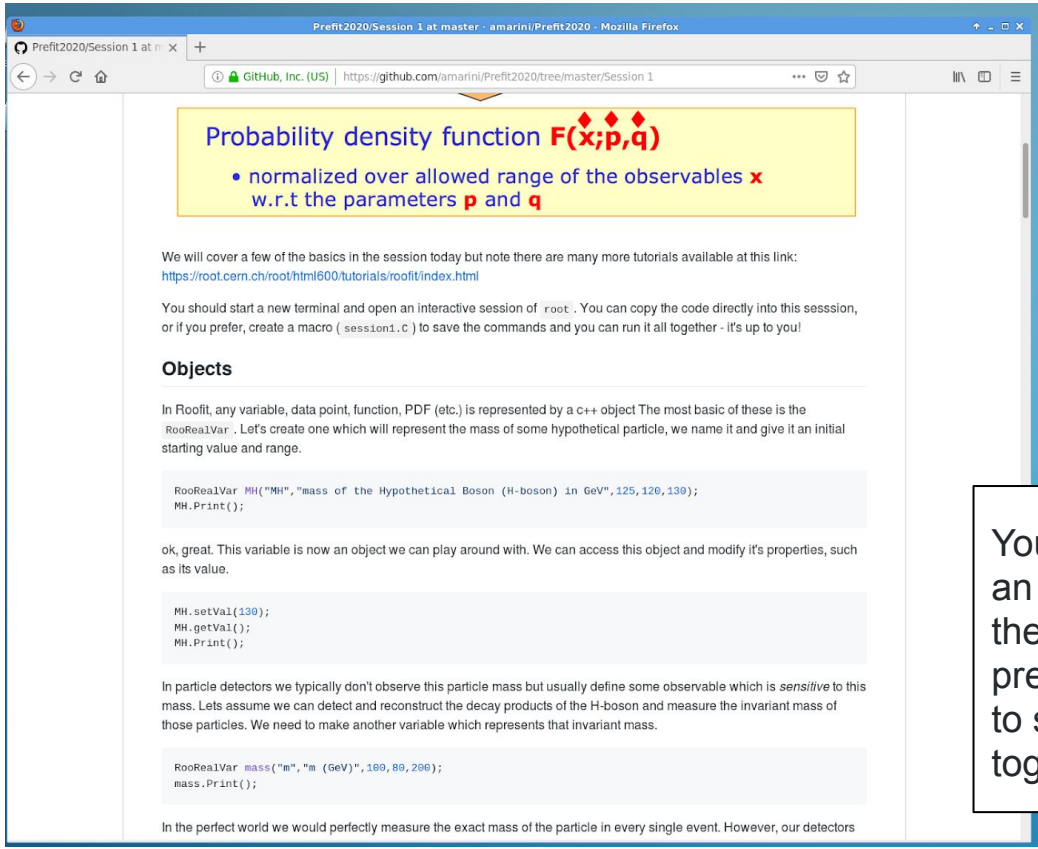
All the instructions and files you need for the hands-on sessions are included there.

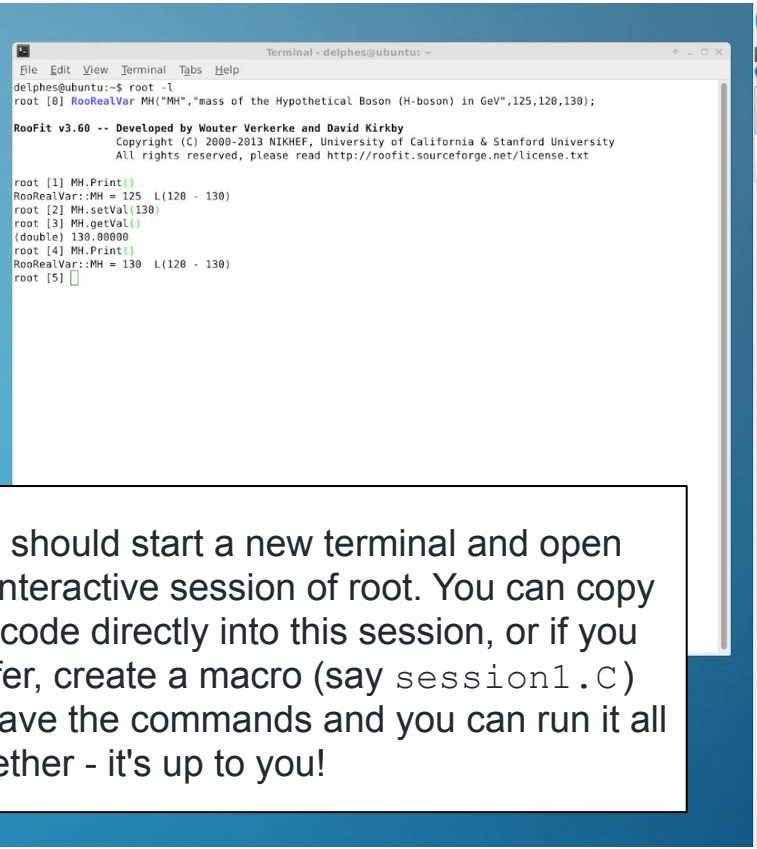
For today's hands-on session, we will cover the basics of `RooFit` - the widely used tool for fitting and statistics in HEP. There are ***many*** other tools for fitting (eg scipy, tensorflow, pyTorch) which can be more useful for large scale (high-dimensionality) fitting, and in particular for use in Global Fits/constraining EFTs (eg HEPfitter, Gambit, MadMiner - see Kyle's ML lectures for examples)

You can find the instructions for this session here :

https://github.com/amarini/Prefit2020/blob/master/Session%201/README.md

Session 1





You should start a new terminal and open an interactive session of root. You can copy the code directly into this session, or if you prefer, create a macro (say `session1.C`) to save the commands and you can run it all together - it's up to you!

Session 1

In Session 1 - you will cover

1. RooFit basics
   a. Creating RooFit objects/functions/pdfs
   b. Creating / using datasets
2. Likelihoods and likelihood fits
3. Nuisance parameters
   a. Profile likelihoods
   b. Marginalisation
4. (Advanced) Feldman-Cousins intervals - if you get there

Hopefully, for those who aren't already familiar with RooFit, this will cover all that you need for the hands-on parts → use these as a reference if you get stuck (and ask the instructors too)!