21/11/2018

MAXIMUM LIKELIHOOD FIT

PARAMETER ESTIMATION

You have a set of measurements $\{x_i\}$ of the same variable and a distribution hypothesis $f(x\,;\,\theta)$. You want to find the value of theta that best represents the measured data.

For this you can use a Maximum Likelihood estimator (ML), defined as

$$\boxed{\boldsymbol{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\theta}\, \mathbb{P}\left(\mathbf{x}, \boldsymbol{\theta}\right)}$$

This estimator is proven to be unbiased and often the best you can use (it reaches the Cramer-Rao limit). We want just to prove that, by comparing the ML estimate to a real analytical solution. Let's use the exponential distribution for that.

$$f(x;\tau) = \frac{1}{\tau} e^{-x/\tau} \Rightarrow \mathcal{L}\left(\tau;\mathbf{x}\right) = \prod_{i=0}^{N-1} f(x_i;\tau) = \prod_{i=0}^{N-1} \frac{1}{\tau} e^{-x_i/\tau} \Rightarrow \ln \mathcal{L}\left(\tau;\mathbf{x}\right) = -N \ln \tau - \frac{1}{\tau} \sum_{i=0}^{N-1} x_i$$

The estimator we want is

$$\frac{d \ln \mathcal{L}}{d\tau} = 0 \Rightarrow \boxed{\tau_{\mathrm{ML}} = \frac{1}{N} \sum_{i=0}^{N-1} x_i}$$

and its variance is

$$\boxed{\mathbb{V}\left[\tau_{\mathrm{ML}}\right] = \mathbb{E}\left[-\frac{d^2 \ln \mathcal{L}}{d\tau^2}\right]^{-1} = \frac{\tau_{\mathrm{ML}}^2}{N}}$$

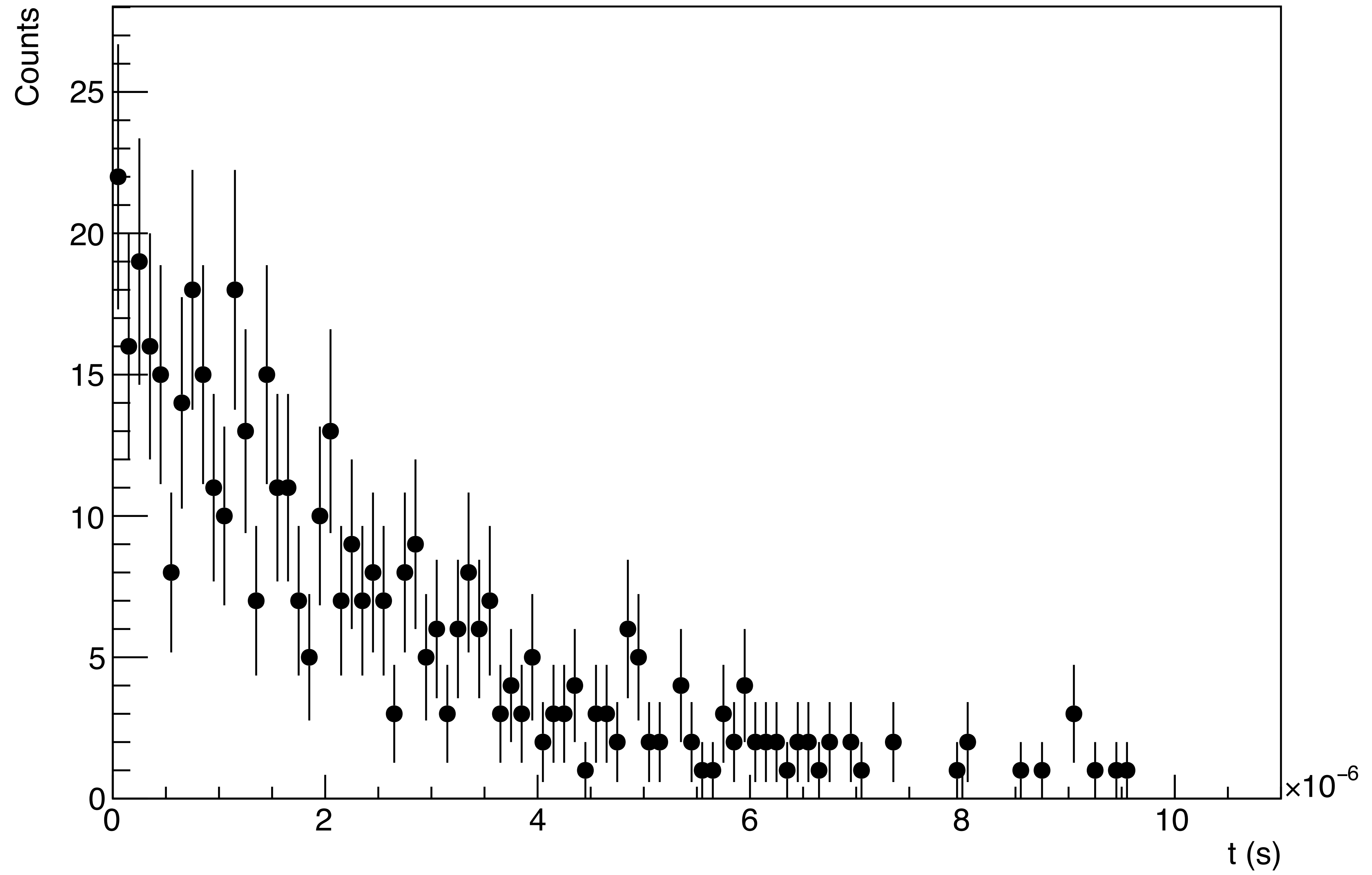
EX01: MUON LIFETIME

Suppose you measured for one hour incoming muons that stopped in your detector and decayed. You get ~500 muons and for each of them you recorded the time passed between the trigger and the detection of the decay products.

$$N \sim \mathrm{Pois}(500)\,; \quad f(x,\tau) = \frac{1}{\tau} e^{-x/\tau} \quad \Rightarrow \quad \boxed{\tau_{\mathrm{ML}} = ? \quad \mathbb{V}\left[\tau_{\mathrm{ML}}\right] = ?}$$

Estimate the muon lifetime and quote an error on the measurement using three different techniques:

1. Analytical solution of the ML problem
2. Scan (and plot) the $\ln\mathcal{L}$ profile and search for its maximum
3. Use a numerical minimisation tool (MINUIT / Minuit2) to find the minimum of $-\ln\mathcal{L}$
4. Build a histogram and use ROOT fit interface (TH1::Fit) to get an estimate of tau (try both the chi-square fit and the binned maximum-likelihood fit; what's the difference?)



Validate the variance you obtain using a toy MC

Bonus question: we don't need to know how much time passed between the muon creation in the atmosphere and it entering the detector. Why?

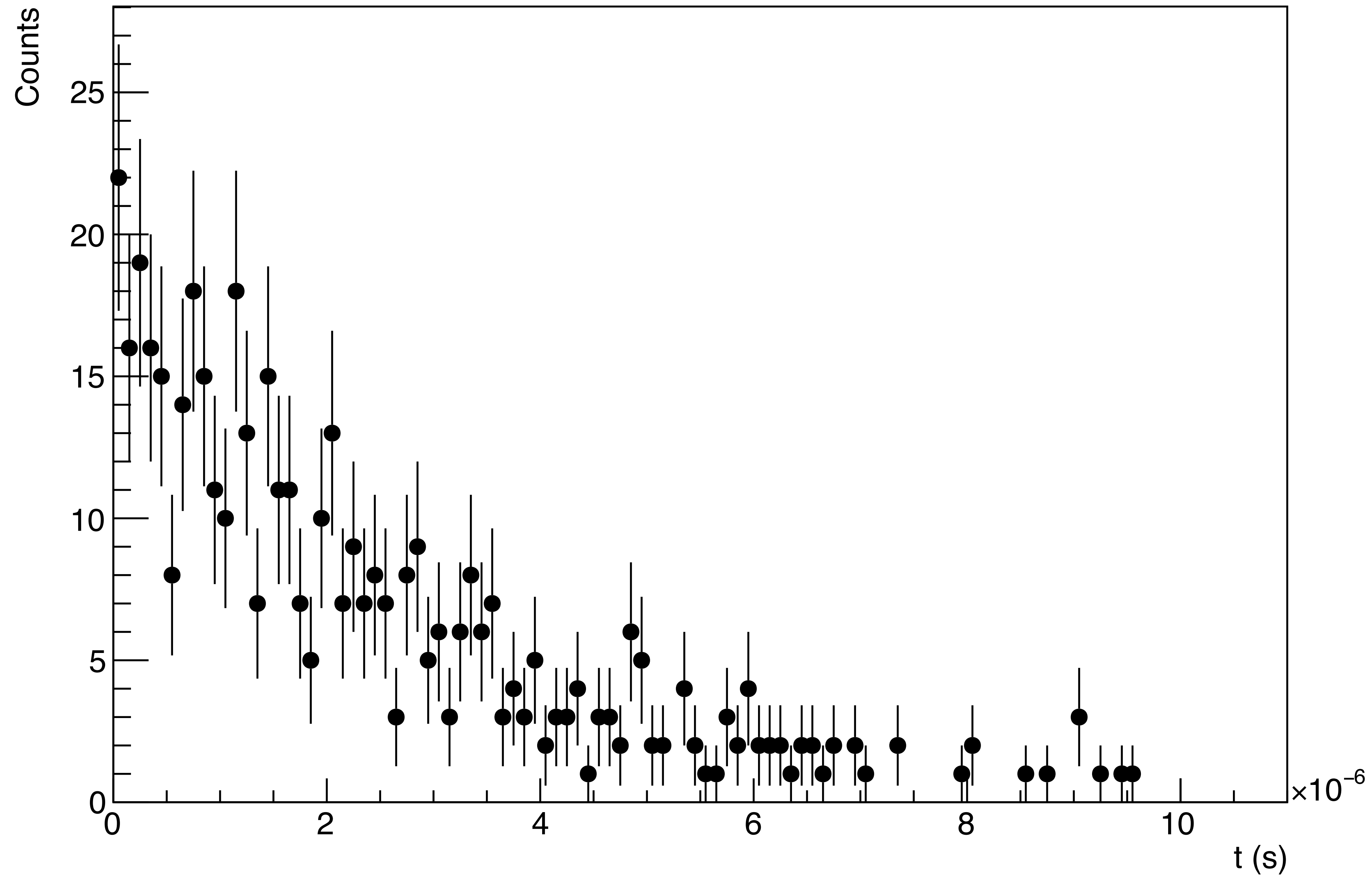
EX01: MUON LIFETIME

Suppose you measured for one hour incoming muons that stopped in your detector and decayed. You get ~500 muons and for each of them you recorded the time passed between the trigger and the detection of the decay products.

$$N \sim \mathrm{Pois}(500)\,; \quad f(x,\tau) = \frac{1}{\tau} e^{-x/\tau} \quad \Rightarrow \quad \boxed{\tau_{\mathrm{ML}} = ? \quad \mathbb{V}\left[\tau_{\mathrm{ML}}\right] = ?}$$

How to access the data: *(you can get it from https://github.com/valerioformato/MLEstimate)*

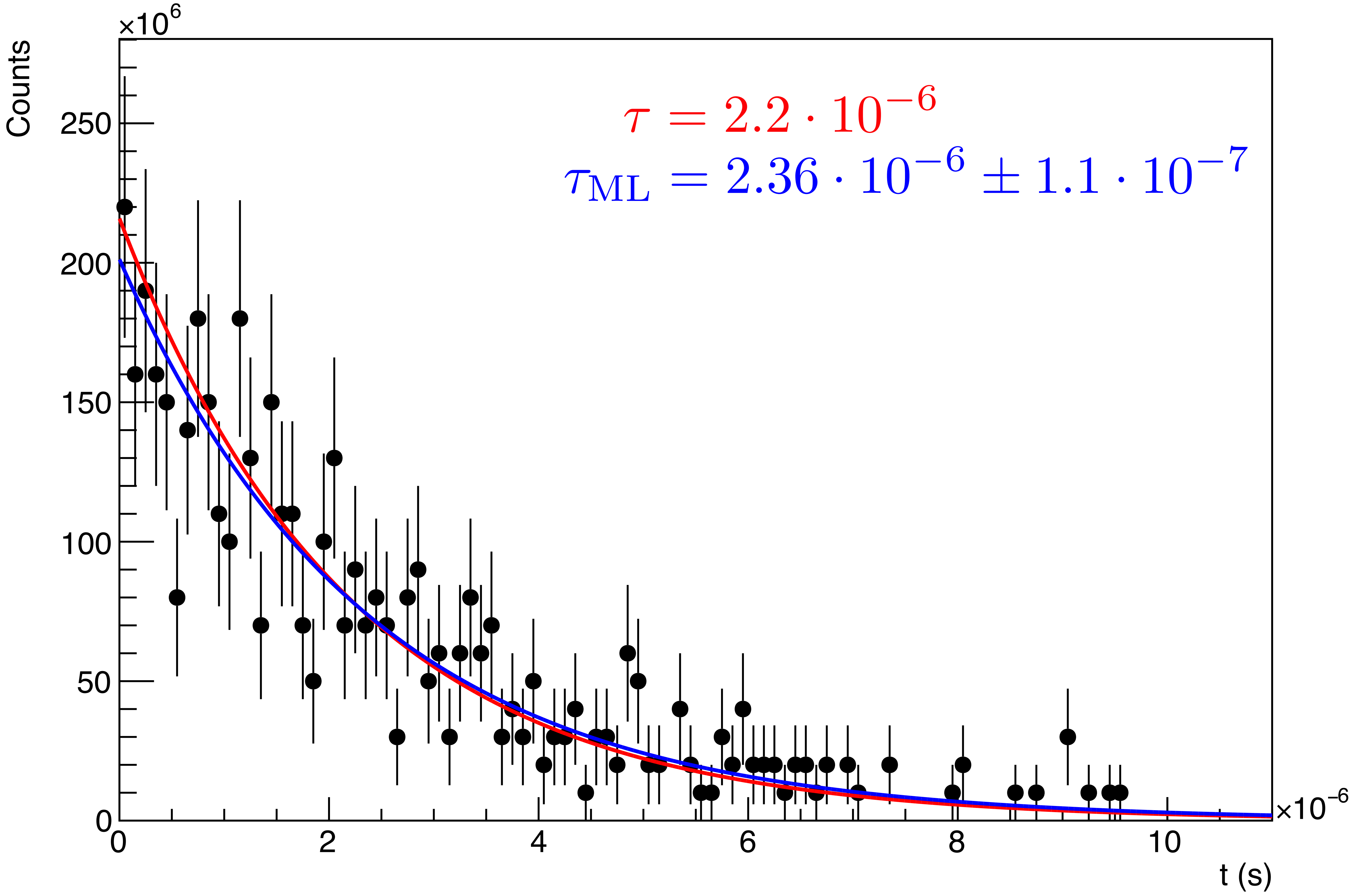
```
auto infile = TFile::Open("data/data.root");
auto tree = (TTree *)infile->Get("Triggers");

float t;
tree->SetBranchAddress("time", &t);

Long64_t nEv = tree->GetEntries();
for (Long64_t iEv = 0; iEv < nEv; iEv++) {
  tree->GetEntry(iEv);
  ...
}
```



EX01: (A) ANALYTICAL SOLUTION

$$\frac{d \ln \mathcal{L}}{d\tau} = 0 \Rightarrow \boxed{\tau_{\mathrm{ML}} = \frac{1}{N}\sum_{i=0}^{N-1} x_i}$$

$$\boxed{\mathbb{V}\left[\tau_{\mathrm{ML}}\right] = \mathbb{E}\left[-\frac{d^2 \ln \mathcal{L}}{d\tau^2}\right]^{-1} = \frac{\tau^2_{\mathrm{ML}}}{N}}$$

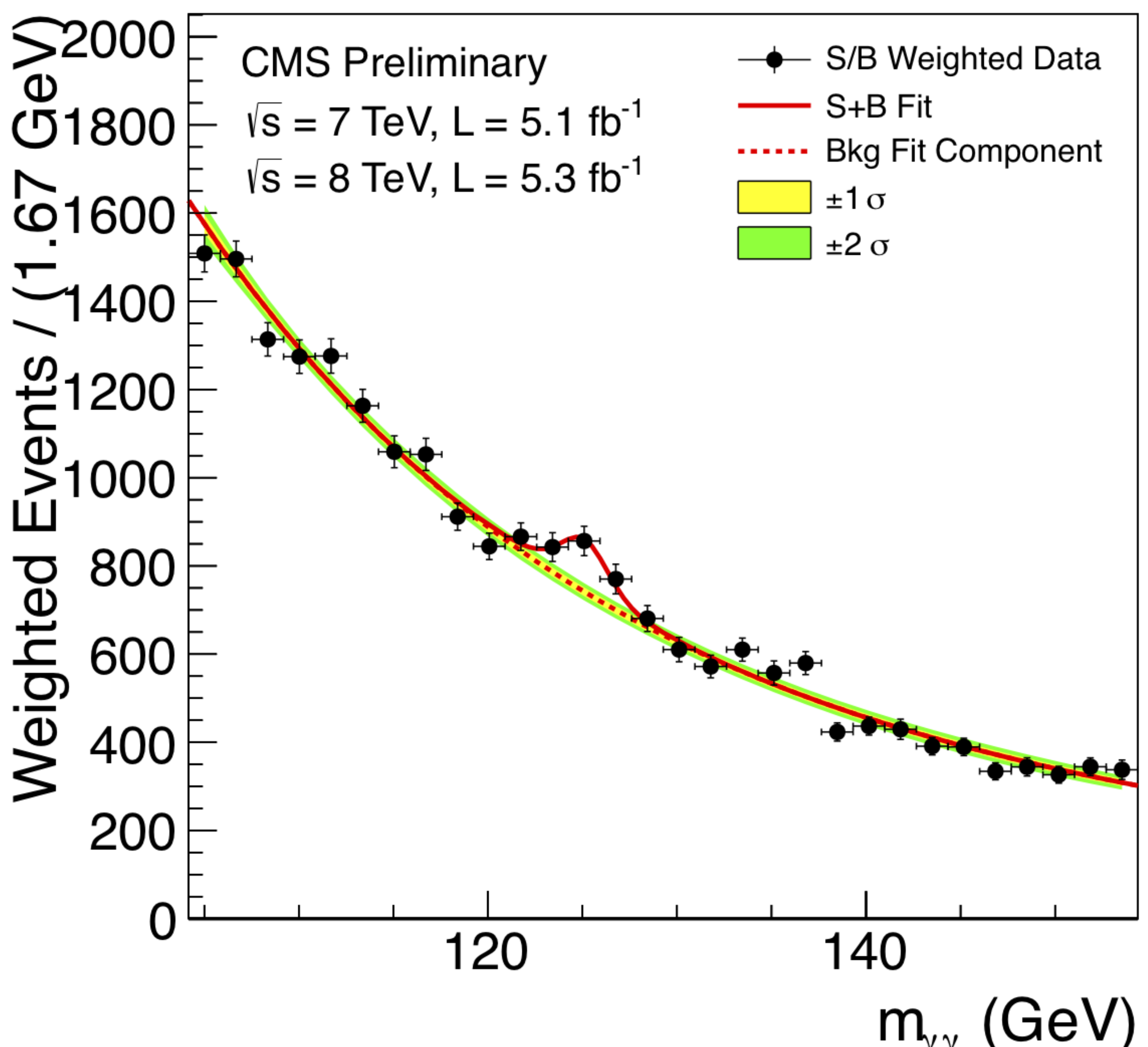


Looks like the easiest and most efficient solution (in terms of computational speed).
And it is.

But almost never you'll be so lucky to have a problem that can be solved analytically.

Real life fits are more complicated:

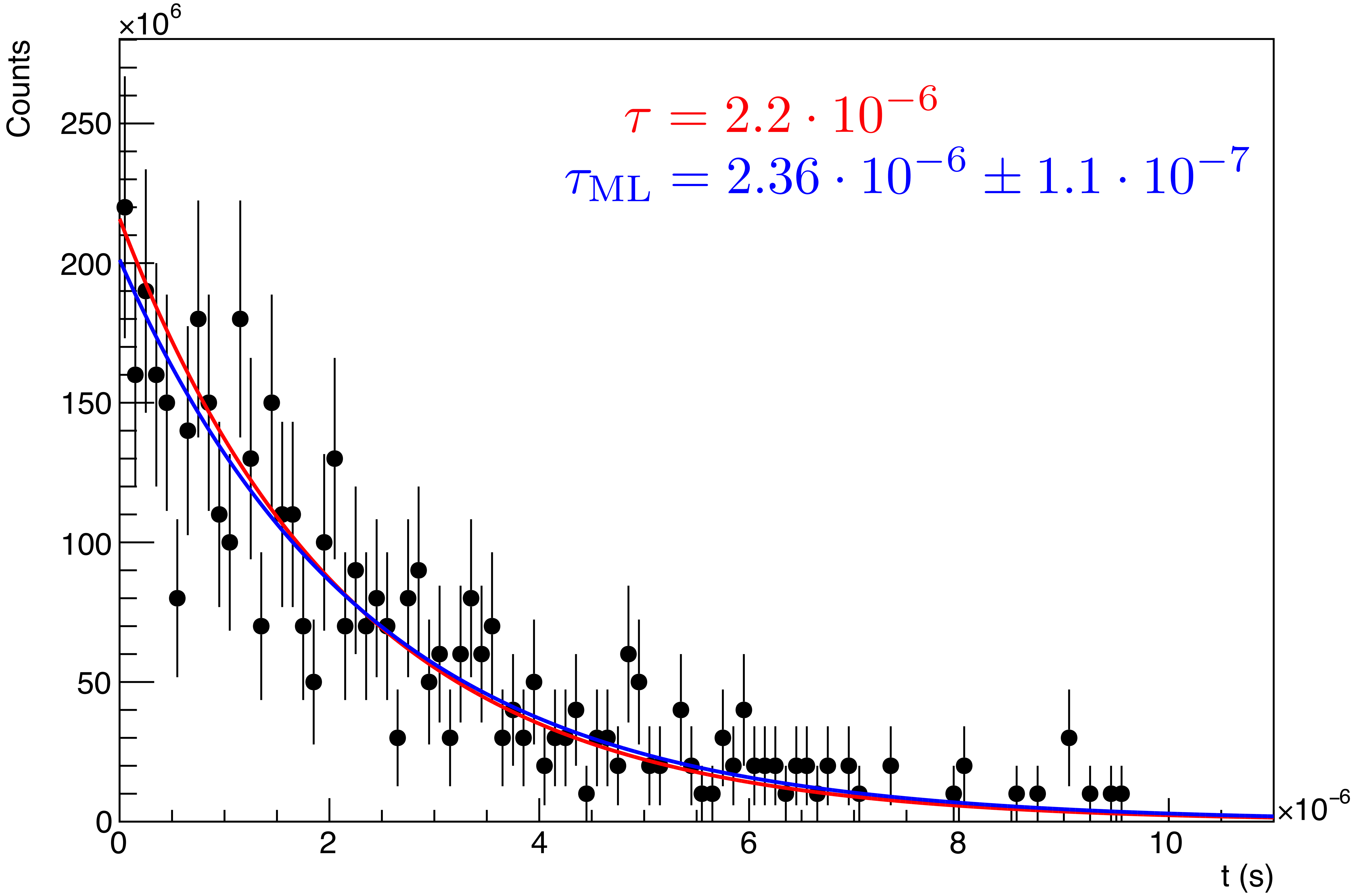
- Sum of p.d.f. from different contributions
- Large number of parameters
- p.d.f. not always analytical (MC templates)

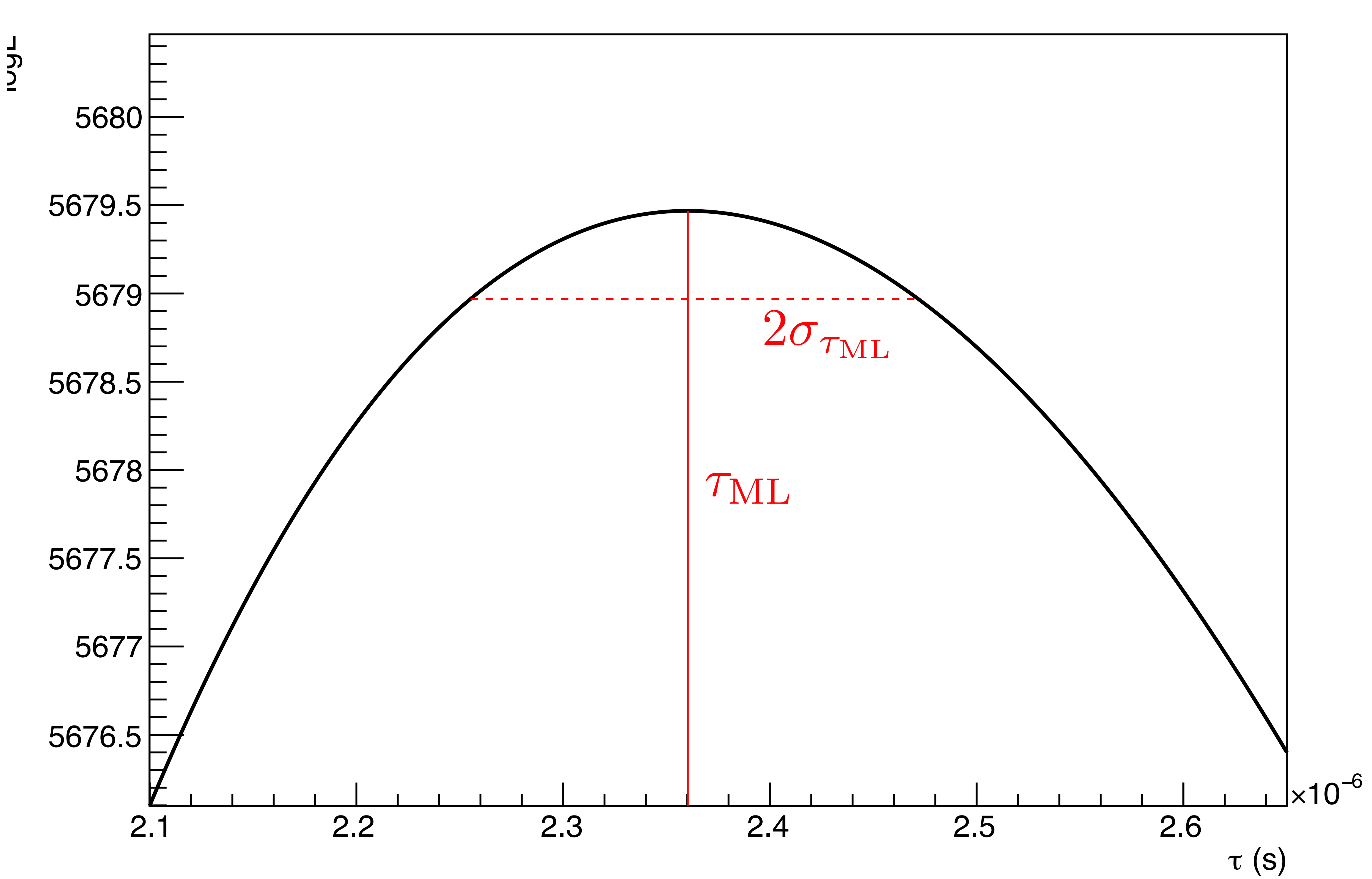


EX01: (B) LIKELIHOOD SCAN

Select a range of values for τ that you think includes the desired value. Scan the range computing $\ln\mathcal{L}$ for each value of τ and look for the maximum.

The 68% confidence level interval can be found as $\left\{\tau \in \mathbb{R} : \ln\mathcal{L}(\tau) > \ln\mathcal{L}(\tau_{\mathrm{ML}}) - \frac{1}{2}\right\}$

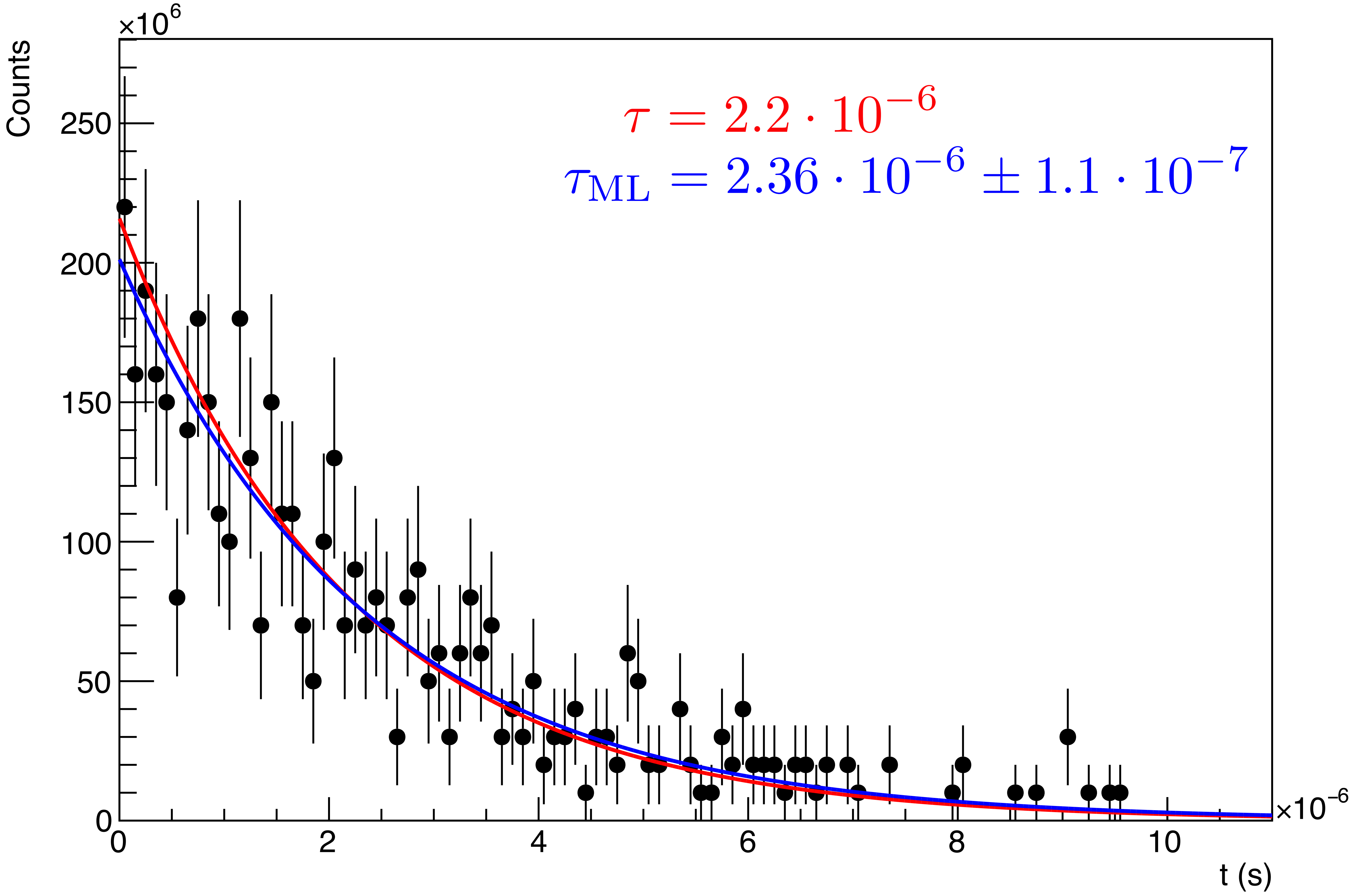




EX01: (C) MINUIT

- Define your $-\ln\mathcal{L}$ to minimize
- Pass it to MINUIT/Minuit2
- Minimize it

You can use the TMinuit interface or the ROOT::Math::Minimizer interface.



(Beware of TMinuit, it's deprecated and quite old)

(Beware of ROOT::Math::Minimizer, it's flexible and powerful, but requires some additional effort to make it work)

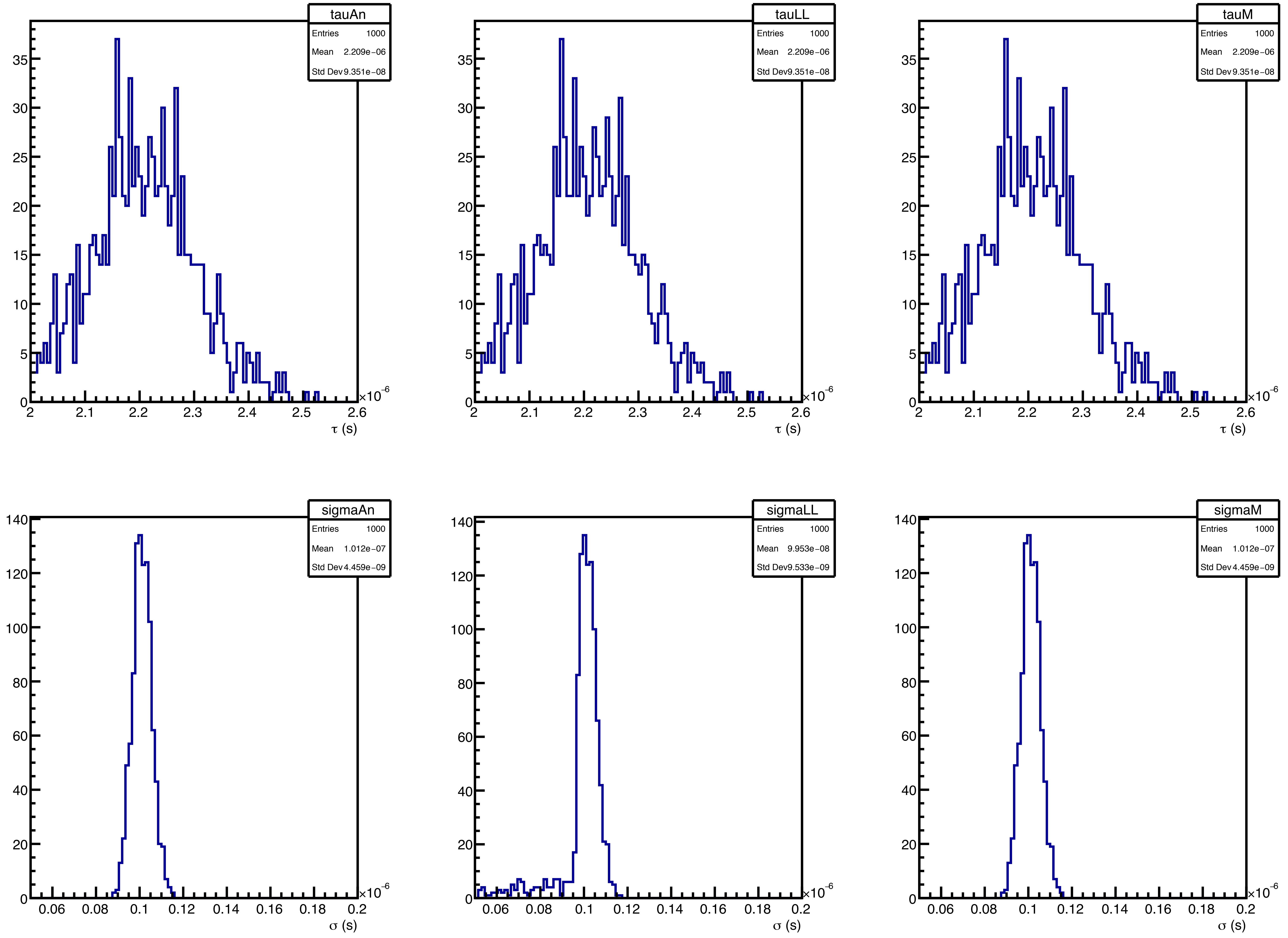
```
**********
**    1 **SET PRINT           0
**********
**********
**    2 **SET NOGRAD
**********
PARAMETER DEFINITIONS:
   NO.    NAME         VALUE      STEP SIZE      LIMITS
     1 tau          1.00000e-06  1.00000e-09     no limits
**********
**    3 **SET ERR          0.5
**********
**********
**    4 **SET PRINT           0
**********
**********
**    5 **SET STR           1
**********
**********
**    6 **MIGRAD       1e+06       0.001
**********
MIGRAD MINIMIZATION HAS CONVERGED.
MIGRAD WILL VERIFY CONVERGENCE AND ERROR MATRIX.
FCN=-5679.47 FROM MIGRAD    STATUS=CONVERGED     35 CALLS          36 TOTAL
                     EDM=2.32486e-14    STRATEGY= 1      ERROR MATRIX ACCURATE
  EXT PARAMETER                                   STEP         FIRST
  NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
   1  tau          2.36017e-06   1.53148e-07   5.63589e-09  -1.40800e+00
```

EX01: (E) TOY MC

Run the same experiment ~1000 times, each time generating a new dataset.

Compute τ and σ each time and check the distribution.

	τ	σ
Analytical	2,209	0.101
LL scan	2,209	0.099
Minuit	2,209	0.101
Minuit2	2,209	0.101



EX02: SCALING

Using toy MC experiments, verify the inverse square root scaling of the error. and how the ML estimator is actually unbiased.