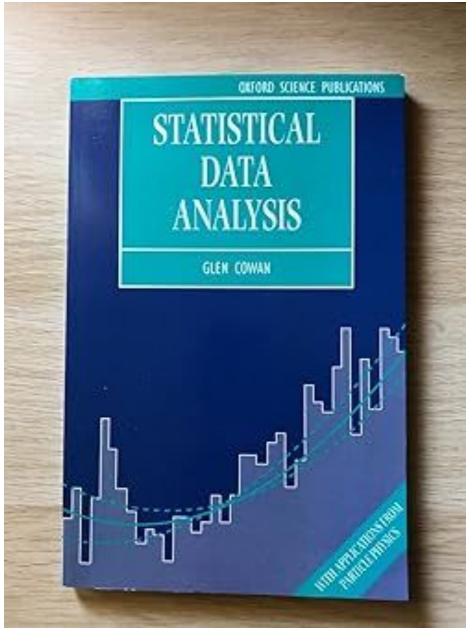
Fitting in HEP for pedestrians

Radek Žlebčík
US Belle II Summer Workshop
Oxford, June 20, 2024

Resources for statistics in HEP

- Statistical methods are getting more and more complex
  → takes time to tame it
  → big experiments have dedicated
  statistical working group
- Glen Cowan's book is an unofficial golden standard
- There are also newer books,
  e.g. from Olaf Behnke et al.
- Look/sign for one of many statistics schools
  → e.g. INFN School of statistics



Stand-alone fitting with Minuit

- The examples in this talk are done without dedicated fitting frameworks, without RooFit, RooStats, zFit…
- We use only Minuit minimizer ported to iminuit Python package
  → for HEP applications Minuit is still superior to scipy.optimize.minimize

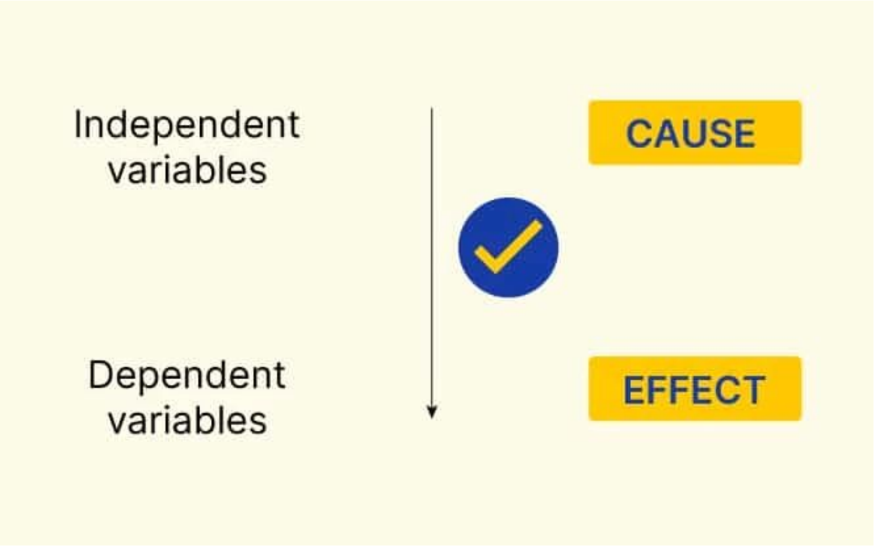
iminuit

Exploring the Black Box

Part 1: Fits with independent and dependent variable

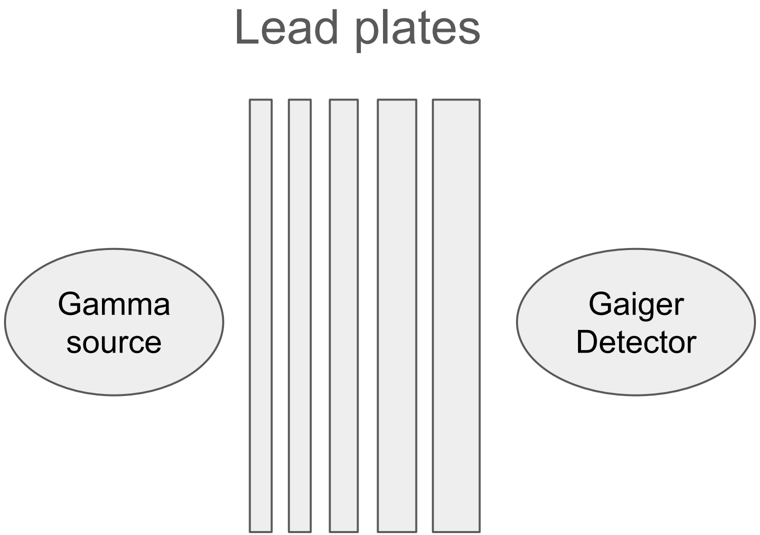
x

y

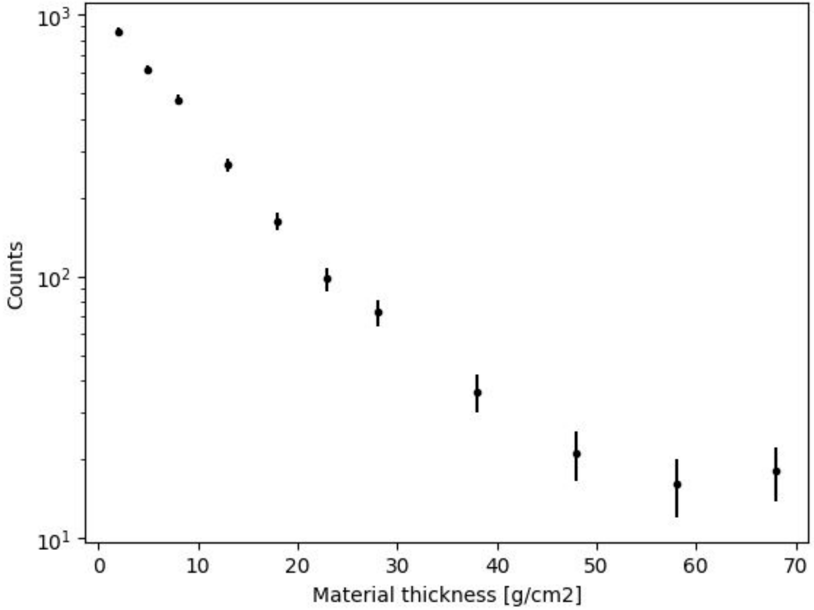


Fitting the absorption curve for photons

- **Independent variable:**
  Thickness of the hindrance [g/cm^2]
- **Dependent variable:**
  Counts in 60s



Example 1: Chi2 fits



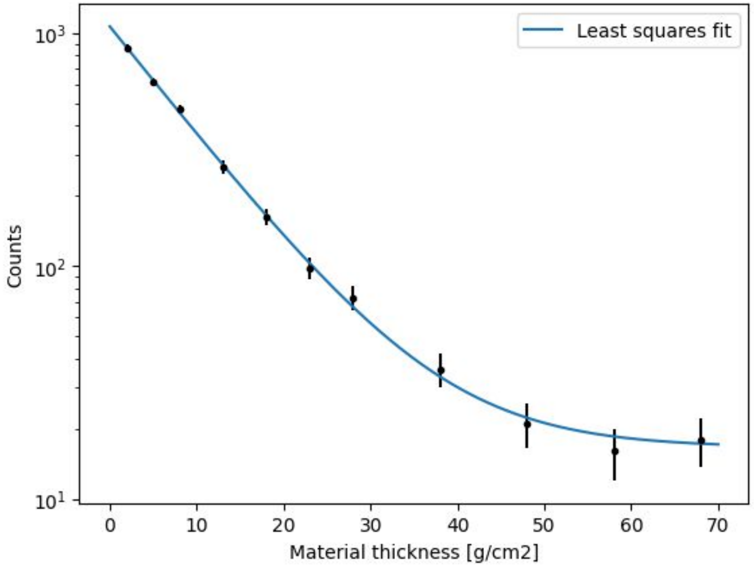
Least square method

Example 1: Chi2 fits

Model: $f(x,p) = N\exp(-\frac{x}{\lambda}) + N_0$

Getting parameters of our model by minimizing sum of deviations in quadrature

$$\mathrm{RSS} = \sum_i (y_i - f(x_i, p))^2$$



$$\lambda = 9.191$$
$$N = 1050.5$$
$$N_0 = 16.6$$

In analogy with the arithmetic mean:

$$\mathrm{RSS} = \sum_i (a_i - \mu)^2 \Leftrightarrow \mu = \frac{1}{N}\sum_i a_i$$

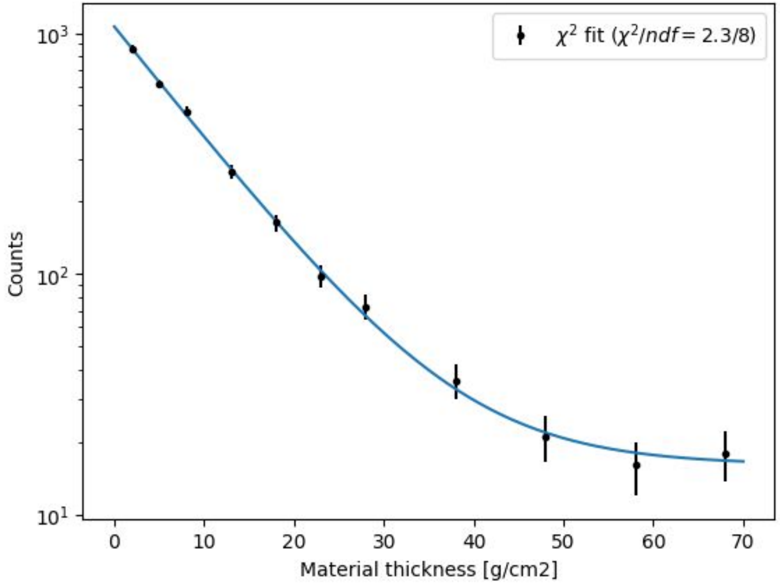
χ^2 fits

Example 1: Chi2 fits

Model: $f(x,p) = N\exp(-\frac{x}{\lambda}) + N_0$

Getting parameters of our model by minimizing the χ^2

$$\chi^2 = \sum_i \left(\frac{y_i - f(x_i,p)}{\sigma_i}\right)^2$$



$$\lambda = 9.237$$
$$N = 1047.9$$
$$N_0 = 16.1$$

Unc.-weighted arithmetic mean:

$$\chi^2 = \sum_i \left(\frac{a_i - \mu}{\sigma_i}\right)^2 \Leftrightarrow \mu = \sum_i \frac{a_i}{\sigma_i^2} \Big/ \sum_i \frac{1}{\sigma_i^2}$$

Parameter uncertainties from **Bootstrap**

[Example 1: Chi2 fits]

Emulate statistical fluctuations of the data sample

1) Generate 1000 statistical replicas of the original dataset

$$y_i^{(r)} = \text{Poisson}(y_i) \qquad \sigma_i^{(r)} = \sqrt{y_i^{(r)}}$$

2) Run the fit on each replica r and calculate standard deviation + bias from all replicas

$$\sigma_{p_j} = \sqrt{\frac{1}{1000}\sum_r (p_j^{(r)} - p_j)^2} \quad B_{p_j} = \frac{1}{\sigma_{p_j}}\left(\frac{1}{1000}\sum_r p_j^{(r)} - p_j\right)$$

$$\lambda = 9.237 \pm 0.328 \qquad B_\lambda = 0.03$$

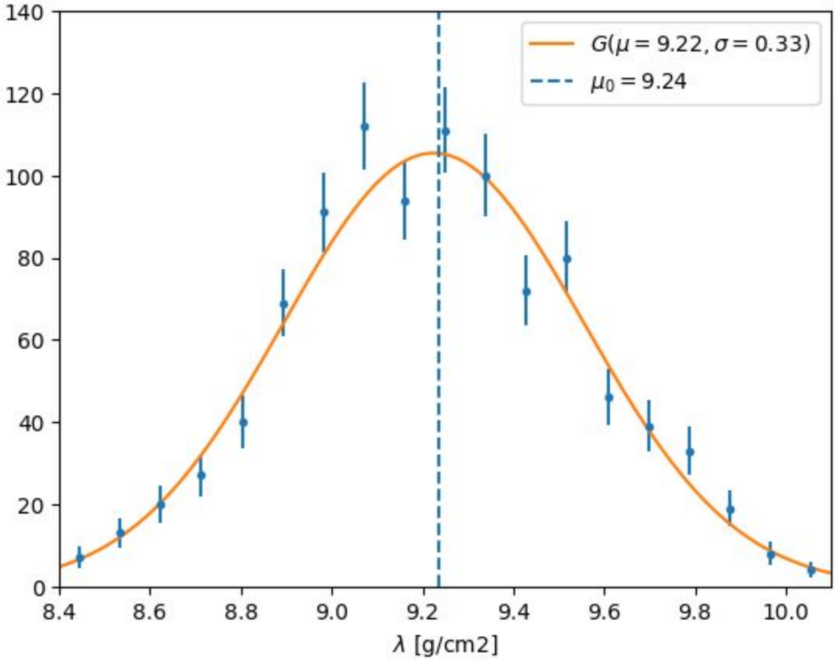
Very popular, it is typically required by the collaboration

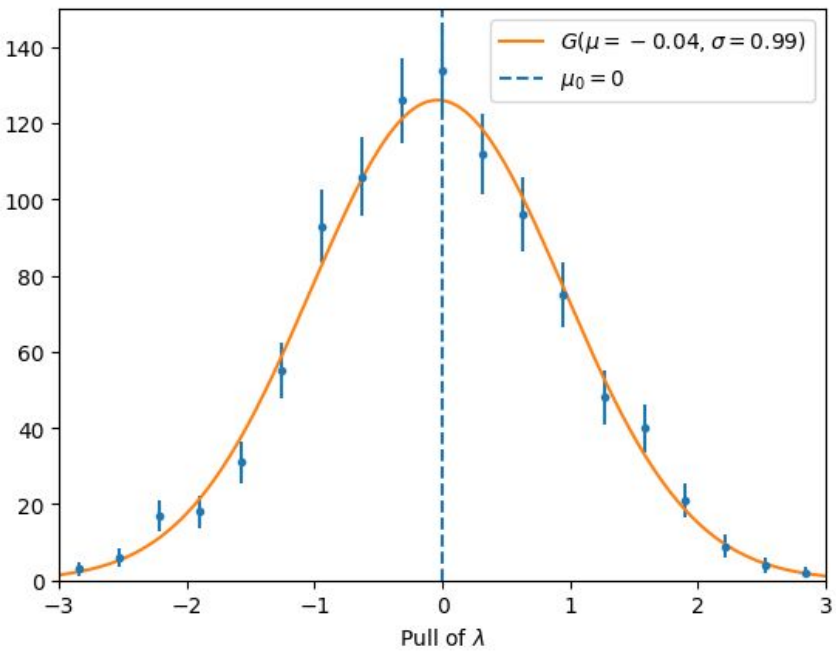
Parameter uncertainties from **Bootstrap**

Example 1: Chi2 fits

Histograms filled from replicas,
ideally they obey Gaussian distribution

$$\frac{\lambda^{(r)} - \lambda}{\sigma_\lambda}$$





Parameter uncertainties from **Error propagation**

Example 1: Chi2 fits

1) We are able to calculate parameters p of the fitted function based on the input data y

$$p_j = F_j(y_1, y_2, \ldots, y_N)$$

2) Applying standard uncertainty propagation formula (derivatives can be evaluated numerically)

$$\sigma_{p_j} = \sqrt{\left(\frac{\partial p_j}{\partial y_1}\sigma_1\right)^2 + \left(\frac{\partial p_j}{\partial y_2}\sigma_2\right)^2 + \cdots + \left(\frac{\partial p_j}{\partial y_N}\sigma_N\right)^2}$$

$$\lambda = 9.237 \pm 0.326$$

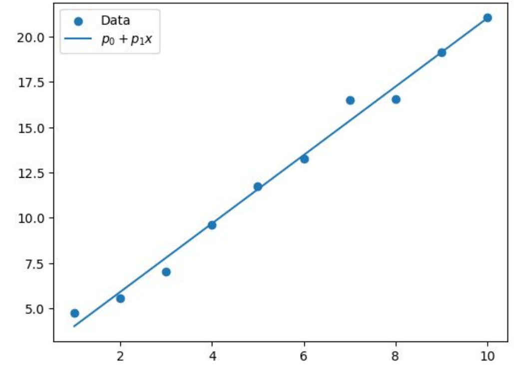
Tedious, not used much!

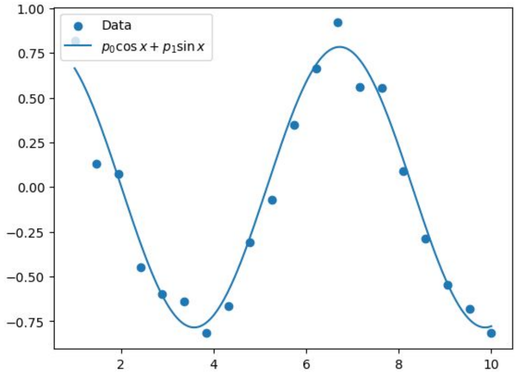
Linear regression

[Example 2: Linear regression]

"Linear" means linear in the fitted parameters, what is linear?

$$
\begin{aligned}
y &= p_0 x \\
y &= p_0 + p_1 x \\
y &= p_0 + p_1 x + p_2 x^2 \\
y &= p_0 + p_1 x + p_2 \exp(-x^2/2) \\
y &= p_0 \cos x + p_1 \sin x \\
y &= p_0 \cos(x - p_1)
\end{aligned}
$$





Linear regression ↔ Linear algebra

1) Let's assume the fitted function is a linear combination of p_j

$$\chi^2 = \sum_i \frac{1}{\sigma_i^2}\left(y_i - \sum_j A_{ij} p_j\right)^2 \quad \Rightarrow \quad \chi^2 = (y - Ap)^T V^{-1} (y - Ap)$$

Covariance matrix of y

2) The χ^2 is a quadratic form of p, it is easy to find minimum

$$\hat{p} = (A^T V^{-1} A)^{-1} A^T V^{-1} y = A^* y$$

3) Error of p can be obtained by standard error propagation

$$V_p = A^* V A^{*T} = (A^T V^{-1} A)^{-1}$$

$$H_{\chi^2} = \frac{\partial^2 \chi^2}{\partial p_i \partial p_j} = 2 A^T V^{-1} A$$

$$V_p = 2\left[\frac{\partial^2 \chi^2}{\partial p_i \partial p_j}\right]^{-1}$$

Parameter uncertainties from **χ^2(p) shape**

Any function is linear in p in the proximity of $\hat{p}$

$$f(x,p) = f(x,\hat{p}) + \sum_j \left(\frac{\partial f}{\partial p_j}\right)_{p=\hat{p}} (p_j - \hat{p}_j)$$

$$\chi^2(p) = \chi^2(\hat{p}) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 \chi^2}{\partial p_i \partial p_j}\right)_{p=\hat{p}} (p_i - \hat{p}_i)(p_j - \hat{p}_j)$$

$$V_p = 2\left[\left(\frac{\partial^2 \chi^2}{\partial p_i \partial p_j}\right)_{p=\hat{p}}\right]^{-1}$$

Example for 1D χ^2

$$\chi^2(p) = \chi^2(\hat{p}) + \frac{1}{\sigma^2}(p - \hat{p})^2$$

Notice that if:

$$\chi^2(\hat{p} \pm \sigma) = \chi^2(\hat{p}) + 1$$

Uncertainties from $\chi^2(p)$: Hesse method

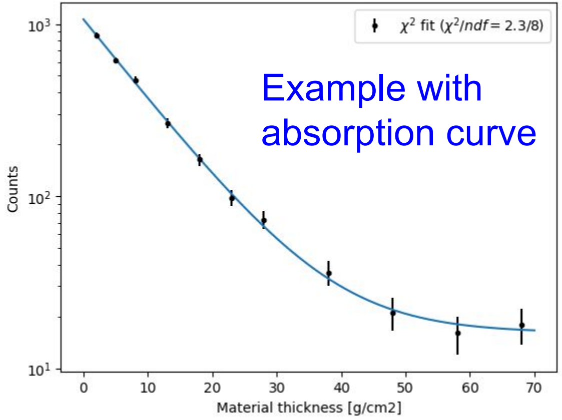
Example 1: Chi2 fits

Hesse method is a derivation of uncertainties from the matrix of the second derivatives
→ Minuit always calculates second derivatives of χ^2 to validate the minimum

$$V_p = 2\left[\left(\frac{\partial^2\chi^2}{\partial p_i \partial p_j}\right)_{p=\hat{p}}\right]^{-1}$$

Uncertainties & correlations are then:

$$\sigma_i = \sqrt{(V_p)_{ii}} \qquad c_{ij} = \frac{(V_p)_{ij}}{\sqrt{(V_p)_{ii}(V_p)_{jj}}}$$



	x0	x1	x2
x0	0.105	-7.83 **(-0.728)**	-0.42 **(-0.513)**
x1	-7.83 **(-0.728)**	1.1e+03	16 **(0.190)**
x2	-0.42 **(-0.513)**	16 **(0.190)**	6.47

	Name	Value	Hesse Error
0	x0	9.24	0.32
1	x1	1.048e3	0.033e3
2	x2	16.1	2.5

Uncertainties from χ^2(p): Minos method

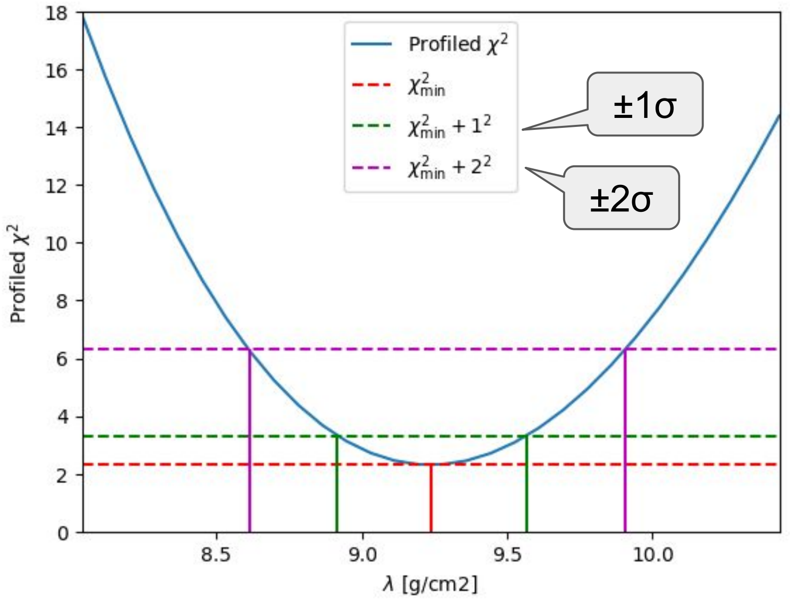
Example 1: Chi2 fits

- The χ^2 around the minima is not necessary Gaussian
  → typically investigated using profile χ^2 and $\Delta\chi^2$=1 rule
- This "graphical" approach is implemented in Minuit as Minos

	val	σ_H	σ^-_M	σ^+_M
x0	9.24	0.32	-0.32	0.33
x1	1.048e3	0.033e3	-0.033e3	0.034e3
x2	16.1	2.5	-2.6	2.5

$$\chi^2_{\text{prof}}(\hat{p} \pm \sigma) = \chi^2(\hat{p}) + 1$$
$$\chi^2_{\text{prof}}(\hat{p} \pm n\sigma) = \chi^2(\hat{p}) + n^2$$

$$\chi^2_{\text{prof}}(\lambda) = \min_{N,N_0} \chi^2(\lambda, N, N_0)$$



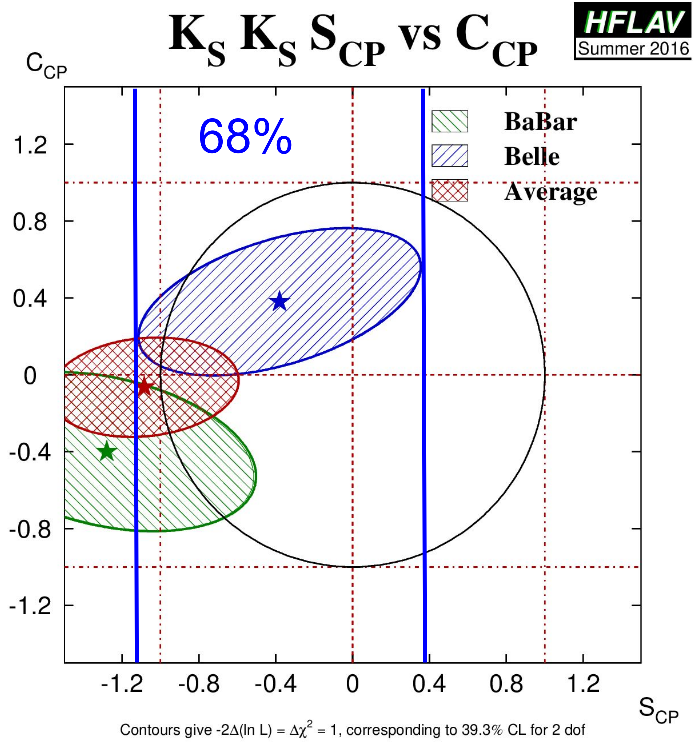
Parameter uncertainties 2D case

HFLAV

- The $\Delta\chi^2$=1 rule is also used for 2D
  →Bevere that 1σ contour
  corresponds to 39% CL

$$\int_0^1 \chi^2_2(x)dx = 0.39$$

- Contour can be also derived from the covariance matrix V_p
  →assumption of gaussian behaviour
- Always check if 39% is 68% contour is plotted



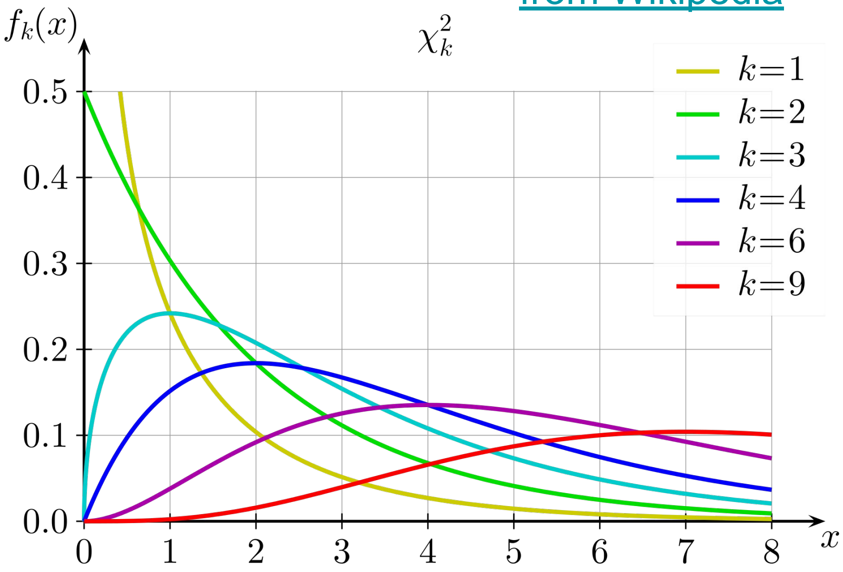
Fit quality and χ^2

- Distribution $X^2 + X^2 + \ldots + X^2$, where X is normally distributed variable
(i.e. sum over residuals)
- With more degrees of freedom it's more and more gaussian
- Some useful properties:

$$\langle \chi^2_n \rangle = n \qquad \mathrm{var}[\chi^2_n] = 2n$$

Example: $\chi^2/\mathrm{ndf} = 70/50$
(variance=100 → 2σ deviation)
scipy.stats.chi2.sf(70, 50) = 3.2%

from Wikipedia



How to judge χ^2 values?

High χ^2/ndf values (low p-values):

- (Systematic) uncertainties are underestimated
- Model does not describe data well
- Some uncertainties not considered in the χ^2 calculation

Low χ^2/ndf values (high p-values):

- (Systematic) uncertainties are overestimated
- Data are derived from the model (e.g. strong regularisation in unfolding)

Example (α_S measurement using CMS & HERA data):

$$\chi^2/\mathrm{ndf} = 1321/1118 = 1.18 \; (p = 2 \times 10^{-5})$$

$$\alpha_S(m_Z) = 0.1170 \pm 0.0014\,(\mathrm{fit}) \pm 0.0007\,(\mathrm{model}) \pm 0.0008\,(\mathrm{scale}) \pm 0.0001\,(\mathrm{param.})$$

Don't trust uncertainties if χ^2/ndf ≫ 1

Example 3: Averaging

Scenario 1: (9 ± 1), (11 ± 1)

Combined = (10.0 ± 0.7)

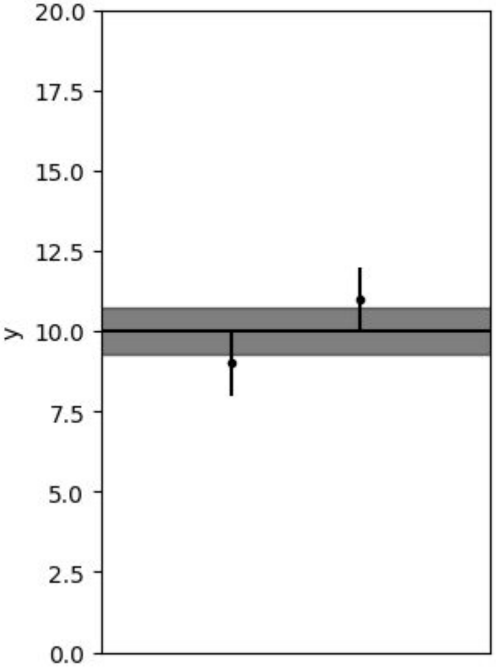
Scenario 2: (5 ± 1), (15 ± 1)

Combined = (10.0 ± 0.7)

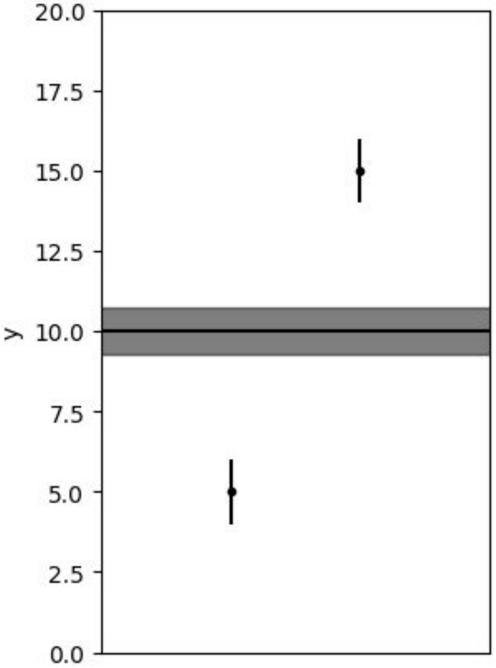
$$\mu = \sum_i \frac{a_i}{\sigma_i^2} / \sum_i \frac{1}{\sigma_i^2}$$

$$\sigma = 1/\sqrt{\sum_i \frac{1}{\sigma_i^2}}$$

χ^2/ndf=2/1



χ^2/ndf=50/1

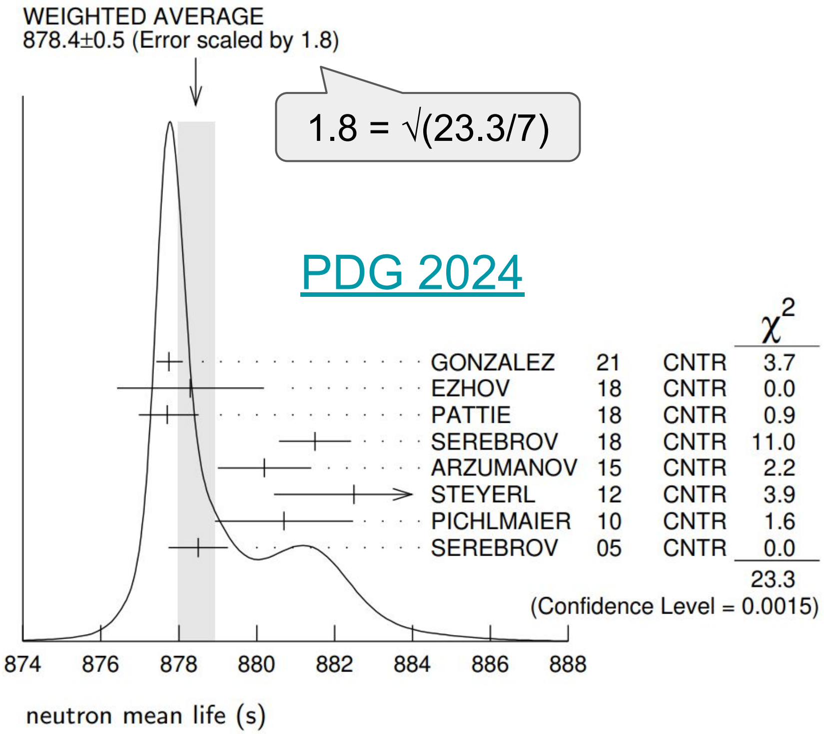


Combination of the measurements: PDG way

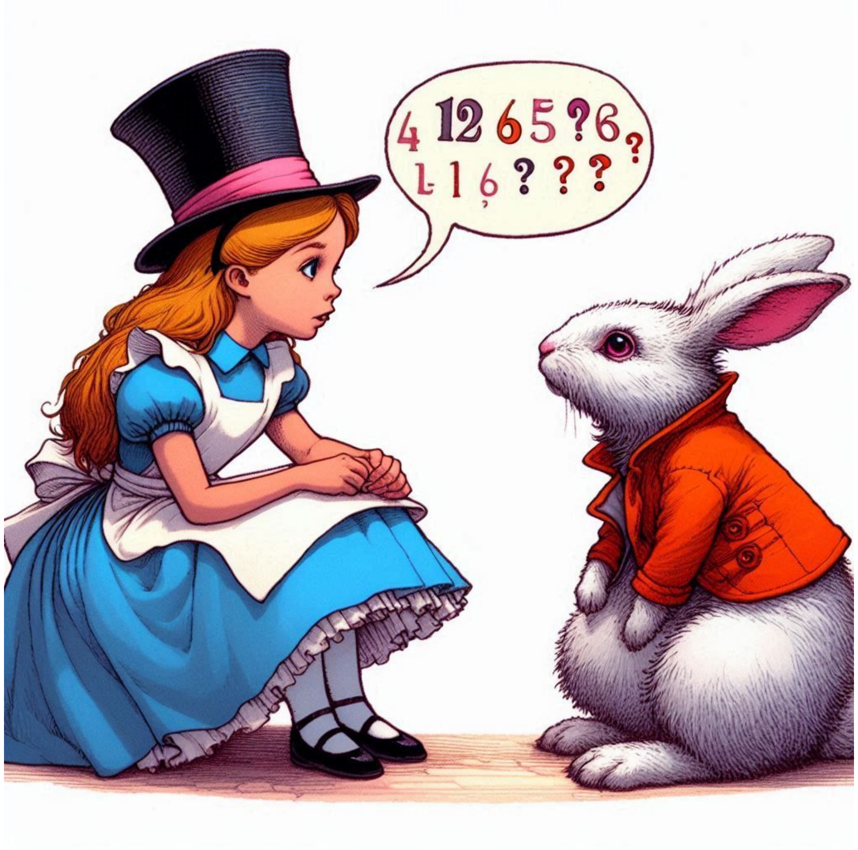
PDG Intro

1) If chi2/ndf ≤ 1, use the standard formula for error propagation of the weighted mean

2) If chi2/ndf ≫ 1, scale the uncertainties of all measurements by identical factor so that chi2'/ndf = 1 (assumption that all measurements underestimated unc. by similar factor)

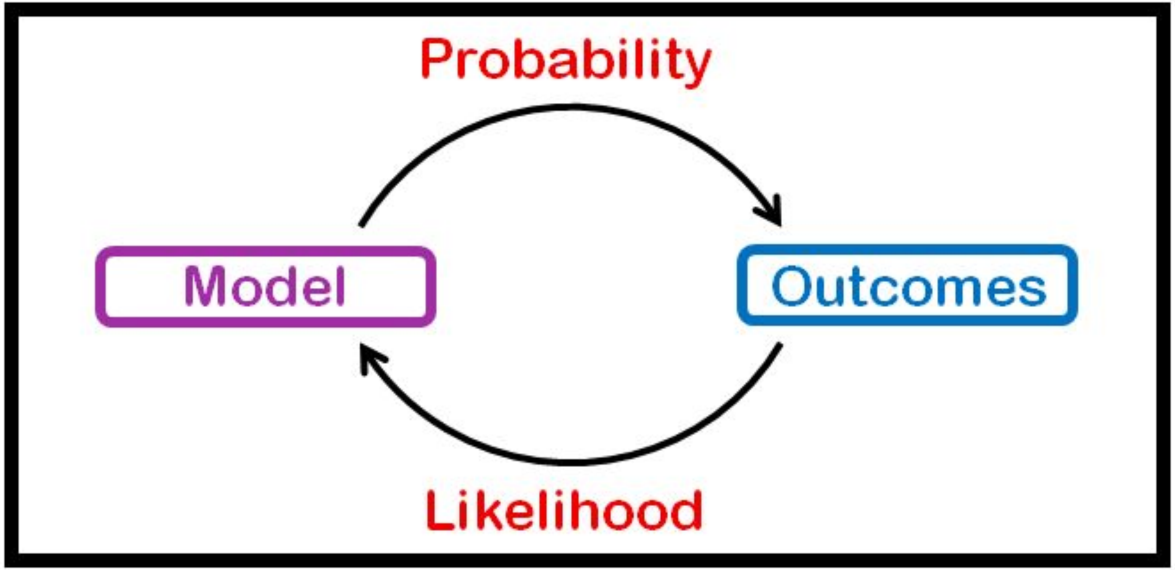


Questions to part 1?



“Alice in Wonderland asking the White Rabbit about probability”

Part 2: Fits of Probability Distribution Function

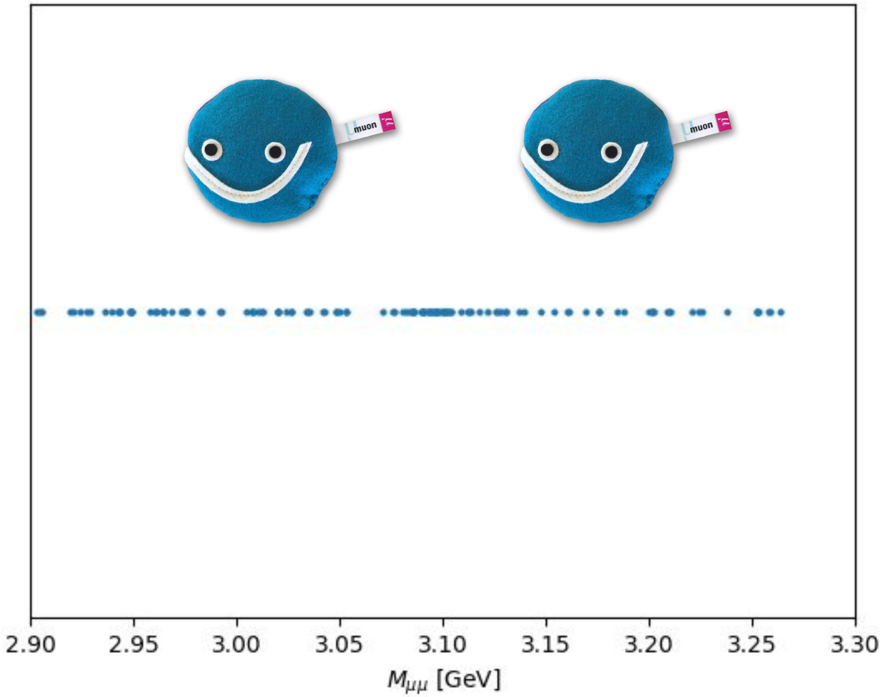


In HEP we often have “random” distribution

Measured values have to be visualized and analysed

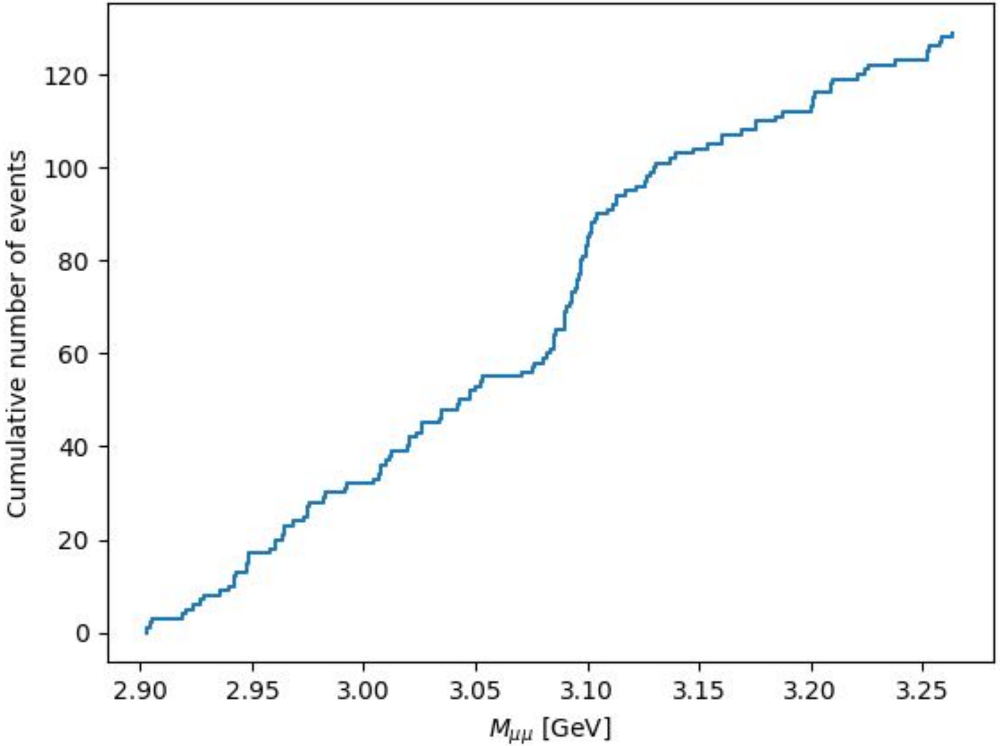
[Example 4: ML fits of mass]

Dimuon invariant mass from Belle II



Visualisation using Empirical cumulative distribution

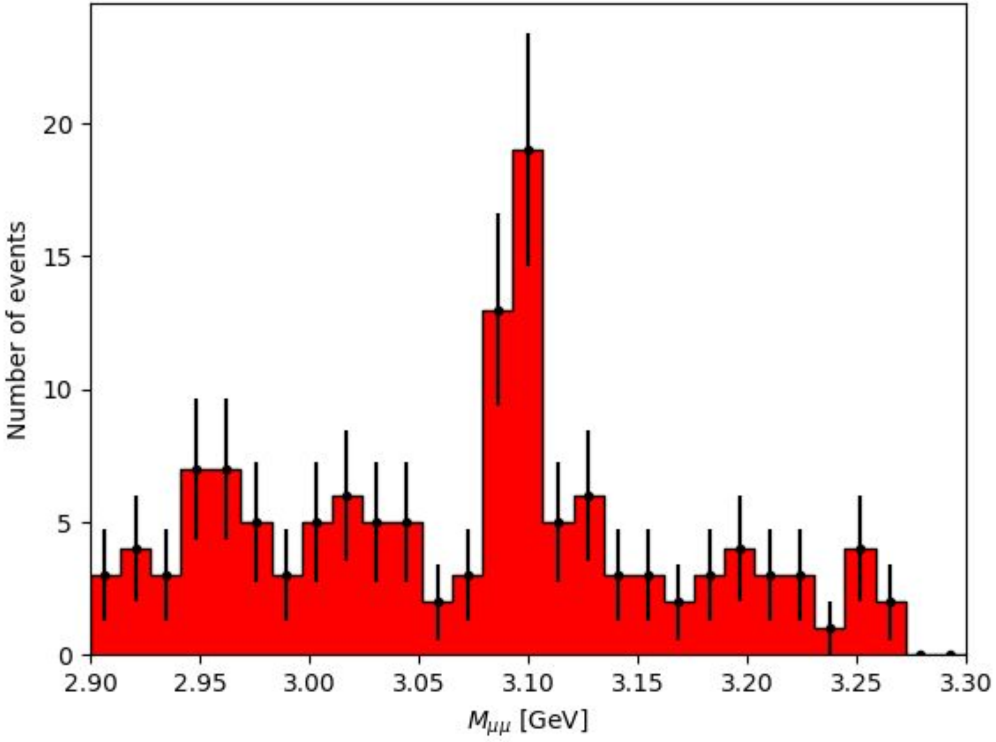
Non local!



Example 4: ML fits of mass

Visualisation using Histogram

Binning dependent!



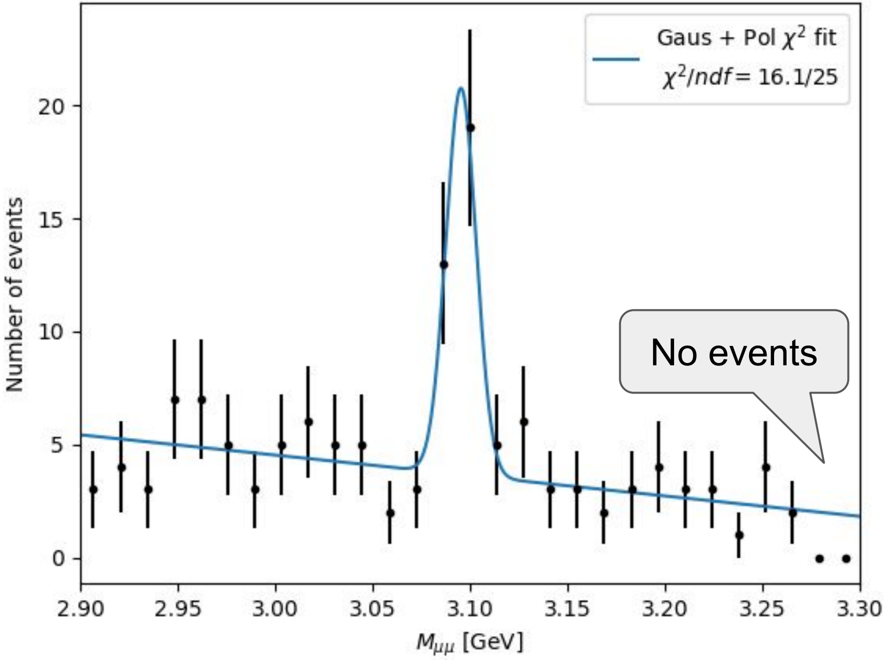
[Example 4: ML fits of mass]

Binned χ^2 fit

Example 4: ML fits of mass

- Binning dependent! Especially problematic when number of events is small
- Fast

$$f(x,p) = fG(x,\mu,\sigma) + (1-f)P_1(x,a)$$



$$\chi^2 = \sum_i \frac{(y_i - f(x_i,p))^2}{f(x_i,p)}$$

Can be also σ^2_i

Getting **maximum** from the measured events

Binning independent!

- Assuming we know from theory and detector simulation the Probability Distribution Function (PDF) of the observable x
$\rightarrow$ we still don't know exact values of parameters p

$$P(x\,|\,p) = f(x,p)$$

- We typically observe many independent events with values **x** = (x_1, x_2,…, x_n):

$$P(\boldsymbol{x}\,|\,p) = f(x_1,p)f(x_2,p)\dots f(x_n,p)$$

What is $P(p\,|\,\boldsymbol{x})$ **?**

Likelihood is all you need

- The Bayes' theorem allows to "swap" the arguments

Likelihood

Prior probability, in frequentist approach $P(p) = 1$

Posterior probability

$$P(p\,|\,\boldsymbol{x}) = \frac{P(\boldsymbol{x}\,|\,p)\;P(p)}{P(\boldsymbol{x})}$$

Normalization

- Probability that parameters have value p, given the observed data points **x** is proportional to the **Likelihood**

$$P(p\,|\,\boldsymbol{x}) \sim P(\boldsymbol{x}\,|\,p) = f(x_1, p) f(x_2, p) \ldots f(x_n, p)$$

Maximum likelihood fits

- Likelihood is defined as:

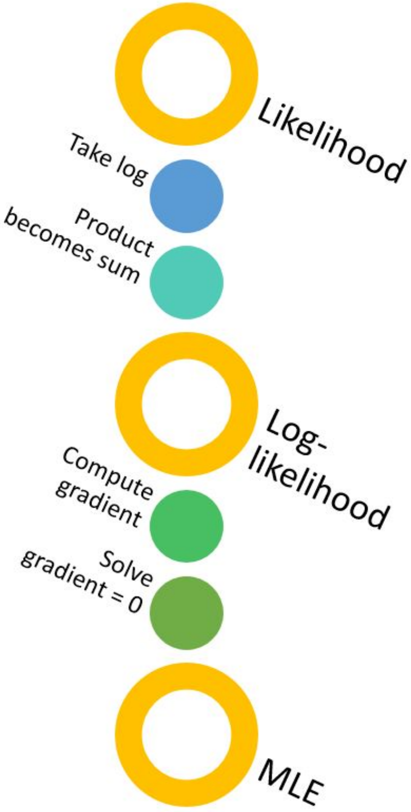
$$L(\boldsymbol{x}, p) = f(x_1, p) f(x_2, p) \dots f(x_n, p)$$

- Likelihood is maximized wrt p to find the most probable value of the parameter p̂
  →Typically done by Minuit,
  it can take time if there are many events

$$\hat{p} = \arg\max_{p} \; L(\boldsymbol{x}, p)$$

Normalize your PDF f(x,p) properly!

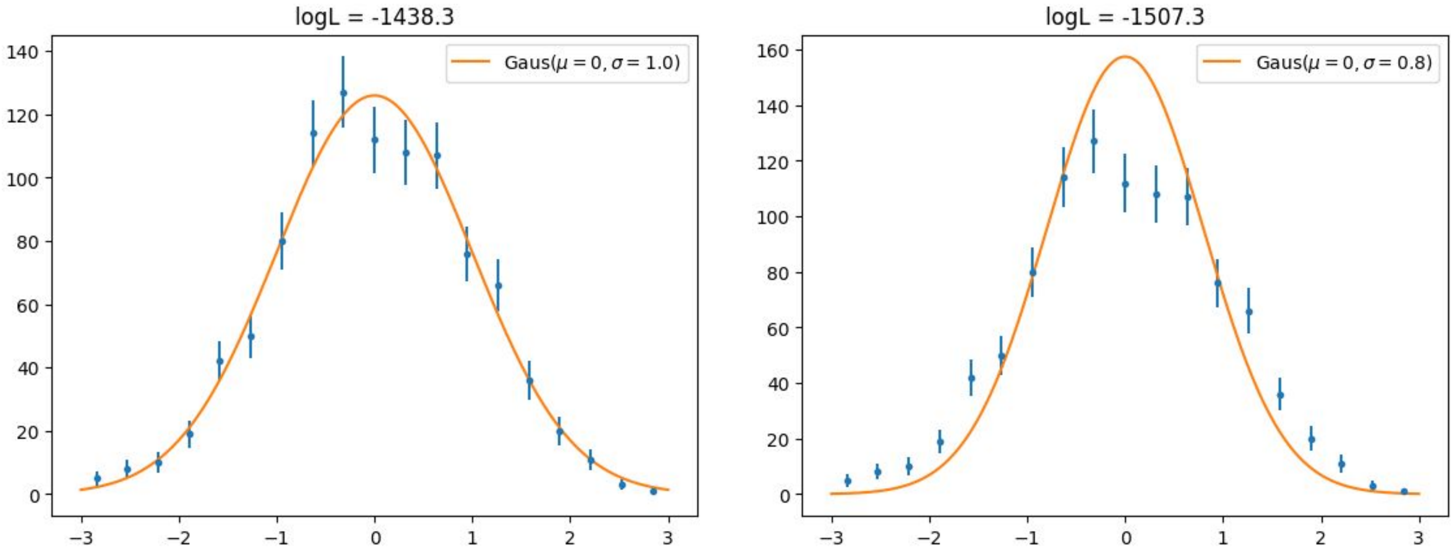
$$\int_a^b f(x,p)dx = 1$$



Likelihood & Data/Model agreement

- When **data match:**
  Higher likelihood → Better model quality

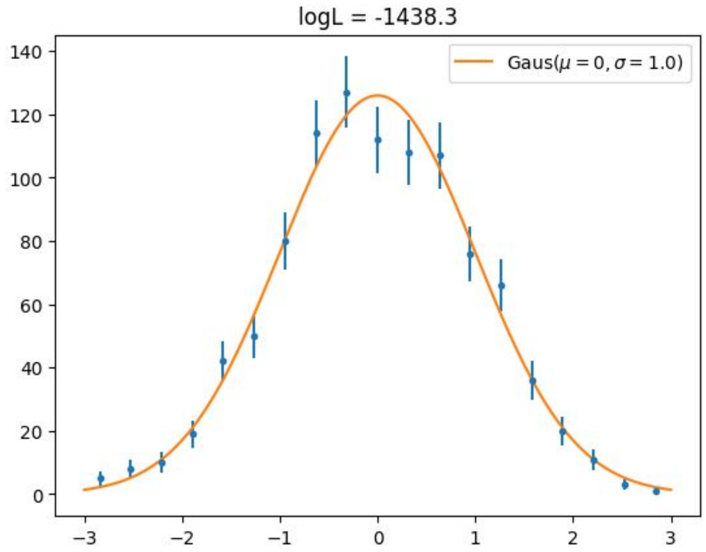
[Example 5: Likelihood values]

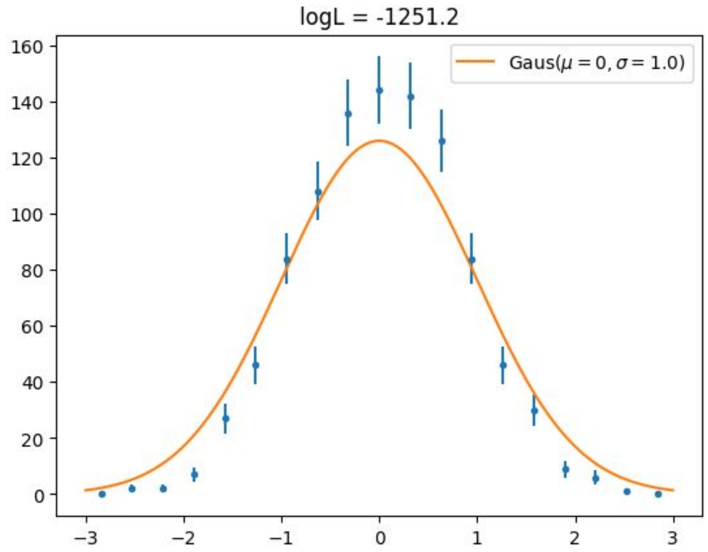


Likelihood & Data/Model agreement

- When **data don't match,**
  don't compare likelihoods

Example 5: Likelihood values





Maximum likelihood estimate: Textbook example

- If measured values obey exponential distribution

$$f(x,\lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$$

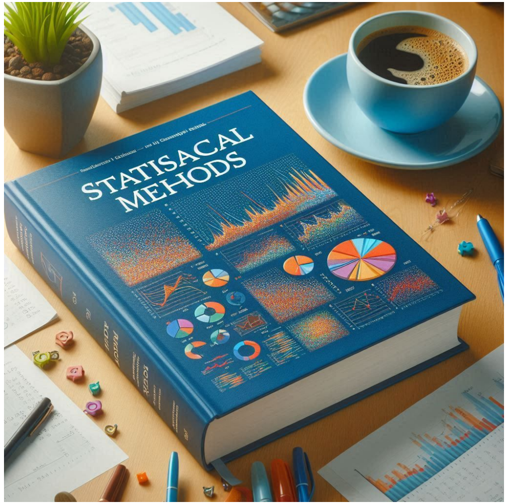
- And we have n measurements x_1, x_2, …, x_n, then

$$L(\lambda) = \frac{1}{\lambda} e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda} e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda} e^{-\frac{x_n}{\lambda}} \qquad \ln L(\lambda) = n \ln \frac{1}{\lambda} - \frac{1}{\lambda} \sum_i x_i$$

$$0 = \frac{\partial}{\partial \lambda} \ln L(\lambda) = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_i x_i \quad \Rightarrow \quad \boxed{\hat{\lambda} = \frac{1}{n} \sum_i x_i}$$

- Bias calculation:

$$\langle \hat{\lambda} - \lambda \rangle = \int dx_1 dx_2 \dots dx_n \left(\frac{1}{n} \sum_i x_i - \lambda \right) \frac{1}{\lambda} e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda} e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda} e^{-\frac{x_n}{\lambda}} = 0$$



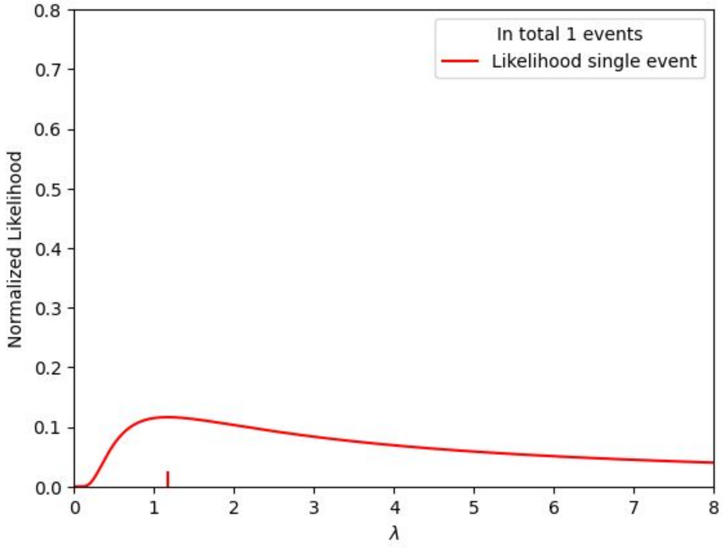
- PDF f must be normalized over the domain
- MLE can be biased, e.g. variance of Gauss

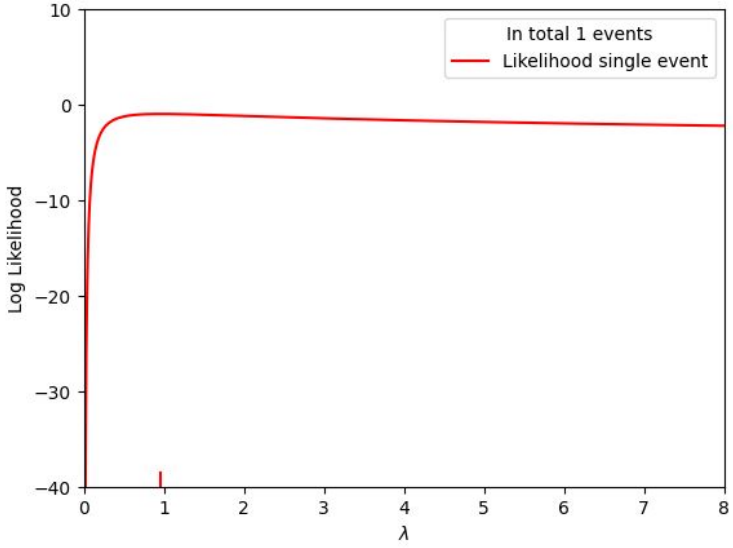
Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

Example 6: Likelihood evolution

$$L(\lambda) = \frac{1}{\lambda}e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda}e^{-\frac{x_2}{\lambda}} \cdots \frac{1}{\lambda}e^{-\frac{x_n}{\lambda}}$$



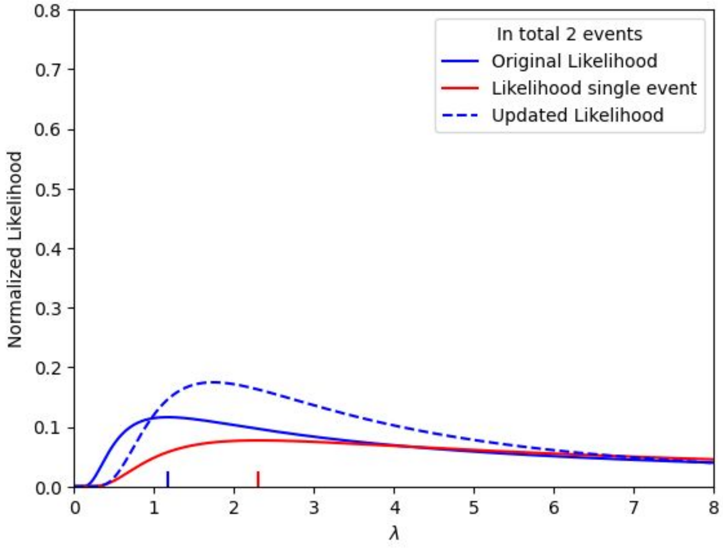


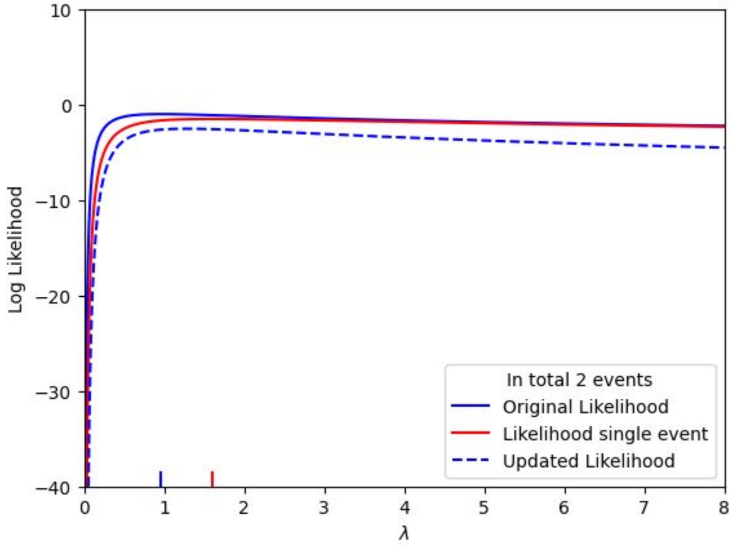
Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

Example 6: Likelihood evolution

$$L(\lambda) = \frac{1}{\lambda}e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda}e^{-\frac{x_2}{\lambda}} \ldots \frac{1}{\lambda}e^{-\frac{x_n}{\lambda}}$$



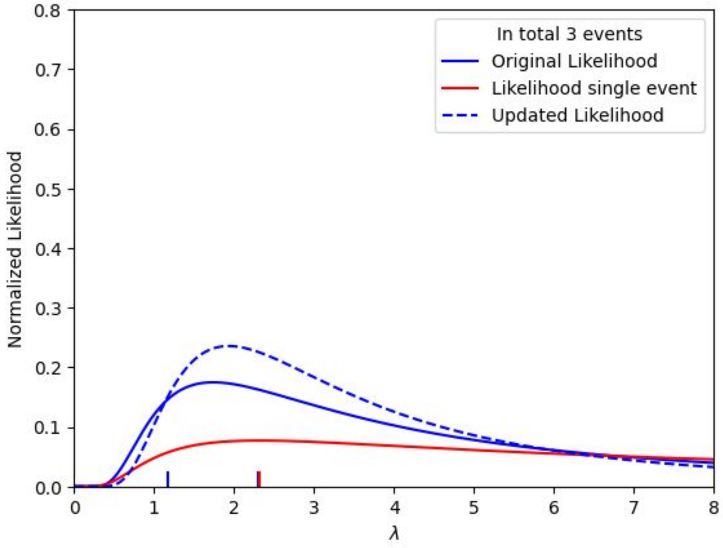


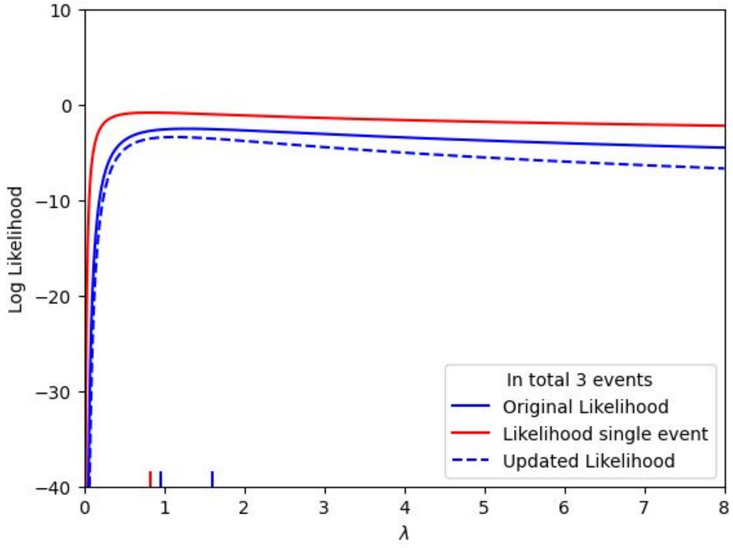
Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

[Example 6: Likelihood evolution]

$$L(\lambda) = \frac{1}{\lambda} e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda} e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda} e^{-\frac{x_n}{\lambda}}$$



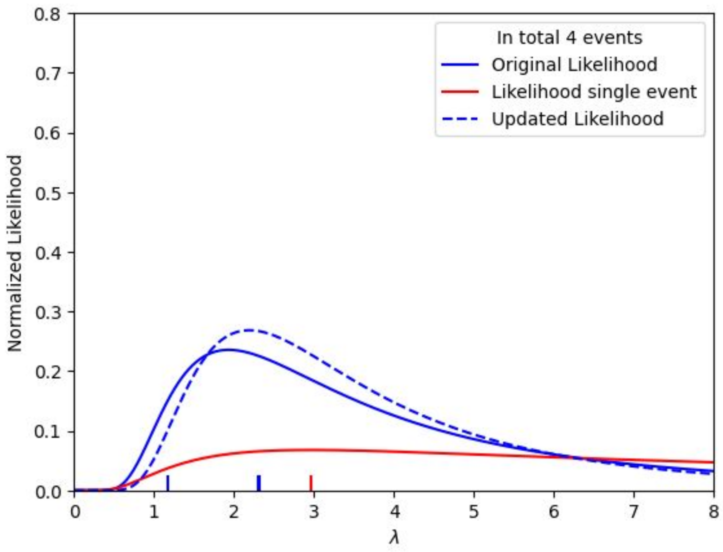


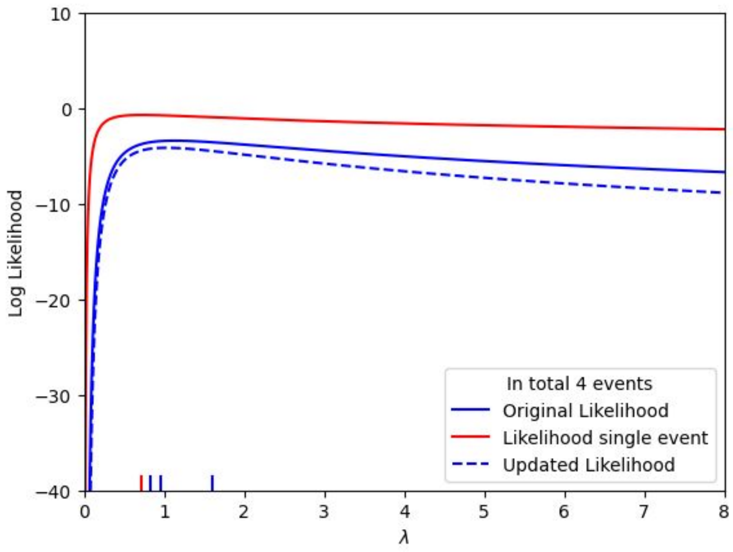
Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

Example 6: Likelihood evolution

$$L(\lambda) = \frac{1}{\lambda} e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda} e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda} e^{-\frac{x_n}{\lambda}}$$



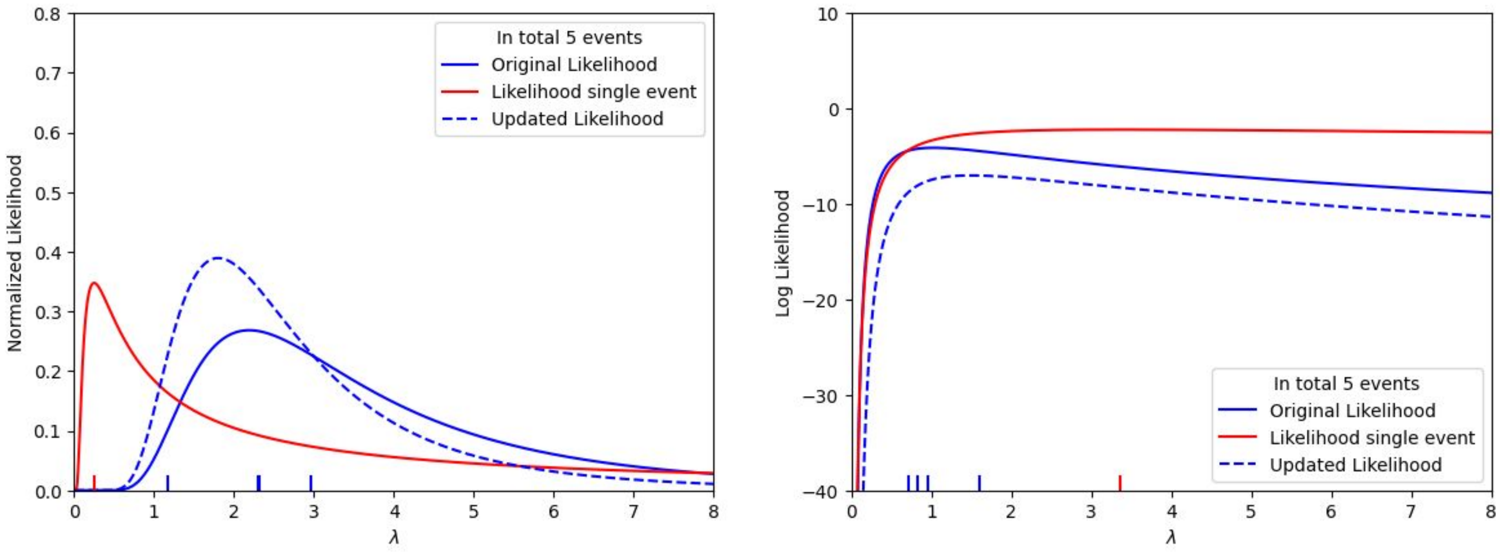


Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

[Example 6: Likelihood evolution](#)

$$L(\lambda) = \frac{1}{\lambda}e^{-\frac{x_1}{\lambda}} \frac{1}{\lambda}e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda}e^{-\frac{x_n}{\lambda}}$$

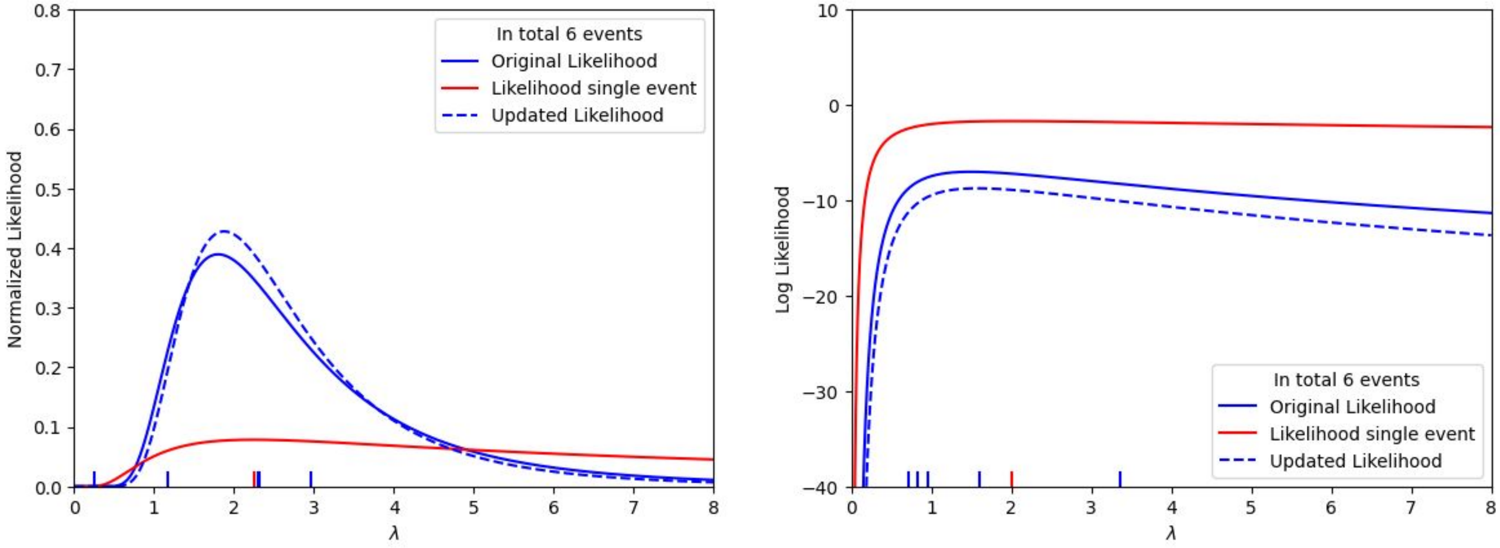


Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

Example 6: Likelihood evolution

$$L(\lambda) = \frac{1}{\lambda}e^{-\frac{x_1}{\lambda}}\ \frac{1}{\lambda}e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda}e^{-\frac{x_n}{\lambda}}$$

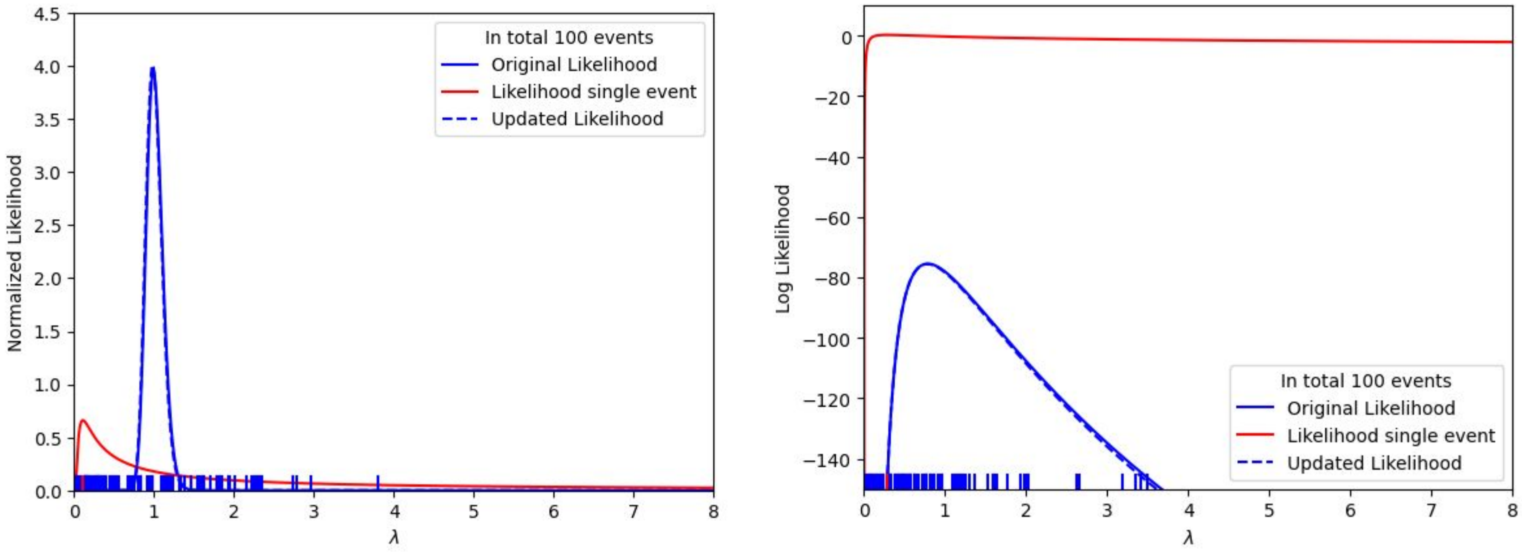


Maximum likelihood estimate: Textbook example

- Likelihood evolution, when events are collected

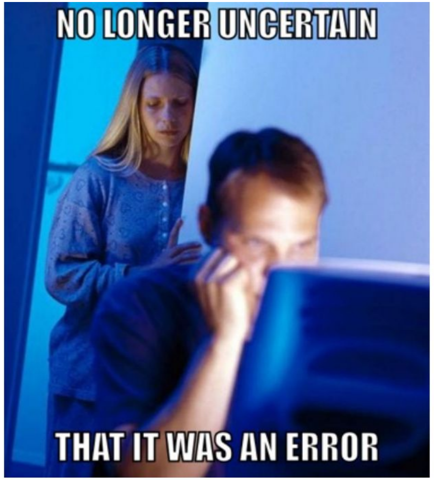
Example 6: Likelihood evolution

$$L(\lambda) = \frac{1}{\lambda}e^{-\frac{x_1}{\lambda}}\ \frac{1}{\lambda}e^{-\frac{x_2}{\lambda}} \dots \frac{1}{\lambda}e^{-\frac{x_n}{\lambda}}$$



Uncertainties of the ML estimates

Uncertainty ~ known unknowns



- Likelihood gets more and more Gaussian with increasing number of events

$$L(p) = L(\hat{p}) \exp\left(-\frac{(p-\hat{p})^2}{2\sigma_p^2}\right)$$

$$\ln L(p) = \ln L(\hat{p}) - \frac{(p-\hat{p})^2}{2\sigma_p^2}$$

From Hesse Matrix

$$V_{\hat{p}} = -\left[\left(\frac{\partial^2 \ln L(p)}{\partial p_i \partial p_j}\right)_{p=\hat{p}}\right]^{-1}$$

"Graphical" method

$$\ln L(\hat{p} \pm \sigma) = \ln L(\hat{p}) - \frac{1}{2}$$

Textbook example: Uncertainty of λ

- Let's assume that the measured values obey exponential distribution

$$f(x,\lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \qquad \ln L(\lambda) = n \ln \frac{1}{\lambda} - \frac{1}{\lambda} \sum_i x_i$$

- ML estimate for λ is from first derivative

$$0 = \frac{\partial}{\partial \lambda} \ln L(\lambda) = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_i x_i \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{n} \sum_i x_i$$

- Uncertainty of lambda using Hesse method

$$V_p = \left(\frac{\partial^2}{\partial \lambda^2} \ln L(\lambda)|_{\lambda=\hat{\lambda}} \right)^{-1} = \frac{1}{n} \hat{\lambda}^2 \qquad \frac{\partial^2}{\partial \lambda^2} \ln L(\lambda) = \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_i x_i$$

$$\boxed{\sigma_{\hat{\lambda}} = \frac{1}{\sqrt{n}} \hat{\lambda}}$$

Relative uncertainty goes like 1/√n

Relation between Maximum likelihood and χ^2 fits

- From the Likelihood of the residuals which are assumed to obey Normal distribution

$$L(p) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y_1 - f(x_1,p))^2}{\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y_2 - f(x_2,p))^2}{\sigma_2^2}} \cdots \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y_n - f(x_n,p))^2}{\sigma_n^2}}$$

$$-2\ln L(p) + C = \sum_i \frac{(y_1 - f(x_1,p))^2}{\sigma_1^2} = \chi^2$$

Hesse method

$$V_p = 2\left[\left(\frac{\partial^2 \chi^2}{\partial p_i \partial p_j}\right)_{p=\hat{p}}\right]^{-1}$$

$$V_{\hat{p}} = 2\left[\left(\frac{\partial^2 -2\ln L}{\partial p_i \partial p_j}\right)_{p=\hat{p}}\right]^{-1}$$

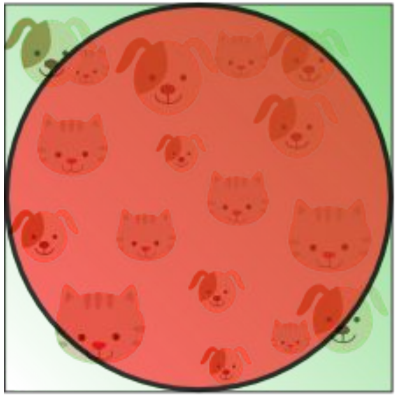
“Graphical” method

$$\chi^2(\hat{p} \pm \sigma) = \chi^2(\hat{p}) + 1$$

$$-2\ln L(\hat{p} \pm \sigma) = -2\ln L(\hat{p}) + 1$$

Binned Maximum likelihood fits

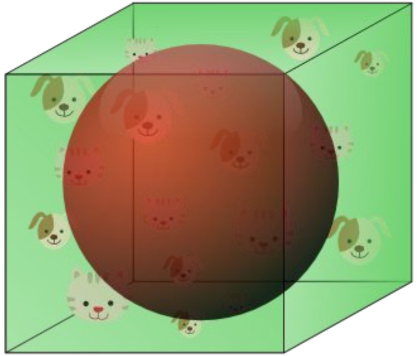
Curse of dimensionality



- When one replace Gauss by Poisson for each bin
- Approaches to unbinned ML for infinity bins
  (bins with zero number of entries are not problem)
- Equivalent to discretisation of observed variable
- Faster fitting → getting prior for unbinned fit

$$L(p) = e^{-f(x_1,p)}\frac{(f(x_1,p))^{y_1}}{y_1!}\, e^{-f(x_2,p)}\frac{(f(x_2,p))^{y_2}}{y_2!} \dots e^{-f(x_n,p)}\frac{(f(x_n,p))^{y_n}}{y_n!}$$

$$\ln L(p) = -\sum_i f(x_i,p) + \sum_i y_i \ln f(x_i,p) + C$$



For both examples (ɣ-absorption fit and $M_{\mu\mu}$ fit) Gaus can be replaced by Poisson

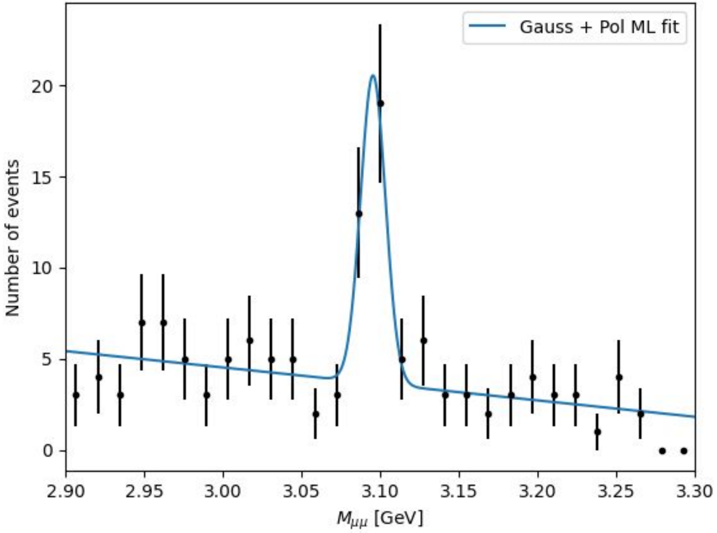
Example 1: Chi2 fits

Fitting $M_{\mu\mu}$ using ML method

[Example 4: ML fits of mass](#)

- Minimizing -2logL using Minuit
- Minos gives asymmetric unc.

$$f(x,p) = fG(x,\mu,\sigma) + (1-f)P_1(x,a)$$



0..μ, 1..σ, 2..a, 3..f

Migrad

FCN = -292 Nfcn = 409

EDM = 3.63e-07 (Goal: 0.0002)

Valid Minimum	Below EDM threshold (goal x 10)
No parameters at limit	Below call limit
Hesse ok	Covariance accurate

	Name	Value	Hesse Error	Minos Error-	Minos Error+	Limit-	Limit+	Fixed
0	x0	3.0949	0.0020	-0.0020	0.0021			
1	x1	0.0076	0.0019	-0.0017	0.0021			
2	x2	-0.291	0.008	-0.006	0.011			
3	x3	0.21	0.05	-0.04	0.05			

Signal fraction

	x0		**x1**		**x2**		**x3**	
Error	-0.0020	0.0021	-0.0017	0.0021	-0.006	0.011	-0.04	0.05
Valid	True	True	True	True	True	True	True	True
At Limit	False	False	False	False	False	False	False	False
Max FCN	False	False	False	False	False	False	False	False
New Min	False	False	False	False	False	False	False	False

	x0	**x1**	**x2**	**x3**
x0	4.01e-06	0e-6 **(0.082)**	-0e-6 **(-0.020)**	2e-6 **(0.020)**
x1	0e-6 **(0.082)**	3.64e-06	0e-6 **(0.004)**	31e-6 **(0.350)**
x2	-0e-6 **(-0.020)**	0e-6 **(0.004)**	6.49e-05	0 **(0.008)**
x3	2e-6 **(0.020)**	31e-6 **(0.350)**	0 **(0.008)**	0.0021

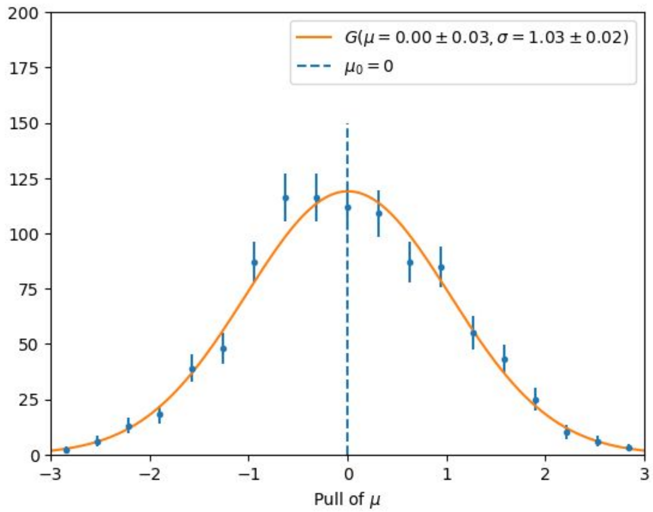
Toy from **PDF** vs Hesse uncertainties

[Example 4: ML fits of mass]

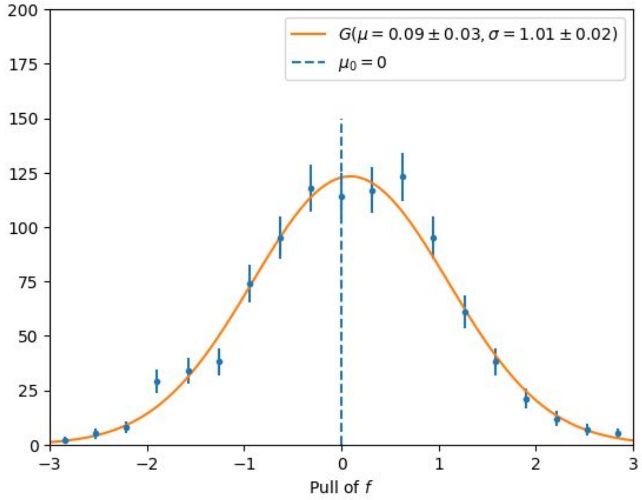
$$\text{pull} = \frac{p^{(r)} - \hat{p}}{\sigma_{\hat{p}}}$$

Toy pseudo-experiment:

1) Generate n from Poisson distribution with $\lambda = n_{\text{Data}}$
2) Generate n event from the PDF obtained at previous slide
3) Run the the ML fit on these events



Measured f_{sig} tends to be slightly higher



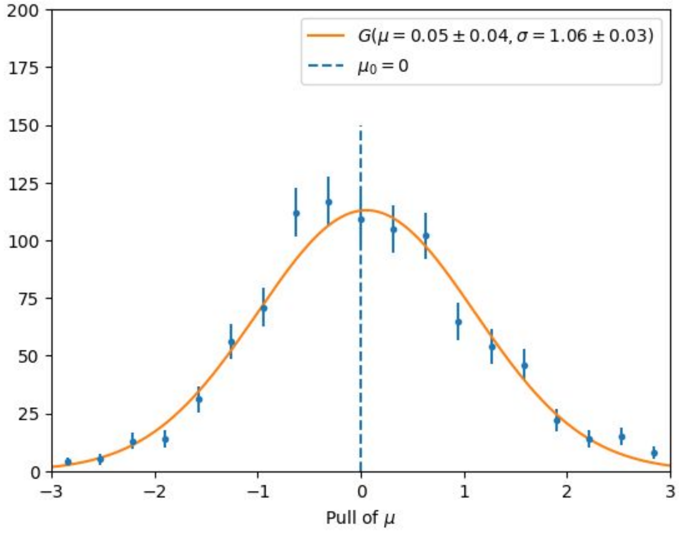
Bootstrap from **Data** vs Hesse uncertainties

Example 4: ML fits of mass

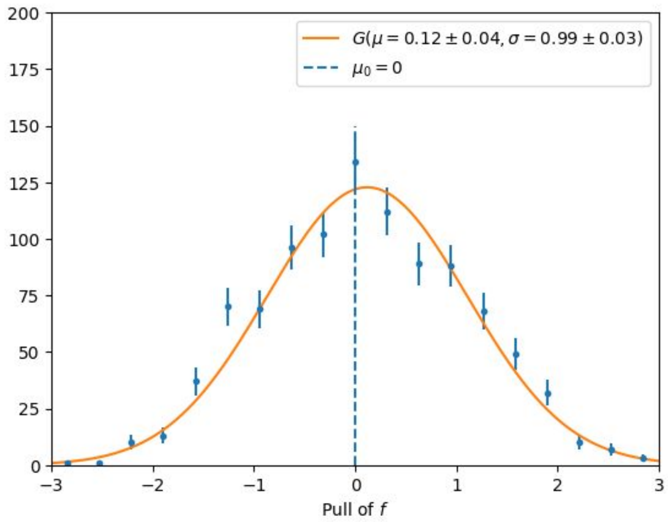
$$\text{pull} = \frac{p^{(r)} - \hat{p}}{\sigma_{\hat{p}}}$$

Bootstrap replica:

1) Generate n from Poisson distribution with $\lambda = n_{Data}$
2) Randomly pick n events from the data set (events can repeat)
3) Run the the ML fit on these events



Measured f_{sig} tends to be slightly higher

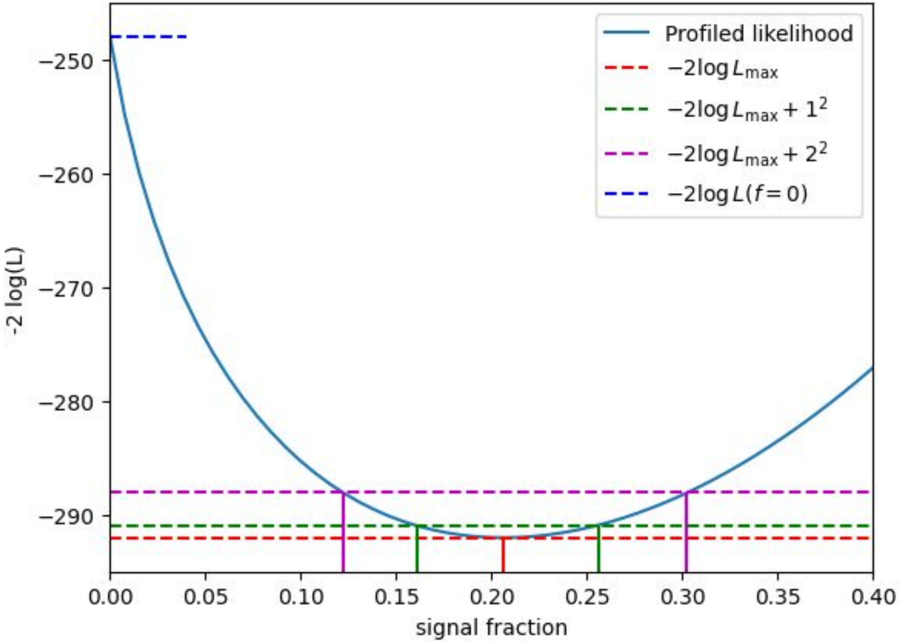


Profile Likelihood scan of f_{Sig}

Example 4: ML fits of mass

- Deriving uncertainties by "graphical method"
  → +1, $+2^2$, $+3^2$… rule for -2 logL
  (in analogy to chi2)
- This approach called Minos in iminuit
- Notice that f=0 is special as profiling is effectively done only over BG parameter a
  (drop in effective N_{df})

$$L_{\text{prof}}(f) = \max_{\mu,\sigma,a} L(\mu, \sigma, a, f)$$



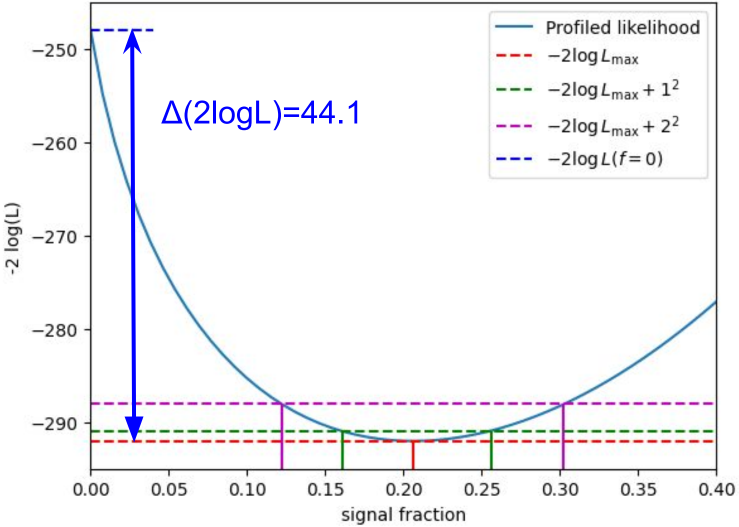
Can we claim discovery: ML ratio test

Example 4: ML fits of mass

- If the PDFs for signal and BG are known, the ML ratio is the most powerful discriminator
- If data obey BG-only hypothesis, the λ_{LR} behaves as $\chi^2_{n_p}$ (for large #events, Wilks theorem)

$$\lambda_{\mathrm{LR}} = 2 \ln \frac{\sup_{p \in \mathrm{Sig+BG}} L(p)}{\sup_{p \in \mathrm{BG}} L(p)}$$

λ_{LR} = 44.11 (n_p = 3)
p = scipy.stats.chi2.sf(44.11, 3) = 1.4e-9
scipy.stats.chi2.sf(6.05^2, 1) = 1.4e-9

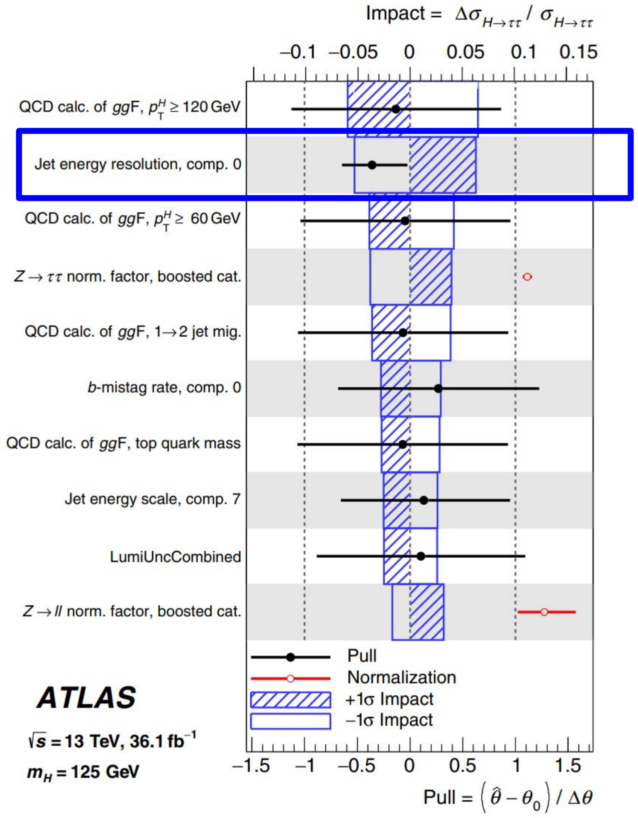


6σ significance→ discovery of J/Ψ

Systematic uncertainties and nuisance parameters

Phys.Rev.D 99 (2019)

- Most parameters of the model are typically detector related and of the technical nature (unpublished)
  → using profiling to maximize L over these parameters
- Systematic variations can be:
  → Treated out of the likelihood
  → Included into the likelihood via nuisance parameters



$$L = \frac{1}{\sqrt{2\pi}\sigma_{K^{\rm jet}}} \exp\left(-\frac{(K^{\rm jet} - K_0^{\rm jet})^2}{2(\sigma_{K^{\rm jet}})^2}\right) \prod_i f(x_i; \sigma_{H\to\tau\tau}, K^{\rm jet})$$

Extended Maximum Likelihood

Example 4: ML fits of mass

- Adding total number of events as parameter into the Likelihood
- Important if there is external constraint for total number of events (e.g. from luminosity)
  → otherwise fit results are identical

$$L = e^{-\nu}\frac{\nu^n}{n!}\prod_i^n f(x_i, p)$$

$$L = e^{-\nu}\frac{1}{n!}\prod_i^n \left[\nu f f_\mathrm{s}(x_i, p) + \nu(1-f) f_\mathrm{b}(x_i, p)\right]$$

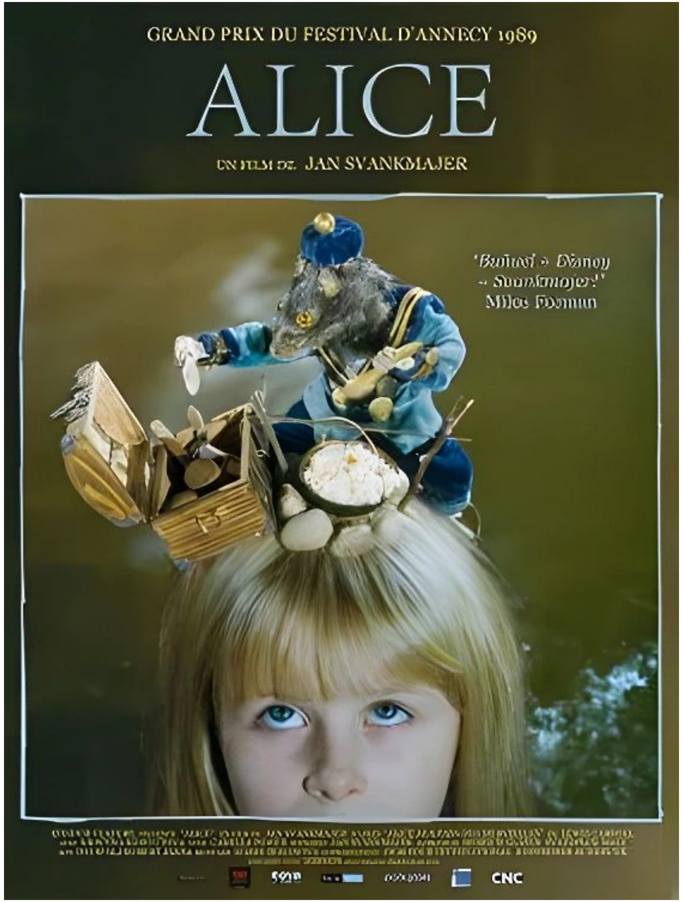
$$\begin{aligned} \nu &= 129.0 \pm 11.4 \\ f &= 0.206 \pm 0.045 \end{aligned}$$

$$L = e^{-(N_\mathrm{s}+N_\mathrm{b})}\frac{1}{n!}\prod_i^n \left[N_\mathrm{s} f_\mathrm{s}(x_i, p) + N_\mathrm{b} f_\mathrm{b}(x_i, p)\right]$$

$$\begin{aligned} N_\mathrm{s} &= 26.6 \pm 6.4 \\ N_\mathrm{b} &= 102.4 \pm 10.8 \end{aligned}$$

[Alice (1988)]

Questions to part 2?



What to remember

- The χ^2 fits used for systematic-dominated measurements and for xy fits
- Likelihood fits are binning-independent → important for small statistics
- Likelihood-ratio test is the standard way to claim discoveries