Statistical Methods in Particle Physics

5. Maximum Likelihood Estimation

Heidelberg University, WS 2023/24

Klaus Reygers, Martin Völkl (lectures)
Ulrich Schmidt, (tutorials)

Reminder: Evaluating Estimator Performance

Consistency:

- Does the estimate converge to the true value?

$$\lim_{n\to\infty} \hat{\theta} = \theta$$

Bias:

- Does the average of many measurements converge towards the true value? Otherwise: bias b

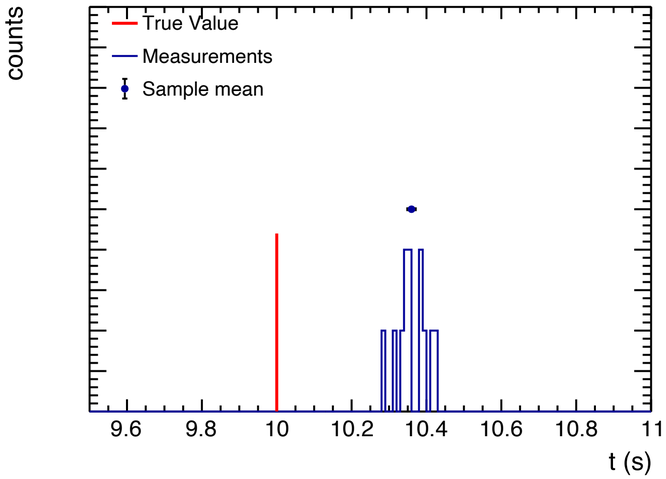
$$E[\hat{\theta}] = \theta$$

Efficiency:

- How small is the uncertainty for a given amount of data and how fast does it decrease with n?

Robustness:

- Does the estimator still work if we are slightly wrong about the assumptions of the data (e.g. in the presence of rare outliers)?



Example: Estimators for the lifetime of a particle

Estimator	Consistent?	Unbiased?	Efficient?
$\hat{\tau} = \frac{t_1+t_2+...+t_n}{n}$	yes	yes	yes
$\hat{\tau} = \frac{t_1+t_2+...+t_n}{n-1}$	yes	no	no
$\hat{\tau} = t_1$	no	yes	no

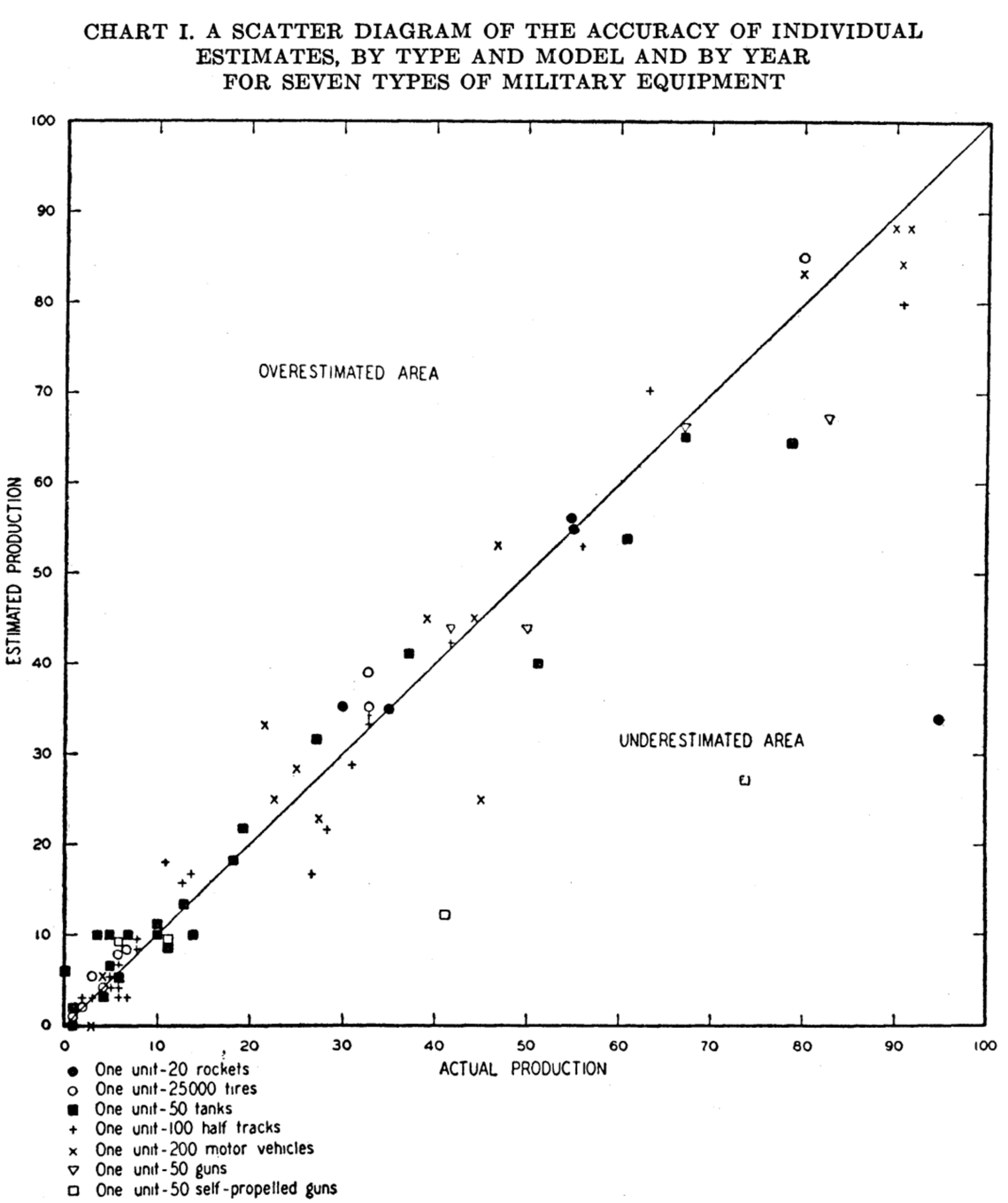
http://www.terascale.de/e149980/index_eng.html

Drawing random numbers from a uniform distribution

- Draw k numbers from an even distribution between 1 and N
- The MVUE is $\hat{N} = (1 + k^{-1})m - 1$, where m is the largest number in the sample
- The variance of the estimator is $V[\hat{N}] \approx \frac{N^2}{k^2}$
- Known as the *German tank problem*
- How to estimate the total number of produced units from the serial numbers of a few investigated ones?

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342

https://en.wikipedia.org/wiki/German_tank_problem



R. Ruggles and H. Brodie, *An Empirical Approach to Economic Intelligence in World War II*

The end of the world

The doomsday argument

- Make N the total number of humans that have been and will ever been born
- Now number them consecutively by date of birth
- You are one random sample from this distribution (m)

- $\hat{N} = (1 + k^{-1})m - 1$

- m is about 100b. So we can estimate that around 200b people will ever exist

- Simple argument: If you are randomly one of all people who are ever born, then you are likely somewhere in the middle of the numbered list (and unlikely to be near the beginning or the end)

Unbiased estimators for mean and variance

Consider n independent and identically distributed measurements x_i drawn from a distribution with mean μ and standard deviation σ:

Estimator for the mean: $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mathrm{E}[\hat{\mu}] = \frac{1}{n}\mathrm{E}[\sum_i x_i] = \frac{1}{n}\sum_i \mathrm{E}[x_i] = \mu \quad \rightarrow \text{ estimator is unbiased}$$

$$\mathrm{V}[\hat{\mu}] = \mathrm{V}[\frac{1}{n}\sum_i x_i] = \frac{1}{n^2}\mathrm{V}[\sum_i x_i] = \frac{1}{n}\mathrm{V}[x] = \frac{\sigma^2}{n}, \text{ i.e., } \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

Unbiased estimator for the variance: $$s^2 := \hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Unbiased estimator of the variance: Derivation (1)

Consider n independent and identically distributed random variables x_i:

$$\mu := E[x_i], \quad \sigma^2 := V[x_i], \quad \bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

We'll use:

$$\sigma^2 = E[x_i^2] - \mu^2 \quad \rightsquigarrow \quad E[x_i^2] = \mu^2 + \sigma^2$$

$$V[\bar{x}] = \frac{1}{n^2} V[\sum_{i=1}^{n} x_i] = \frac{1}{n} V[x_i] = \frac{\sigma^2}{n} \overset{!}{=} E[\bar{x}^2] - \mu^2 \quad \rightsquigarrow \quad E[\bar{x}^2] = \frac{\sigma^2}{n} + \mu^2$$

Now we calculate the expectation value of $\sum_{i=1}^{n}(x_i - \bar{x})^2$:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2$$

$$E[\sum_{i=1}^{n}(x_i - \bar{x})^2] = E[\sum_{i=1}^{n} x_i^2] - E[n\bar{x}^2] = n(\mu^2 + \sigma^2) - \sigma^2 - n\mu^2 = (n-1)\sigma^2$$

Unbiased estimator of the variance: Derivation (2)

This means that

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is an unbiased estimator of the variance, i.e., $E[s^2] = \sigma^2$

Multiplying the sample variance by $n/(n-1)$ is known as Bessel's correction.

Note that s is not an unbiased estimator of the standard deviation:
https://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation

Unbiased estimator for the standard deviation for the normal distribution ($E[\hat{\sigma}] = \sigma$):

$$\hat{\sigma} = c_4(n)\sqrt{s^2}, \quad c_4(n) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} = 1 - \frac{1}{4n} - \frac{7}{32n^2} + \ldots,$$

Rule of thumb:

$$\hat{\sigma} \approx \sqrt{\frac{1}{n-1.5}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Reminder: Probability of the data ⇄ likelihood

- $p(\vec{d}\,|\,\vec{\theta})$ is the probability distribution of the data for different parameters
- When considered as a function of $\vec{\theta}$ instead, it is called the *likelihood*
- Often called $\mathscr{L}$ or L with $\mathscr{L}(\vec{\theta}\,|\,\vec{d}) \equiv p(\vec{d}\,|\,\vec{\theta})$

Likelihood function is not a probability density function

The integral of $L(\vec{x}, \vec{\theta})$ with respect to the parameter is not necessarily equal to unity ($L(\vec{x}, \vec{\theta})$ might not be integrable at all).

This is why $L(\vec{x}, \vec{\theta})$ is not a probability density function.

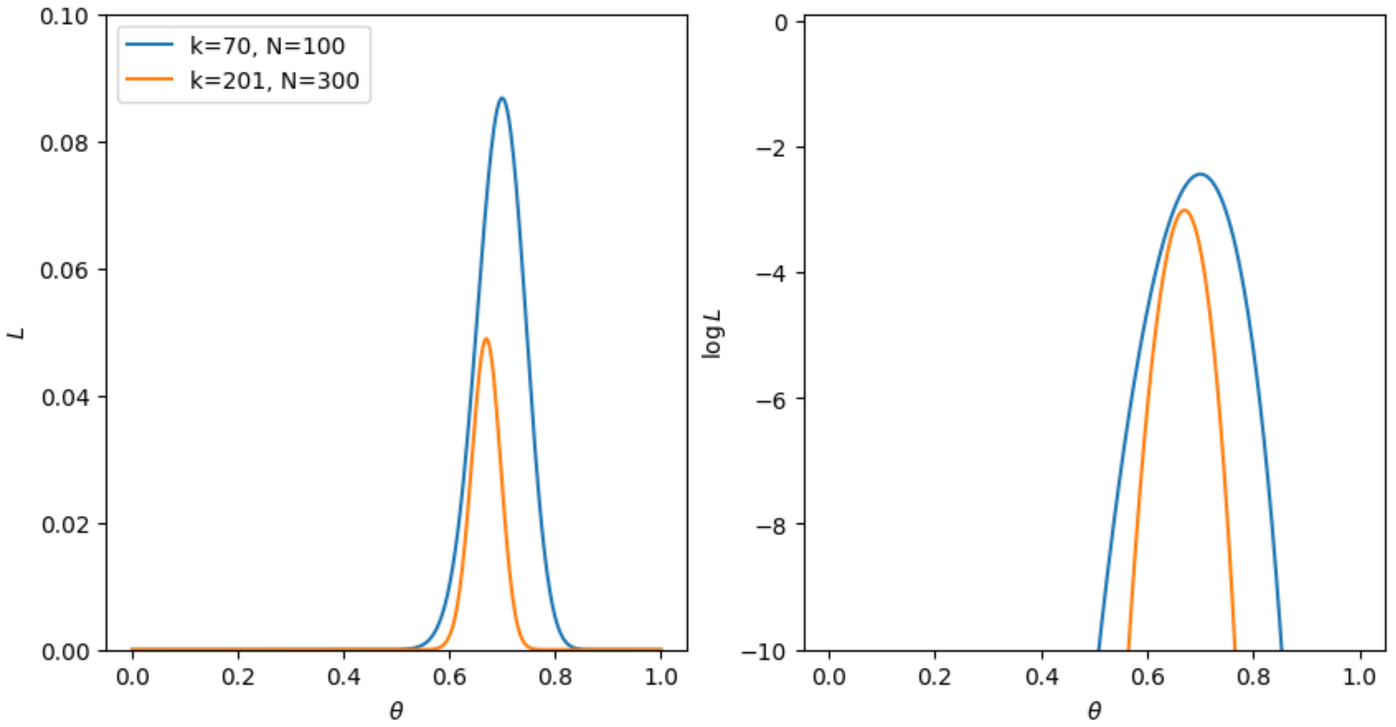
Example: exponential decay, one measurement at t = 1h.

$$L(\tau) = \frac{1}{\tau} e^{-t/\tau} \approx \frac{1}{\tau} \quad \text{as} \quad \tau \to \infty, \qquad \int_0^\infty L(\tau)\, d\tau \quad \text{not defined}$$

Note: With Jeffreys' prior 1/τ the posterior $L(\tau)\ \pi(\tau)$ is normalizable.

How much do we learn from a result?

- For Bayes:
  - Narrower posterior distribution means more knowledge
  - Happens when likelihood distribution is narrow
- Generally: Narrower likelihood distributions contain more information
- For large statistics, likelihood becomes Gaussian, $\log L$ becomes parabola
- $\partial_\theta^2 \log L(\theta)$ relates to width
- For frequentist approach, interested in fluctuations when experiment is repeated
- $E[-\partial_\theta^2 \log L(\theta)]$ is thus a measure for the information content (evaluated at the true θ)



Likelihood for binomial distribution

The Fisher Information Matrix

$$\mathbb{I}(\theta) = E[-\partial_\theta^2 \log L(\theta)] = E[(\partial_\theta \log L(\theta))^2]$$

is called the *Fisher Information*

- It is larger for narrower likelihood distributions and fulfils

$$\mathbb{I}_{1+2} = \mathbb{I}_1 + \mathbb{I}_2,$$

which is what we want from an information measure.

- For more than one parameter $(\theta_1, \ldots, \theta_n)$ this generalises to

$$\mathbb{I}_{i,j} = E[-\partial_i \partial_j \log L(\vec{\theta})] = E[(\partial_i \log L(\vec{\theta}))\ (\partial_j \log L(\vec{\theta}))],$$

the *Fisher Information Matrix.*

Minimum Variance Bound

- Consider an unbiased estimator:

$$\langle \hat{\theta} \rangle = E[\hat{\theta}] = \int \hat{\theta}(x)\, p(x\,|\,\theta)\, \mathrm{d}x = \int \hat{\theta}(x)\, L(\theta\,|\,x)\, \mathrm{d}x = \theta$$

- What can we learn about the variance of the estimator?
- Use $\dfrac{\mathrm{d}\log L}{\mathrm{d}\theta} = \dfrac{\mathrm{d}L}{\mathrm{d}\theta}\dfrac{1}{L}$

1. $\displaystyle\int L\,\mathrm{d}x = 1 \implies \int \frac{\mathrm{d}L}{\mathrm{d}\theta}\,\mathrm{d}x = 0 \implies \int \theta \frac{\mathrm{d}\log L}{\mathrm{d}\theta} L\,\mathrm{d}x = 0$

2. $\displaystyle\int \hat{\theta}\, L\,\mathrm{d}x = \theta \implies \int \hat{\theta}\frac{\mathrm{d}L}{\mathrm{d}\theta}\,\mathrm{d}x = \int \hat{\theta}\frac{\mathrm{d}\log L}{\mathrm{d}\theta} L\,\mathrm{d}x = 1$

and thus

$$\int (\hat{\theta} - \theta)\frac{\mathrm{d}\log L}{\mathrm{d}\theta} L\,\mathrm{d}x = \left\langle (\hat{\theta} - \theta)\frac{\mathrm{d}\log L}{\mathrm{d}\theta} \right\rangle = 1$$

Minimum Variance Bound (II)

$$\int (\hat{\theta} - \theta) \frac{\mathrm{d} \log L}{\mathrm{d}\theta} L \, \mathrm{d}x = \left\langle (\hat{\theta} - \theta) \frac{\mathrm{d} \log L}{\mathrm{d}\theta} \right\rangle = 1$$

with $u = (\hat{\theta} - \theta)\sqrt{L}$ and $v = \frac{\mathrm{d} \log L}{\mathrm{d}\theta}\sqrt{L}$, we can now use the Cauchy-Schwarz inequality:

$$\left(\int (\hat{\theta} - \theta)^2 L \, \mathrm{d}x \right) \left(\int \left(\frac{\mathrm{d} \log L}{\mathrm{d}\theta} \right)^2 L \, \mathrm{d}x \right) \geq 1 \text{ , or}$$

$$\boxed{V[\hat{\theta}] \geq \frac{1}{\left\langle (\mathrm{d} \log L / \mathrm{d}\theta)^2 \right\rangle} = \frac{1}{\mathbb{I}[\theta]}}$$

- This is the Cramér-Rao minimum variance bound
- No unbiased estimator can be better than this

Cauchy-Schwarz inequality:

$$\int u^2 \mathrm{d}x \int v^2 \mathrm{d}x \geq \left(\int uv \, \mathrm{d}x \right)^2$$

Likelihood function and maximum likelihood

Principle of maximum likelihood:

- The best estimate of the parameters $\vec{\theta}$ is that value which maximizes the likelihood function
- This is called the *maximum likelihood estimator*

In the limit of infinite statistics, the maximum likelihood estimator:

- Is unbiased
- Achieves the Cramer-Rao bound (meaning it is maximally *efficient*)

- In most practical cases it is thus prudent to simply take the maximum likelihood estimator, rather than look for a MVUE

Properties of the ML estimator

- The ML Estimator is invariant under parameter transformation:

$$\psi = g(\theta) \quad \Rightarrow \quad \hat{\psi} = g(\hat{\theta})$$

- ML does not provide a goodness-of-fit measure
  - It tells you which parameter set fits best with the data
  - It does not tell you if it fits well in an absolute sense (no goodness-of-fit measure)
- Sometimes the maximum has to be found numerically

Maximum likelihood example 1: Exponential Decay

Consider exponential pdf: $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$

Independent measurements drawn from this distribution: $t_1, t_2, \ldots, t_n$

Likelihood function: $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

$L(\tau)$ is maximum when $\ln L(\tau)$ is maximum:

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

Find maximum:

$$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \quad \rightsquigarrow \quad \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \rightsquigarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

MVB example: Exponential decay

Minimum variance bound (MVB):

$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n} \left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3} \right) = \frac{n}{\tau^2} - \frac{2}{\tau^3} \sum_{i=1}^{n} t_i = \frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right)$$

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right) \right]} = \frac{1}{-\frac{n}{\tau^2} \left(1 - \frac{2E[\hat{\tau}]}{\tau} \right)} = \frac{\tau^2}{n}$$

Variance of ML estimator:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i \qquad V[\hat{\tau}] = V\left[\frac{1}{n} \sum_{i=1}^{n} t_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n}$$

So $\hat{\tau}$ reaches the MVB for any n.

Maximum likelihood example 2: Gaussian (I)

Consider x_1, x_2, …, x_n drawn from Gaussian(μ, σ²)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Log-likelihood function:

$$\ln L(\mu,\sigma^2) = \sum_{i=1}^{n} \ln f(x_i;\mu,\sigma^2) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln\sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

Derivatives w.r.t. μ and σ²:

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} \qquad \frac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

Maximum likelihood example 2: Gaussian (II)

Setting the derivatives w.r.t. μ and σ^2 to zero and solving the equations:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

We find that the ML estimator for σ^2 is biased!

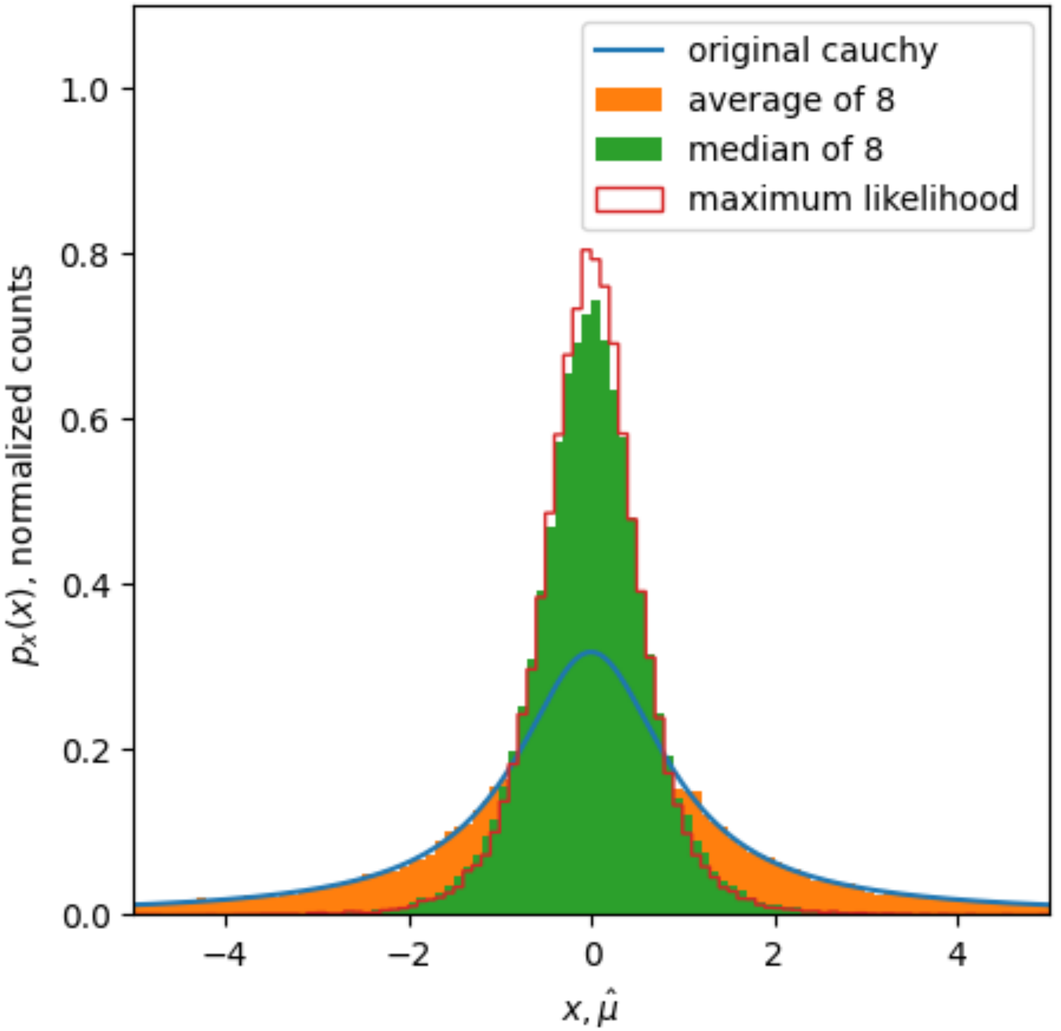
Maximum likelihood example 3: Cauchy distribution

Cauchy: $p(x) = \dfrac{1/(\pi\gamma)}{1 + (x-\mu)^2/\gamma^2}$

- For 8 measurements, maximise:

$$L = \prod_i \frac{1/(\pi\gamma)}{1 + (x_i-\mu)^2/\gamma^2}$$

- Solving derivative for $\hat{\mu}$ gives complicated polynomial → solve numerically instead
- A bit better than the median estimator



Uncertainty of the ML estimator: Approximating the minimum variance bound

In many cases it is impractical to calculate the MVB analytically. Instead, one uses the following approximation which is good for large n:

$$E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right] \approx -\left.\frac{\partial^2 \ln L}{\partial^2 \theta}\right|_{\theta=\hat{\theta}}$$

The variance of the ML estimator is given by:

$$V[\hat{\theta}] = -\frac{1}{\left.\frac{\partial^2 \ln L}{\partial^2 \theta}\right|_{\theta=\hat{\theta}}}$$

Example: Exponential decay

$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3}\right) = \frac{n}{\tau^2} - \frac{2}{\tau^3}\sum_{i=1}^{n} t_i = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

$$V[\hat{\tau}] = -\left(\frac{\partial^2 \ln L}{\partial^2 \theta}\right)^{-1}_{\tau=\hat{\tau}} = \frac{\hat{\tau}^2}{n}$$

Maximum likelihood uncertainty

Consider maximum likelihood estimate of a parameter θ. Methods to estimate Uncertainty of θ:

1. $\sigma_{\hat{\theta}}$ **from Monte Carlo**
   Generate pseudo-data by sampling the assumed distribution using the ML estimate $\hat{\theta}$ as parameter

2. **Use minimum variance bound**

$$\sigma_{\hat{\theta}} = \frac{1}{\sqrt{-\frac{\partial^2}{\partial^2 \theta} \ln L(\theta)}}$$

3. $\Delta \ln L = -1/2$ **method**:

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

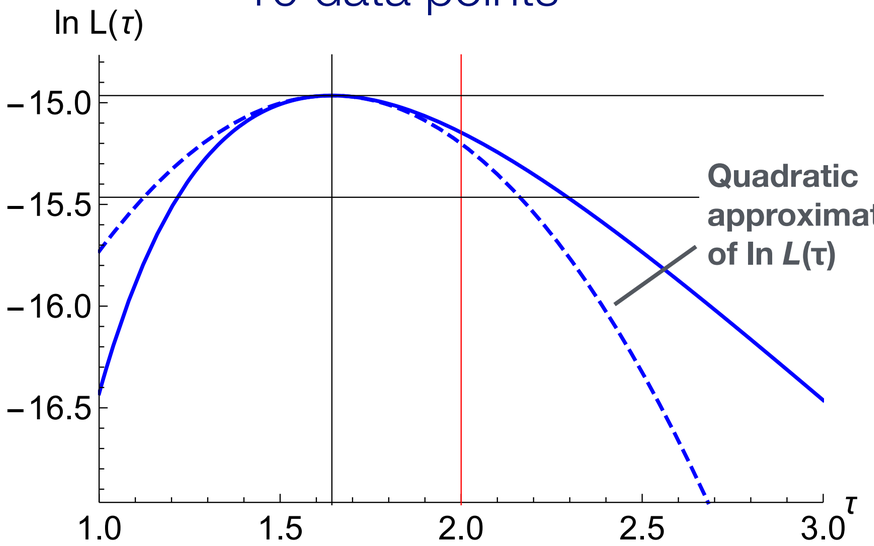
For a Gaussian likelihood function all methods agree.

Method 3 usually gives asymmetric uncertainties (which are messy).

Asymptotic normality of the likelihood function

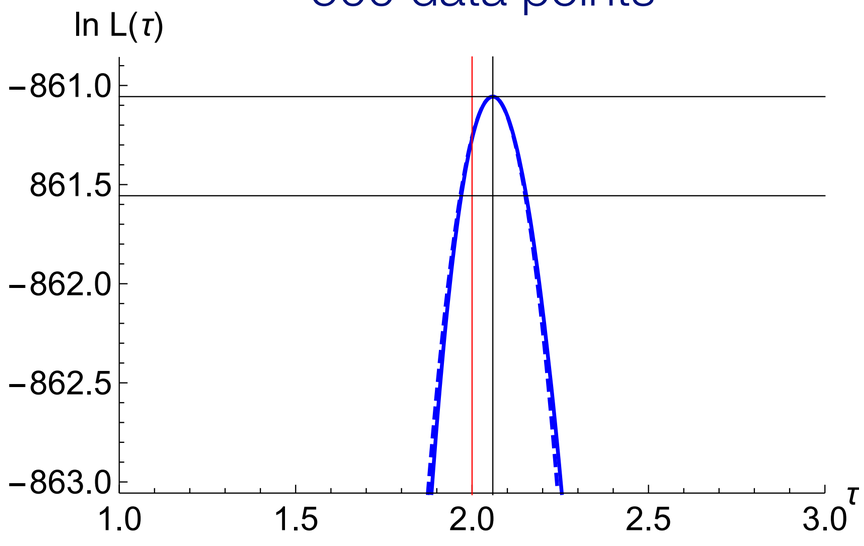
For any probability function $f(x;\theta)$ the likelihood function L approaches a Gaussian for large n, i.e., for a large number of events, and the variance of the ML estimator reaches the minimum variance bound.

10 data points



quadratic approximation of ln $L(\tau)$ is not very good

500 data points



quadratic approximation of ln $L(\tau)$ is excellent

Data points sampled from $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$ with $\tau = 2$

Uncertainty of the ML estimator: $\Delta \ln L = -1/2$ method

Taylor expansion of ln L around the maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \underbrace{\left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})}_{=0} + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial^2 \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

$\left[\frac{\partial^2 \ln L}{\partial^2 \theta}\right]_{\theta=\hat{\theta}} \approx -\frac{1}{\sigma^2}$ [from MVB, or from assuming Gaussian shape]

If $L(\theta)$ is approximately Gaussian (ln $L(\theta)$ then is a approximately a parabola):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2_{\hat{\theta}}}}$$

good approximation in the large sample limit

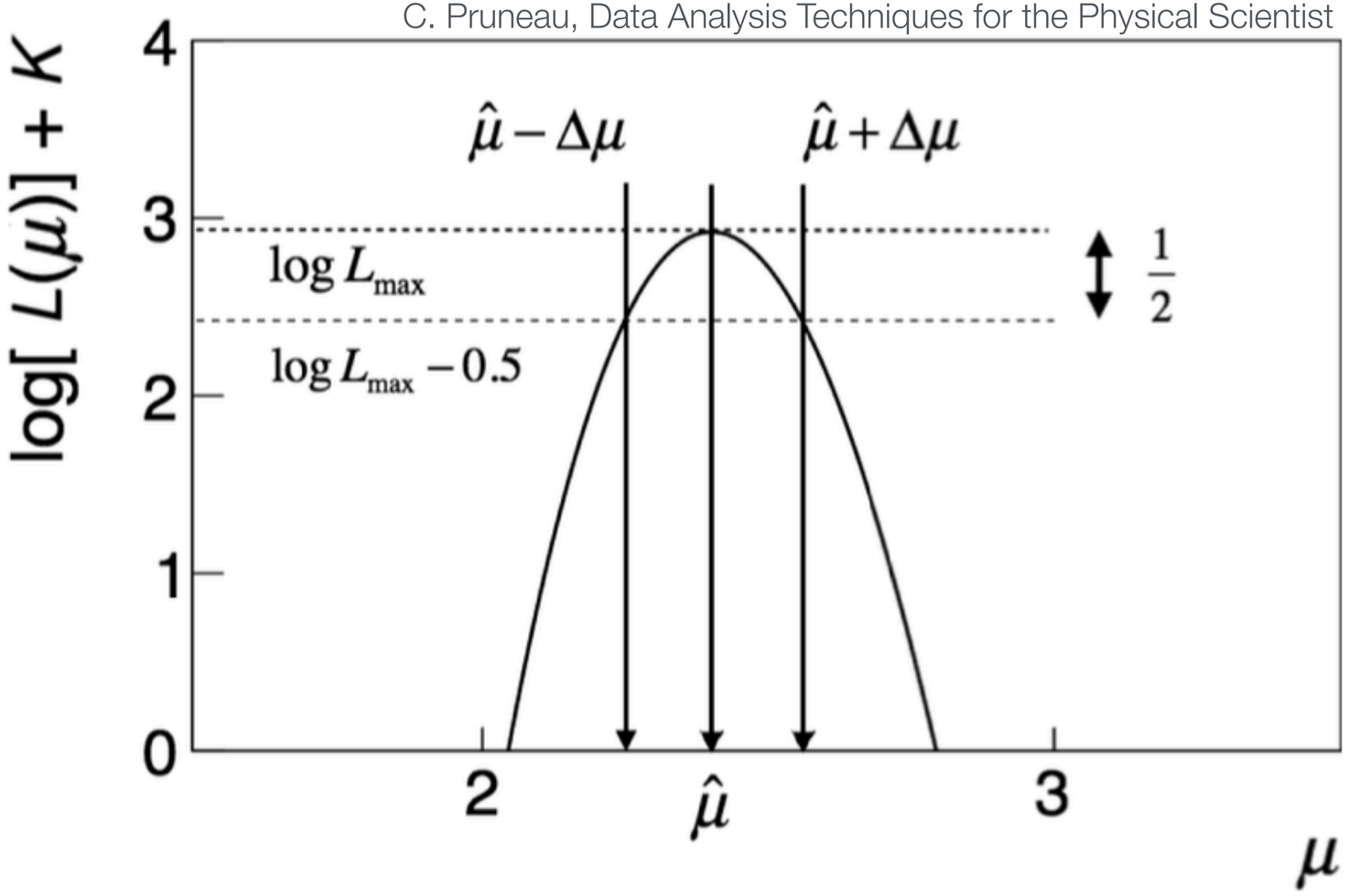
One can then estimate the uncertainties from the points where ln L has dropped by 1/2 from its maximum:

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

Illustration of the $\Delta \ln L = -1/2$ method

L is Gaussian $\leftrightarrow$ $\ln L$ is a parabola

C. Pruneau, Data Analysis Techniques for the Physical Scientist



Averaging measurements with Gaussian uncertainties

pdf for measurement (same mean, different σ):

$$f(x;\mu,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}} \qquad \ln L(\mu) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_i - \frac{(x_i-\mu)^2}{2\sigma_i^2} \right)$$

Weighted average = ML estimate

$$\left.\frac{\partial \ln L(\mu)}{\partial \mu}\right|_{\mu=\hat{\mu}} = \sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\sigma_i^2} \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

Uncertainty? In this case L is Gaussian and we can write it as

$$L(\mu) \propto e^{-\frac{(\mu-\hat{\mu})^2}{2\sigma_{\hat{\mu}}^2}} \quad \text{with} \quad \sigma_{\hat{\mu}}^2 = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

We obtain the formula for the weighted average:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \quad \pm \quad \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}$$

Minimum variance bound for m parameters

$$f(x;\vec{\theta}), \quad \vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$$

Fisher information matrix $I(\vec{\theta})$ ($m \times m$ matrix):

$$I_{jk}[\vec{\theta}] = -E\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \ln L(x, \vec{\theta})\right]$$

Covariance matrix of the parameters: $V_{ij} := \mathrm{cov}[\theta_i, \theta_j]$

Cramér-Rao-Frechet bound for an unbiased estimator then states that $V - I^{-1}$ is a positive-semidefinite matrix.

In particular one obtained for the variance:

$$V[\hat{\theta}_j] \geq (I(\vec{\theta})^{-1})_{jj}$$

Variance of the ML estimator for m parameters

For any probability function $f(x;\vec{\theta})$ the likelihood function L approaches a multi-variate Gaussian for large n

$$L(\vec{\theta}) \propto e^{-\frac{1}{2}(\vec{\theta}-\widehat{\vec{\theta}})^{\mathsf{T}}\, V^{-1}[\widehat{\vec{\theta}}]\,(\vec{\theta}-\widehat{\vec{\theta}})}$$

The variance of the ML estimator then reaches the MVB:

$$V[\widehat{\vec{\theta}}] \to I(\vec{\theta})^{-1}$$

Covariance matrix of the estimated parameters:

$$V[\widehat{\vec{\theta}}] \approx \left[-\left. \frac{\partial^2 \ln L(\vec{x};\vec{\theta})}{\partial^2 \vec{\theta}} \right|_{\vec{\theta}=\widehat{\vec{\theta}}} \right]^{-1}$$

or equivalently:

$$(V^{-1}[\widehat{\vec{\theta}}])_{ij} = -\left. \frac{\partial^2 \ln L(\vec{x};\vec{\theta})}{\partial\theta_i \partial\theta_j} \right|_{\vec{\theta}=\widehat{\vec{\theta}}}$$

Standard deviation of a single parameters:

$$\hat{\sigma}_{\hat{\theta}_j} = \sqrt{(V[\widehat{\vec{\theta}}])_{jj}}$$

Example: Two-parameter ML fit (from Cowan's book)

Scattering angle distribution, $x = \cos\theta$:

$$f(x; a, b) = \frac{1 + ax + bx^2}{2 + 2b/3}$$

Normalization:

$$\int_{x_{\text{min}}}^{x_{\text{max}}} f(x; a, b)\, \mathrm{d}x = 1$$

Example: $a = 0.5,\ b = 0.5;\ x_{\text{min}} = -0.95,\ x_{\text{max}} = 0.95,\ 1000\,\text{MC events}$

Numerical minimization with MINUIT:

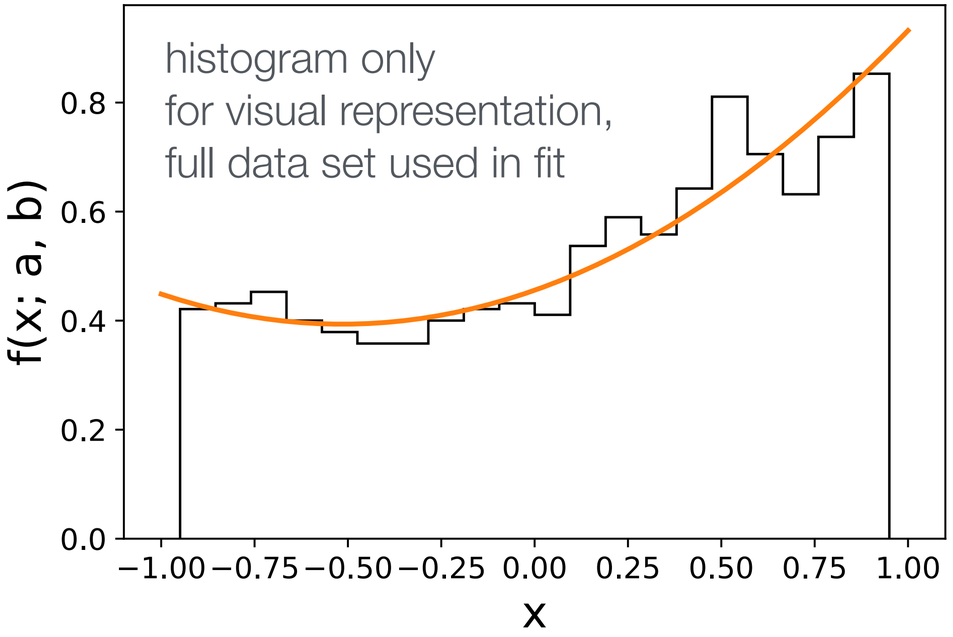
[link to jupyter notebook]

$$\hat{a} = 0.53 \pm 0.08$$

$$\hat{b} = 0.51 \pm 0.16$$

$$\text{cov}[\hat{a}, \hat{b}] = 0.006$$

$$\rho = 0.48$$



Uncertainties and covariance from inverse of Hessian matrix H:

$$\widehat{V} = -H^{-1},\ (H)_{ij} = \left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$$

Example: Two-parameter ML fit (iminuit)

```python
import numpy as np
import matplotlib.pyplot as plt
from iminuit import Minuit
```

```python
x = np.loadtxt("data.txt")
```

```python
def f(x, a, b):
    """normalized fit function"""
    xmin = -0.95
    xmax = 0.95
    return (6 * (1 + a * x + b * x * x)) /
           ((xmax - xmin) * (3 * a * (xmax + xmin) +
           2 * (3 + b * (xmax * xmax + xmax * xmin + xmin * xmin))))
```

```python
def negative_log_likelihood(a, b):
    p = np.log(f(x, a, b))
    return -np.sum(p)
```

iminuit uses introspection to detect the parameter names of your function

```python
m = Minuit(negative_log_likelihood,
           a=1, b=1, error_a=0.01, error_b=0.01, errordef=Minuit.LIKELIHOOD)
```

```python
m.migrad()
```

Example: Two-Parameter ML Fit (iminuit)

```
m.migrad()
```

FCN = 606.5	Ncalls = 10 (146 total)
EDM = 1.33e-10 (Goal: 0.0001)	up = 0.5

Valid Min.	Valid Param.	Above EDM	Reached call limit
True	True	False	False

Hesse failed	Has cov.	Accurate	Pos. def.	Forced
False	True	True	True	False

	Name	Value	Hesse Error	Minos Error-	Minos Error+	Limit-	Limit+	Fixed
0	a	0.53	0.08					
1	b	0.51	0.16					

https://iminuit.readthedocs.io/en/stable/

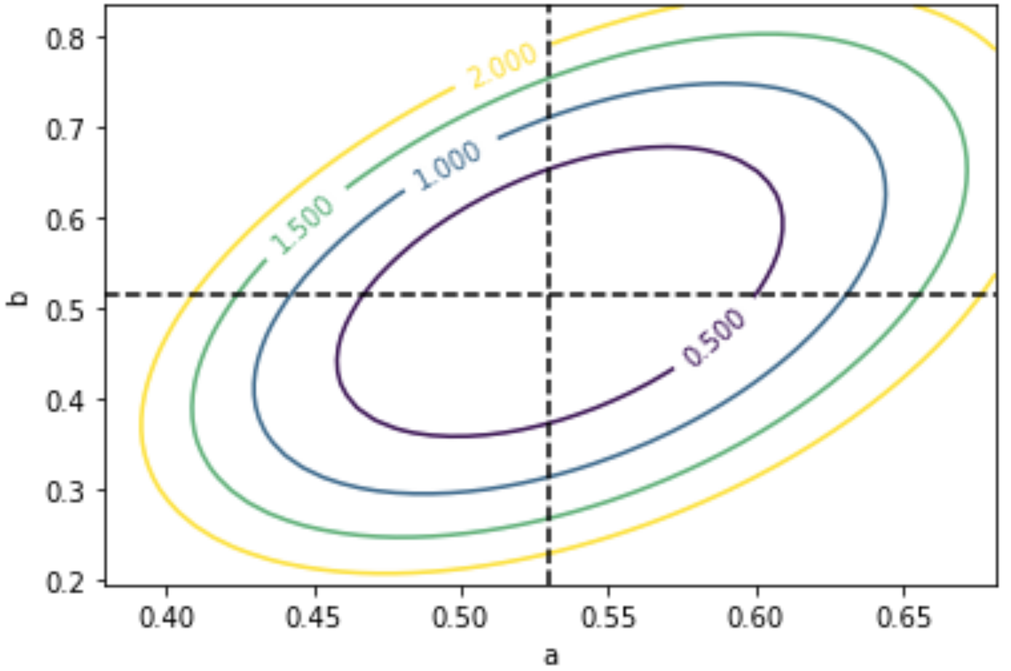
https://nbviewer.jupyter.org/github/scikit-hep/iminuit/blob/master/tutorial/basic_tutorial.ipynb

Example: Two-Parameter ML Fit (iminuit)

```
# covariance matrix
m.matrix()
```

	a	b
a	0.006	0.006
b	0.006	0.026

```
m.draw_contour('a','b');
```



Extended maximum likelihood method (1)

Standard ML fit: information is in the shape of the distribution of the data x_i.

Extended ML fit: normalization becomes a fit parameter

> Sometimes the number of observed events contains information about the parameters of interest, e.g., when we measure a rate.

Normal ML method:

$$\int f(x, \vec{\theta})\, \mathrm{d}x = 1$$

Extended ML method:

$$\int q(x, \vec{\theta})\, \mathrm{d}x = \nu(\vec{\theta}) = \text{ predicted number of events}$$

Extended maximum likelihood method (2)

Normalized pdf:

$$\int f(x, \vec{\theta})\, \mathrm{d}x = 1$$

Likelihood function:

$$L(\vec{\theta}) = \frac{\nu^n e^{-\nu}}{n!} \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{where} \;\; \nu \equiv \nu(\vec{\theta})$$

Log-Likelihood function:

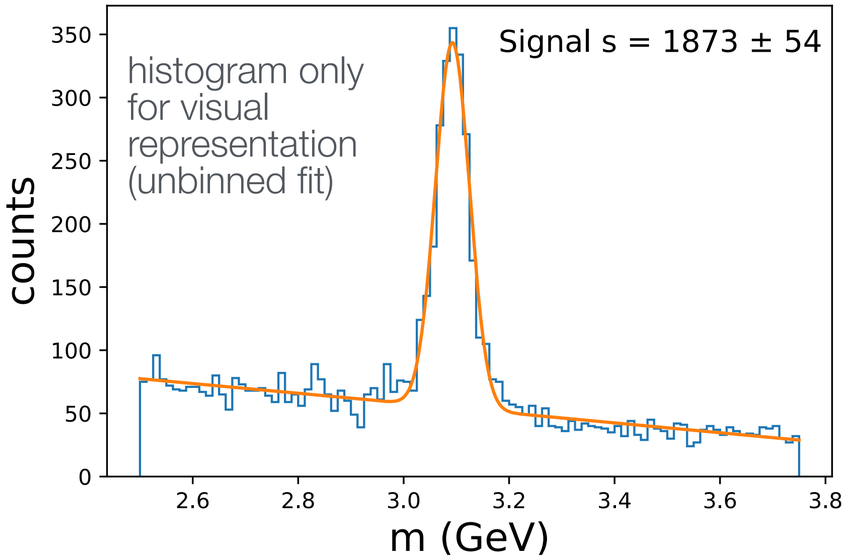
$$\ln L(\vec{\theta}) = -\ln(n!) - \nu(\vec{\theta}) + \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

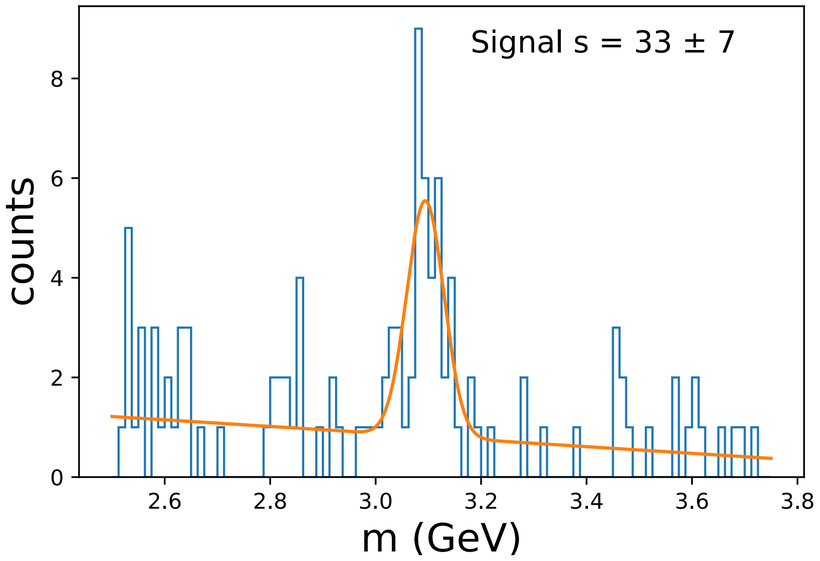
ln(n!) does not depend on the parameters. So we need to minimize:

$$-\ln \tilde{L}(\vec{\theta}) = \nu(\vec{\theta}) - \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

prediction for total number of events

Application of the extended ML method: Linear combination of signal and background PDF (1)





Two-component fit (signal + linear background)

Parameters:
- signal counts *s*
- background counts *b*
- linear background (slope, intercept)
- Gaussian peak: μ, σ

Normalized pdf:

$$f(x; r, \vec{\theta}) = r\, f_s(x, \vec{\theta}) + (1 - r)\, f_b(x, \vec{\theta})$$

negative log-likelihood:

$$-\ln \tilde{L}(\vec{\theta}) = s + b - n \ln(s + b) - \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})]$$

$$\nu(s, b) = s + b, \quad r = \frac{s}{s + b}$$

Unbinned ML fit works fine also in case of low statistics

Application of the extended ML method: Linear combination of signal and background PDF (2)

Discussion:

We could have just fitted the normalized pdf:

$$f(x; r_s, \vec{\theta}) = r\, f_s(x, \vec{\theta}) + (1 - r)\, f_b(x, \vec{\theta})$$

Good estimate of the number of signal events: $n_{\text{signal}} = r\, n$

However, $\sigma_r\, n$ is not a good estimate of the variation of the number of signal events (ignores fluctuations of n)

[C. Blocker, Maximum Likelihood Primer]

(Trivial) example (L. Lyons):
96 protons and 4 heavy nuclei have been measured in a cosmic ray experiment

	protons	heavy nuclei
ML estimate	96 ± 2	4 ± 2
Extended ML estimate	96 ± 10	4 ± 2

Maximum likelihood fits with binned data (1)

Common practice: data put into a histogram: $\vec{n} = (n_1, \ldots, n_k), \; n_{\text{tot}} = \sum_{i=1}^{k} n_i$

Model prediction for the expected counts in bin i for fixed n_{tot}:

$$\nu_i(\vec{\theta}) = n_{\text{tot}} \int_{\text{bin } i} f(x; \vec{\theta}) \, \mathrm{d}x \qquad \nu_i(\vec{\theta}) = (\nu_1, \ldots, \nu_k)$$

If n_{tot} is fixed the probability to get a certain $\vec{n}$ is given by the multinomial distribution.

Multinomial distribution (generalization of binomial distribution):

→ k different possible outcomes, probability for outcome i is p_i, $\sum_{i=1}^{k} p_i = 1$

$$f(\vec{n}; n_{\text{tot}}, \vec{p}) = \frac{n_{\text{tot}}!}{n_1! \cdot \ldots \cdot n_k!} \, p_1^{n_1} \cdot \ldots \cdot p_k^{n_k} \qquad \vec{p} = (p_1, \ldots, p_k)$$

Maximum likelihood fits with binned Data (2)

With $p_i = \nu_i / n_{\text{tot}}$ we write the likelihood of a certain $n_1, \ldots, n_k$ outcome as:

$$L(\vec{\theta}) = \frac{n_{\text{tot}}!}{n_1! \cdot \ldots \cdot n_k!} \left(\frac{\nu_1}{n_{\text{tot}}}\right)^{n_1} \cdot \ldots \cdot \left(\frac{\nu_k}{n_{\text{tot}}}\right)^{n_k} \qquad \nu_i(\vec{\theta}) = (\nu_1, \ldots, \nu_k)$$

Log-likelihood function:

$$\ln L(\vec{\theta}) = \sum_{i=1}^{k} n_i \ln \nu_i(\vec{\theta}) + C$$

Limit of zero bin width → usual unbinned maximum likelihood method

Maximum likelihood fits with binned Data (3)

Extended log-likelihood fit for binned data:

n_{tot} fluctuates, predicted average: ν_{tot}

$$\nu_{\text{tot}} = \sum_{i=1}^{k} \nu_i, \qquad n_{\text{tot}} = \sum_{i=1}^{k} n_i$$

Likelihood function:

$$L(\vec{\theta}) = \frac{\nu_{\text{tot}}^{n_{\text{tot}}}}{n_{\text{tot}}!} e^{-\nu_{\text{tot}}} \frac{n_{\text{tot}}!}{n_1! \cdot \ldots \cdot n_k!} \left(\frac{\nu_1}{\nu_{\text{tot}}}\right)^{n_1} \cdot \ldots \cdot \left(\frac{\nu_k}{\nu_{\text{tot}}}\right)^{n_k}$$

$$= \prod_{i=1}^{k} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Function that needs to be maximized (dropping terms that do not depend on the parameters):

$$\ln L(\vec{\theta}) = \sum_{i=1}^{k} n_i \ln \nu_i - \nu_i = -\nu_{\text{tot}} + \sum_{i=1}^{k} n_i \ln \nu_i, \qquad \nu_i(\vec{\theta}) = (\nu_1, \ldots, \nu_k)$$

Maximum likelihood fits with binned Data (4)

- Can also be understood with Poisson distribution: There is an expectation value of the number of particles in each bin ν_i
- When repeating the experiment, the counts of particles in each bin fluctuate as described by Poisson distribution
- Poisson: $p(n_i \mid \nu_i) = \frac{\nu_i^{n_i} \exp(-\nu_i)}{n_i!}$ for each bin
- So the likelihood is $L = \prod \frac{\nu_i^{n_i} \exp(-\nu_i)}{n_i!}$
- And the log-likelihood is $\log L = \sum n_i \log \nu_i - \nu_i$

The likelihood principle

The likelihood $L(\vec{\theta}\,|\,\vec{d})$ contains all information in the data $\vec{d}$ that is relevant for the parameters $\vec{\theta}$ within the context of this model.

Optional stopping

- Option 1: Perform an experiment N times and count number of successes
- Option 2: Perform the experiment until k successes are reached

- Probability distribution is different
- But likelihoods are the same (apart from constant prefactors)
- Do we draw the same conclusions?

- If we think the experimenter used one rule, but later find they used another, does this change our conclusions for the same data?

Binomial: $p_b(k \,|\, N, \phi) = \binom{N}{k} \phi^k (1-\phi)^{N-k}$

Negative Binomial:

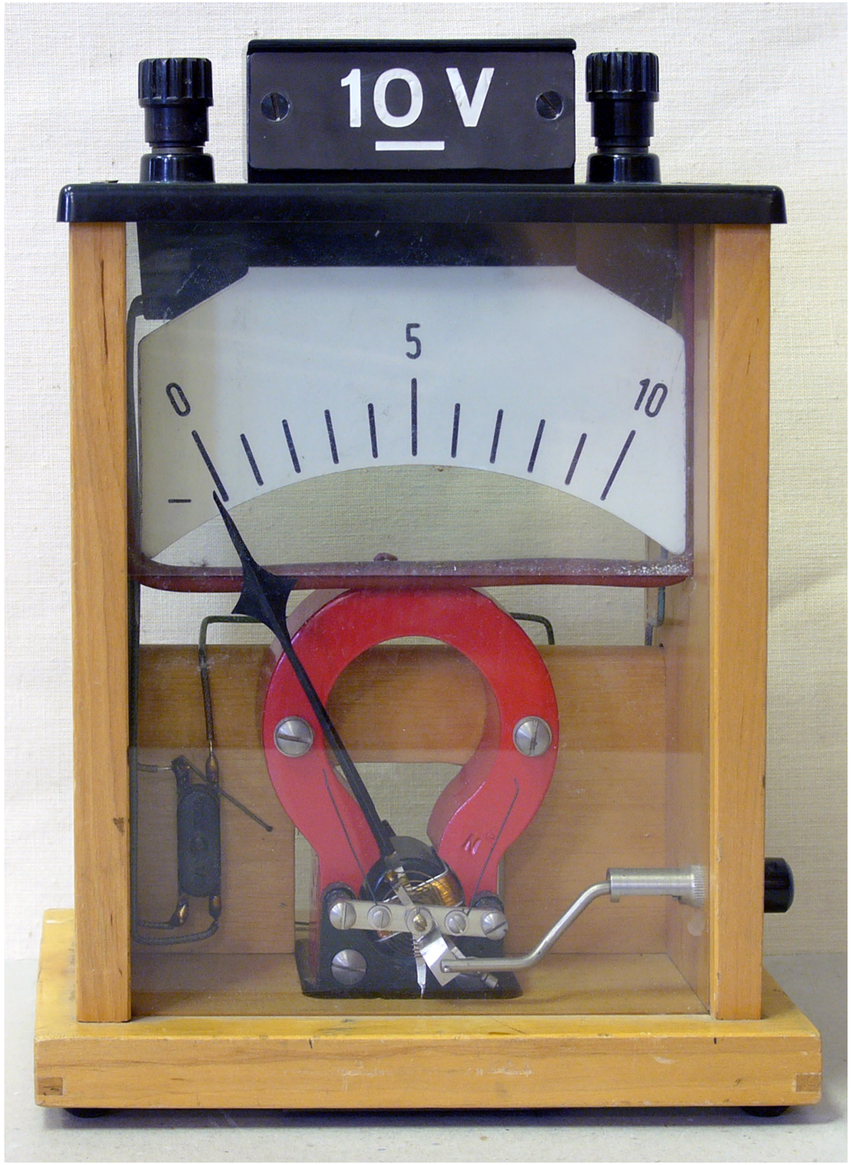
$$p_b(m \,|\, k, \phi) = \binom{m+k-1}{m} \phi^k (1-\phi)^m$$

$$(m = N - k)$$

Example

- Measurements were made by a student on a voltmeter going to 10V
- All measurements below
- Average gives unbiased estimator of some quantity
- But since there is a maximum, suddenly the measurement is biased
- Should this affect the conclusions even if the value was never reached?

- What if there was a second voltmeter that could be used in such cases?
- What if afterwards you find out that the second voltmeter was broken?



Birnbaums argument

Consider two possible experiments. We now flip a coin to decide which one we perform. The result of our analysis can only depend on the experiment we actually did, not the potential other one.