Karlsruher Institut für Technologie

Modern Methods of Statistical Data Analysis

From parameter estimation to deep learning — A guided tour of probability

Lecture 4

Introduction to Parameter Estimation & the Method of Maximum Likelihood

P.-D. Dr. Roger Wolf

roger.wolf@kit.edu

Dr. Pablo Goldenzweig

pablo.goldenzweig@kit.edu

Dr. Jan Kieseler

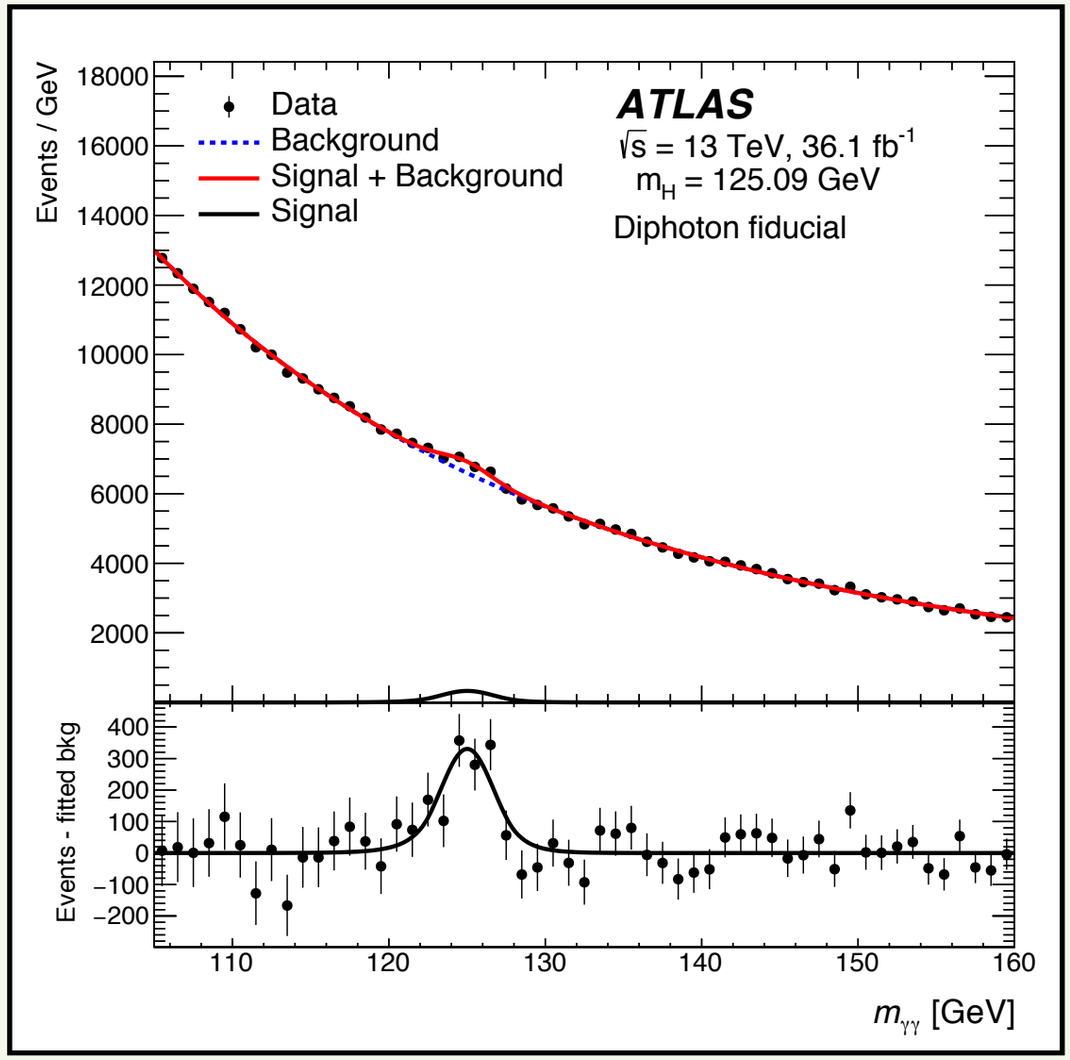
jan.kieseler@kit.edu

Dr. Slavomira Stefkova

slavomira.stefkova@kit.edu

Today

- What goes into such fits?
  - **First** part of lecture: general concepts of parameter estimation
  - **Second** part: the method of maximal likelihood



i.i.d. random variables

- Consider n variables $X_1, X_2, \ldots, X_n$
  - $X_1, X_2, \ldots, X_n$ are **independent and identically distributed (i.i.d.)** if
    - $X_1, X_2, \ldots, X_n$ are **independent**, and
    - all have the **same PMF** *(if discrete)* or **PDF** *(if continuous)*
      - $E[X_i] = \mu$ for $i = 1, \ldots, n$
      - $\text{Var}[X_i] = \sigma^2$ for $i = 1, \ldots, n$

Quick check: Are $X_1, X_2, \ldots, X_n$ i.i.d. with the following distributions?

1. $X_i \sim \text{Exp}(\tau)$, X_i independent ✓
2. $X_i \sim \text{Exp}(\tau_i)$, X_i independent ✗ (unless τ_i equal)
3. $X_i \sim \text{Exp}(\tau)$, $X_1 = X_2 = \cdots = X_n$ ✗ dependent! ($x_1 = x_2 = \ldots = x_n$)
4. $X_i \sim \text{Bin}(n_i, p)$, X_i independent ✗ (unless n_i equal)

Working with the CLT

Let $X_1, X_2, \ldots, X_n$ be i.i.d., where $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$.

As $n \to \infty$:

Sum of i.i.d. RVs

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Average of i.i.d. RVs (sample mean)

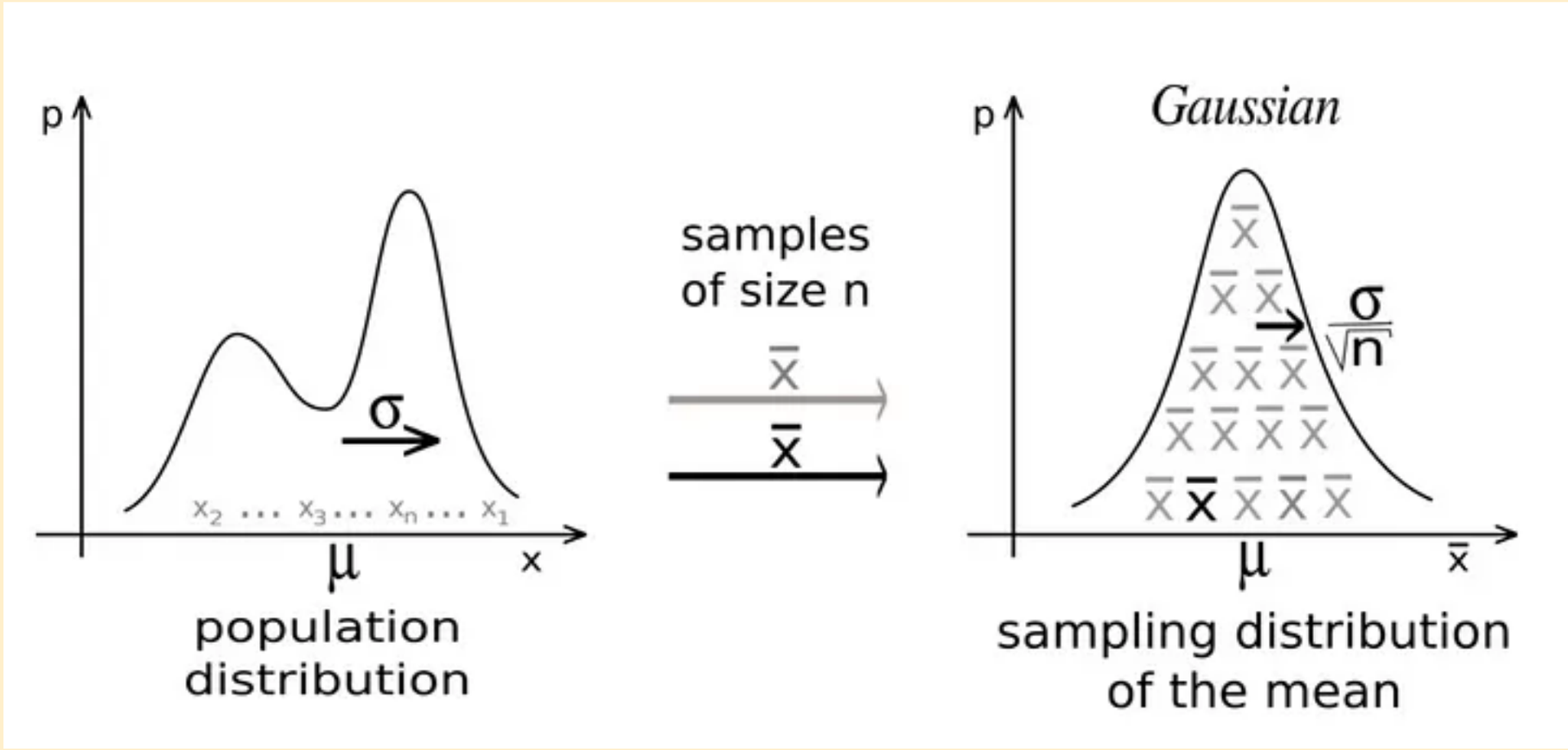
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Interpret: As we increase n (the size of our sample):

- The variance of our sample mean σ^2/n decreases
- The probability that our sample mean $\bar{X}$ is close to the true mean μ increases

CLT: *Key take home message*

No matter what the distribution of the population is, the distribution of mean samples from the population will always be Normally distributed



i.e., No matter what the distribution of the sample is, if you sample batches of data from that distribution and take the mean of each batch, *the mean values from those batches will be Normally distributed*

The CLT and real life

Review

- **Central Limit Theorem:** ✓
  - Sample mean $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
  - If we know μ and σ^2, we can compute probabilities on the sample mean $\bar{X}$ of a given sample size n

- **In real life:**
  - Yes, the CLT still holds…
  - But we often don't know μ or σ^2 of our original distribution
  - However, we can collect data (a sample of size n)
  - **Question: How can we estimate the values μ or σ^2 from our sample?**
    - *Answer: Parameter estimation!*

Parameter estimation: General concepts (i)

Take RV X described by PDF $f(x)$ $\quad$ S = the set of all possible values of X

n independent observations of X
= sample of size n

S = set of all possible values for the n-D vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$

Can be understood as a single random experiment with n measurements

The joint PDF of such a sample

$$f_{\text{sample}}(x_1, \ldots, x_n) = f(x_1) f(x_2) \ldots f(x_n)$$

Since it is assumed that the observations are all independent and that each X_i are described by the same PDF $f(x)$

What if we don't know $f(x)$?

Central problem of statistics: use x to learn properties of $f(x)$

Parameter estimation: General concepts (ii)

The RVs we have been discussing are parametric models:

$$\textit{Distribution} = \textit{model} + \textit{parameter } \theta$$

For each of the distributions below, what are the parameters θ ?

1. Ber(p) $\quad \theta = p$
2. Poi(λ) $\quad \theta = \lambda$
3. Uni(α, β) $\quad \boldsymbol{\theta} = (\alpha, \beta)$
4. $\mathcal{N}(\mu, \sigma^2)$ $\quad \boldsymbol{\theta} = (\mu, \sigma^2)$
5. $Y = mX + b$ $\quad \boldsymbol{\theta} = (m, b)$

In the real world, we don't know the *true parameters,* but we do get to observe **data**

In general: consider a random variable X distributed according to a PDF $f(x; \theta)$

- The *functional form* of $f(x; \theta)$ is ***known***
- The value of at least one parameter θ (or parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$) is ***not known***

Goal: construct a function of observed x to estimate θ

Key definitions: Statistic & Estimator

def statistic ≡ A function of the observed measurements which contains no unknown parameters.

def estimator ≡ A statistic used to estimate some property of a PDF (e.g., μ, σ, or other parameter).

(Cowan notation)

Use observations to construct functions that estimate properties of PDFs

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

'hatted' *sample* parameter =
Parameter we determine using observations

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

If limit converges, estimator is consistent

***Population* parameter = true parameter in PDF**
(May forever remain unknown)

Parameter fitting

Estimating $\hat{\theta}$ from given data x: **parameter fitting**

Since $\hat{\theta}$ is a function of RVs $\Rightarrow$ **it is *itself* a RV**

i.e. if the entire experiment $x = (x_1, \ldots x_n)$ is repeated, $\hat{\theta}$ would change and be described according to PDF $\boldsymbol{g(\hat{\theta}; \theta)}$

Depends on true value

Sampling distribution
(The PDF of a statistic)

Expectation value of Estimator $\hat{\theta}$

General: $a(x)$ is a function of RVs distributed according to $f(x)$

$$E[a(x)] = \int_{-\infty}^{\infty} a g(a) da \overset{!}{=} \int_{-\infty}^{\infty} a(x) f(x) dx$$ Eqn. 1.44, Cowan

Since

$$g(a)da = \int_{dS} f(x)dx \xrightarrow{\text{multiply by } a} ag(a)da = a\int_{dS} f(x)dx \xrightarrow{\text{Integrate over } dS} \int_{-\infty}^{\infty} ag(a)da = \int_{-\infty}^{\infty} af(x)dx$$

Thus (replacing a with $\hat{\theta}$)

$$E[\hat{\theta}(\boldsymbol{x})] = \int \hat{\theta} g(\hat{\theta};\theta) d\hat{\theta} = \int \cdots \int \hat{\theta}(\boldsymbol{x}) f(x_1;\theta) \cdots f(x_n;\theta) dx_1 \cdots dx_n$$

Interpret: This is the expected mean of $\hat{\theta}$ from an infinite # of similar experiments, each with a sample of size n

Bias & mean squared error (MSE)

Measures of the quality of an estimator

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

↑ Bias of an estimator

— does *NOT* depend on the measured values of the sample

— does depend on the sample size (n), functional form of $\hat{\theta}$, and properties of $f(x)$

If b independent of n, and $b = 0$ $\Rightarrow \hat{\theta}$ unbiased

If b dependent of n, and $\lim_{n\to\infty} b = 0$ $\Rightarrow \hat{\theta}$ asymptotically unbiased

Note: $\hat{\theta}$ can be biased, even if it's consistent

$$\mathrm{MSE} = E\left[\left(\hat{\theta} - \theta\right)^2\right] = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] + \left(E[\hat{\theta} - \theta]\right)^2$$

$$= \quad V[\hat{\theta}] \quad + \quad b^2$$

i.e., sum of variance and bias^2

Interpret: sum of squares of statistical and systematic uncertainties

Estimator for mean

Suppose: sample of size n of RV X with values $x_1, x_2, \ldots, x_n$

Assume: X is distributed according to some PDF $f(x)$

↑ **unknown**
(even as a parameterization)

Need: a function of the x_i to be an estimator for the expectation value (population mean) of x, μ.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

↑ **sample mean**
(not population mean)

Important property given by the weak law of large #'s:
*If the variance of x exists, then $\bar{x}$ is a **consistent** estimator for the population mean μ*

i.e., for $n \to \infty$, $\bar{x}$ converges to μ

Expectation value of estimator $\bar{x}$

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

↑ $= n\mu$

Estimators for variance

If the mean μ is unknown, the statistic s^2 is an unbiased estimator of the variance σ^2

population mean — sample variance — population variance

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

sample mean

Included so that $E[s^2] = \sigma^2$

i.e., so that the sample variance is an unbiased estimator for the population variance

If the mean μ is known, the statistic S^2 is an unbiased estimator of the variance σ^2

big S

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Proof of unbiased estimators for variance

1. Show that the sample variance s^2 and the statistic S^2 are unbiased estimators of the population variance σ^2 with

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{and} \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \,. \tag{1}$$

with $\bar{x}$ denoting the sample mean and μ being the population mean. I.e. if one does not know the population mean and needs to estimate it from the sample via $\bar{x}$, one has to use s^2 to obtain an unbiased estimator for the population variance.

Proofs: (we omitted all indices in the sums for brevity)

a) $E[x_i^2] - 2\mu E[x_i] + \mu^2 = E[x_i^2] - \mu^2 = \sigma^2 \rightarrow E[x_i^2] = \sigma^2 + \mu^2$ □

b) $V[\bar{x}] = E[\left(\frac{1}{n}\sum_i x_i\right)\left(\frac{1}{n}\sum_j x_j\right)] - \mu^2 = \frac{1}{n^2}\sum_{ij} E[x_i x_j] - \mu^2 = \frac{1}{n^2}\left((n^2-n)\,\mu^2 + n\,(\mu^2+\sigma^2)\right) - \mu^2 = \frac{1}{n^2}\left((n^2 - n^2 + n - n)\,\mu^2 + n\sigma^2\right) \rightarrow V[\bar{x}] = \frac{\sigma^2}{n}$ □

c) $E[\sum_i (x_i - \bar{x})^2] = E[\sum_i x_i^2 - 2\bar{x}\sum_i x_i + n\bar{x}] = E[\sum_i x_i^2] - nE[\bar{x}^2]$.

With the results from a) and b) we now get $nE[\bar{x}^2] = n\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2 + n\mu^2$ and $E[\sum_i x_i^2] = \sum_i E[x_i^2] = n\,(\sigma^2 + \mu^2)$.

→ $E[\sum_i (x_i - \bar{x})^2] = n\,(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2 = (n-1)\,\sigma^2$. Thus

$$E[s^2] = E[\frac{1}{n-1}\sum_i (x_i - \bar{x})^2] = \sigma^2 \quad \square$$

d) Using all results we find $E[\sum (x_i - \mu)^2] = \sum E[x_i^2] - n\mu^2 = n\,(\sigma^2 + \mu^2) - n\mu^2 = n\sigma^2$. Thus

$$E[S^2] = E[\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2] = \sigma^2 \quad \square$$

Now you've proved to yourself when to use n and $n-1$:

n if population mean is known

$n-1$ if population mean is unknown and has to be estimated

Review proofs at home: ILIAS: /Reading material / L04 / UnbiasedEstimators

Estimator for covariance

- Similarly one can show that this expression is an unbiased estimator for the covariance of two random variables x and y:

$$\widehat{V}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{n}{n-1}\left(\overline{xy} - \overline{x}\,\overline{y}\right)$$

- This can be normalized by the square-root of the estimators of the sample variance s_x and s_y

$$r = \frac{\widehat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{j=1}^{n}(x_j - \overline{x})^2 \cdot \sum_{k=1}^{n}(y_k - \overline{y})^2\right)^{1/2}}$$

$$= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}}.$$

http://www.pp.rhul.ac.uk/~cowan/sda/

Variance of estimators

Now were talking about the variance of the estimators we just defined

- Variance of sample mean $\bar{x}$

This was step b) in the proof on slide 15

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] - \mu^2$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} E[x_i x_j] - \mu^2$$

$$= \frac{1}{n^2}\left[(n^2-n)\mu^2 + n(\mu^2+\sigma^2)\right] - \mu^2 = \frac{\sigma^2}{n},$$

This expresses the well-known result that the standard deviation of the mean of n measurements of x is equal to the standard deviation of $f(x)$ itself divided by $\sqrt{n}$.

- where σ^2 is the variance of $f(x)$ and we have used $E[x_i x_j] = \mu^2$ for $i \neq j$ and $E[x_i^2] = \mu^2 + \sigma^2$

- In a similar way one can show that the variance of s^2 is

$$V[s^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right),$$

Central moments

$$\int_{-\infty}^{\infty} (x-\mu)^n f(x)\,dx = \mu_n,$$

http://www.pp.rhul.ac.uk/~cowan/sda/

Expectation value and variance of correlation

- The expectation value and variance of the correlation coefficient r depend on higher moments of the joint PDF $f(x, y)$

$$E[r] = \rho - \frac{\rho(1-\rho^2)}{2n} + O(n^{-2})$$

$$V[r] = \frac{1}{n}(1-\rho^2)^2 + O(n^{-2}).$$

http://www.pp.rhul.ac.uk/~cowan/sda/

Answer Time: Quiz 3

Monte Carlo method recap

How would you write an algorithm to integrate a multi-dimensional function (of dimensionality N) using the 'acceptance-rejection' method? Address in particular: what ingredients you need and sketch out the algorithm explicitly.

1) Generate N uniform random numbers in intervals of $[x^i_{\min}, x^i_{\max}]$

- Denote these as $u = (u_1, u_2, \ldots, u_N)$

2) Evaluate the function we wish to integrate: $f(u_1, u_2, \ldots, u_N)$

3) Generate another random number ν between 0 and $f_{\max}$

- Reject the event if $\nu > f(u_1, u_2, \ldots, u_N)$; otherwise accept

4) Repeat 1- 3

The integral is then given by total number of accepted events divided by the total number of tried events

Take 5

Likelihood (function)

PDF → Likelihood (Interpretation)

Suppose a measurement of the mass m of an elementary particle yields the value m_0, and it is known that the measuring apparatus yields values normally distributed about the unknown true mass m_t, with a known rms deviation σ_m

The probability density for obtaining the value m given the true mass m_t

$$P(m \mid m_t) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{\frac{-i(m-m_t)^2}{2\sigma_m^2}} = N(m_t, \sigma_m)$$

t=true

By writing $\mathscr{L}$ instead of P we draw attention to the fact that we are considering its behavior *for different values of* m_t *given the particular measured datum* $m = m_0$

$$\mathscr{L}(m_0 \mid m_t) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{\frac{-i(m_0-m_t)^2}{2\sigma_m^2}}$$

t=theory

Interpret:
$\mathscr{L}$ is the probability, under the assumption of the theory (m_t), to observe the data which were actually observed (m_0)

ILIAS /Reading material /L04 /WhyIsntEveryPhysicistABayesian_RCousins.pdf

https://www.astro.princeton.edu/~strauss/AST303/bayesian_paper.pdf

Recall Bayes' Theorem terminology

Review

Where's $\mathscr{L}$?

likelihood *prior*

posterior

$$P(F \mid E) = \frac{P(E \mid F)\ P(F)}{P(E)}$$

normalization constant

Likelihood function

Under the assumption of the hypothesis $f(x;\theta)$, including the value of θ

The probability that x_i in $[x_i, x_i + dx_i]$ for all $i = \prod_{i=1}^{n} f(x_i;\theta)dx_i$

Since the dx_i do not depend on the parameter θ

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$ ← PDF of x

↑ parameter(s) we wish to estimate

This is the ***joint PDF*** for the x_i, although it is treated as a function of the parameter θ (i.e., it's really $\mathscr{L}(x_i;\theta)$). The x_i, on the other hand, are treated as **fixed** (i.e., the experiment is over).

Maximum likelihood estimation (MLE)

Maximum likelihood estimation is just a systematic way of searching for the parameter values of our *chosen distribution* that maximize the probability of observing the data that we observe

i.e., we have the data but want to learn about the model, specifically the model's parameters.

Maximum likelihood estimators

def maximum likelihood estimators for the parameters ≡ those which maximize the likelihood function

$$\frac{\partial \mathscr{L}}{\partial \theta_i} = 0, \quad i = 1,\ldots,m$$

Often not so easy…

Solution: Take the logarithm

→ *Monotonically increasing* *(param. val. which maximizes $\mathscr{L}$, maximizes $\log \mathscr{L}$)*

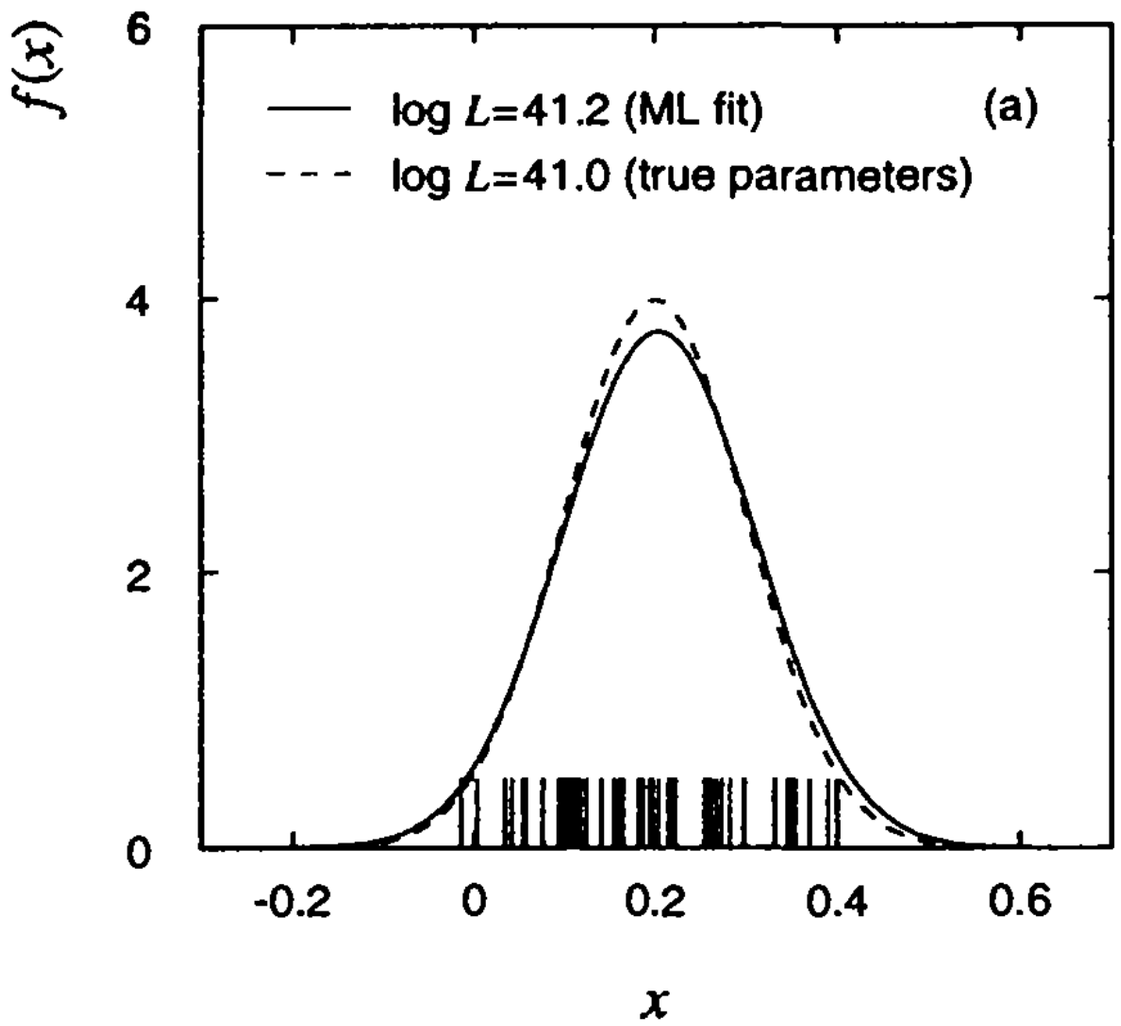
→ *exponentials in f are converted into simple factors*

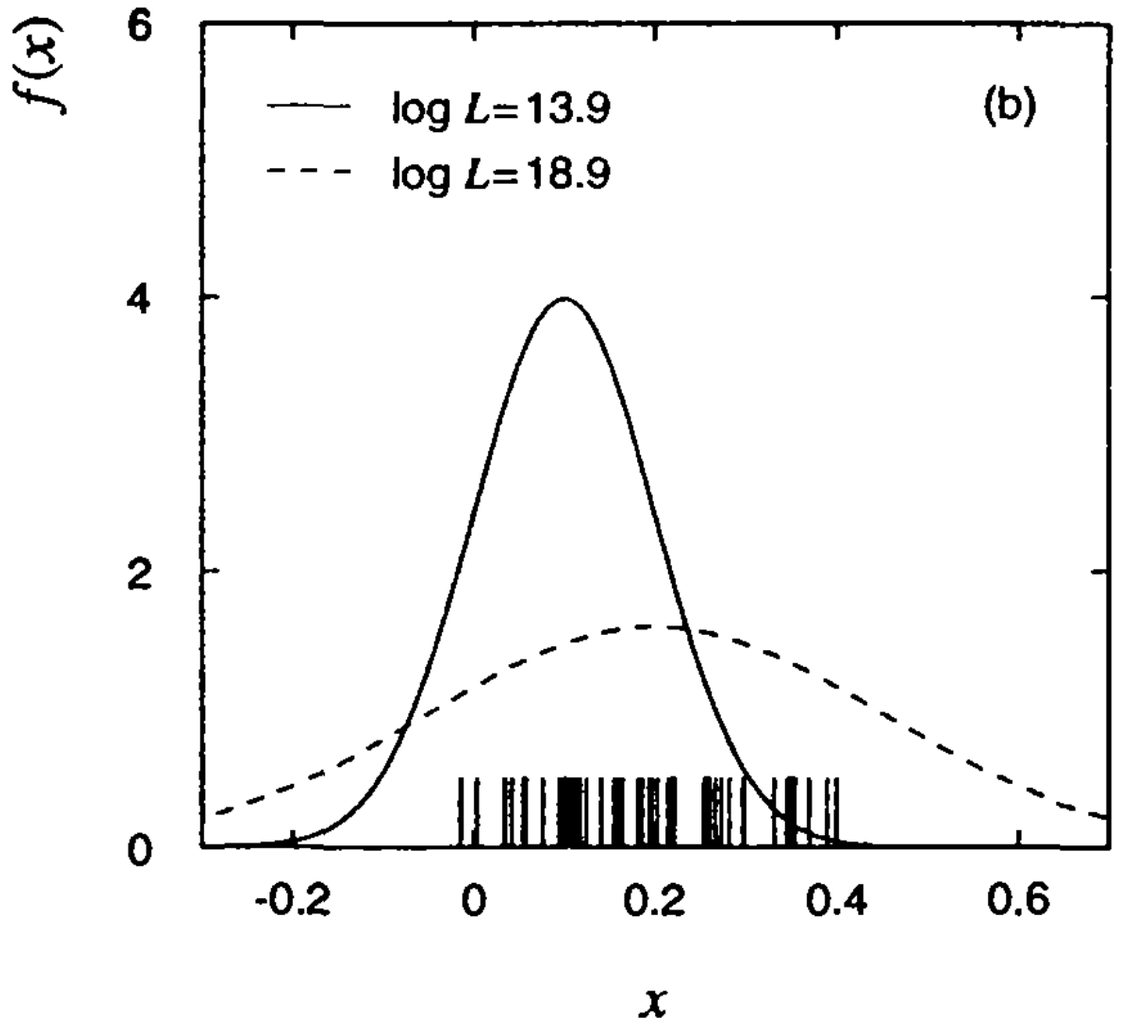
$\rightarrow \prod \Rightarrow \sum$

$$\log \mathscr{L}(\theta) = \log \left(\prod_{i=1}^{n} f(x_i; \theta) \right) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

Good vs. bad ML estimates

- **A sample of 50 observations (shown as tick marks) of a Gaussian random variable.**
  - Left: mean and width parameters that maximize the ML (solid shows estimated, dashed shows true PDF)
  - Right: The PDF evaluated with parameters far from the true values, giving lower likelihoods





http://www.pp.rhul.ac.uk/~cowan/sda/

Maximum likelihood estimator

...an example

Exponential distribution

Given: The proper decay times of unstable particles of a certain type have been **measured** for n decays, yielding values $t_i, \ldots t_n$

Choose: The exponential PDF with mean τ as a **hypothesis** for the distribution of t

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

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

Maximize $\log \mathscr{L}$ w.r.t. τ →

$$\frac{\partial \log \mathscr{L}(\tau)}{\partial \tau} = 0$$

Gives the ML estimator $\hat{\tau}$ →

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

✓

$$E[\hat{\tau}(t_1, t_2, \ldots, t_n)] = \frac{1}{n} \sum_{i=1}^{n} \tau = \tau$$

$\hat{\tau}$ is an unbiased estimator for τ

Read up on Gaussian example at home (pg74, Cowan)

Variance of ML estimators

So far: Method to determine θ via ML and get $\hat{\theta}$ ✓

Now: What is the variance of $\hat{\theta}$?

*i.e., if we **repeat our experiment** a large number of times, how widely spread out would $\hat{\theta}$ be?*

3 methods:

1. Analytical calculation of $V[\hat{\theta}]$
2. MC method
3. RCF bound / graphical technique

Continue using the exponential function as an example since *applicable to all 3 methods*

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

Variance of ML estimators: analytic method

For a **very limited** number of cases, one can compute the variances of the ML estimators **analytically**

$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2$$

Recall: $E[x] = \int_{-\infty}^{\infty} x f(x) ds$ (1.39, Cowan)

$$= \int \dots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right)^2 \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} \, dt_1 \dots dt_n$$

$$- \left(\int \dots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} \, dt_1 \dots dt_n \right)^2$$

$$= \frac{\tau^2}{n}$$

No surprise here: Recall we derived this on slide 17

How would we report this?

$$\hat{\tau} \pm \hat{\sigma}_{\hat{\tau}}$$

Estimate from one experiment

Estimated standard deviation one expects $\hat{\tau}$ to vary by if experiment is repeated many times with the same number of measurements per experiment

Variance of ML estimators: MC method (i)

Often too difficult to compute the variances analytically

Use the **MC method** to investigate the distribution of the ML estimates

1. **Simulate a large # of experiments** → **"true" parameter = the estimated value from the real experiment**
2. **Compute the ML estimates each time**
3. **Look at how the resulting values are distributed**

Unbiased estimator for the variance of a PDF

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

→ Compute s for the ML estimates *obtained from the MC experiments* and give this as the statistical error of the parameter estimated from the real measurement

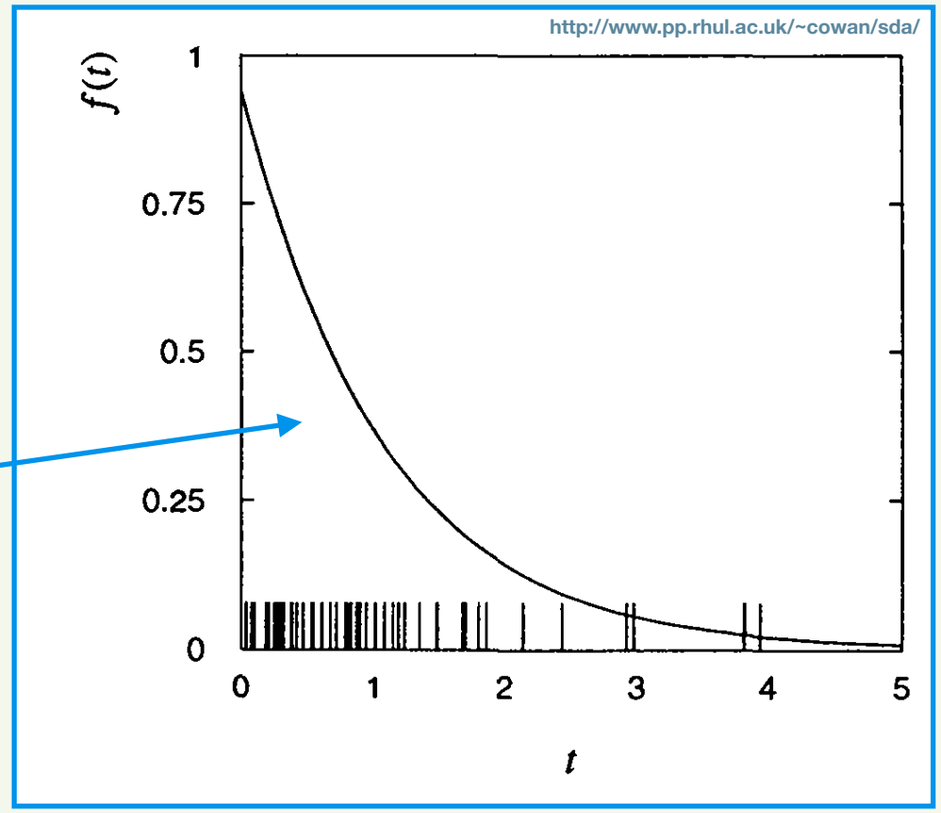
Variance of ML estimators: MC method (ii)

For arguments sake, pretend these 50 observations (tick marks) **are from a real experiment**

In reality, these are 50 MC generated observations of an exponential variable t with mean $\tau = 1.0$

The curve shows the exponential PDF evaluated with the ML estimate $\hat{\tau} = 1.062$

Use this as the "true" parameter τ going forward

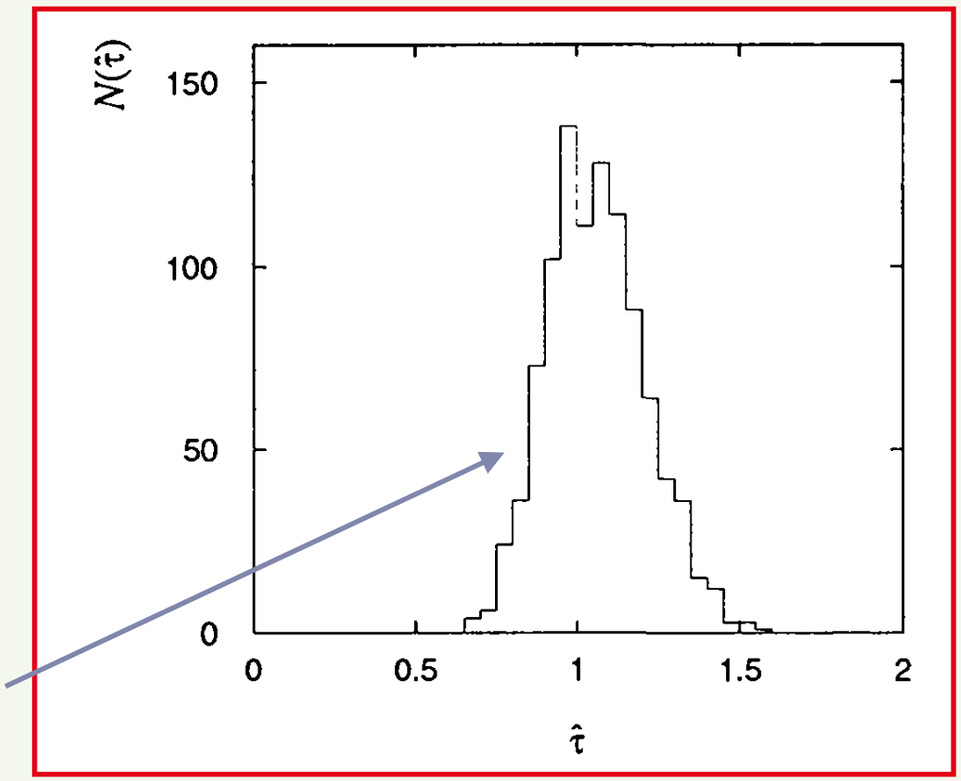


Now simulate 1000 experiments with 50 measurements each using $\hat{\tau} = 1.062$

- Sample mean of the estimates $\bar{\hat{\tau}} = 1.059$
- Sample std. dev. of the 1000 experiments is $s = 0.151$

Analytical std. dev. (3 slides back) is $\hat{\sigma}_{\hat{\tau}} = \hat{\tau}/\sqrt{n} = 1.062/\sqrt{50} = 0.150$

$\approx$



Note that the distribution is approximatively Gaussian in shape. This is a general property of ML estimators for the large sample limit: **asymptotic normality**

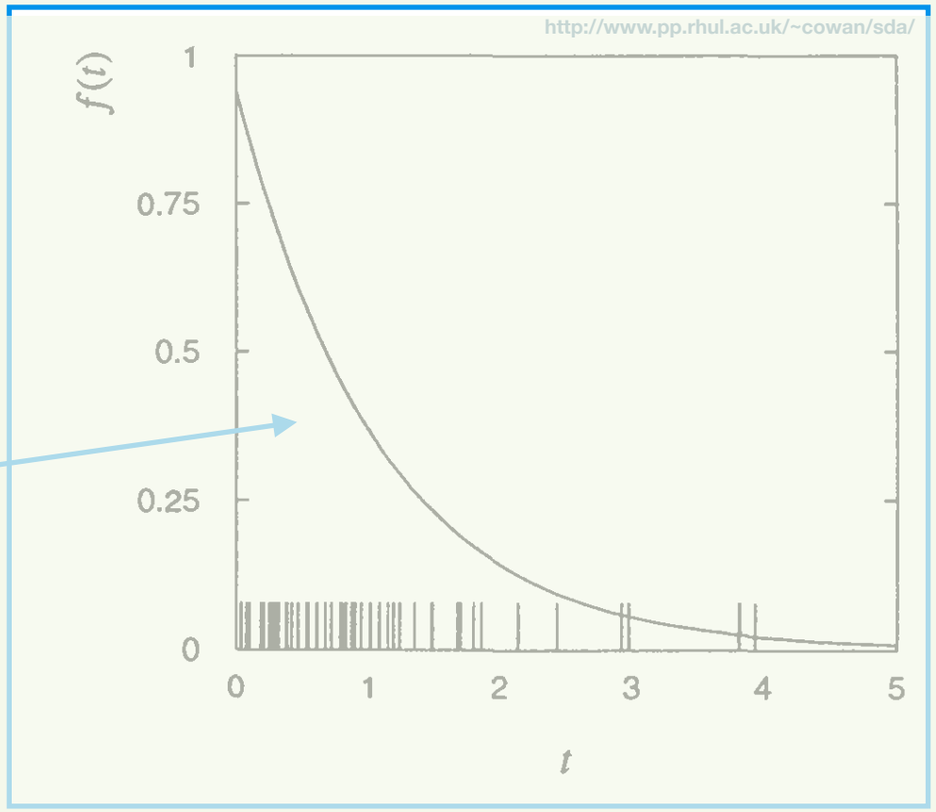
Variance of ML estimators: MC method (iii)

For arguments sake, pretend these 50 observations (tick marks) are from a real experiment

In reality, these are 50 MC generated observations of an exponential variable t with mean $\tau = 1.0$

The curve shows the exponential PDF evaluated with the ML estimate $\hat{\tau} = 1.062$

Use this as the "true" parameter τ going forward



But what if we want the decay constant instead of the mean lifetime?

$$\lambda = \frac{1}{\tau}$$

In general: transformational invariance

$\partial L/\partial\theta = 0$ implies $\partial L/\partial a = 0$ at $a = a(\theta)$ unless $\partial a/\partial\theta = 0$.

$$\frac{\partial L}{\partial\theta} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial\theta} = 0.$$

But with

$$\hat{\lambda} = 1/\hat{\tau} = n/\sum_{i=1}^{n} t_i.$$

$$E[\hat{\lambda}] = \lambda\frac{n}{n-1}$$

…biased for small n

Variance of ML estimators: RCF bound (i)

Often too difficult to compute the variances analytically and a MC study involves a significant effort

Proof: ILIAS /Reading material / L04 /RCF_proof.pdf

Use the Rao-Cremer-Frechet (RCF) inequality (aka information inequality) to compute the lower bound on an estimator's variance

$$V[\hat{\theta}] \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{E\left[-\frac{\partial^2 \log \mathscr{L}}{\partial \theta^2}\right]}$$

For the **exponential distribution** with mean τ

Recall we chose this distribution since it ***can be solved analytically***

$$\frac{\partial^2 \log \mathscr{L}}{\partial \tau^2} = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

$$\partial b/\partial \tau = 0, \quad b = 0 \quad \text{(slide 29)}$$

$$V[\hat{\tau}] \geq \frac{\tau}{n}$$ ✓

In the case of equality (minimum variance), the estimator is said to be **efficient**

Key points:

1) If efficient estimators exist for a given problem, **the ML method will find them**
2) **ML estimators are always efficient in the large sample limit** (exept when the extent of the sample space depends on the estimated parameter)

…but what if you ***<u>can't</u>*** *compute the RCF bound analytically?*

Variance of ML estimators: RCF bound (ii)

One can estimate V^{-1} by evaluating the second derivative with the measured data and the ML estimates $\hat{\theta}$:

$$\left(\hat{V^{-1}}\right)_{ij} = \left.\frac{\partial^2 \log \mathscr{L}}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}}$$

For a single parameter θ this reduces to:

$$\hat{\sigma^2}_{\hat{\theta}} = \left.\left(-1 \Big/ \frac{\partial^2 \log \mathscr{L}}{\partial\theta^2}\right)\right|_{\theta=\hat{\theta}}$$

*The routines **MIGRAD** and **HESSE** in **MINUIT** determine numerically the matrix of second derivatives of $\log \mathscr{L}$ using finite differences, evaluate it at the ML estimates, and invert to find the covariance matrix.*

Intuition

Why is this how we calculate the standard errors?

The easy way to think about this is to recognize that the ***curvature of the likelihood function*** *tells us how certain we are about our estimate of our parameters. The* ***more curved*** *the likelihood function, the* ***more certainty*** *we have that we have estimated the right parameter. The second derivative of the likelihood function is a measure of the likelihood function's curvature - this is why it provides our estimate of the uncertainty with which we have estimated our parameters.*

If the curvature is small, then the likelihood surface is flat around its maximum value (the MLE). If the curvature is large and thus the variance is small, the likelihood is strongly curved at the maximum.

https://www.math.arizona.edu/~jwatkins/n-mle.pdf

Variance of ML estimators: graphical technique (i)

Key ingredient: Taylor expand log likelihood function around maximum:

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

$\log\mathscr{L}(\hat{\theta}) = \log\mathscr{L}_{\max}$ by definition of $\hat{\theta}$

$= 0$ at maximum

$\hat{\sigma}^2_{\hat{\theta}} = \left(-1 \Big/ \frac{\partial^2 \log\mathscr{L}}{\partial\theta^2}\right)\Big|_{\theta=\hat{\theta}}$ RCF

$$\log\mathscr{L}(\theta) = \log\mathscr{L}_{\max} - \frac{(\theta-\hat{\theta})^2}{2\hat{\sigma}^2_{\hat{\theta}}} \qquad \textit{or}$$

$$\log\mathscr{L}(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log\mathscr{L}_{\max} - \frac{1}{2}$$

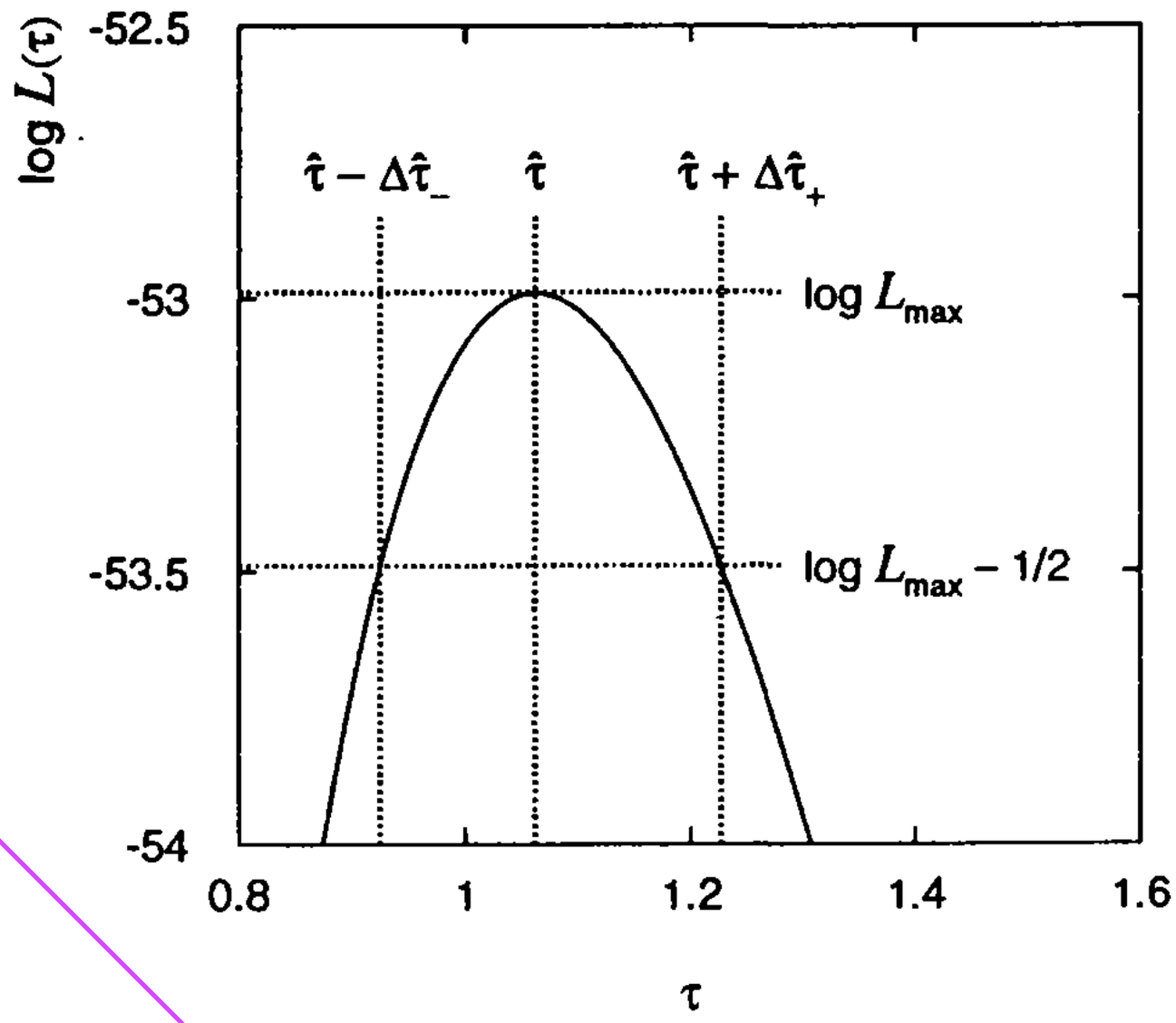
With known maximum, can determine estimator of variance by solving what value of σ_θ gives a likelihood value of $\log\mathscr{L}_{\max} - 1/2$

Variance of ML estimators: graphical technique (ii)

Return to our previous example using the exponential distribution

- Reading off from the curve
  - $\Delta\hat{\tau}_- = 0.137$
  - $\Delta\hat{\tau}_+ = 0.165$
- Both reasonably close and we find
  - $\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$
- We will later make a reinterpretation of the interval $[\tau - \sigma_\tau, \tau + \sigma_\tau]$ as an approximation of the **68.3% central confidence interval**

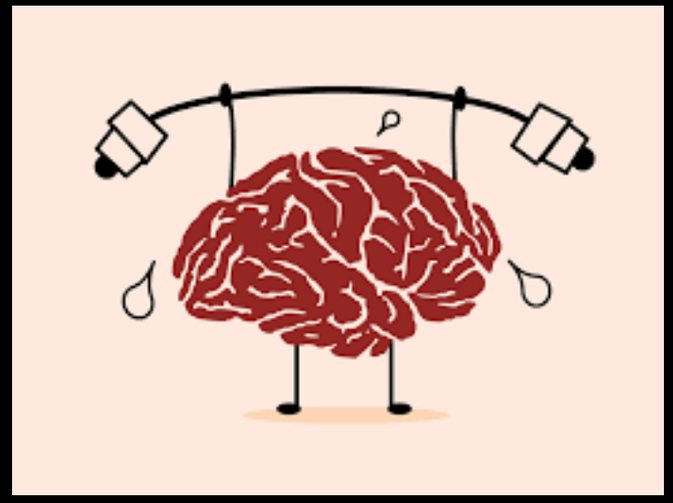
http://www.pp.rhul.ac.uk/~cowan/sda/



*This leads to approximately the **same result** as from the exact standard deviation $\tau/\sqrt{n}$ evaluated with $\tau = \hat{\tau}$*

(slides 31 & 34)

Take a second



Another ML example: 2 parameters (i)

- Consider a particle reaction where each scattering event is characterized by a certain scattering angle Θ (or equivalently $x = \cos\Theta$)
  - Suppose now a theory predicts that this follows an angular distribution given by

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+2\beta/3}.$$

e.g. $\alpha = 0$, $\beta = 1$ is the **LO QED expectation** for

$$e^+e^- \to \mu^+\mu^-$$

- To make this slightly more complicated, let's assume we have a finite detector acceptance, such that $x_{min} < x < x_{max}$:

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{(x_{\max}-x_{\min})+\frac{\alpha}{2}(x_{\max}^2-x_{\min}^2)+\frac{\beta}{3}(x_{\max}^3-x_{\min}^3)}.$$

Ensures proper normalization and makes f a PDF over $[x_{min}, x_{max}]$

Another ML example: 2 parameters (ii)

- MC experiment with 2000 events according to
  - $\alpha = 0.5,\ \beta = 0.5$
  - $x_{min} = -0.95,\ x_{max} = 0.95$

- Numerically maximizing the log-likelihood function we find

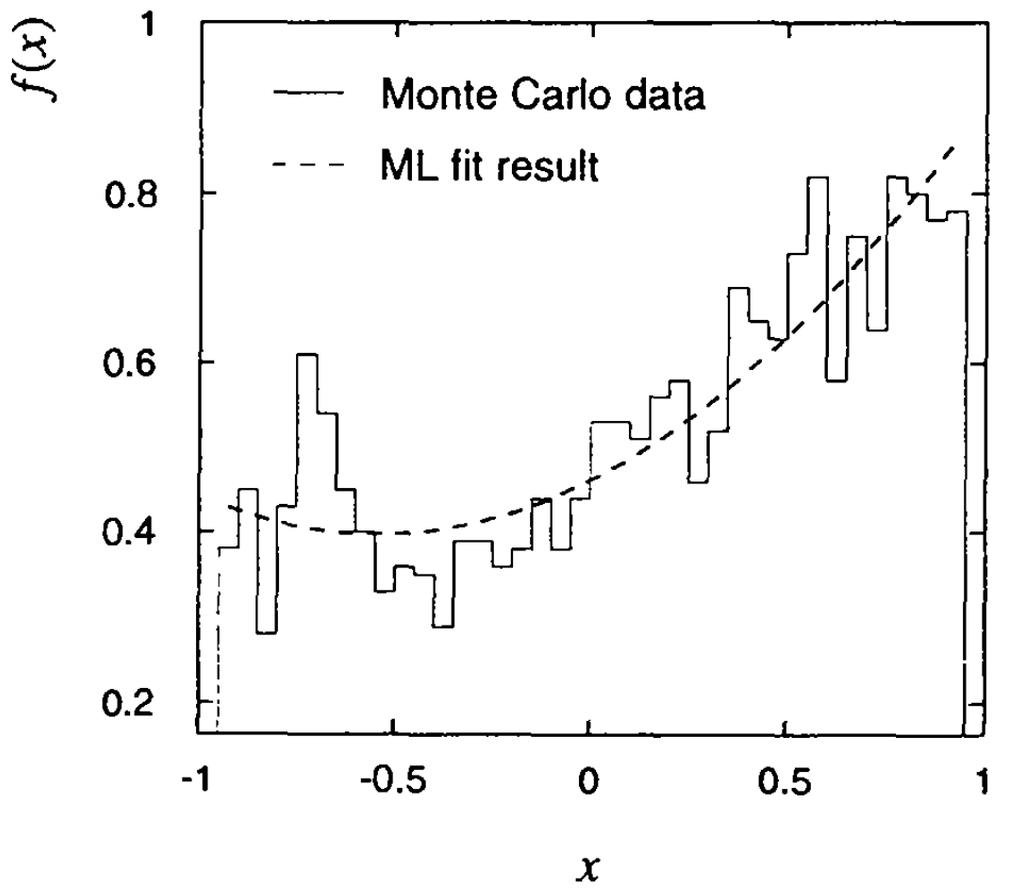
$$\hat{\alpha} = 0.508 \pm 0.052,$$

$$\hat{\beta} = 0.47 \pm 0.11,$$

from ML maximum

from second derivatives

http://www.pp.rhul.ac.uk/~cowan/sda/



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

$$\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0026 \qquad r = 0.46.$$

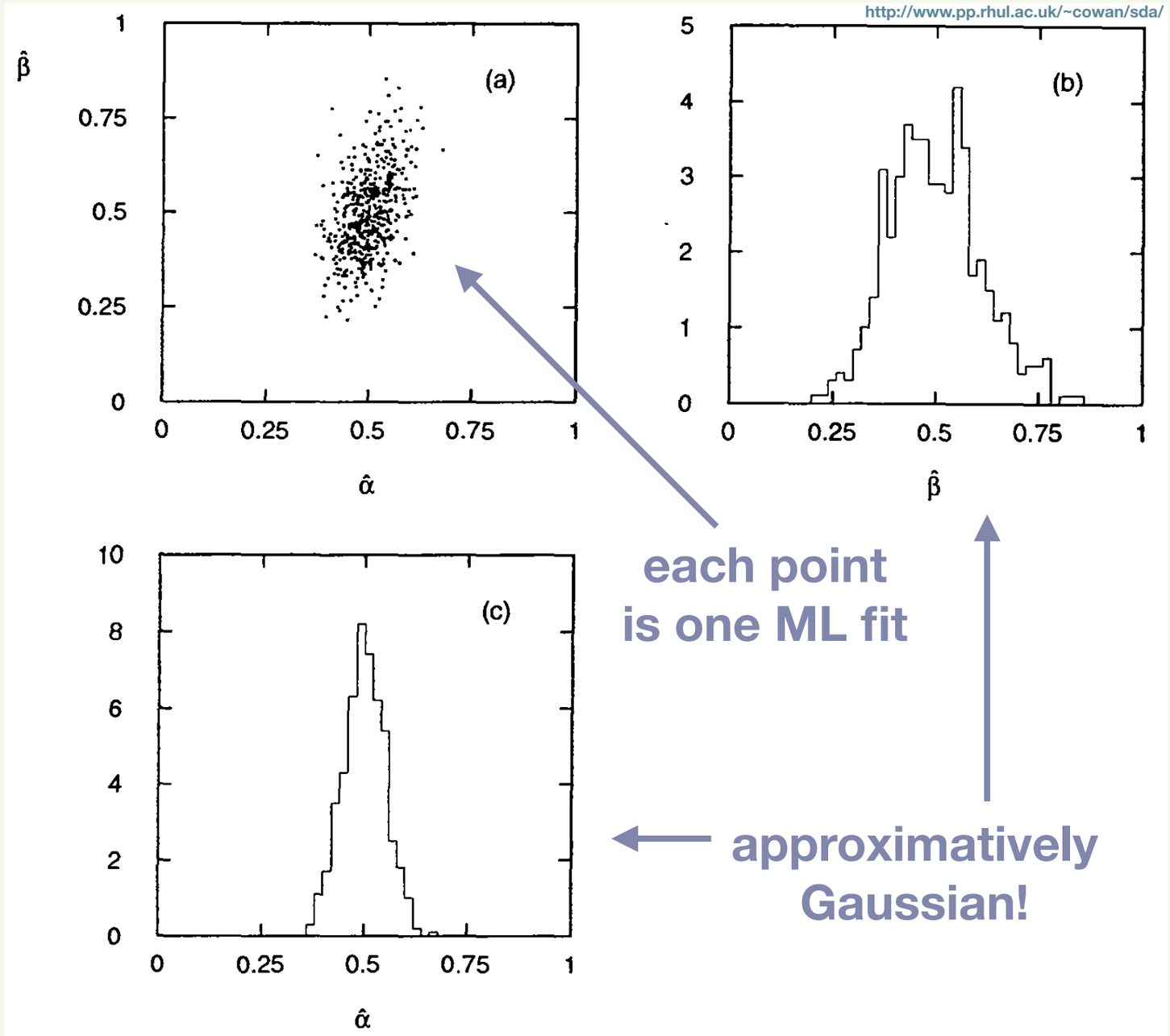
Another ML example: 2 parameters (iii)

- Can validate this result by using MC techniques
  - Let's produce 500 similar experiments, all with 2000 events with the true values for $\alpha = 0.5,\ \beta = 0.5$
  - Sample means, standard deviations, covariance and correlation

In good agreement with true values

$$\begin{aligned} \overline{\hat{\alpha}} &= 0.499 & \overline{\hat{\beta}} &= 0.498 \\ s_{\hat{\alpha}} &= 0.051 & s_{\hat{\beta}} &= 0.111 \\ \widehat{\mathrm{cov}}[\hat{\alpha},\hat{\beta}] &= 0.0024 & r &= 0.42. \end{aligned}$$

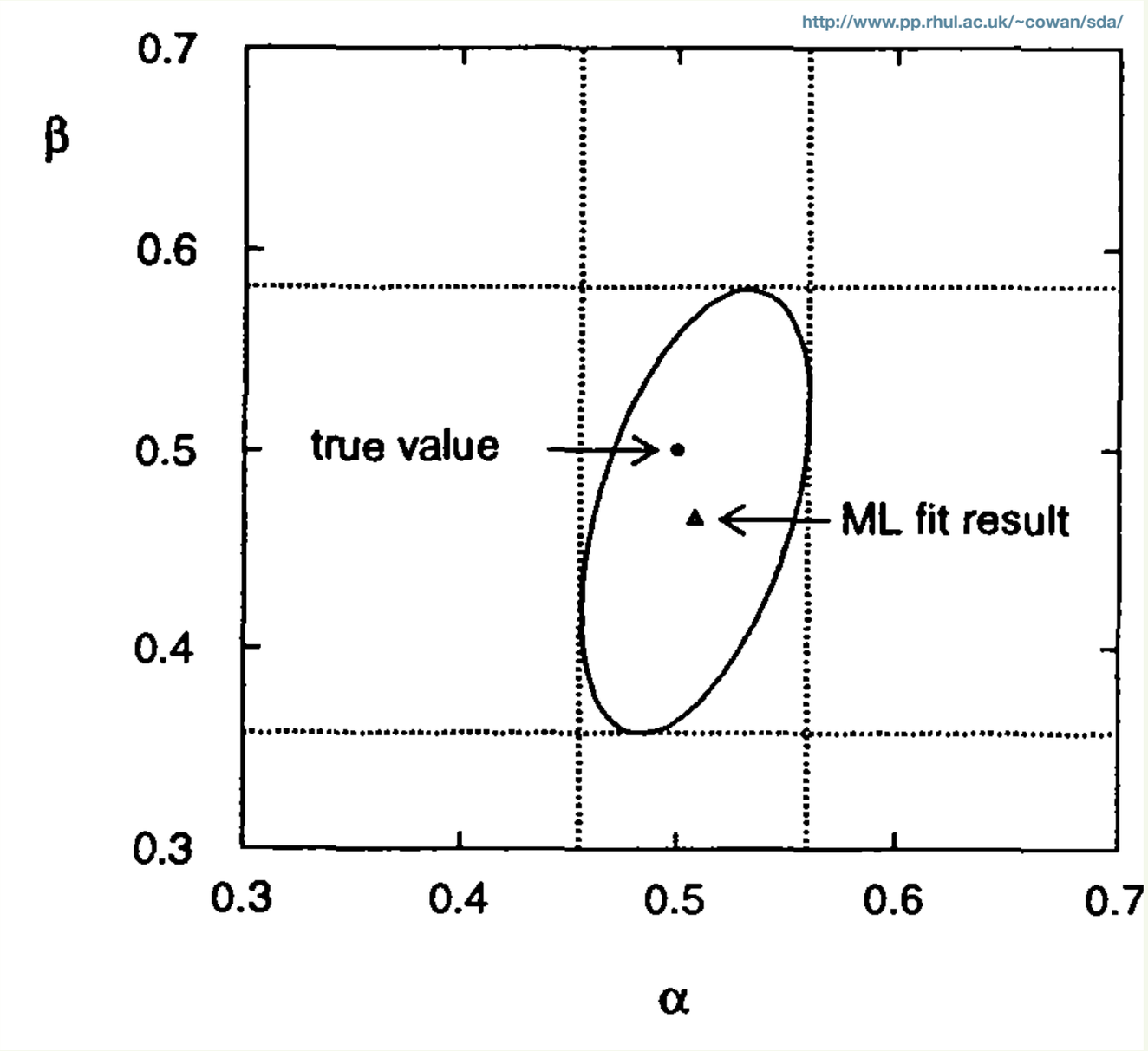
http://www.pp.rhul.ac.uk/~cowan/sda/



Sample variance, covariance, correlation: Good agreement with RCF bound

Another ML example: 2 parameters (iv)

- Can also draw likelihood contour in 2D
  - Contour of $\log\mathscr{L} = \log\mathscr{L}_{\max} - \frac{1}{2}$



- Large sample limit contour given by

$$\frac{1}{1-\rho^2}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)^2 + \left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)\right] = 1.$$

- I.e. ellipse with center at $(\hat{\alpha}, \hat{\beta})$ and angle ϕ with respect to the α-axis

$$\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}.$$

For next time

- Required reading
  - Cowan textbook: chapters 5 & 6.1-6.8
  - Reading material / L04 /
    - VarianceOfMLEstimators
    - UnbiasedEstimators
    - RCF_proof

- Extra reading for fun: /Reading material / L04 /
  - WhyIsntEveryPhysicistABayesian_RCousins

*Listed as 'fun' but you should **really** read it!*

Next time

- Extended maximum likelihood
- Maximum likelihood with binned data
- Testing goodness-of-fit
- χ^2 method

Quiz Time: 4th Round

2. Let us assume you have a simple PDF of the form

$$f(x;\lambda) = 1 + \lambda\,(x - 0.5)\ , \tag{2}$$

with a sample space S spanning the interval $[0, 1]$ such that $\int_S f(x;\lambda)\mathrm{d}x = \int_0^1 f(x;\lambda)\mathrm{d}x = 1$.

a) First sketch the PDF for $\lambda = 1, 0, -1$.

b) Five measurements were done giving $x = (0.89, 0.03, 0.50, 0.31, 0.49)$. Calculate the log-likelihood function for three different values of $\lambda = 1, -0.5, -1$ by hand.

c) The log-likelihood function is in good approximation a parabolic function, i.e. can be described by a polynomial of second order as a function of the tested value λ. Calculate the coefficients a, b, c of $\log L(\lambda) = a\lambda^2 + b\lambda + c$. For what value of λ is $\log L(\lambda)$ maximal?

d) Sketch the log-likelihood function. Using the graphical method, i.e.

$$\log L(\hat{\lambda} \pm \hat{\sigma}_{\hat{\lambda}}) = \log L_{\max} - \frac{1}{2}\,, \tag{3}$$

determine the uncertainty $\hat{\sigma}_{\hat{\lambda}}$ of the estimated parameter $\hat{\lambda}$ (with $\hat{\lambda}$ denoting the value where $\log L(\hat{\lambda}) = \log L_{\max}$ and $L_{\max}$ is the maximal likelihood value).

Bibliography

- Part of the material presented in this lecture is adapted from the following sources. See the active links (when available) for a complete reference
  - **Probability for CS** (Stanford): slides 3-4
  - **Statistical Data Analysis** textbook by G. Cowan (U. London): all figures with white background