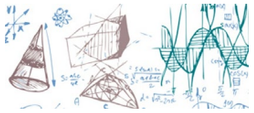
Introduction to probability, statistics and data handling

Agnieszka Obłąkowska-Mucha
based on lectures by Tomasz Szumlak

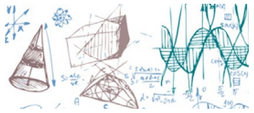
Faculty of Physics and Applied Computer Science
AGH University of Krakow

Model and real life



- So, we collected the data – we are going to be interested in a procedure, which basing on the observed variation gives the best value (we could also ask about the range of values) for the corresponding underlaying model parameter(s)
- Again, using the stat lingo we want to get **the best possible estimate of the value of the parameter**(s)
- That is what the point estimation is all about
- BTW, it may also be useful to estimate the range of „good” parameter values – that is yet another story called estimation with confidence – we are going to look at this next time!

Point Estimators-remainder

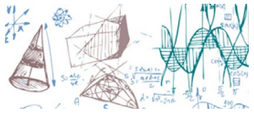


- Let's set a **generic procedure** using this simple example
- First, we **pick the parameter** to be estimated
- Next, we need to **collect data** and **compute a sampling statistics** using a formula corresponding to the parameter we are interested in
- In our example that is a **sample mean**

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

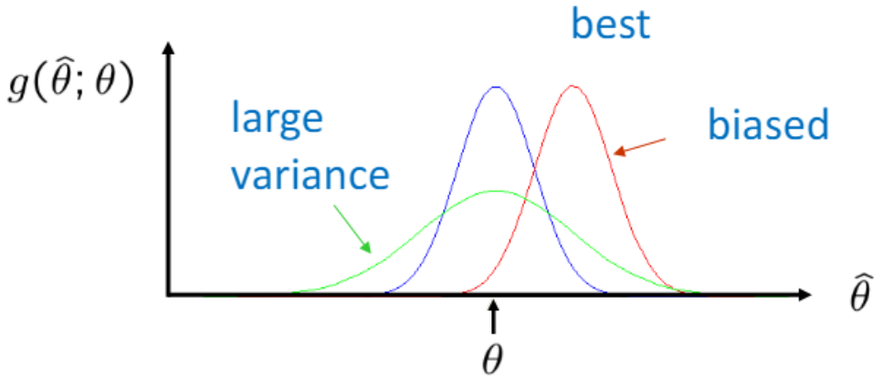
- This, in turn, we call an **estimator** of true parameter, in our case this would be: $\mu \rightarrow \bar{X} = \hat{\mu}$ (we use the caret symbor "^")
- Remember – the estimator is a random variable, for different sample we are going to get different value
- The estimator will follow its own distribution – **sampling distribution of the estimator**

Short recap



- We learned about parameter(s) point estimation
- We set up a general recipe and considered some more intuitive and some not so intuitive examples (slopes)
- We discussed two methods that are popular and often used when estimation theory is needed:
  - Least squares
  - Method of moments

- We also learned about bias and efficiency of the estimators, which are related not to the obtained values from the samples but rather to the formulas that represent the said estimators.
- We want small bias (systematic error): $b = E(\hat{\theta}) - \theta$
- We want a small variance (statistical error): $V(\hat{\theta})$



small bias & variance are in general conflicting criteria

Maximum Likelihood Estimation (MME)

- Let's begin our tale on the **Maximum Likelihood** technique then
- Consider a R.V. X that is described by a P.D.F. $f(x)$. The support set $supp(f) = \{x \in X, f(x) \neq 0\}$ is also the sample space. Then we take n independent observations which form a data sample $\vec{\boldsymbol{x}}_{(1)} = (x_1, x_2, \dots, x_n)$. Now, we can take another sample, and another, and…
- By this „innocent" operation, we redefined the sample space to be a space of all $n - \dim$ vectors $\vec{x}_{(i)}$ and the first experiment is just **a single measurement** in that space
- Now, something very important, by our assumption of I.D.R.V. and independence we can construct a **joint P.D.F.**

$$f_{\vec{x}_{(1)}}(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) f(x_2|\theta) \cdots f(x_n|\theta)$$

- If the sample is large this formula is seemingly complicated but remember each R.V. is identically distributed!
- This is the function with the data obtained and regard it as a function of the parameters. It is called the **likelihood function**:

$$L(\theta) = P(\vec{x}/\theta)$$

Maximum Likelihood Estimation (MLE)

- Before an experiment is performed the outcome is unknown.
- Probability allows us to predict ***unknown*** outcomes based on ***known*** parameters:

$$P(data/\theta)$$

for example: $P(x/n,p)=\binom{n}{x}p^x(1-p)^{n-x}$

- After the experiment is performed the outcome is ***known***.
- Now we talk about the ***likelihood*** that a parameter would generate the observed data:

$$P(\theta/data)$$

like here: $P(p/n,x)=\binom{n}{x}p^x(1-p)^{n-x}$

so: $$P(data/\theta)=P(\theta/data)$$

Estimation proceeds by finding the value θ of that makes the observed data most likely

Likelihood Ratio Test (LRT)

Example: Ratio of Likelihoods – comparison of hypothesis

Suppose we have a coin which – as we happen to know – is not a fair one. Namely – one side is likely to happen twice as frequently as the second one but ... we do not know which one.
we perform an experiment: flipping the coin 6 times and we get 4 heads (and 2 tails). We have two possibilities:

First case	Second case
(1) $\mathcal{P}(H) = 2/3; \quad \mathcal{P}(T) = 1/3$	(2) $\mathcal{P}(H) = 1/3; \quad \mathcal{P}(T) = 2/3$
$W_4^6 = \binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2$	$W_4^6 = \binom{6}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2$

$$\frac{L_1}{L_2} = \frac{\binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2}{\binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2} = \ldots = 4.$$

Obviously, the first hypothesis about the $\mathcal{P}$'s is the better one.

A.Lenda

Maximum Likelihood Estimation (MLE)

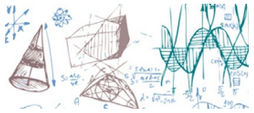
- The game is still the same: using the observed x_i we want to **infer the properties of the parent P.D.F**. (its parameters). We already seen, that by constructing special functions of R.V.s we can estimate parameters
- Here, these functions (**statistics**) are called estimators (e.g. the sample mean)
- One more thing – we need to propose a hypothesis regarding the concrete form of the parent function in a form $f = f(\vec{x}|\vec{\theta}), \vec{x} = (x_1, x_2, \dots, x_n), \vec{\theta} = (\theta_1, \theta_2, \dots, \theta_n)$
- Then, when taking the data sample we can use our joint P.D.F. to evaluate the probability of observing this particular sample as (let's take the discrete case for starters):

$$p(X_1 = x_1, \dots, X_n = x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \prod_i f(x_i|\theta)$$

$$\mathcal{L} = \prod_i f(x_i|\theta)$$

Likelihood function

Statement of the MLE



- When a sample of n independent and I.D.R.V. is collected it is possible to construct the **likelihood function** of unknown parameter θ for the random sample $\vec{x}_{(1)}$, the value $\hat{\theta}$ of the unknown parameter at which the **likelihood is maximised** is called the **maximum likelihood estimator** for that parameter
- And for the C.R.V. we get pretty much the same but now we need to formally requested that the value observed during the experiment is within an interval $[x_i, x_i + dx_i]$, and the prob.:

$$p(x_i \in [x_i, x_i + dx_i], \ldots | \theta) = f(x_1|\theta)dx_1 \cdots f(x_n|\theta)dx_n = \prod_i f(x_i|\theta)\, dx_i$$

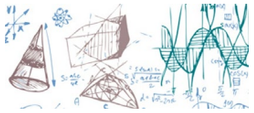
- Since any of the dx_i do not depend on parameter, we are going to end up with the same likelihood function!
- Note, that the intuitive reasoning here is as follow: if the P.D.F. we proposed is the **correct** one we should observe a **high probability** for **observing the sample** for the right value of θ

Statement of the MLE

- Now, with such intuitive interpretation, the ML estimator(s) for the parameter(s) are defined as those that maximise the likelihood function (we silently assumed that $\mathcal{L}$ is differentiable):

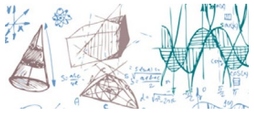
$$\frac{\partial \mathcal{L}}{\partial \theta_j} = 0, j = 1,2, \dots, m$$

Statement of the MLE



- Ok, now we are finished! Just a short note on the **properties of the ML estimators** (MLEs)
- It can be shown that the MLE are often unbiased: $E\left[\hat{\theta}\right] = \theta$, or at least asymptotically unbiased: $E\left[\hat{\theta}\right] \to \theta, n \to \infty$
- MLEs are consistent: $V\left[\hat{\theta}\right] \to 0, n \to \infty$
- MLEs are asymptotically normally distributed (this feature now pertains to the sampling distribution of the estimators, we know that it is important – remember that when we start talking about statistical tests)

Example – geometric P.D.F.



- We are going to discuss the properties of the ML using several examples. Let's start with an easy one
- Some phenomena is known to be governed by a geometric density function:

$$p(x,\theta) = (1-\theta)^{x-1}\theta, x = 1, 2, \ldots$$

A random sample of three observations was taken: $\vec{x} = \{3, 4, 8\}$

- The likelihood function is then:

$$\mathcal{L}(\boldsymbol{\theta}|\vec{\boldsymbol{x}}) = p(3,\theta) \cdot p(4,\theta) \cdot p(8,\theta) =$$

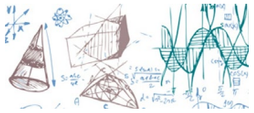
$$= [(1-\theta)^{3-1}\theta] \cdot [(1-\theta)^{4-1}\theta] \cdot [(1-\theta)^{8-1}\theta] = \cdots = (1-\theta)^{12}\theta^{3}$$

- A technical note: it is much easier to handle sums than products, especially when differentiation is required:

$$\mathcal{L}(\theta|\vec{x}) = \prod_i f(x_i|\theta) \rightarrow \ln\big(\mathcal{L}(\theta|\vec{x})\big) = \sum_i f(x_i|\theta)$$

$$\ln\big(\mathcal{L}(\theta|\vec{x})\big) = 12 \cdot \ln(1-\theta) + 3 \cdot \ln\theta \rightarrow \frac{d(\ln\mathcal{L})}{d\theta} = 0$$

Example – geometric P.D.F.



- For the record we can write out both ways

$$\frac{d(\ln\mathcal{L})}{d\theta} = -\frac{12}{1-\theta} + \frac{3}{\theta} \rightarrow \hat{\theta} = 0.2$$

$$\frac{d(\mathcal{L})}{d\theta} = -12 \cdot (1-\theta)^{11} \cdot \theta^3 + 3 \cdot (1-\theta)^{12} \cdot \theta^2 \rightarrow \hat{\theta} = 0.2$$

- **The log likelihood much more „user friendly”**
- We know, that in case of the geometric distribution $f(x;p)$ its mean value can be used to estimate the p:

$$E[x] = \frac{1}{p} \rightarrow \hat{p} = \frac{1}{E[x]}, x \in N, p \equiv \theta$$

- So, using the method of moments, we get the same result:

$$E[x] = \frac{3+4+8}{3} = \frac{15}{3} = 5, \hat{p} = \frac{1}{5} = 0.2$$

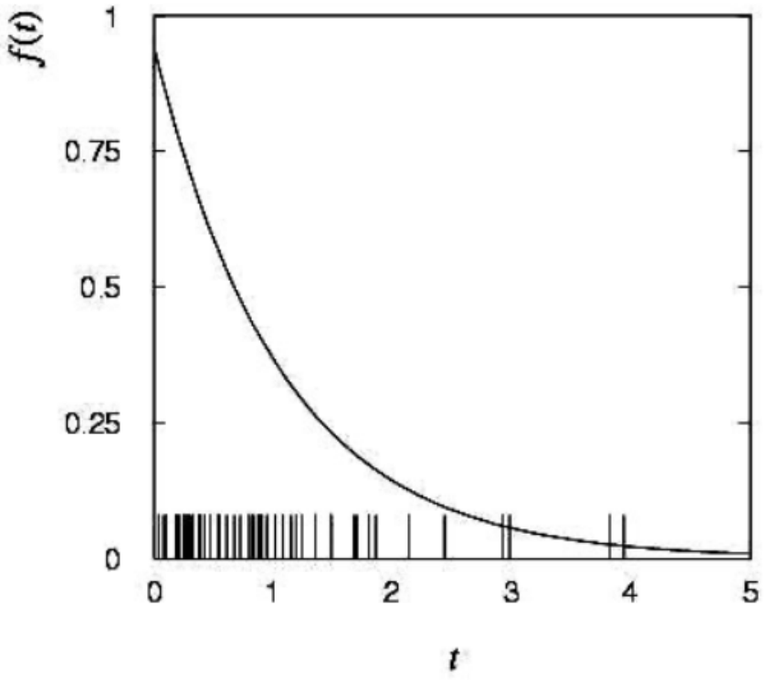
Exponential distribution

- Let's try to find a maximum of exponential distribution $f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$
- Try guessing: $\hat{\tau} = \frac{1}{n}\sum_t t_i$
- Now set:

$$\frac{d\left(\ln \mathcal{L}(\tau; t)\right)}{d\tau} = 0$$

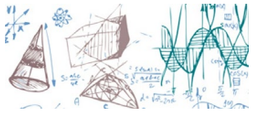
- Compare to Monte Carlo test – generation of 50 values with $\tau = 1$

$$\hat{\tau} = 1.062$$



G.Cowan

Exponential distribution



- It is a very common practice to model spontaneous particle decays using exponential model, say we did an experiment and observed n decays obtaining a sample $\vec{t}_{(1)} = \{t_1, t_2, \dots, t_n\}$
- We attempt to describe the variability of measured times t_i using the following formula

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}, E[f] = \tau$$

- The likelihood function:

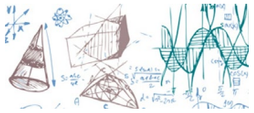
$$\mathcal{L}(\tau;t) = \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} = \frac{1}{\tau^n} e^{-\sum_i \frac{t_i}{\tau}}$$

$$\ln\mathcal{L}(\tau;t) = \ln\frac{1}{\tau^n} - \sum_i \frac{t_i}{\tau} = n \cdot \ln\frac{1}{\tau} - \sum_i \frac{t_i}{\tau} = \sum_i \left(\ln\frac{1}{\tau} - \frac{t_i}{\tau}\right)$$

$$\frac{d\left(\ln\mathcal{L}(\tau;t)\right)}{d\tau} = \sum_i \left(\frac{t_i}{\tau^2} - \frac{1}{\tau}\right) = \frac{n}{\tau}\sum_i \left(\frac{t_i}{\tau} - 1\right) = 0$$

- And finally we get: $\hat{\tau} = \frac{1}{n}\sum_t t_i$ - just the same as for the method of moments!

Exponential distribution



- It is also very instructive to check for the possible biases in our estimator – let's evaluate its expectation value:

$$E[\hat{\tau}(t_1, t_2, \dots, t_n)] = \int \cdots \int \hat{\tau}(t_1, t_2, \dots, t_n) \cdot f_{\text{joint}}(t_1, t_2, \dots, t_n; \tau) \prod_i dt_i$$

$$= \int \cdots \int \hat{\tau}(t_1, t_2, \dots, t_n) \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \prod_i dt_i =$$

$$= \frac{1}{n} \sum_i \left(\int t_i \frac{1}{\tau} e^{-\frac{t_i}{\tau}} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-\frac{t_j}{\tau}} dt_j \right) = \frac{1}{n} \sum_i \tau = \tau$$

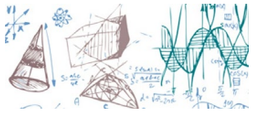
- This formula may look a bit intimidating, but it is not so bad, let's take a detailed look at the case with just two time values

$$\vec{t} = \{t_1, t_2\}, n = 2$$

$$E[\hat{\tau}(t_1, t_2)] = \int \int \left(\frac{1}{n}(t_1 + t_2) \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} dt_1 dt_2 \right)$$

- Not so bad!

Exponential distribution



- Let's expand the sum

$$E[\hat{\tau}(t_1, t_2)]$$

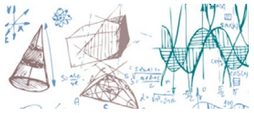
$$= \int\int \left(\frac{1}{n} t_1 \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} dt_1 dt_2 + \frac{1}{n} \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot t_2 \frac{1}{\tau} e^{-\frac{t_2}{\tau}} dt_1 dt_2 \right)$$

- In each component there will be exactly one element with matching exponent this is how we obtained the third line on the previous slide. Also, the respective measurements are independent and we can separate integrals now:

$$E[\hat{\tau}(t_1, t_2)] = \frac{1}{n} \sum_i \left(\int t_i \frac{1}{\tau} e^{-\frac{t_i}{\tau}} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-\frac{t_j}{\tau}} dt_j \right)$$

- Here, $i = 1,2$ and the product will have just one component
- Let's think about integrating this: we are going to have two cases: $\int x e^x dx$ (integration by parts) and $\int e^x dx$

Exponential distribution

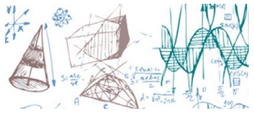


- The latter: $\int e^x dx \rightarrow \left\{ x = -\frac{t_i}{\tau}, dx = -\frac{1}{\tau} dt_i, dt_i = -\tau dx \right\}$
- Integration will yield: $I = -e^{-\frac{t_i}{\tau}}|_0^\infty = -e^{-\frac{\infty}{\tau}} + e^{-\frac{0}{\tau}} = 1$
- And the former: $-\tau e^{-\frac{t_i}{\tau}} \left(\frac{t_i}{\tau} + 1 \right) |_0^\infty = \tau$
- We get after plug in these results:

$$E[\hat{\tau}(t_1, t_2)] = \frac{1}{n} \sum_i (\tau \cdot 1) = \tau$$

- The point here is that the ML estimator for an exponential distribution is not biased – that is, if we are after τ.

Functions of ML estimators



- What would happen, if instead of the mean life-time we would be interested in its reciprocal (called decay constant): $\lambda = \frac{1}{\tau}$?
- In this case we can treat the λ as a function of our model parameter, if we represent such function in general as ω, the likelihood maximisation formula will be as follow:

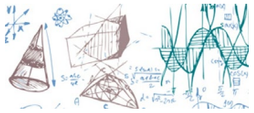
$$\frac{d\mathcal{L}(\theta|\vec{x})}{d\theta} = \frac{\partial\mathcal{L}(\theta|\vec{x})}{\partial\omega}\frac{\partial\omega}{\partial\theta} = 0$$

$$\frac{\partial\mathcal{L}(\theta|\vec{x})}{\partial\theta} = 0 \rightarrow \frac{\partial\mathcal{L}(\theta|\vec{x})}{\partial\omega} = 0, \frac{\partial\omega}{\partial\theta} \neq 0$$

- Or in words: we can evaluate the ML estimator for any function of the original estimator just by inserting the original estimator in

$$\hat{\omega} = \omega(\hat{\theta}) \qquad \hat{\lambda} = \frac{1}{\hat{\tau}} = \frac{n}{\sum_i t_i}$$

Normal distribution



- Similar analysis can be performed for the normal distribution to estimate its mean and variance (this time we are not going to do the exact calculations just cite the results, again you are encouraged to repeat them)
- The log likelihood function for $\mathcal{N}(x;\mu,\sigma^2)$:

$$\ln\mathcal{L}(\mu,\sigma^2)=\sum_i\left(\ln\frac{1}{\sqrt{2\pi}}+\frac{1}{2}\ln\frac{1}{\sigma^2}-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

- The minimisation will give us:

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

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

$$E\left[\widehat{\sigma^2}\right]=\frac{n-1}{n}\sigma^2$$

- ML estimator for the $\mathcal{N}(x;\mu,\sigma^2)$ variance is biased

The MLE's distribution becomes Gaussian

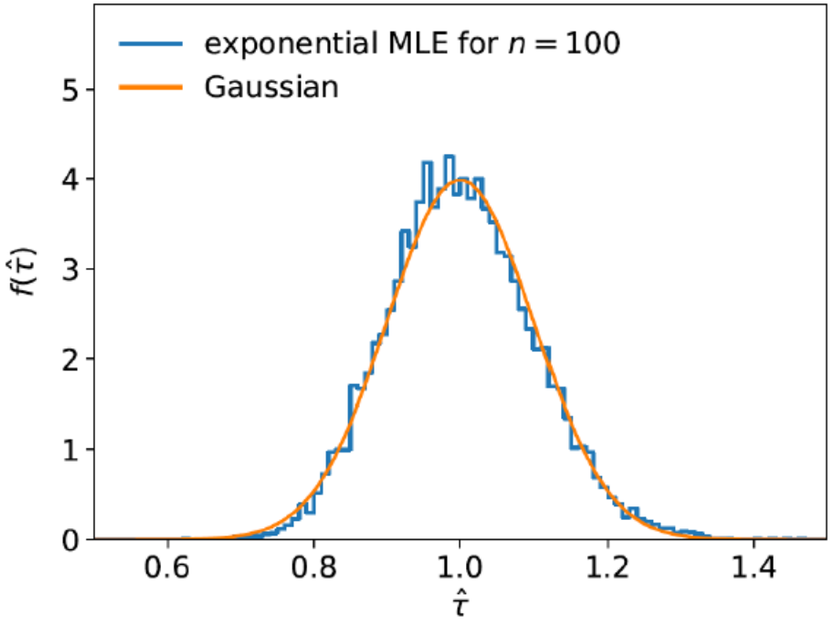
In the large-sample limit, the pdf of the MLE becomes Gaussian,

$$f(\hat{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\theta}}} e^{-(\hat{\theta}-\theta)^2/2\sigma^2_{\hat{\theta}}}$$

where $\sigma^2_{\hat{\theta}}$ is the minimum variance bound (note bias is zero).

For example, exponential MLE with sample size $n = 100$.

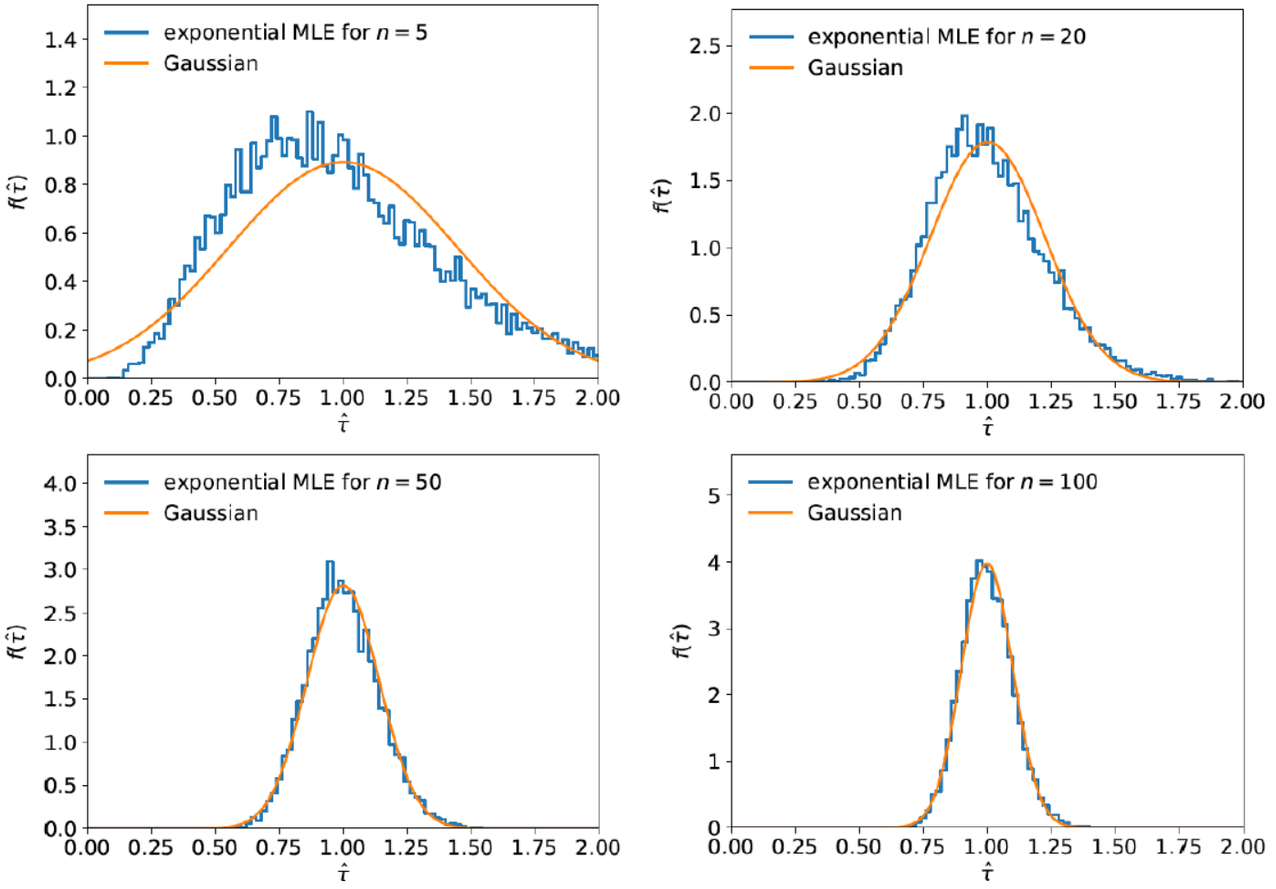
Note that for exponential, MLE is arithmetic average, so Gaussian MLE seen to stem from Central Limit Theorem.



G. Cowan / RHUL Physics

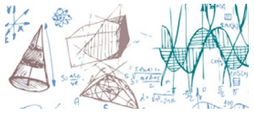
The MLE's distribution becomes Gaussian

- Distribution of MLE exponential parameter



G. Cowan / RHUL Physics

Variance of ML estimators



- So far our procedure of ML estimation requires to collect a data sample and make a hypothesis for the P.D.F. $f(x, \theta)$
- If we keep repeating experiments we also get different values for the estimated parameters – sampling distribution
- That would be the way to analyse the properties of this S.D.o.E. (sampling distribution of estimator) and evaluate its variance
- One approach, called analytic method, would be to just make an exact computations, for instance the exponential model would give us:

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

$$= \int \cdots \int \left(\frac{1}{n}\sum_i t_i\right)^2 \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \prod_i dt_i$$

$$- \int \cdots \int \left(\frac{1}{n}\sum_i t_i\right) \cdot \frac{1}{\tau} e^{-\frac{t_1}{\tau}} \cdot \frac{1}{\tau} e^{-\frac{t_2}{\tau}} \cdots \frac{1}{\tau} e^{-\frac{t_n}{\tau}} \prod_i dt_i = \frac{\tau^2}{n}$$

Variance of ML estimators –MC method

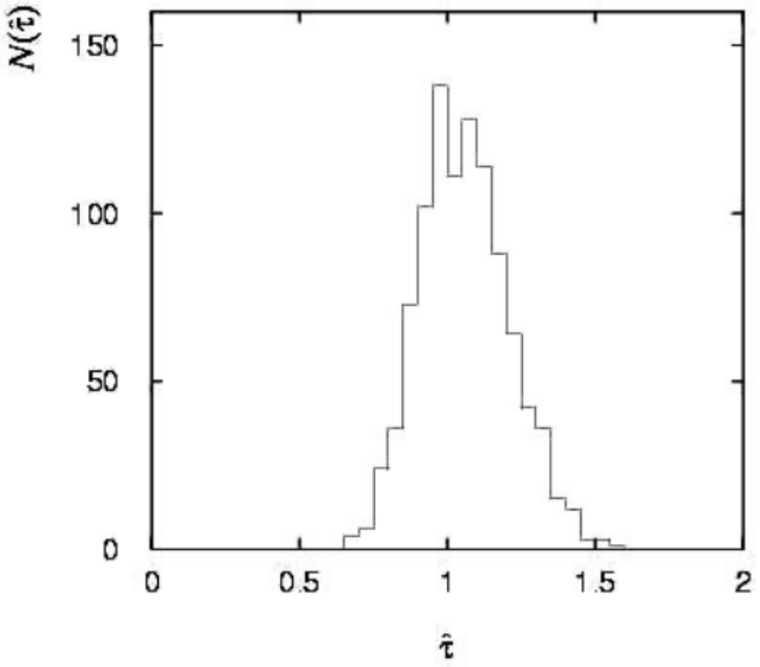
Having estimated our parameter we now need to report its 'statistical error', using e.g. the estimator's standard deviation, or (co)variance.

It is usually not possible to do this with an exact calculation.

Another way is to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

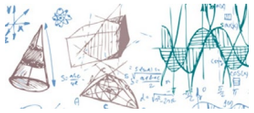
For exponential example (n=50), from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$



G. Cowan / RHUL Physics

Variance of ML estimators



- The result above is universal in a sense, the variance of the sample mean is $1/n$ times smaller than the variance of P.D.F. used to model the variation of t R.V.
- Since our result depends on an unknown parameter (τ) we need to use our estimated value to calculate the variance:

$$\sigma_{\hat{\tau}}^2 = \frac{\tau^2}{n} \rightarrow \sigma_{\hat{\tau}}^2 = \frac{\hat{\tau}^2}{n}, \sigma_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

- Formally, we used the transformation invariance property of the ML estimators
- So, how should one interpret an experimental result like this:

$$\tau = 12.38 \pm 0.72$$

- Here, the ML estimate is 12.38 and the statistical uncertainty means that if the experiment would be repeated many times, the standard deviation of its S.D.o.E. would be 0.72

Variance of ML estimators

- And what if we do not know, even in principle, the P.D.F. (say we are looking for a new phenomena)?
- We could use the following reasoning, let's start with expanding the likelihood function in a Taylor series about the estimate $\hat{\theta}$:

$$\ln\mathcal{L}(\theta) = \ln\mathcal{L}(\hat{\theta}) + \left[\frac{\partial \ln\mathcal{L}}{\partial\theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \left[\frac{\partial^2 \ln\mathcal{L}}{\partial\theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \cdots$$

- We also define:

$$\widehat{\sigma_\theta^2} = \left(\frac{-1}{\frac{\partial^2 \ln\mathcal{L}}{\partial\theta^2}}\right)_{\theta=\hat{\theta}}$$

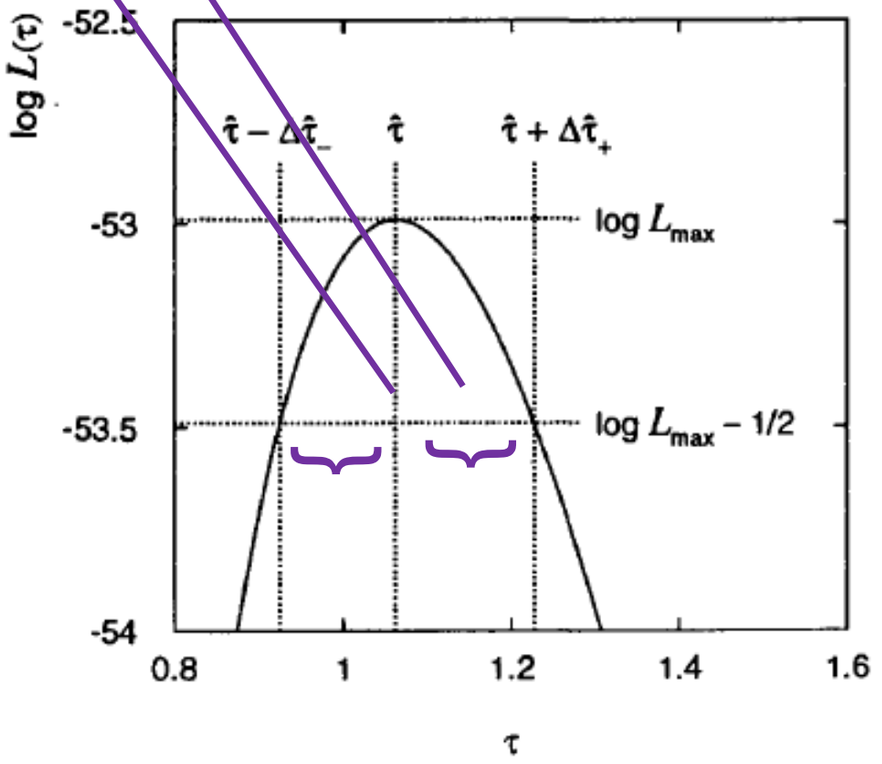
- Since, the first derivative by definitione should be zero for the estimate and $\ln\mathcal{L}(\hat{\theta}) = \ln\mathcal{L}_{Max}$

$$\ln\mathcal{L}(\hat{\theta}) = \ln\mathcal{L}_{Max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma_{\hat{\theta}}^2}}$$

Variance of ML estimators – graphical method

- What is the effect of varying this formula by one standard deviation?

$$\ln\mathcal{L}\left(\hat{\theta} \pm \widehat{\sigma_{\hat{\theta}}}\right) = \ln\mathcal{L}_{Max} - \frac{1}{2}$$



In order to obtain $\widehat{\sigma_{\hat{\theta}}}$, change θ away from $\hat{\theta}$ until $\ln\mathcal{L}$ decreases by $\frac{1}{2}$

Variance of ML estimators – graphical method

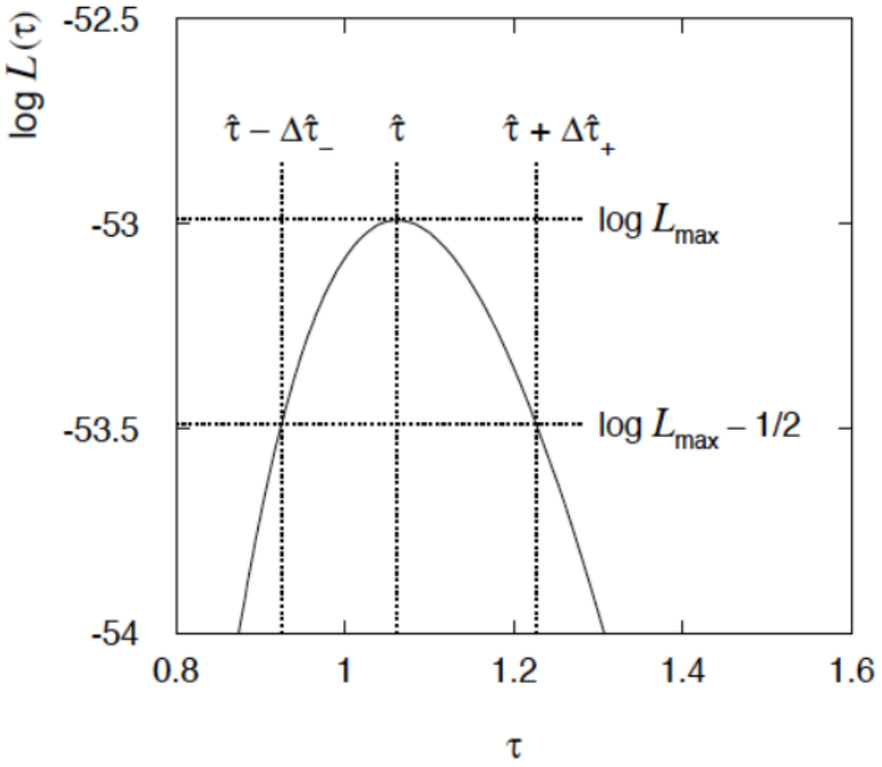
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



Not quite parabolic $\ln L$ since finite sample size $(n = 50)$.