Statistics

or “How to find answers to your questions”

Pietro Vischia[1]

[1] CP3 — IRMP, Université catholique de Louvain

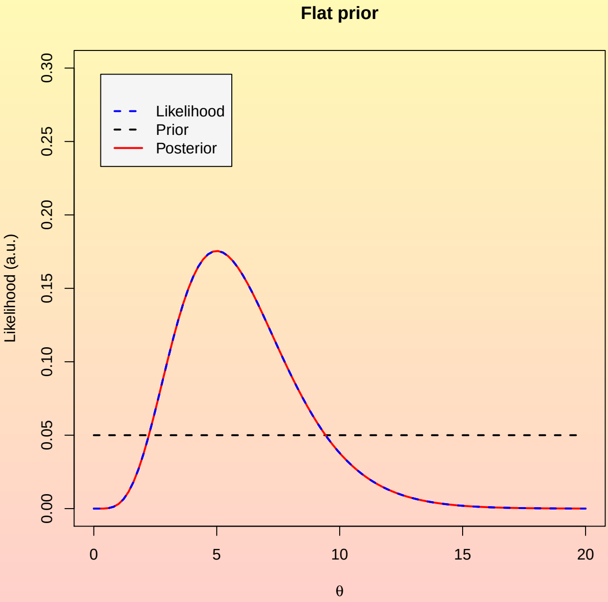
UCLouvain
Institut de recherche
en mathématique et physique

CP3, Lectures on Statistics for HEP

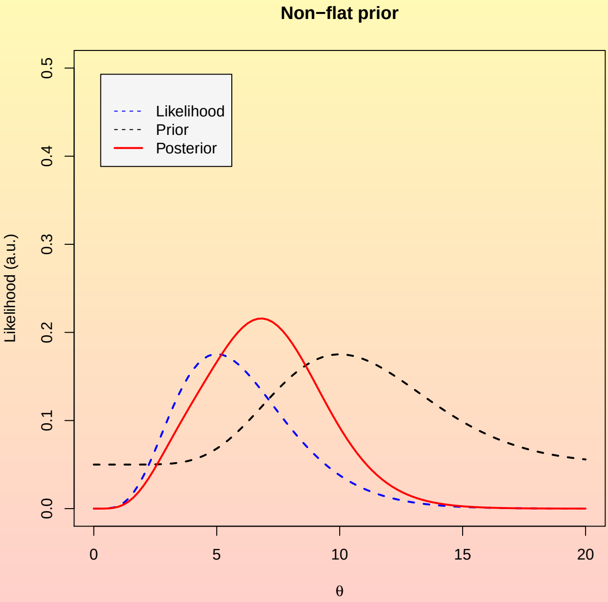
Practicalities

- We start now (10:00), and will stop at 11:45 to give room to a one-hour seminar
  - It has been called at the last minute and there was no other option compatible with the speaker's plans
  - *The quantum enhanced Virgo interferometer* by Dr. Marco Vardaro, abstract at https://agenda.irmp.ucl.ac.be/event/3415/
- As announced yesterday by email, you can choose among yourselves:
  - Restarting at 12:00 until 13:45
  - Restarting at 13:00 until 14:45 (in case you prefer to have lunch at about 12:00)
- Some of you asked for certificate of attendance with explicit mention of the amount of hours (for PhD courses credits)
  - It will be provided on the last day
  - Please let me know (now) if you need it, so I can pass the list to Carinne

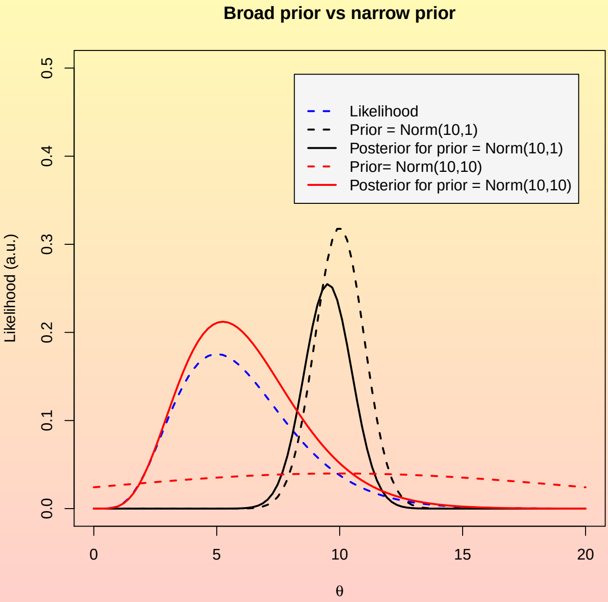
Throwback Tuesday — Flat prior



Throwback Tuesday — Non-flat prior



Throwback Tuesday — Broad and narrow non-flat priors



Estimating a physical quantity

Information, the Fisher way

- The information of a set of observations should increase with the number of observations
  - Double the data should result in double the information if the data are independent
- Information should be conditional on what we want to learn from the experiment
  - Data which are irrelevant to our hypothesis should carry zero information relative to our hypothesis
- Information should be related to precision
  - The greatest the information carried by the data, the better the precision of our result

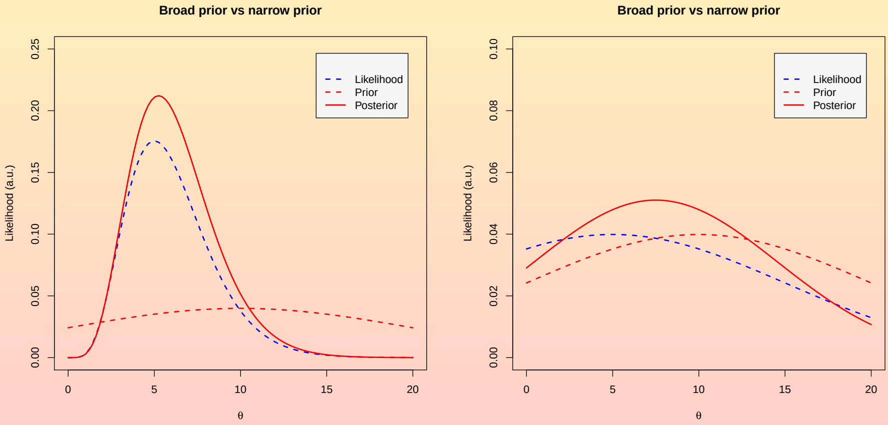
Likelihood and Fisher Information

- The narrowness of the likelihood can be estimated by looking at its curvature
- The curvature is the second derivative with respect to the parameter of interest
- A very narrow (peaked) likelihood is characterized by a very large and positive $-\frac{\partial^2 lnL}{\partial\theta^2}$
- The second derivative of the likelihood is linked to the Fisher Information

$$I(\theta) = -E\left[\frac{\partial^2 lnL}{\partial\theta^2}\right] = E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right]$$

Likelihood and Fisher Information

- A very narrow likelihood will provide much information about θ_{true}
  - The posterior probability will be more localized than the prior in the regimen in which the likelihood function dominates the product $L(\vec{x};\vec{\theta}) \times \pi$
  - The Fisher Information will be large
- A very broad likelihood will not carry much information, and in fact the computed Fisher Information will turn out to be small



Fisher Information and Jeffreys priors

- When changing variable, the change of parameterization must not result in a change of the information
  - The information is a property of the data only, through the likelihood—that summarizes them completely (likelihood principle)
- Search for a parametrization $\theta'(\theta)$ in which the Fisher Information is constant
- Compute the prior as a function of the new variable

$$\pi(\theta) = \pi(\theta')\left|\frac{d\theta'}{d\theta}\right| \propto \sqrt{E\left[\left(\frac{\partial lnN}{\partial\theta'}\right)^2\right]\left|\frac{\partial\theta'}{\partial\theta}\right|}$$

$$= \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta'}\frac{\partial\theta'}{\partial\theta}\right)^2\right]}$$

$$= \sqrt{E\left[\left(\frac{\partial lnL}{\partial\theta}\right)^2\right]}$$

$$= \sqrt{I(\theta)}$$

- For any θ, $\pi(\theta) = \sqrt{I(\theta)}$; with this choice, the information is constant under changes of variable
- Such priors are called Jeffreys priors, and assume different forms depending on the type of parametrization
  - Location parameters: uniform prior
  - Scale parameters: prior $\propto \frac{1}{\theta}$
  - Poisson processes: prior $\propto \frac{1}{\sqrt{\theta}}$

Sufficient statistic and data reduction

- A test statistic is a function of the data (a quantity derived from the data sample)
- A statistic $T = T(\boldsymbol{X})$ is sufficient for θ if the density function $f(\boldsymbol{X}|T)$ is independent of θ
  - If T is a sufficient statistic for θ, then also any strictly monotonic $g(T)$ is sufficient for θ
- The statistic T carries as much information about θ as the original data $\boldsymbol{X}$
  - No other function can give any further information about θ
  - Same inference from data $\boldsymbol{X}$ with model M and from sufficient statistic $T(\boldsymbol{X})$ with model M'

Example: is it sufficient?

- Example: data $1, 2, 3, 4, 5$; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic?

Example: is it sufficient?

- Example: data $1, 2, 3, 4, 5$; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic?
  - Since the sample mean is 3, we also estimate the population mean to be 3
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean

Example: is it sufficient?

- Example: data $1, 2, 3, 4, 5$; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic?
  - Since the sample mean is 3, we also estimate the population mean to be 3
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Binomial test in coin toss
  - Record heads and tails, with their order: $HTTHHHTHHTTTHTHTH$
  - Can we somehow improve by identifying a sufficient statistic?

Example: is it sufficient?

- Example: data $1, 2, 3, 4, 5$; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic?
  - Since the sample mean is 3, we also estimate the population mean to be 3
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Binomial test in coin toss
  - Record heads and tails, with their order: $HTTHHHTHHTTTHTHTH$
  - Can we somehow improve by identifying a sufficient statistic?
  - What happens if we record only the number of heads? (remember that the binomial p.d.f. is: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$, $r = 0, 1, \ldots, N$

Example: is it sufficient?

- Example: data $1, 2, 3, 4, 5$; sample mean (estimate of population mean) $\hat{x} = \frac{1+2+3+4+5}{5} = 3$
  - Imagine we don't have the data; we only know that the sample mean is 3
  - Is the sample mean a sufficient statistic?
  - Since the sample mean is 3, we also estimate the population mean to be 3
  - Knowing the data (the set $1, 2, 3, 4, 5$) or knowing only the sample mean does not improve our estimate for the population mean
- Binomial test in coin toss
  - Record heads and tails, with their order: *HTTHHHTHHTTTHTHTH*
  - Can we somehow improve by identifying a sufficient statistic?
  - What happens if we record only the number of heads? (remember that the binomial p.d.f. is: $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$, $r = 0, 1, \ldots, N$
  - Recording only the number of heads (no tails, no order) gives exactly the same information
  - Data can be reduced; we only need to store a sufficient statistic
  - Storage needs are reduced

The Likelihood Principle

- Common enunciation: given a set of observed data $\vec{x}$, the likelihood function $L(\vec{x};\theta)$ contains all the information relevant to the measurement of θ contained in the data sample
  - The likelihood function is seen as a function of θ, for a fixed set (a particular realization) of observed data $\vec{x}$
  - As we have seen, the likelihood is used to define the information contained in a sample
- Bayesian statistics normally complies, frequentist statistics usually does not, because a frequentist has to consider the hypothetical set of data that might have been obtained.
- This on one side implies that a frequentist always needs multiple sets of observations
  - Even in forecasts: computer simulations of the day of tomorrow, or counting the past frequency of correct forecasts by the grandpa feeling arthritis in the shoulder
- On the other side a Bayesian would say “Probably tomorrow will rain”, a frequentist “the sentence -tomorrow it will rain- is probably true”

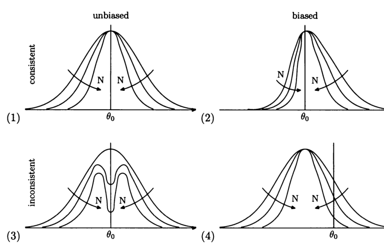
Estimators

- Set $\vec{x} = (x_1, ..., x_N)$ of N statistically independent observations x_i, sampled from a p.d.f. $f(x)$.
- Mean and width of $f(x)$ (or some parameter of it: $f(x;\vec{\theta})$, with $\vec{\theta} = (\theta_1, ..., \theta_M)$ unknown)
  - In case of a linear p.d.f., the vector of parameters would be $\vec{\theta} = (intercept, slope)$
- We call estimator a function of the observed data $\vec{x}$ which returns numerical values $\hat{\vec{\theta}}$ for the vector $\vec{\theta}$.

- $\hat{\vec{\theta}}$ is (asymptotically) consistent if it converges to $\vec{\theta}_{true}$ for large N:

$$\lim_{N\to\infty} \hat{\vec{\theta}} = \vec{\theta}_{true}$$

- $\hat{\vec{\theta}}$ is unbiased if its bias is zero, $\vec{b} = 0$
  - Bias of $\hat{\vec{\theta}}$: $\vec{b} := E[\hat{\vec{\theta}}] - \vec{\theta}_{true}$
  - If bias is known, can redefine $\hat{\vec{\theta}}' = \hat{\vec{\theta}} - \vec{b}$, resulting in $\vec{b}' = 0$.



Plot from James, 2nd ed.

- $\hat{\vec{\theta}}$ is efficient if its variance $V[\hat{\vec{\theta}}]$ is the smallest possible
- An estimator is robust when it is insensitive to small deviations from the underlying distribution (p.d.f.) assumed (ideally, one would want distribution-free estimates, without assumptions on the underlying p.d.f.)

The Maximum Likelihood Method 1/

- Let $\vec{x} = (x_1, ..., x_N)$ be a set of N statistically independent observations x_i, sampled from a p.d.f. $f(x; \vec{\theta})$ depending on a vector of parameters
- Under independence of the observations, the likelihood function factorizes to the individual p.d.f. s

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{N} f(x_i, \vec{\theta})$$

- The maximum-likelihood estimator is the $\vec{\theta}_{ML}$ which maximizes the joint likelihood

$$\vec{\theta}_{ML} := argmax_{\theta}\left(L(\vec{x}, \vec{\theta})\right)$$

  - The maximum must be global
- Numerically, it's usually easier to minimize

$$-lnL(\vec{x}; \vec{\theta}) = -\sum_{i=1}^{N} lnf(x_i, \vec{\theta})$$

  - Easier working with sums than with products
  - Easier minimizing than maximizing
- If the minimum is far from the range of permitted values for $\vec{\theta}$, then the minimization can be performed by finding solutions to

$$-\frac{lnL(\vec{x}; \vec{\theta})}{\partial\theta_j} = 0$$

  - It is assumed that the p.d.f. s are correctly normalized, i.e. that $\int f(\vec{x}; \vec{\theta})dx = 1$ ($\rightarrow$ integral does not depend on $\vec{\theta}$)

- Solutions to the likelihood minimization are found via numerical methods such as MINOS
  - Fred James' Minuit: https://root.cern.ch/root/htmldoc/guides/minuit2/Minuit2.html
- $\vec{\theta}_{ML}$ is an estimator $\rightarrow$ let's study its properties!
  1. **Consistent:** $\lim_{N\rightarrow\infty} \vec{\theta}_{ML} = \vec{\theta}_{true}$;
  2. **Unbiased:** only asymptotically. $\vec{b} \propto \frac{1}{N}$, so $\vec{b} = 0$ only for $N \rightarrow \infty$;
  3. **Efficient:** $V[\vec{\theta}_{ML}] = \frac{1}{I(\theta)}$
  4. **Invariant:** for change of variables $\psi = g(\theta)$; $\hat{\psi}_{ML} = g(\vec{\theta}_{ML})$
- $\vec{\theta}_{ML}$ is only asymptotically unbiased, and therefore it does not always represent the best trade-off between bias and variance
- Remember that in frequentist statistics $L(\vec{x};\vec{\theta})$ is not a p.d.f. . In Bayesian statistics, the posterior probability is a p.d.f.:

$$P(\vec{\theta}|\vec{x}) = \frac{L(\vec{x}|\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})d\vec{\theta}}$$

  - Note that if the prior is uniform, $\pi(\vec{\theta}) = k$, then the MLE is also the maximum of the posterior probability, $\vec{\theta}_{ML} = maxP(\vec{\theta}|\vec{x})$.

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N; \tau) = \prod_i f(t_i; \tau)$$

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N;\tau) = \prod_i f(t_i;\tau)$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i;\tau)$

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau}e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N;\tau) = \prod_i f(t_i;\tau)$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i;\tau)$
- Now all you need to do is to maximize the likelihood

Nuclear Decay with Maximum Likelihood Method

- A nuclear decay with half-life τ is described by the p.d.f., expected value, and variance

$$f(t;\tau) = \frac{1}{\tau}e^{-\frac{t}{\tau}}$$
$$E[f] = \tau$$
$$V[f] = \tau^2$$

- Sampling N independent measurements t_i from the same p.d.f. results in a set of measurements identically distributed
- Exercise: compute the MLE for this p.d.f.
- The joint p.d.f. can be factorized

$$f(t_1, ...t_N;\tau) = \prod_i f(t_i;\tau)$$

- For a particular set of N measurements t_i, the p.d.f. can be written as a function of τ only, $L(\tau) := f(t_i;\tau)$
- Now all you need to do is to maximize the likelihood
- The logarithm of the likelihood, $lnL(\tau) = \sum \left(ln\frac{1}{\tau} - \frac{t_i}{\tau} \right)$, can be maximized analytically

$$\frac{\partial lnL(\tau)}{\partial\tau} = \sum_i \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) \equiv 0$$

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient?

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient?
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient?
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table!

	Consistente	Insesgado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+...+t_N}{N}$			
$\hat{\tau} = \frac{t_1+...+t_N}{N-1}$			
$\hat{\tau} = t_i$			

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient?
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table!

	Consistente	Insesgado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+\ldots+t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1+\ldots+t_N}{N-1}$			
$\hat{\tau} = t_i$			

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N}\sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient?
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N}\sum_i t_i\right] = \frac{1}{N^2}\sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table!

	Consistente	Insesgado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+...+t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1+...+t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$			

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

Nuclear Decay with Maximum Likelihood Method

- The maximum-likelihood estimator is

$$\hat{\tau}(t_1, ..., t_N) = \frac{1}{N} \sum_i t_i$$

- It's the simple arithmetical mean of the individual measurements!
- What's the expected value? Is the estimator unbiased?
- The expected value is $E[\hat{\tau}] = \tau$, and the estimator is unbiased:

$$b = E[\hat{\tau}] - E[f] = \tau - \tau = 0$$

- What is the variance? Which is its relationship to N? Is the estimator efficient?
- The variance interestingly decreases when N increases, and it is possible to demonstrate that the estimator is efficient

$$V[\hat{\tau}] = V\left[\frac{1}{N} \sum_i t_i\right] = \frac{1}{N^2} \sum_i V[t_i] = \frac{\tau^2}{N}$$

- The MLE is not the only estimator we can think of. Fill the table!

	Consistente	Insesgado	Eficiente
$\hat{\tau} = \hat{\tau}_{ML} = \frac{t_1+...+t_N}{N}$	✓	✓	✓
$\hat{\tau} = \frac{t_1+...+t_N}{N-1}$	✓	✗	✗
$\hat{\tau} = t_i$	✗	✓	✗

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

Why $\hat{\tau} = t_i$ is unbiased

- Bias: $b = E[\hat{\tau}] - \tau$
  - Note: if you don't know the true value, you must simulate the bias of the method
  - Generate toys with known parameters, and check what is the estimate of the parameter for the toy data
  - If there is a bias, correct for it to obtain an unbiased estimator
- t_i is an individual observation, which is still sampled from the original factorized p.d.f. $f(t_i;\tau) = \frac{1}{\tau} e^{-\frac{t_i}{\tau}}$
- The expected value of t_i is therefore still $E[\hat{\tau}] = E[t_i] = \tau$
- $\hat{\tau} = t_i$ is therefore unbiased!

	Consistente	Insesgado	Eficiente
$\hat{\tau} = t_i$	✗	✓	✗

Table: Propiedades de diferentes estimadores de la vida media de un decaimiento nuclear.

UCLouvain
Institut de recherche
en mathématique et physique

- We usually want to optimize both bias $\vec{b}$ and variance $V[\hat{\vec{\theta}}]$
- While we can optimize each one separately, optimizing them simultaneously leads to none being optimally optimized, in genreal
  - Optimal solutions in two dimensions are often suboptimal with respect to the optimization of just one of the two properties
- The variance is linked to the width of the likelihood function, which naturally leads to linking it to the curvature of $L(\vec{x}; \vec{\theta})$ near the maximum
- However, the curvature of $L(\vec{x}; \vec{\theta})$ near the maximum is linked to the Fisher information, as we have seen
- Information is therefore a limiting factor for the variance (no data set contains infinite information, variance cannot collapse to zero)
- Variance of an estimator satisfies the Rao-Cramér-Frechet (RCF) bound

$$V[\hat{\theta}] \geq \frac{1}{\hat{\theta}}$$

Information Inequality – 1

- Rao-Cramer-Frechet (RCF) bound
  $V[\hat{\theta}] \geq \frac{(1+\partial b/\partial\theta)^2}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
  - In multiple dimensions, this is linked with the Fisher Information Matrix:
    $I_{ij} = E\left[\partial^2 lnL/\partial\theta_i\partial\theta_j\right]$
- Approximations
  - Neglect the bias ($b = 0$)
  - Inequality is an approximate equality (true for large data samples)
- $V[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]}$
- Estimate of the variance of the estimate of the parameter!
- $\hat{V}[\hat{\theta}] \simeq \frac{1}{-E\left[\partial^2 lnL/\partial\theta^2\right]|_{\theta=th\hat{e}ta}}$

Bias-variance tradeoff and optimal variance 2/

UCLouvain
Institut de recherche
en mathématique et physique

- For multidimensional parameters, we can build the information matrix with elements:

$$I_{jk}(\vec{\theta}) = -E\Big[\sum_{i}^{N} \frac{\partial^2 ln f(x_i;\vec{\theta})}{\partial\theta_k \partial\theta_k}\Big]$$
$$= N \int \frac{1}{f}\frac{\partial f}{\partial\theta_j}\frac{\partial f}{\partial\theta_k} dx$$

- (the last equality is due to the integration interval not being dependent on $\vec{\theta}$)

Estimating variance non-analytically

- We have calculated the variance of the MLE in the simple case of the nuclear decay
- Analytic calculation of the variance is not always possible
- Write the variance approximately as:

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

- This expression is valid for any estimator, but if applied to the MLE then we can note $\vec{\theta}_{ML}$ is efficient and asymptotically unbiased
- Therefore, when $N \rightarrow \infty$ then $b = 0$ and the variance approximate to the RCF bound, and $\geq$ becomes $\simeq$:

$$V[\vec{\theta}_{ML}] \simeq \frac{1}{-E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right]\Big|_{\theta=\vec{\theta}_{ML}}}$$

How to extract an interval from the likelihood function 1/

- For a Gaussian p.d.f., $f(x;\vec{\theta}) = N(\mu,\sigma)$, the likelihood can be written as:

$$L(\vec{x};\vec{\theta}) = ln\Big[-\frac{(\vec{x}-\vec{\theta})^2}{2\sigma^2}\Big]$$

- Moving away from the maximum of $L(\vec{x};\vec{\theta})$ by one unit of σ, the likelihood assumes the value $\frac{1}{2}$, and the area enclosed in $[\vec{\theta}-\sigma,\vec{\theta}+\sigma]$ will be—because of the properties of the Normal distribution—equal to 68.3%.

How to extract an interval from the likelihood function 2/

UCLouvain
Institut de recherche
en mathématique et physique

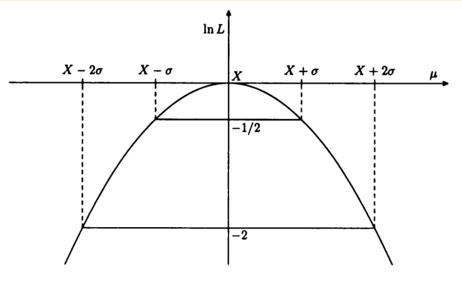
- We can therefore write

$$P\left((\vec{x}-\vec{\theta})^2 \leq \sigma)\right) = 68.3\%$$

$$P(-\sigma \leq \vec{x}-\vec{\theta} \leq \sigma) = 68.3\%$$

$$P(\vec{x}-\sigma \leq \vec{\theta} \leq \vec{x}+\sigma) = 68.3\%$$

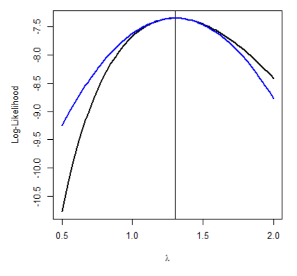
- Taking into account that it is important to keep in mind that probability is a property of sets, in frequentist statistics
  - Confidence interval: interval with a fixed probability content
- This process for computing a confidence interval is exact for a Gaussian p.d.f.
  - Pathological cases reviewed later on (confidence belts and Neyman construction)
- Practical prescription:
  - Point estimate by computing the Maximum Likelihood Estimate
  - Confidence interval by taking the range delimited by the crossings of the likelihood function with $\frac{1}{2}$ (for 68.3% probability content, or 2 for 95% probability content—2σ, etc)

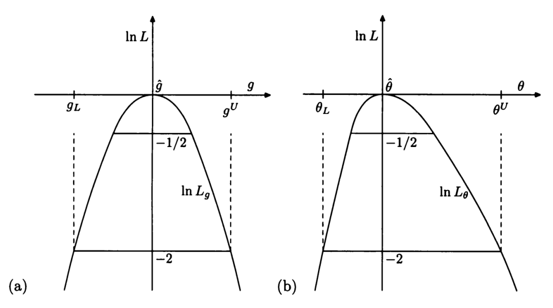


Plot from James, 2nd ed.

How to extract an interval from the likelihood function 3/

- MLE is invariant for monotonic transformations of θ
  - This applies not only to the maximum of the likelihood, but to all relative values
  - The likelihood ratio is therefore an invariant quantity (we'll use it for hypothesis testing)
  - Can transform the likelihood such that $log(L(\vec{x};\vec{\theta}))$ is parabolic, but not necessary (MINOS/Minuit)
- When the p.d.f. is not normal, either assume it is, and use symmetric intervals from Gaussian tails...
  - This yields symmetric approximate intervals
  - The approximation is often good even for small amounts of data
- ...or use asymmetric intervals by just looking at the crossing of the $log(L(\vec{x};\vec{\theta}))$ values
  - Naturally-arising asymmetrical intervals
  - No gaussian approximation
- In any case (even asymmetric intervals) still based on asymptotic expansion
  - Method is exact only to $\mathcal{O}(\frac{1}{N})$





Plot from James, 2nd ed.

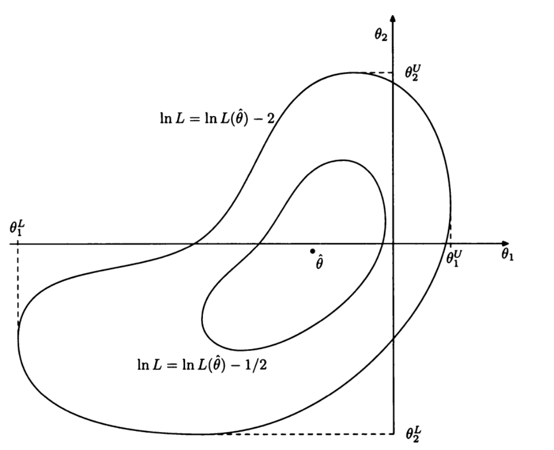
And in many dimensions...

- Construct $log\mathcal{L}$ contours and determine confidence intervals by MINOS
- Elliptical contours correspond to gaussian Likelihoods
  - The closer to MLE, the more elliptical the contours, even in non-linear problems
  - All models are linear in a sufficiently small region
- Nonlinear regions not problematic (no parabolic transformation of $log\mathcal{L}$ needed)
  - MINOS accounts for non-linearities by following the likelihood contour

- Confidence intervals for each parameter

$$\max_{\theta_j, j\neq i} log\mathcal{L}(\theta) = log\mathcal{L}(\hat{\theta}) - \lambda$$

- $\lambda = \frac{Z^2_{1-\beta}}{2}$
  - $\lambda = 1/2$ for $\beta = 0.683$ ("1σ")
  - $\lambda = 2$ for $\beta = 0.955$ ("2σ")



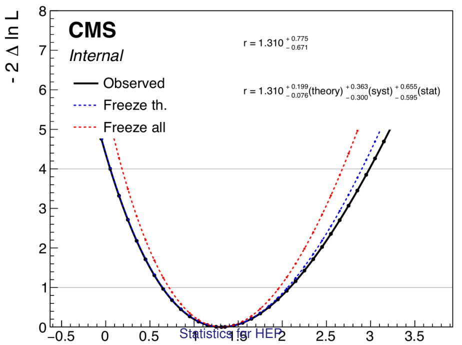
Plot from James, 2nd ed.

What if I have systematic uncertainties? /1

- Parametrize them into the likelihood function; conventional separation of parameters in two classes
  - the Parameter(s) of Interest (POI), often representing σ/σ_{SM} and denoted as μ (*signal strength*)
  - the parameters representing uncertainties, *nuisance parameters* θ
- H_0: $\mu = 0$ (Standard Model only, no Higgs)
- H_1: $\mu = 1$ (Standard Model + Standard Model Higgs)
- Find the maximum likelihood estimates (MLEs) $\hat{\mu}$, $\hat{\theta}$
- Find the conditional MLE $\hat{\hat{\theta}}(\mu)$, i.e. the value of θ maximizing the likelihood function for each fixed value of μ

What if I have systematic uncertainties? /2

- Write the test statistics as $\lambda(\mu) = \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}$
  - Independent on the nuisance parameters (profiled, i.e. their MLE has been taken as a function of each value of μ)
  - Can even freeze them one by one to extract their contribution to the total uncertainty
- Conceptually, you can run the experiment many times (e.g. toys) and record the value of the test statistic
- The test statistic can therefore be seen as a distribution
- Asymptotically, $\lambda(\mu) \sim exp\Big[-\frac{1}{2}\chi^2\Big]\Big(1+\mathcal{O}(\frac{1}{\sqrt{N}})\Big)$ (Wilks Theorem, under some regularity conditions—continuity of the likelihood and up to 2nd derivatives, existence of a maximum, etc)
  - The χ^2 distribution depends only on a single parameter, the number of degrees of freedom
  - It follows that the test statistic is independent of the values of the nuisance parameters
  - Useful: you don't need to make toys in order to find out how is $\lambda(\mu)$ distributed!



How to extract an interval from the likelihood function

- Theorem: for any p.d.f. $f(x|\vec{\theta})$, in the large numbers limit $N \to \infty$, the likelihood can always be approximated with a gaussian:

$$L(\vec{x};\vec{\theta}) \propto_{N\to\infty} e^{-\frac{1}{2}(\vec{\theta}-\vec{\theta}_{ML})^T H(\vec{\theta}-\vec{\theta}_{ML})}$$

- where H is the information matrix $I(\vec{\theta})$.
- Under these conditions, $V[\vec{\theta}_{ML}] \to \frac{1}{I(\vec{\theta}_{ML})}$, and the intervals can be computed as:

$$\Delta lnL := lnL(\theta') - lnL_{max} = -\frac{1}{2}$$

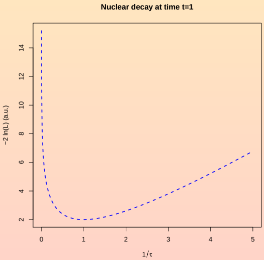
- The resulting interval has in general a larger probability content than the one for a gaussian p.d.f., but the approximation grows better when N increases
  - The interval overcovers the true value $\vec{\theta}_{true}$

How to extract an interval from the likelihood function—interpretation

- $\vec{\theta}_{true}$ is therefore stimated as $\hat{\theta} = \vec{\theta}_{ML} \pm \sigma$. This is another situation in which frequentist and Bayesian statistics differ in the interpretation of the numerical result
- Frequentist: $\vec{\theta}_{true}$ is fixed
  - “if I repeat the experiment many times, computing each time a confidence interval around $\vec{\theta}_{ML}$, on average 68.3% of those intervals will contain $\vec{\theta}_{true}$”
  - Coverage: “the interval covers the true value with 68.3% probability”
  - Direct consequence of the probability being a property of <u>data sets</u>
- Bayesian: $\vec{\theta}_{true}$ is not fixed
  - “the true value $\vec{\theta}_{true}$ will be in the range $[\vec{\theta}_{ML} - \sigma, \vec{\theta}_{ML} + \sigma]$ with a probabilty of 68.3%”
  - This corresponds to giving a value for the posterior probability of the parameter $\vec{\theta}_{true}$

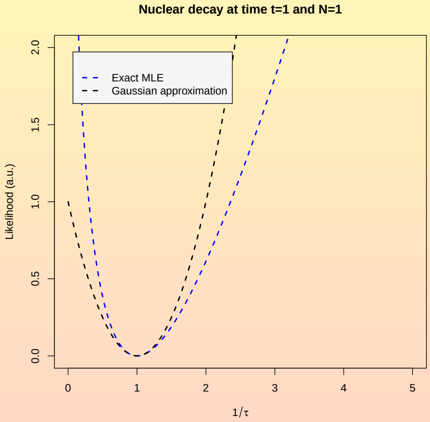
Non-normal likelihoods and Gaussian approximation — 1

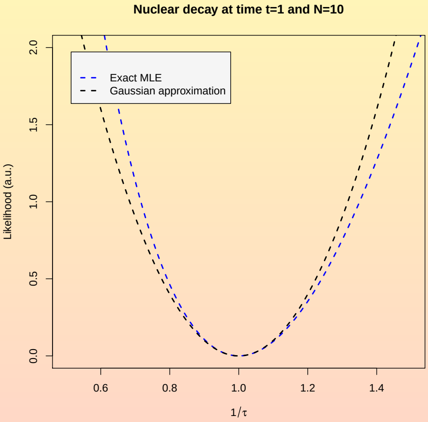
- How good is the approximation $L(\vec{x};\vec{\theta} \propto exp\Big[-\frac{1}{2}(\vec{\theta}-\vec{\theta}_{MLE})^T H(\vec{\theta}-\vec{\theta}_{ML})\Big]$?
  - Here H is the information matrix $I(\vec{\theta})$
  - True only to $\mathcal{O}(\frac{1}{N})$
  - In these conditions, $V[\vec{\theta}_{ML}] \to \frac{1}{I(\vec{\theta}_{ML})}$
  - Intervals can be derived by crossings: $\Delta lnL = lnL(\theta') - lnL_{max} = k$
- Convince yourselves of how good is this approximation in case of the nuclear decay (simplified case of N measurements in which $t_i = 1$)!
  wget https://raw.githubusercontent.com/vischia/statex/master/nuclearDecay.R



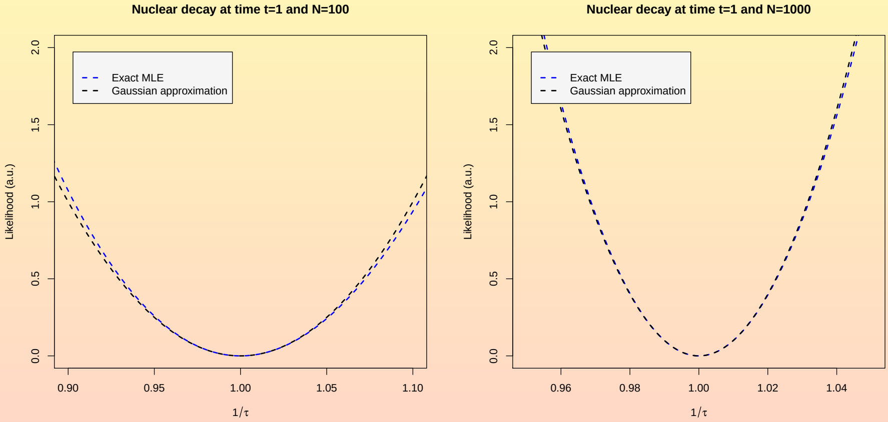
Non-normal likelihoods and Gaussian approximation — 2

UCLouvain
Institut de recherche
en mathématique et physique





Non-normal likelihoods and Gaussian approximation — 3



- The convergence of the likelihood $L(\vec{x};\vec{\theta})$ to a gaussian is a direct consequence of the central limit theorem
- Take a set of measurements $\vec{x} = (x_i, ..., x_N)$ affected by experimental errors that results in uncertainties $\sigma_1, ..., \sigma_N$ (not necessarily equal among each other)
- In the limit of a large number of events, $M \to \infty$, the random variable built summing M measurements is gaussian-distributed:

$$Q := \sum_{j=1}^{M} x_j \sim N\Big(\sum_{j=1}^{M} x_j, \sum_{j=1}^{M} \sigma_j^2\Big), \qquad \forall\, f(x,\vec{\theta})$$

- The demonstration runs by expanding in series the characteristic function $y_i = \frac{x_j - \mu_j}{\sqrt{\sigma_j}}$
- The theorem is valid for any p.d.f. $f(x,\vec{\theta})$ that is reasonably peaked around its expected value.
  - If the p.d.f. has large tails, the bigger contributions from values sampled from the tails will have a large weight in the sum, and the distribution of Q will have non-gaussian tails
  - The consequence is an alteration of the probability of having sums Q outside of the gaussian

Asymptoticity of the Central limit theorem

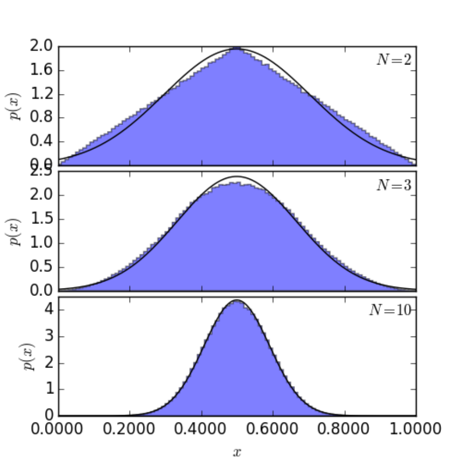
- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions.
- How large does M need to be for the approximation to be reasonably good?

Asymptoticity of the Central limit theorem

- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions.
- How large does M need to be for the approximation to be reasonably good?
- Download the file and check!
  wget https://raw.githubusercontent.com/vischia/statex/master/centralllimit.py

Asymptoticity of the Central limit theorem

- The condition $M \to \infty$ is reasonably valid if the sum is of many small contributions.
- How large does M need to be for the approximation to be reasonably good?
- Download the file and check!
  wget https://raw.githubusercontent.com/vischia/statex/master/centrallimit.py
- Not much!



UCLouvain
Institut de recherche
en mathématique et physique

- Measure N times the same quantity: values x_i and uncertainties σ_i. MLE and variance are:

$$\hat{x}_{ML} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

$$\frac{1}{\hat{\sigma}_x^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

- The MLE is obtained when each measurement is weighted by its own variance
  - This is because the variance is essentially an estimate of how much information lies in each measurement
- This works if the p.d.f. is known
  - Compare this method with an alternative one that does not assume knowledge of the p.d.f.
  - The second method will be the only one applicable to cases in which the p.d.f. is unknown

Combination of measurements: alternative method 1/

- Take a set of measures sampled from an unknown p.d.f. $f(\vec{x}, \vec{\theta})$
- Compute the expected value and variance of a combination of such measurements described by a function $g(\vec{x})$.
- The expected value and variance of x_i are elementary:

$$\mu = E[x] V_{ij} = E[x_i x_j] - \mu_i \mu_j$$

- If we want to extract the p.d.f. of $g(\vec{x})$, we would normally use the jacobian of the transformation of f to g, but in this case we assumed $f(\vec{x})$ is unknown.

UCLouvain
Institut de recherche
en mathématique et physique

- We don't know f, but we can still write an expansion in series for it:

$$g(\vec{x}) \simeq g(\vec{\mu}) + \sum_{i=1}^{N} \left(\frac{\partial g}{\partial x_i} \right)\Big|_{x=\mu} (x_i - \mu_i)$$

- We can compute the expected value and variance of g by using the expansion:

$$E\big[g(\vec{x})\big] \simeq g(\mu), \qquad (E[x_i - \mu_i] = 0)$$

$$\sigma_g^2 = \sum_{ij=1}^{N} \left[\frac{\partial g}{\partial x_i} \frac{\partial g}{\partial x_j} \right]\Big|_{\vec{x}=\vec{\mu}} V_{ij}$$

- The variances are propagated to g by means of their jacobian!
- For a sum of measurements, $y = g(\vec{x}) = x_1 + x_2$, the variance of y is $\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$, which is reduced to the sum of squares for independent measurements

UCLouvain
Institut de recherche
en mathématique et physique

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the time taken for a laser light pulse to go from the Earth to the Moon and back (in units of Earth-to-Moon-Time EMT)
  - On the Moon we have a receiver built by NASA. It's very good but placed in unfavourable conditions, yielding only a 2% precision on Earth-to-Moon
  - On Earth we have a receiver made out of scrap material. It is however placed in favourable conditions, yielding a 5% precisionon Moon-to-Earth

$$N_{EM} = 0.99 \pm 0.02\ EMT$$

$$N_{ME} = 1.05 \pm 0.05\ EMT$$

- Evidently, the time to moon and back is $N_{EME} = N_{EM} + N_{ME}$, and we can apply Eq. 42: Do it!

UCLouvain
Institut de recherche
en mathématique et physique

- Let's compare the two ways of combining measurements, and check the role of the Fisher Information
- Let's estimate the time taken for a laser light pulse to go from the Earth to the Moon and back (in units of Earth-to-Moon-Time EMT)
  - On the Moon we have a receiver built by NASA. It's very good but placed in unfavourable conditions, yielding only a 2% precision on Earth-to-Moon
  - On Earth we have a receiver made out of scrap material. It is however placed in favourable conditions, yielding a 5% precisionon Moon-to-Earth

$$N_{EM} = 0.99 \pm 0.02 \; EMT$$

$$N_{ME} = 1.05 \pm 0.05 \; EMT$$

- Evidently, the time to moon and back is $N_{EME} = N_{EM} + N_{ME}$, and we can apply Eq. 42: Do it!
- Resulting estimate:
  - $N_{EME} = 0.99 + 1.05 \pm \sqrt{0.02^2 + 0.05^2} \; EMT = 2.05 \pm 0.05 \; EMT$, corresponding to a precision of $\frac{\sigma_{N_{EME}}}{N_{EME}} \sim 2.4\%$.

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information?

Combination of measurements: example 2/

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information?
- We can use this additional information to note that the two estimates N_{EM} and N_{ME} are independent estimates of the same physical quantity $\frac{N_{EME}}{2}$
- Compute N_{EME} and $\sigma(N_{EME})$based on this reasonment

UCLouvain
Institut de recherche
en mathématique et physique

- We now however can argue that over the time it takes for light to go to the Moon and back any environment condition would be roughly constant
- How can we exploit this additional information?
- We can use this additional information to note that the two estimates N_{EM} and N_{ME} are independent estimates of the same physical quantity $\frac{N_{EME}}{2}$
- Compute N_{EME} and $\sigma(N_{EME})$based on this reasonment
- We can therefore use Eq. 40 to compute $\frac{N_{EME}}{2}$ and multiply the result by 2, obtaining

$$N_{EME} = 2.00 \pm 0.03\ EMT$$

- This estimate corresponds to a precision of only 1.5%!!!
- The dramatic improvement in the precision of the measurement, from 2.4% to 1.5%, is a direct consequence of having used additional information under the form of a relationship (constraint) between the two available measurements.
- A good physicist exploits as many constraints as possible in order to improve the precision of a measurement
  - Sometimes the contraints are arbitrary or correspond to special cases
  - Is is very important to explicitly mention any constraint used to derive a measurement, when quoting the result.

What about asymmetric uncertainties?

UCLouvain
Institut de recherche
en mathématique et physique

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate N_{EM} and N_{ME}
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example, $N_{EMT} = 2.09^{+0.06}_{-0.03}$

What about asymmetric uncertainties?

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate N_{EM} and N_{ME}
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example, $N_{EMT} = 2.09^{+0.06}_{-0.03}$
- No!
- Why?

What about asymmetric uncertainties?

- Now suppose my receivers operate by taking data and performing a maximum likelihood fit to estimate N_{EM} and N_{ME}
- Can I combine these two measurements with the two methods seen above?
  - $N_{EM} = 0.99 \pm 0.03$
  - $N_{ME} = 1.10^{+0.05}_{-0.01}$
- For example, $N_{EMT} = 2.09^{+0.06}_{-0.03}$
- No!
- Why?
- The naïve quadrature of the two uncertainties is wrong!
  - The naïve combination is an expression of the Central Limit Theorem
  - The resulting combination is expected to be more symmetric than the measurements it originates from
  - Symmetric uncertainties usually assume a Gaussian approximation of the likelihood
  - Asymmetric uncertainties? One would need a study of the non-linearity (large biases might be introduced if ignoring this)
- Intrinsic difference between averaging and most probable value
  - Averaging results in average value and variance that propagate linearly
  - Taking the mode (essentially what MLE does) does not add up linearly!
- With asymmetric uncertainties from MLE fits, always combine the likelihoods (better in an individual simultaneous fit)

- Throwback Tuesday: what happens to the posterior for different broadness/narrowness of the likelihood and the prior
- Information and estimates of physical parameters
- Sufficient statistic
- The Likelihood Principle
- The Maximum Likelihood Method
- Uncertainties: how to get them from the likelihood
- Combining measurements: use all the available information

Non-exhaustive list of references

- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- Pearl, Judea: Causal inference etc etc, a Primer (add full details)
- R.J.Barlow: A Guide to the Use of Statistical Methods in the Physical Sciences - Wiley
- Kyle Cranmer: Lessons at HCP Summer School 2015
- Kyle Cranmer: Practical Statistics for the LHC - http://arxiv.org/abs/1503.07622
- Harrison Prosper: Practical Statistics for LHC Physicists - CERN Academic Training Lectures, 2015 https://indico.cern.ch/category/72/

THANKS FOR THE ATTENTION!

Backup