Bayesian inference and Monte Carlo methods

Lecture 2: Parameter inference and likelihood

Susana J. Landau

8 de octubre de 2025

Likelihood

- The likelihood L_i corresponding to a hypothesis H_i, which is associated with a probability density function $f_i(x) = f(x|H_i)$ or a discrete probability distribution $W_i(k) = p(k|H_i)$, after the observation/measurement x or k has been realized, is given by:

$$L_i = L(i|x) = f_i(x) = f(x|H_i)$$

or

$$L_i = L(i|k) = W_i(k) = p(k|H_i)$$

Likelihood

- The likelihood L_i corresponding to a hypothesis H_i, which is associated with a probability density function $f_i(x) = f(x|H_i)$ or a discrete probability distribution $W_i(k) = p(k|H_i)$, after the observation/measurement x or k has been realized, is given by:

$$L_i = L(i|x) = f_i(x) = f(x|H_i)$$

or

$$L_i = L(i|k) = W_i(k) = p(k|H_i)$$

When we replace the hypothesis H_i with a continuous parameter θ or a discrete probability distribution $W(k|\theta)$, the corresponding likelihood is:

$$L(\theta) = L(\theta|x) = f(x|\theta)$$
$$L(\theta) = L(\theta|k) = W(k|\theta)$$

While the likelihood is related to the validity of a hypothesis given an observation, the p.d.f. is related to the probability to observe a variate for a given hypothesis.

- Usually, experiments or observations provide a sample of N independent measurements x_i, which all follow independently the same probability density function $f(x|\theta)$ which depends on the parameter θ (i.i.d variates). The probability density for all the observations can be written as:

$$\bar{f}(x_1, x_2, ..., x_N) = \prod_{i=1}^{N} f(x_i|\theta)$$

For any value of θ, the function $\bar{f}(x_1, ..., x_N)$ evaluated at the observations $x_1, ..., x_N$ is equal to the likelihood $\bar{L}$:

$$\begin{aligned} \bar{L}(\theta) &= \bar{L}(\theta|x_1, ..., x_n) \\ &= \bar{f}(x_1, ..., x_N|\theta) \\ &= \prod_{i=1}^{N} f(x_i|\theta) \end{aligned}$$

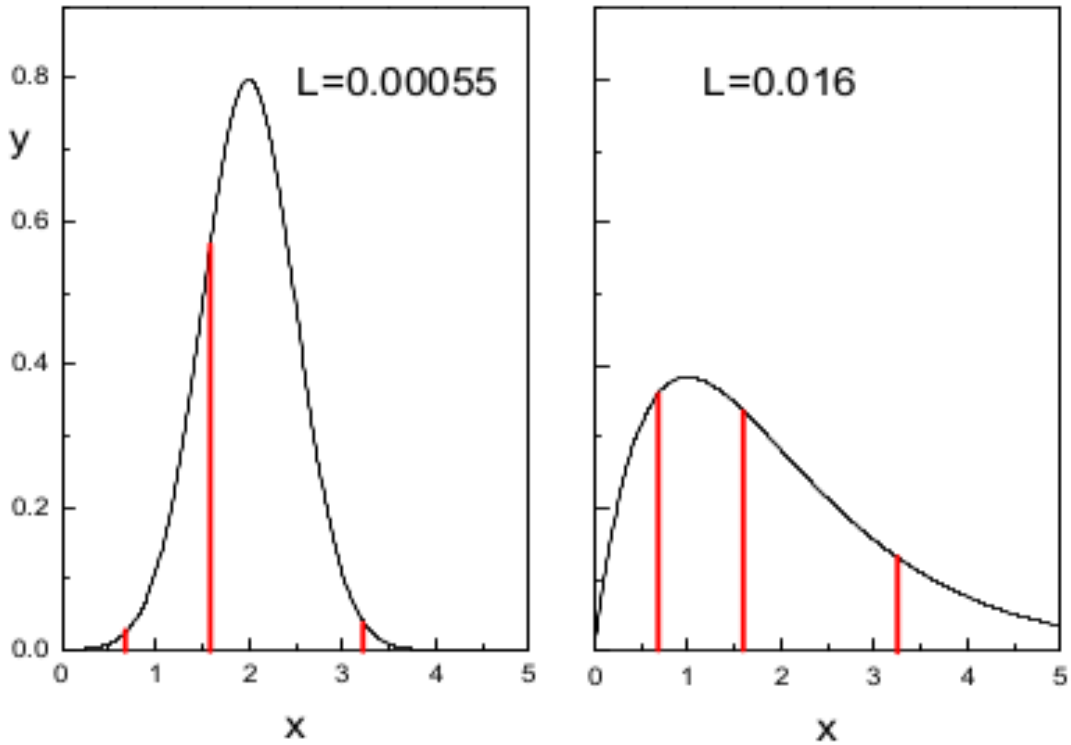


FIGURA: Likelihood of 3 observaciones and 2 hypothesis with different pdfs.

The likelihood is not a probability density function (pdf) of the observations x_i but a function of the parameter θ. It is used to compare different hypotheses or models based on the observed data. It measures how strongly a hypothesis is supported by the data.

Likelihood and Probability

- The likelihood has a close relationship with probability, but requires careful interpretation:
  - Probability: $P(A|B)$ is the probability of event A occurring given that hypothesis B is true.
  - Likelihood: $L(\theta|D)$ is the likelihood of parameters θ given observed data D.
- For fixed data: the likelihood is a function of the parameters (θ), but **is NOT a probability distribution over the parameter space**.
- It doesn't satisfy the normalization requirement: $\int P(D|\theta)d\theta$ is generally not equal to 1.
- It doesn't represent probabilities of the parameters, but rather compatibility of those parameters with the data.

Likelihood of independent experiments

- Consider the case where we have two independent experiments A and B, both measuring the same quantity x. The combined likelihood is the product of the individual likelihoods $L_A(x_1|\theta) = f_A(x_1|\theta)$ and $L_B(x_2|\theta) = f_B(x_2|\theta)$:

$$\begin{aligned} L(\theta) &= f(x_1, x_2|\theta) = f_A(x_1|\theta) f_B(x_2|\theta) \\ &= L_A(x_1|\theta)\, L_B(x_2|\theta) \end{aligned}$$

The likelihood for independent experiments is equal to the product of the likelihoods for each experiment.

Example 1

- To discriminate between different hypotheses, we can use the likelihood ratio:

Example 1

- To discriminate between different hypotheses, we can use the likelihood ratio:
- Five events are observed, and we want to calculate the relative probabilities for three hypotheses. Hypothesis H_1 assumes a Poisson distribution with an expected value of 2, while H_2 and H_3 assume the same distribution but with expected values of 9 and 20, respectively.

Example 1

- To discriminate between different hypotheses, we can use the likelihood ratio:
- Five events are observed, and we want to calculate the relative probabilities for three hypotheses. Hypothesis H_1 assumes a Poisson distribution with an expected value of 2, while H_2 and H_3 assume the same distribution but with expected values of 9 and 20, respectively.

$$\begin{aligned} L_1 &= P_2(5) \sim 0{,}036 \\ L_2 &= P_9(5) \sim 0{,}061 \\ L_3 &= P_{20}(5) \sim 0{,}00005 \end{aligned}$$

Example 1

- To discriminate between different hypotheses, we can use the likelihood ratio:
- Five events are observed, and we want to calculate the relative probabilities for three hypotheses. Hypothesis H_1 assumes a Poisson distribution with an expected value of 2, while H_2 and H_3 assume the same distribution but with expected values of 9 and 20, respectively.

$$\begin{aligned} L_1 &= P_2(5) \sim 0{,}036 \\ L_2 &= P_9(5) \sim 0{,}061 \\ L_3 &= P_{20}(5) \sim 0{,}00005 \end{aligned}$$

- If we are interested in hypothesis H_2, the ratio $L_2/(L_1 + L_2 + L_3) \sim 0{,}63$ is relevant. Now, suppose a second measurement is made in the same time interval and 8 events are observed:

Example 1

- To discriminate between different hypotheses, we can use the likelihood ratio:
- Five events are observed, and we want to calculate the relative probabilities for three hypotheses. Hypothesis H_1 assumes a Poisson distribution with an expected value of 2, while H_2 and H_3 assume the same distribution but with expected values of 9 and 20, respectively.

$$\begin{aligned} L_1 &= P_2(5) \sim 0{,}036 \\ L_2 &= P_9(5) \sim 0{,}061 \\ L_3 &= P_{20}(5) \sim 0{,}00005 \end{aligned}$$

- If we are interested in hypothesis H_2, the ratio $L_2/(L_1 + L_2 + L_3) \sim 0{,}63$ is relevant. Now, suppose a second measurement is made in the same time interval and 8 events are observed:

$$\begin{aligned} L_1 &= P_2(5)P_2(8) \sim 6{,}4 \times 10^{-3} \\ L_2 &= P_9(5)P_9(8) \sim 5{,}1 \times 10^{-2} \\ L_3 &= P_{20}(5)P_{20}(8) \sim 6{,}1 \times 10^{-7} \end{aligned}$$

- The ratio $L_2/(L_1 + L_2 + L_3) \sim 0{,}89$ is much more significant.

Example 2

- We have samples drawn from one of the following normal distributions:

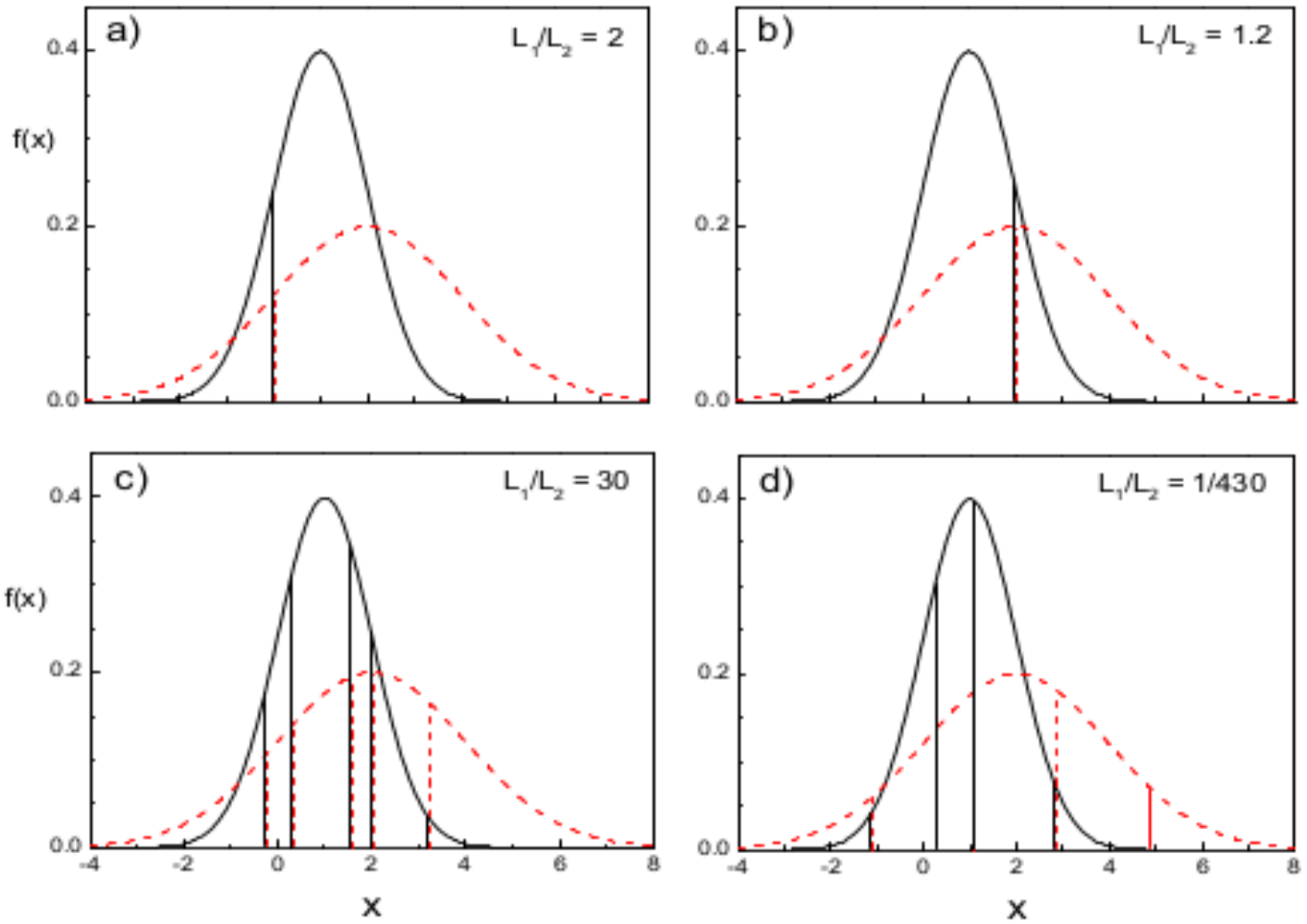
$$f_1 = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}$$

$$f_2 = \frac{1}{\sqrt{2\pi}\,2} e^{-\frac{(x-1)^2}{8}}$$

- We calculate the ratio $\frac{L_1}{L_2}$ for the following cases:
  - Initially, the sample consists of a single observation at $x = 0$.
  - Then, a second observation is made at $x = 2$.
  - Finally, we consider five observations drawn from f1 and five from f2

In both examples discussed, the likelihood allows us to make statements about whether the observed data are consistent with a given hypothesis. We do not make statements about the probability of obtaining a specific value in the measurement.

Example 2



Maximum likelihood estimator (MLE)

- Principle: The likelihood function exhausts all the information contained in the observations related to the parameters.

Maximum likelihood estimator (MLE)

- Principle: The likelihood function exhausts all the information contained in the observations related to the parameters.
- Maximum likelihood principle: The value of the parameters that maximizes the likelihood (MLE) can be chosen.

Maximum likelihood estimator (MLE)

- Principle: The likelihood function exhausts all the information contained in the observations related to the parameters.
- Maximum likelihood principle: The value of the parameters that maximizes the likelihood (MLE) can be chosen.
- If we are interested in a range for the parameters, we select those intervals where the likelihood is higher than outside them.

Maximum likelihood estimator (MLE)

- Principle: The likelihood function exhausts all the information contained in the observations related to the parameters.
- Maximum likelihood principle: The value of the parameters that maximizes the likelihood (MLE) can be chosen.
- If we are interested in a range for the parameters, we select those intervals where the likelihood is higher than outside them.
- Both the MLE and the confidence intervals are invariant under transformations of the parameters. The likelihood is not a probability density function, but a function of the parameters. $L(\theta) = L'(\theta')$ (For example, we obtain the same result for the mass of a particle as for the estimate of the squared mass.)

Likelihood with a single parameter

- Given a sample of N observations $x_1, ..., x_N$ that follow a probability density function $f(x|\theta)$, the likelihood can be written as:

$$L(\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

Likelihood with a single parameter

- Given a sample of N observations $x_1, ..., x_N$ that follow a probability density function $f(x|\theta)$, the likelihood can be written as:

$$L(\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

- The MLE is the value $\hat{\theta}$ that maximizes L (the maximum of $\ln L$ is the same). Therefore:

$$\frac{d \ln L}{d\theta}\Big|_{\hat{\theta}} = 0$$

- The point estimate has to be accompanied by an error interval. This error is generally estimated assuming that the likelihood function can be approximated by a Gaussian distribution around the MLE.

Likelihood with a single parameter

- Given a sample of N observations $x_1, ..., x_N$ that follow a probability density function $f(x|\theta)$, the likelihood can be written as:

$$L(\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

- The MLE is the value $\hat{\theta}$ that maximizes L (the maximum of $\ln L$ is the same). Therefore:

$$\frac{d \ln L}{d\theta}\Big|_{\hat{\theta}} = 0$$

- The point estimate has to be accompanied by an error interval. This error is generally estimated assuming that the likelihood function can be approximated by a Gaussian distribution around the MLE.
- However, this is not always the case, and the likelihood function can be asymmetric or have multiple maxima. In such cases, other methods are needed to estimate the error interval.

Confidence intervals and MLE

- When the number of data is sufficiently large, we can approximate the likelihood function by a Gaussian distribution centered at the MLE $\hat{\theta}$:

$$L(\theta) \approx L(\hat{\theta}) e^{-\frac{1}{2} \frac{\left(\theta - \hat{\theta}\right)^2}{\sigma_\theta^2}}$$

Confidence intervals and MLE

- When the number of data is sufficiently large, we can approximate the likelihood function by a Gaussian distribution centered at the MLE $\hat{\theta}$:

$$L(\theta) \approx L(\hat{\theta}) e^{-\frac{1}{2} \frac{\left(\theta - \hat{\theta}\right)^2}{\sigma_\theta^2}}$$

- Now we consider the taylor expansion of the likelihood function around the MLE $\hat{\theta}$:

Confidence intervals and MLE

- When the number of data is sufficiently large, we can approximate the likelihood function by a Gaussian distribution centered at the MLE $\hat{\theta}$:

$$L(\theta) \approx L(\hat{\theta}) e^{-\frac{1}{2}\frac{\left(\theta-\hat{\theta}\right)^2}{\sigma_\theta^2}}$$

- Now we consider the taylor expansion of the likelihood function around the MLE $\hat{\theta}$:

$$\begin{aligned}
\ln L(\theta) &= \ln L(\hat{\theta}) + \frac{d(\ln L)}{d\theta}|_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right) + \frac{1}{2}\frac{d^2(\ln L(\theta))}{d\theta^2}|_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right)^2 \\
&= \ln L(\hat{\theta}) - \frac{1}{2}\frac{\left(\theta - \hat{\theta}\right)^2}{\sigma_\theta^2}
\end{aligned}$$

Therefore if $\theta - \hat{\theta} = \sigma_\theta$, we have:

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

$$\frac{L(\theta)}{L(\hat{\theta})} = e^{-1/2}$$

Confidence intervals and MLE

- The limits for the $68\,\%$ (1σ) confidence interval are those for which the likelihood function $L(\theta)$ satisfies

$$L(\theta) = e^{-1/2} L_{\max} \tag{1}$$

or equivalently

$$\ln L_{\max} - \ln L(\theta) = \ln L(\hat{\theta}) - \ln L(\theta) = \frac{1}{2} \tag{2}$$

For $95\,\%$ and $99\,\%$ (2 and 3 standard deviations), the factors are e^{-2} and $e^{-4,5}$, respectively.

- Remember that $L(\theta)$ is not a probability density function, but a function of the parameters. The confidence region is defined by the values of the parameters for which the likelihood is higher than outside them. ($L(\theta) > L(\hat{\theta})$).

Example: Lifetime of a particle

- Suppose we have N measurements of the decay times t_i of an unstable particle. The probability density function for the decay times follows an exponential distribution.

$$f(t|\gamma) = \gamma e^{-\gamma t}$$

where $\gamma = \frac{1}{\tau}$, with τ being the mean lifetime of the particle. Use the MLE method to find the estimated value of the particle's lifetime.

Example: Lifetime of a particle

- Suppose we have N measurements of the decay times t_i of an unstable particle. The probability density function for the decay times follows an exponential distribution.

$$f(t|\gamma) = \gamma e^{-\gamma t}$$

  where $\gamma = \frac{1}{\tau}$, with τ being the mean lifetime of the particle. Use the MLE method to find the estimated value of the particle's lifetime.
- We can write the likelihood corresponding to the N measurements as:

$$L(\gamma) = \prod_{i=1}^{N} f(t_i|\gamma)$$

Example: Lifetime of a particle

- Suppose we have N measurements of the decay times t_i of an unstable particle. The probability density function for the decay times follows an exponential distribution.

$$f(t|\gamma) = \gamma e^{-\gamma t}$$

where $\gamma = \frac{1}{\tau}$, with τ being the mean lifetime of the particle. Use the MLE method to find the estimated value of the particle's lifetime.

- We can write the likelihood corresponding to the N measurements as:

$$\begin{aligned} L(\gamma) &= \prod_{i=1}^{N} f(t_i|\gamma) \\ &= \prod_{i=1}^{N} \gamma e^{-\gamma t_i} = \gamma^N \prod_{i=1}^{N} e^{-\gamma t_i} = \gamma^N e^{-\sum_{i=1}^{N} \gamma t_i} \end{aligned}$$

Taking the logarithm, we obtain:

$$\ln L = N \ln \gamma - \gamma \sum_{i=1}^{N} t_i$$

Example: Lifetime of a particle

- Suppose we have N measurements of the decay times t_i of an unstable particle. The probability density function for the decay times follows an exponential distribution.

$$f(t|\gamma) = \gamma e^{-\gamma t}$$

where $\gamma = \frac{1}{\tau}$, with τ being the mean lifetime of the particle. Use the MLE method to find the estimated value of the particle's lifetime.

- We can write the likelihood corresponding to the N measurements as:

$$\begin{aligned} L(\gamma) &= \prod_{i=1}^{N} f(t_i|\gamma) \\ &= \prod_{i=1}^{N} \gamma e^{-\gamma t_i} = \gamma^N \prod_{i=1}^{N} e^{-\gamma t_i} = \gamma^N e^{-\sum_{i=1}^{N} \gamma t_i} \end{aligned}$$

Taking the logarithm, we obtain:

$$\ln L = N \ln \gamma - \gamma \sum_{i=1}^{N} t_i$$

- We want to use the MLE method to find the estimated value of the particle's lifetime:

$$\frac{d \ln L}{d\gamma}\Big|_{\hat{\gamma}} = \frac{N}{\hat{\gamma}} - \sum_{i=1}^{N} t_i = 0$$

Examples

- MLE of the mean value of a normal distribution with known width: Given N observation x_i drawn from a normal distribution of known width σ and mean value μ to be estimated.
- MLE of the width of a normal distribution with given mean value μ and unknown width σ to be estimated.
- MLEs of the mean value and the width of a normal distribution where both parameters are unknown

Likelihood for Several Parameters

- Given a sample of N observations $x_1, \ldots, x_N$ that follow a probability density function $f(x|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the likelihood can be written as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta})$$

Likelihood for Several Parameters

- Given a sample of N observations $x_1, \ldots, x_N$ that follow a probability density function $f(x|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the likelihood can be written as:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta}) \\
\ln L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} f(x_i|\boldsymbol{\theta})
\end{aligned}
$$

Likelihood for Several Parameters

- Given a sample of N observations $x_1, \ldots, x_N$ that follow a probability density function $f(x|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the likelihood can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta}) \\ \ln L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} f(x_i|\boldsymbol{\theta}) \end{aligned}$$

- The MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ maximizes L (the maximum of $\ln L$ is the same). This can be written as a system of equations:

$$\begin{aligned} \frac{d \ln L}{d\theta_1}\Big|_{(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} &= 0 \\ \frac{d \ln L}{d\theta_2}\Big|_{(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} &= 0 \\ \ldots &= 0 \\ \frac{d \ln L}{d\theta_k}\Big|_{(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} &= 0 \end{aligned}$$

Likelihood for Several Parameters

- Given a sample of N observations $x_1, \ldots, x_N$ that follow a probability density function $f(x|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the likelihood can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta}) \\ \ln L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} f(x_i|\boldsymbol{\theta}) \end{aligned}$$

- The MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ maximizes L (the maximum of $\ln L$ is the same). This can be written as a system of equations:

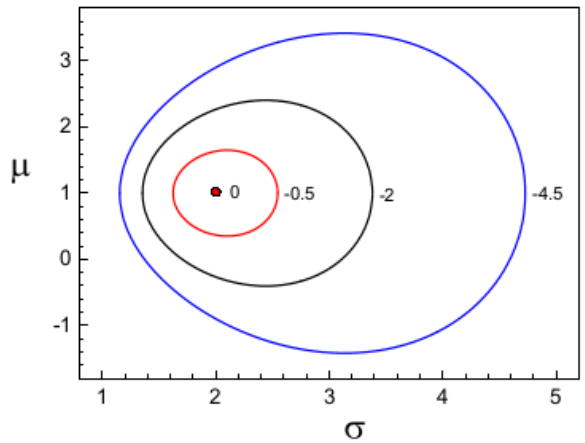
$$\begin{aligned} \frac{d \ln L}{d\theta_1}\Big|_{(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} &= 0 \\ \frac{d \ln L}{d\theta_2}\Big|_{(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} &= 0 \\ \ldots &= 0 \\ \frac{d \ln L}{d\theta_k}\Big|_{(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} &= 0 \end{aligned}$$

Likelihood for Several Parameters

- The error interval is now to be replaced by an error volume with its surface defined again by the likelihood function. The boundaries for the $68\,\%$ (1σ) confidence region are those for which the likelihood function satisfies:

$$L(\boldsymbol{\theta}) = e^{-1/2} L(\hat{\boldsymbol{\theta}}) \tag{3}$$

- This defines a closed surface in the parameter space, in two dimensions just a closed contour.
- For $95\,\%$ and $99\,\%$ (2 and 3 standard deviations), the factors are e^{-2} and $e^{-4,5}$, respectively.



Normally distribute variates and χ^2

- Let us assume that N observations x_i each following a normal distribution with variance δ_i^2 are to be compared with the function $g_i(\theta)$. (here each x_i follows a different distribution)

Normally distribute variates and χ^2

- Let us assume that N observations x_i each following a normal distribution with variance δ_i^2 are to be compared with the function $g_i(\theta)$. (here each x_i follows a different distribution)

$$f(x_1, ..., x_n|\theta) = L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\delta_i} \exp\left[-\frac{(x_i - g_i(\theta))^2}{2\delta_i^2}\right]$$

Normally distribute variates and χ^2

- Let us assume that N observations x_i each following a normal distribution with variance δ_i^2 are to be compared with the function $g_i(\theta)$. (here each x_i follows a different distribution)

$$f(x_1, ..., x_n|\theta) = L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\delta_i} \exp\left[-\frac{(x_i - g_i(\theta))^2}{2\delta_i^2}\right]$$

- Therefore,

$$\ln L(\theta) \quad = \quad \sum_{i=1}^{N} - \ln\left[\sqrt{2\pi\delta_i^2}\right] + \sum_{i=1}^{N} - \frac{(x_i - g_i(\theta))^2}{2\delta_i^2}$$

Normally distribute variates and χ^2

- Let us assume that N observations x_i each following a normal distribution with variance δ_i^2 are to be compared with the function $g_i(\theta)$. (here each x_i follows a different distribution)

$$f(x_1, ..., x_n|\theta) = L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\delta_i} \exp\left[-\frac{(x_i - g_i(\theta))^2}{2\delta_i^2}\right]$$

- Therefore,

$$\begin{aligned} \ln L(\theta) &= \sum_{i=1}^{N} -\ln\left[\sqrt{2\pi\delta_i^2}\right] + \sum_{i=1}^{N} -\frac{(x_i - g_i(\theta))^2}{2\delta_i^2} \\ &= -\frac{1}{2}\sum_{i=1}^{N} \ln\left[2\pi\delta_i^2\right] - \frac{1}{2}\frac{(x_i - g_i(\theta))^2}{\delta_i^2} \end{aligned}$$

Normally distribute variates and χ^2

- Let us assume that N observations x_i each following a normal distribution with variance δ_i^2 are to be compared with the function $g_i(\theta)$. (here each x_i follows a different distribution)

$$f(x_1, ..., x_n|\theta) = L(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\delta_i} \exp\left[-\frac{(x_i - g_i(\theta))^2}{2\delta_i^2}\right]$$

- Therefore,

$$\begin{aligned} \ln L(\theta) &= \sum_{i=1}^{N} -\ln\left[\sqrt{2\pi\delta_i^2}\right] + \sum_{i=1}^{N} -\frac{(x_i - g_i(\theta))^2}{2\delta_i^2} \\ &= -\frac{1}{2}\sum_{i=1}^{N} \ln\left[2\pi\delta_i^2\right] - \frac{1}{2}\frac{(x_i - g_i(\theta))^2}{\delta_i^2} \end{aligned}$$

- The last term in the previous equation is called χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - g_i(\theta))^2}{\delta_i^2}$$

Normally distribute variates and χ^2

- For parameter inference we can omit the constant terms and therefore

$$\ln L = -\frac{1}{2}\chi^2$$

- Minimizing χ^2 is equivalent to maximizing the log likelihood $\ln L(\theta)$.
- This problem is related to the case when we have to compare measurements with normally distributed errors to a parameter dependent prediction, for instance when we fit a curve to measurements..

Least Squares Method

- A common problem arises when we have a set of N data points $(x_i, y_i \pm \delta_i)$ and we want to fit these data to a theoretical function $y = f(x, \theta)$, where θ represents the vector of free parameters in the model. The standard solution to this problem is the least squares method, which consists of minimizing the following function:

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \theta))^2}{\delta_i^2}$$

Least Squares Method

- A common problem arises when we have a set of N data points $(x_i, y_i \pm \delta_i)$ and we want to fit these data to a theoretical function $y = f(x, \theta)$, where θ represents the vector of free parameters in the model. The standard solution to this problem is the least squares method, which consists of minimizing the following function:

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \theta))^2}{\delta_i^2}$$

- This is one of the best methods for fitting data when the variance of the errors is known. Moreover, this method is closely related to the maximum likelihood method when the errors follow a Gaussian distribution. In that case:

$$\begin{aligned} f(y_1, ..., y_n|\theta) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\delta_i} \exp\left[-\frac{(y_i - f(x_i, \theta))^2}{2\delta_i^2}\right] \\ \ln L(\theta) &= -\frac{1}{2}\left[\sum_{i=1}^{N} \ln(2\pi\delta_i^2) + \frac{(y_i - f(x_i, \theta))^2}{\delta_i^2}\right] \\ &= -\frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\delta_i^2) - \frac{1}{2}\chi^2(\theta) \end{aligned}$$

Least Squares Method

- If the errors are known and follow a Gaussian distribution, minimizing χ^2 is equivalent to maximizing the likelihood.
- To determine the standard error, recall that in the maximum likelihood method, the error is defined where the likelihood reaches $e^{-1/2}$ of its maximum value. This is equivalent to requiring that $\chi^2 - \chi^2_{\text{min}} = 1$.

Least Squares Method

- If the errors are known and follow a Gaussian distribution, minimizing χ^2 is equivalent to maximizing the likelihood.
- To determine the standard error, recall that in the maximum likelihood method, the error is defined where the likelihood reaches $e^{-1/2}$ of its maximum value. This is equivalent to requiring that $\chi^2 - \chi^2_{\text{min}} = 1$.
- The function $\chi^2(\theta)$ follows a χ^2 distribution with $\nu = N - P$ degrees of freedom, where N is the number of data points and P is the number of free parameters in the model. The mean value of the χ^2 distribution with ν degrees of freedom is ν, and its variance is 2ν.

Table to calculate confidence intervals

Number of Parameters	Confidence level (probability contents desired inside hypercontour of $\chi^2 = \chi^2_{\min} + \texttt{UP}$)				
	50%	70%	90%	95%	99%
1	0.46	1.07	2.70	3.84	6.63
2	1.39	2.41	4.61	5.99	9.21
3	2.37	3.67	6.25	7.82	11.36
4	3.36	4.88	7.78	9.49	13.28
5	4.35	6.06	9.24	11.07	15.09
6	5.35	7.23	10.65	12.59	16.81
7	6.35	8.38	12.02	14.07	18.49
8	7.34	9.52	13.36	15.51	20.09
9	8.34	10.66	14.68	16.92	21.67
10	9.34	11.78	15.99	18.31	23.21
11	10.34	12.88	17.29	19.68	24.71
	If FCN is $-\log(\text{likelihood})$ instead of χ^2, all values of UP should be divided by 2.				

Table 7.1: Table of UP for multi-parameter confidence regions

Example: Fitting a Straight Line

- Fit a set of N observations (x_i, y_i) with uncertainties δ_i to a straight line $y = ax + b$.

Example: Fitting a Straight Line

- Fit a set of N observations (x_i, y_i) with uncertainties δ_i to a straight line $y = ax + b$.
- The quantity to minimize is:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{\delta_i^2}$$

Example: Fitting a Straight Line

- Fit a set of N observations (x_i, y_i) with uncertainties δ_i to a straight line $y = ax + b$.
- The quantity to minimize is:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{\delta_i^2}$$

- That is, the following conditions must be satisfied:

$$\frac{d\chi^2}{da} = \sum_{i=1}^{N} \frac{2x_i(-y_i + ax_i + b)}{\delta_i^2} = 0$$

$$\frac{d\chi^2}{db} = \sum_{i=1}^{N} \frac{2(-y_i + ax_i + b)}{\delta_i^2} = 0$$

Example: Fitting a Straight Line

- We define the following quantities, with $w_i = \frac{1}{\delta_i^2}$:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad \bar{y} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}$$

$$\overline{x^2} = \frac{\sum_{i=1}^{N} w_i x_i^2}{\sum_{i=1}^{N} w_i} \qquad \overline{xy} = \frac{\sum_{i=1}^{N} w_i x_i y_i}{\sum_{i=1}^{N} w_i}$$

Example: Fitting a Straight Line

- We define the following quantities, with $w_i = \frac{1}{\delta_i^2}$:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad \bar{y} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}$$

$$\overline{x^2} = \frac{\sum_{i=1}^{N} w_i x_i^2}{\sum_{i=1}^{N} w_i} \qquad \overline{xy} = \frac{\sum_{i=1}^{N} w_i x_i y_i}{\sum_{i=1}^{N} w_i}$$

- The conditions to minimize χ^2 can be written as:

$$\begin{aligned} \hat{b} &= \bar{y} - \hat{a}\bar{x} \\ \overline{xy} - \hat{a}\overline{x^2} - \hat{b}\bar{x} &= 0 \end{aligned}$$

- The values of the parameters a and b that minimize χ^2 and maximize the likelihood are:

$$\hat{a} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$\hat{b} = \frac{\overline{x^2}\bar{y} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2}$$

Generalization for Correlated Errors

- When the errors are correlated, the χ^2 function can be generalized as:

$$\chi^2(\theta) = \sum_{i,j=1}^{N} (y_i - f(x_i, \theta))^T \; C_{ij}^{-1} \; (y_j - f(x_j, \theta))$$

where C_{ij} is the covariance matrix, with the variances δ_i^2 on the diagonal and the correlation coefficients between data points in the off-diagonal elements.

Generalization for Correlated Errors

- When the errors are correlated, the χ^2 function can be generalized as:

$$\boxed{\chi^2(\theta) = \sum_{i,j=1}^{N} (y_i - f(x_i,\theta))^T \; C_{ij}^{-1} \; (y_j - f(x_j,\theta))}$$

where C_{ij} is the covariance matrix, with the variances δ_i^2 on the diagonal and the correlation coefficients between data points in the off-diagonal elements.

- Likewise, the likelihood for the case of correlated errors can be written as:

$$\boxed{L(\theta) = \frac{1}{(2\pi)^{N/2}\det(\mathbf{C})^{1/2}} \; \exp\left[-\frac{1}{2}(\mathbf{y} - f(\mathbf{x},\theta))^T \; \mathbf{C^{-1}} \; (\mathbf{y} - f(\mathbf{x},\theta))\right]}$$

where

$$\mathbf{y} = (y_1, y_2, \cdots, y_N)$$

is the vector of observations and

$$f(\mathbf{x},\theta) = (f(x_1,\theta), f(x_2,\theta), \cdots, f(x_n,\theta))$$

is the model prediction. The relation $\boxed{\ln L \approx -\frac{1}{2}\chi^2}$ still holds.

Nuisance Parameters and Profile Likelihood

- Sometimes, we are interested in only a subset of the parameters in a problem. Parameters that are not of direct interest are called nuisance parameters (ν). In the frequentist approach, the way to deal with this nuisance parameters is the so called Profile Likelihood.

Nuisance Parameters and Profile Likelihood

- Sometimes, we are interested in only a subset of the parameters in a problem. Parameters that are not of direct interest are called nuisance parameters (ν). In the frequentist approach, the way to deal with this nuisance parameters is the so called Profile Likelihood.
- The likelihood function is maximized with respect to the nuisance parameter ν as a function of the parameter of interest θ, The function $\hat{\nu}(\theta)$ which maximized the likelihood is given by:

$$\frac{dL(\theta, \nu)}{d\nu}|_{\hat{\nu}} = 0$$

Nuisance Parameters and Profile Likelihood

- Sometimes, we are interested in only a subset of the parameters in a problem. Parameters that are not of direct interest are called nuisance parameters (ν). In the frequentist approach, the way to deal with this nuisance parameters is the so called Profile Likelihood.
- The likelihood function is maximized with respect to the nuisance parameter ν as a function of the parameter of interest θ, The function $\hat{\nu}(\theta)$ which maximized the likelihood is given by:

$$\frac{dL(\theta,\nu)}{d\nu}|_{\hat{\nu}} = 0$$

- The profile likelihood is defined as:

$$L_p(\theta) = L(\theta, \hat{\nu}(\theta))$$

- The profile likelihood is a function of the parameter of interest θ only, and it is used to construct confidence intervals for θ.
- The following video1 and video2 explain the basics concepts of profile likelihood

Bayesian Approach

- We recall Bayes' theorem for PDFs

$$p(\theta|D,I) = \frac{p(\theta|I)p(D|\theta,I)}{p(D|I)}$$

  - $p(\theta|D,I)$ is the posterior probability of the parameters θ given the data D and the prior information I.
  - $p(\theta|I)$ is the prior probability of the parameters θ given the prior information I.
  - $p(D|\theta,I)$ is the likelihood of the data given the parameters θ and the prior information I.
  - $p(D|I)$ is a normalization constant, known as baysian evidence, which ensures that the posterior probability integrates to 1.

$$p(D|I) = \int d\theta\, p(\theta|I)p(D|\theta,I)$$

Prior Probability or PRIORS

- A uniform prior is defined as:

$$p(\theta|I) = \begin{cases} K & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

Prior Probability or PRIORS

- A uniform prior is defined as:

$$p(\theta|I) = \begin{cases} K & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

- Since this is a probability, it must be normalized:

$$\int_{-\infty}^{\infty} p(\theta|I)d\theta = \int_{\theta_{\min}}^{\theta_{\max}} K d\theta = 1$$

Prior Probability or PRIORS

- A uniform prior is defined as:

$$p(\theta|I) = \begin{cases} K & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

- Since this is a probability, it must be normalized:

$$\int_{-\infty}^{\infty} p(\theta|I)d\theta = \int_{\theta_{\min}}^{\theta_{\max}} K d\theta = 1$$

and therefore $K = \frac{1}{\theta_{\max} - \theta_{\min}} = \frac{1}{R_\theta}$

Jeffreys Prior

- The Jeffreys prior is equivalent to a uniform prior on a logarithmic scale:

$$p(\theta|I) = \begin{cases} \frac{K}{\theta} & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

Jeffreys Prior

- The Jeffreys prior is equivalent to a uniform prior on a logarithmic scale:

$$p(\theta|I) = \begin{cases} \frac{K}{\theta} & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

- Since this is a probability, it must be normalized:

$$\int_{-\infty}^{\infty} p(\theta|I)d\theta = \int_{\theta_{\min}}^{\theta_{\max}} \frac{K}{\theta} d\theta = K \ln\left(\frac{\theta_{\max}}{\theta_{\min}}\right) = 1$$

Jeffreys Prior

- The Jeffreys prior is equivalent to a uniform prior on a logarithmic scale:

$$p(\theta|I) = \begin{cases} \frac{K}{\theta} & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

- Since this is a probability, it must be normalized:

$$\int_{-\infty}^{\infty} p(\theta|I)d\theta = \int_{\theta_{\min}}^{\theta_{\max}} \frac{K}{\theta} d\theta = K \ln\left(\frac{\theta_{\max}}{\theta_{\min}}\right) = 1$$

and therefore $K = \frac{1}{\ln(\frac{\theta_{\max}}{\theta_{\min}})}$ and

$$p(\theta|I) = \begin{cases} \frac{1}{\theta \ \ln\left(\frac{\theta_{\max}}{\theta_{\min}}\right)} & \text{if } \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases}$$

Gaussian prior

- A Gaussian prior is defined as:

$$p(\theta|I) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

Gaussian prior

- A Gaussian prior is defined as:

$$p(\theta|I) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

- where μ is the mean value and σ is the standard deviation of the prior distribution. This prior is normalized, and it is often used when we have some prior knowledge about the parameter θ.
- In cosmology, sometimes we use a Gaussian prior for the baryon density given bounds obtained from Big Bang Nucleosynthesis (BBN): $\Omega_b h^2 = 0{,}0225 \pm 0{,}0015$

Parameter Estimation

- The Bayesian solution to the parameter estimation problem is not a point estimate, but rather the posterior distribution $p(\theta|D, I)$ itself.
- The information can be summarized by the most probable value of the parameter or by the posterior mean:

$$\langle\theta\rangle = \int p(\theta|D, I)\theta d\theta$$

  These two values do not necessarily coincide.
- It is also possible to calculate the range R of a parameter that contains a certain probability C:

$$\int_R p(\theta|D, I)d\theta = C$$

Nuisance Parameters in the Bayesian Approach

- In the bayesian approach, the way to deal with nuisanuince parameters is to integrate them out (marginalization). For example, if we have a model with two parameters θ and ϕ, and we consider that ν is a nuisance parameter, we can calculate the posterior distribution for θ by marginalizing over ϕ:

$$p(\theta|D) = \int d\phi \, P(\theta, \phi|D)$$

we refer to ϕ as a nuisance parameter.

Confidence Intervals in the Bayesian Approach

- Let us assume that we have a posterior distribution that is a function of many parameters:

$$p(\boldsymbol{\theta}|D, I) = p(\theta_1, \theta_2, \ldots, \theta_k|D, I)$$

Confidence Intervals in the Bayesian Approach

- Let us assume that we have a posterior distribution that is a function of many parameters:

$$p(\boldsymbol{\theta}|D, I) = p(\theta_1, \theta_2, \ldots, \theta_k|D, I)$$

- We can define the $C\,\%$ confidence region R for parameters θ_i and θ_j:

$$\int \int_R p(\theta_1, \theta_2, \ldots, \theta_k|D, I) d\theta_i d\theta_j = C$$

Confidence Intervals in the Bayesian Approach

- Let us assume that we have a posterior distribution that is a function of many parameters:

$$p(\boldsymbol{\theta}|D, I) = p(\theta_1, \theta_2, \ldots, \theta_k|D, I)$$

- We can define the $C\,\%$ confidence region R for parameters θ_i and θ_j:

$$\int \int_R p(\theta_1, \theta_2, \ldots, \theta_k|D, I) d\theta_i d\theta_j = C$$

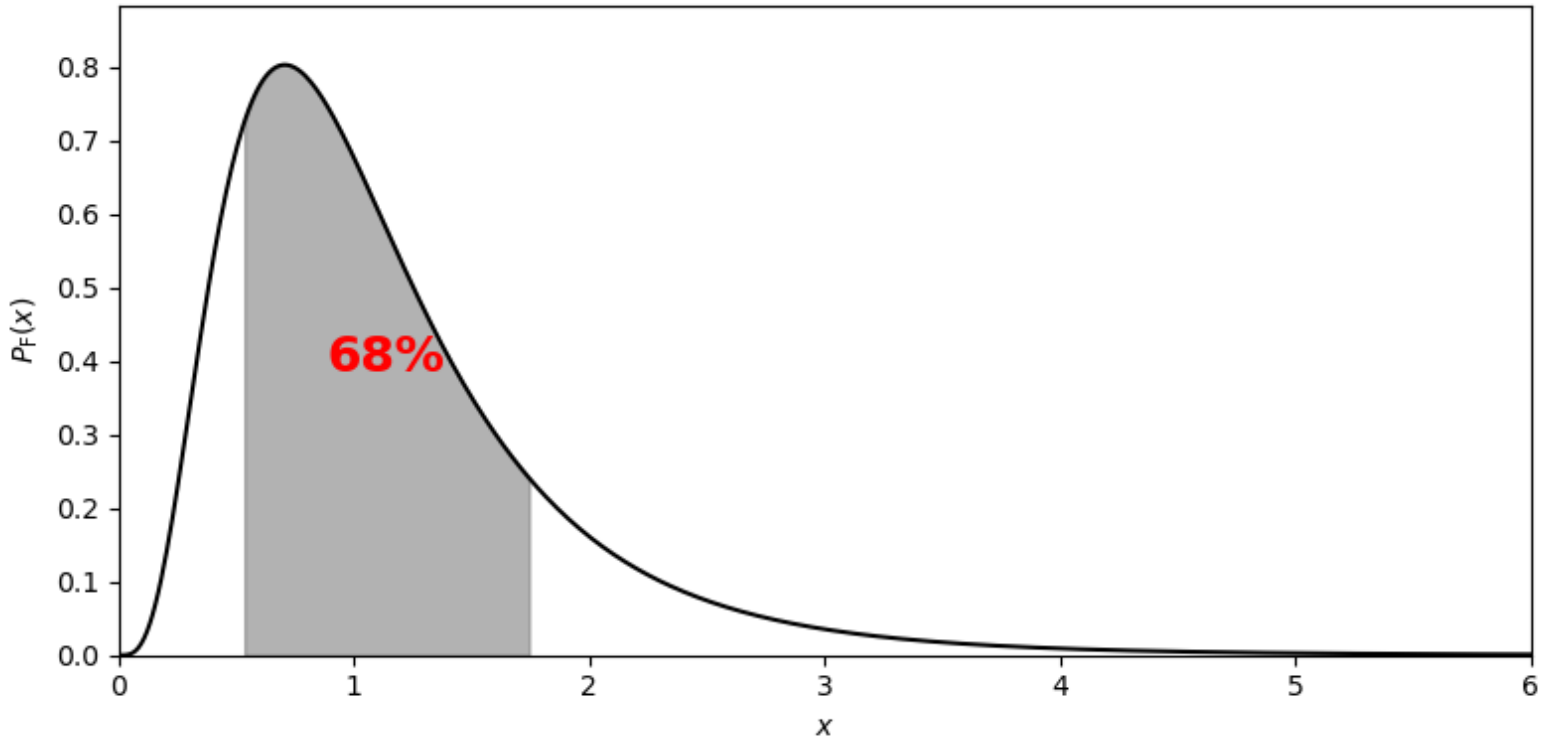
- We can define the $C\,\%$ confidence interval $(-a, a)$ for parameter θ_i:

$$\int_{-a}^{a} p(\theta_1, \theta_2, \ldots, \theta_k|D, I) d\theta_i = C$$

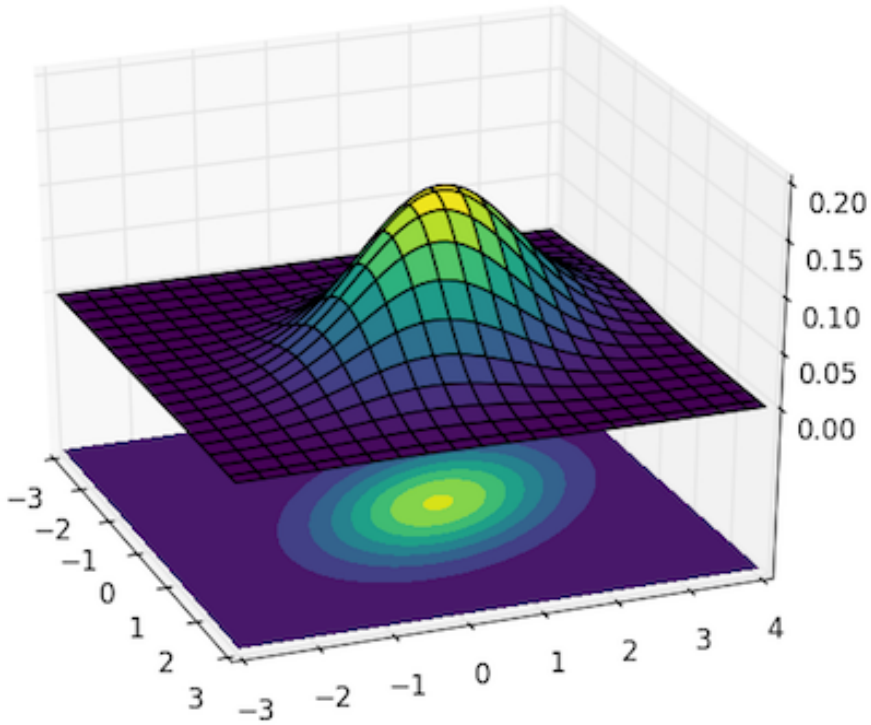
Remarks

- The Bayesian approach does not provide a point estimate for the parameters, but rather a probability distribution for them.
- The Bayesian approach allows us to incorporate prior information about the parameters, which can be useful when we have some knowledge about their values.
- The Bayesian approach is more flexible than the frequentist approach, as it allows us to use different priors and to deal with nuisance parameters in a more natural way.
- The meaning of confidence intervals in the Bayesian approach is different from the frequentist approach. In the Bayesian approach, the confidence interval is a region where the parameter has a certain probability of being found, while in the frequentist approach, it is a region where the parameter would be found in a certain percentage of repeated experiments.

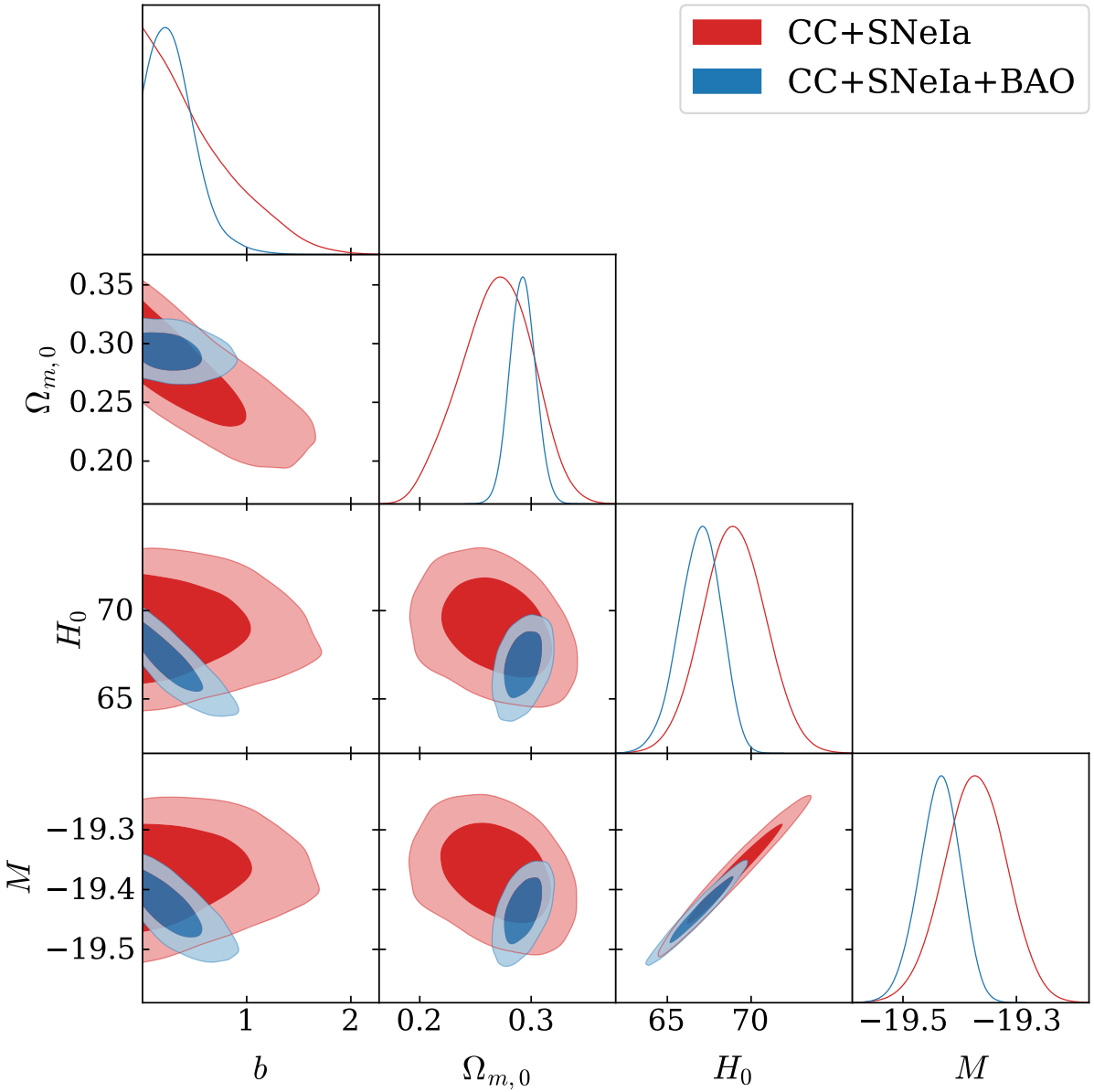
Confidence Intervals



2D Confidence Regions



Triangular plots



Example: The Hubble Constant

- In 1929, Edwin Hubble discovered a linear relationship between the distance to a galaxy d and its recession velocity v, given by $v = H_0 d$, where H_0 is known as the Hubble constant. A usual value for $H_0 = 70 \pm 5$ km/s/Mpc. The recession velocity of a galaxy has been measured as $v_m = 100 \pm 5 \times 10^3$ km/s. Determine the posterior probability for the distance to the galaxy in the following cases:
  1. The value of H_0 is 70 km/s/Mpc.
  2. The prior for H_0 is:

$$p(H_0|I) = k\, e^{-\frac{(H_0-70)^2}{2\times 10^2}}$$

  3. Assume a uniform prior for H_0 given by:

$$p(H_0|I) = \begin{cases} \frac{1}{90-50} & \text{when} \quad 50 < H_0 < 90 \\ 0 & \text{otherwise} \end{cases}$$

  4. Assume a Jeffreys prior for H_0 given by:

$$p(H_0|I) = \begin{cases} \frac{1}{H_0 \ \ln(\frac{90}{50})} & \text{when} \quad 50 < H_0 < 90 \\ 0 & \text{otherwise} \end{cases}$$

Model comparison in the bayesian framework

- In the bayesian framework, model comparison is done using the bayesian evidence, which is defined as:

$$p(D|M,I) = \int d\theta\, p(\theta|M,I)p(D|\theta,M,I)$$

  where M is the model being considered.
- The bayesian evidence is used to calculate the bayes factor, which is defined as the ratio of the evidences of two models:

$$B_{12} = \frac{p(D|M_1,I)}{p(D|M_2,I)}$$

  where M_1 and M_2 are the two models being compared.
- The bayes factor can be used to determine which model is more likely given the data. A value of $B_{12} > 1$ indicates that model M_1 is more likely than model M_2, while a value of $B_{12} < 1$ indicates that model M_2 is more likely than model M_1.

Model comparison in the bayesian framework

- The interpretation of the bayes factor can be done using the Jeffreys scale, which is a table that relates the value of the bayes factor to the strength of evidence for one model over another.
- The Jeffreys scale is as follows:
  - $1 < B_{12} < 3$: weak evidence for M_1 over M_2
  - $3 < B_{12} < 10$: moderate evidence for M_1 over M_2
  - $B_{12} > 10$: strong evidence for M_1 over M_2
- Bayesian evidence is typically obtained through nested sampling or thermodynamic integration methods.
- As a fully Bayesian approach, the value of the Bayes Factor depends strongly on the choice of prior density,

Model comparison

There are other methods based in the Likelihood that allow to compare between models:

- Akaike Information Criterion (AIC):

$$\mathrm{AIC} = 2k - 2\ln(L_{\max})$$

  where k is the number of parameters in the model and $L_{\max}$ is the maximum value of the likelihood function for the model.
- Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = k\ln(n) - 2\ln(L_{\max})$$

  where n is the number of data points.
- Deviance Information Criterion (DIC)

$$\mathrm{DIC} = 2\overline{D(\theta)} - D(\hat{\theta})$$

  where $D(\theta) = -2\ln(L(\theta))$ is the deviance, $\overline{D(\theta)}$ is the average deviance over the posterior distribution, and $D(\hat{\theta})$ is the deviance at the posterior mean.

Model comparison

- The AIC, BIC, and DIC are all based on the likelihood function and penalize models with more parameters. The model with the lowest value of AIC, BIC, or DIC is considered the best model.

Model comparison

- The AIC, BIC, and DIC are all based on the likelihood function and penalize models with more parameters. The model with the lowest value of AIC, BIC, or DIC is considered the best model.
- The AIC is more appropriate for small sample sizes, while the BIC is more appropriate for large sample sizes. The DIC is a fully Bayesian approach that takes into account the uncertainty in the parameter estimates.

Model comparison

- The AIC, BIC, and DIC are all based on the likelihood function and penalize models with more parameters. The model with the lowest value of AIC, BIC, or DIC is considered the best model.
- The AIC is more appropriate for small sample sizes, while the BIC is more appropriate for large sample sizes. The DIC is a fully Bayesian approach that takes into account the uncertainty in the parameter estimates.
- When comparing a test model versus a baseline one (e.g., ΛCDM) using ICs (AIC and BIC), we compute the difference in criterion values as

$$\Delta \mathrm{IC}_{\mathrm{test}} = \mathrm{IC}_{\mathrm{baseline}} - \mathrm{IC}_{\mathrm{test}}$$

- $\Delta IC > 0$ favors the test model, while $\Delta IC < 0$ favors the baseline model. The value of $|\Delta IC|$ indicates the strength of the preference: $|\Delta IC| \geq 2$ (weak), $|\Delta IC| \geq 6$ (medium), $|\Delta IC| \geq 10$ (strong)

Model comparison

- Unlike AIC and BIC, DIC uses the entire posterior distribution rather than just the maximum likelihood estimate, making it particularly suitable for Bayesian hierarchical models where the effective number of parameters may be less than the actual number due to prior constraints.

Suggested Lectures

- Introduction to Statistics and Data Analysis for Physicists Gerhard Bohm, Günter Zech: Chapter 6 until 6.5.7, Chapter 7, Chapter 8 until 8.3
- Bayesian Logical Data Analysis for the Physical Sciences Phil Gregory: Chapter 3 until 3.7
- To solve numerical integrals: `https://www.wolframalpha.com/`