Probability of the data versus likelihood of the parameters

- Suppose you are counting how many cars pass in front of your window on Sundays between 9:00 and 9:02 am. Counting experiments are generally well described by the Poisson distribution. Therefore, if the mean counts are λ, the probability of counting n cars follows the distribution:

$$P(n \mid \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

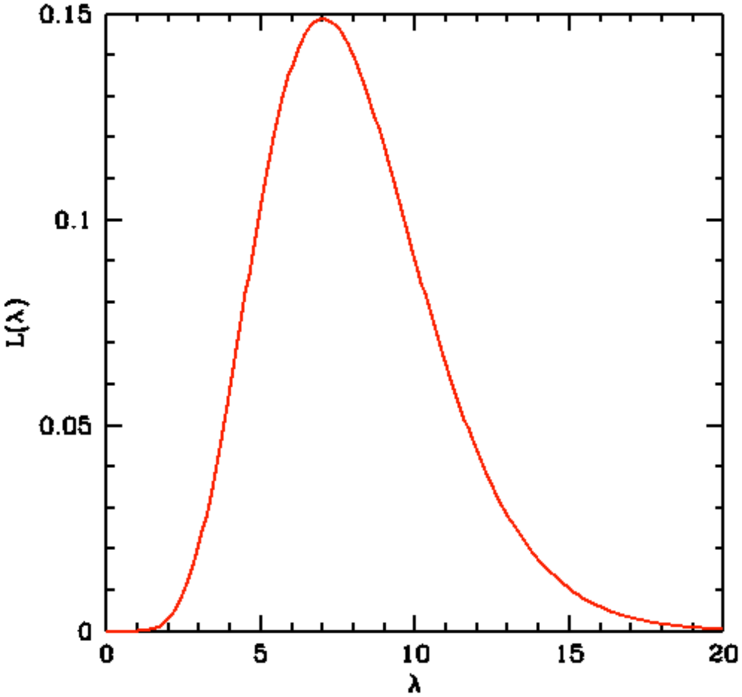
- This means that if you repeat the experiment many times, you will measure different values of n following the frequency P(n). Note that the sum over all possible n is unity.

- Now suppose that you actually perform the experiment once and you count 7. Then, the likelihood for the model parameter λ GIVEN the data is:

$$L(\lambda) = P(7 \mid \lambda) = \frac{\lambda^7 e^{-\lambda}}{5040}$$

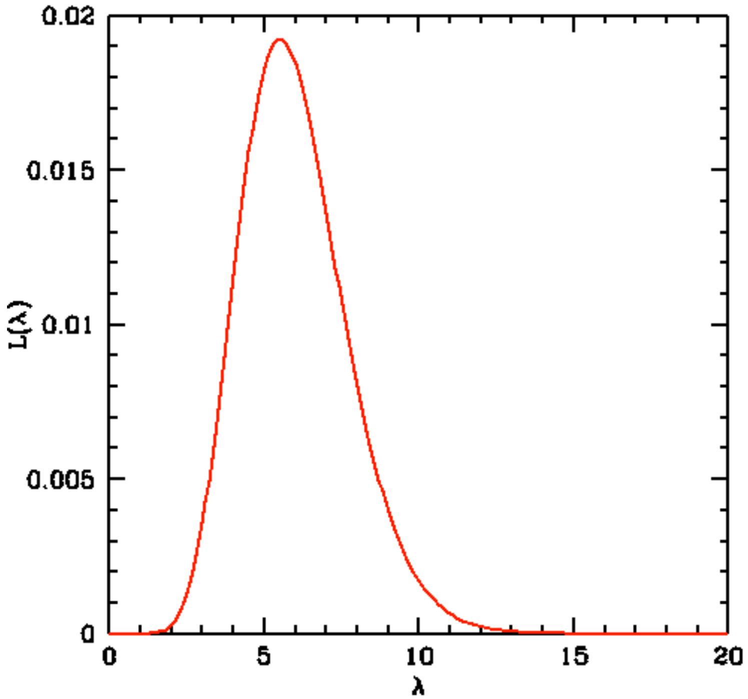
The likelihood function

- This is a function of λ only but it is NOT a probability distribution for λ! It simply says how likely it is that our measured value of n=7 is obtained by sampling a Poisson distribution of mean λ. It says something about the model parameter GIVEN the observed data.



The likelihood function

- Let us suppose that after some time you repeat the experiment and count 4 cars. Since the two experiments are independent, you can multiply the likelihoods and obtain the curve below. Note that now the most likely value is λ=5.5 and the likelihood function is narrower than before, meaning that we know more about λ.



Likelihood for Gaussian errors

- Often statistical measurement errors can be described by Gaussian distributions. If the errors σ_i of different measurements d_i are independent:

$$L(\theta) = P(d \mid \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\left(d_i - m_i(\theta)\right)^2}{2\sigma_i^2} \right]$$

$$-\ln L(\theta) = \sum_{i=1}^{N} \frac{(d_i - m_i(\theta))^2}{2\sigma_i^2} + \text{const.} = \frac{\chi^2(\theta,d)}{2} + \text{const.}$$

- Maximizing the likelihood corresponds to finding the values of the parameters $\theta = \{\theta_1,\ldots,\theta_n\}$ which minimize the χ^2 function (weighted least squares method).

The general Gaussian case

- In general, errors are correlated and

$$-\ln L(\theta) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[d_i - m_i(\theta)\right] C_{ij}^{-1} \left[d_j - m_j(\theta)\right] + \text{const.} = \frac{\chi^2(\theta,d)}{2} + \text{const.}$$

where C_{ij}=< ε_i ε_j> is the covariance matrix of the errors.

- For uncorrelated errors the covariance matrix is diagonal and one reduces to the previous case.

- Note that the covariance matrix could also derive from a model and then depend on the model parameters. We will encounter some of these cases in the rest of the course.

The Likelihood function: a summary

- In simple words, the likelihood of a model given a dataset is proportional to the probability of the data given the model

- The likelihood function supplies an order of preference or plausibility of the values of the free parameters θ_i by how probable they make the observed dataset

- The likelihood ratio between two models can then be used to prefer one to the other

- Another convenient feature of the likelihood function is that it is functionally invariant. This means that any quantitative statement about the θ_i implies a corresponding statements about any one to one function of the θ_i by direct algebraic substitution

Maximum Likelihood

- The likelihood function is a statistic (i.e. a function of the data) which gives the probability of obtaining that particular set of data, given the chosen parameters θ_1, ... , θ_k of the model. It should be understood as a function of the unknown model parameters (but it is NOT a probability distribution for them)

- The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimates or MLE's.

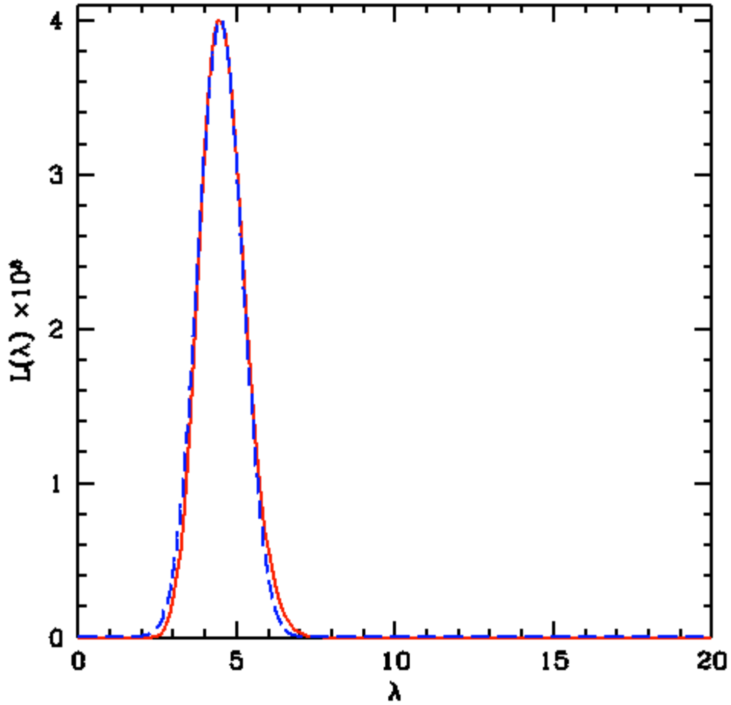
- Assuming that the likelihood function is differentiable, estimation is done by solving

$$\frac{\partial L(\theta_1,...,\theta_k)}{\partial \theta_i} = 0 \qquad \text{or} \qquad \frac{\partial \ln L(\theta_1,...,\theta_k)}{\partial \theta_i} = 0$$

- On the other hand, the maximum value may not exists at all.

Back to counting cars

- After 9 experiments we collected the following data: 7, 4, 2, 6, 4, 5, 3, 4, 5. The new likelihood function is plotted below, together with a Gaussian function (dashed line) which matches the position and the curvature of the likelihood peak (λ =4.44). Note that the 2 curves are very similar (especially close to the peak), and this is not by chance.



Score and information matrix

- The first derivative of the log-likelihood function with respect to the different parameters is called the Fisher score function:

$$S_i = \frac{\partial \ln L(\theta)}{\partial \theta_i}$$

- The Fisher score vanishes at the MLE.

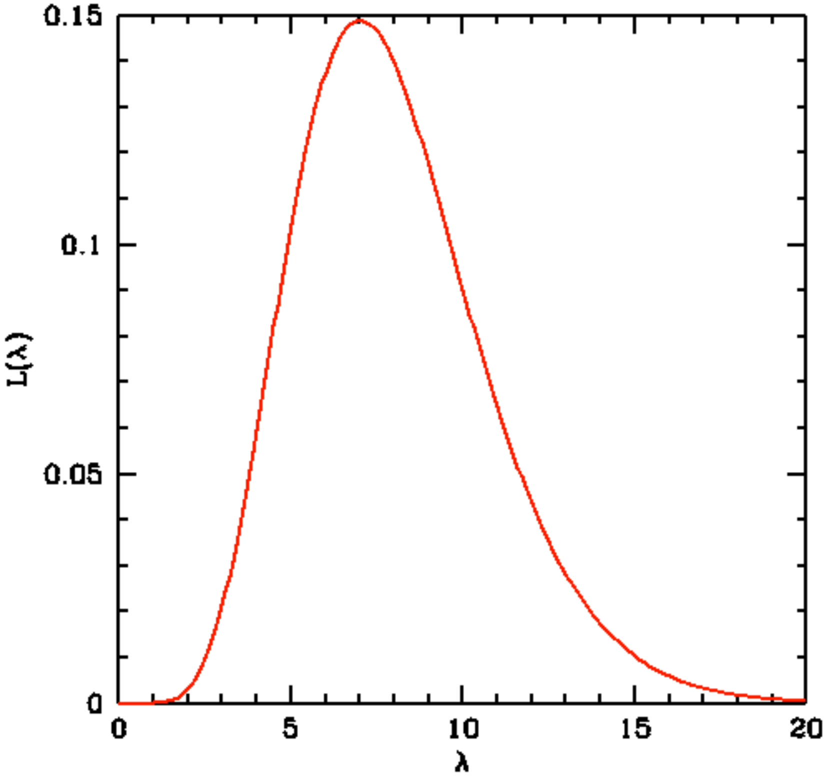
- The negative of the Hessian matrix of the log-likelihood function with respect to the different parameters is called the observed information matrix:

$$O_{ij} = -\frac{\partial^2 \ln L(\theta)}{\partial \theta_i \, \partial \theta_j}$$

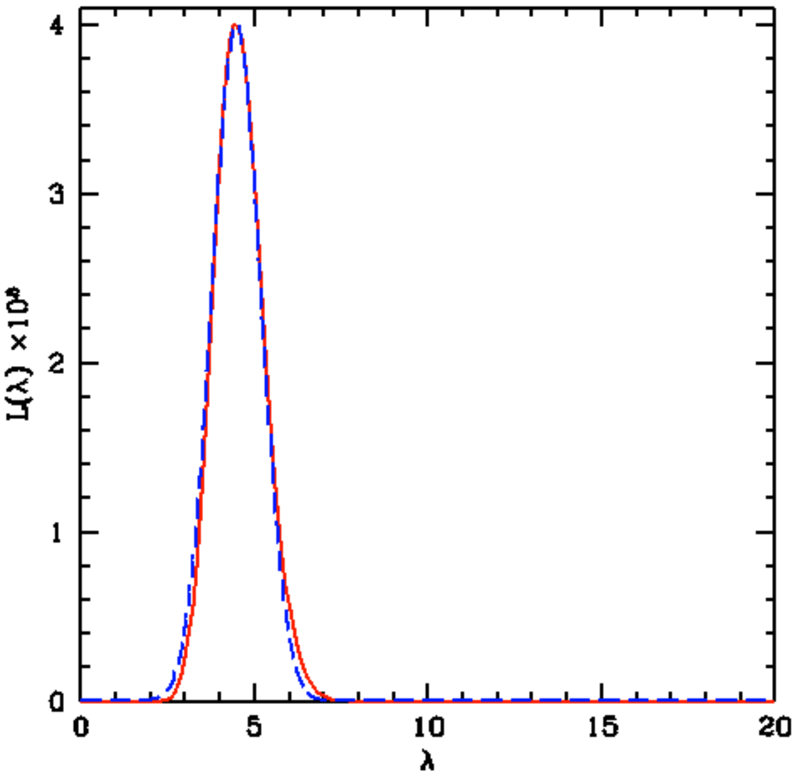
- The observed information matrix is definite positive at the MLE. Its elements tell us how broad is the likelihood function close to its peak and thus with what accuracy we determined the model parameters.

Example

1 datapoint
Low information
Large uncertainty in λ



9 datapoints
High information
Small uncertainty in λ



Fisher information matrix

- If we took different data, then the likelihood function for the parameters would have been a bit different and so its score function and the observed information matrix.
- Fisher introduced the concept of information matrix by taking the ideal ensemble average (over all possible datasets of a given size) of the observed information matrix (evaluated at the true value of the parameters).

$$F_{ij} = -\left\langle \frac{\partial^2 \ln L(\theta)}{\partial\theta_i \, \partial\theta_j} \right\rangle$$

- Under mild regularity conditions, it can been shown that the Fisher information matrix also corresponds to

$$F_{ij} = \left\langle \frac{\partial \;\ln L(\theta)}{\partial\theta_i} \frac{\partial \;\ln L(\theta)}{\partial\theta_j} \right\rangle$$

i.e. to the covariance matrix of the scores at the MLE's.

Cramér-Rao bound

- The Cramér-Rao bound states that, for ANY unbiased estimator of a model parameter θ_i, the measurement error (keeping the other parameters constant) satisfies

$$\Delta\theta_i \geq \frac{1}{\sqrt{F_{ii}}}$$

- For marginal errors that also account for the variability of the other parameters (see slide 75 for a precise definition), instead, it is the inverse of the Fisher information matrix that matters and

$$\Delta\theta_i \geq \sqrt{F^{-1}_{ii}}$$

Fisher matrix with Gaussian errors

- For data with Gaussian errors, the Fisher matrix assumes the form (the notation is the same as in slide 43)

$$F_{ij} = \frac{1}{2}\mathrm{Tr}\left[\underbrace{C^{-1}C_{,i}\,C^{-1}C_{,j}}_{\text{Information from the noise}} + \underbrace{C^{-1}M_{ij}}_{\text{Information from the signal}}\right]$$

where

$$M_{ij} = m_{,i}\,m_{,j}^{T} + m_{,j}\,m_{,i}^{T}$$

(note that commas indicate derivatives with respect to the parameters while data indices are understood)

Properties of MLE's

As the sample size increases to infinity (under weak regularity conditions):

- MLE's become asymptotically efficient and asymptotically unbiased
- MLE's asymptotically follow a normal distribution with covariance matrix (of the parameters) equal to the inverse of the Fisher's information matrix (that is determined by the covariance matrix of the data).

However, for small samples,

- MLE's can be heavily biased and the large-sample optimality does not apply

Maximizing likelihood functions

- For models with a few parameters, it is possible to evaluate the likelihood function on a finely spaced grid and search for its minimum (or use a numerical minimisation algorithm).

- For a number of parameters >>2 it is NOT feasible to have a grid (e.g. 10 point in each parameter direction, 12 parameters = 10^{12} likelihood evaluations!!!)

- Special statistical and numerical methods needs to be used to perform model fitting.

- Note that typical cosmological problems consider models with a number of parameters ranging between 6 and 20.

Forecasting

- Forecasting is the process of estimating the performance of future experiments for which data are not yet available

- It is a key step for the optimization of experimental design (e.g. how large must be my survey if I want to determine a particular parameter to 1% accuracy?)

- The basic formalism has been developed by Fisher in 1935

Figure of merit

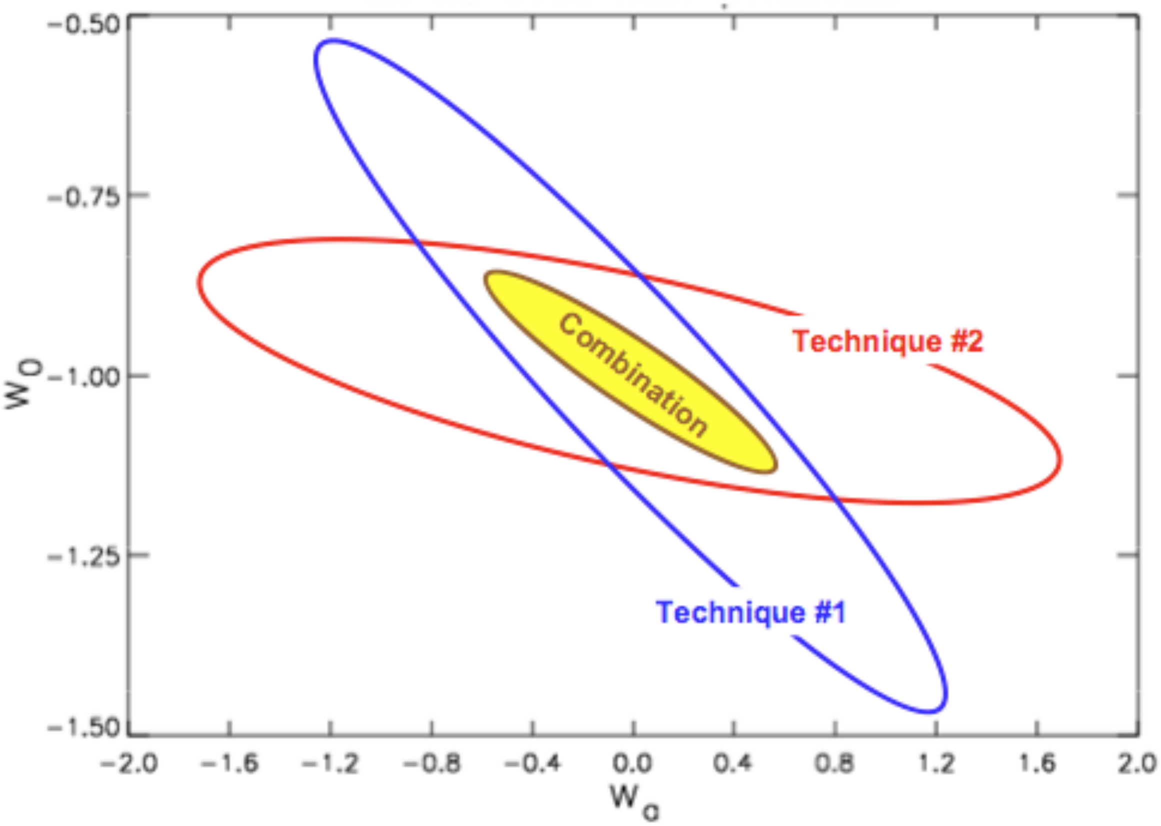
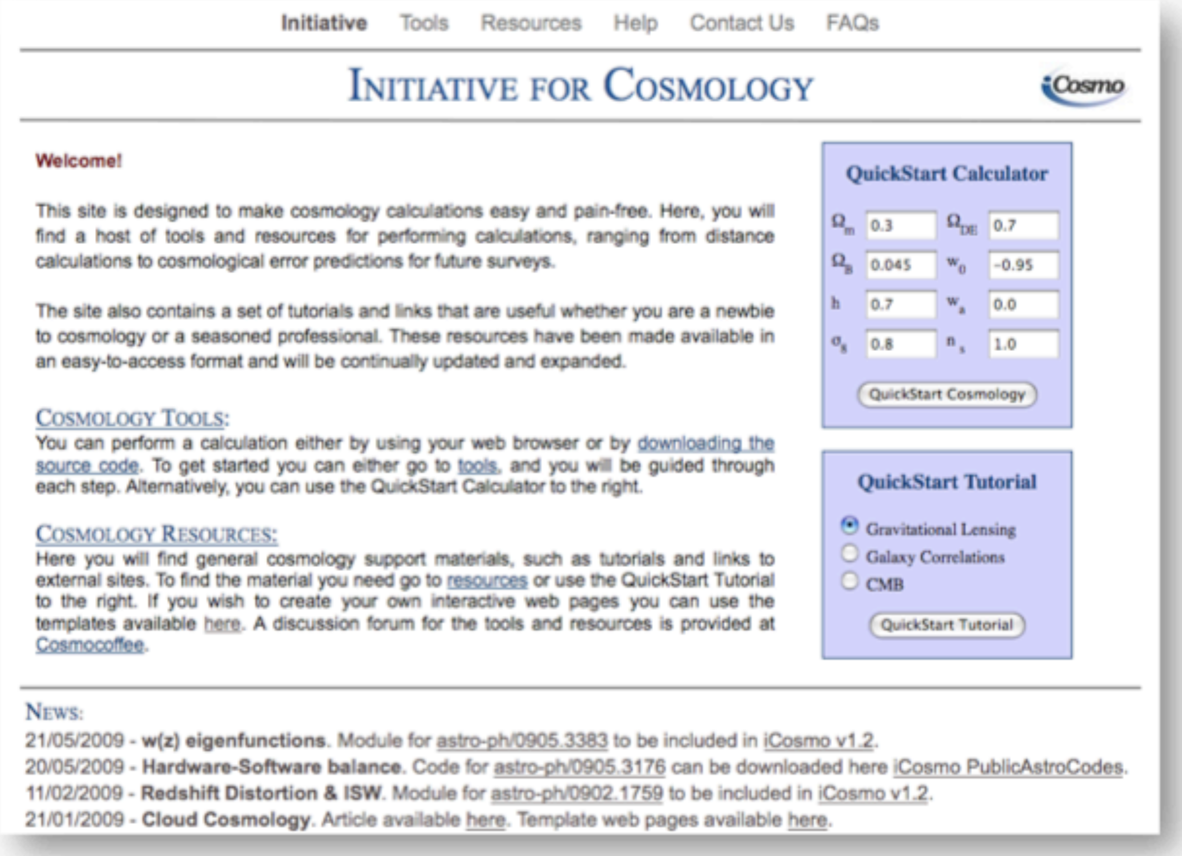


Figure of merit = 1 / (area of the ellipse)

iCOSMO.org



Initiative for Cosmology

Welcome!

This site is designed to make cosmology calculations easy and pain-free. Here, you will find a host of tools and resources for performing calculations, ranging from distance calculations to cosmological error predictions for future surveys.

The site also contains a set of tutorials and links that are useful whether you are a newbie to cosmology or a seasoned professional. These resources have been made available in an easy-to-access format and will be continually updated and expanded.

Cosmology Tools:

You can perform a calculation either by using your web browser or by downloading the source code. To get started you can either go to tools, and you will be guided through each step. Alternatively, you can use the QuickStart Calculator to the right.

Cosmology Resources:

Here you will find general cosmology support materials, such as tutorials and links to external sites. To find the material you need go to resources or use the QuickStart Tutorial to the right. If you wish to create your own interactive web pages you can use the templates available here. A discussion forum for the tools and resources is provided at Cosmocoffee.

QuickStart Calculator

Ω_m	0.3	Ω_{DE}	0.7
Ω_B	0.045	w_0	-0.95
h	0.7	w_a	0.0
σ_8	0.8	n_s	1.0

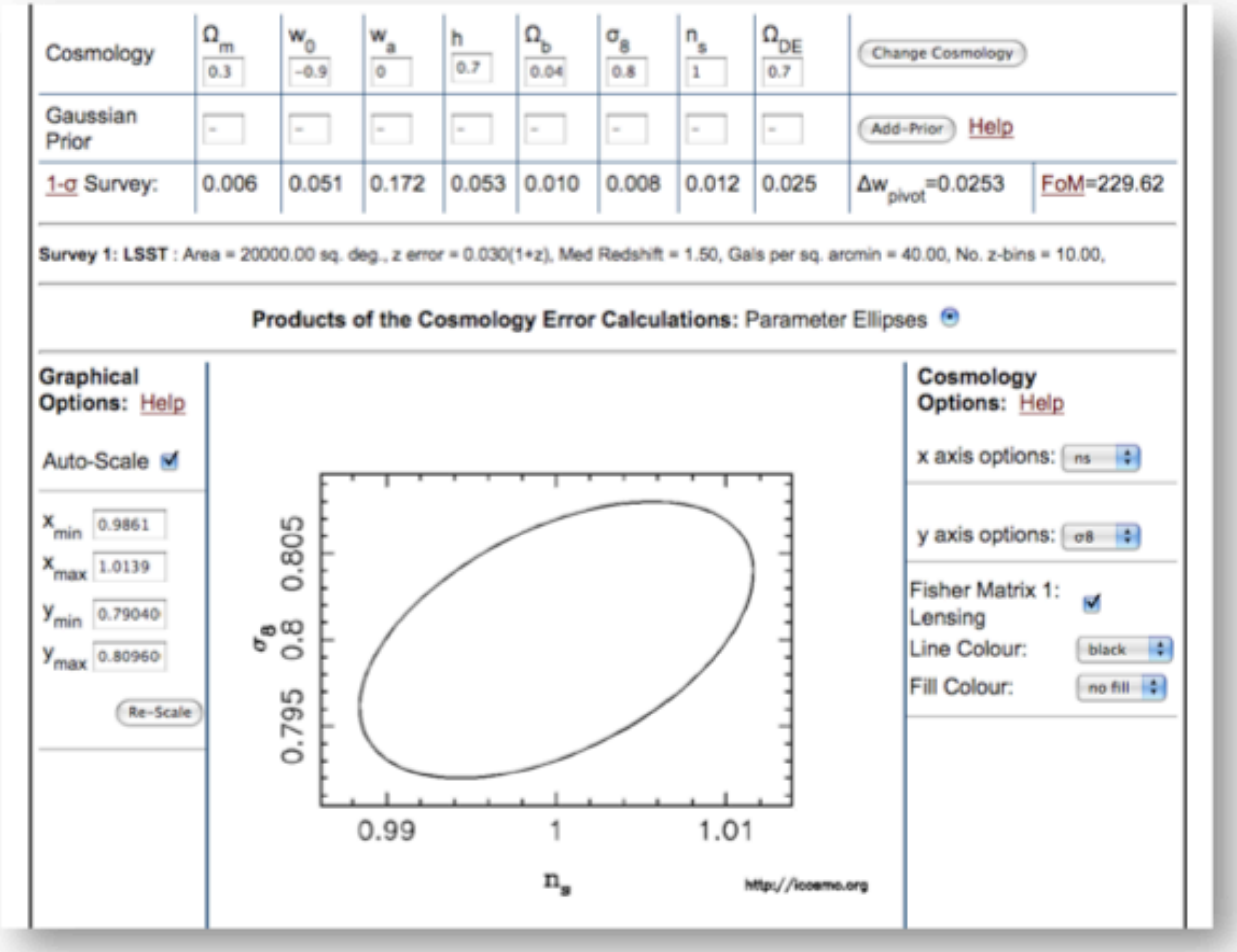
QuickStart Tutorial

● Gravitational Lensing
○ Galaxy Correlations
○ CMB

News:

21/05/2009 - **w(z) eigenfunctions**. Module for astro-ph/0905.3383 to be included in iCosmo v1.2.
20/05/2009 - **Hardware-Software balance**. Code for astro-ph/0905.3176 can be downloaded here iCosmo PublicAstroCodes.
11/02/2009 - **Redshift Distortion & ISW**. Module for astro-ph/0902.1759 to be included in iCosmo v1.2.
21/01/2009 - **Cloud Cosmology**. Article available here. Template web pages available here.

Open source Fisher matrices



	Ω_m	w_0	w_a	h	Ω_b	σ_8	n_s	Ω_{DE}		
Cosmology	0.3	-0.9	0	0.7	0.04	0.8	1	0.7	Change Cosmology	
Gaussian Prior	-	-	-	-	-	-	-	-	Add-Prior Help	
1-σ Survey:	0.006	0.051	0.172	0.053	0.010	0.008	0.012	0.025	Δw_{pivot}=0.0253	FoM=229.62

Survey 1: LSST : Area = 20000.00 sq. deg., z error = 0.030(1+z), Med Redshift = 1.50, Gals per sq. arcmin = 40.00, No. z-bins = 10.00,

Products of the Cosmology Error Calculations: Parameter Ellipses ●

Graphical Options: Help

Auto-Scale ☑

x_{min} 0.9861

x_{max} 1.0139

y_{min} 0.79040

y_{max} 0.80960

Re-Scale

Cosmology Options: Help

x axis options: ns

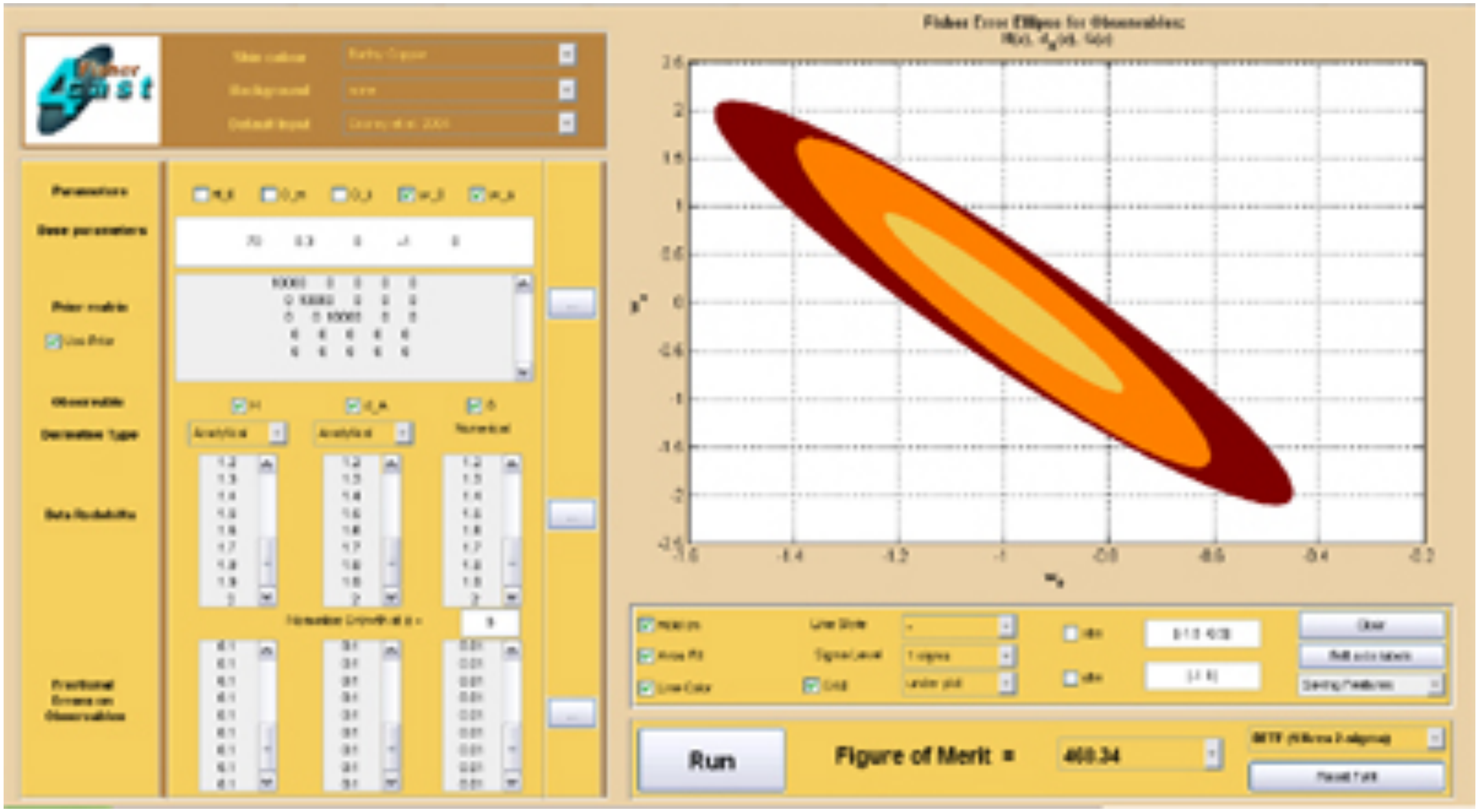
y axis options: σ8

Fisher Matrix 1: Lensing ☑

Line Colour: black

Fill Colour: no fill

Fisher 4cast (Matlab toolbox)



Counting cars, again

- In our study of the car counts we implicitly assumed that all the values of λ are equally likely a priori (i.e. before we started taking the data). However, we didn't consider that an automatic gate regulates the traffic in our street and does not allow more than 8 cars to enter every 10 minutes. Therefore λ cannot be larger than 8 and the likelihood derived from our counts should have been truncated at $\lambda=8$.

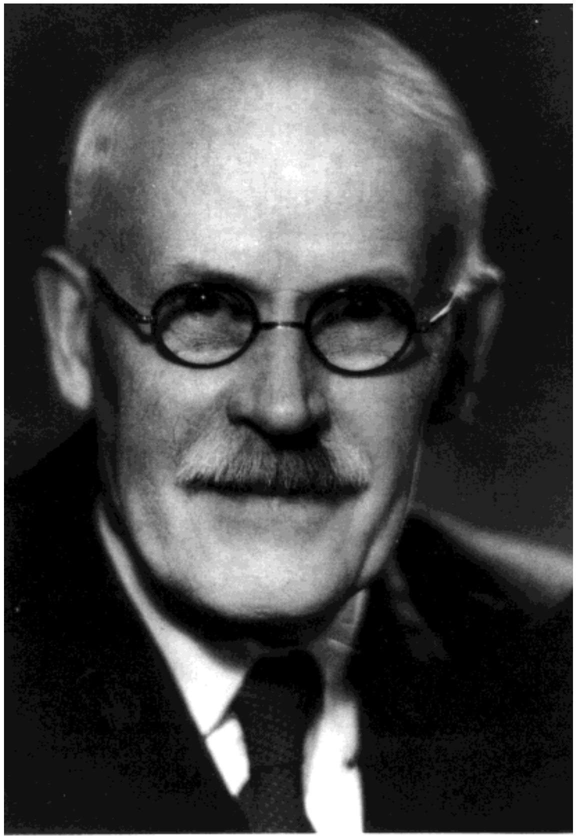
- Also, we live close to a church and whenever there is a wedding the traffic is more intense than usual. This means that on wedding days a higher value of λ is more likely than on non-wedding days.

- Moreover, a fellow that had been living in our flat before us did the same exercise and told us that he obtained $\lambda=4.2\pm0.5$.

- Is there a way to account for all this information in our study?

The Bayesian way

Bruno de Finetti (1906-1985)



Harold Jeffreys (1891-1989)

What is probability?

- Probability is a modern concept first discussed in a correspondence between Blaise Pascal and Pierre de Fermat in 1654

- There is no unique definition of probability, statisticians are divided into different schools with contrasting views

- Classic definition: The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible. (Pierre-Simon de Laplace, *Théorie analytique des probabilités*, 1812)

- This is based on the "principle of insufficient reason" (or principle of indifference) which states that when cases are only distinguishable by their name they should be assigned the same probability

What is probability?

- Frequentist: the long-run expected frequency of occurrence of a random event

- Axiomatic: given a sample space Ω, a σ-algebra F of events E (a set of subsets of Ω), we call probability measure a real function on F such that $P(E)\geq0$, $P(\Omega)=1$, and for any countable series of pairwise disjoint events $P(E_1 \cup E_2 \cup \ldots \cup E_N)=P(E_1)+P(E_2)+\ldots+P(E_N)$. These are known as Kolmogorov axioms.

- Bayesian: a measure of the degree of belief (the plausibility of an event given incomplete knowledge)

Reasoning with beliefs

- There is 90% chance that today it will rain
- There is a 30% chance that my favourite football team will win the league this year
- There is a 10% chance that I will fail the observational cosmology examination
- There is a 0.1% chance that I will die before being 30
- There is 68.3% chance that H_0 lies between 67 and 73 km/s/Mpc

De Finetti’s game
Can you measure degree of belief?

Suppose we are on a trip and you say that you are “pretty sure” you locked the door of your flat. I want to determine how sure you are.

- I offer you to play a game: I propose you to draw a marble from a bag containing 95 red and 5 blue marbles. If you pick at random a red marble, I give you one million euros. Alternatively, I offer you to go back home and check the door. If you choose this option and the door is locked indeed, I give you one million euros.

- If you choose to pick a marble, it means that your degree of belief is lower than 95%

- I can then propose many other rounds of the game by progressively reducing the fraction of red marbles until you choose to go back. This would measure your degree of belief.

Posterior probability

$P(x|\theta)$: old name "direct probability"
It gives the probability of contingent events (i.e. observed data) for a given hypothesis (i.e. a model with known parameters θ)

$L(\theta)=P(x|\theta)$: modern name "likelihood function" or simply "likelihood"
It quantifies the likelihood that the observed data would have been observed as a function of the unknown model parameters (it can be used to rank the plausibility of model parameters but it is not a probability density for θ)

$P(\theta|x)$: old name "inverse probability"
modern name "posterior probability"
Starting from observed events and a model, it gives the probability of the hypotheses that may explain the observed data (i.e. of the unknown model parameters)

Rev. Thomas Bayes (1702-1761)

Bayes theorem

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

Prior probability for the parameters (what we know before performing the experiment)

Posterior probability for the parameters given the data

Likelihood function

$$p(x \mid \theta) = L(x \mid \theta)$$

Evidence (normalization constant useful for Bayesian model selection)

$$p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta$$

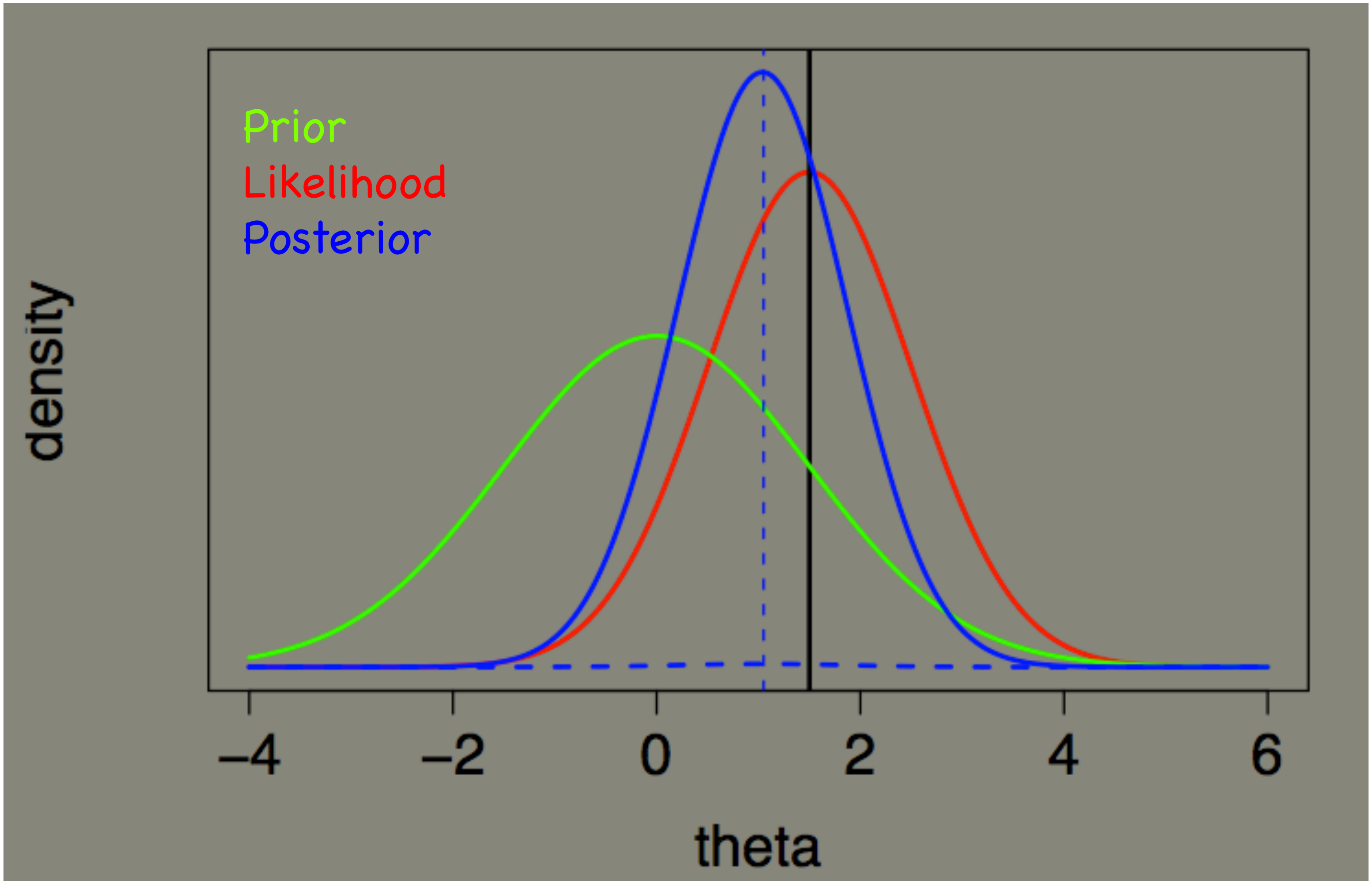
Pierre Simon (Marquis de) Laplace (1749-1827)

Bayes theorem

Let us re-write Bayes theorem emphasizing that all the probabilities that we generically called “p” are actually different functions:

$$P(\theta \mid x) = \frac{L(x \mid \theta)\ \pi(\theta)}{E(x)}$$

Bayes theorem visually



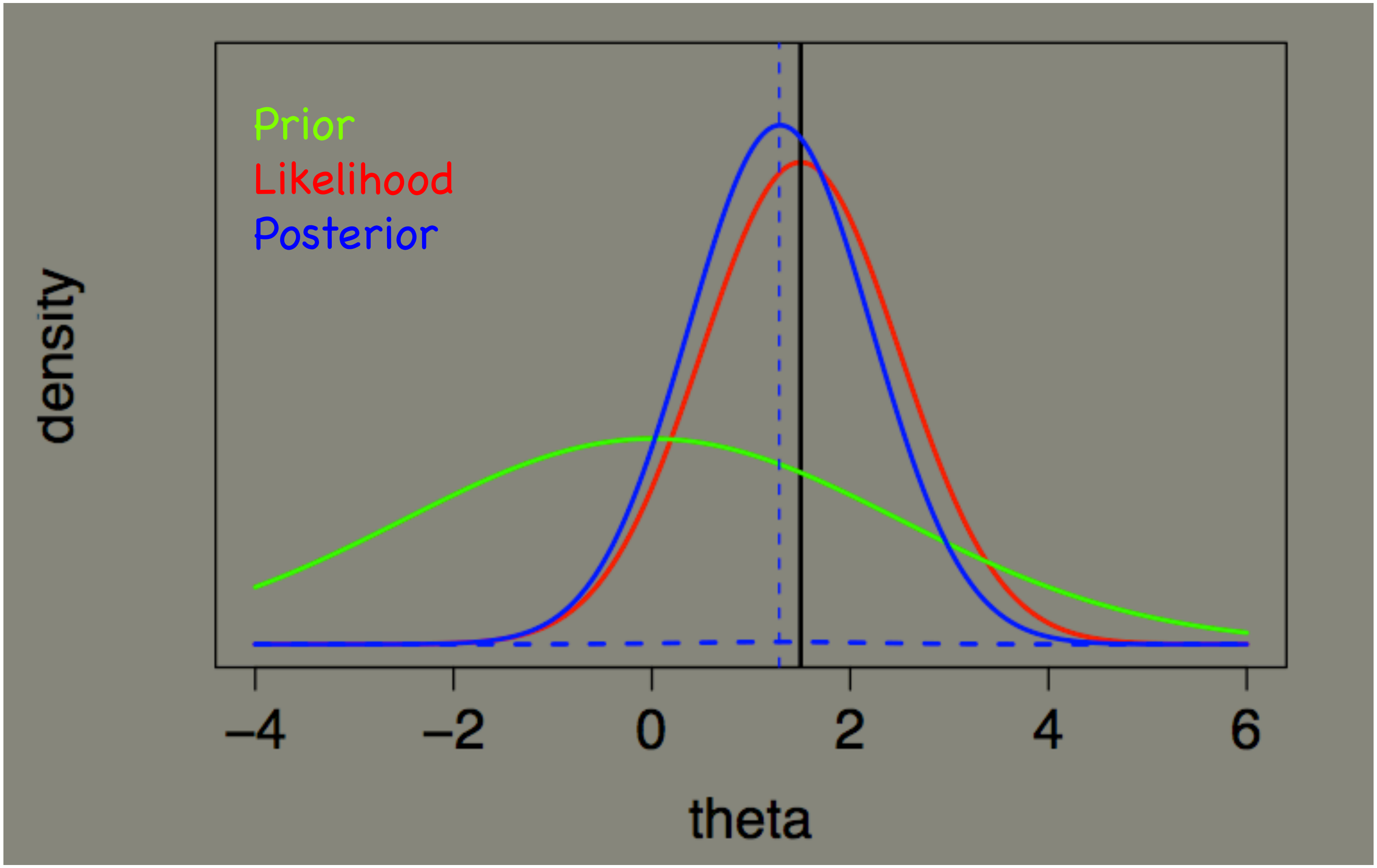
Priors

- Non-informative: if it has a minimal impact on the posterior distribution (i.e. it is “flat” with respect to the likelihood function). Non-informative priors are also called “vague”, “diffuse” or “flat”.

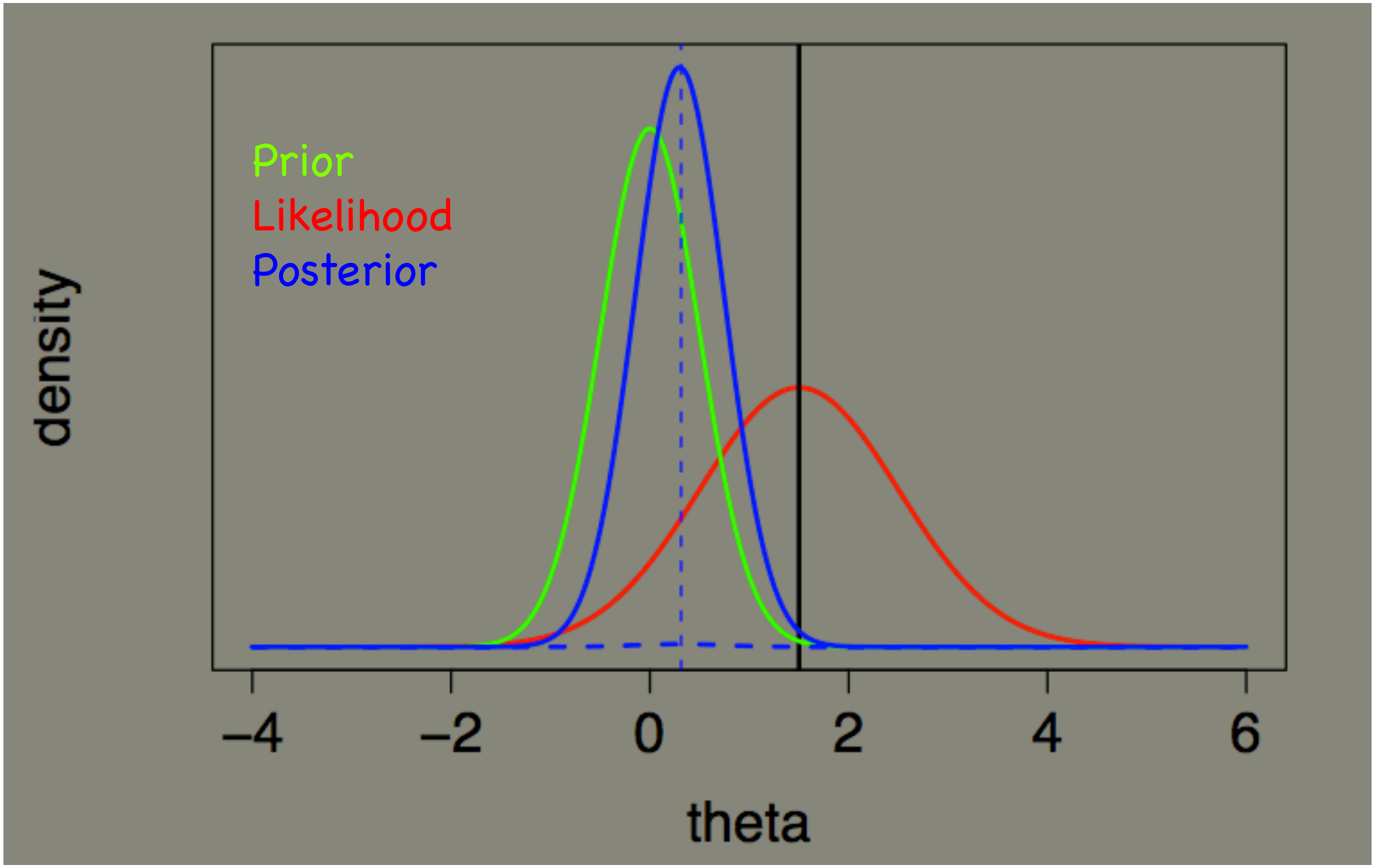
- Improper: if $\int \pi(\vartheta) d\theta = \infty$ e.g. uniform prior on the real line. Generally leads to a proper posterior but, if also the posterior is improper, inference is invalid.

- Informative: a prior which is not dominated by the likelihood. Must be handled with care in actual practice. On the other hand, illustrates the power of the Bayesian method: information gathered from previous study, past experience, or expert opinion can be combined with current information in a natural way

Not very informative prior



Very informative prior



Bayesian estimation

- In the Bayesian approach to statistics, population parameters are associated with a posterior probability which quantifies our DEGREE OF BELIEF in the different values

- Sometimes it is convenient to introduce estimators obtained by minimizing the posterior expected value of a loss function

- For instance one might want to minimize the mean square error, which leads to using the mean value of the posterior distribution as an estimator

- If, instead one prefers to keep functional invariance, the median of the posterior distribution has to be chosen

- Remember, however, that whatever choice you make is somewhat arbitrary as the relevant information is the entire posterior probability density.

Estimation: frequentist vs Bayesian

- Frequentist: there are TRUE population parameters that are unknown and can only be estimated by the data

- Bayesian: only data are real. The population parameters are an abstraction, and as such some values are more believable than others based on the data and on prior beliefs.

Confidence vs. credibility intervals

- Confidence intervals (Frequentist): measure the variability due to sampling from a fixed distribution with the TRUE parameter values. If I repeat the experiment many times, what is the range within which 95% of the results will contain the true values?

- Credibility interval (Bayesian): For a given significance level, what is the range I believe the parameters of a model can assume given the data we have measured?

- They are profoundly DIFFERENT things even though they are often confused. Sometimes practitioners tend use the term "confidence intervals" in all cases and this is ok because they understand what they mean but this might be confusing for the less experienced readers of their papers. PAY ATTENTION!