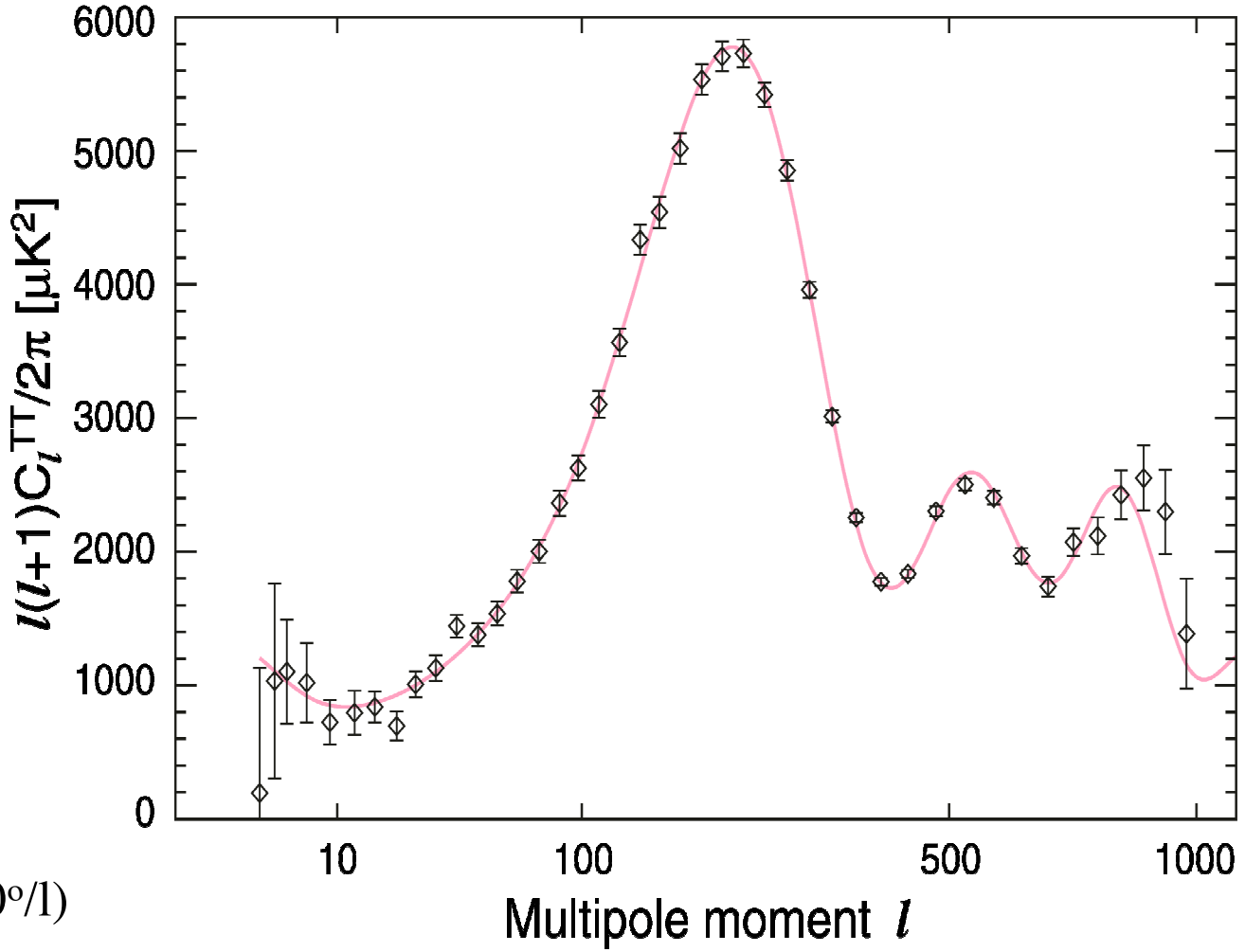
Parameter estimation and forecasting

Cristiano Porciani

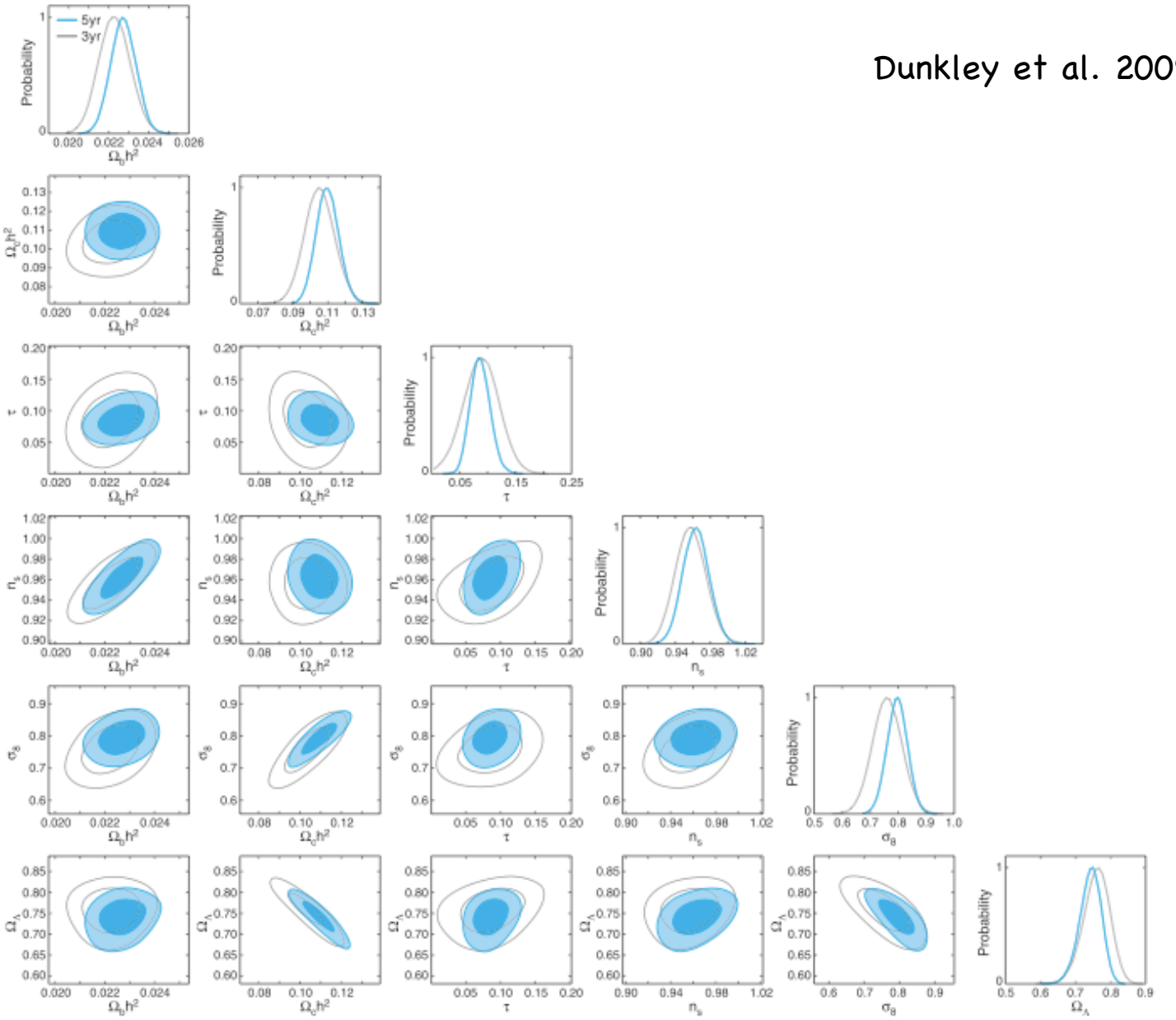
AIfA, Uni-Bonn

Questions?

Temperature fluctuations

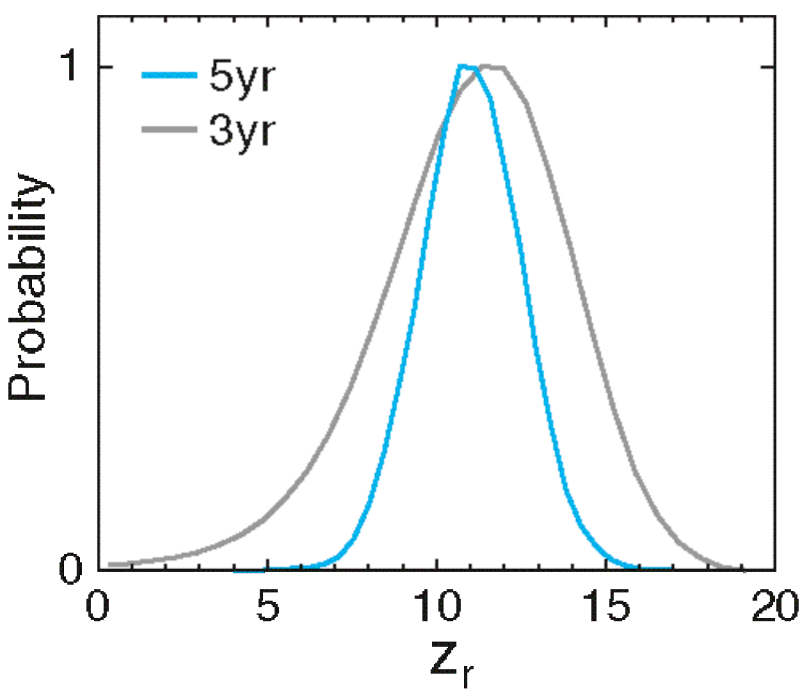


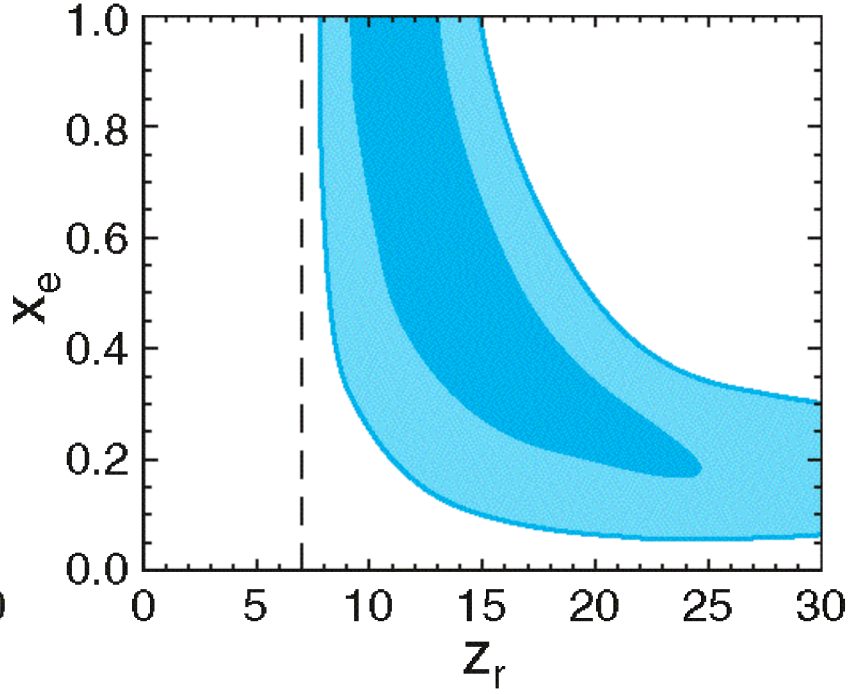
Variance at multipole l (angle ~180°/l)



Dunkley et al. 2009

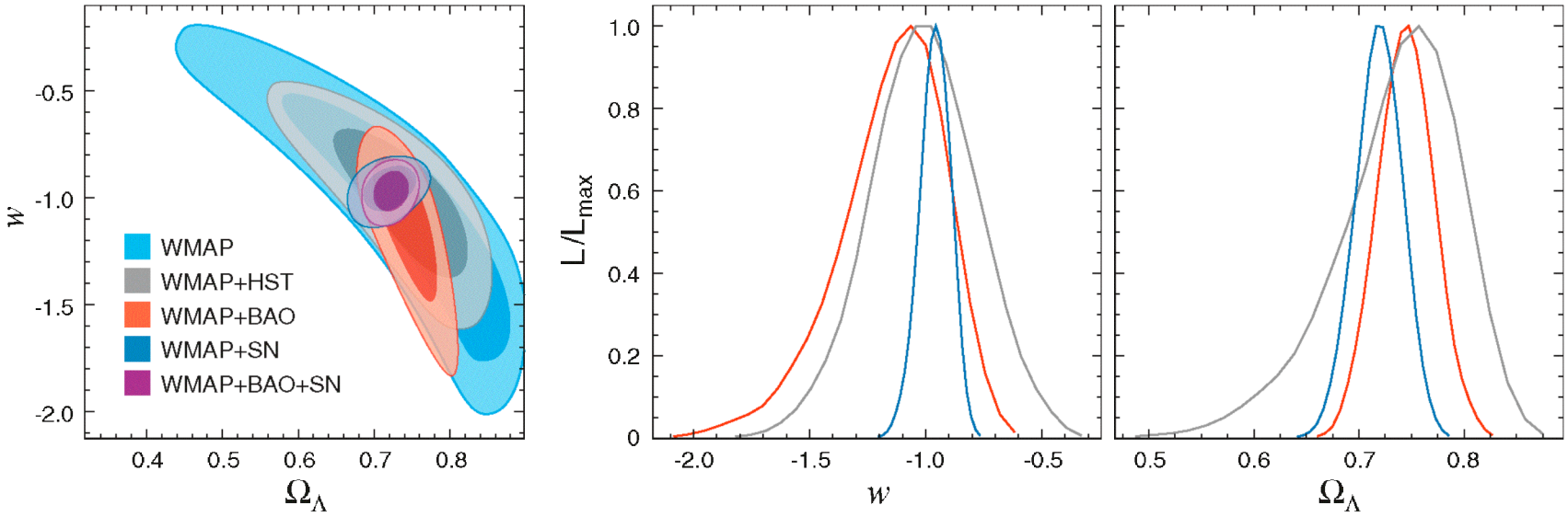
Dunkley et al. 2009





The current state of the art

Bennett 2006



What is the meaning of these plots?

- What's the difference between the 1D and the 2D plots?
- What is a confidence interval?
- What is a credibility interval?
- What does marginalization mean?
- What's the difference between the frequentist and the Bayesian interpretation of statistics?

What is probability?

- Frequentist: the long-run expected frequency of occurrence

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

Estimation

- 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

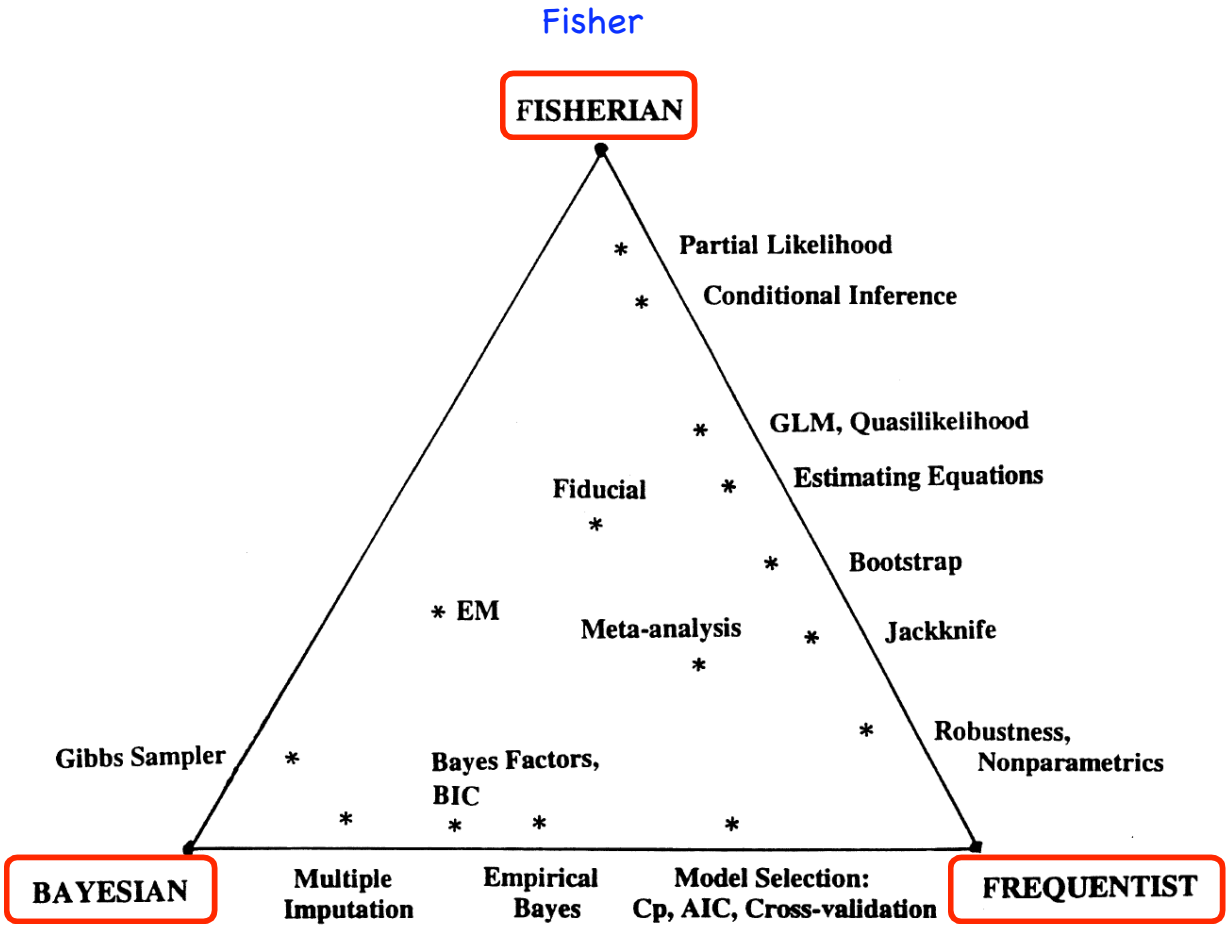
- Confidence intervals (Frequentist): measure the variability due to sampling from a fixed distribution with the TRUE parameter values. If I repeat the experiment many times, what is the range within which 95% of the results will contain the true values?

- 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!

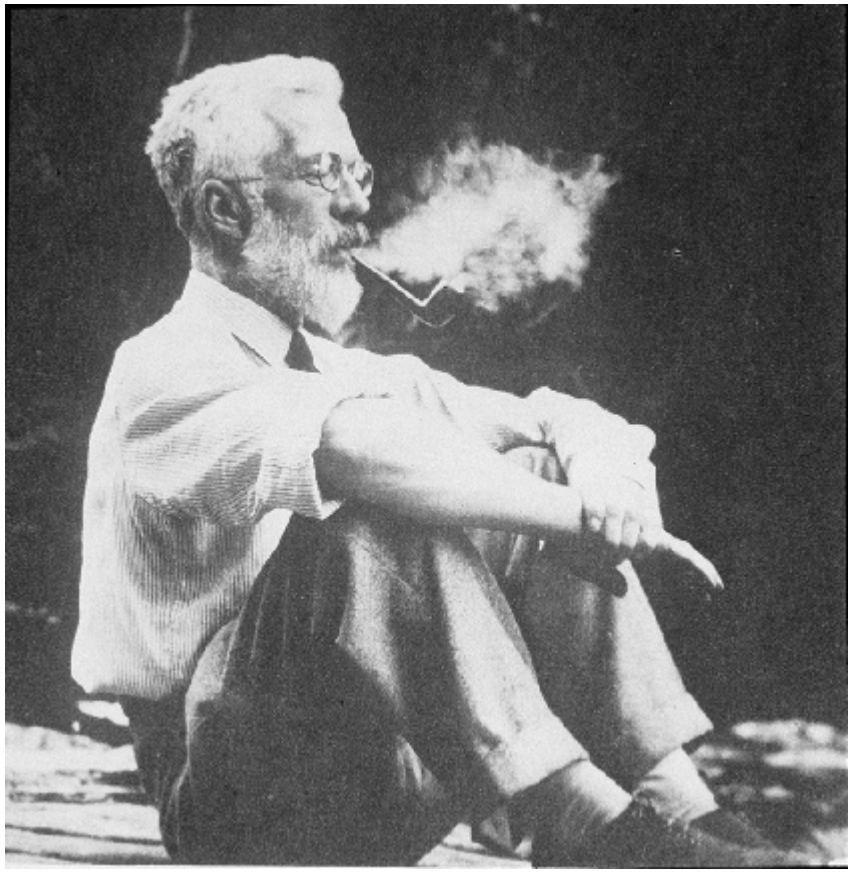
The coordinates of statistics

Bradley Efron's triangle (1998)



R.A. Fisher (1890-1962)

"Fisher was to statistics what Newton was to Physics" (R. Kass)



"Even scientists need their heroes, and R.A. Fisher was the hero of 20th century statistics" (B. Efron)

Fisher's concept of likelihood

- "Two radically distinct concepts have been confused under the name of 'probability' and only by sharply distinguishing between these can we state accurately what information a sample does give us respecting the population from which it was drawn." (Fisher 1921)

- "We may discuss the probability of occurrence of quantities which can be observed…in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of the hypotheses…We may ascertain the likelihood of the hypotheses…by calculation from observations:…to speak of the likelihood…of an observable quantity has no meaning." (Fisher 1921)

- "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed." (Fisher 1922)

The Likelihood function

- 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

$$\frac{\partial L(\theta_1,\ldots,\theta_k,x_1,\ldots,x_n)}{\partial \theta_i}=0 \qquad \text{or} \qquad \frac{\partial \ln L(\theta_1,\ldots,\theta_k,x_1,\ldots,x_n)}{\partial \theta_i}=0$$

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

Properties of MLE's

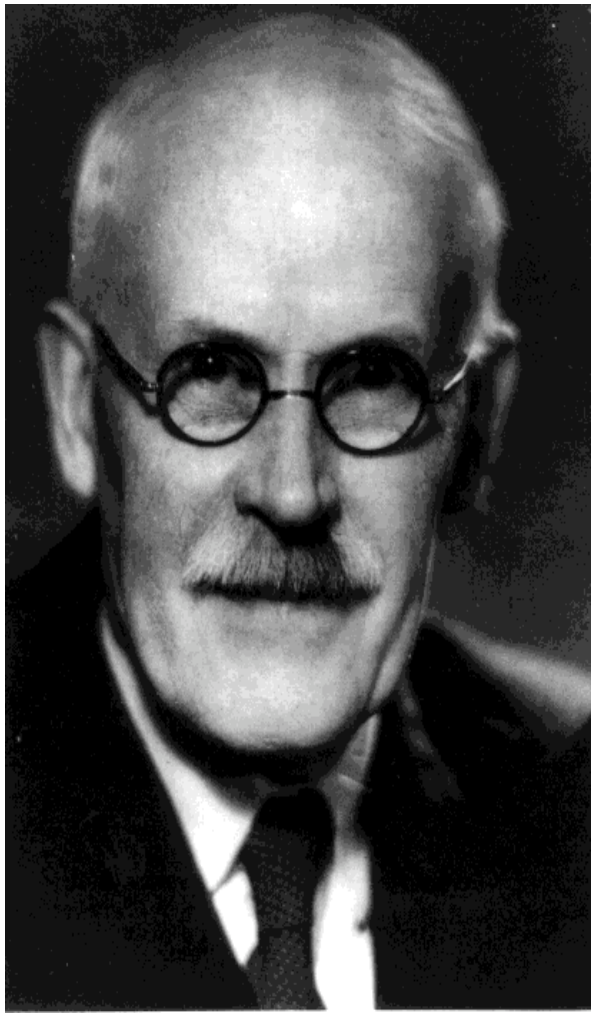
As the sample size increases to infinity (under weak regularity conditions):

- MLE's become asymptotically efficient and asymptotically unbiased
- MLE's asymptotically follow a normal distribution with covariance matrix equal to the inverse of the Fisher's information matrix

However, for small samples,

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

The Bayesian way



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)

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

Bayesian estimation

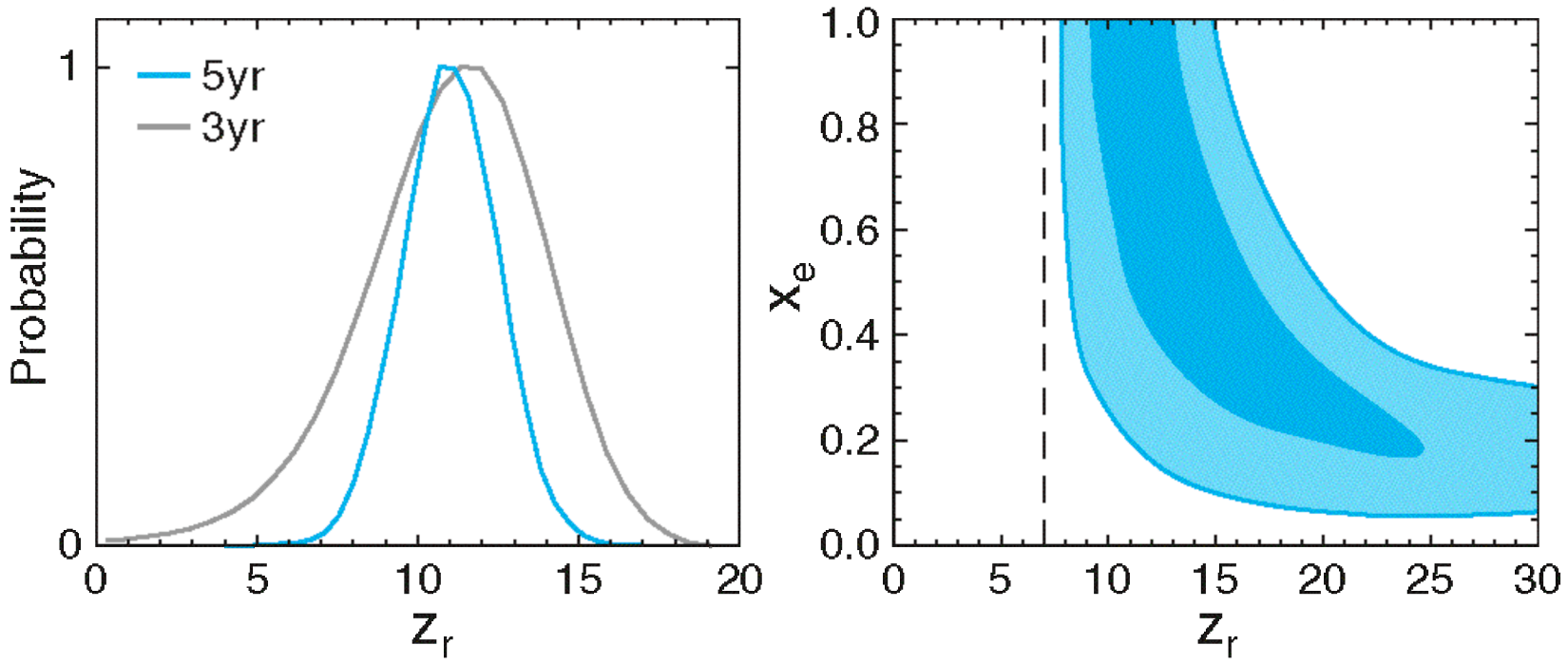
- In the Bayesian approach to statistics, population parameters are associated with a posterior probability which quantifies our DEGREE OF BELIEF in the different values
- Sometimes it is convenient to introduce estimators obtained by minimizing the posterior expected value of a loss function
- For instance 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.

Marginalization

Marginal probability: probability of a given parameter regardless of the value of the others

$$p(\vartheta_2 \mid x) = \int p(\theta \mid x)\, d\theta_1 d\theta_3 \ldots d\theta_n$$

Dunkley et al. 2009



Computing likelihoods and posteriors

- For 2 or 3 parameters, a grid is usually possible and marginalization is performed by integrating along each axis of the grid

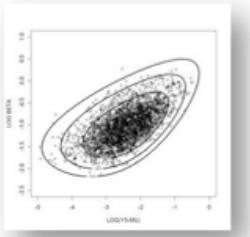
- For a number of parameters >>2 it is NOT feasible to have a grid (e.g. 10 point in each parameter direction, 12 parameters = 10^{12} likelihood evaluations!!!)

- Special numerical methods are used to sample the posterior distribution for the parameters

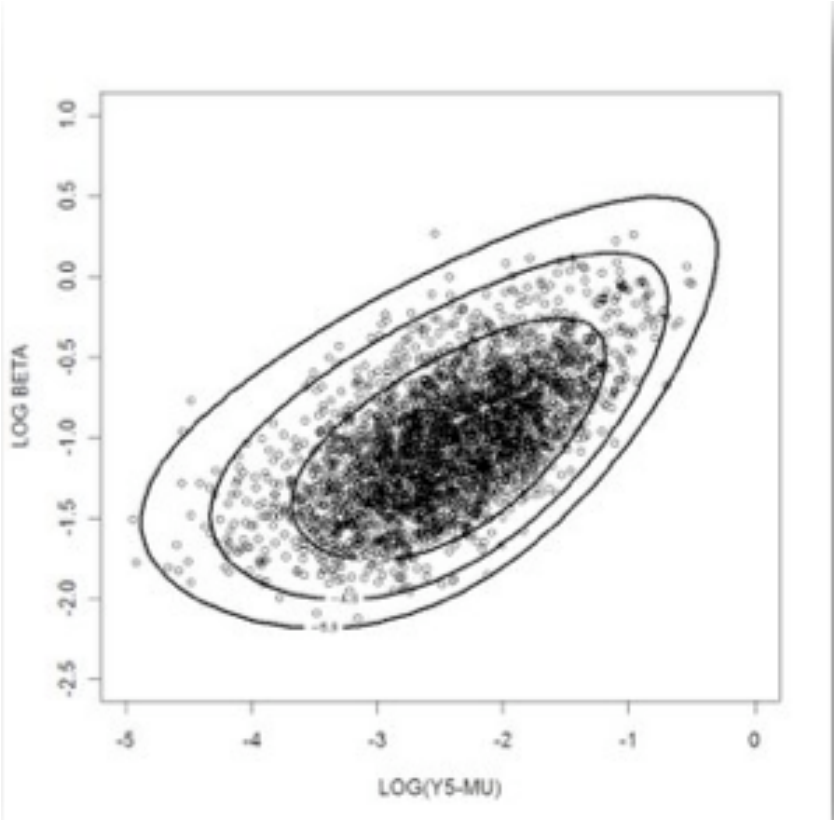
Markov Chain Monte Carlo

MCMC

Aim of MCMC: generate a set of points in the parameter space whose distribution function is the same as the target density.



MCMC follows a Markov process - i.e. the next sample depends on the present one, but not on previous ones.


LOG BETA
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-5
-4
-3
-2
-1
0
LOG(Y5-MU)

MCMC algorithm

- Choose a random initial starting point in parameter space, and compute the target density

- Repeat:
- ✓ Generate a step in parameter space from a proposal distribution, generating a new trial point for the chain.
- ✓ Compute the target density at the new point, and accept it or not with the Metropolis-Hastings algorithm.
- ✓ If the point is not accepted, the previous point is repeated in the chain.
- End Repeat

Metropolis-Hastings algorithm

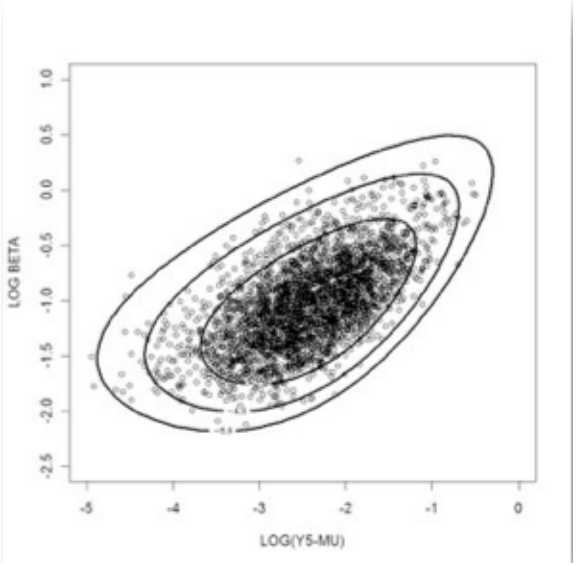
$$p(acceptance) = min\left[1, \frac{p(\theta^*)q(\theta^*|\theta)}{p(\theta)q(\theta|\theta^*)}\right]$$

Metropolis algorithm (special case): (for symmetric proposal distributions)

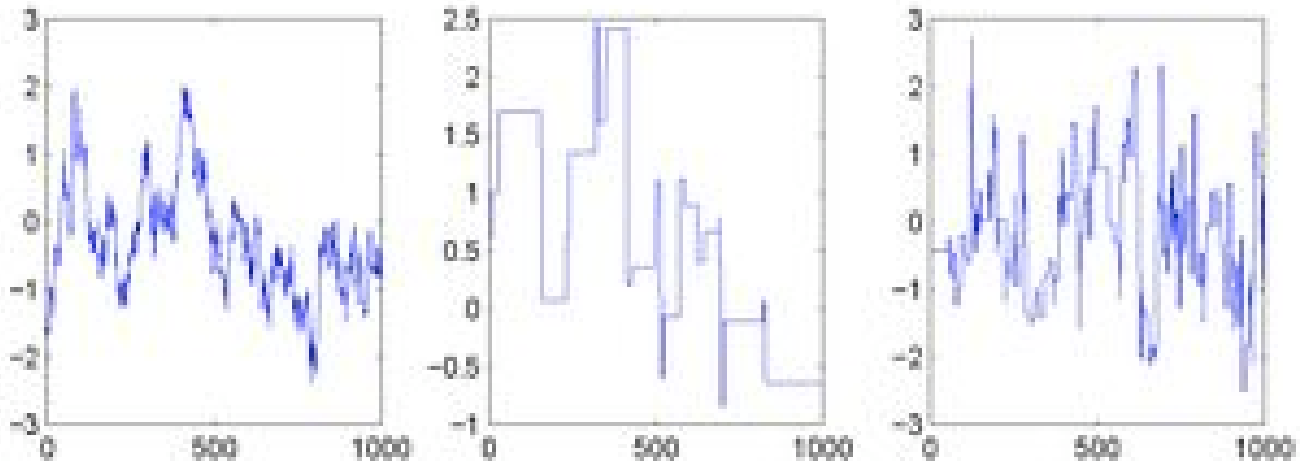
$$min\left[1, \frac{p(\theta^*)}{p(\theta)}\right]$$

The proposal distribution

- Too small steps, and it takes long time to explore the target
- Too large steps and almost all trials are rejected
- q ~ "Fisher size" is good

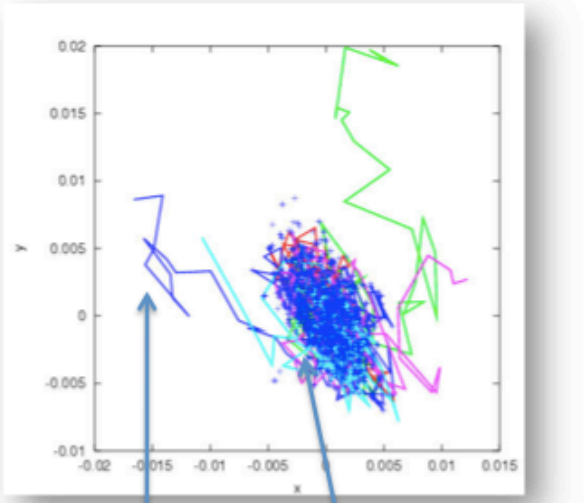


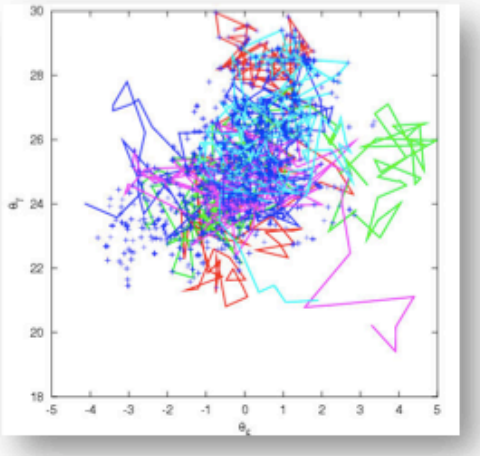
The proposal distribution



It can be shown thrat the optimal mean acceptance rate is 0.234

Burn-in and convergence



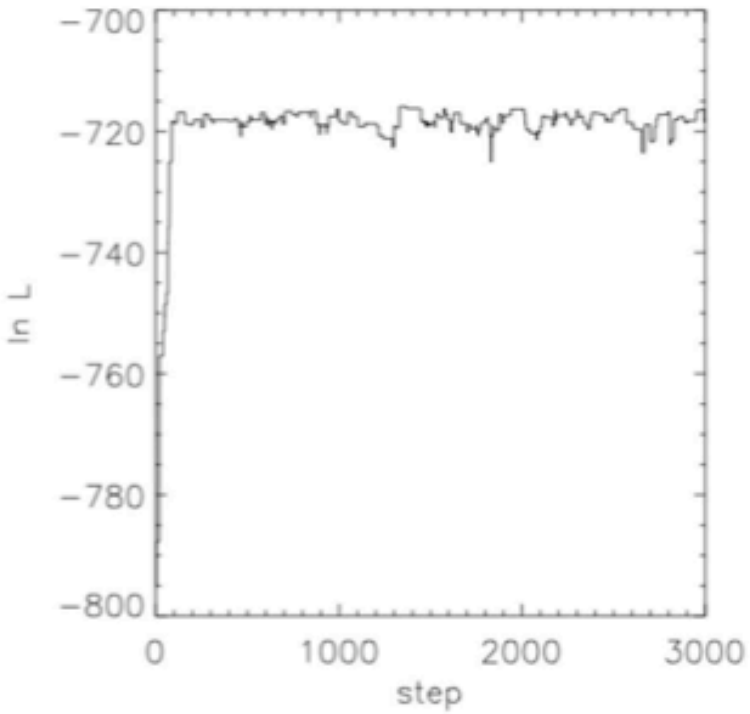


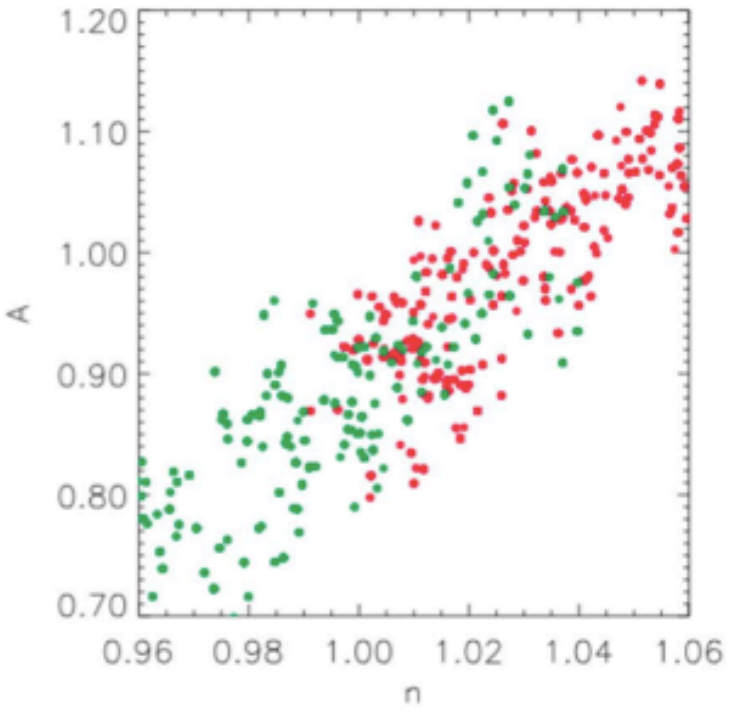
You must use a convergence test.
Gelman-Rubin test is most common (see notes)

"Burn-in"

Points are correlated

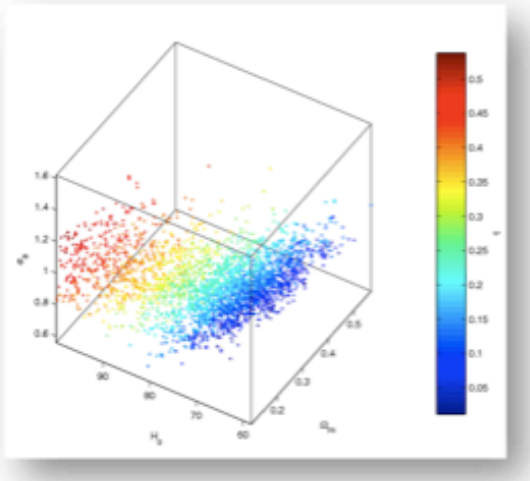
Unconverged chains

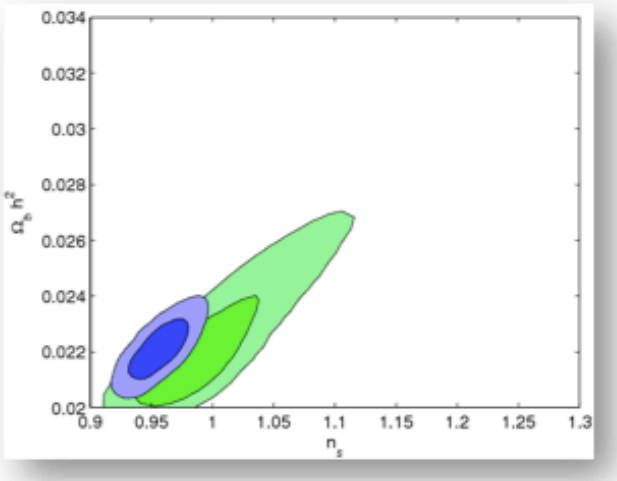




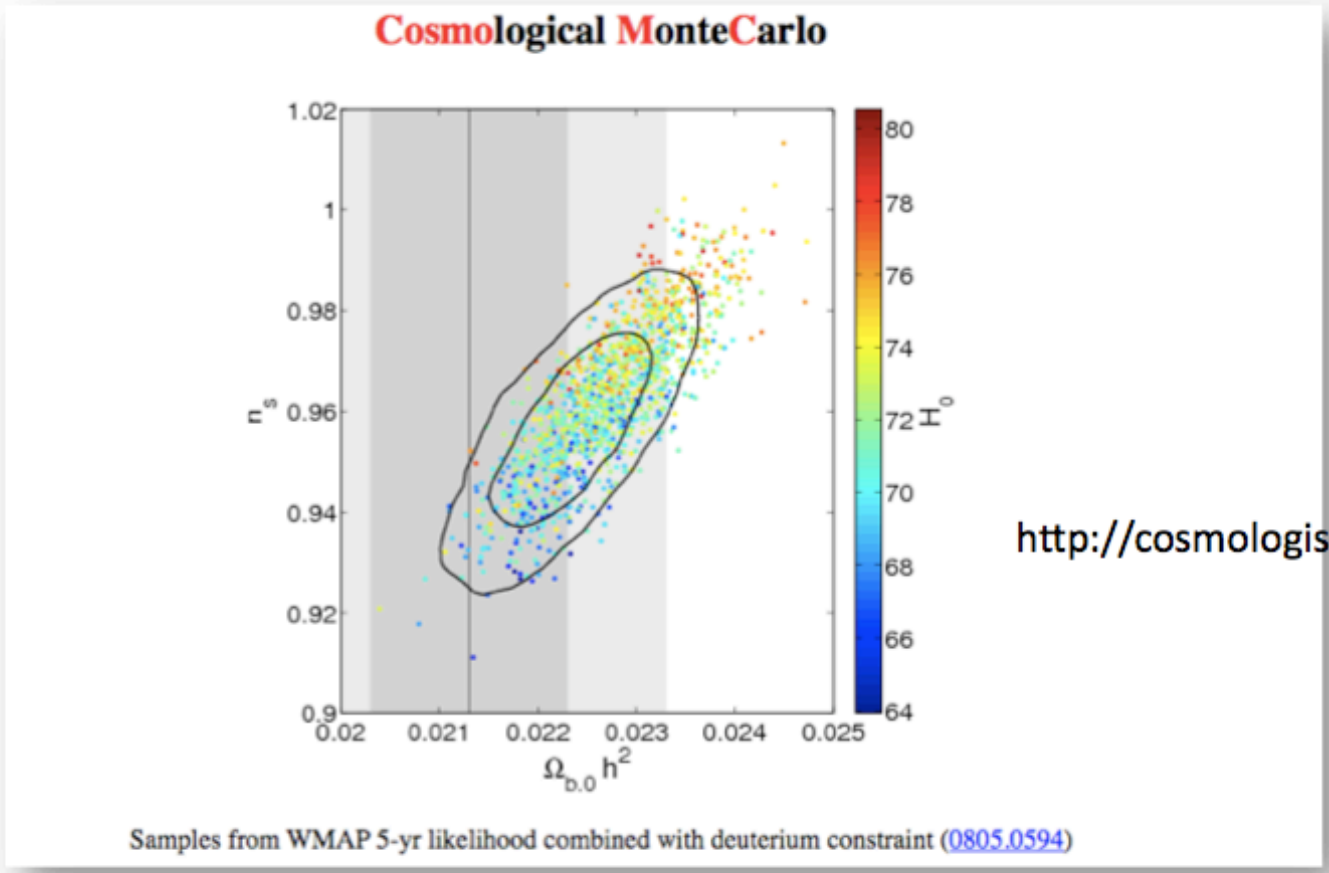
Marginalisation

- Marginalisation is trivial
  - Each point in the chain is labelled by all the parameters
  - To marginalise, just ignore the labels you don't want





CosmoMC



http://cosmologist.info/cosmomc/

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

Fisher information matrix

$$F_{ij} = \left\langle \frac{\partial^2 \Lambda}{\partial\theta_i \, \partial\theta_j} \right\rangle, \qquad where \qquad \Lambda = -\ln L$$

The Cramer-Rao inequality states that, for any unbiased estimator:

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

For Gaussian data, the Fisher matrix is

$$F_{ij} = \frac{1}{2} Tr\left[C^{-1} C_{,i} \, C^{-1} C_{,j} + C^{-1} M_{ij} \right] \quad with \quad M_{ij} = \mu_{,i} \, \mu_{,j}^T + \mu_{,j} \, \mu_{,i}^T$$

Figure of merit

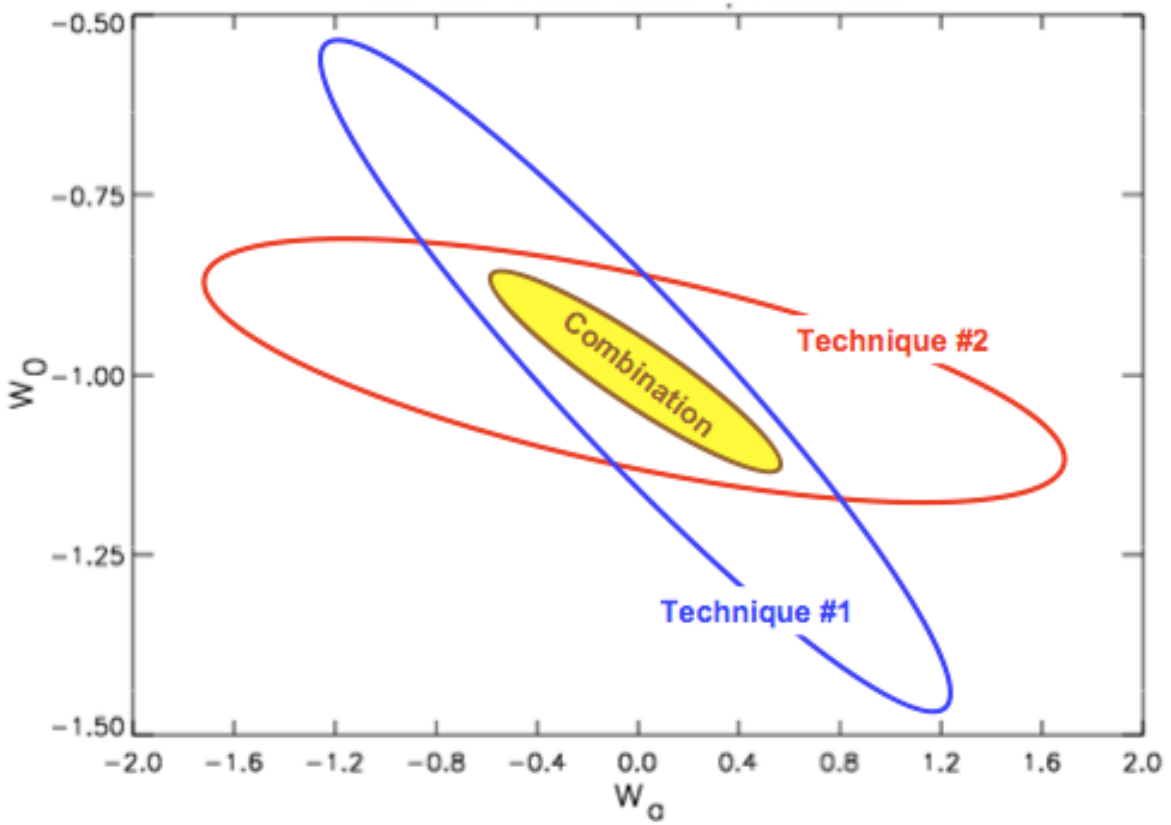
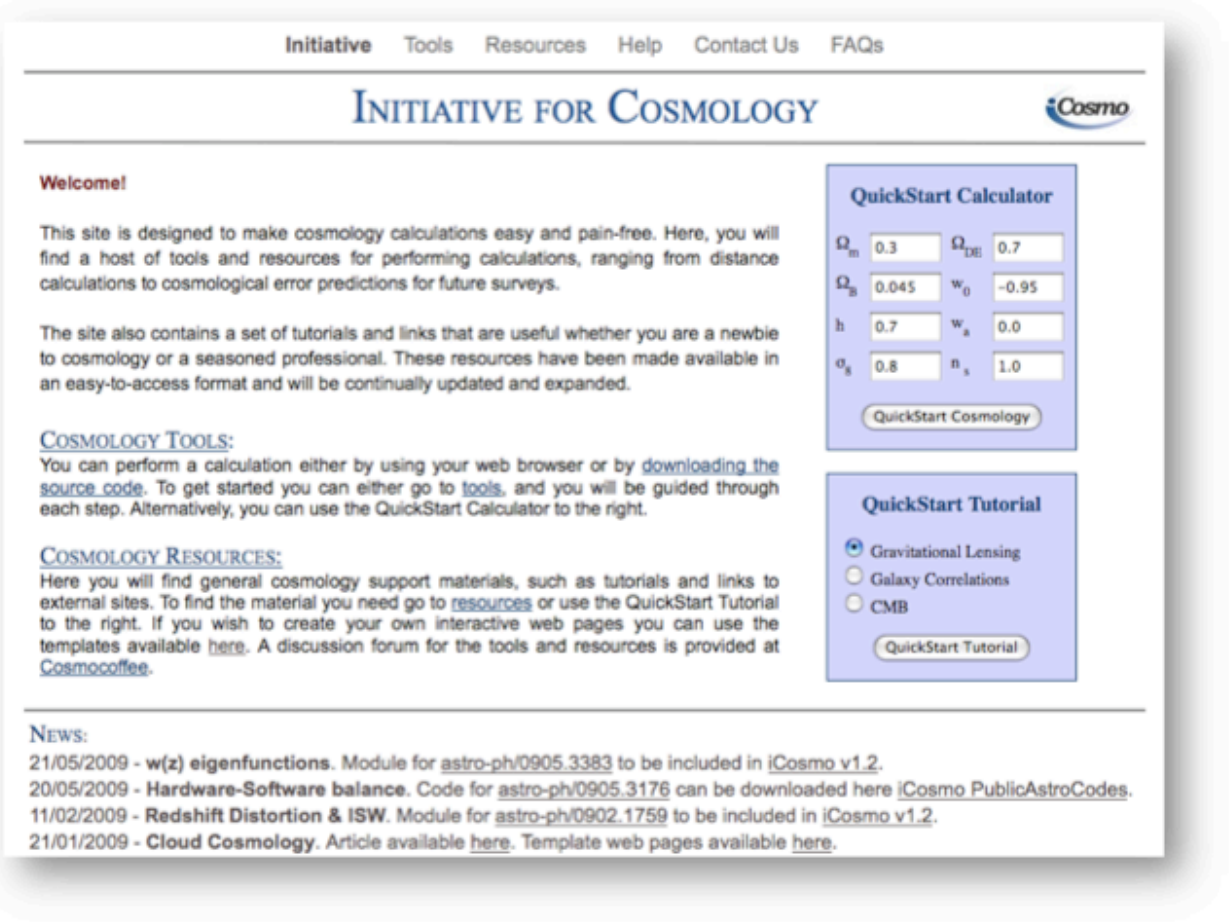
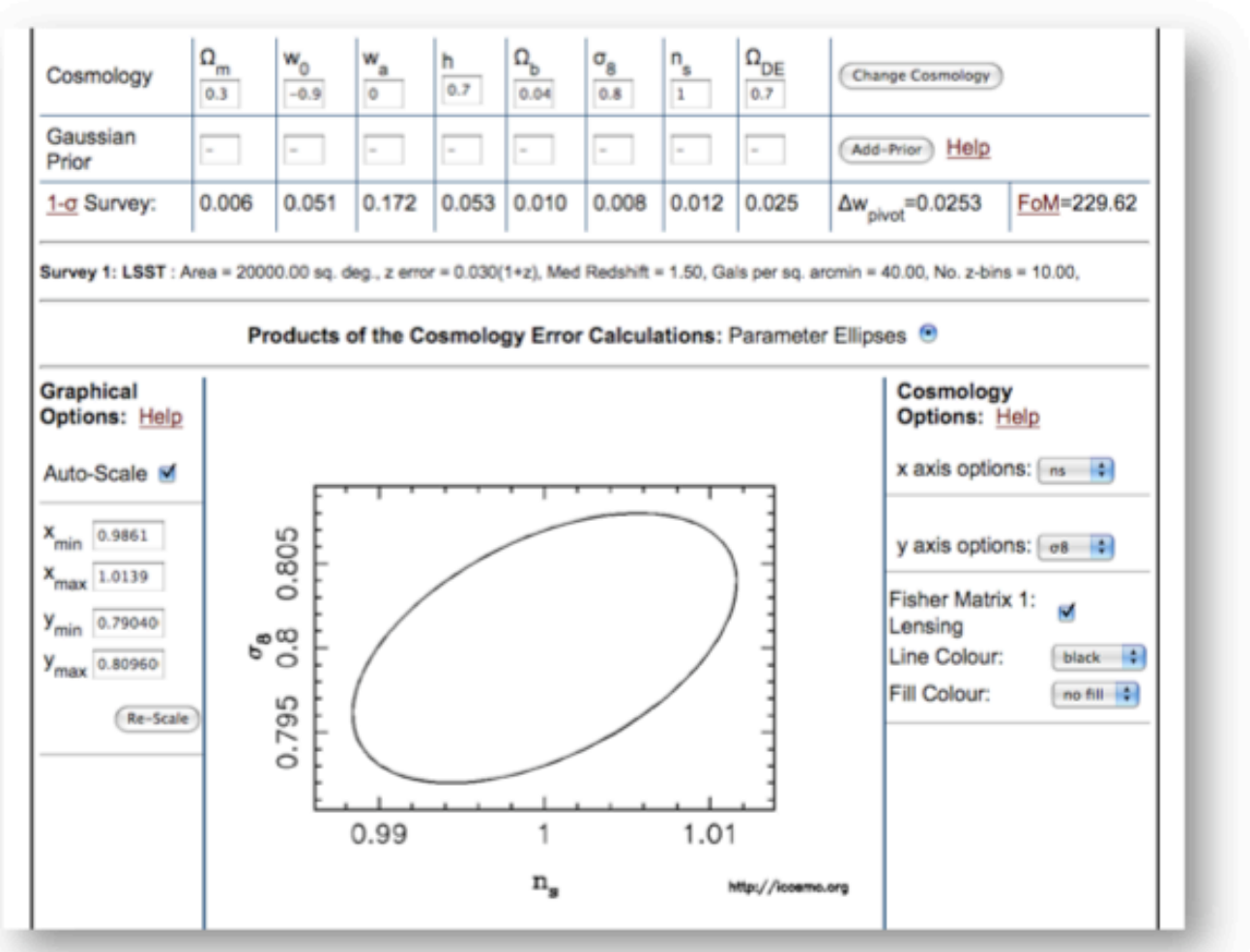


Figure of merit = 1 / (area of the ellipse)

iCOSMO.org


Initiative Tools Resources Help Contact Us FAQs
INITIATIVE FOR COSMOLOGY
iCosmo
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 Cosmology
QuickStart Tutorial
Gravitational Lensing
Galaxy Correlations
CMB
QuickStart Tutorial
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



Hypothesis testing

(Neyman & Pearson 1933)

1. State the null hypothesis H_0 (usually, that the observations are due to pure chance). This hypothesis is tested against possible rejection under the assumption that it is true.

2. Formulate an alternate hypothesis H_A which is mutually exclusive with H_0 (usually, that the observation are due to a combination of a real effect and chance)

3. Identify a test statistic to assess the truth of the null hypothesis and evaluate the statistic using the sample data.

4. Assuming that the null hypothesis were true, compute the probability p that the test statistic assumes a value at least as significant as the one observed. This requires knowledge of the PDF of the statistic (the sampling distribution).

5. Draw a conclusion by comparing p with a significance value (or confidence level) $1-\alpha$ ($0 \leq \alpha \leq 1$). If $p<1-\alpha$ the observed effect is statistically significant and the null hypothesis is rejected in favour of H_A. If $p> \alpha$ there is not enough evidence to reject H_0 .