ICIC Data Analysis Workshop: the Bayesics

Alan Heavens

Imperial College London

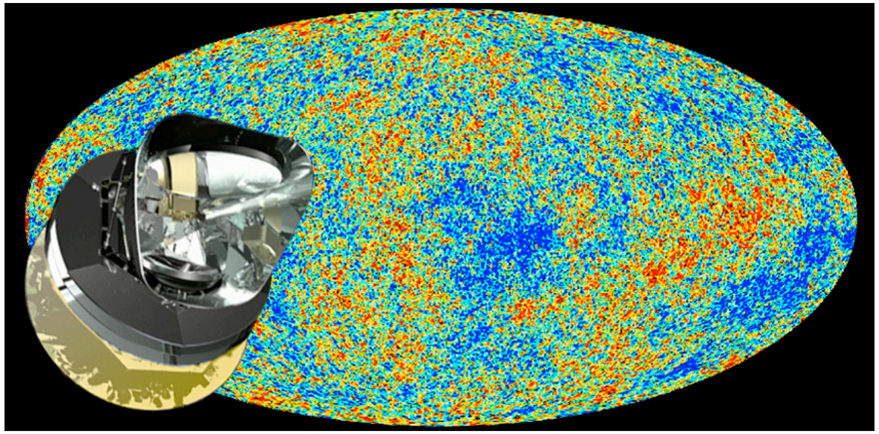
ICIC Data Analysis Workshop

11 September 2013

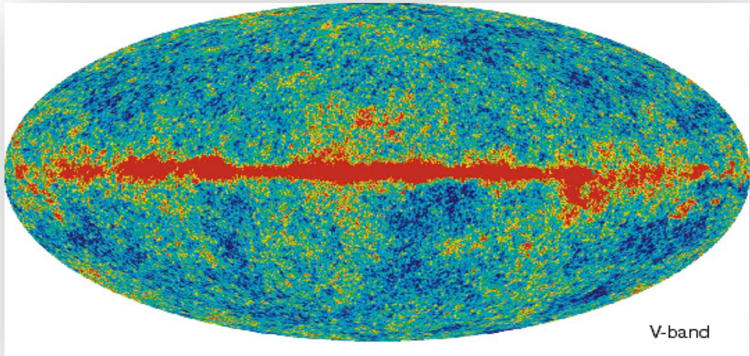
ICIC

Outline

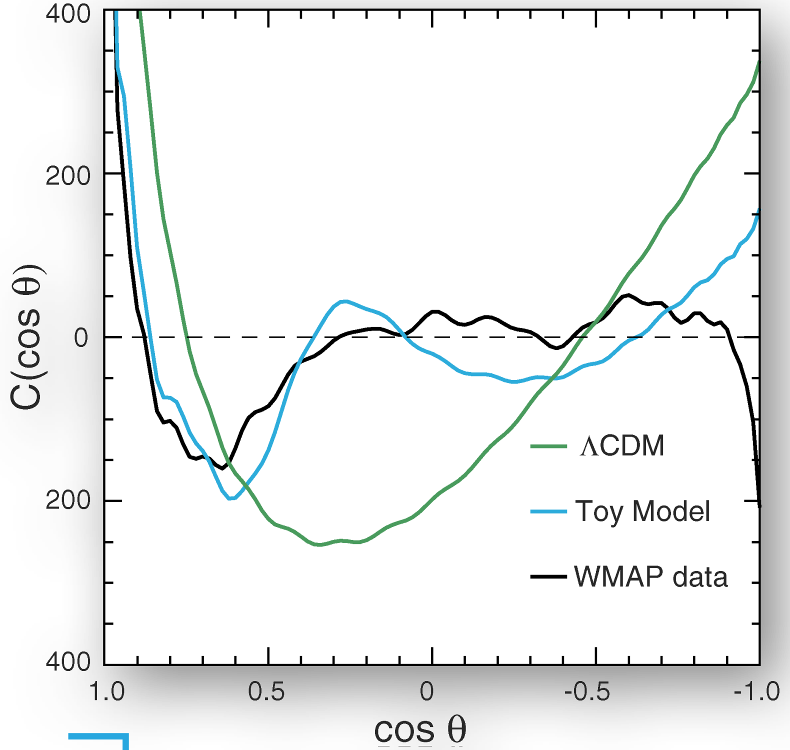
- Inverse problems: from data to theory
- Probability review, and Bayes' theorem
- Parameter Estimation
- Priors
- Marginalisation
- Errors
- Error prediction and experimental design: Fisher Matrices



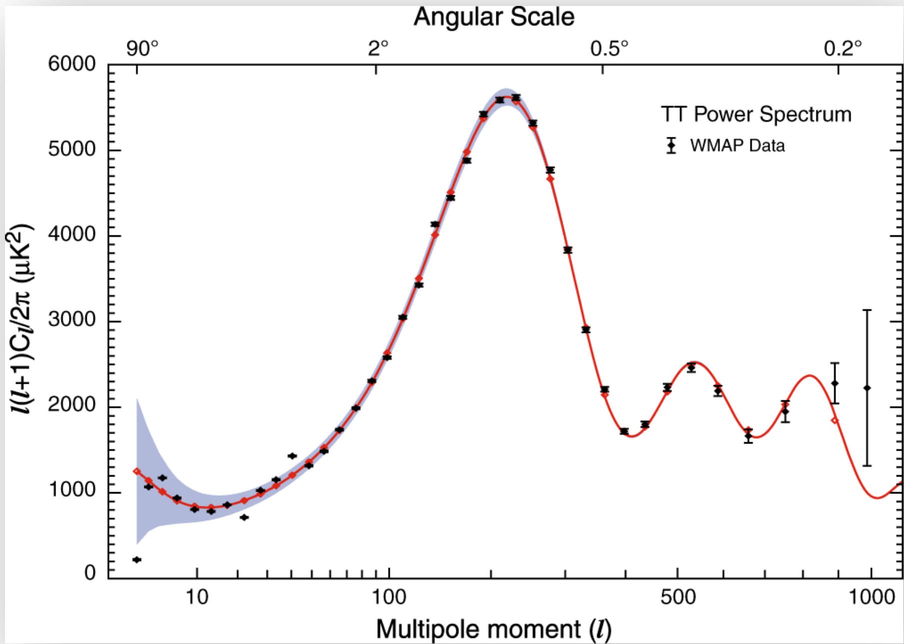
LCDM fits the WMAP data well.



Correlation Function $\langle \frac{\Delta T}{T}(\phi) \frac{\Delta T}{T}(\phi + \theta) \rangle$



Power Spectrum C_l



ICIC

Inverse problems

- Most cosmological problems are *inverse problems*, where you have a set of data, and you want to infer something.
- This is harder than predicting the outcomes when you know the model and its parameters
- Examples
  - Hypothesis testing
  - Parameter estimation
  - Model selection

ICIC

Examples

- Hypothesis testing
  - Is the CMB radiation consistent with (initially) gaussian fluctuations?
- Parameter estimation
  - In the Big Bang model, what is the value of the matter density parameter?
- Model selection
  - Do cosmological data favour the Big Bang theory or the Steady State theory?
  - Is the gravity law General Relativity or a different higher-dimensional theory?

ICIC

What is probability?

- Frequentist view: p describes the relative *frequency of outcomes* in infinitely long trials
- Bayesian view: p expresses our *degree of belief*
- Bayesian view is what we seem to want from experiments: e.g. *given the Planck data, what is the probability that the density parameter of the Universe is between 0.9 and 1.1?*
- Cosmology is in good shape for inference because we have decent model(s) with parameters – well-posed problem

ICIC

Bayes' Theorem

- Rules of probability:
- $p(x)+p(\text{not } x) = 1$ sum rule
- $p(x,y) = p(x|y)\, p(y)$ product rule
- $p(x) = \Sigma_k\, p(x,y_k)$ marginalisation
- Sum⟶ integral continuum limit (p=pdf)

$$p(x) = \int dy\, p(x,y)$$

- p(x,y)=p(y,x) gives *Bayes' theorem*

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

ICIC

p(x|y) is not the same as p(y|x)

- x = female, y=pregnant
- $p(y|x) = 0.03$
- $p(x|y) = 1$

The Monty Hall problem:

An exercise in using Bayes' theorem

You choose this one

?

Do you change your choice?

This is the Monty Hall problem

ICIC

Twist: After you make your first choice, an earthquake opens another door. Should you change your choice?

Bayes' Theorem and Inference

- If we accept *p* as a degree of belief, then what we often want to determine is*

$$p(\theta|x)$$

θ: model parameter(s), *x*: the data

To compute it, use Bayes' theorem $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$

Note that these probabilities are all conditional on a) prior information I, b) a model M

$$p(\theta|x) = p(\theta|x, I, M) \text{ or } p(\theta|x\, I\, M)$$

ICIC

*This is RULE 1: start by writing down what it is you want to know

RULE 2: There is no RULE n, n>1

Posteriors, likelihoods, priors and evidence

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

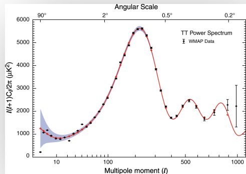
Posterior   Likelihood L   Evidence   Prior

Remember that we interpret these in the context of a model M, so all probabilities are conditional on M (and on any prior info I). E.g. $p(\theta) = p(\theta|M)$

The *evidence* looks rather odd – what is the probability of the data? For parameter estimation, we can ignore it – it simply normalises the posterior. If you need it,

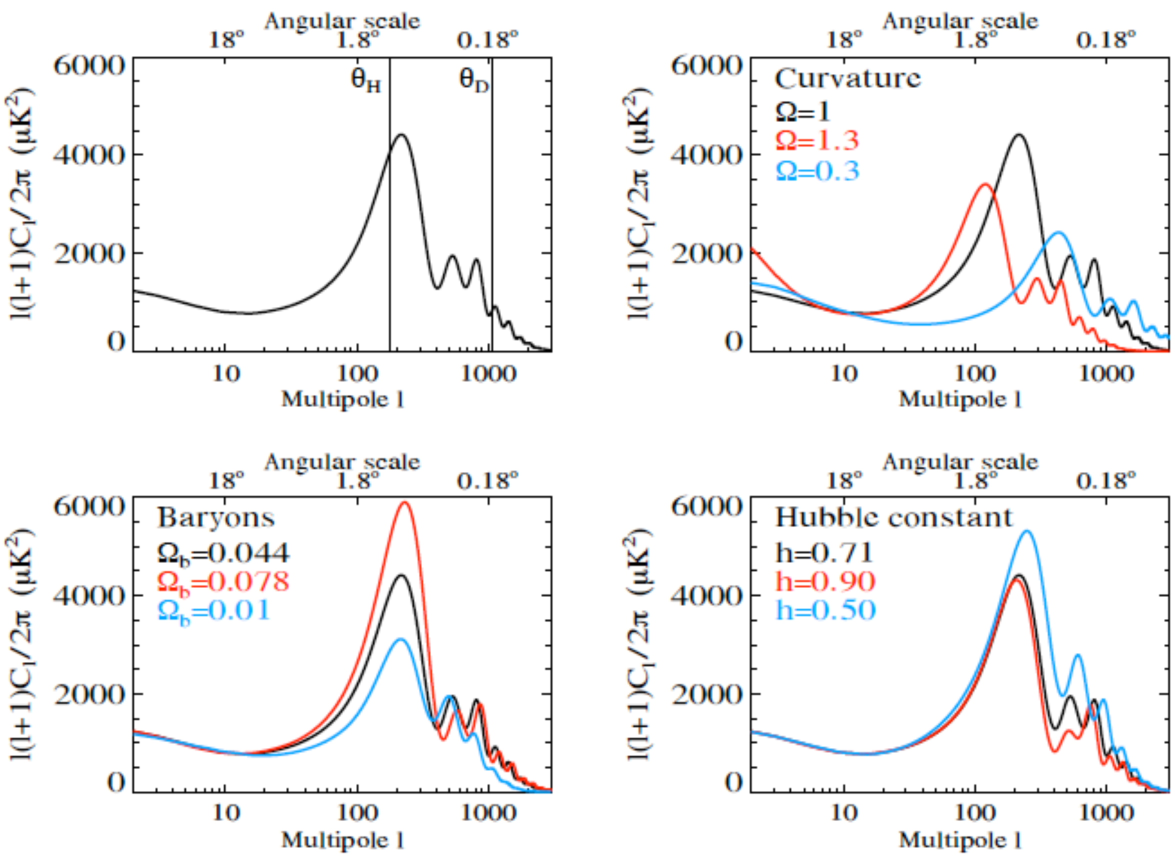
$$p(x) = \sum_k p(x|\theta_k)p(\theta_k) \text{ or } p(x) = \int d\theta \, p(x|\theta)p(\theta)$$

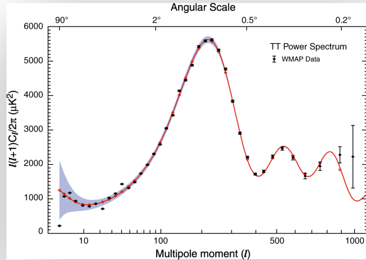
Noting that $p(x) = p(x|M)$ makes its role clearer.
In *model selection* (from M and M'), $p(x|M) \neq p(x|M')$



ICIC

Forward modelling $p(x|\theta)$





With noise properties we can predict the *Sampling Distribution* (the probability of obtaining a general set of data).

The *Likelihood* refers to the specific data we have) - it isn't a probability, strictly.

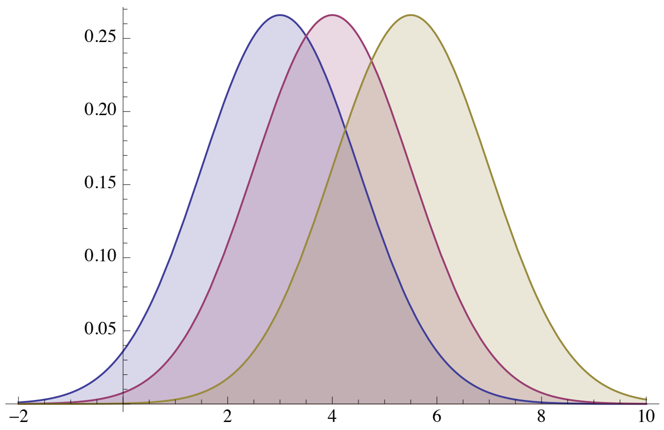
ICIC

Note: this is just the expectation value of x; the distribution is needed

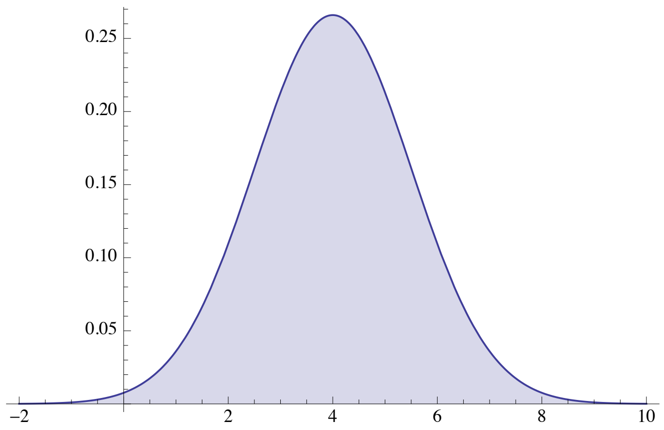
Case study: the mean

- Given a set of N independent samples $\{x_i\}$ from the same distribution, with gaussian dispersion σ, what is the mean of the distribution $\mu = \langle x \rangle$?
- Bayes: compute the *posterior probability* $p(\mu|\{x_i\})$
- Frequentist: devise an *estimator* $\hat{\mu}$ for μ. Ideally it should be *unbiased*, so $\langle \hat{\mu} \rangle = \mu$ and have as small an error as possible (*minimum variance*).
- These lead to apparently identical results (although they aren't), but the interpretation is very different

ICIC



0.25
0.20
0.15
0.10
0.05
−2
2
4
6
8
10



0.25
0.20
0.15
0.10
0.05
−2
2
4
6
8
10

State your priors

- In easy cases, the effect of the prior is simple
- As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes
- Rule of thumb: if changing your prior† to another reasonable one changes the answers significantly, you need more data
- Reasonable priors? Noninformative* – constant prior
- scale parameters in $[0, \infty)$; uniform in log of parameter (Jeffreys' prior*)
- Beware: in more complicated, multidimensional cases, your prior may have subtle effects…

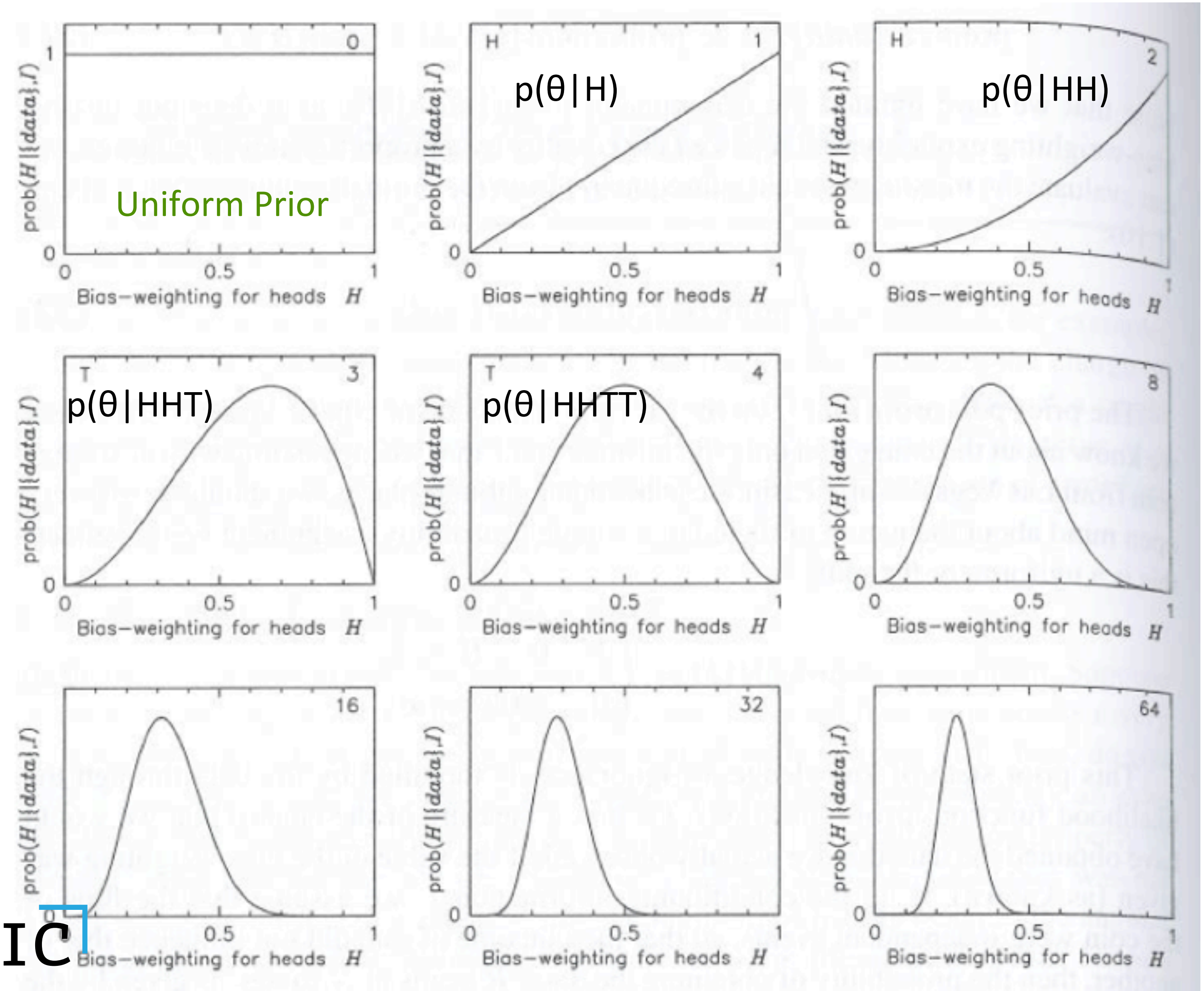
† I mean the raw theoretical one, not modified by an experiment

* Actually, it's better not to use these terms – other people use them to mean different things – just say what your prior is!

ICIC

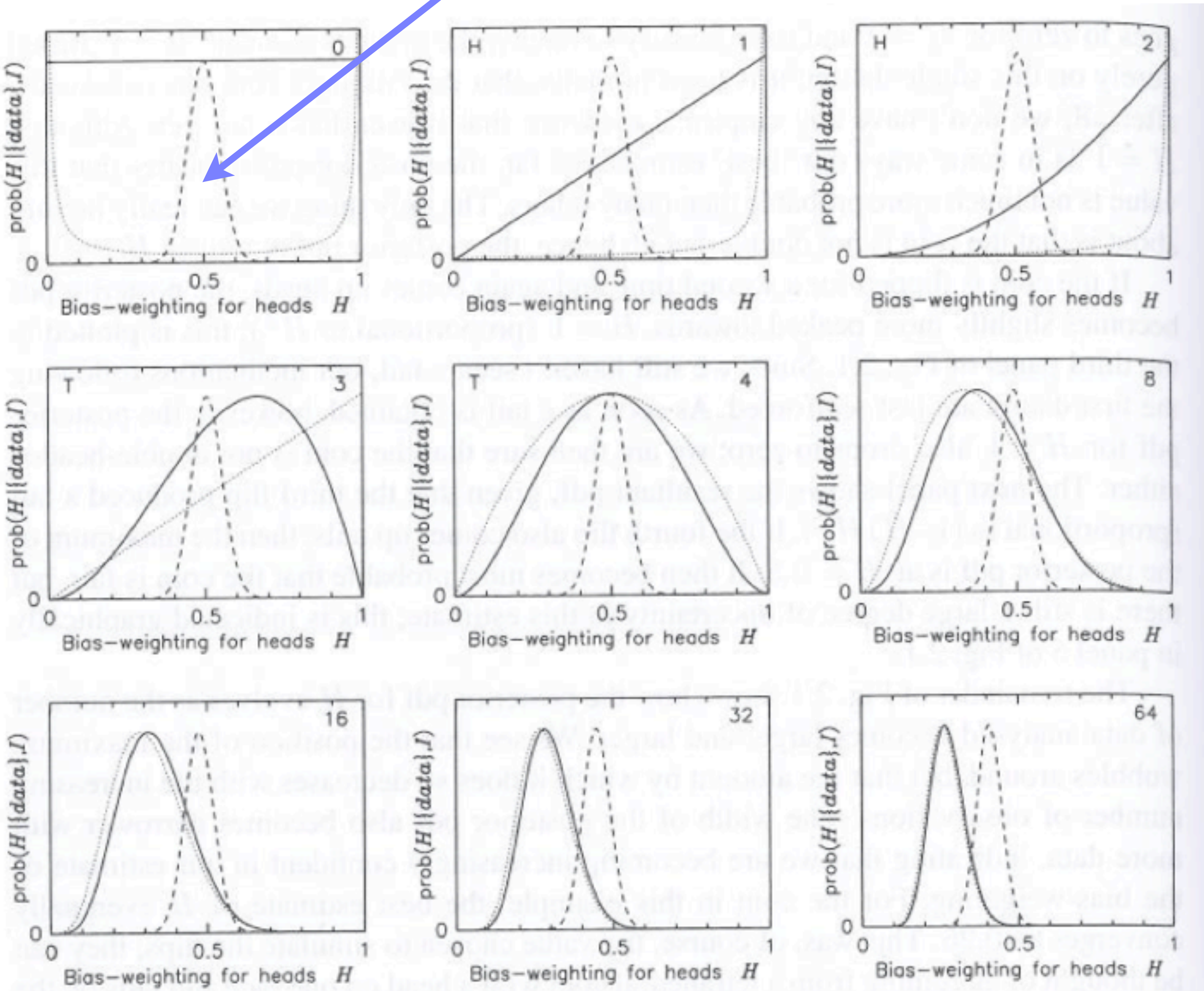
From Sivia & Skilling's *Data Analysis* book. IS THE COIN FAIR?

Model: independent throws of coin. Parameter θ = probability of H



The effect of priors

Priors = "It's likely to be nearly fair", "It's likely to be very unfair"



Sivia & Skilling

- VSA CMB experiment

(Slosar et al 2003)

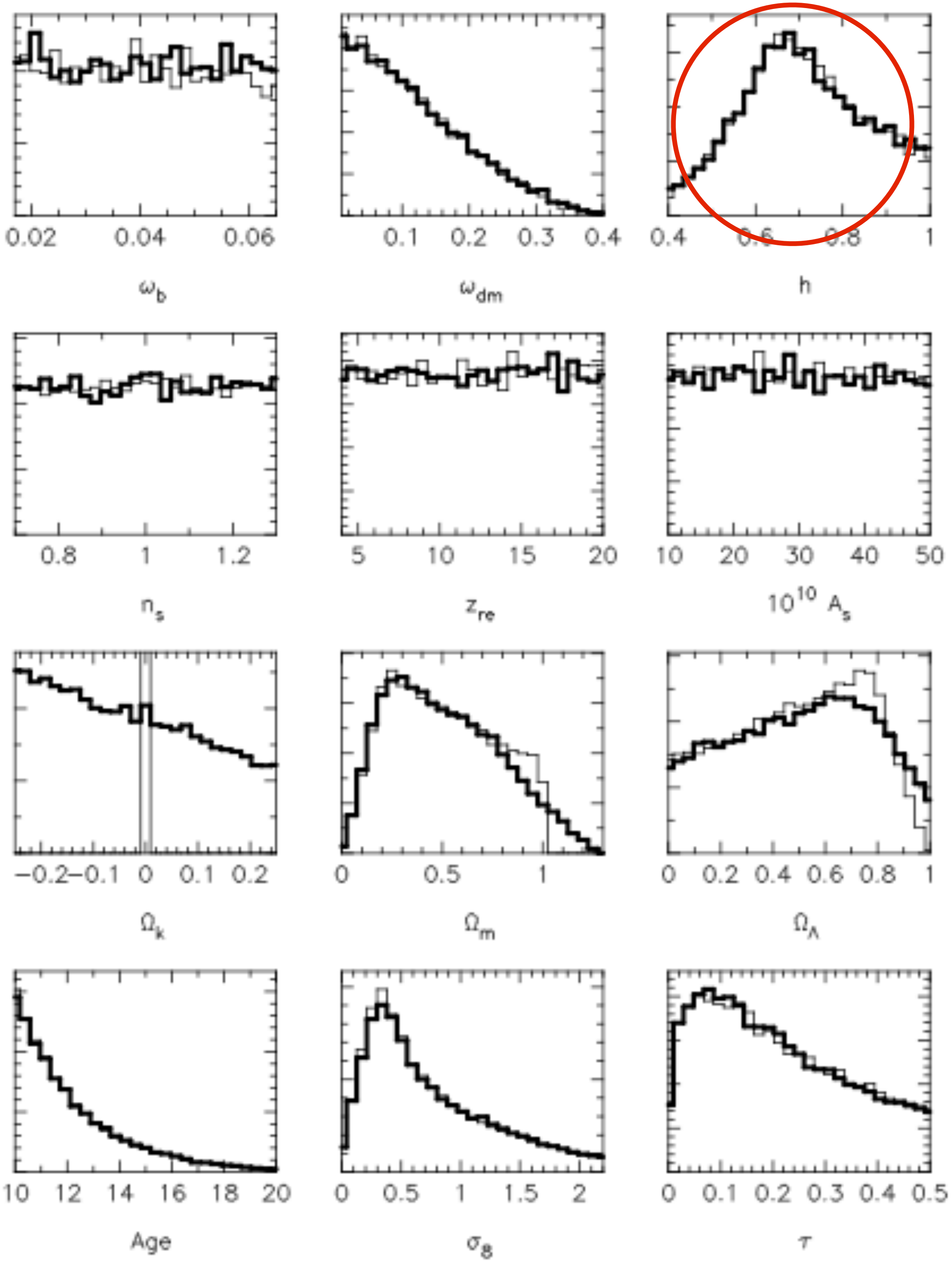
Priors: $\Omega_\Lambda \geq 0$

$10 \leq \text{age} \leq 20$ Gyr

h ≈ 0.7 ± 0.1

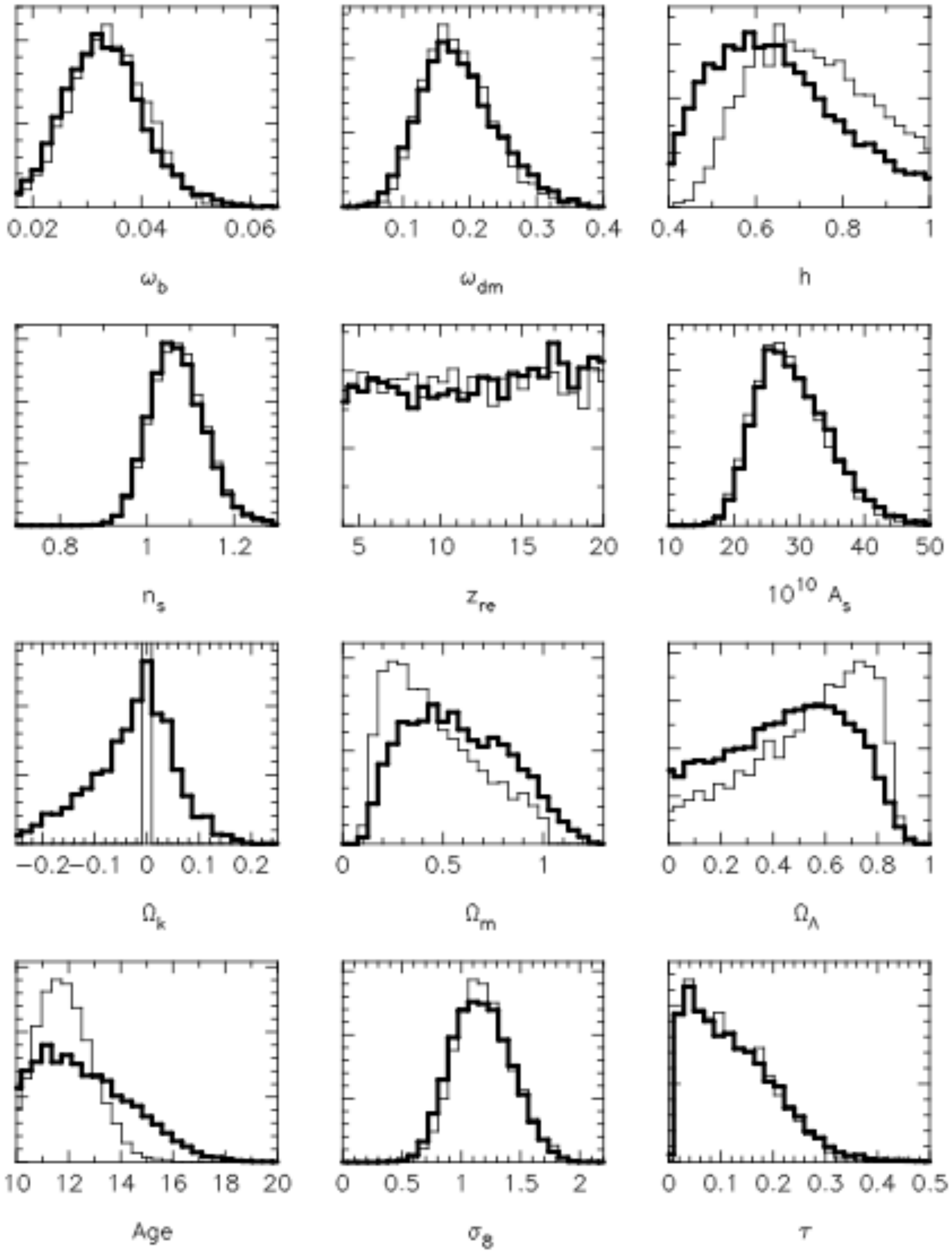
There are no data in these plots – it is all coming from the prior!

$$p(\theta_1) = \int d\theta_{j\neq 1}\, p(x|\theta)\, p(\theta)$$



ICIC

VSA posterior



Estimating the parameter(s)

- Commonly the mode is used (the peak of the posterior)
- Mode = *Maximum Likelihood Estimator*, if the priors are uniform
- The *posterior mean* may also be quoted, but beware
- Ranges containing x% of the posterior probability of the parameter are called *credibility intervals* (or *Bayesian confidence intervals*)

ICIC

Errors

- If we assume uniform priors, then the posterior is proportional to the likelihood.

If further, we assume that the likelihood is single-moded (one peak at θ_0) , we can make a Taylor expansion of lnL:

$$\ln L(x;\theta) = \ln L(x;\theta_0) + \tfrac{1}{2}(\theta_\alpha - \theta_{0\alpha})\frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta}(\theta_\beta - \theta_{0\beta}) + \ldots$$

$$L(x;\theta) = L_0 \exp\left[-\tfrac{1}{2}(\theta_\alpha - \theta_{0\alpha})H_{\alpha\beta}(\theta_\beta - \theta_{0\beta}) + \ldots\right]$$

where the Hessian matrix is defined by these equations. Comparing this with a gaussian, the *conditional error* (keeping all other parameters fixed) is

$$\sigma_\alpha = \frac{1}{\sqrt{H_{\alpha\alpha}}}$$

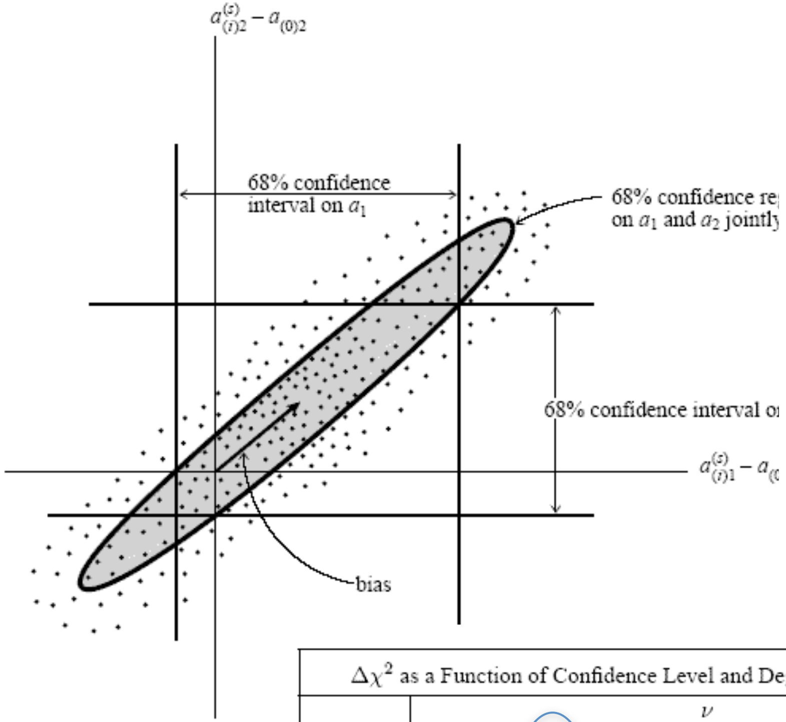
Marginalising over all other parameters gives the *marginal error*

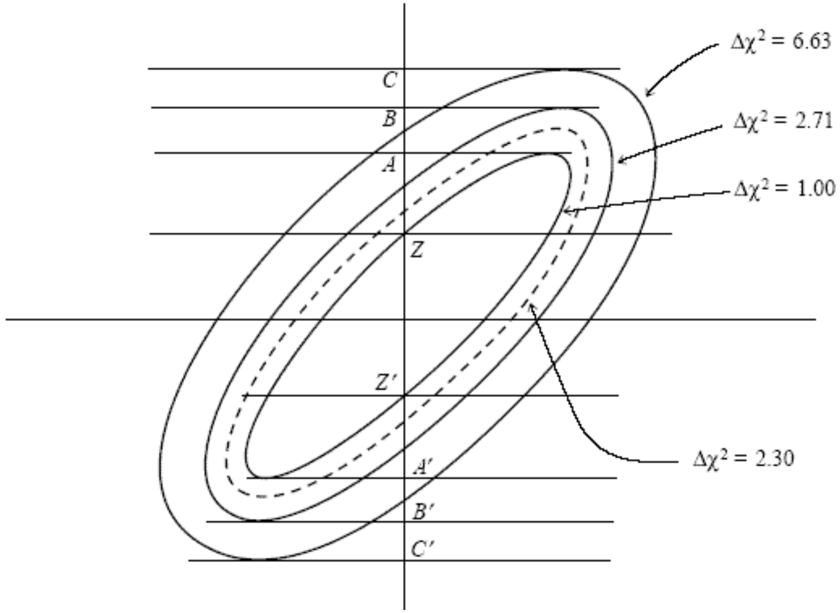
$$\sigma_\alpha = \sqrt{(H^{-1})_{\alpha\alpha}}$$

ICIC

How do I get error bars in several dimensions?

- Read Numerical Recipes, Chapter 15.6





$$L \propto e^{-\frac{1}{2}\chi^2}$$

$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom

p	ν = 1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

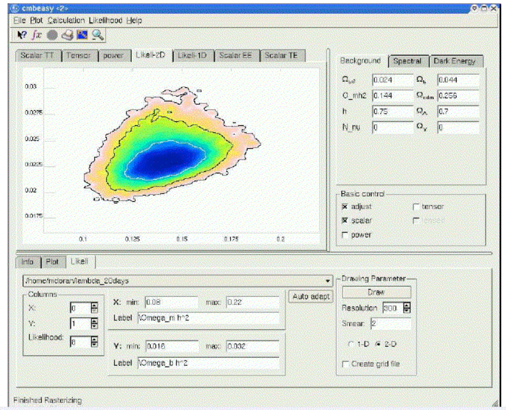
Beware! Assumes gaussian distribution

Say what your errors are!
e.g. 1σ, 2 parameter

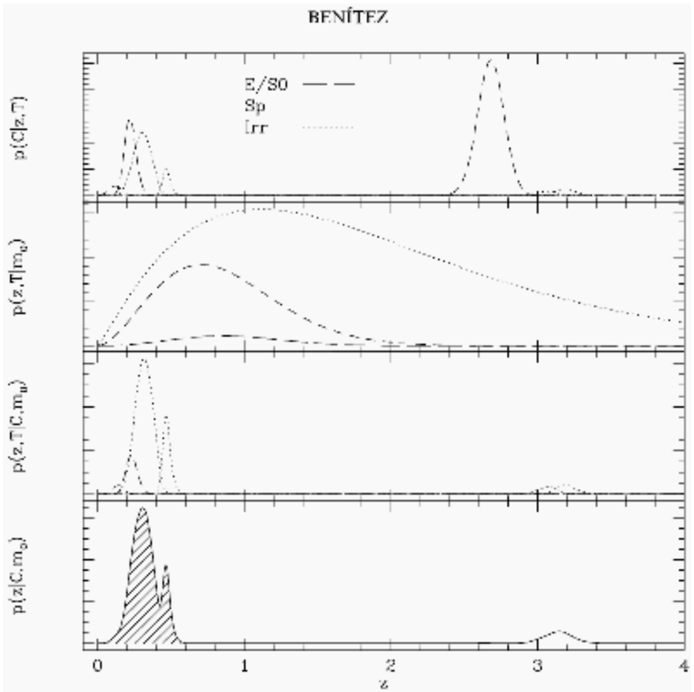
ICIC

Multimodal posteriors etc

- Peak may not be gaussian
- Multimodal? Characterising it by a mode and an error is probably inadequate. May have to present the full posterior.
- Mean posterior may not be useful in this case – it could be very unlikely, if it is a valley between 2 peaks.



From CMBEasy MCMC



From BPZ

ICIC

Non-gaussian likelihoods: number counts

- A radio source is observed with a telescope which can detect sources with fluxes above S_0. The radio source has a flux $S_1 = 2S_0$.

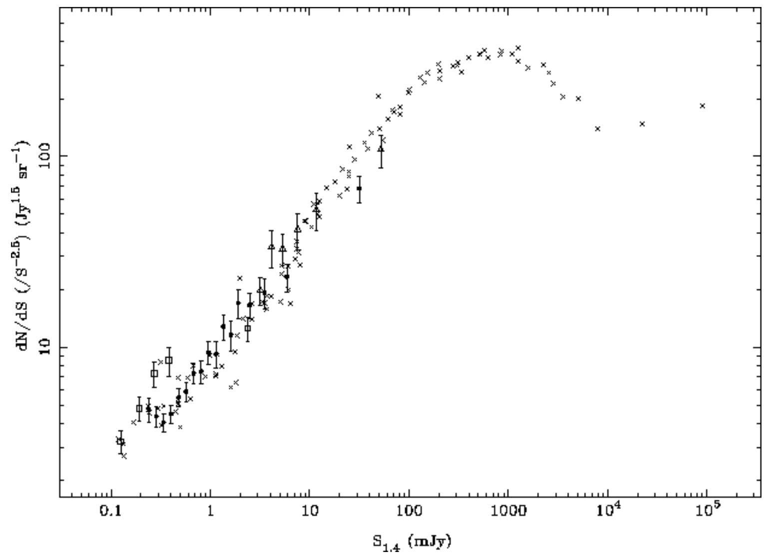
What is the slope of the number counts? (Assume $N(S)dS \propto S^{-\alpha} dS$)

Possible answers:

Pretty steep ($\alpha>1.5$)

Pretty shallow ($\alpha<1.5$)

We can't tell from one point.

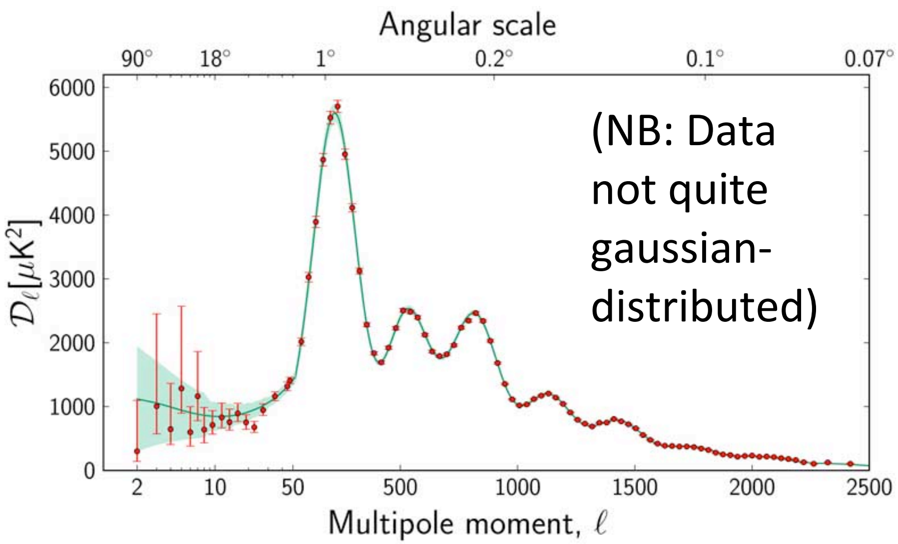


Fisher Matrices

- Useful for forecasting errors, and experimental design
- The likelihood depends on the data collected. Can we estimate the errors before we do the experiment?
- With some assumptions, yes, using the Fisher matrix

$$F_{\alpha\beta} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle$$

For gaussian data, we need to know only:
1. The expectation value of the data, μ(θ)
2. The covariance matrix of the data, C(θ)



(NB: Data not quite gaussian-distributed)

Gaussian errors

- If the data have gaussian errors (which may be correlated) then we can compute the Fisher matrix easily:

$$F_{\alpha\beta} = \frac{1}{2} Tr[C^{-1} C_{,\alpha} C^{-1} C_{,\beta} + C^{-1} M_{\alpha\beta}],$$

e.g. Tegmark, Taylor, Heavens 1997

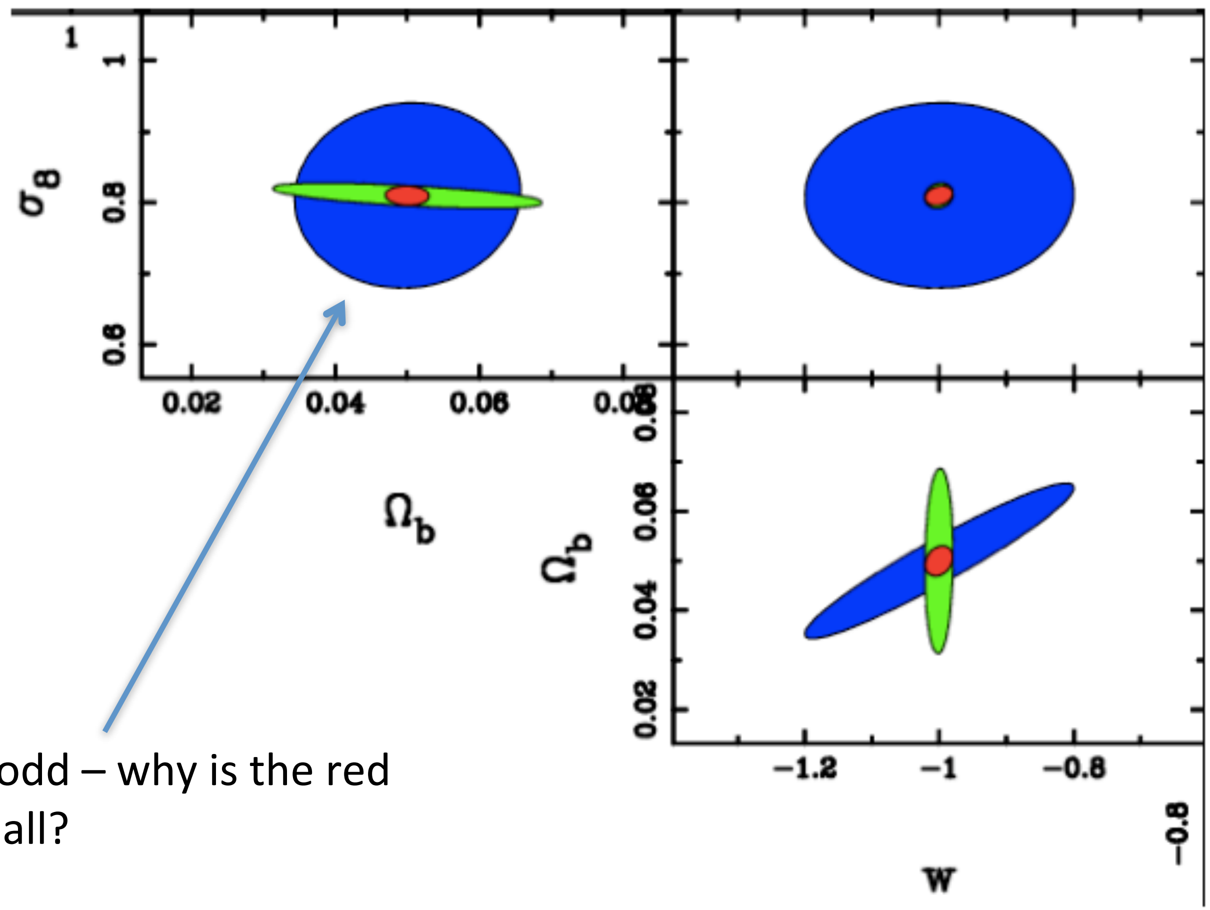
$\mu_\alpha = \langle x_\alpha \rangle$ $\quad C_{\alpha\beta} = \langle (x-\mu)_\alpha (x-\mu)_\beta \rangle$ $\quad M_{\alpha\beta} = \mu_{,\alpha}\mu^T_{,\beta} + \mu^T_{,\alpha}\mu_{,\beta}$

Forecast marginal error on parameter α : $\sigma_\alpha = \sqrt{(F^{-1})_{\alpha\alpha}}$

- For independent experiments, the Fisher Matrices add (the inverse may pleasantly surprise you)

ICIC

Combining datasets



This looks odd – why is the red blob so small?

Summary

- Write down what you want to know. Typically:

$$p(\theta|xIM)$$

- What is θ ?
- What is I?
- What is M?

- You might want p(M|xI)...Model Selection