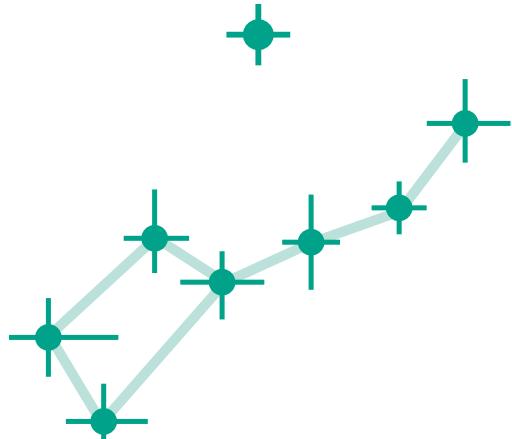




Statistical Inference for Astronomy and Particle Physics

Alan Heavens



ICIC

Imperial Centre
for Inference & Cosmology

Weizmann Institute
March 8-12 2015

Introduction to Bayesian Inference



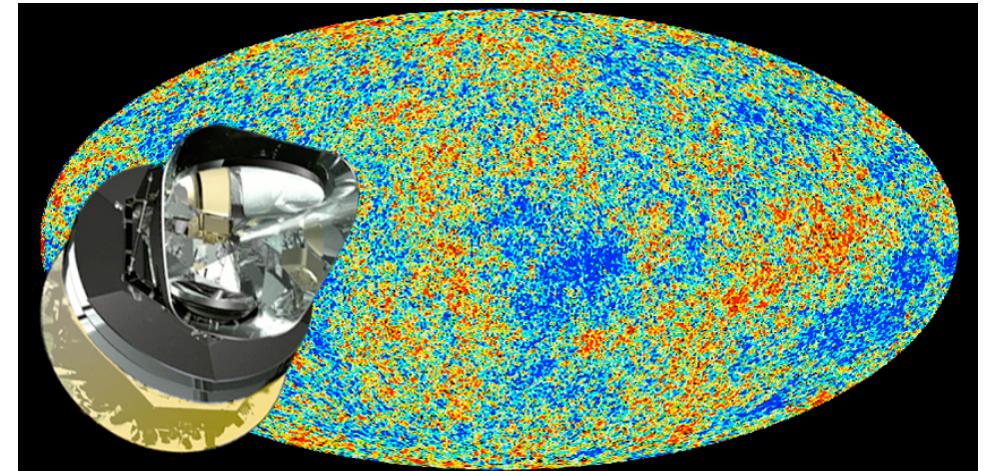
Alan Heavens
Imperial College London

Further reading

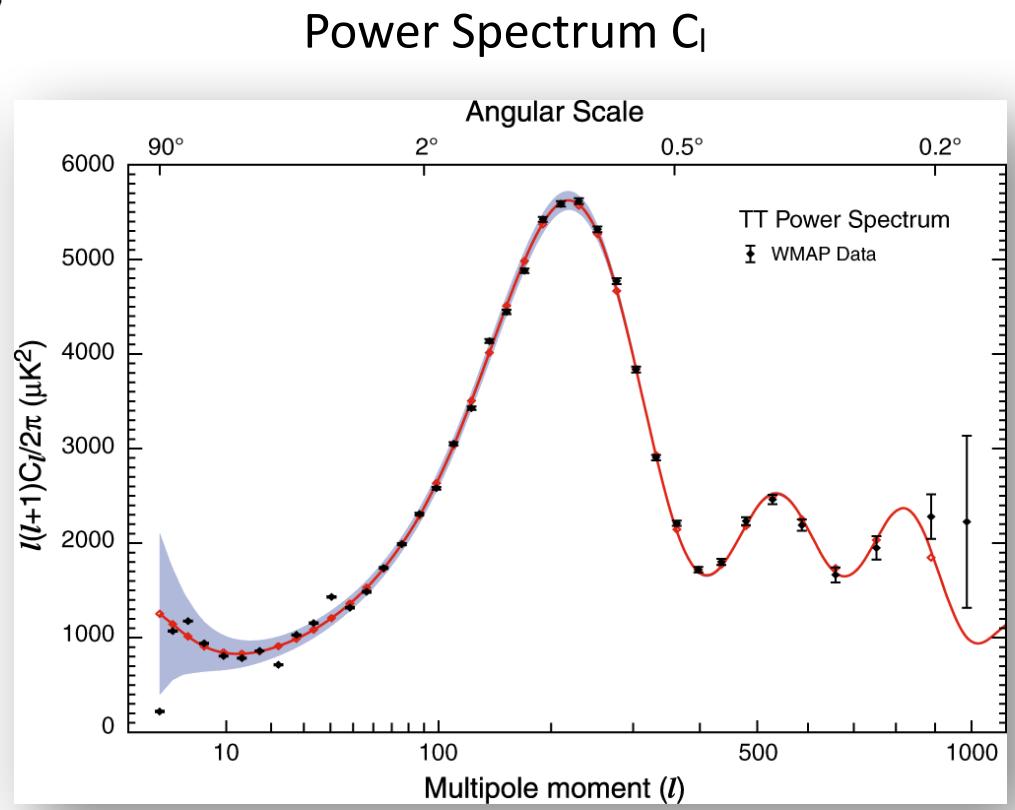
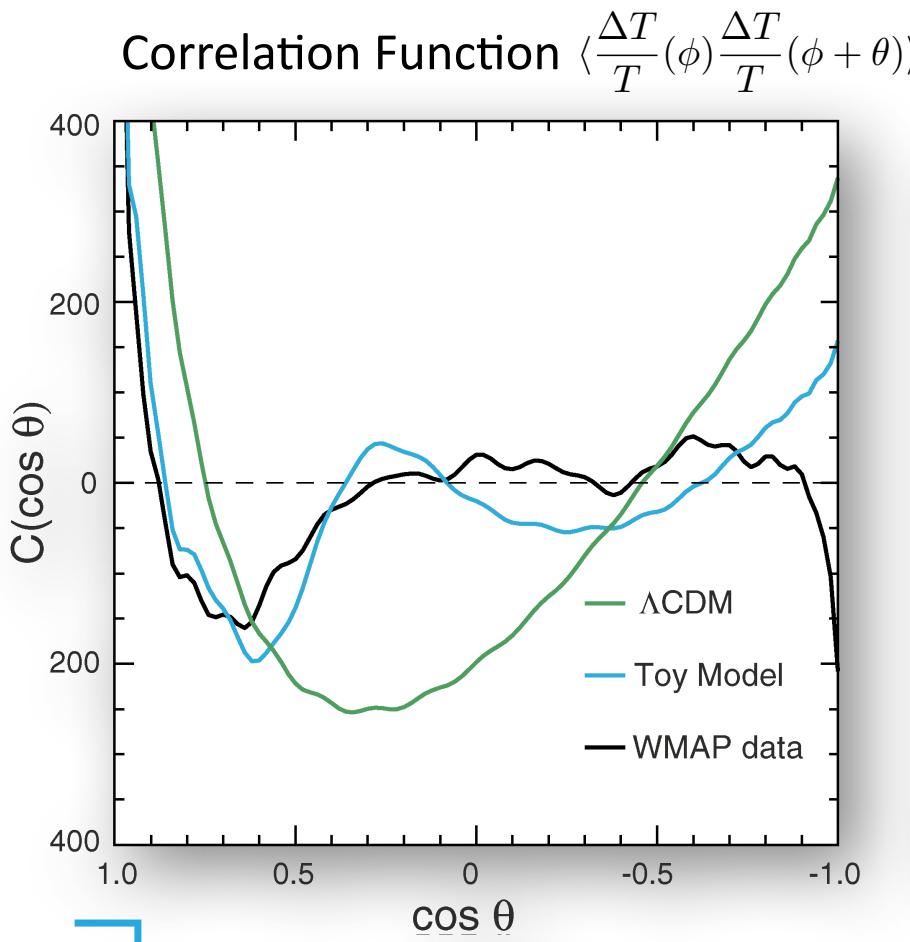
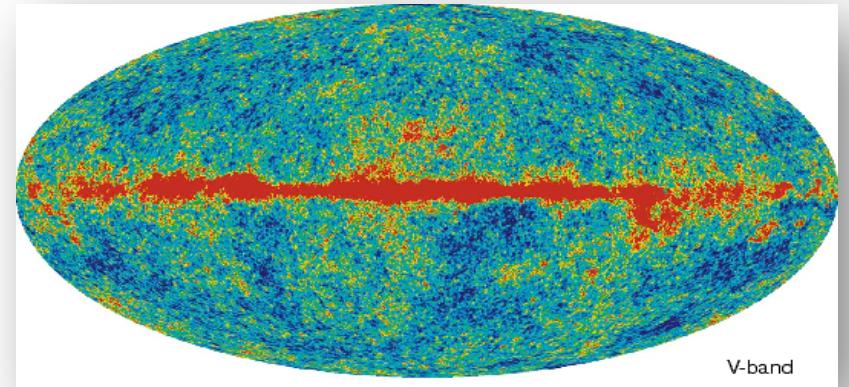
- Loredo's *Bayesian Inference in the Physical Sciences*:
 - <http://astrosun.tn.cornell.edu/staff/loredo/bayes>
 - “The Promise of Bayesian Inference for Astrophysics” & “From Laplace to SN 1987a”
- MacKay, *Information Theory, Inference & Learning Algorithms*
- Jaynes, *Probability Theory: the Logic of Science*
 - And other refs at <http://bayes.wustl.edu>
- Hobson et al, *Bayesian Methods in Cosmology*
- Sivia, *Data Analysis: A Bayesian Tutorial*
Later edition: Sivia and Skilling

Outline

- Inverse problems: from data to theory
- Probability review, and Bayes' theorem
- Parameter inference
- Priors
- Marginalisation
- Posteriors
- Sampling - MCMC, HMC, Gibbs
- Model selection
- Hypothesis testing
- Bayesian Hierarchical Models



LCDM fits the WMAP data well.



Inverse problems

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

Examples

- Hypothesis testing
 - Is the CMB radiation consistent with (initially) gaussian fluctuations?
- Parameter inference
 - 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 theory?

What is probability?

- Frequentist view: p describes the relative *frequency of outcomes* in infinitely long trials
- Bayesian view: p expresses our *degree of belief*
- logical proposition: statement of fact that could be true or false
- $p(A|B) = \text{degree to which truth of some logical proposition } B \text{ implies that the logical proposition } A \text{ is also true.}$
- 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?*

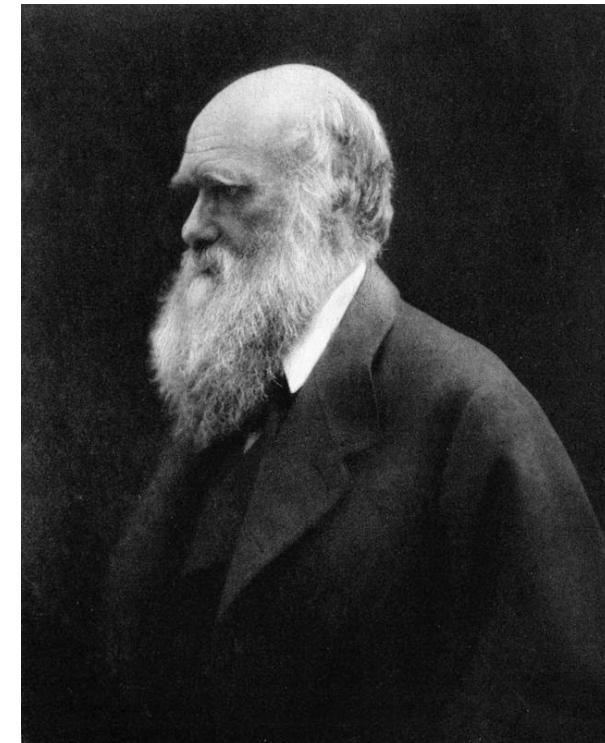
Bayes' Theorem

- Rules of probability:
- $p(x) + p(\text{not } x) = 1$ sum rule
- $p(x,y) = p(x|y) p(y)$ product rule
- $p(x) = \sum_k p(x,y_k)$ marginalisation
- Sum \rightarrow integral continuum limit ($p=\text{pdf}$)
$$p(x) = \int dy p(x,y)$$
- $p(x,y)=p(y,x)$ gives *Bayes' theorem*

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

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

- $x = \text{is male}; y = \text{has beard}$
- $p(y|x) = 0.1$
- $p(x|y) = 1$



Julia Margaret Cameron



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



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

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

*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

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

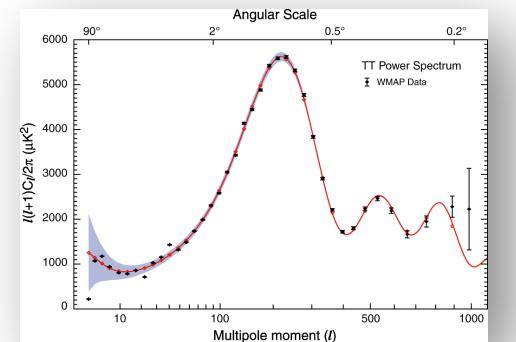
Posterior Likelihood L Evidence or Model Likelihood Prior

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 inference, 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'), $p(x|M) \neq p(x|M')$

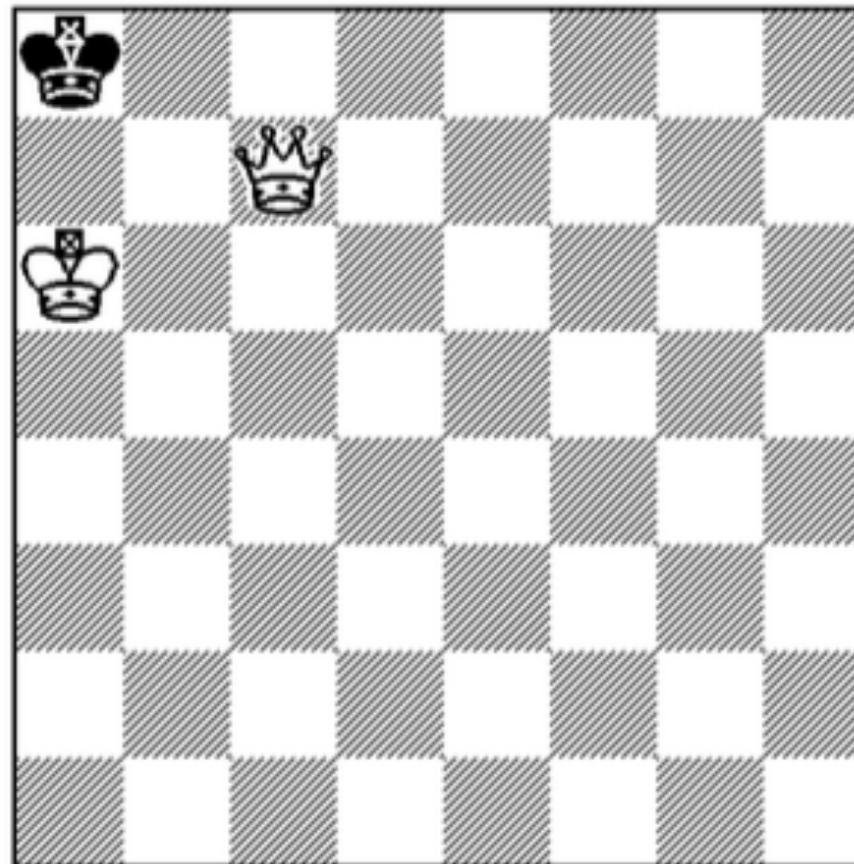


Self-consistent?

- Yes. Consider data from 2 experiments. We can do one of 3 things:
 - Define prior; obtain posterior from dataset 1; update the prior, then analyse dataset 2
 - As above, but swap 1 and 2
 - Define prior; obtain posterior from datasets 1 and 2 combined
- These have to (and do) give the same answers

A diversion on priors

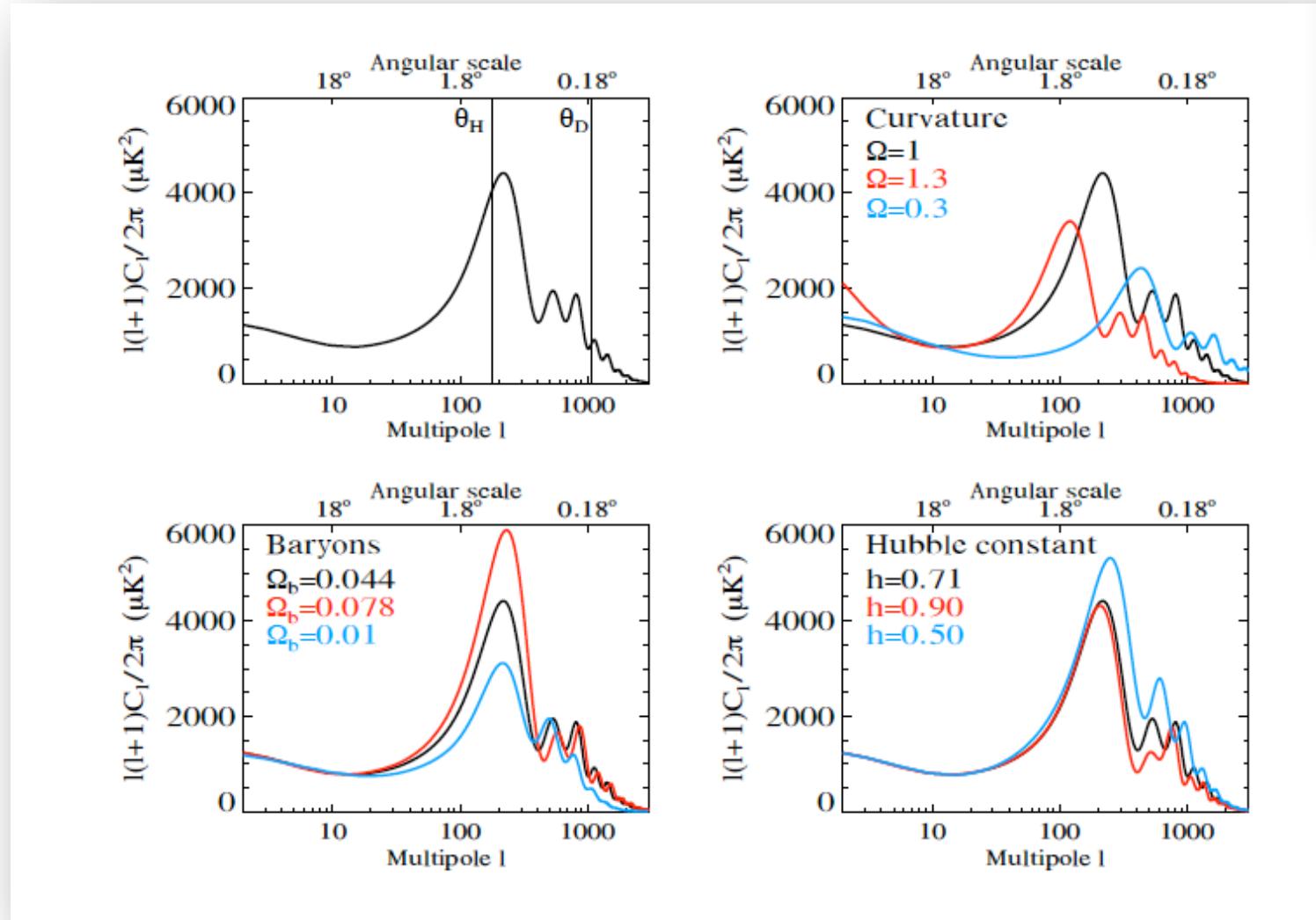
- You bring more prior information than you may think...



Credit: Daniel Mortlock

What is the probability that White wins?

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.

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 superficially identical results (although they aren't), but the interpretation is very different
- Bayesian: no estimators - just posteriors

Set up the problem

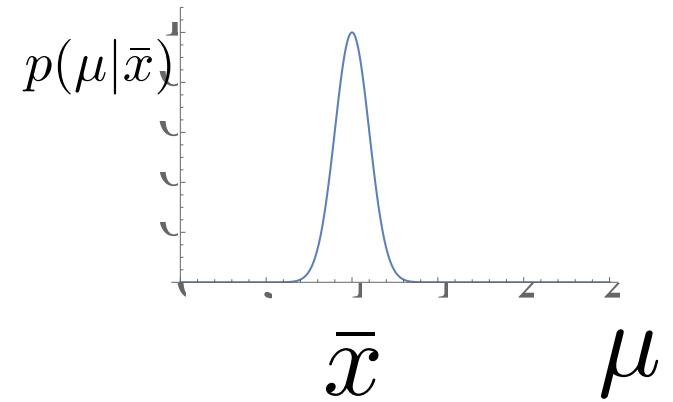
- What is the model for the data, M ?
- $M: x = \mu + n$
- Data: a set of values $\{x_i\}$, $i=1\dots N$
- Prior info I : noise $\langle n \rangle = 0$ $\langle n^2 \rangle = \sigma^2$ (known); gaussian distributed
- θ : the mean, μ
- Rule 1: what do we want?
- $p(\mu | \{x_i\})$

Discussion

- Bayesian: Posterior for μ is

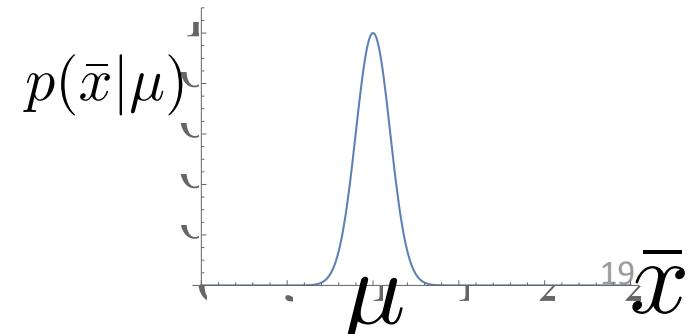
$$p(\mu|\{x_i\}) \propto \exp \left[-\frac{(\mu - \bar{x})^2}{2\sigma^2/N} \right]$$

– where $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$



- Frequentist: the *estimator* $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ is unbiased, has minimum variance, and has a distribution

$$p(\bar{x}|\mu) \propto \exp \left[-\frac{(\mu - \bar{x})^2}{2\sigma^2/N} \right]$$



State your priors

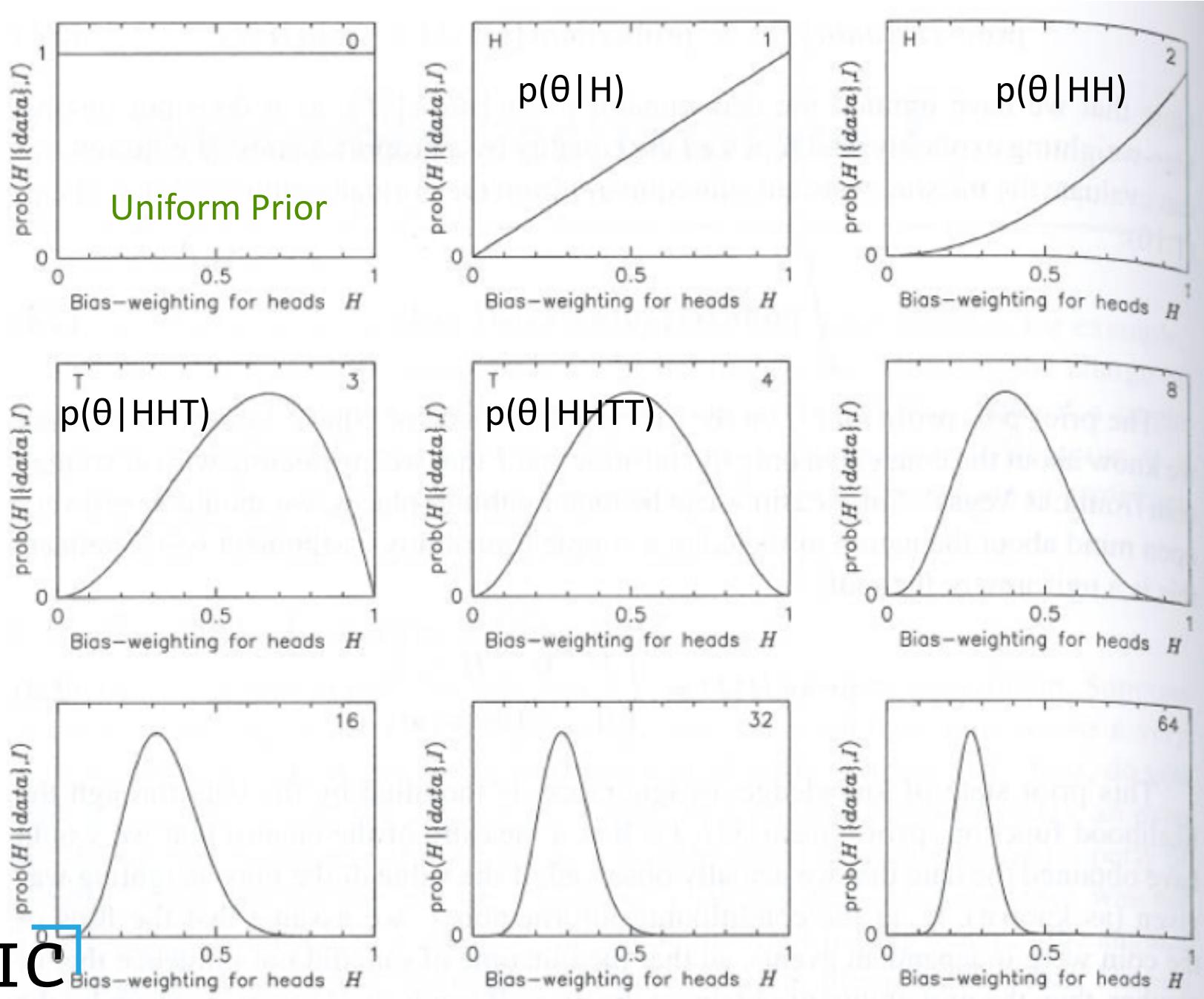
- In easy cases, the effect of the prior is simple
- As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes
- Rule of thumb: if changing your prior[†] to another reasonable one changes the answers a lot, you could do with more data
- Reasonable priors? Noninformative^{*} – constant prior (can usually do this even if infinite interval).
- scale parameters in $[0, \infty)$; uniform in log of parameter (Jeffreys' prior)
- Bayesian reasoning is NOT subjective - posterior is determined unambiguously from the prior and likelihood

[†] 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!

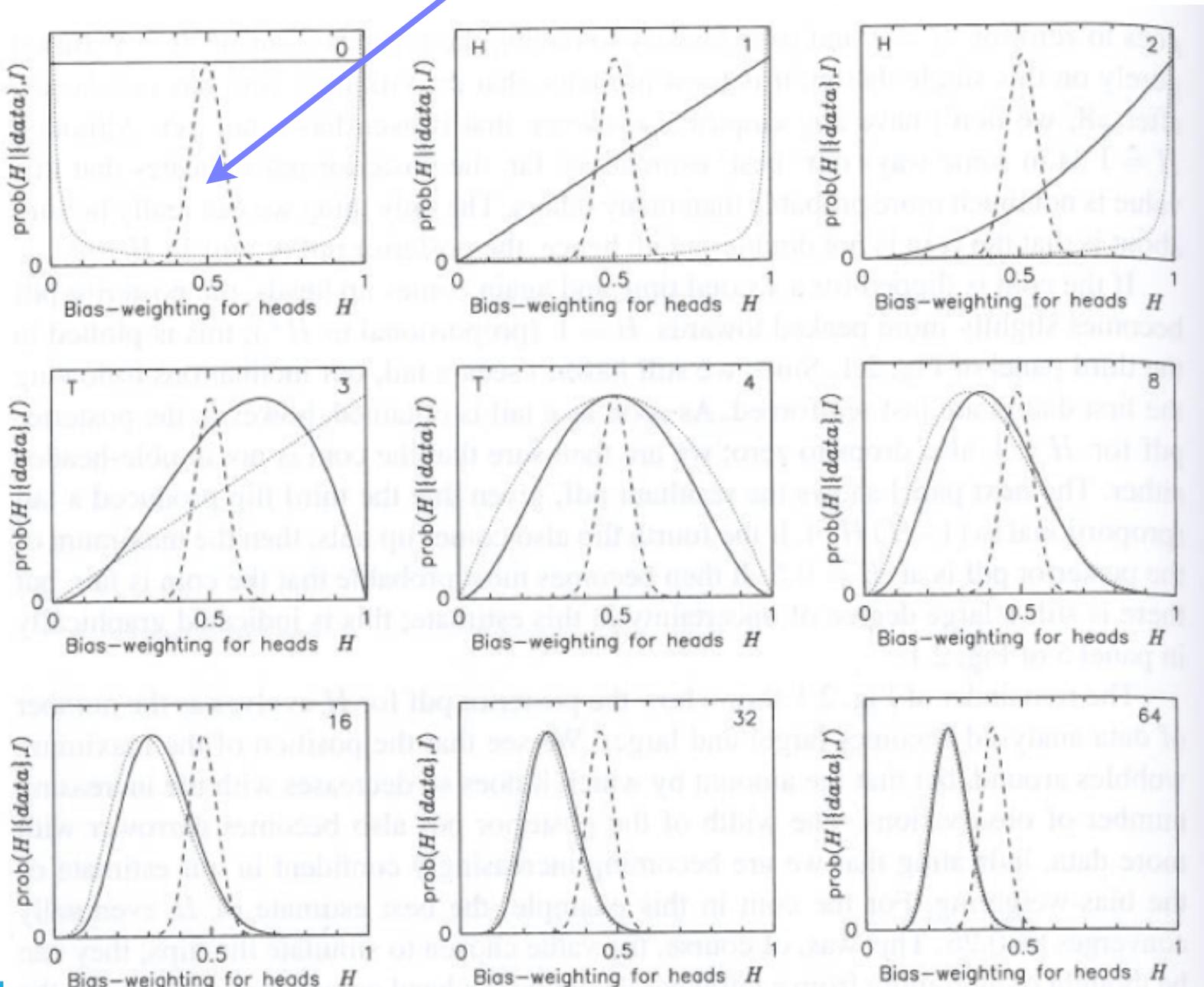
From Sivia & Skilling's *Data Analysis* book. IS THE COIN FAIR?

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”

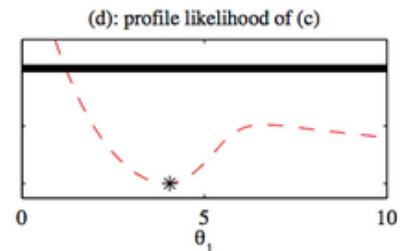
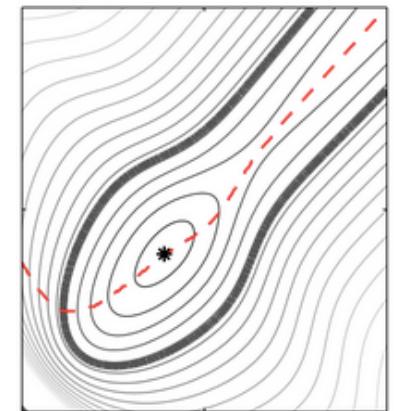


Marginalisation

- Important concept: the *marginal distribution* of θ_1 is

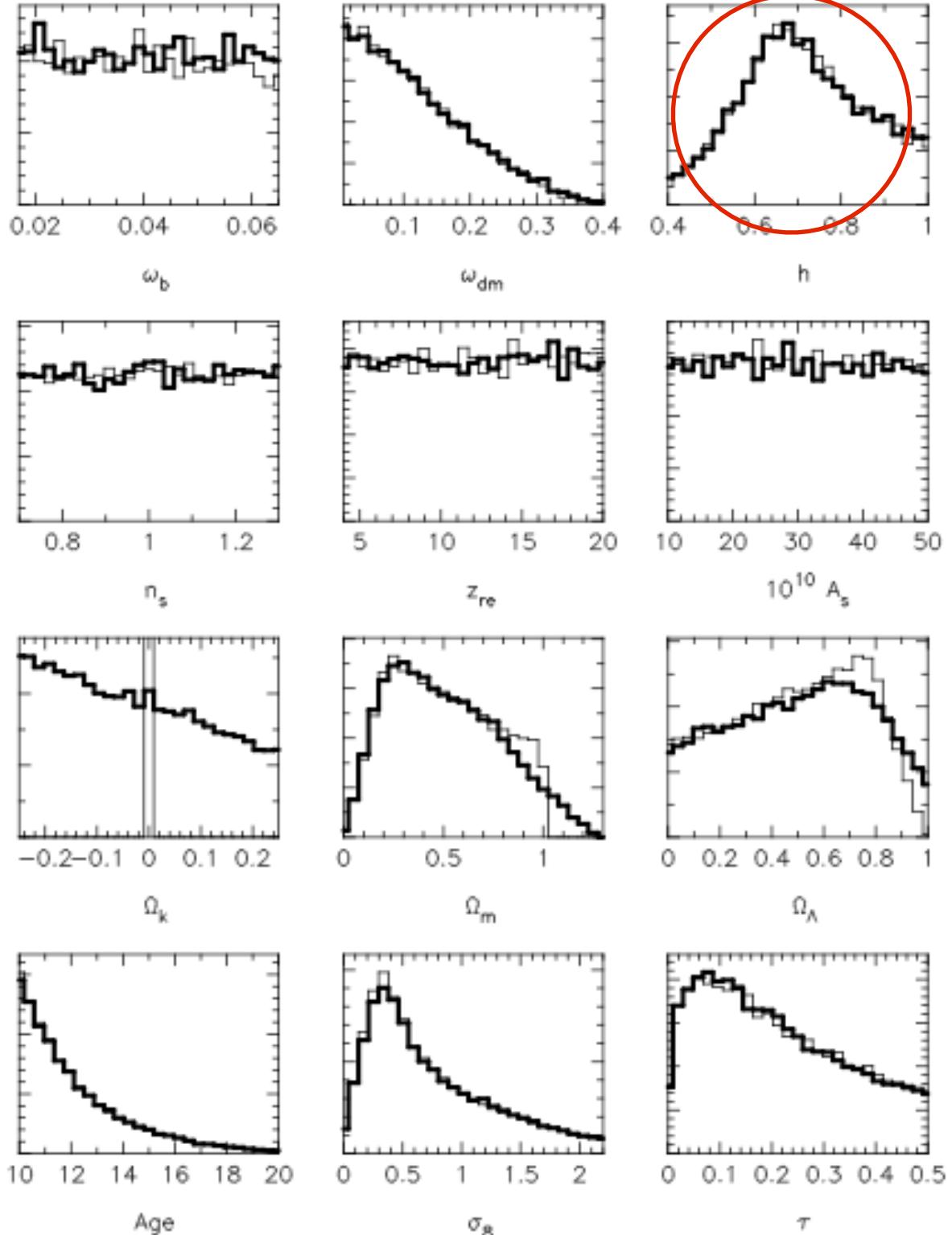
$$p(\theta_1|x) = \int p(\theta_1, \theta_2, \dots | x) d\theta_2 d\theta_3 \dots$$

- Posterior for each parameter includes the uncertainty in the other parameters
- Profile likelihood* is something different: maximise w.r.t. some of the parameters.
- From a Bayesian point-of-view, the profile likelihood is unsatisfactory, as it does not include the uncertainties in the other parameters



- VSA CMB experiment

(Slosar et al 2003)



Priors: $\Omega_\Lambda \geq 0$

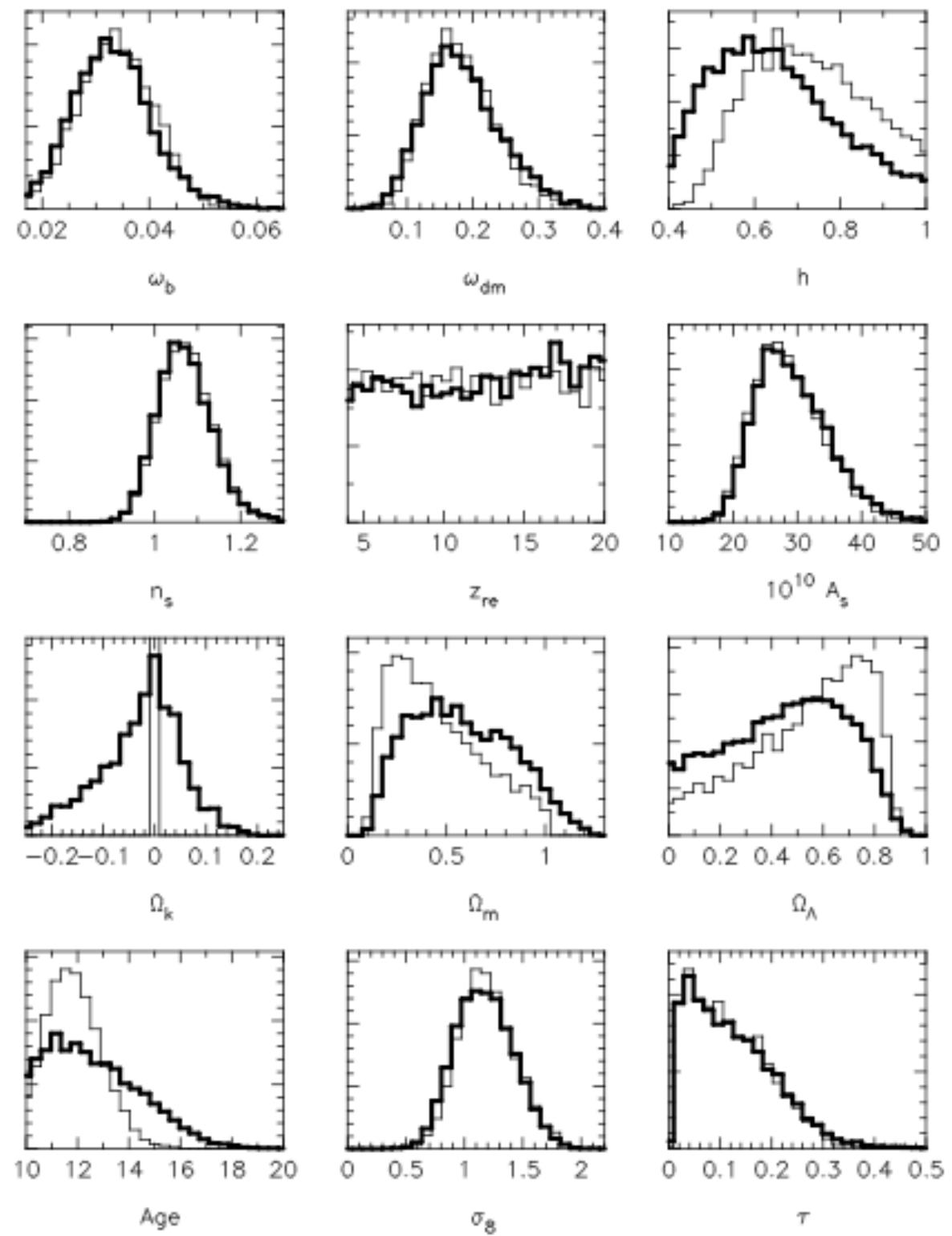
$10 \leq \text{age} \leq 20 \text{ Gyr}$

$$h \approx 0.7 \pm 0.1$$

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

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

VSA posterior



Inferring the parameter(s)

- What to report, when you have the posterior?
- 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
$$\bar{\theta} = \int \theta p(\theta|x)d\theta$$
- Ranges containing x% of the posterior probability of the parameter are called *credibility intervals* (or *Bayesian confidence intervals*)

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 $\ln L$:

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

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

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

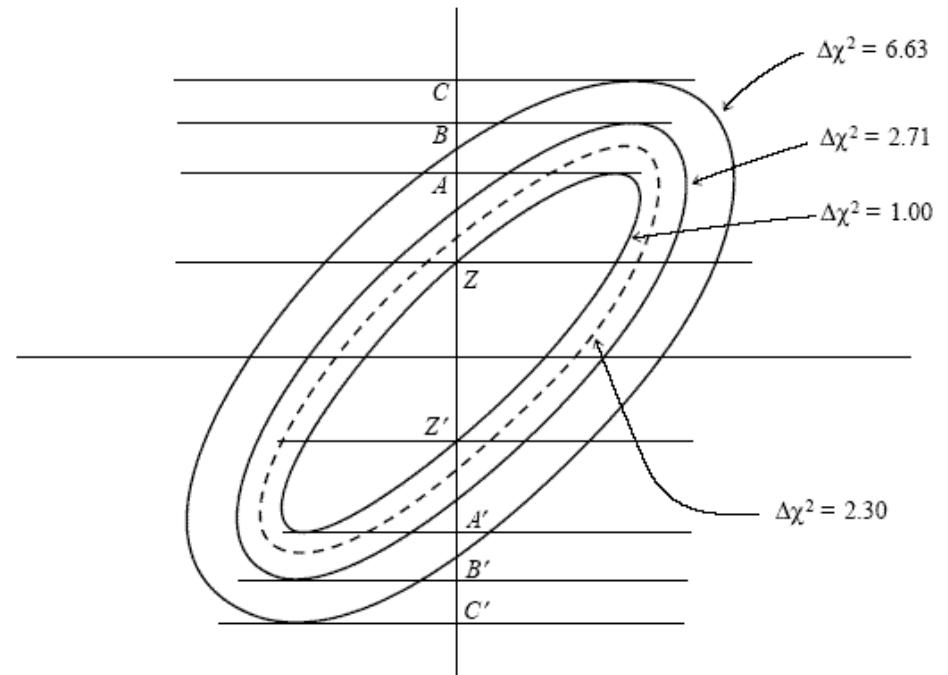
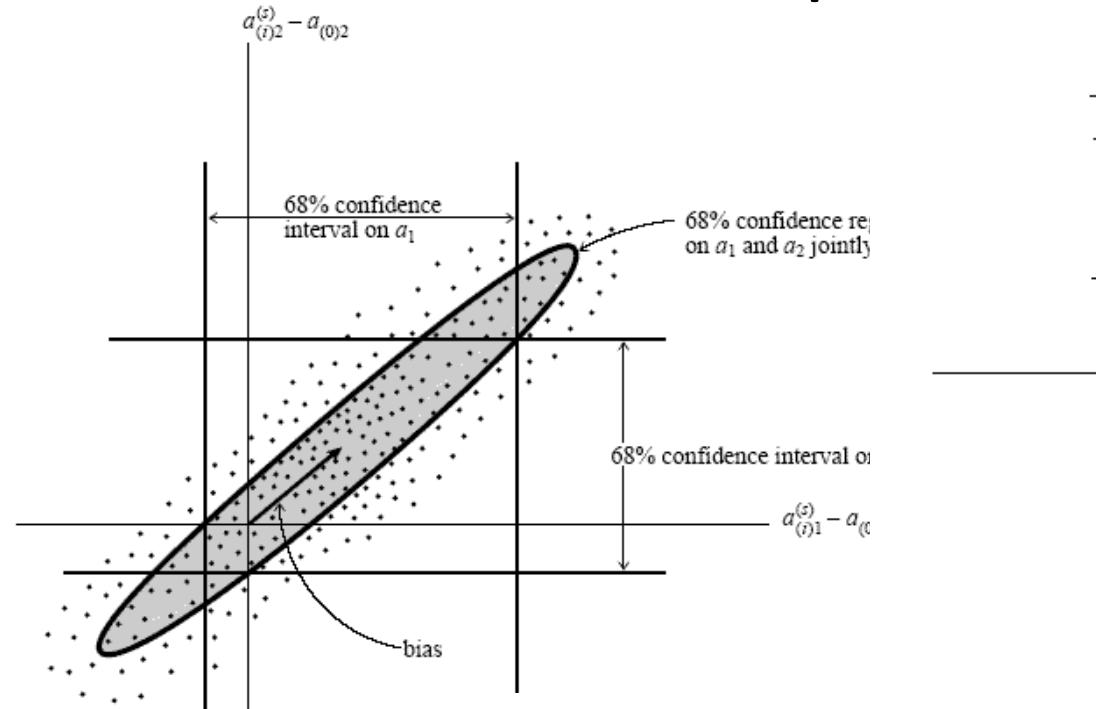
$$\sigma_\alpha = \frac{1}{\sqrt{H_{\alpha\alpha}}}$$

Marginalising over all other parameters gives the *marginal error*

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

How do I get error bars in several dimensions?

- Read Numerical Recipes, Chapter 15.6



Beware! Assumes gaussian distribution

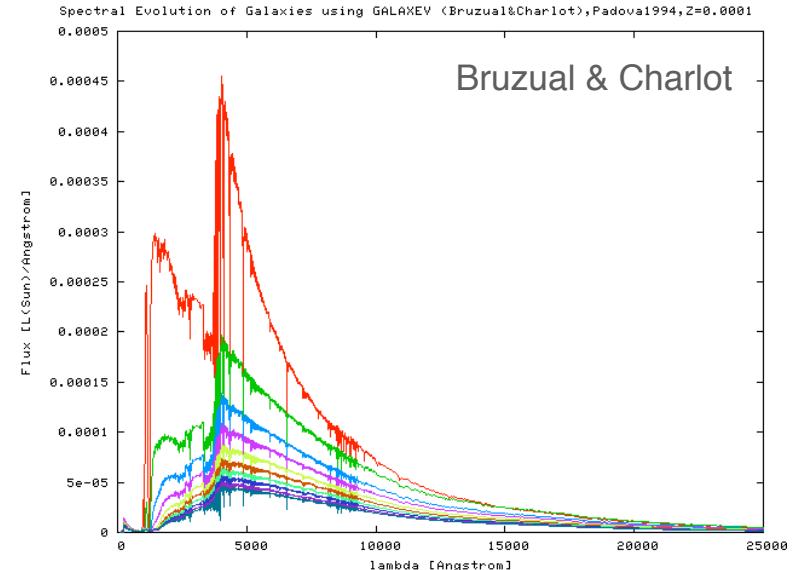
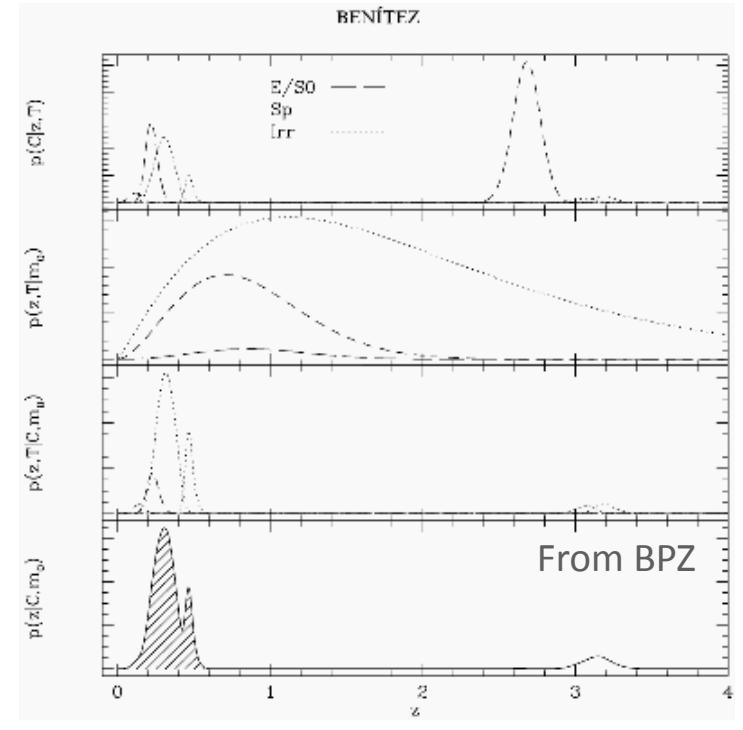
$$L \propto e^{-\frac{1}{2}\chi^2}$$

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

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

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.



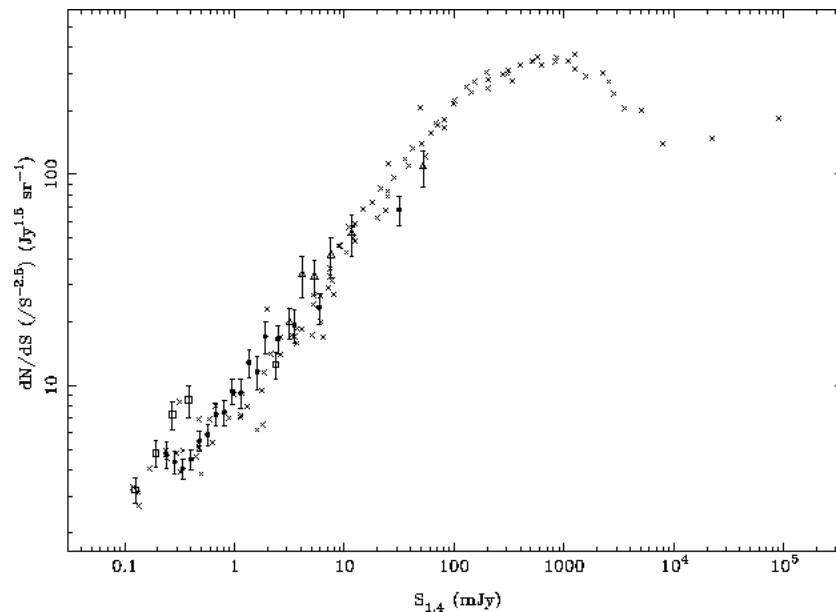
Non-gaussian likelihoods: number counts

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

What is the slope of the number counts?

(Assume $N(S)dS \propto S^{-\alpha} dS$)

Can you tell anything?



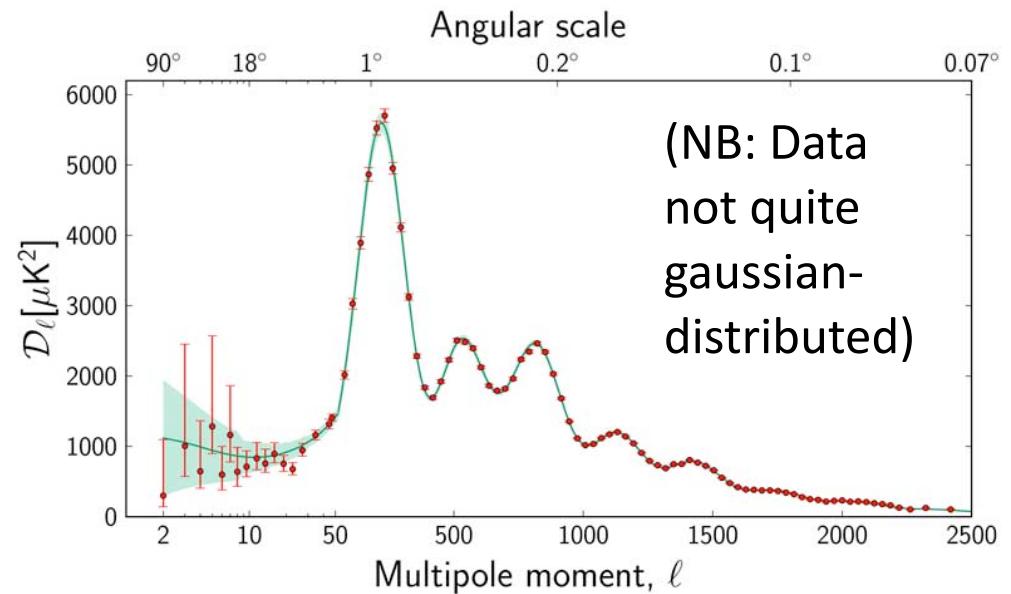
Fisher Matrices

- 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, $\mu(\theta)$
2. The covariance matrix of the data, $C(\theta)$



Gaussian errors

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

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

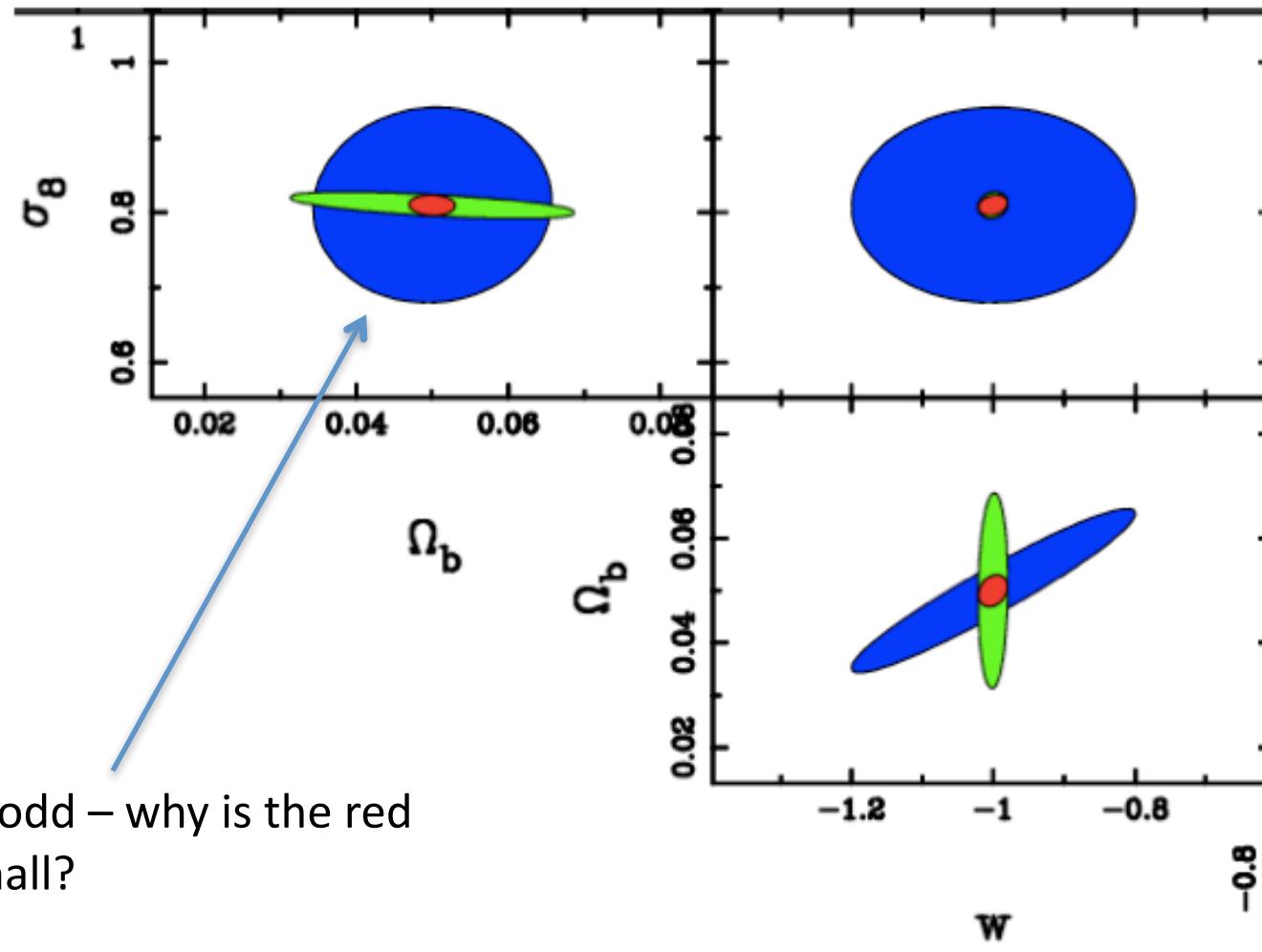
e.g. Tegmark, Taylor, Heavens 1997

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

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)

Combining datasets



Summary

- Write down what you want to know. For *parameter inference* it is typically:

$$p(\theta|xIM)$$

- What is M ?
- What is/are θ ?
- What is I ?
- You might want $p(M|x I)$...this is *Model Selection* - see later