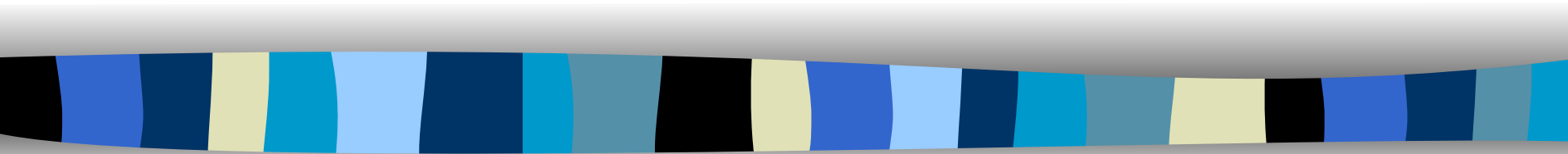


Probabilities and Parameter Estimation

Observational Cosmology



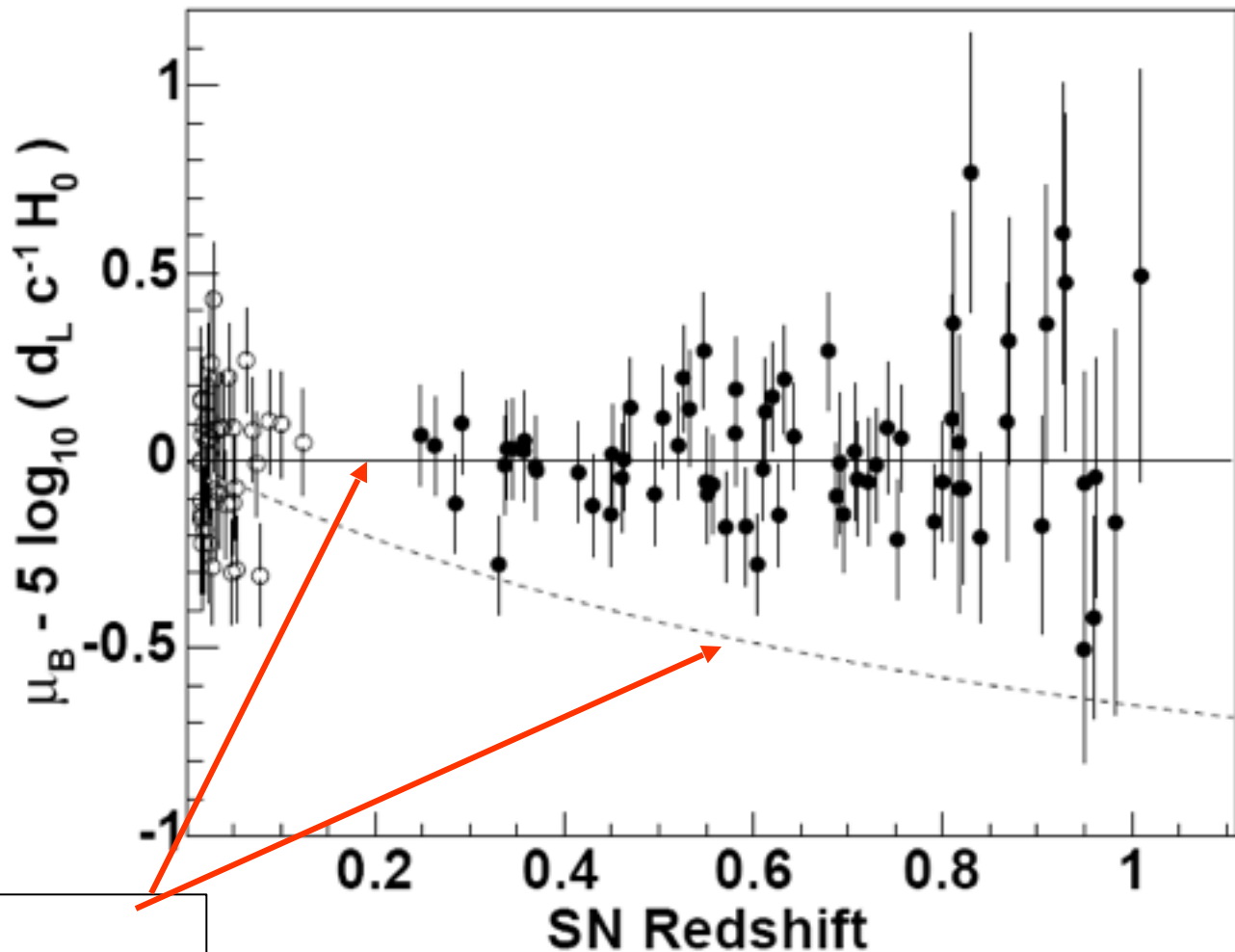
Bruce Bassett (bruce@sao.ac.za)

AIMS, SAAO and UCT

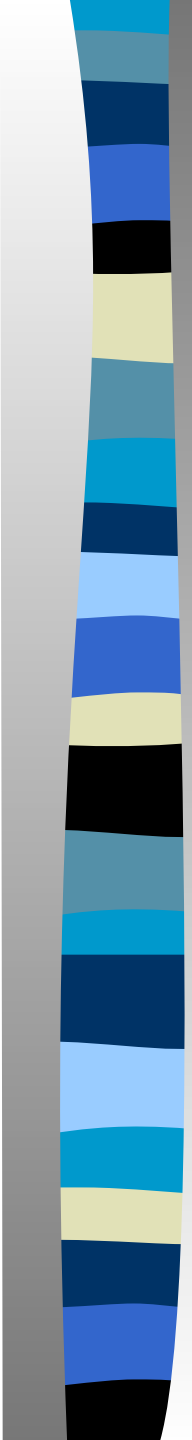
Why focus on statistics and why do it now?

- There is a shortage of modern, applied, statistical skills in many arenas in South African academia and industry
- These skills are essential for **e-Science** and for progress in the modern 'Moore's-Law world' where large data sets are becoming the norm.
- This is particularly true in cosmology (e.g. LOFAR will deliver 1 Terabyte/second, KAT will probably deliver 100 MB/s).

A standard problem – parameter estimation



How do we find the Best-fitting curve to the data?

- 
- We have introduced the χ^2 statistic in the last lecture as a way of estimating parameters.
 - However, we will need to be more sophisticated in our treatment in general to do a good job of parameter estimation in cosmology...



Some basic theory...

Random variables

- We start by assuming that X is a random variable taking values denoted x , either **continuously** or discretely distributed.
- The probability distribution function (**pdf**), $f_X(x)$ is the key function containing the information about how likely any given value, x , is.

Basic Probability Theory

- The pdf controls the probability of observing a given value x for X

(1)

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

Some value must be found with 100% probability

And the probability that $a < x < b$ is

(2)

$$P(a < x < b) = \int_a^b f_X(x) dx$$

Expectations

- For any function $g(X)$, the expectation value of $g(X)$ is defined to be:

$$\langle g(X) \rangle \equiv \int_{-\infty}^{\infty} f_X(x) g(x) dx$$

- This is also sometimes denoted $\mathbf{E}(g(X))$ or $\overline{g(X)}$. The mean of X is often denoted μ :

$$\mu = \langle X \rangle$$

Example: the uniform distribution.

- For the uniform distribution between the limits a and b , the pdf is just (why?):

$$f_X(x) = 1/(b - a)$$

- Hence the expected value (mean) is:

$$\langle X \rangle \equiv \int_{-\infty}^{\infty} f_X(x) x dx = \frac{1}{(b - a)} \int_a^b x dx = \frac{1}{2} (b + a)$$

Moments

- It is of interest to look at fluctuations around the mean, μ

$$\mu_n = \langle (X - \mu)^n \rangle \equiv \int_{-\infty}^{\infty} f_X(x)(x - \mu)^n dx$$

- $n=2$ corresponds to the usual variance, σ^2 .
- Exercise: compute the variance for the uniform distribution.

Skewness and Kurtosis

- The **skewness** (γ_1) and **Kurtosis** (γ_2) are defined to be:

$$\begin{aligned}\gamma_1 &\equiv \frac{\mu_3}{\mu_2^{3/2}} \\ \gamma_2 &\equiv \frac{\mu_4}{\mu_2^2} - 3\end{aligned}$$

- They vanish for a Gaussian pdf (all odd moments vanish for a Gaussian and all even moments can be written in terms of σ^2).



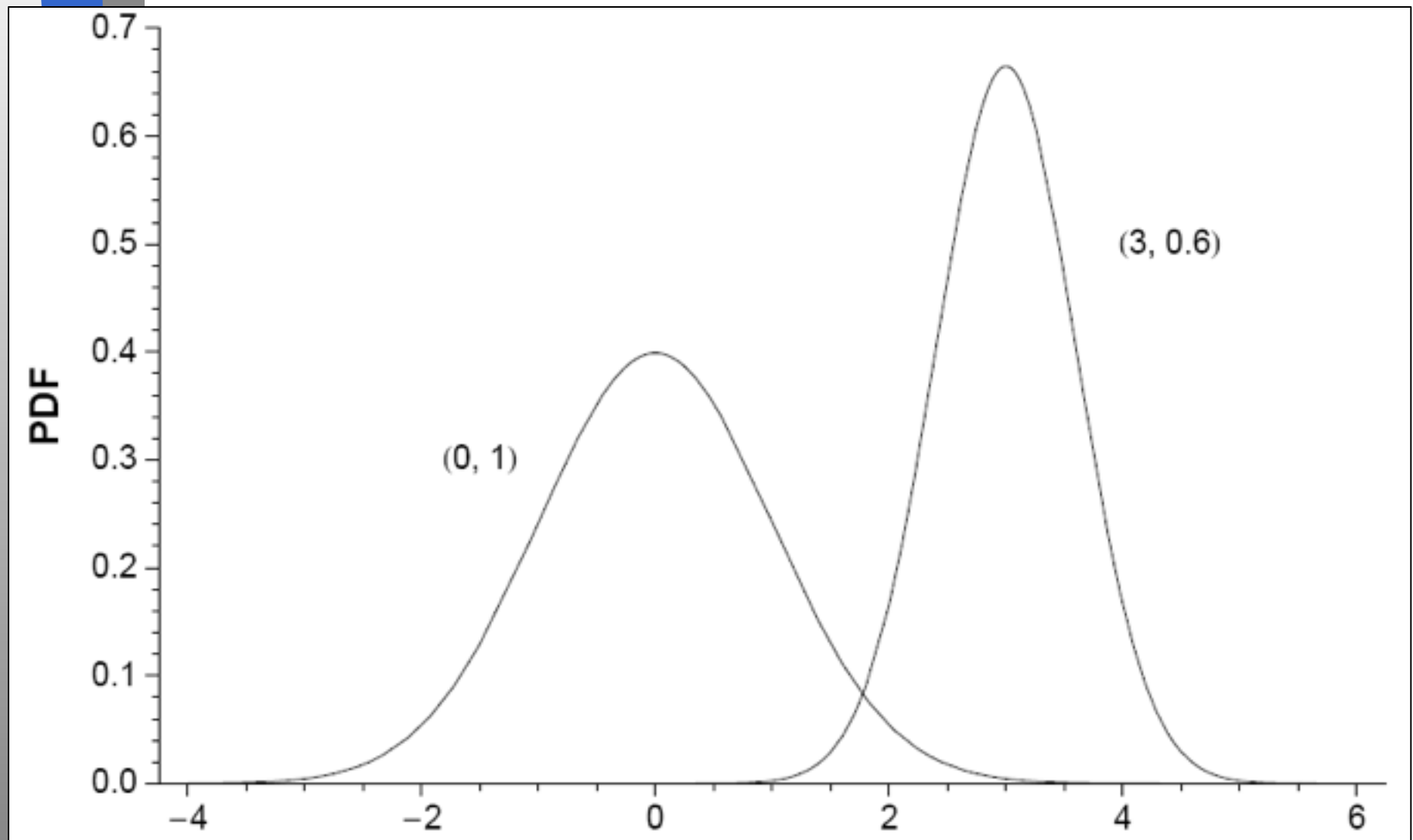
Some standard probability distributions...

The Gaussian (Normal) distribution

- It has the pdf:

$$f_X(x) \equiv \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right]$$

- Exercise 2: show that the Gaussian has mean and variance given by μ , σ^2
- It is often denoted $N(\mu, \sigma^2)$



The Gaussian Central Limit Theorem

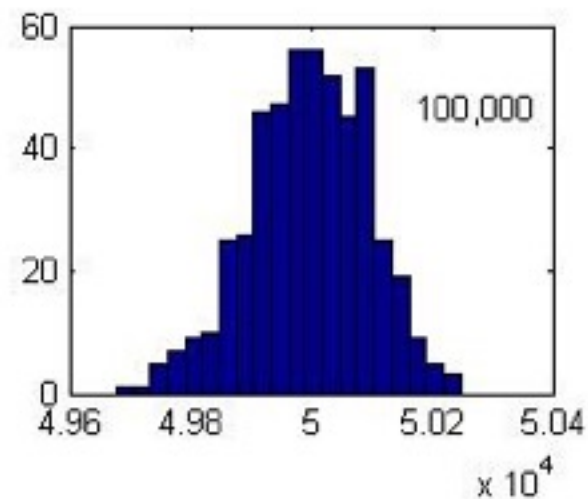
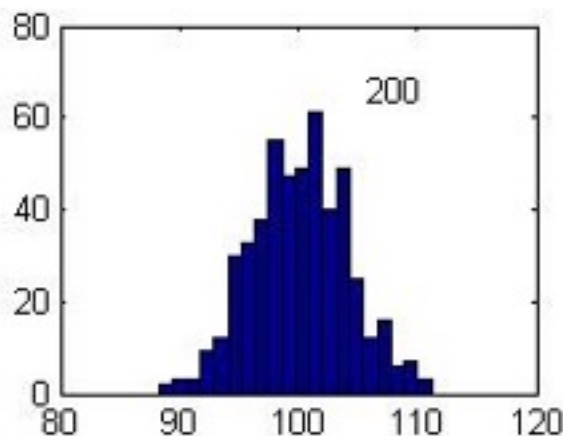
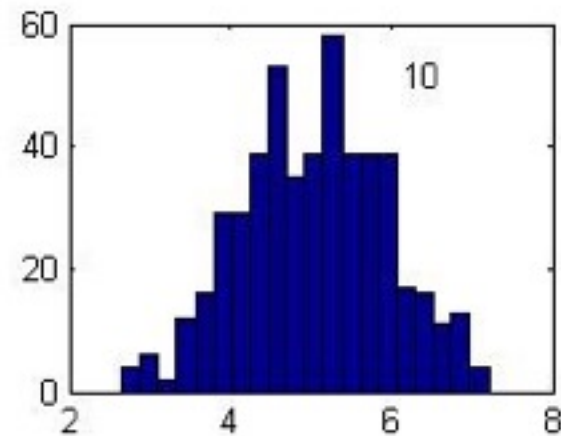
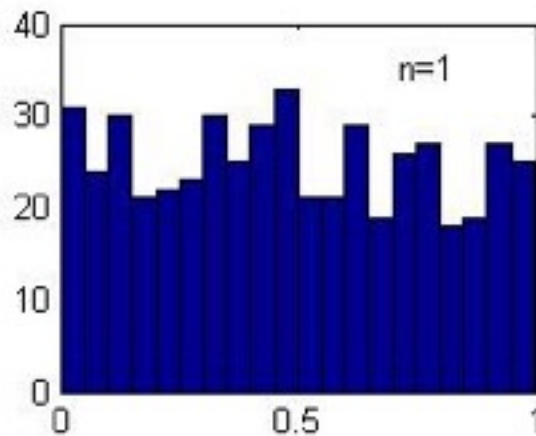
- The sum

$$U_1 + U_2 + \dots U_n$$

of a large number of *independent, identically distributed (iid)* variables, U_i , is Gaussian distributed in the limit $n \rightarrow \infty$.

(True within reason – must have finite mean and variance)

Example: distribution of sums of n uniformly distributed numbers (500)...





Some notation....

Conditional probability

- $P(A|B)$ should be read: the probability of A **assuming (given)** that B is true.

Joint probability

- This is in contrast to: $P(A,B)$ which is the probability that both A *and* B are true.

Relating the two..the product rule

- If X and Y are independent then (*by definition*):

$$P(X, Y) = P(X)P(Y)$$

- But also

$$P(X | Y) = P(X)$$

(X doesn't depend on Y at all)

- Hence we have:
(NB: this is generally true)

$$P(X | Y) = \frac{P(X, Y)}{P(Y)}$$

Bayes theorem

- Since $P(X, Y) = P(Y, X)$ Bayes' theorem follows immediately, which states:

$$\frac{P(X | Y)}{P(Y | X)} = \frac{P(X)}{P(Y)}$$

- Now on to Bayesian statistics...

Bayesian statistics

- We will use Bayes' theorem to evaluate the likelihood of a theory/model/parameter given some data (D)
- Hence we will talk of

$P(D|T)$ or $P(D|\theta)$ or $P(T|D)$

Data

Theory

Parameters

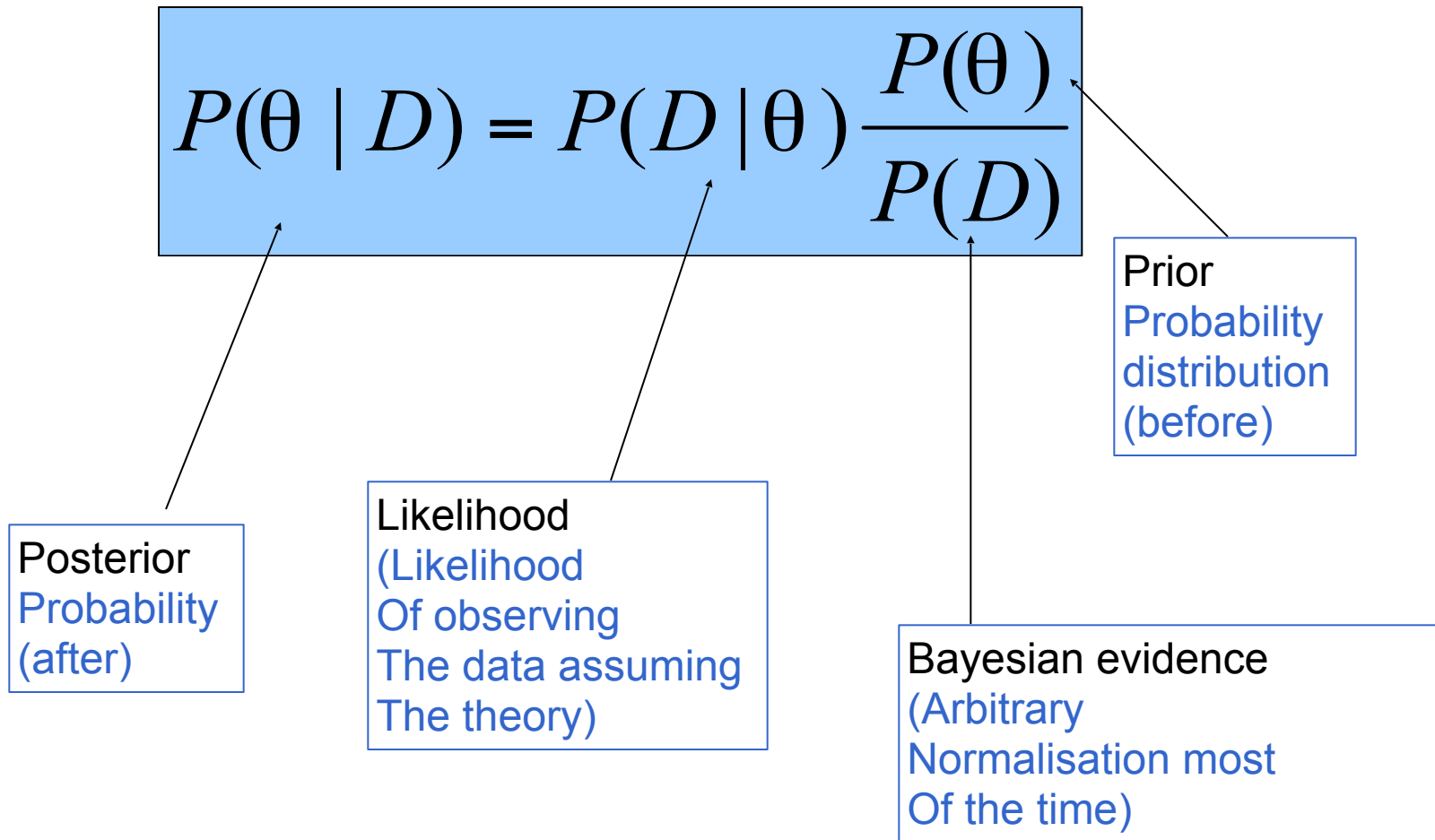
Bayesian statistics II

- If we take $X = \theta$ and $Y = D$ then Bayes' theorem becomes:

$$P(\theta \mid D) = P(D \mid \theta) \frac{P(\theta)}{P(D)}$$

- This allows us to evaluate the probability of the parameters (θ) given the observed data (D), which is what we want to do in science often.

Notation: Priors, likelihoods and posteriors



Priors – $P(T)$

- Bayesian statistics differs from “old-style” frequentist statistics in its use of prior knowledge:
- If you toss a coin 10 times and get 10 heads, what is the probability of getting heads on the 11th toss?
- Frequentist answer: 50%
- Bayesian answer: >95% because the coin is obviously biased!

Maximum Likelihood (posterior) Parameter estimation

- Idea is simple: the parameters that are most likely to be correct are the ones that maximise the posterior probability distribution.
- I.e., one finds the parameter values that maximises

$$P(\theta \mid D) = P(D \mid \theta) \frac{P(\theta)}{P(D)}$$

Max L parameter estimation II

- In practice, since $P(D)$ doesn't depend on the theory, we simply need to maximise:

$$P(D | \theta)P(\theta)$$

- But how do we compute this in reality?
- The prior, $P(\theta)$ is chosen by you, so that is easy.
- The part that needs some discussion is the likelihood, $P(D|\theta)$.

Computing the likelihood

- Let us *assume* that our data consists of N independent data-points, d_i
- Their independence **implies** that:

$$P(d_1, \dots, d_N | \theta) = P(d_1 | \theta) \times P(d_2 | \theta) \times \dots \times P(d_N | \theta)$$

- So the problem reduces to finding the likelihood for single points.

Gaussianity to the rescue...

- Let us assume (often a good approximation) that for each point $d_i = \mu_i + n_i$: the data (d) is the true value (μ) plus noise (n) which has a Gaussian pdf with mean 0 and variance σ_i^2
- Then

$$P(d_i | \theta) \propto \exp \left[-\frac{(d_i - \mu_i)^2}{2\sigma_i^2} \right]$$

- Hence

$$P(d_1, \dots, d_N | \theta) = P(d_1 | \theta) \times P(d_2 | \theta) \times \dots \times P(d_N | \theta)$$

is simply:

$$P(d_1 \dots d_N | \theta) \propto \prod_i \exp \left[-\frac{(d_i - \mu_i)^2}{2\sigma_i^2} \right]$$
$$\propto \exp \left[-\sum_i \frac{(d_i - \mu_i)^2}{2\sigma_i^2} \right]$$

Great, but what now?

- What are the true values μ_j ?
- Well, we are doing **parameter estimation** so we assume that our theory is the correct description of reality, so the μ_j are just the predictions of the theory for the given parameter values.
- Let's denote the theoretical predictions by t_j (θ_α) where θ_α denote the parameters we are trying to estimate.
- Then...

Chi-squared...

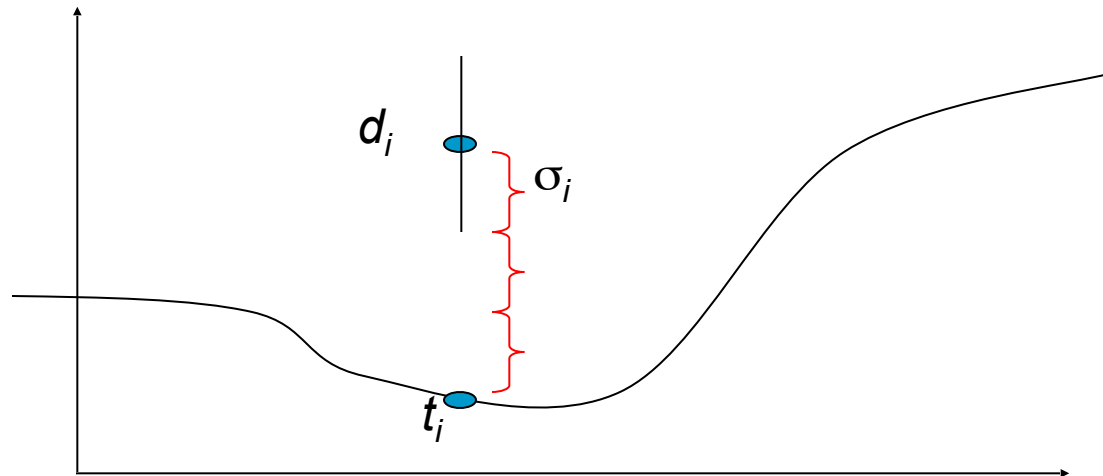
- The likelihood then becomes:

$$L \equiv P(D | \theta_{\alpha}) \propto \exp \left[- \sum_i \frac{(d_i - t_i(\theta_{\alpha}))^2}{2\sigma_i^2} \right]$$
$$\equiv \exp(-\chi^2(\theta_{\alpha}) / 2)$$

- Where χ^2 is the standard chi-squared statistic.

What is the χ^2 ?

$$\chi^2(\theta_\alpha) \equiv \sum_i \frac{(d_i - t_i(\theta_\alpha))^2}{\sigma_i^2}$$



- This point has $\chi^2 \sim 9$

Minimisation

- Maximising the likelihood is **equivalent** to minimising the χ^2 statistic (in the case of a trivial prior)
- As before, the best parameters therefore solve:

$$\frac{\partial \chi^2}{\partial \theta_{\alpha}} = 0 \dots \alpha = 1 \dots n$$

- Note that the constants of proportionality (e.g. $P(D)$) in the posterior don't matter...

Errors on parameters

- In general we don't just want the best-fitting parameters (those that maximise the likelihood) we also want to give **error bars** on the parameters.
- For example we want to be able to say: at 95% confidence level a given parameter θ lies between -1.2 and 0.4 with a best fit of -0.7 .
- How do we do this?

1-sigma, 2-sigma...

- If the random variable, X , is Gaussian distributed with standard deviation, σ , then we can easily calculate probabilities
- $|x - \mu| < \sigma$ with about 68% probability
- $|x - \mu| < 2\sigma$ with about 95% probability
- $|x - \mu| < 3\sigma$ with about 99.7% probability

Typically results are quoted in the format:

$$\Omega_{DE} = 0.71^{+0.1}_{-0.1}$$

1-sigma error
bars

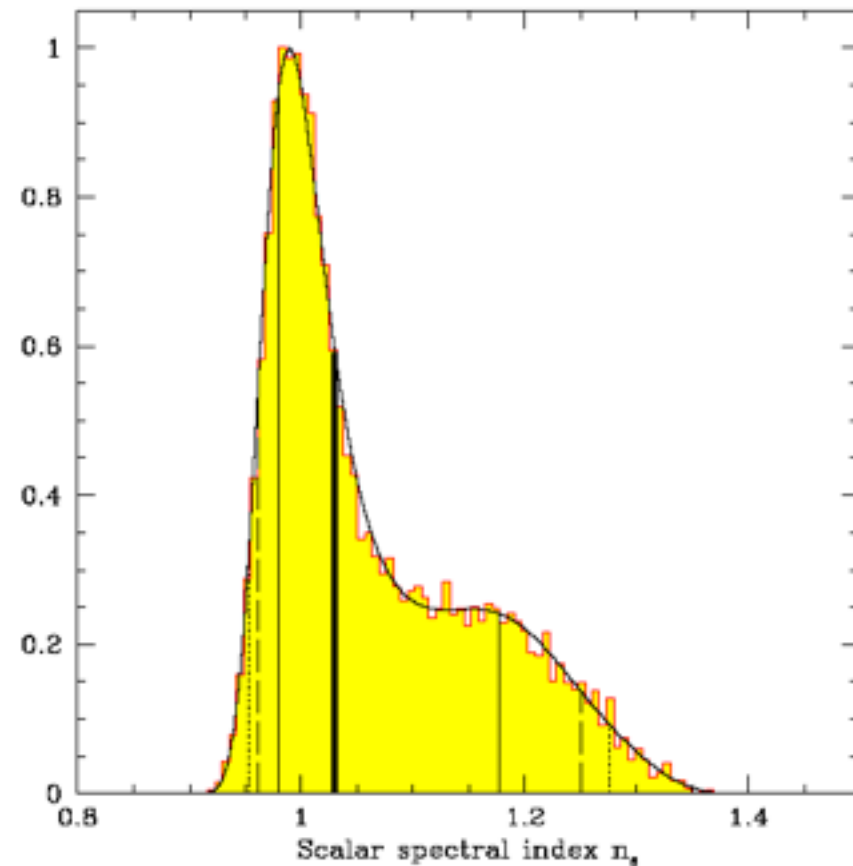
Asymmetric probabilities

What happens if the distribution is very non-Gaussian?

Then we simply compute confidence limits numerically as *percentiles*.

In these cases we can end up with asymmetric error bars...e.g.

Tegmark et al 04

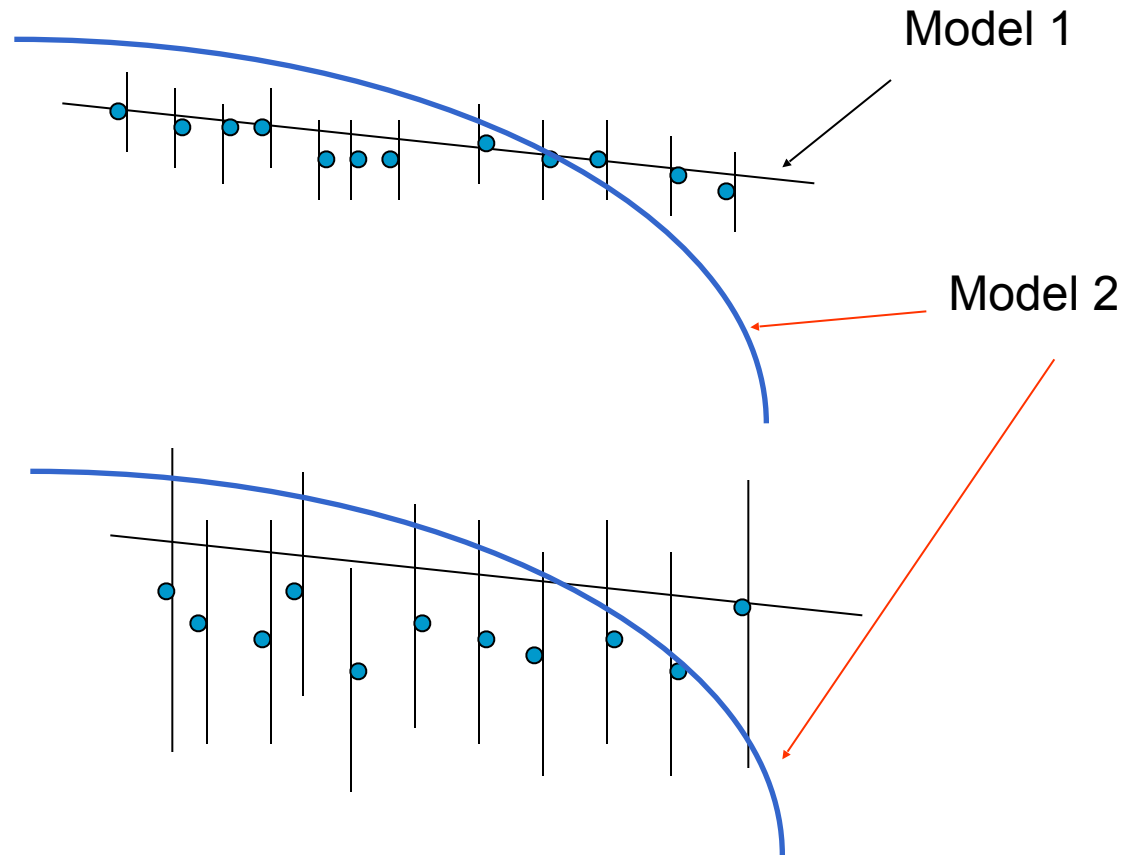


$$\Omega_{DE} = 0.71^{+0.21}_{-0.05}$$

Error bar sizes

Here model 2 is ruled out by the data at high confidence.

Here model 2 is **not** ruled out by the data.



Some basic points about errors...

- Why do error bars matter at all?
- Consider the recent WMAP result that the best-fit curvature is:

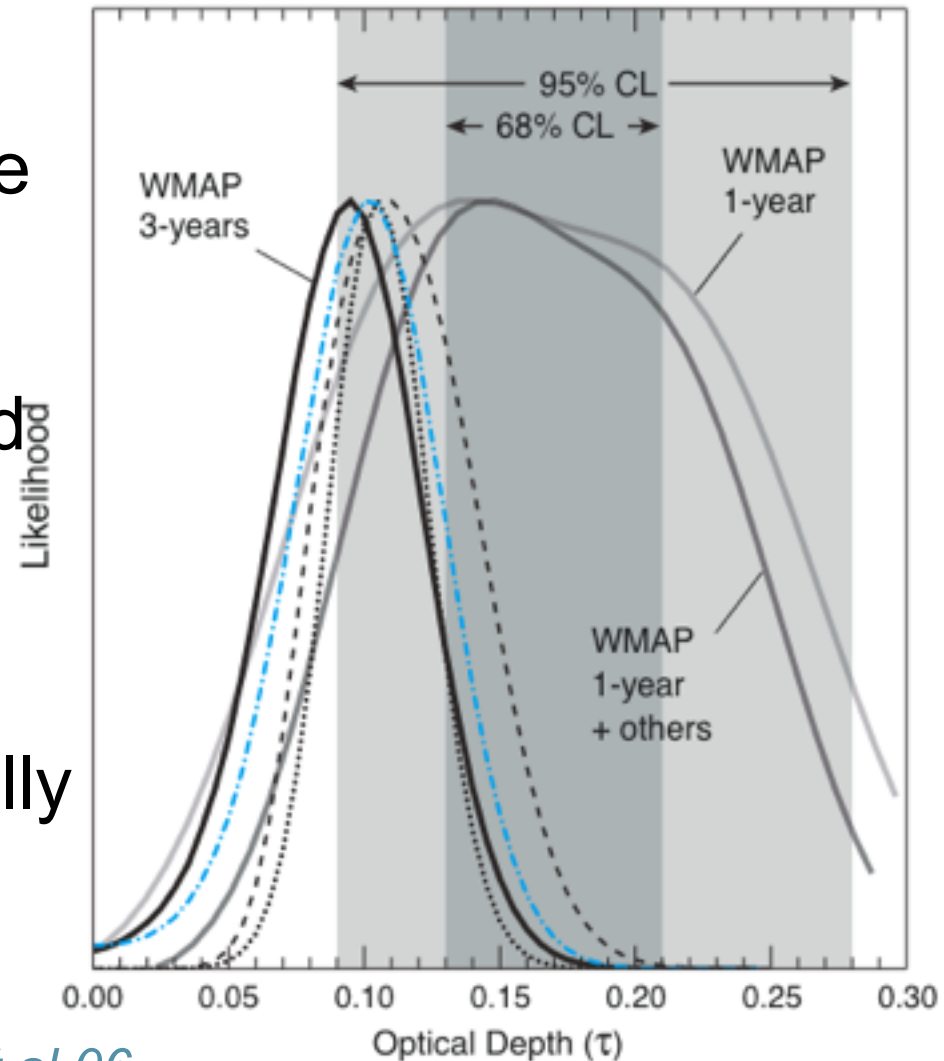
$$\Omega_{tot} = 1.02$$

- Is this a “real detection” of spatial curvature?
- This depends on the error bar. The full result is approximately: $\Omega_{tot} = 1.02^{+0.02}_{-0.02}$
- Hence, only 1σ away from 1 – assuming Gaussianity, this means that with $\sim 16\%$ probability the universe is open and $\sim 84\%$ closed.
- This may seem significant but it is actually very weak. Usually we require 3σ ($>99\%$)

Why even 2-sigma is not enough...

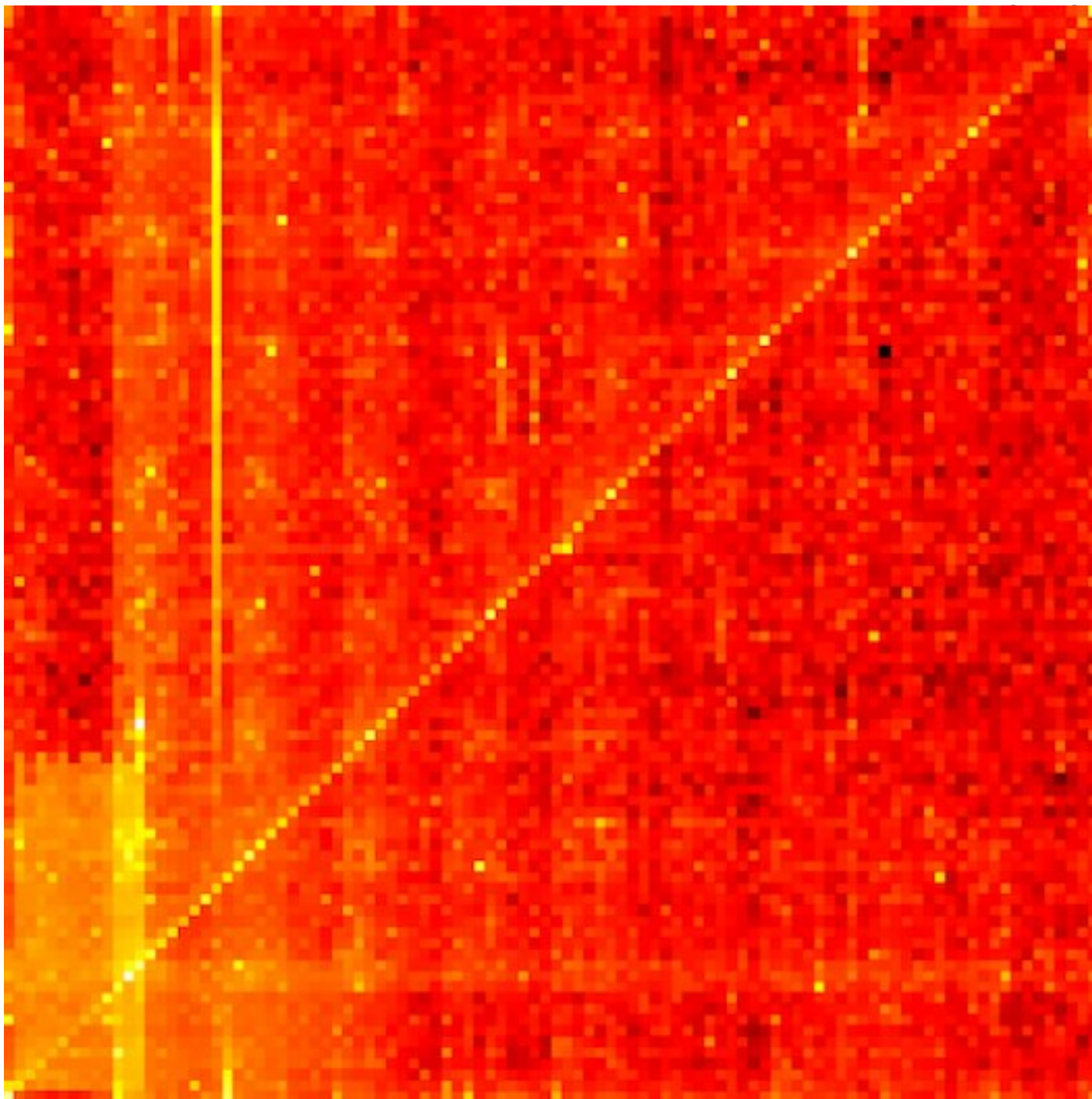
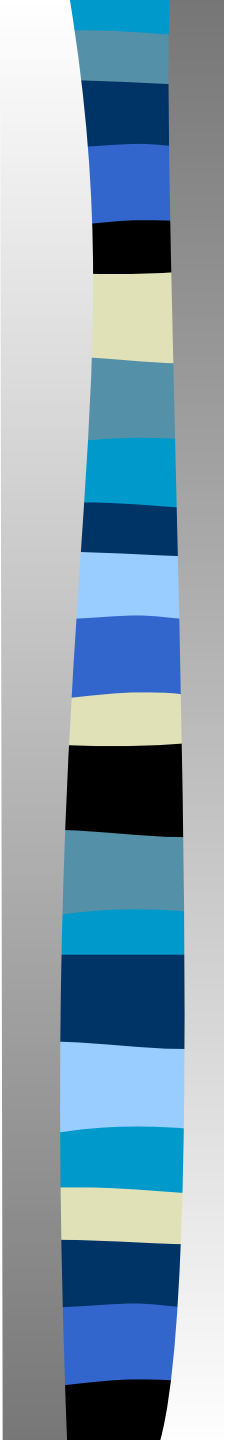
✦ The improved polarisation measurements in the 3rd year WMAP results have significantly changed the best-fit optical depth (by 2-sigma)

✦ Moral: 2-sigma is usually not statistically significant (systematics!)



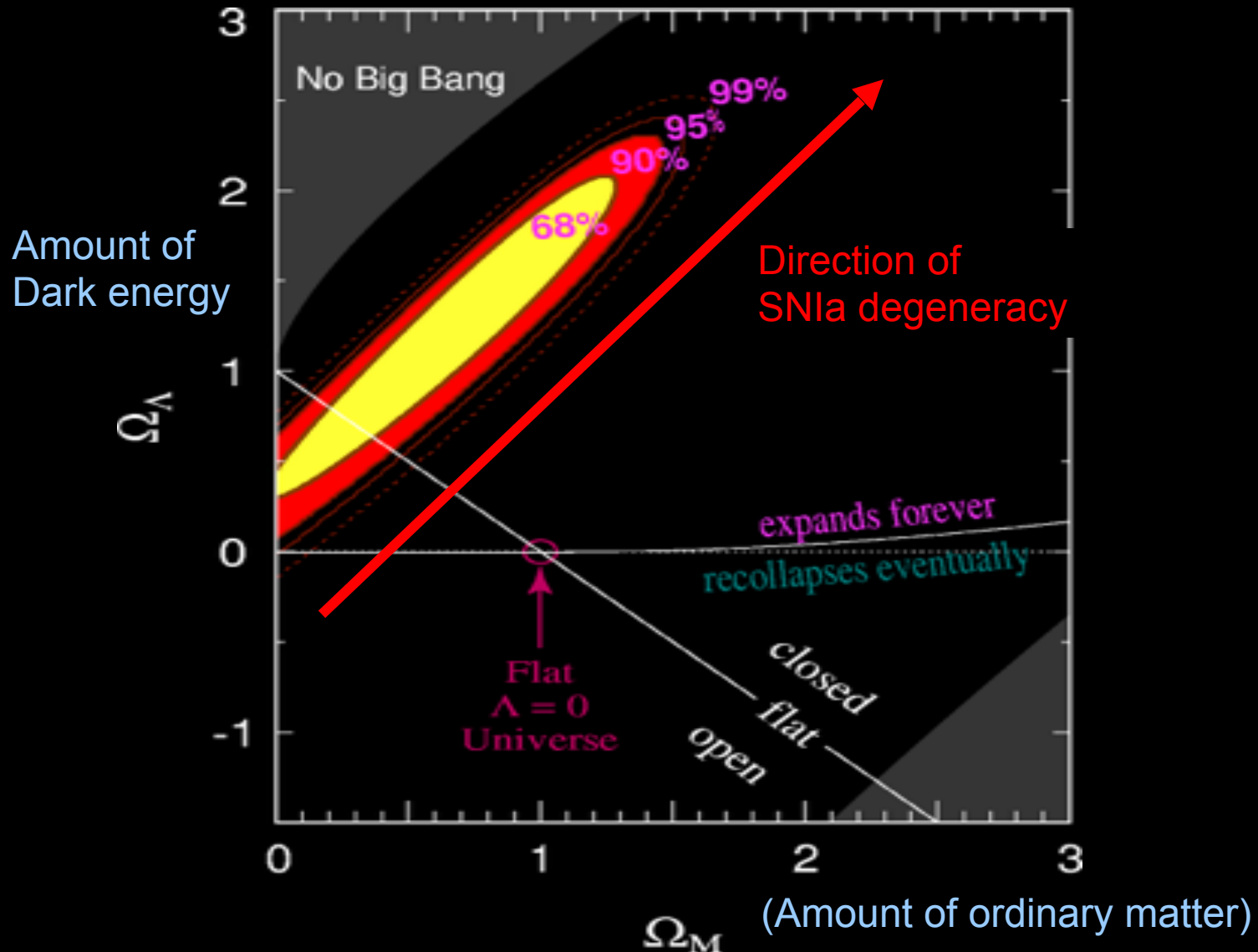
Multiple parameters

- If we consider two parameters we don't have error bars, but **error ellipses** which give the region within which the parameters lie with a given probability (e.g. 95%, 99% etc...)
- Often however the effect of changing one parameter can be exactly offset by changing another one...
- This implies a **degeneracy** – one can only measure accurately a combination of the parameters. E.g. given the equation $xy = 1$, one cannot measure x or y *separately*.
- As a result, the error on a parameter is always larger (**or at best =**) if other parameters are simultaneously being estimated from the same data.



OS

Example of error ellipses – SNIa data



How do we combine data from multiple experiments?

- If datasets are statistically independent simply compute the χ_i^2 for each experiment and then simply **sum them** to compute a total chi-squared:

$$\chi_{tot}^2 = \sum_i \chi_i^2$$

- Then minimise this total chi-squared to estimate the best-fit parameters...
- If the data **are** correlated, then combining them is non-trivial - one needs to know how they are correlated.

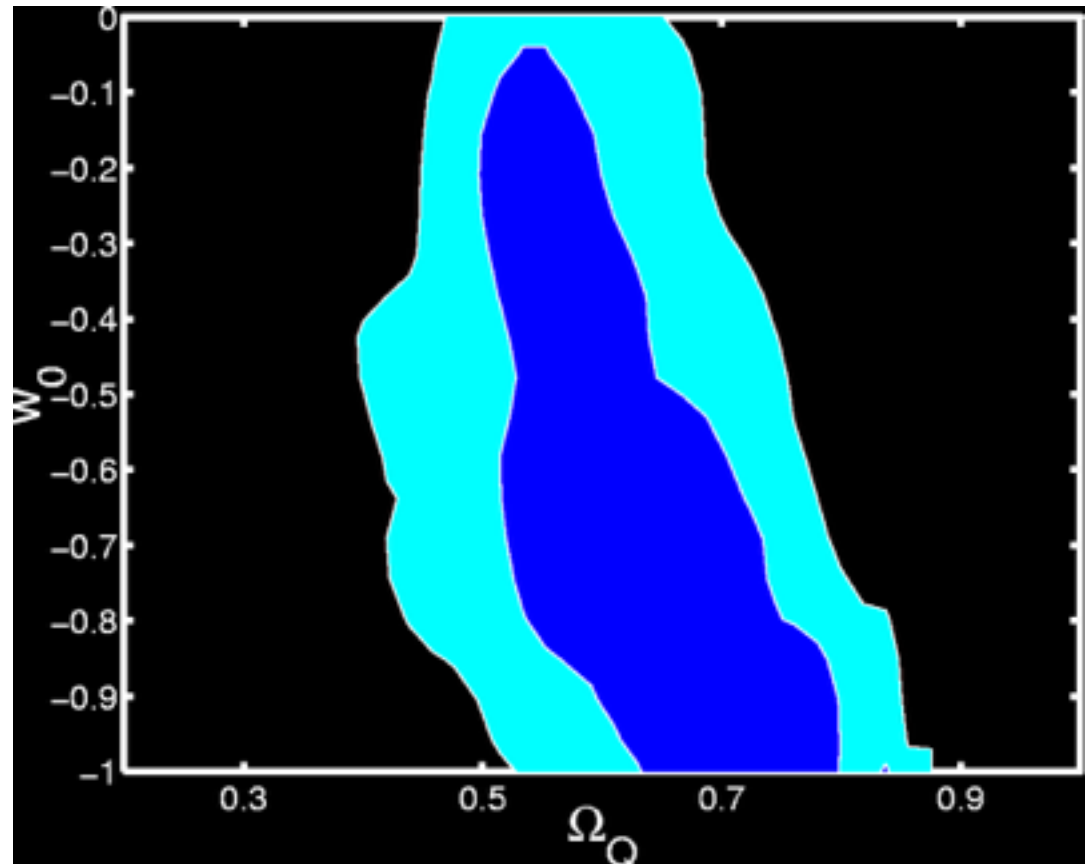
Degeneracies in the CMB

Corasaniti et al. 2005

We can decrease $w_0 = p/\rho$
By lowering Ω_{DE} and H_0
without changing the CMB.

→ degeneracy in the CMB

→ WMAP alone constrains
dark energy models
badly



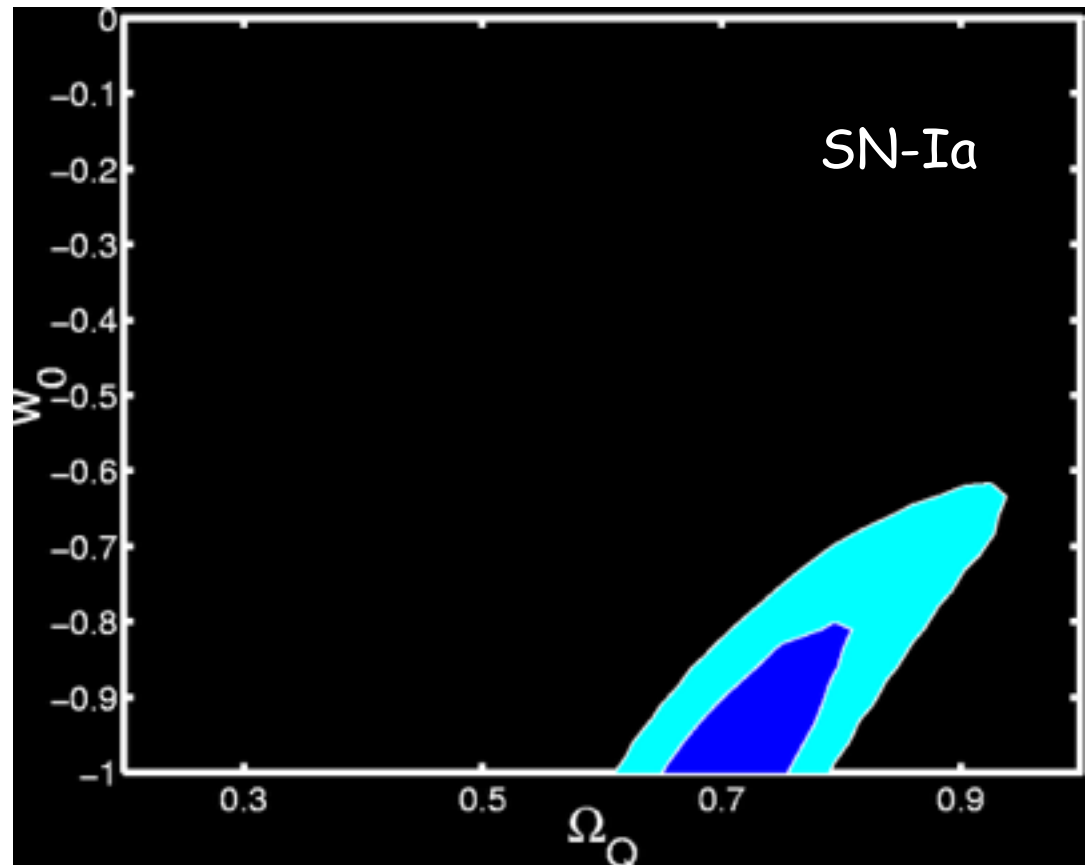
(Fraction of total energy density in DE)⁴³

degeneracies in the SNIa

Kunz

A different degeneracy exists if we use the supernova Data.

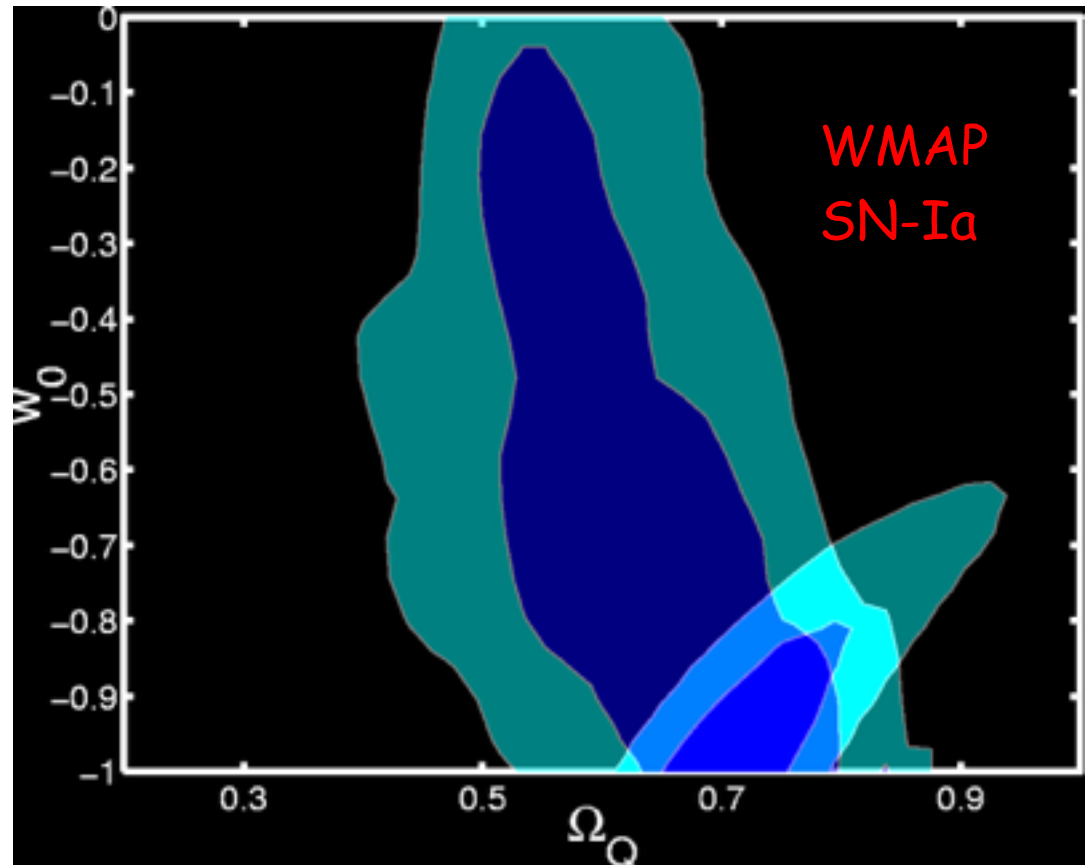
But it runs in another direction



Breaking Degeneracies by combining different data

Kunz

By combining the two data sets, we can constrain the parameters much more tightly.

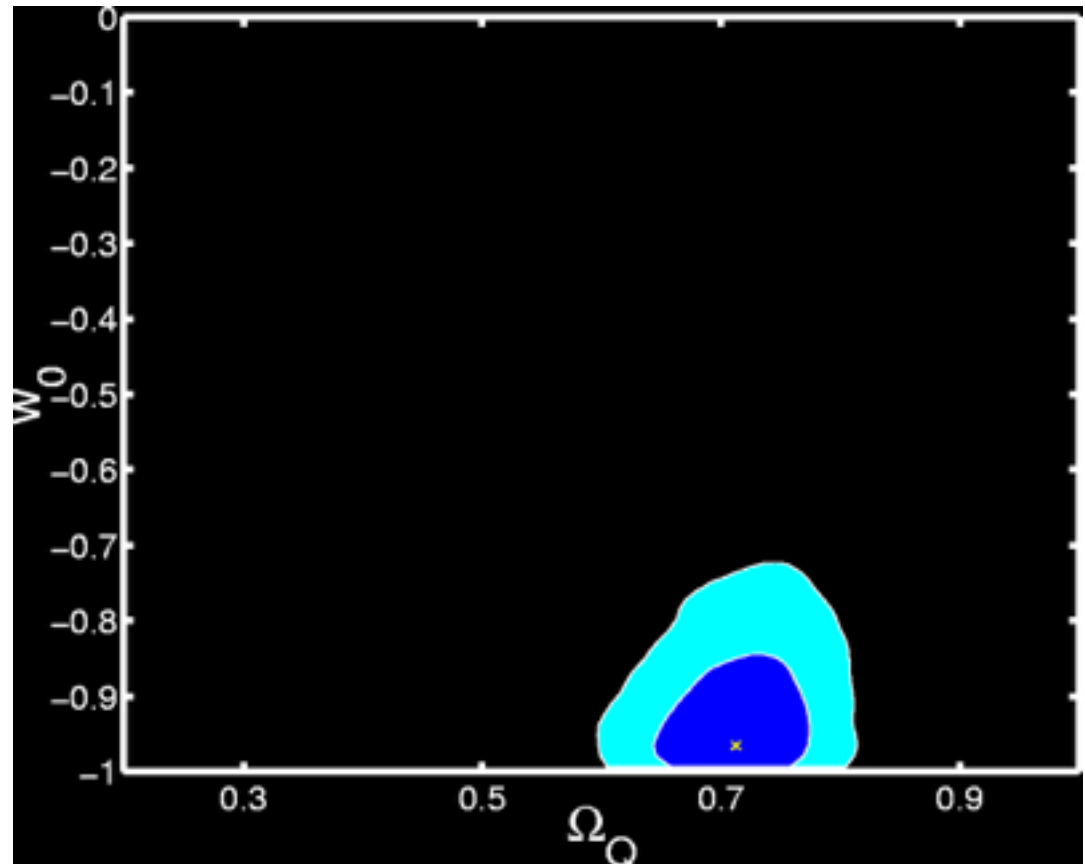


The accelerating cosmos

Kunz

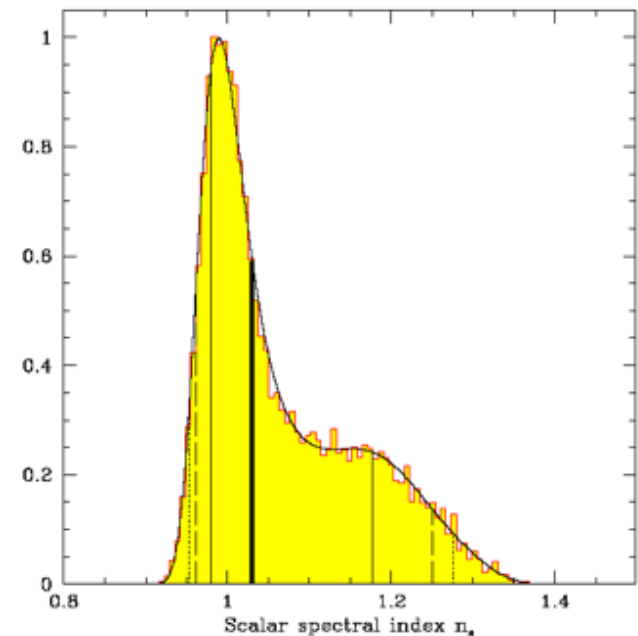
The combined data gives good evidence for the Cosmos being dominated By energy with significant negative pressure ($w < -1/3$)

Hence good evidence for Cosmic acceleration



Marginalisation

- If we are estimating 100 parameters we end up with a **100 dimensional** posterior distribution.
- But if we want to estimate errors we need a **one-dimensional** likelihood such as:



Marginalisation II

- The process of moving from the full posterior to just a **one or two dimensional** likelihood is called **marginalisation** and is given by integrating the full posterior over all other parameters. E.g.:

$$P(\theta_2) = \int P(\theta_1, \theta_2, \dots, \theta_n) d\theta_1 \cdot d\theta_3 \dots d\theta_n$$

Marginalisation III

- This is actually just the natural generalisation of a standard law of probability:
- $P(X) = P(X,Y) + P(X,Z) + \dots$
where the sum is over all possible joint probabilities involving X.
- E.g. the probability of finding my keys is the sum of the joint probabilities that I find it with my papers, my laptop, my pens etc... (including all possible combinations)

Marginalisation IV

- The problem with marginalisation is that for high-dimensional posteriors, it is computationally very intensive (n-dimensional numerical integration).
- Again MCMC is a great way of doing marginalisation as we later discuss...