Evaluating cosmological tensions using posterior predictions

David Parkinson
KASI

In collaboration with
Shahab Joudaki (Oxford)

Outline

- Introduction
- Statistical Inference
- Posterior predictions
- Example using WL and CMB data
- Conclusions

Quotes about statistics

- “There are lies, damned lies and statistics.”
  - Mark Twain
- “Statistics are used much like a drunk uses a lamppost: for support, not illumination.”
  - Vic Scully
- “There are two ways of lying. One, not telling the truth and the other, making up statistics.”
  - Josefina Vazquez Mota
- “Definition of Statistics: The science of producing unreliable facts from reliable figures.”
  - Evan Esar
- “Statistics are no substitute for judgment”
  - Henry Clay

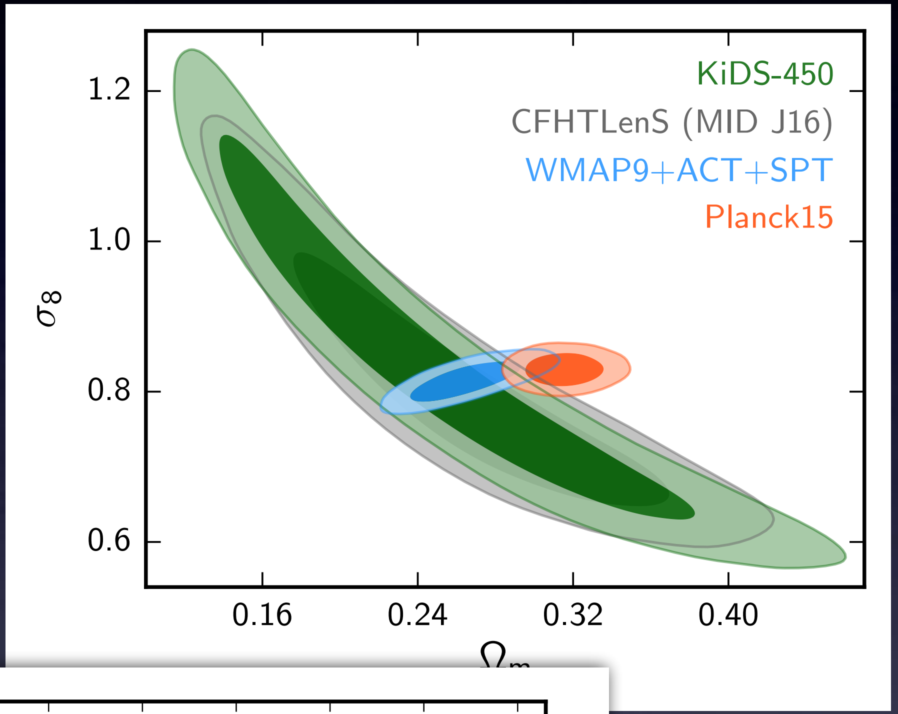
What is Probability?

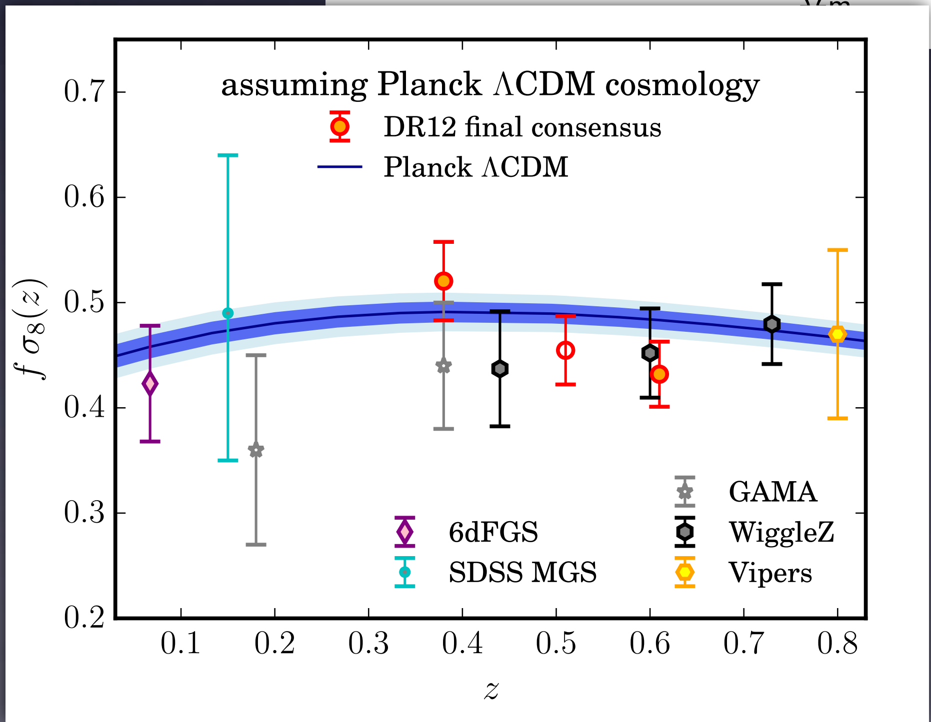
- In 1812 Laplace published *Analytic Theory of Probabilities*
- He suggested the computation of *"the probability of causes and future events, derived from past events"*
- *"Every event being determined by the general laws of the universe, there is only probability relative to us."*
- *"Probability is relative, in part to [our] ignorance, in part to our knowledge."*
- So to Laplace, probability theory is applied to our level of knowledge

Pierre-Simon Laplace

Comparing datasets

- As there is only one Universe (setting aside the Multiverse), we make observations of un-repeatable ‘experiments’
- Therefore we have to proceed by inference
- Furthermore we cannot check or probe for biases by repeating the experiment - we cannot ‘restart the Universe’ (however much we may want to)
- If there is a tension (i.e. if two data sets don’t agree), can’t take the data again. Need to instead make inferences with the data we have





Alam et al 2016

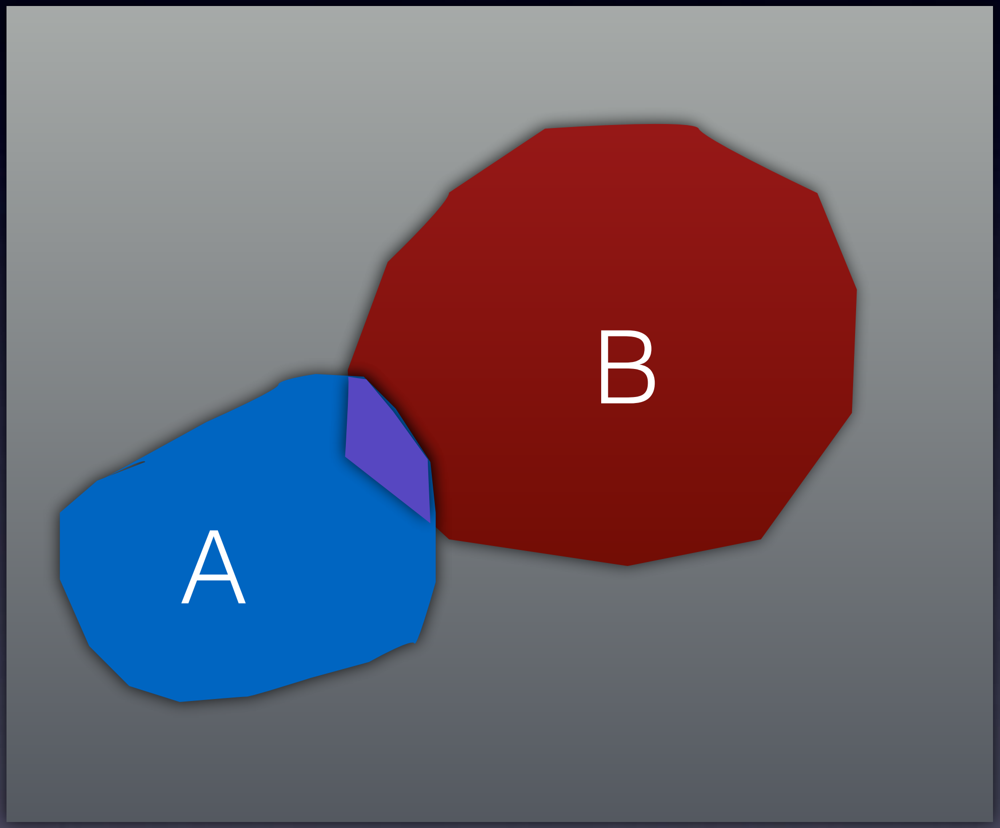
Types of questions

- There are three types of questions we can use statistics to answer

1. *The probability of data*, given some causes.
2. *The probability of parameter values*, given some model, which can be updated through observations.
3. *The probability of the model*, which can also be updated by observation.

Rules of Probability

- We define Probability to have numerical value
- We define the lower bound, of logical absurdities, to be zero, $P(\varnothing)=0$
- We normalize it so the sum of the probabilities over all options is unity, $\sum P(Ai)\equiv 1$



Sum Rule: $P(A \cup B)=P(A)+P(B)-P(A \cap B)$

Product Rule: $P(A \cap B)=P(A)P(B|A)=P(B)P(A|B)$

Bayes Theorem

- Bayes theorem is easily derived from the product rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- We have some model M, with some unknown parameters θ, and want to test it with some data D

likelihood

prior

posterior

$$P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}$$

evidence

- Here we apply probability to models and parameters, as well as data

Model Selection

- If we marginalize over the parameter uncertainties, we are left with the marginal likelihood, or evidence

evidence likelihood prior

$$E = P(D|M) = \int P(D|\theta, M) P(\theta|M) d\theta$$

- If we compare the evidences of two different models, we find the Bayes factor

Model posterior evidence Model prior

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)}$$

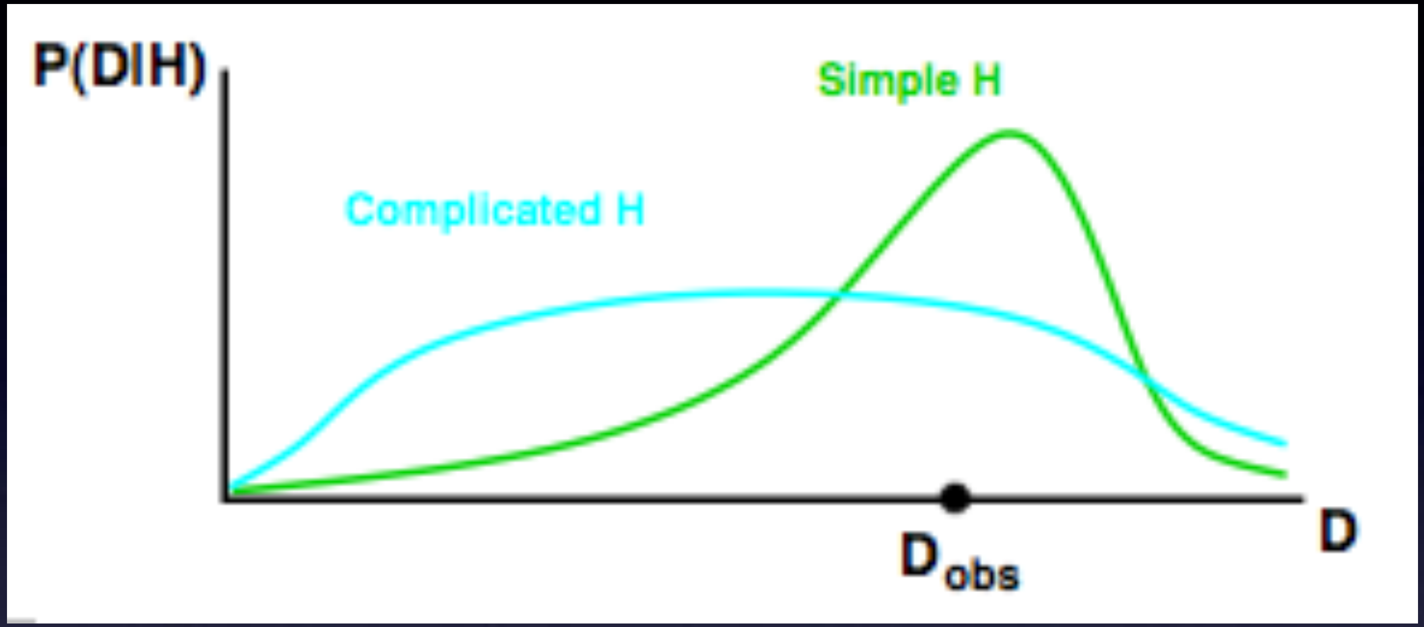
- Bayes theorem provides a consistent framework for choosing between different models

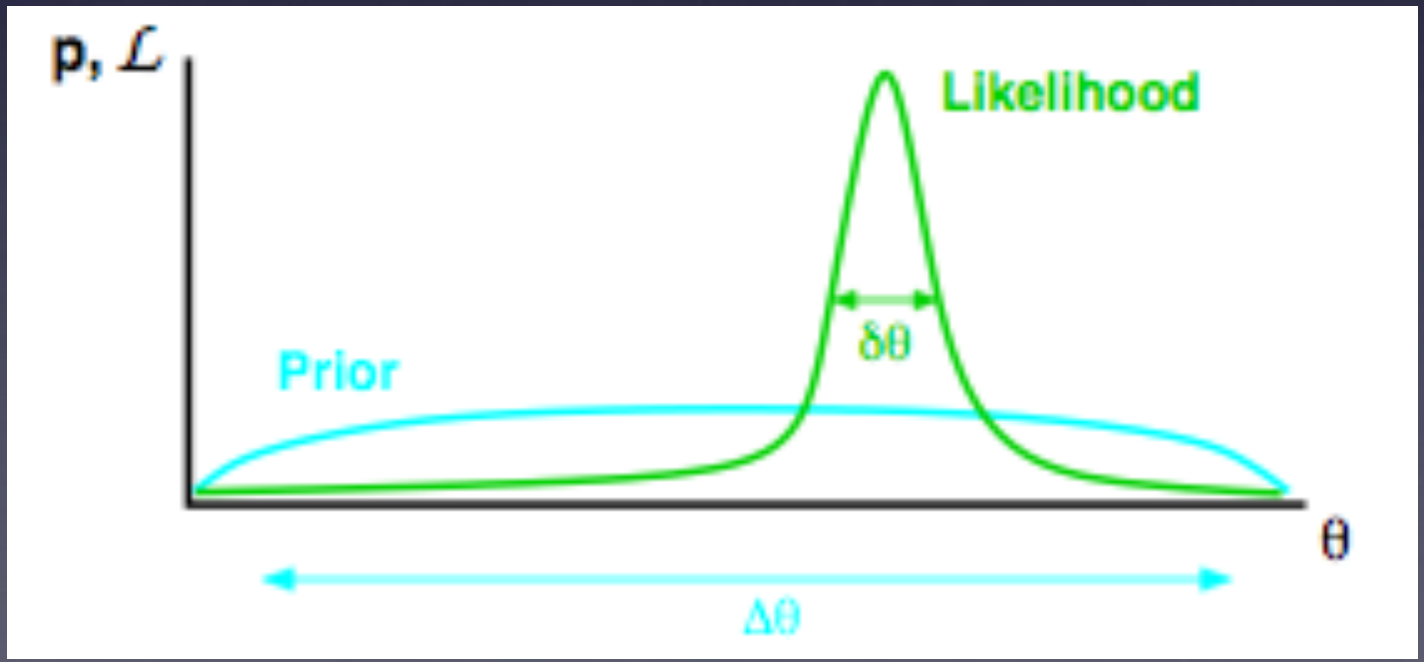
Occam's Razor

$$E = \int d\theta P(D|\theta, \mathcal{M}) P(\theta|\mathcal{M})$$

$$\approx \underbrace{P(D|\hat{\theta}, \mathcal{M})}_{\text{Best fit likelihood}} \times \underbrace{\frac{\delta\theta}{\Delta\theta}}_{\text{Occam factor}}$$

- Occam factor rewards the model with the least amount of wasted parameter space ("most predictive")





Bayesian Model Comparison

- Jeffrey's (1961) scale:

Difference	Jeffrey	Trotta	Odds
$\Delta\ln(E)<1$	No evidence	No	3:1
$1<\Delta\ln(E)<2.5$	substantial	weak	12:1
$2.5<\Delta\ln(E)<5$	strong	moderate	150:1
$\Delta\ln(E)>5$	decisive	strong	>150:

- If model priors are equal, evidence ratio and Bayes factor are the same

Information Criteria

- Instead of using the Evidence (which is difficult to calculate accurately) we can approximate it using an Information Criteria statistic
- Ability to fit the data (chi-squared) penalised by (lack of) predictivity
- Smaller the value of the IC, the better the model
- Bayesian Information Criterion (Schwarz, 1978) - point estimate approximation to the evidence

$$\mathrm{BIC} = \chi^2(\hat{\theta}) + k \ln N$$

  - k is the number of free parameters and N is the number of data points

Complexity

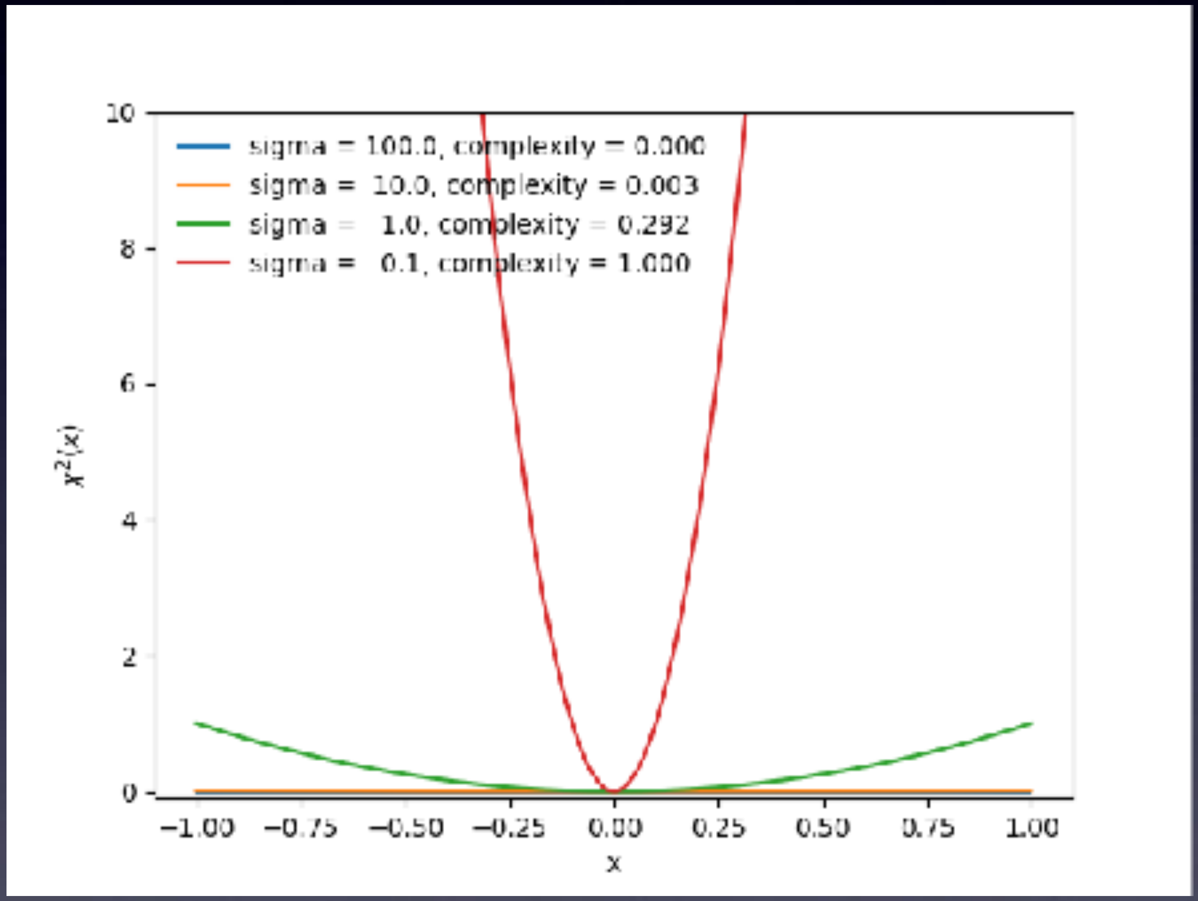
- The DIC penalises models based on the *Bayesian complexity*, the number of well-measured parameters
- This can be computed through the information gain (KL divergence) between the prior and posterior, minus a point estimate

$$\mathcal{C}_b = -2\left(D_{\mathrm{KL}}\left[P(\theta|D,\mathcal{M})P(\theta|\mathcal{M})\right] - \widehat{D_{\mathrm{KL}}}\right)$$

- For the simple gaussian likelihood, this is given by

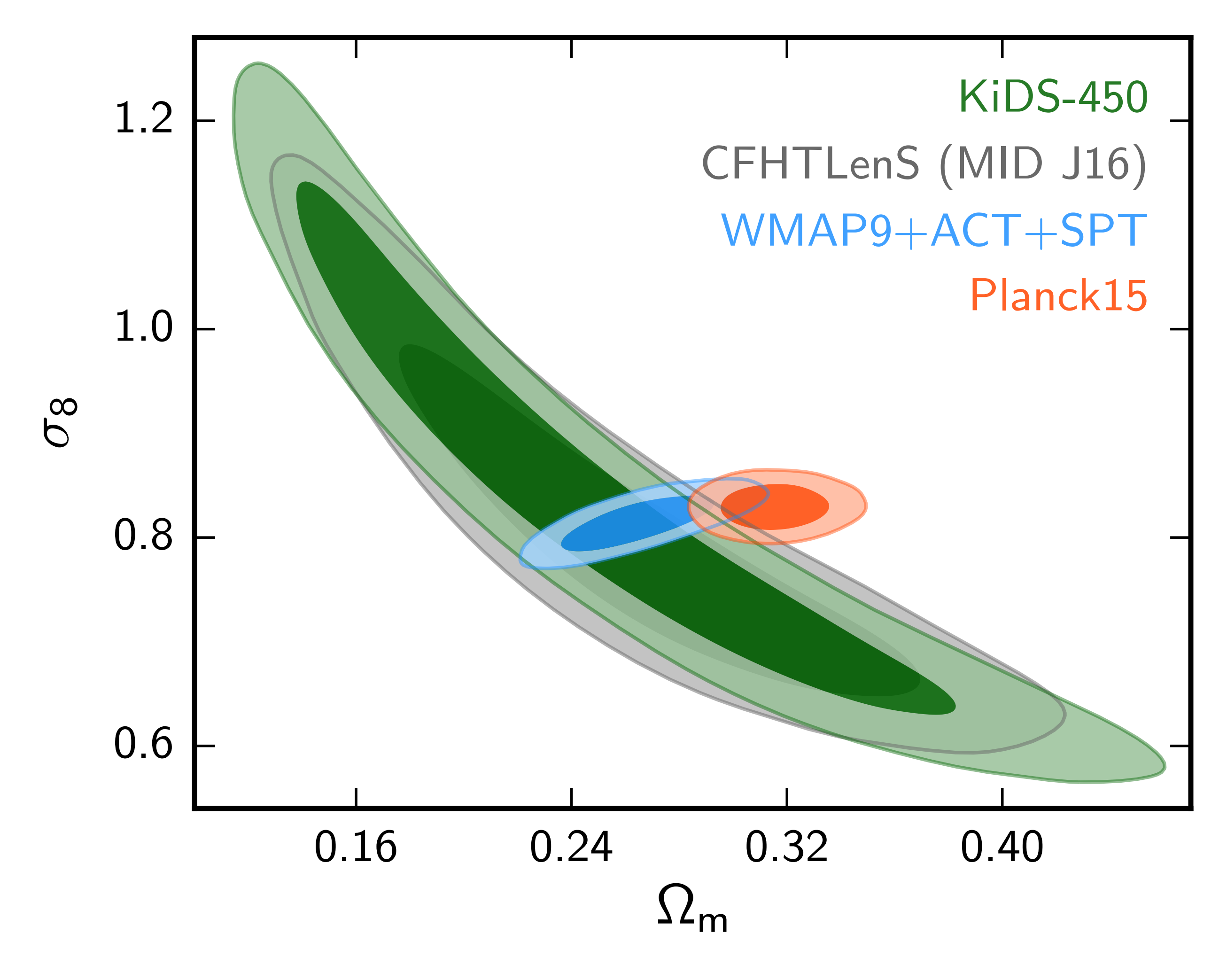
$$\mathcal{C}_b = \overline{\chi^2(\theta)} - \chi^2(\bar{\theta})$$

- Average is over posterior



Tensions

- Tensions occur when two datasets have different preferred values (posterior distributions) for some common parameters
- This can arise due to
  - random chance
  - systematic errors
  - undiscovered physics



Forward modelling

- The goal of the game is to "extract" the plastic teeth from a crocodile toy's mouth by pushing them down into the gum. If the "sore tooth" is pushed, the mouth will snap shut on the player's finger
- Bayes theorem allows for forward modelling of the data
- Based on our previous experience (how many teeth have been pushed down), and model (how many teeth remain), we update our probability of a new outcome

Data validation

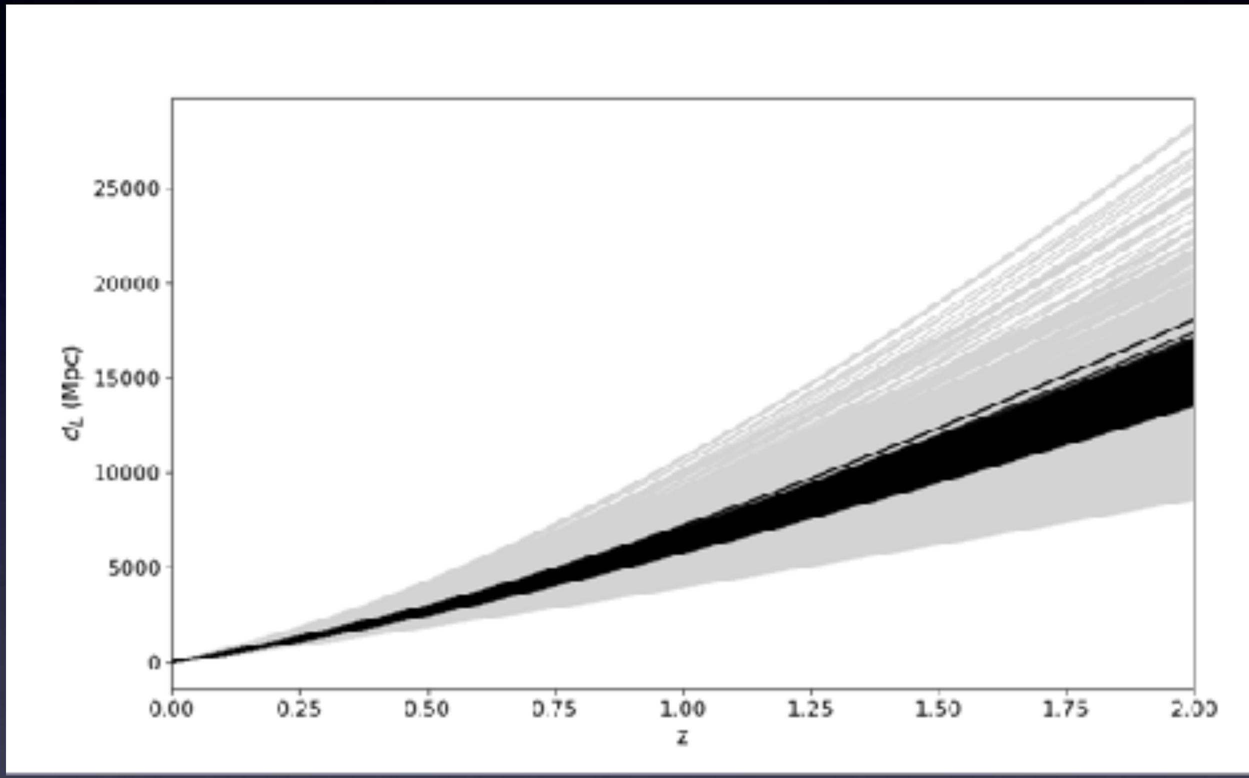
- How can we use Bayesian statistics to make inferences about the data itself?
- Prior predictive distribution

$$P(\{\tilde{D}\}|\mathcal{M}) = \int P(\{\tilde{D}\}|\theta, \mathcal{M}) P(\theta|\mathcal{M}) d\theta$$

- Posterior predictive distribution

$$P(\{\tilde{D}_2\}|D_1, \mathcal{M}) = \int P(\{\tilde{D}_2\}|\theta, \mathcal{M}) P(\theta|D_1, \mathcal{M}) d\theta$$

- We can compare predictive data to actual repetitions or further observations to validate data



Posterior predictive p-value

- Consider some test statistic *T(D)*, which we use for checking for discrepancy
- For the next observation or repetition, the posterior predictive distribution for $T(D_2)$ is given by
$$P\left(T(\tilde{D}_2|D_1)\right) = \int P\left(T(\tilde{D}_2|\theta)\right) P(\theta|D_1)d\theta$$
- The posterior predictive p-value is the cumulative probability for which the predicted value of the test statistic exceeds the actual measured value (using the new data)
$$p = P\left(T(\tilde{D}_2) > T(D_2)\Big|D_1, \theta\right)$$

Procedure

1. Make predictions for data using prior, and current data
2. Take new data
3. Validate data against prior
    a. If bad match, either check analysis pipeline, or reconsider prior (and return to step 1)
4. Validate data against previous data
    a. If tension exists, either check analysis pipeline for both datasets or reconsider prior (and return to step 1)
5. If current and new data are in good agreement, then make posterior inferences and model selection

Diagnostic statistics

- Simple test χ^2 per degree of freedom
  - Equivalent to frequentist p-value test on data, but weighted by posterior predictions
- Raveri (2015): the evidence ratio

$$\mathcal{C}(D_1, D_2, \mathcal{M}) = \frac{P(D_1 \cup D_2|\mathcal{M})}{P(D_1|\mathcal{M})P(D_2|\mathcal{M})}$$

  - Posterior predicted p-value of the normalised likelihood of the second dataset D_2, tested with respect to D_1.
- Joudaki et al (2016): change in DIC

$$\Delta\mathrm{DIC} = \mathrm{DIC}(D_1 \cup D_2) - \mathrm{DIC}(D_1) - \mathrm{DIC}(D_2)$$

Simple linear model

- Frequentist p-value - evaluate χ^2 at best fit, and compare with cumulative density function
- Bayesian p-value - average new χ^2 over while range, weighted by previous posterior
- In simple linear case, with wide (and flat) priors, reduces to difference in χ^2 between first dataset and second, averaged over prior range

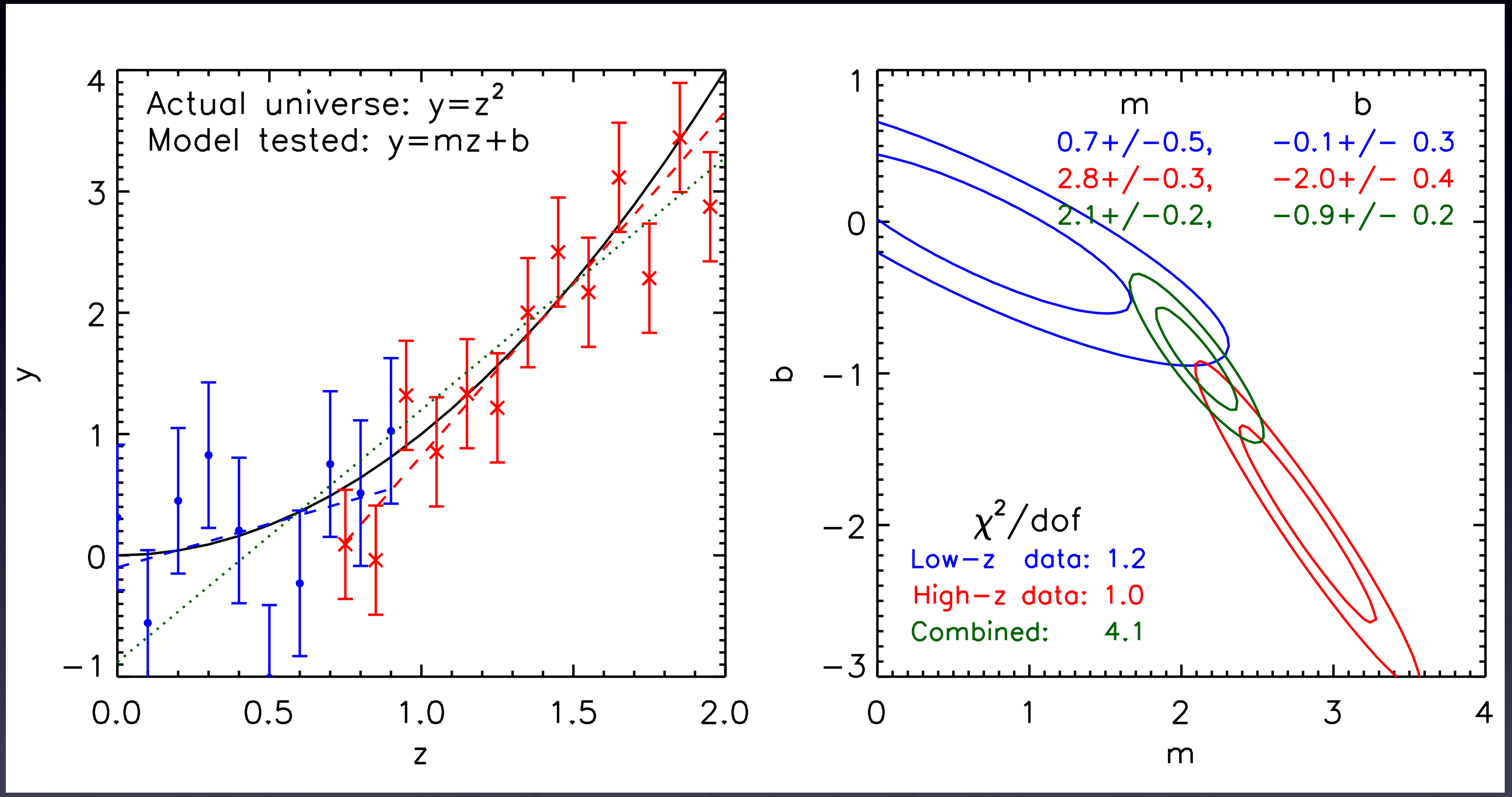


Image credit: Tamara Davis

Diagnostics II: The Surprise

- Seehars et al (2016): the 'Surprise' statistic, based on cross entropy of two distributions
- Cross entropy given by KL divergence

$$D_{\mathrm{KL}}\left(P(\theta|D_2)||P(\theta|D_1)\right) = \int P(\theta|D_2)\log\left[\frac{P(\theta|D_2)}{P(\theta|D_1)}\right]$$

- Surprise is difference of observed KL divergence relative to expected
  - where expected assumes consistency

$$S \equiv D_{\mathrm{KL}}\left(P(\theta|D_2)||P(\theta|D_1)\right) - \langle D\rangle$$

- Not a posterior prediction test - average is over new posterior

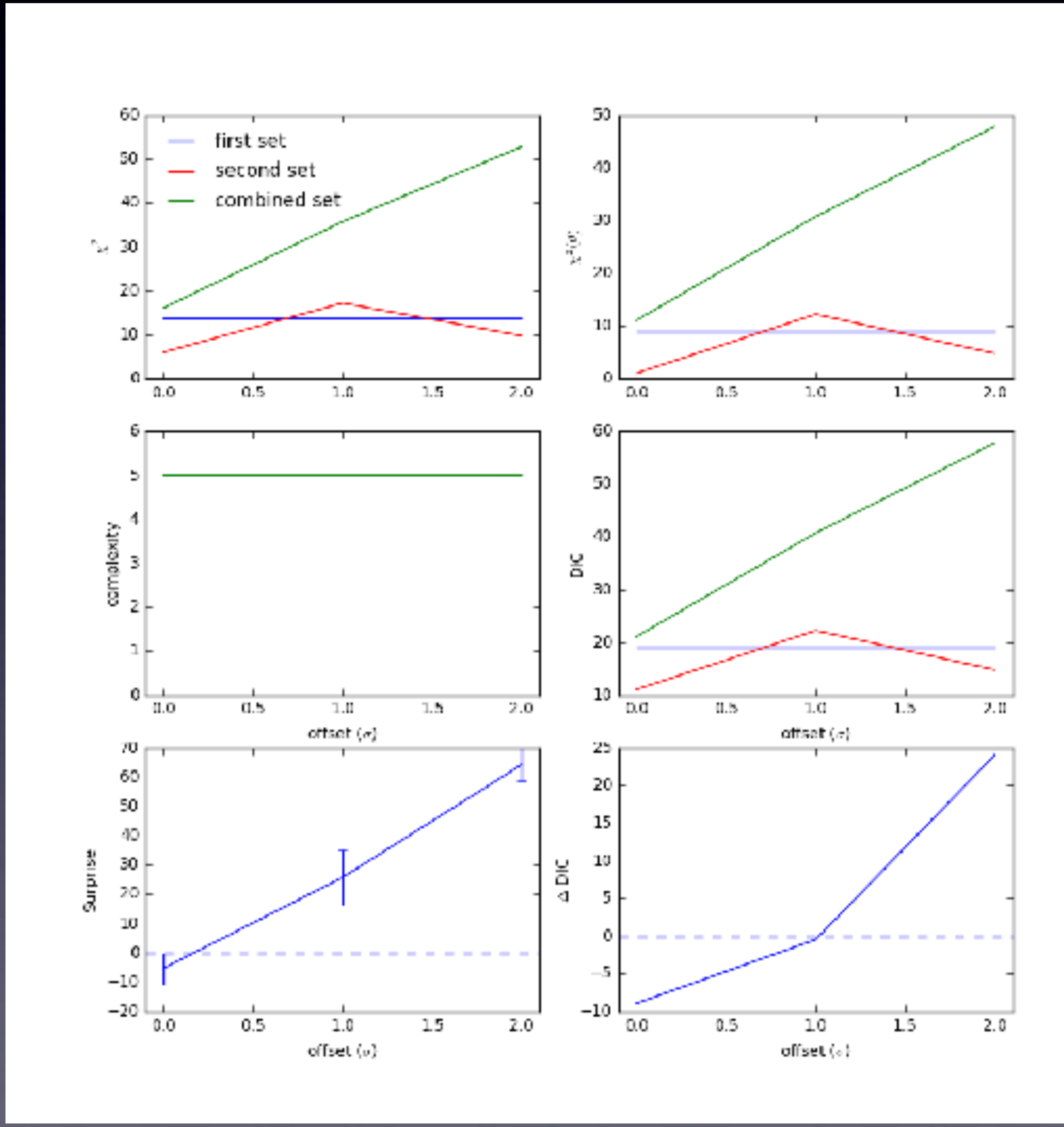
Pros and Cons

Approach	Like ratio	Evidence	DIC	Surprise
Average over parameters	(Yes)	Yes	Yes	Yes
From MCMC chain	Yes	No	Yes	Yes
Probabalistic	Yes	Yes	Yes	No
Symmetric	Yes	Yes	Yes	No

DIC

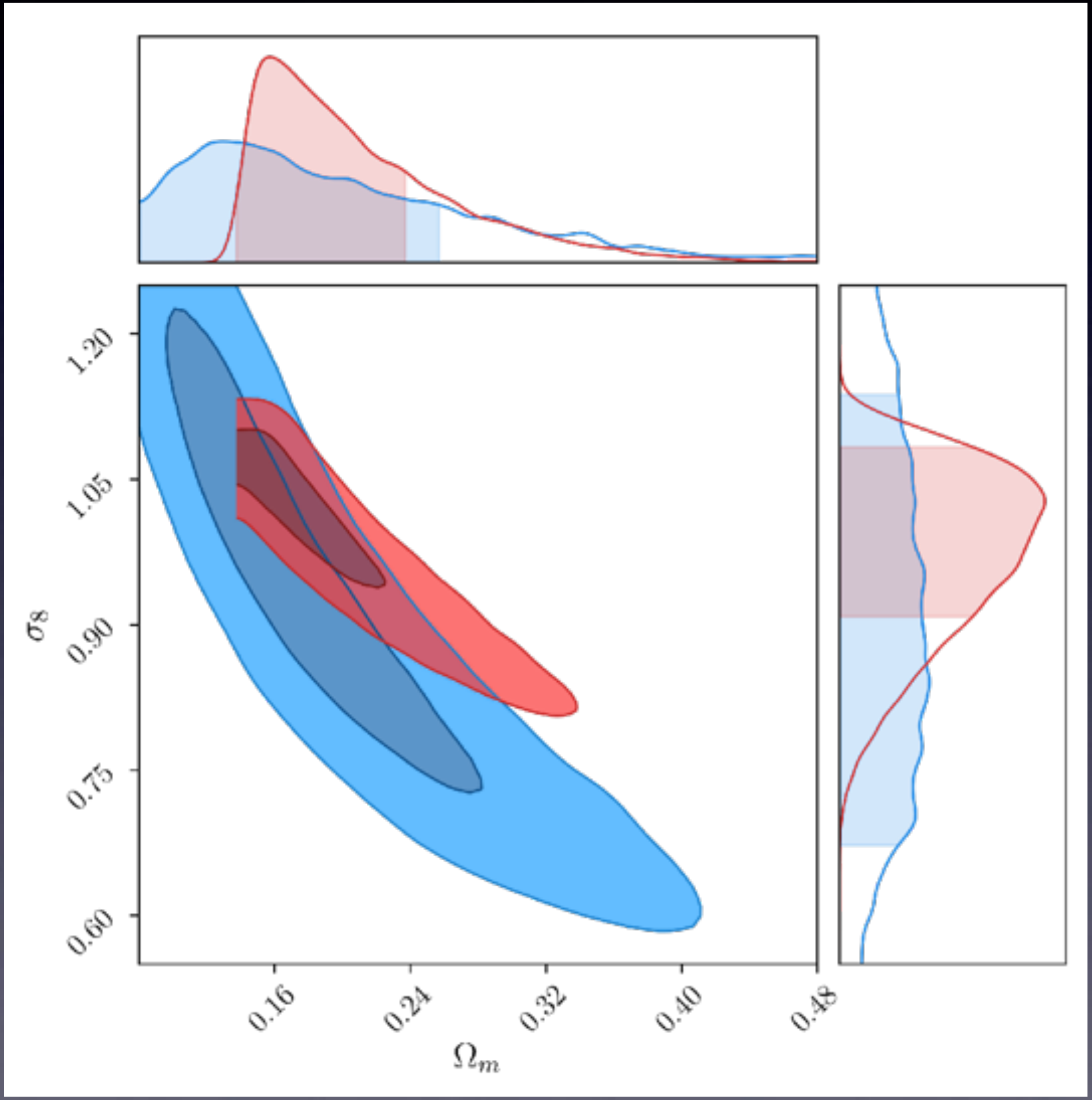
- Simple 5th order polynomial model, with second data set offset from the first
- Complexity of each individual data, and also combined data, is the same
  - Both measure the 5 free parameters well
- DIC only changes due to worsening of χ^2
- The ΔDIC goes from negative (agreement) to positive (tension) as the offset increases
- Odds ratio of agreement

$$\mathcal{I}(D_1, D_2) \equiv \exp\{-\Delta\mathrm{DIC}(D_1, D_2)/2\}$$



KiDS vs Planck

- All tensions considered here are in light of a particular model
- If the model is changed, the tension may be alleviated
- This is not the same as model selection

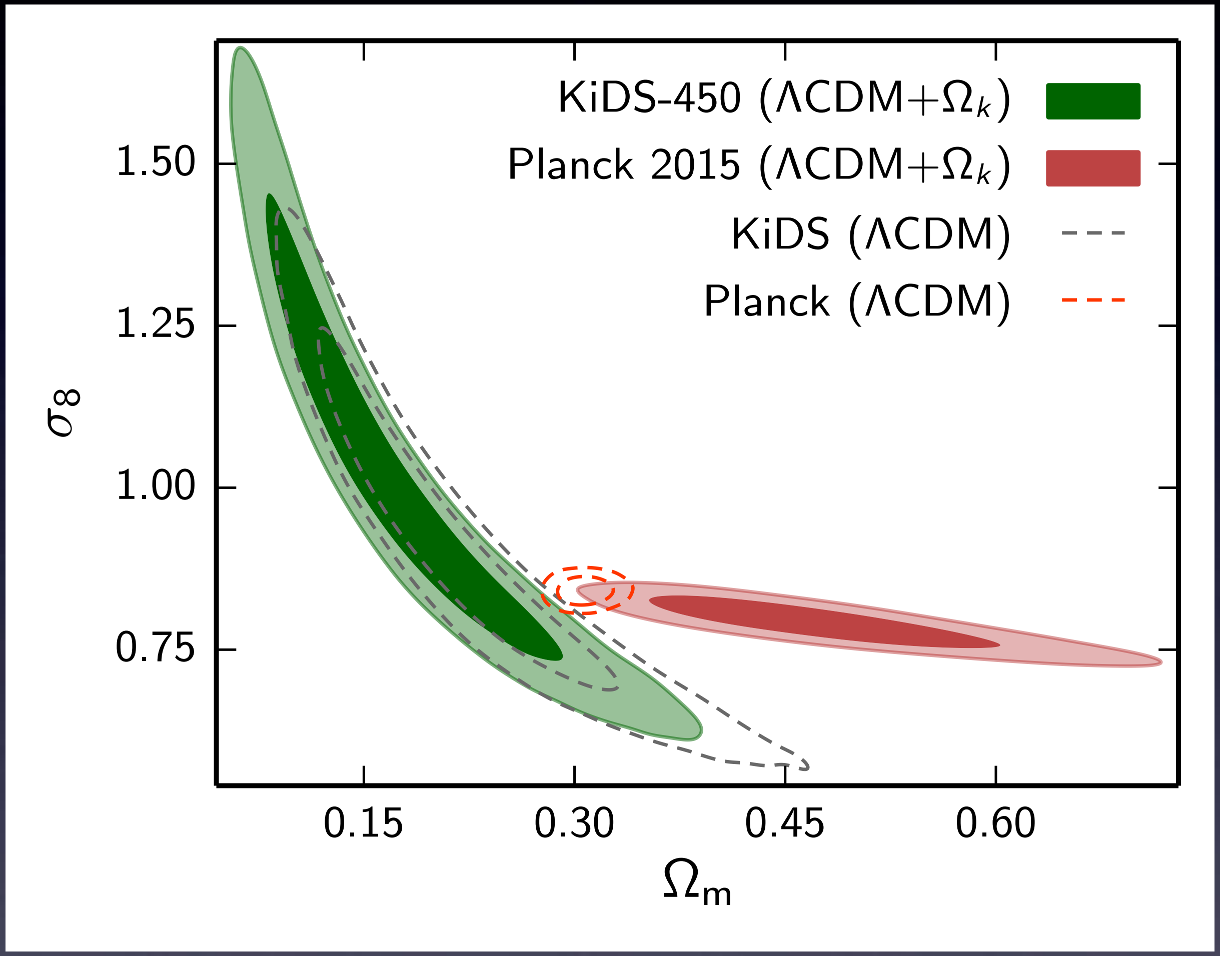


Application to lensing data

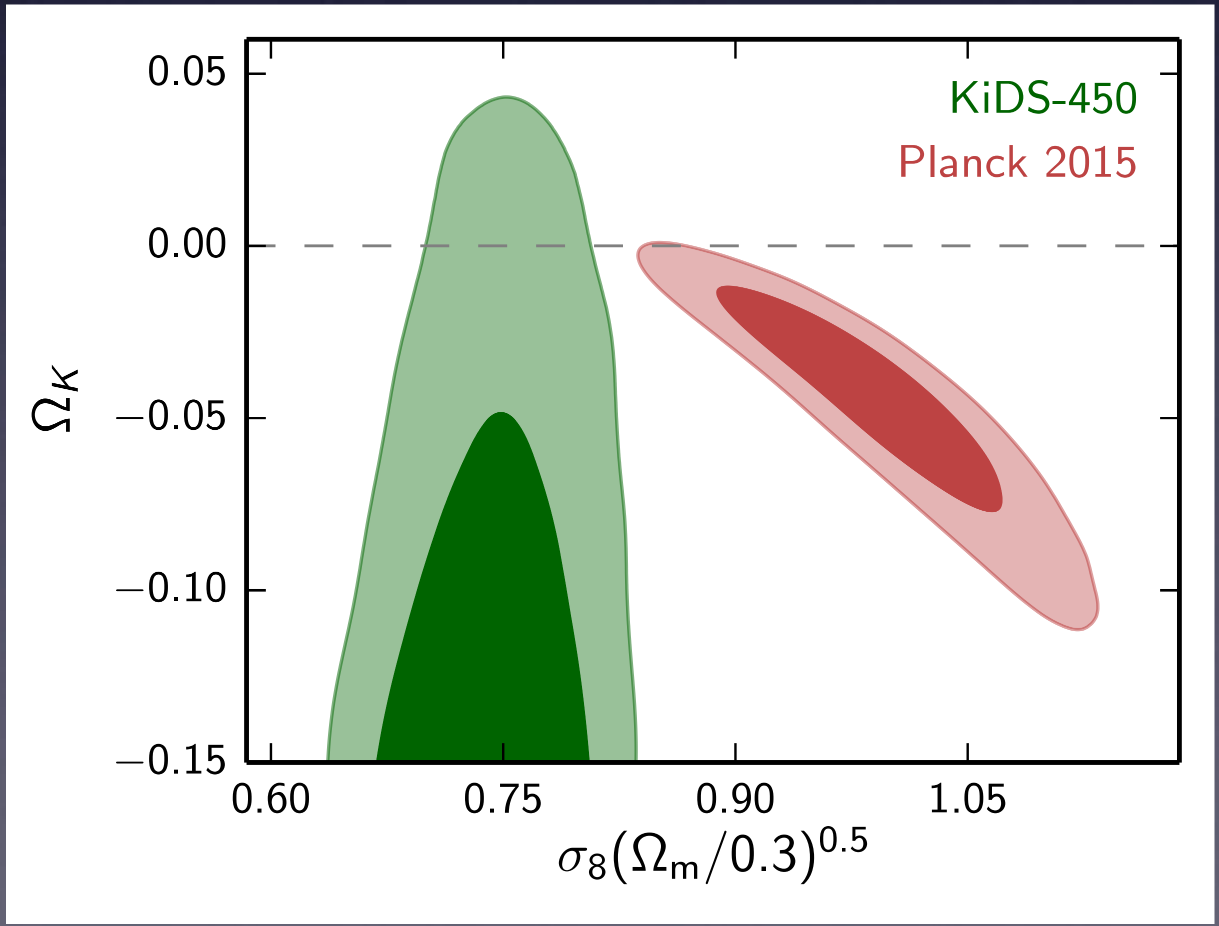
- In Joudaki et al (2016) they compared the cosmological constraints from Planck CMB data with KiDS-450 weak lensing data
- Including curvature worsened tension, but allowing for dynamical dark energy improved agreement

Model	$T(S_8)$	ΔDIC	
ΛCDM			
— fiducial systematics	2.1σ	1.26	Small tension
— extended systematics	1.8σ	1.4	Small tension
— large scales	1.9σ	1.24	Small tension
Neutrino mass	2.4σ	0.022	Marginal case
Curvature	3.5σ	3.4	Large tension
Dark Energy (constant w)	0.89σ	-1.98	Agreement
Curvature + dark energy	2.1σ	-1.18	Agreement

Curvature



KiDS-450 (ΛCDM+Ω_k)
Planck 2015 (ΛCDM+Ω_k)
KiDS (ΛCDM)
Planck (ΛCDM)
σ_8
1.50
1.25
1.00
0.75
0.15
0.30
0.45
0.60
Ω_m



KiDS-450
Planck 2015
Ω_K
0.05
0.00
−0.05
−0.10
−0.15
0.60
0.75
0.90
1.05
$\sigma_8(\Omega_m/0.3)^{0.5}$

Summary

- We can estimate the probability of a new dataset given the prior predictive distribution, or posterior predictive distribution from a previous dataset
- The posterior predictive p-value gives us the probability of some discrepancy statistic evaluated relative to some posterior prediction
- A number of tension statistics exist, including the simple likelihood, surprise, and DIC
- We can also estimate the relative probability of tensions between data sets using ratios of model likelihood (evidence)
- The Deviance Information Criteria is a simple method to evaluate tensions, being sensitive to likelihood ratio, but calibrated against parameter confidence regions (of individual and combined posteriors)
- Tension between the CMB and weak lensing shear tomography data exists, but seems to be alleviated through changing the model to include some dynamical dark energy