

# Evaluating the tension on $H_0$ using the Bayesian posterior predictive distribution

David Parkinson  
KASI

In collaboration with  
Shahab Joudaki (Oxford)

# Outline

- Introduction
- Statistical Inference
- Posterior predictions
- Application to  $H_0$  data
- Conclusions



# Quotes about statistics



# Quotes about statistics

- “There are lies, damned lies and statistics.”
  - Mark Twain



# 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



# 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



# 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



# 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?



Pierre-Simon Laplace



# What is Probability?

- In 1812 Laplace published *Analytic Theory of Probabilities*



Pierre-Simon Laplace



# 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"*



Pierre-Simon Laplace



# 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."*



Pierre-Simon Laplace



# 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."*



Pierre-Simon Laplace



# 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

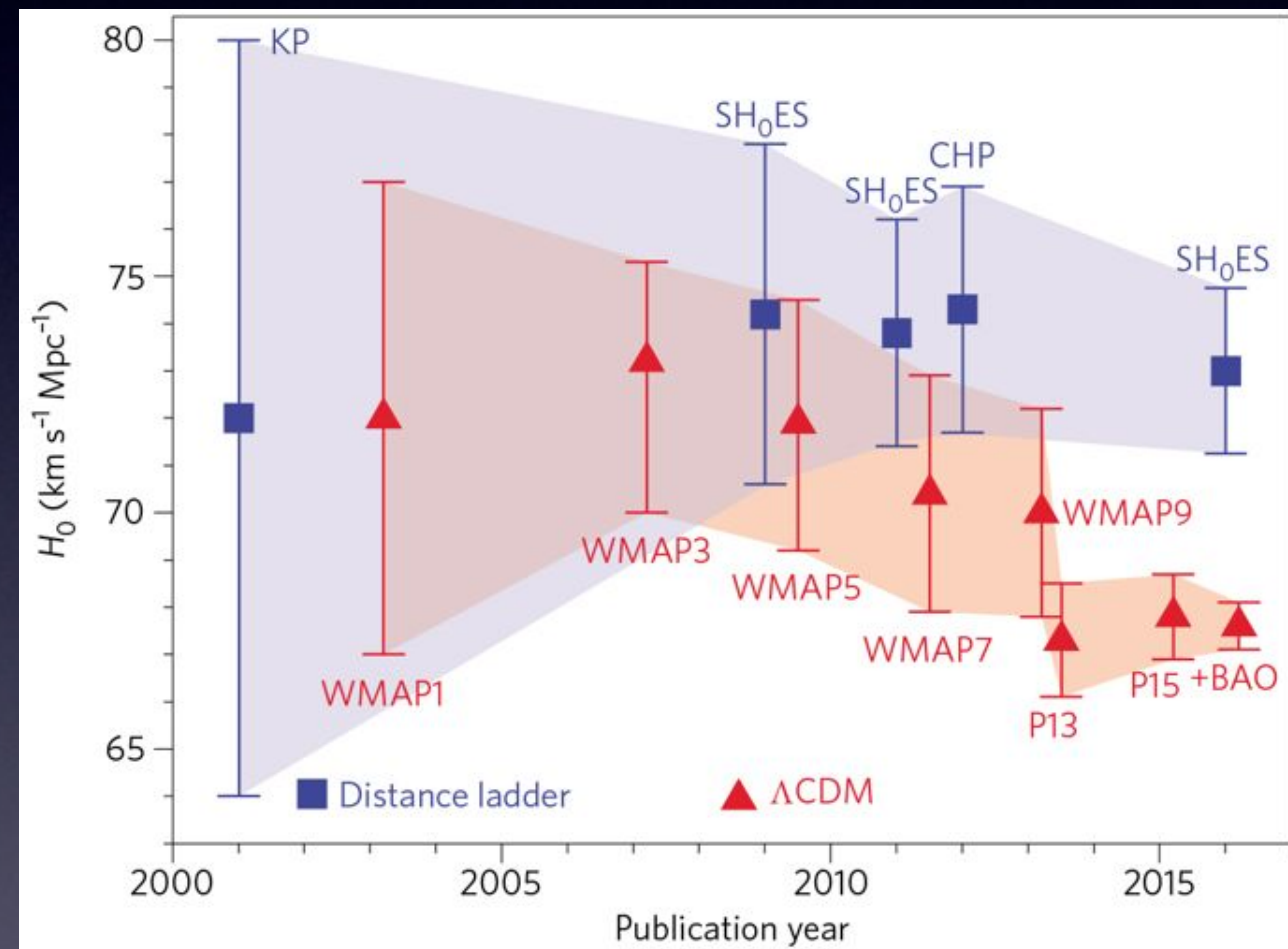


Pierre-Simon Laplace



# 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 now



Freedman 2017



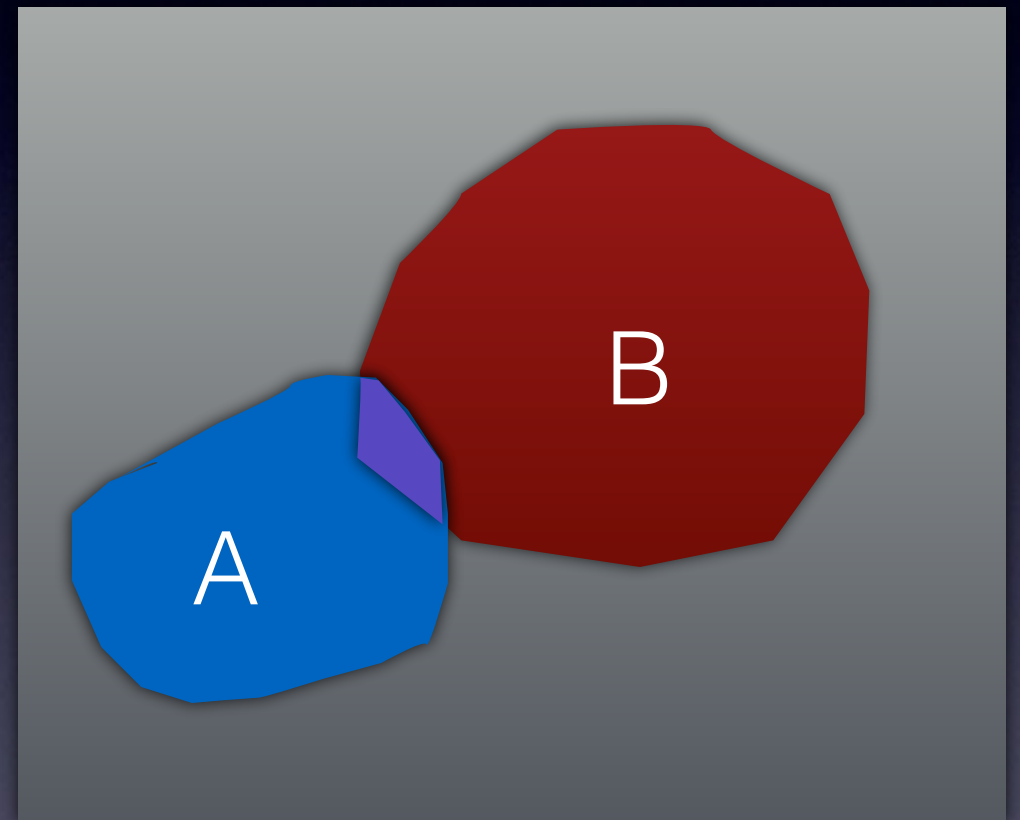
# Types of questions

- 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,  $P(\emptyset)=0$
- We normalize it so the sum of the probabilities over all options is unity,  $\sum P(A_i) \equiv 1$



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  $\theta$ , and want to test it with some data  $D$

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

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

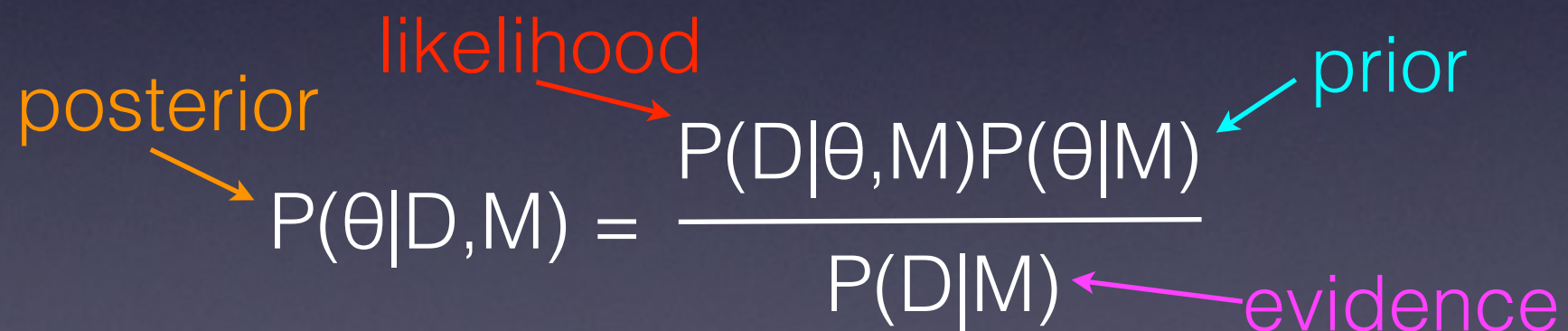


# Bayes Theorem

- 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  $\theta$ , and want to test it with some data  $D$



A diagram illustrating the components of Bayes' Theorem for model parameters. The equation is  $P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}$ . Colored arrows point from labels to parts of the equation: an orange arrow from 'posterior' to  $P(\theta|D,M)$ , a red arrow from 'likelihood' to  $P(D|\theta,M)$ , a cyan arrow from 'prior' to  $P(\theta|M)$ , and a magenta arrow from 'evidence' to  $P(D|M)$ .

$$\text{posterior} \rightarrow P(\theta|D,M) = \frac{\text{likelihood} \rightarrow P(D|\theta,M) \text{prior} \rightarrow P(\theta|M)}{\text{evidence} \rightarrow P(D|M)}$$

- 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

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

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

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

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



# Model Selection

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

$$\begin{array}{c} \text{evidence} \quad \quad \quad \text{likelihood} \quad \quad \quad \text{prior} \\ \swarrow \quad \quad \quad \downarrow \quad \quad \quad \swarrow \\ E = P(D|M) = \int P(D|\theta, M) P(\theta|M) d\theta \end{array}$$

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

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

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



# Model Selection

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

$$\text{evidence} \rightarrow E = P(D|M) = \int \text{likelihood} P(D|\theta, M) \text{prior} P(\theta|M) d\theta$$

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

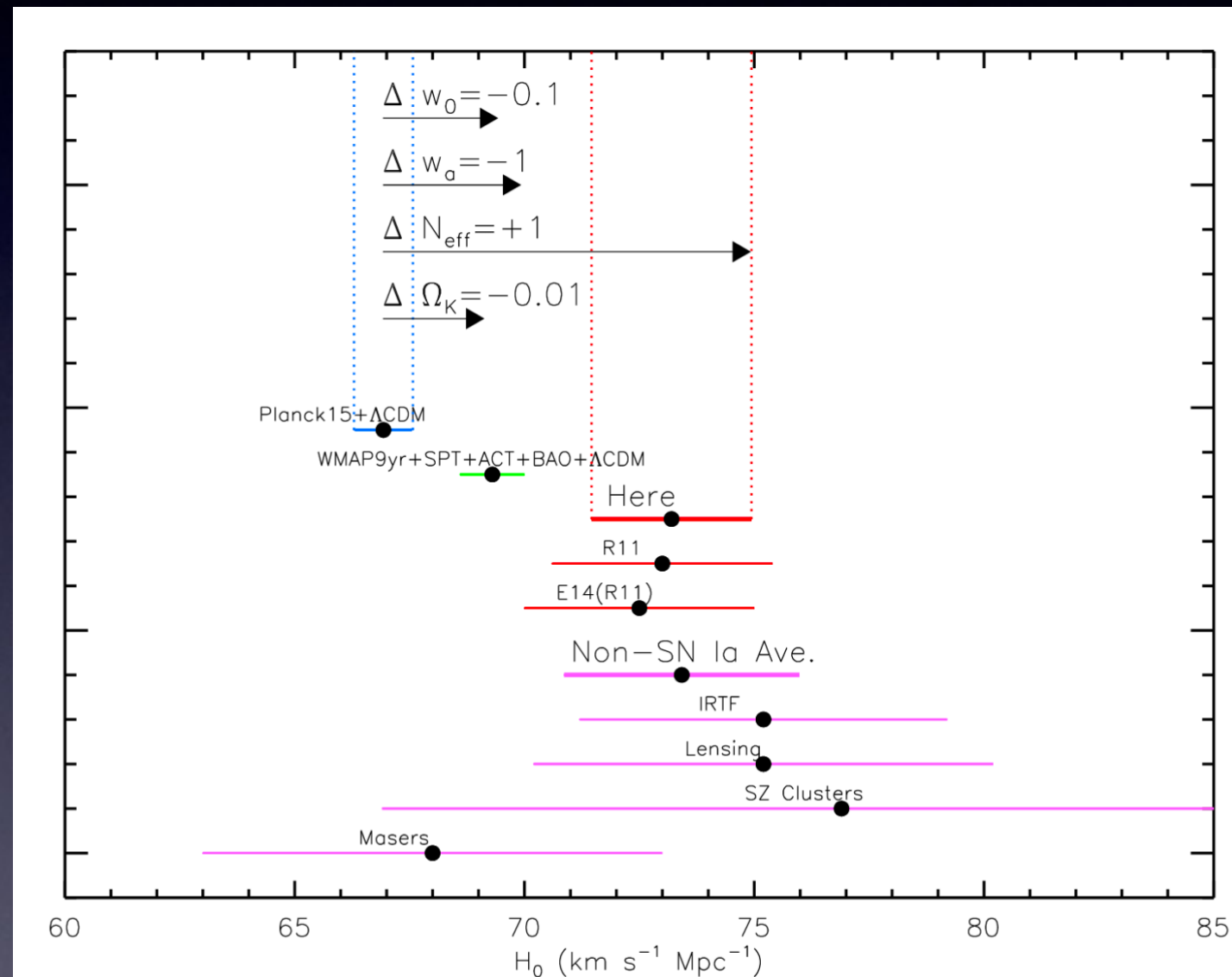
$$\text{Model posterior} \rightarrow \frac{P(M_1|D)}{P(M_2|D)} = \frac{\text{evidence} P(D|M_1) \text{Model prior} P(M_1)}{P(D|M_2)P(M_2)}$$

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



# 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
- Need to evaluate probability of data





# 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?

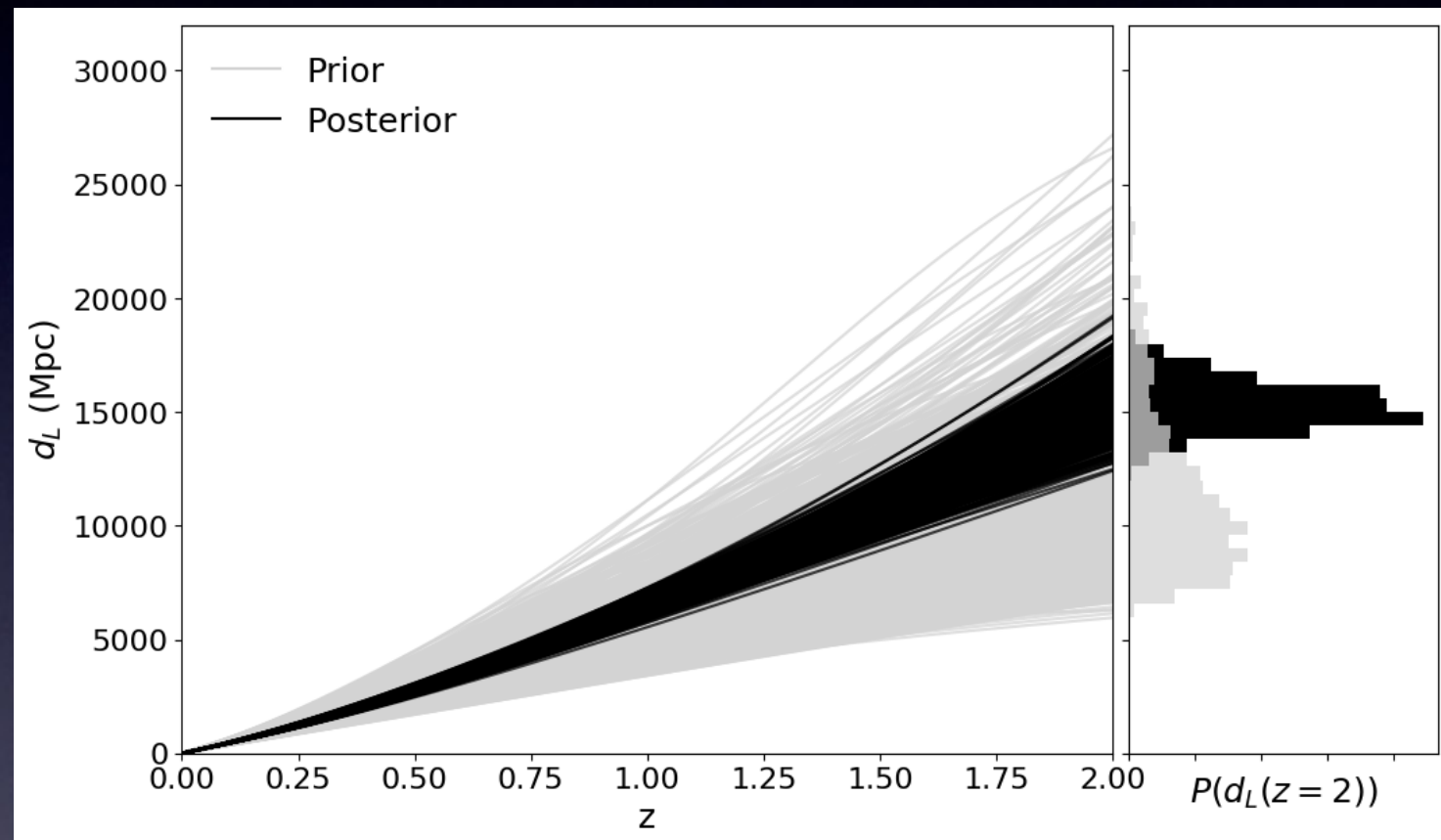
- Prior predictive distribution

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

- Posterior predictive distribution

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

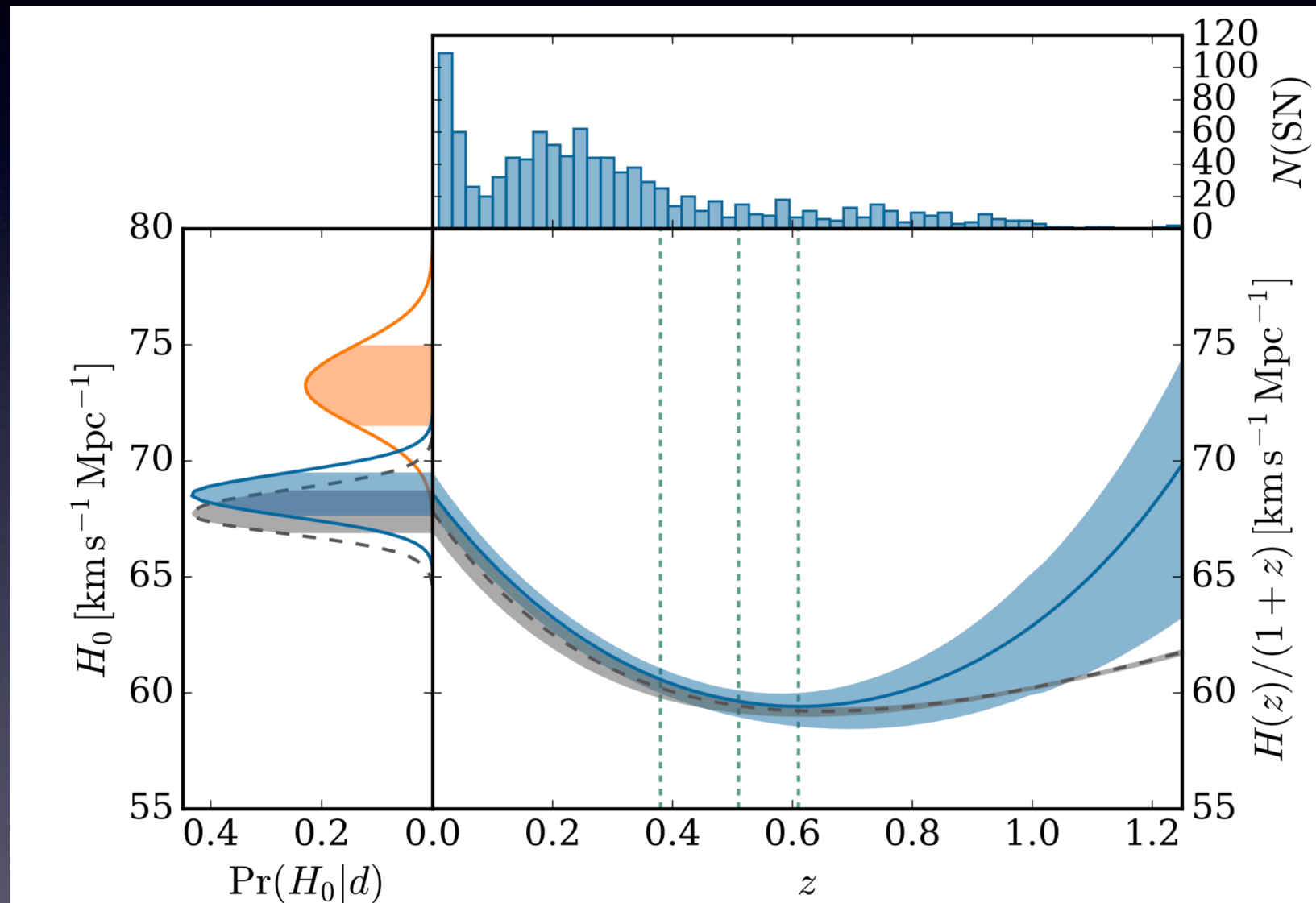
- We can compare predictive data to actual repetitions or further observations to validate data





# Planck measurement

- Planck ‘measurement’ is just posterior predictive distribution
- And will change, depending on assumptions
- (Right:) Black is LCDM, Blue is most general, orange is Cepheid measurement



Feeney et al 2018



# 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)

$$p = P\left(T(\tilde{D}_2) > T(D_2) \middle| D_1, \theta\right)$$



# 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  $\chi^2$  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  $D_1$ .

# 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
- Joudaki et al (2016): change in DIC

$$\mathcal{G}(D_1, D_2) = \text{DIC}(D_1 \cup D_2) - \text{DIC}(D_1) - \text{DIC}(D_2)$$



# Complexity

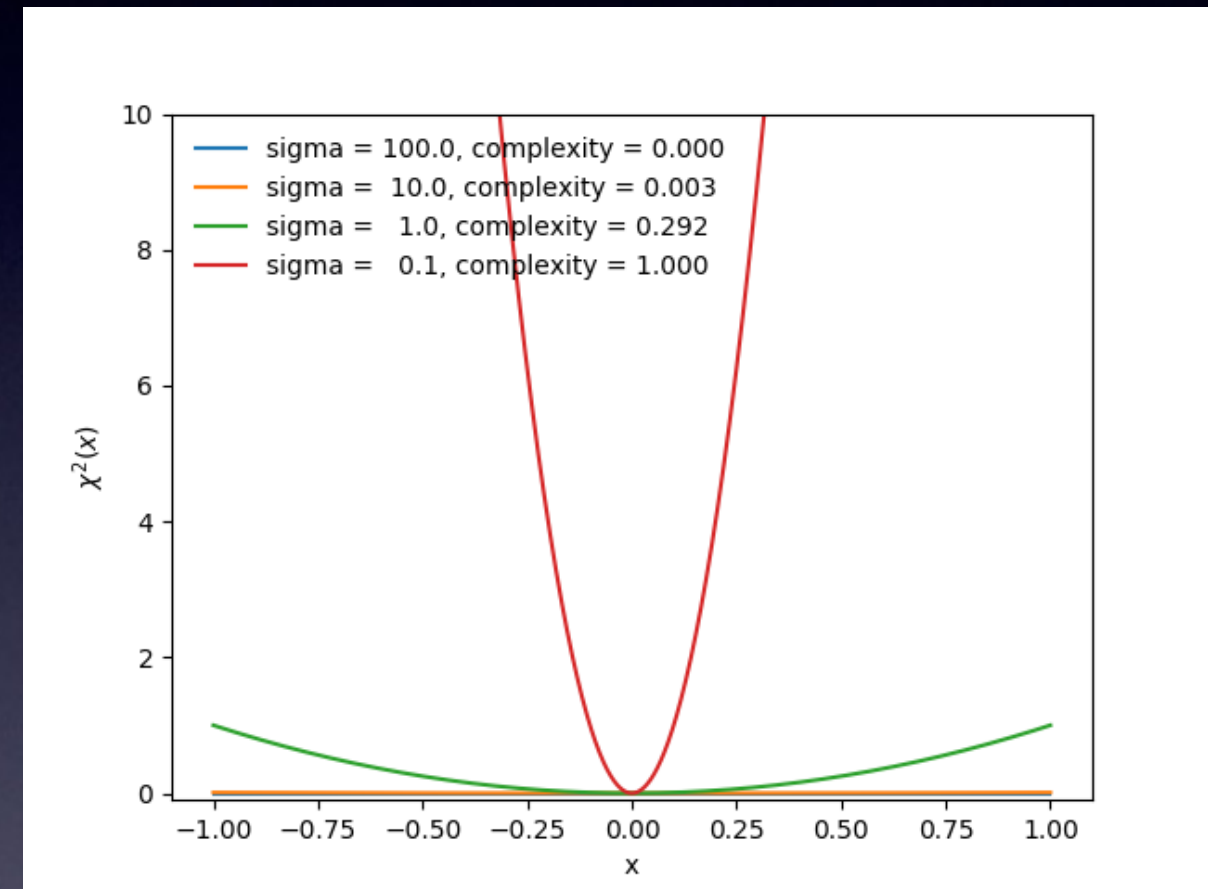
- The DIC penalises models based on the *Bayesian complexity*, the number of well-measured parameters
- This can be computed through the information gain (KL divergence) between the prior and posterior, minus a point estimate

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

- For the simple gaussian likelihood, this is given by

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

- Average is over posterior



# Diagnostics II: The Surprise

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

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

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

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

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

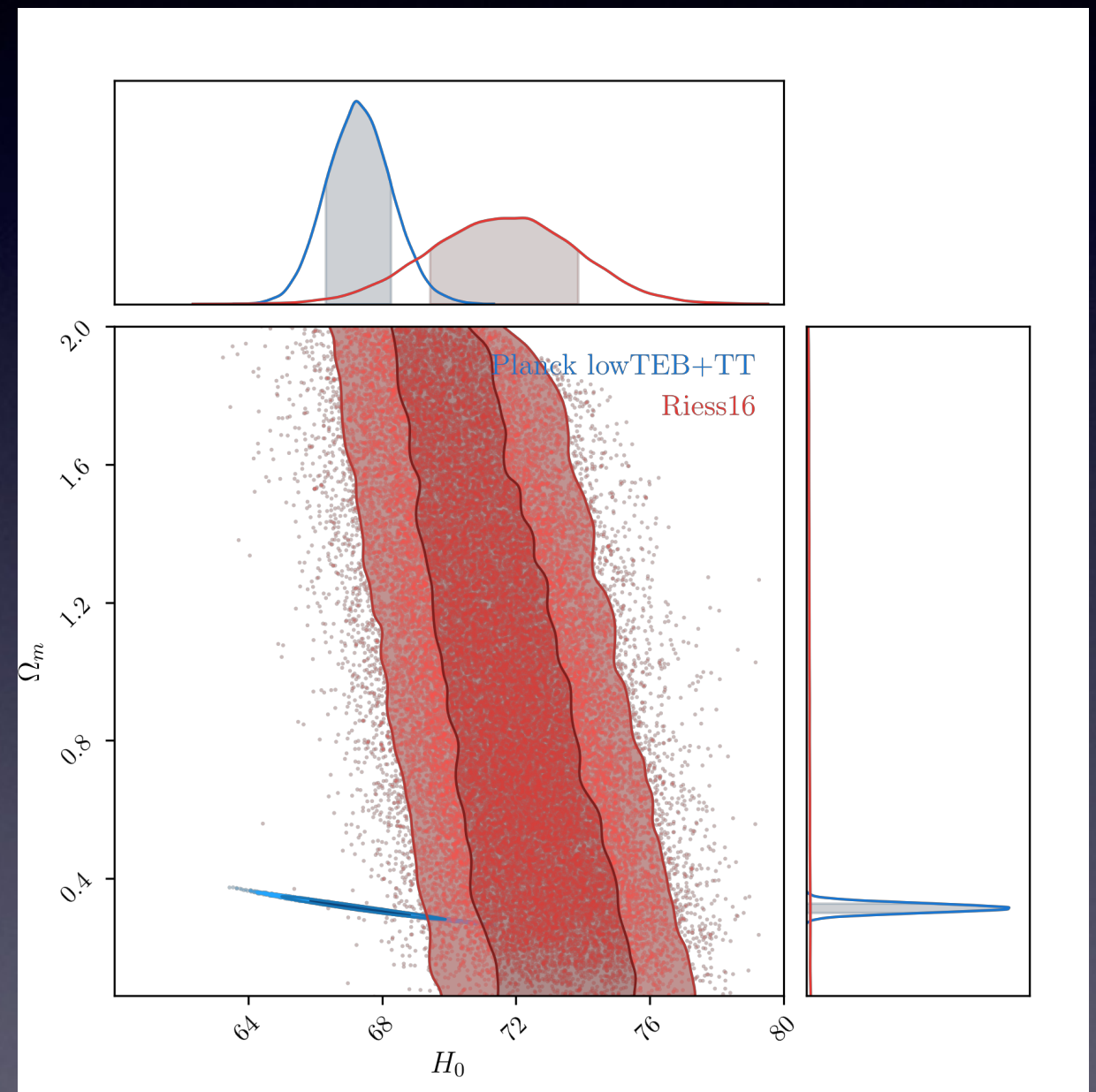


# Pros and Cons

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

# H0 data

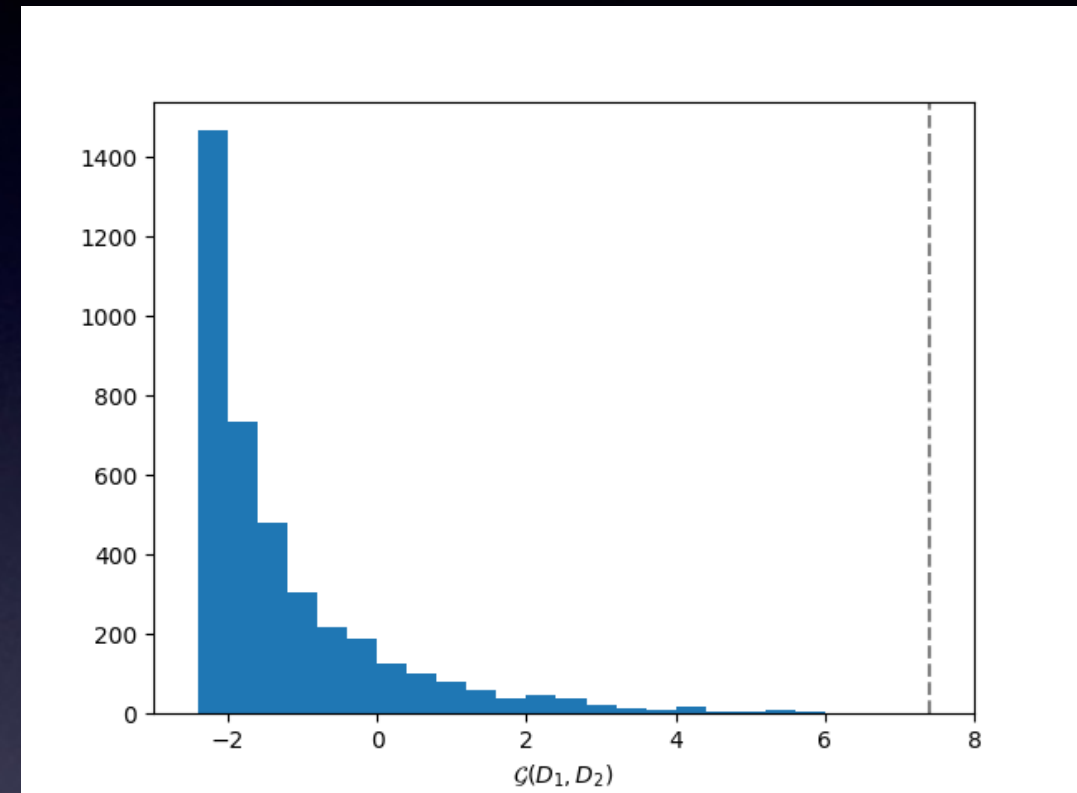
- SHOES does not measure  $H_0$ , it measures a luminosity and angular diameter distances of some objects
- Even at very low redshift, the measurement is not completely independent of the other parameters





# Tension: Planck vs SHOES

- We found a change in DIC of 7.4, mainly driven by a change in average  $\chi^2$
- Corresponding posterior predictive p-value  $\sim 0.0025$



Dataset	best-fit $\chi^2$	average $\chi^2$	Complexity	DIC
Planck-2015	11261.9	11281.9	20	11301.9
Reiss 2016 $H_0$	0	1	1	2
Planck + $H_0$	11269.9	11290.5	20.7	11311.3

# 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 prior or posterior prediction
- A number of tension statistics exist, including the simple likelihood, surprise, and DIC
- The tension statistic is different to model selection, as it can be applied to a single physical model
- Using the DIC as a tension statistic, we find that the posterior predictive probability of the  $H_0$  measurement, assuming Planck and a LambdaCDM cosmology, to be roughly 0.0025, or roughly 400:1 against