Statistical Tools for Data Analysis

Justin Ellis
IPTA Student Workshop
June 18, 2014

(Many slides courtesy of Joe Romano)

This talk has been colored by personal experience and is likely to be unbiased

The Great Battle of Our Time

Membership pre-test

- An astronomer measures the mass of a neutron star in a binary pulsar system to be:

  "$M = (1.39 \pm .02) M_\odot$ with 90% confidence"

  where the uncertainty is measurement noise (assumed Gaussian) in the observing apparatus.

- Q: How do you interpret the quoted result?

- A1: You are 90% confident that the true mass of the NS lies in the interval $[1.37 M_\odot, 1.41 M_\odot]$

- A2: 90% is the long-term relative frequency with which the true mass of the neutron star lies in the set of intervals $\{[\widehat{M} - .02 M_\odot, \widehat{M} + .02 M_\odot]\}$ where $\{\widehat{M}\}$ is the set of measured masses.

Affiliation

- If you chose answer A1, you belong to the Bayesian church.
- If you chose answer A2, you belong to the Frequentist church.

Goal of science

"Infer nature's state from observations"

- Observations are:

  (i) incomplete (problem of induction)

  (ii) imprecise (measurement noise)

  → Conclusions uncertain!!

- Statistical inference (a.k.a. plausible inference, probabilistic inference) is a way to quantify and manipulate uncertainty

Algebra of probability

$$0 \leq p(X) \leq 1$$

$$p(X = \text{true}) = 1$$

$$p(X = \text{false}) = 0$$

$$p(X) + p(\overline{X}) = 1$$

sum rule

$$p(X, Y) = p(X|Y)p(Y)$$

product rule

joint probability

conditional probability

Bayes' theorem
("learning from experience")

likelihood

prior

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}$$

posterior

normalization factor

$$p(D) = p(D|H)p(H) + p(D|\overline{H})p(\overline{H})$$

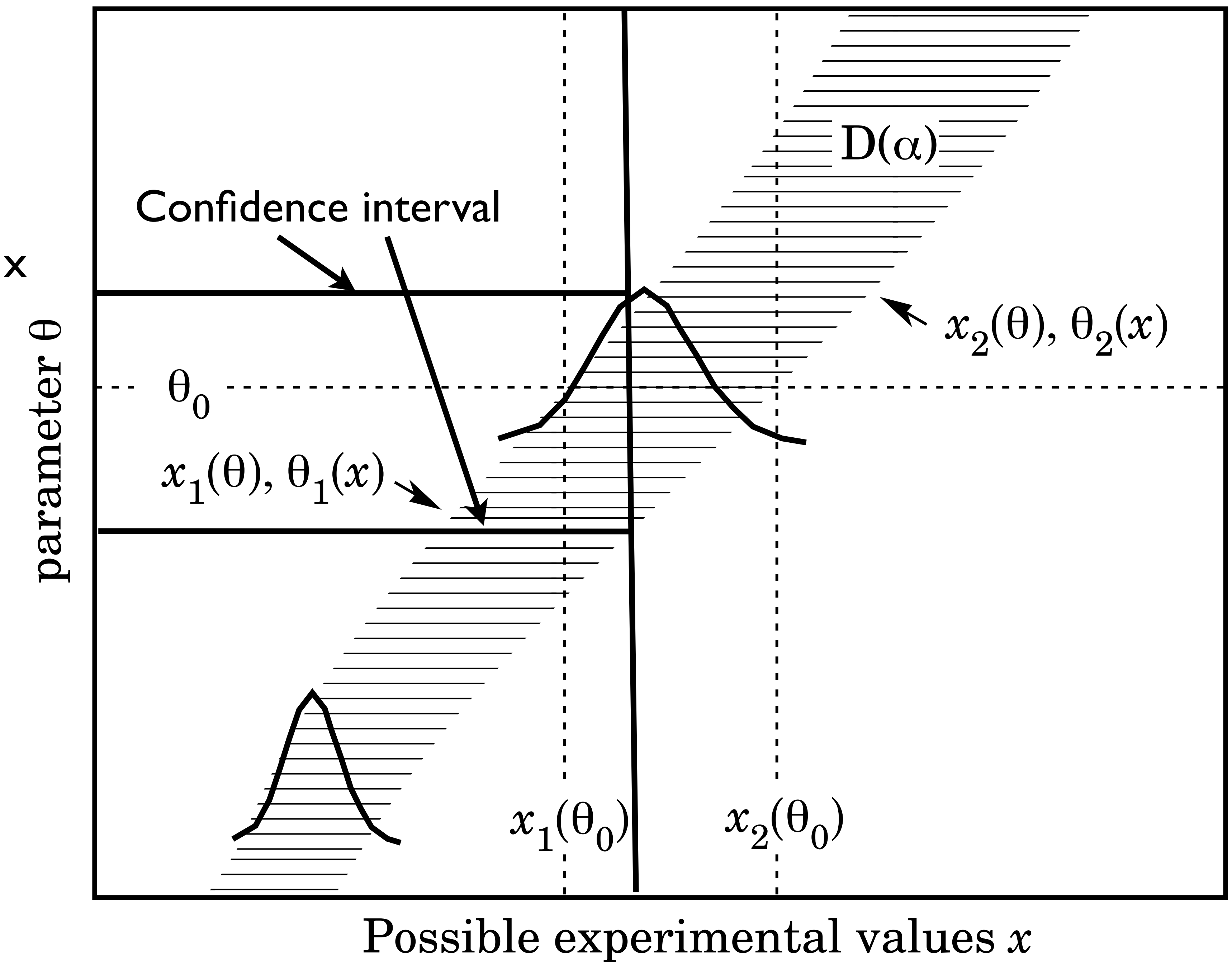
A frequentist...	A Bayesian...
probability = long-run relative freq	probability = degree of belief
Assumes that the data are random and that the hypothesis (parameters) are fixed but unknown. Makes use of the likelihood function p(D|H)	Assumes that the data are fixed and that the hypothesis (parameters) are random. Makes use of the posterior probability distribution p(H|D)
constructs a statistic to estimate a parameter, or see if the data are consistent with a model	needs to specify prior degree of belief in a particular hypothesis or parameter
calculates the probability distribution of the statistic (sampling distribution)	uses Bayes' theorem to update prior degree of belief in light of new data
constructs confidence intervals and p-values (for parameter estimation and hypothesis testing)	constructs credible sets and odds ratios (for parameter estimation and hypothesis testing)

Frequentist Confidence Intervals

- Coverage: Does the "true" value of the parameter lie in the x% confidence interval in x % of experiments

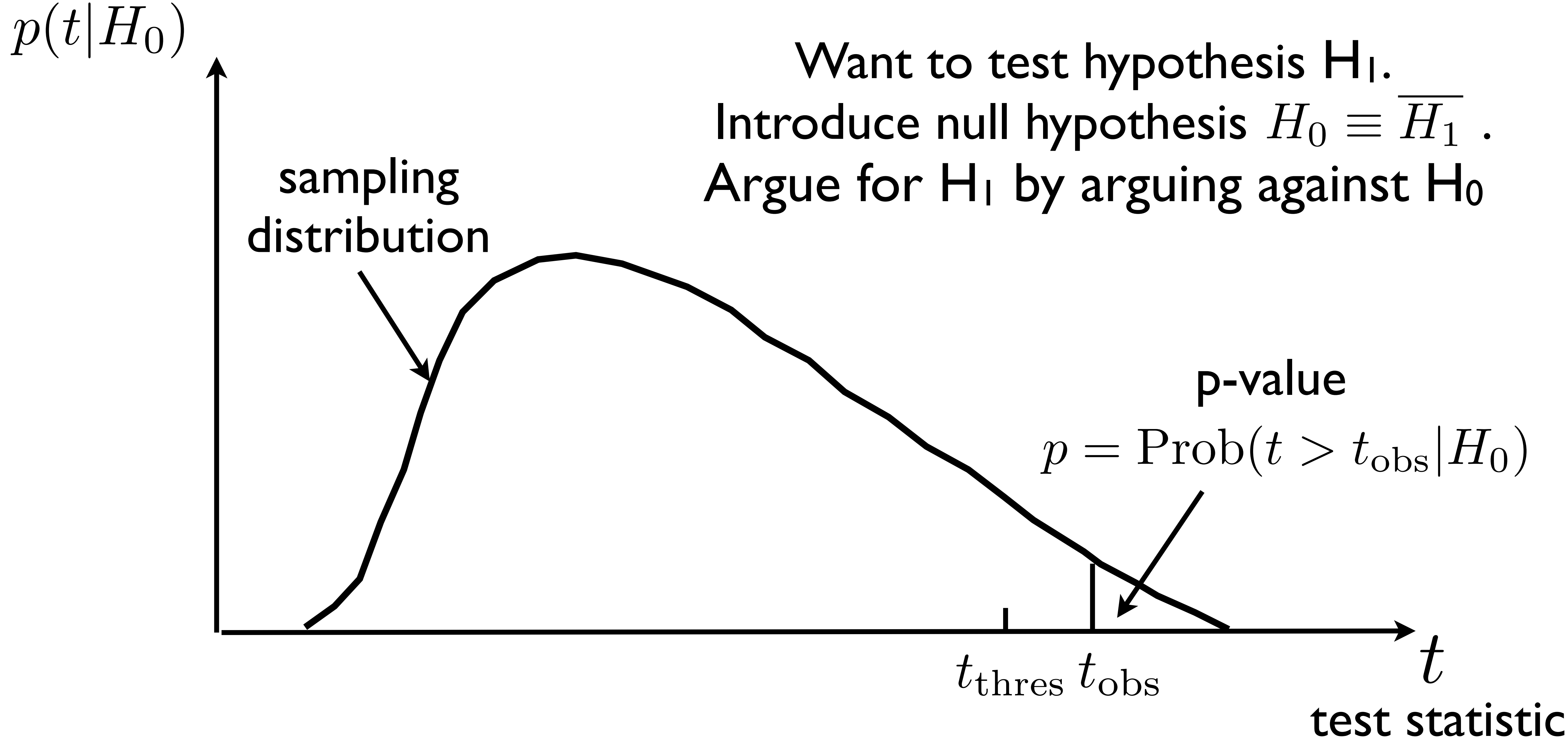
- Neyman construction guarantees correct coverage.

- Over coverage is ok, under-coverage is bad



Frequentist hypothesis testing

Want to test hypothesis H_1.
Introduce null hypothesis $H_0 \equiv \overline{H_1}$.
Argue for H_1 by arguing against H_0



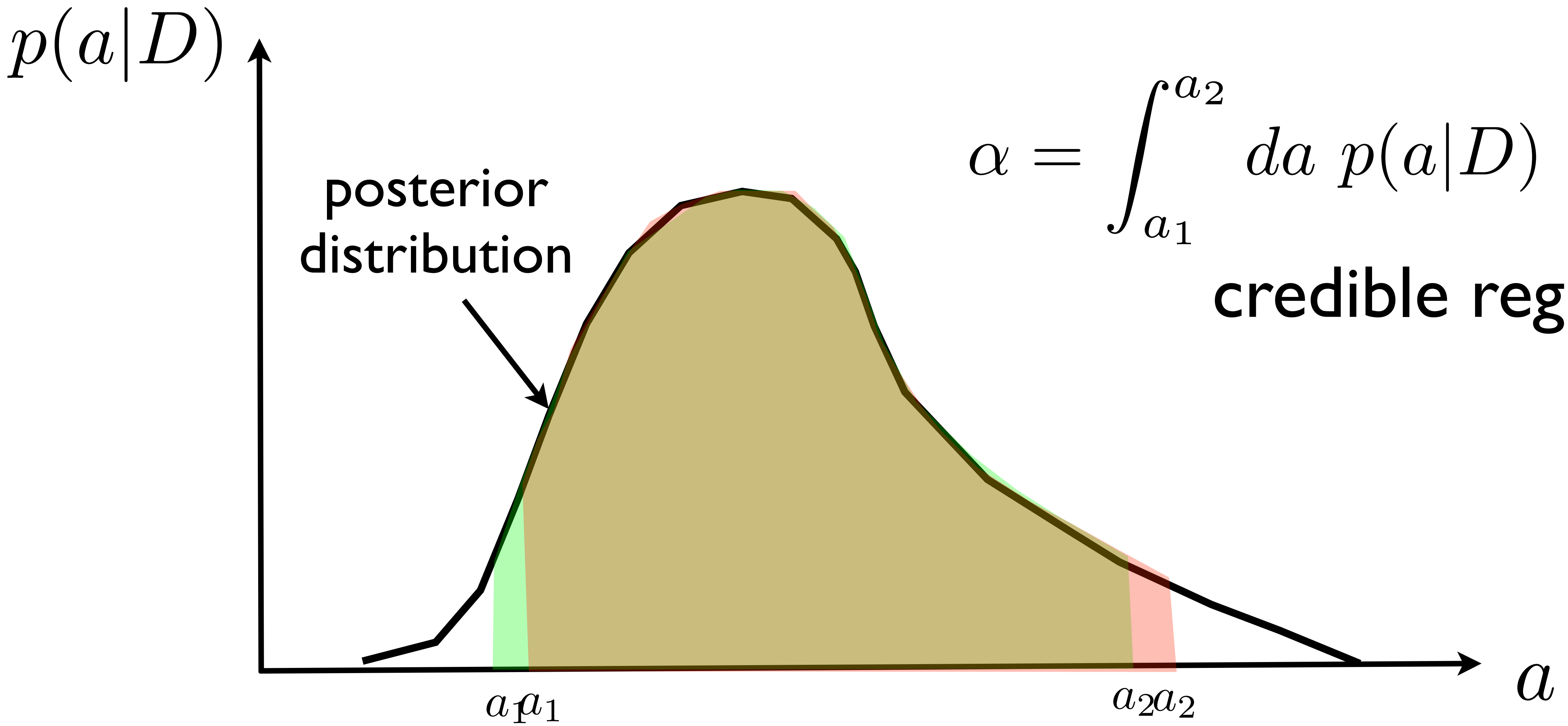
If t_{obs}>t_{thres}, reject H_0 (accept H_1) at p$*$100% confidence level.

CAUTION! $(1-p) \neq \mathrm{Prob}(H_1|t > t_{\mathrm{obs}})$

Frequentist hypothesis testing

- The p-value needed to reject the null hypothesis is the *threshold* for acceptance of H_1
- There are two types of errors:
  - *False alarm*: Reject null hypothesis when true
  - *False dismissal*: Accept null hypothesis when false
- Different test statistics are judged according to their false alarm and false dismissal probabilities
- In GW data analysis, one fixes the false alarm probability at some tolerably low level, then finds the test statistic that minimizes the false dismissal probability (maximize detection probability)

Bayesian parameter estimation



$$\alpha = \int_{a_1}^{a_2} da\ p(a|D)$$

credible region

- credible region not unique

$$\frac{1-\alpha}{2} = \int_{-\infty}^{a_1} d\lambda\ p(a|D)$$

$$p(a_1|D) = p(a_2|D)$$

$$\frac{1-\alpha}{2} = \int_{a_2}^{\infty} d\lambda\ p(a|D)$$

Bayesian hypothesis testing

- It doesn't make sense to talk about a single hypothesis without reference to alternative hypotheses since

$$p(D) = \sum_i p(D|H_i)p(H_i)$$

Compare two hypotheses:

$$\frac{p(H_1|D)}{p(H_0|D)} = \frac{p(D|H_1)}{p(D|H_0)} \frac{p(H_1)}{p(H_0)}$$

posterior odds

marginalized likelihood ratio
(Bayes factor)

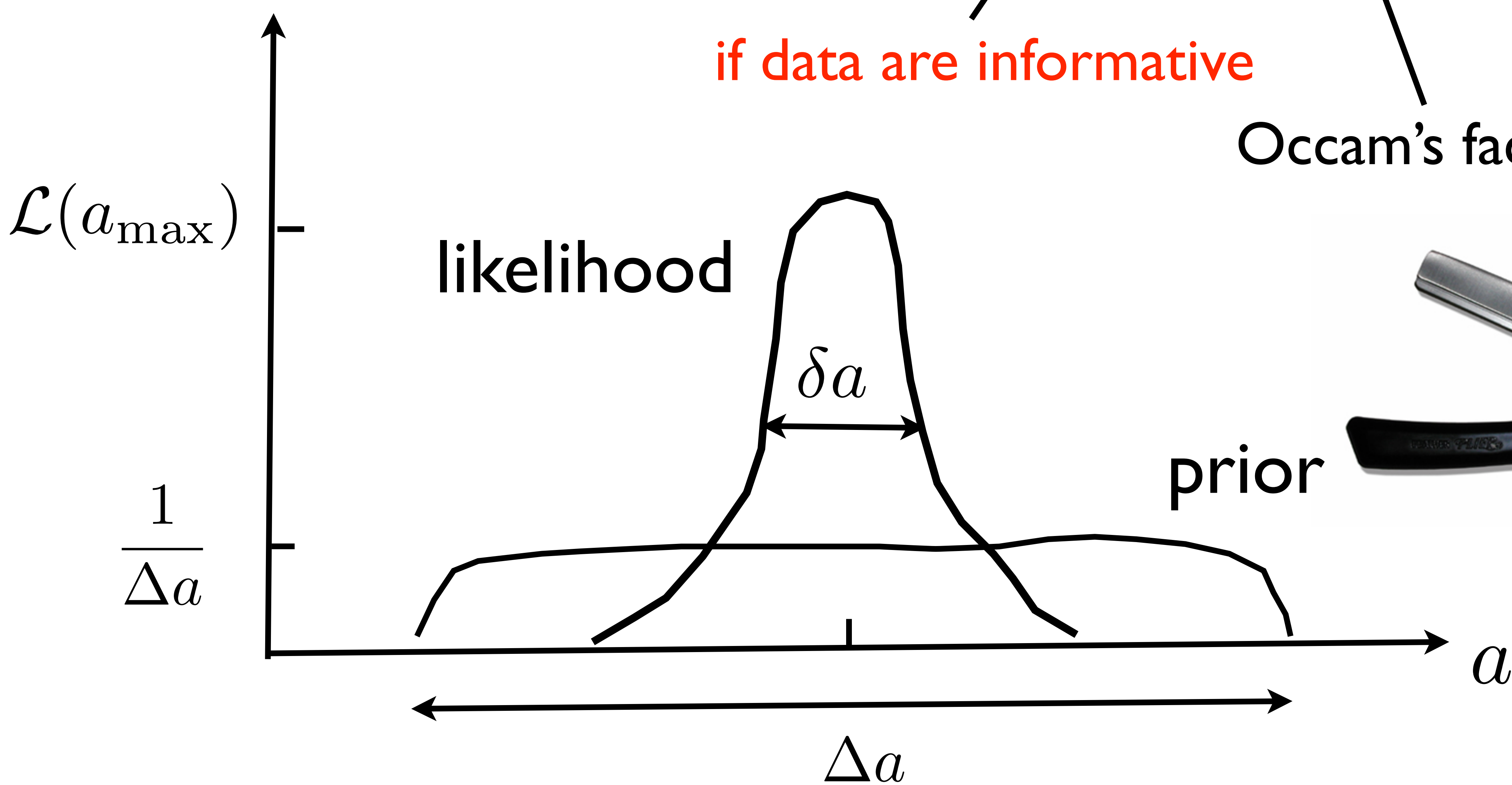
prior odds

Marginalized likelihood

$$p(D|H) = \int da\, p(D|a,H)p(a|H) \approx \mathcal{L}(a_{\max})\frac{\delta a}{\Delta a}$$

if data are informative

Occam's factor



A frequentist...	A Bayesian...
probability = long-run relative freq	probability = degree of belief
Assumes that the data are random and that the hypothesis (parameters) are fixed but unknown. Makes use of the likelihood function p(D|H)	Assumes that the data are fixed and that the hypothesis (parameters) are random. Makes use of the posterior probability distribution p(H|D)
constructs a statistic to estimate a parameter, or see if the data are consistent with a model	needs to specify prior degree of belief in a particular hypothesis or parameter
calculates the probability distribution of the statistic (sampling distribution)	uses Bayes' theorem to update prior degree of belief in light of new data
constructs confidence intervals and p-values (for parameter estimation and hypothesis testing)	constructs credible sets and odds ratios (for parameter estimation and hypothesis testing)

Mathematical problem

instrument response

measured data

intrinsic signal

noise

$$\mathbf{d}(t) = (\underbrace{\mathbf{R} * \mathbf{h}}_{\text{measured signal}})(t; \boldsymbol{\theta}) + \mathbf{n}(t)$$

signal parameters

- Given the data, infer the value of the signal parameters
- Simplify: $d(t) = s(t; a, b) + n(t)$
- Begin by characterizing the noise

Noise is ...

- Anything that interferes with identification of the signal
- Typically associated with measuring apparatus, but could be a foreground signal
- Usually easier to characterize than the signal (point off-source, estimate from other data stretches, ...)
- Typically associated with random processes (otherwise, subtract it out)
- Characterized statistically (probability distribution or ensemble averages over all possible measurements)

$$\langle n(t_1)\rangle \,, \quad \langle n(t_1)n(t_2)\rangle \,, \quad \langle n(t_1)n(t_2)n(t_3)\rangle \,, \cdots$$

Gaussian processes

- Noise is often described as a Gaussian random process
- Why?
  - histogram of samples is approximately Gaussian
  - Central Limit Theorem (sum of a large number of random disturbances)
  - given knowledge of only 1st and 2nd moments, a Gaussian is the least informative (maximum entropy) probability distribution

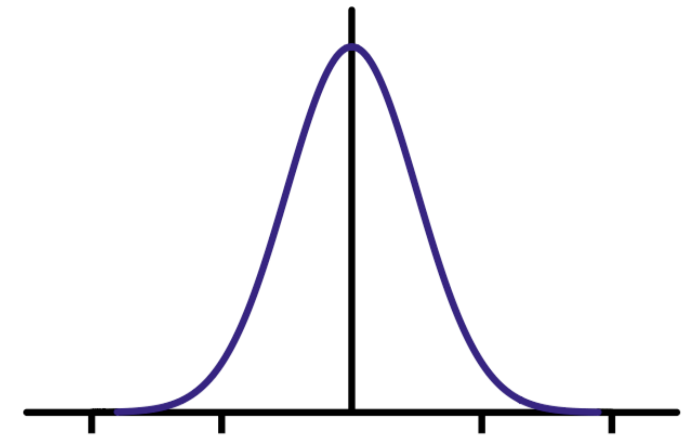
Gaussian distribution

- Single sample:

mean

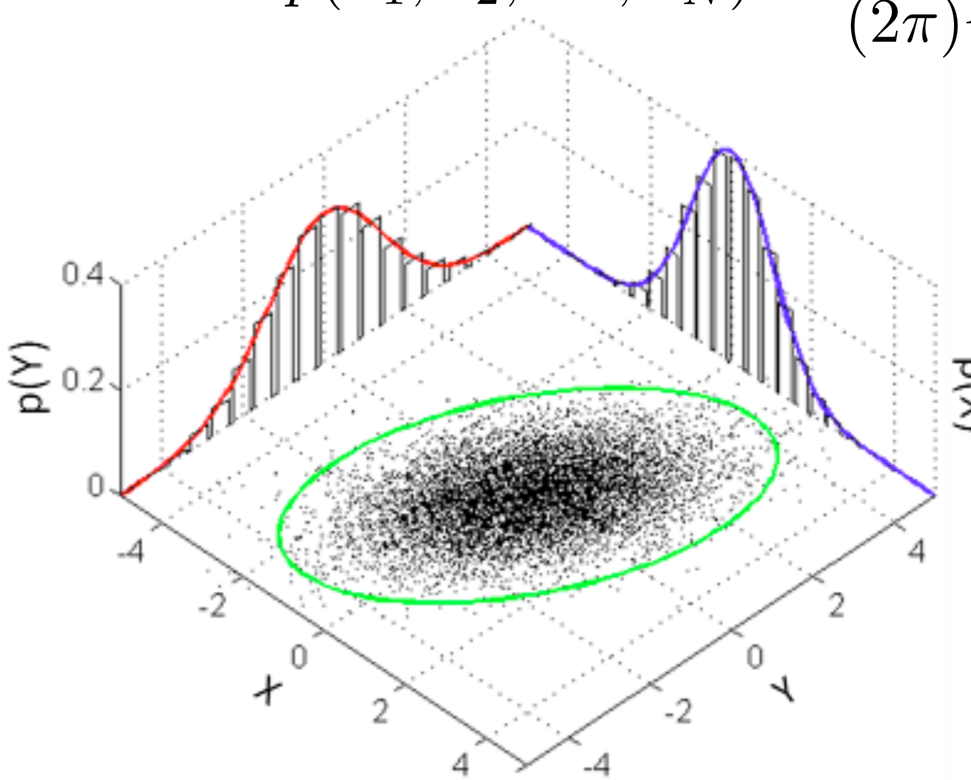
$$p(n) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2}\frac{(n-\mu_n)^2}{\sigma_n^2}\right]$$

variance



- Multivariate Gaussian:

$$p(n_1, n_2, \ldots, n_N) = \frac{1}{(2\pi)^{N/2}\sqrt{\det C_n}} \exp\left[-\frac{1}{2}\sum_{i,j=0}^{N-1}(n_i - \mu_{ni})C_{nij}^{-1}(n_j - \mu_{nj})\right]$$



covariance matrix

$$C_{nij} := \langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle$$

Likelihood function

Probability of noise: $p(n|\theta) = \frac{1}{\sqrt{\det(2\pi C_n)}} \exp\left(-\frac{1}{2} n^T C^{-1} n\right)$

Measured data is: $d(t) = s(t, \lambda) + n(t, \theta) \;\; \Rightarrow n(t, \theta) = d(t) - s(t, \lambda)$

Probability of data: $p(d|\lambda, \theta) = \frac{1}{\sqrt{\det(2\pi C_n)}} \exp\left(-\frac{1}{2}(d - s(\lambda))^T C^{-1}(d - s(\lambda))\right)$

Null Hypothesis: $d(t) = n(t, \theta)$ with likelihood $p_0(d|\theta) = p(n|\theta)$

Log-likelihood ratio: $\ln \Lambda(d|\lambda, \theta) = \frac{p(d|\lambda, \theta)}{p_0(d|\theta)} = (d|s) - \frac{1}{2}(s|s)$

Noise-weighted inner product: $(x|y) = x^T C_n^{-1} y$

Maximum Likelihood Estimators (MLEs)

- It is common in frequentist statistics to compute the maximum likelihood estimators of the signal parameters by maximizing the likelihood ratio.

$$\frac{\partial \ln \Lambda(d|\lambda, \theta)}{\partial \lambda_i} = 0$$

generally must be done numerically but can be done analytically in some cases

- Covariance matrix Γ of parameters is defined through:

$$(\Gamma^{-1})_{ij} = -\frac{\partial^2 \ln \Lambda(d|\lambda, \theta)}{\partial \lambda_i \partial \lambda_j}\bigg|_{\hat{\lambda}}$$

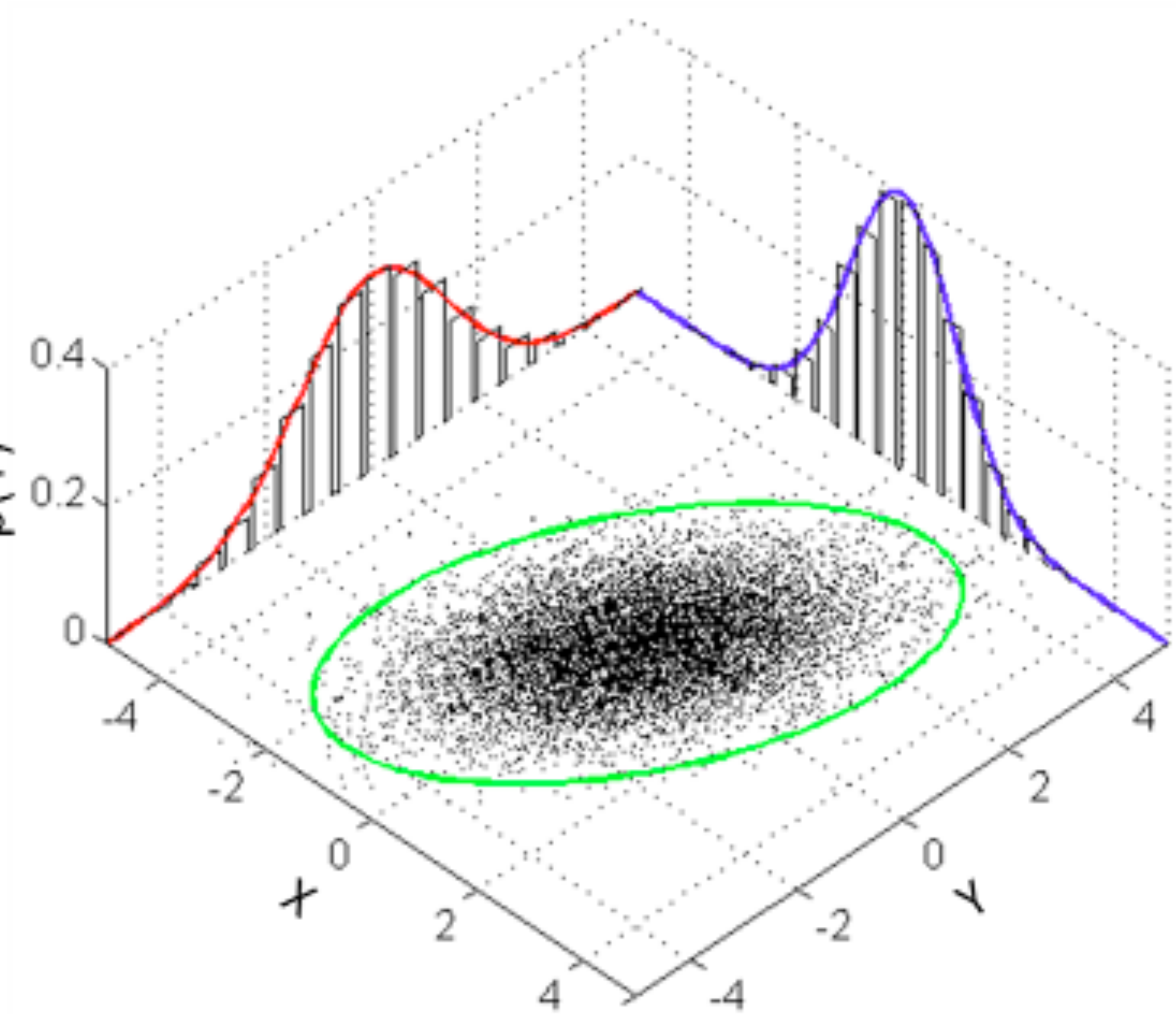
- Use maximum likelihood estimators $\hat{\lambda}$ to construct confidence intervals on “true” parameters

- This is exactly what tempo2 did when you hit “fit”

Nuisance Parameters: What about that θ?

- In many cases our likelihood depends on parameters that must be included but are not of particular interest
- Want our parameter estimates and detection statements to be independent of θ
- Frequentist statistics have no robust way of dealing with nuisance parameters. Common strategies are:
  - Fix the nuisance parameters to their maximum likelihood value and perform all analysis using these values
  - Construct profile likelihood which maximizes the likelihood function over the nuisance parameters for each true value of parameters of interest
  - Use frequentist-Bayesian hybrid method

Ba ed



Joint posterior

if θ and λ are independent

rior distribution

ginalized likelihood

• marginalized l

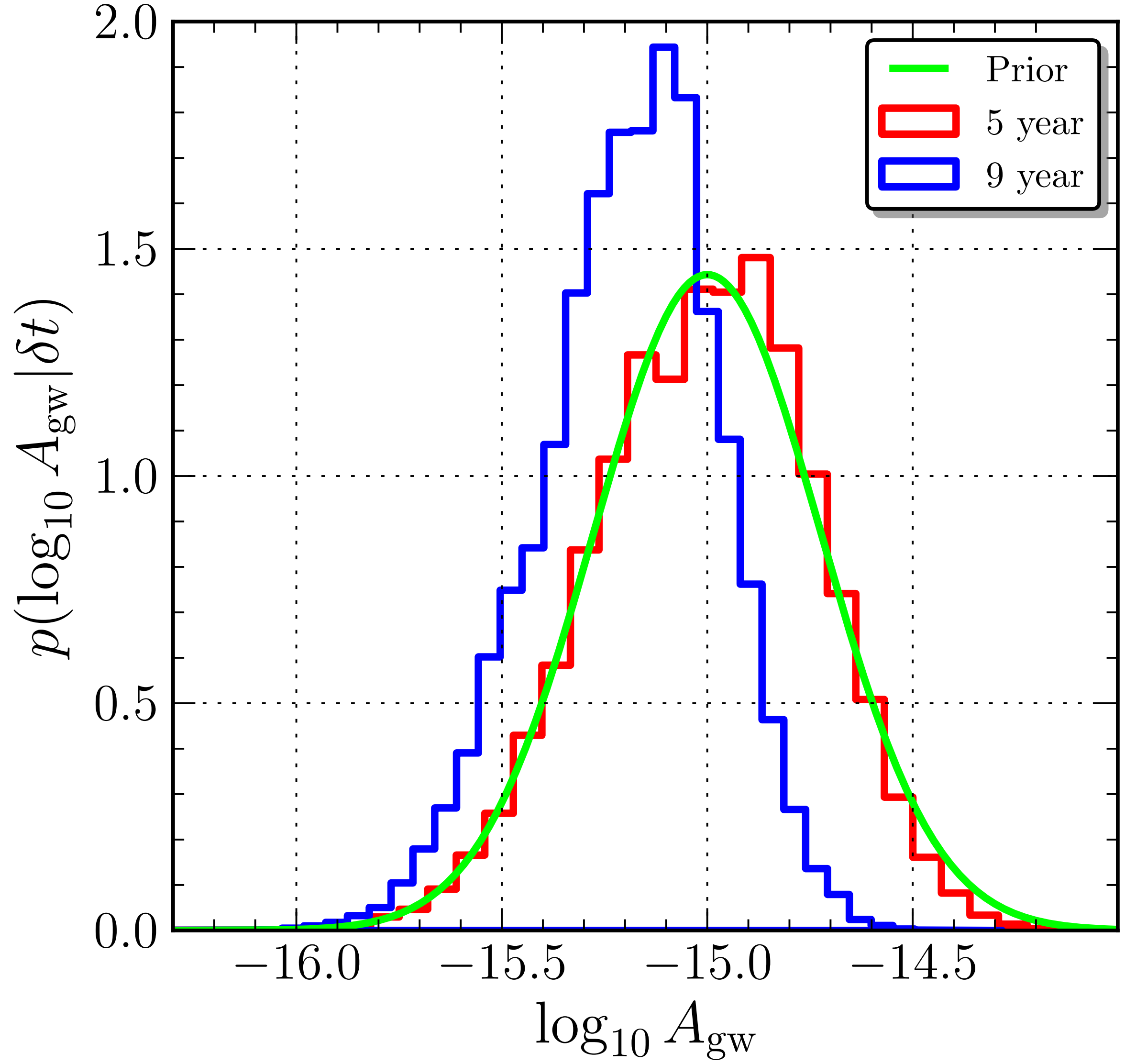
$\lambda d\theta$

J

• nuisance parameters are trivially marginalized: $p(\lambda|d) = \int p(\lambda, \theta|d) d\theta$

• Map out entire parameter space and then construct credible regions using marginalized posterior distributions.

Bayesian Example



Real NANOGrav data.

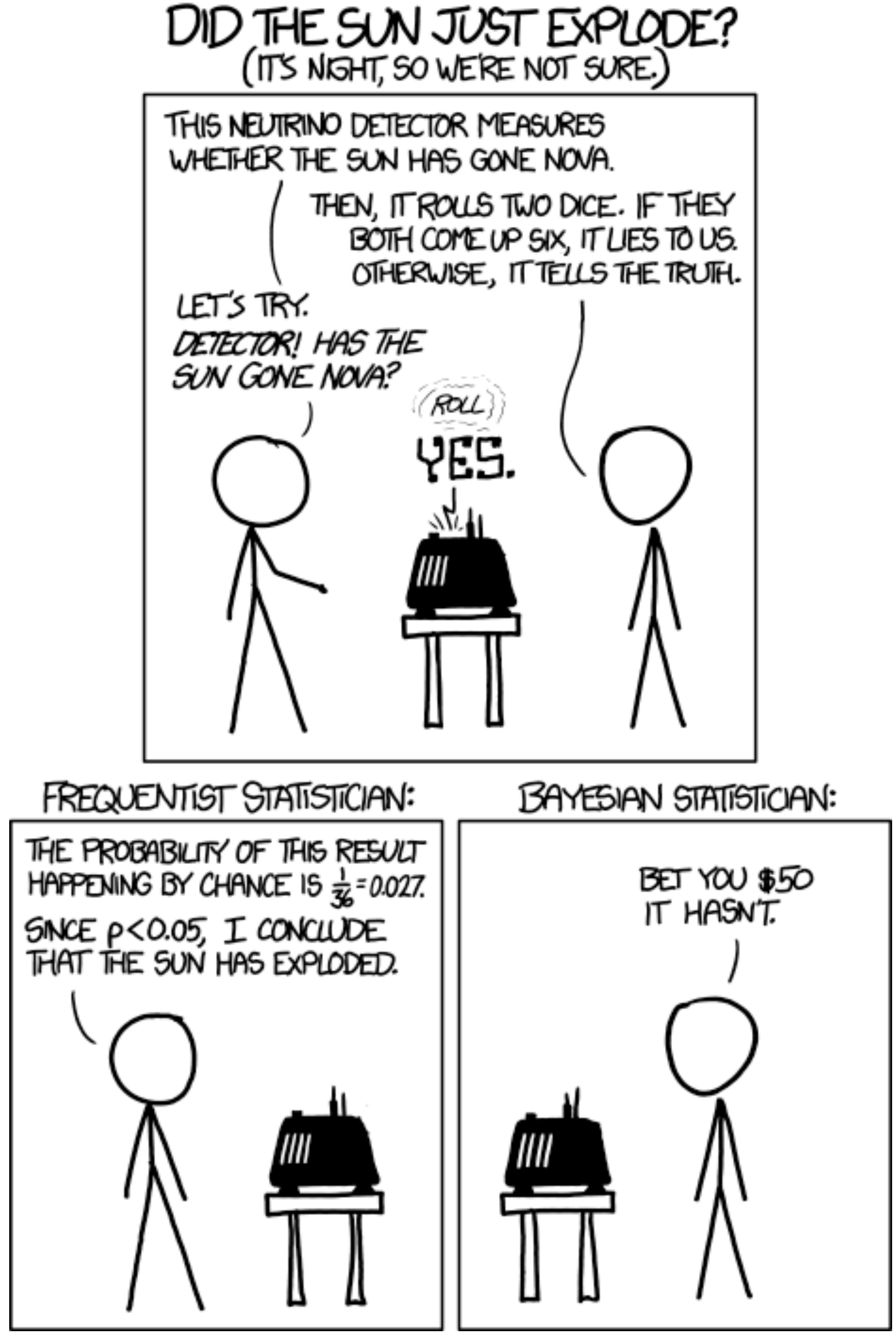
Green distribution is prior on GWB amplitude from simulations

Red is the posterior on the GWB amplitude using 5-year data.

Blue is posterior on the GWB amplitude using 9-year data.

Summary

- Frequentist:
  - pros:
    - usually fast to compute
    - usually easy to implement
  - cons:
    - relies on simulations to perform parameter estimation and hypothesis testing
    - no robust way to deal with nuisance parameters
- Bayesian
  - pros:
    - Does not rely on simulations, only data we have measured
    - Maps out entire parameter space not just peak
    - Robust way to deal with nuisance parameters
    - Directly measures "evidence" for a model
  - cons:
    - Not as easy to implement (especially in large parameter spaces)
    - Final results have some dependence on possibly subjective prior information


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.