Supplement

Statistical vs Bayesian Hypothesis Testing

We analyzed the data using Bayesian methods because several of the questions to be answered in our analysis concern model selection. Model selection involves identifying, from a group of appropriate mathematical models, those most strongly indicated by the data. When models are not nested (as is the case here), we cannot use traditional, frequentist methods.

Although probabilistic and traditional statistical algorithms produced similar results in many cases, the two types of analysis are based on very different calculations. Probability-based algorithms for parameter estimation and model selection calculate the joint posterior probability distribution over combinations of unknown parameter values, integrating over any unwanted, or nuisance, parameters. For example, in comparing the three possible slope models for the data of Fig. 5 (Location Experiment: Results and Analysis), this meant first calculating the 2-D joint posterior probability distribution for all possible pairs of slope and variance values allowed within each model. Calculating the probabilities of the models in the model comparison problem requires integrating over all possible slope and variance values separately for each model (the three models all assume homoscedastic Gaussian error with variance unknown). Because we have integrated over all possible parameter values to calculate the model probabilities, these probabilities reflect how well each model captures the data, regardless of the parameters' values (within the allowed range for that model).

When comparing models we will, following Jaynes (2003), present the evidence for a specific model relative to some specific alternative or class of alternative models, where evidence is defined as $e\left(\theta \mid DI\right) = 10\log\left[O\left(\theta \mid DI\right)\right]$. It is a function of the odds, $O\left(\theta \mid DI\right) = p\left(\theta \mid DI\right)\big/p\left(\overline{\theta} \mid DI\right)$; where θ is the hypothesis of interest, D is the data, and I is background information which specifies the appropriate error model and bounds on prior probability distributions. Evidence is measured in decibels (dB), a measure familiar to most behavioral and neuroscientists. An additional advantage of using a logarithmic measure like evidence is that the terms in the equations simply add. For example, if we were interested to know the effect of changing the prior evidence for the hypothesis θ_1 from 0 to 10 dB, we simply add that to the posterior evidence, which in turn changes by the same 10 dB.

All sampling distributions are assigned a Gaussian distribution unless otherwise stated. All prior distributions are assigned using Jeffreys rule (1946). This method of assigning prior distributions has the advantage that posterior inferences made using a Jeffreys prior are invariant under reparameterizations of the prior. Jeffreys rule assigns priors proportional to the square root of the Fisher information (Fisher, 1925). The Jeffreys prior for some parameter ϑ is

$$p(\vartheta \mid I) \propto \left[-E\left(\frac{\partial^2}{\partial \vartheta^2}\ln L(\vartheta)\right)\right]^{1/2}. \tag{1}$$

For example, using a Gaussian likelihood function on the space of possible samples, the Jeffreys prior for the location parameter (μ) is a uniform distribution:

$$\begin{aligned} p(\mu \mid I) &\propto \left[-E_x\left(\frac{\partial^2}{\partial \mu^2}\ln L(\mu)\right)\right]^{1/2} \\ &\propto \left[-E_x\left(\frac{\partial^2}{\partial \mu^2}\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right)\right)\right]^{1/2} \\ &\propto 1/\sigma. \end{aligned} \tag{2}$$

Given bounds (a b), the normalized prior probability distribution becomes $p(\mu \mid I) = (b - a)^{-1}$.

Confidence Intervals for Expected Hit Rates

Here, we provide 95% confidence intervals for the expected hit rates from both the Location and Scale experiments. We are providing these values as many readers are used to confidence intervals as a means of judging the accuracy of the data. However, we must emphasize that these intervals are *not* relevant to the hypotheses we compare in the paper. The frequentist *p*=.05 criterion cannot be translated into a fixed number of dB of evidence for or against some hypothesis. To calculate odds (and therefore evidence), one needs two mutually exclusive hypotheses (e.g., sets of possible parameter values) that bear on the comparison of interest. That is, we would like the ratio of the probability of the hypothesis being correct vs. incorrect. The .05 probability in the frequentist criterion is the probability of obtaining a parameter estimate equal to or more extreme than that obtained experimentally, conditioned on the truth of the null hypothesis. Since that calculation assumed the null hypothesis was true, the result says nothing about the probability of the null hypothesis being incorrect, and odds can’t properly be calculated for the question of interest.

TABLE 1 – Location Experiment: Expected Hit Rates and Confidence Intervals

	Strategy 1	**Strategy 2**	**Strategy 3**	**Strategy 4**	**Strategy 5**
Condition 1	0.51 [0.48 - 0.55]	0.47 [0.44 - 0.51]	0.28 [0.25 - 0.31]	0.35 [0.32 - 0.38]	0.38 [0.35 - 0.42]
Condition 2	0.47 [0.44 - 0.51]	0.58 [0.54 - 0.61]	0.56 [0.52 - 0.59]	0.48 [0.44 - 0.51]	0.29 [0.26 - 0.33]
Condition 3	0.41 [0.37 - 0.45]	0.56 [0.52 - 0.59]	0.58 [0.55 - 0.62]	0.46 [0.42 - 0.49]	0.40 [0.37 - 0.44]
Condition 4	0.28 [0.25 - 0.32]	0.49 [0.46 - 0.53]	0.49 [0.46 - 0.53]	0.56 [0.53 - 0.60]	0.47 [0.43 - 0.50]
Condition 5	0.15 [0.13 - 0.18]	0.37 [0.33 - 0.40]	0.24 [0.21 - 0.27]	0.48 [0.44 - 0.51]	0.49 [0.45 - 0.52]

TABLE 2 – Scale Experiment: Expected Hit Rates and Confidence Intervals

	Strategy 1	**Strategy 2**	**Strategy 3**
Condition 1	0.58 [0.56 - 0.61]	0.62 [0.60 - 0.65]	*0.57* [0.54 - 0.60]
Condition 2	0.49 [0.46 - 0.52]	0.55 [0.52 - 0.58]	0.54 [0.51 - 0.57]
Condition 3	0.37 [0.35 - 0.40]	0.48 [0.45 - 0.51]	0.48 [0.45 - 0.50]