What to believe: Bayesian methods for data analysis

John K. Kruschke

Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, USA

Although Bayesian models of mind have attracted great interest from cognitive scientists, Bayesian methods for data analysis have not. This article reviews several advantages of Bayesian data analysis over traditional null-hypothesis significance testing. Bayesian methods provide tremendous flexibility for data analytic models and yield rich information about parameters that can be used cumulatively across progressive experiments. Because Bayesian statistical methods can be applied to any data, regardless of the type of cognitive model (Bayesian or otherwise) that motivated the data collection, Bayesian methods for data analysis will continue to be appropriate even if Bayesian models of mind lose their appeal.

Cognitive science should be Bayesian even if cognitive scientists are not

An entire issue of Trends in Cognitive Sciences was devoted to the topic of Bayesian models of cognition [1] and there has been a surge of interest in Bayesian models of perception, learning and reasoning [2–6]. The essential premise of the Bayesian approach is that the rational, normative way to adjust knowledge when new data are observed is to apply Bayes' rule (i.e. the mathematically correct formula) to whatever representational structures are available to the reasoner. The promise that spontaneous human behavior might be normatively Bayesian on some to-be-discovered representation has driven a surge in theoretical and empirical research.

Ironically, the behavior of researchers themselves has often not been Bayesian. There are many examples of a researcher positing a Bayesian model of how people perform a cognitive task, collecting new data to test the predictions of the Bayesian model, and then using non-Bayesian methods to make inferences from the data. These researchers are usually aware of Bayesian methods for data analysis, but the mortmain of 20th century methods compels adherence to traditional norms of behavior.

Traditional data analysis has many well-documented problems that make it a feeble foundation for science, especially now that Bayesian methods are readily accessible [7–9]. Chief among the problems is that the basis for declaring a result to be 'statistically significant' is ill defined: the so-called p value has no unique value for any set of data. Another problem with traditional analyses is that they produce impoverished estimates of parameter values, with no indication of trade-offs among parameters and with confidence intervals that are ill defined because they are based on p values. Traditional methods also often impose many computational constraints and assumptions into which data must be inappropriately squeezed.

The death grip of traditional methods can be broken. Bayesian methods for data analysis are now accessible to all, thanks to advances in computer software and hardware. Bayesian analysis solves the problems of traditional methods and provides many advantages. There are no p values in Bayesian analysis, inferences provide rich and complete information regarding all the parameters, and models can be readily customized for different types of data. Bayesian methods also coherently estimate the probability that an experiment will achieve its goal (i.e. the statistical power or replication probability).

It is important to understand that Bayesian methods for data analysis are distinct from Bayesian models of mind. In Bayesian data analysis, any useful descriptive model of the data has parameters estimated by normative, rational methods. The descriptive models have no necessary relation or commitment to particular theories of the natural mechanisms that actually generated the data. Thus, every cognitive scientist, regardless of his or her preferred model of cognition, should use Bayesian methods for data analysis. Even if Bayesian models of mind lose favor, Bayesian data analysis remains appropriate.

Null hypothesis significance testing (NHST)

In NHST, after collecting data, a researcher computes the value of a summary statistic such as t or F or χ^2, and then determines the probability that so extreme a value could have been obtained by chance alone from a population with

Glossary

Analysis of variance (ANOVA): when metric data (e.g. response times) are measured in each of several groups, traditional ANOVA decomposes the variance among all data into two parts: the variance between group means and the variance among data within groups. The underlying descriptive model can be used in Bayesian data analysis.

Bayes' rule: a simple mathematical relationship between conditional probabilities that relates the posterior probability of parameter values, on the one hand, to the probability of the data given the parameter values, and the prior probability of the parameter values, on the other hand. The formula is named after Thomas Bayes (1702–1761), an English minister and mathematician.

Chi-square(χ^2): the Pearson χ^2 value is a measure of the discrepancy between the frequencies observed for nominal values and what would be expected according to a (null) hypothesis. In NHST, the sampling distribution of the Pearson χ^2 distribution is approximated by the continuous χ^2 distribution when the expected frequencies are not too small.

Descriptive versus explanatory model: descriptive models summarize relations between variables without ascribing mechanistic meaning to the functional form or to the parameters, whereas explanatory models do make such semantic ascriptions. Bayesian inference for descriptive models of data is desirable regardless of whether Bayesian explanatory models account for cognition.

***p* value**: in NHST, the p value is the probability of obtaining the observed value of a sample statistic (such as t, F, χ^2) or a more extreme value if the data were generated from a null-hypothesis population and sampled according to the intention of the experimenter, where the intention could be to stop at a pre-specified sample size or after a pre-specified sampling duration, or to check after every 10 observations and stop either when significance is achieved or at the end of the week, whichever is sooner.

Corresponding author: Kruschke, J.K. (kruschke@indiana.edu).

1364-6613/$ – see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2010.05.001

no effect if the experiment were repeated many times. If the probability of obtaining the observed value is small (e.g. $p < 0.05$), then the null hypothesis is rejected and the result is deemed significant.

Friends do not let friends compute p values

The crucial problem with NHST is that the p value is defined in terms of repeating the experiment, and what constitutes the experiment is determined by the experimenter's intentions. The single set of data could have arisen from many different experiments, and therefore the single set of data has many different p values. In all the conventional statistical tests, it is assumed that the experimenter intentionally fixed the sample size in advance, so that repetition of the experiment means using the same fixed sample size. But if the experiment were instead run for a fixed duration with subjects arriving randomly in time, then repetition of the experiment means repeating a run of that duration with randomly different sample sizes, and therefore the p value is different [10]. If data collection stops for some other reason, such as not being able to find any more subjects of the necessary type (e.g. with a specific brain lesion) or because a research assistant unexpectedly quits, then it is unclear how to compute a p value at all, because we do not know what it means to repeat the experiment. This dependence of p on the intended stopping rule for data collection is well known [11,12,9], but rarely if ever acknowledged in applied textbooks on NHST.

The only situation in which standard NHST textbooks explicitly confront the dependence of p on experimenter intention is when multiple comparisons are made. When there are several conditions to be compared, each comparison inflates the probability of spuriously declaring a difference to be non-zero. To compensate for this inflation of false alarms, different 'corrections' can be made to the p-value criterion used to declare significance. These corrections go by the names of Bonferroni, Scheffe, Tukey, Dunnett, Hsu or a variation called the false discovery rate (FDR) [for an excellent review, see reference 13, Ch. 5]. Each correction specifies a penalty for multiple comparisons that is appropriate to the intended set of comparisons. This penalty for exploring the data provides incentive for researchers to feign interest in only a few comparisons and even to pretend that various subsets of conditions are in different 'experiments'. Indeed, it is trivial to make any observed difference non-significant merely by conceiving of many other conditions with which to compare the present data and having the intention to eventually collect the data and make the comparisons.

Poverty of point estimates

NHST summarizes a data set with a value such as t or F, which in turn is based on a point estimate from the data, such as the mean and standard deviation for each group. The point estimate is the value for the parameter that makes the model most consistent with the data in the sense of minimizing the sum squared deviation or maximizing the likelihood (or some other measure of consistency).

Unfortunately, the point estimate provides no information about the range of other parameter values that are reasonably consistent with the data. Some researchers use confidence intervals for this purpose. But some NHST analyses do not easily provide confidence intervals, such as χ^2 analyses of contingency-table cell probabilities. More fundamentally, confidence intervals are as fickle as p values because a confidence interval is simply the range of parameter values that would not be rejected by a significance test (and significance tests depend on the intentions of the analyst). In recognition of this fact, many computer programs automatically adjust the confidence intervals they produce, depending on the set of intended comparisons selected by the analyst.

Point estimates of parameters also provide no indication of correlations between plausible parameter values. Consider simple linear regression of response latency on stimulus contrast, assuming that RT decreases as contrast increases. Many different regression lines fall plausibly close to the data, but lines with higher y intercepts must have steeper (negative) slopes; hence, the intercept and slope are (negatively) correlated. There are methods for approximating these correlations at point estimates, but these approximations rely on assumptions about asymptotic distributions for large samples.

Impotence of power computation

Statistical power in NHST is the probability of rejecting the null hypothesis when an alternative hypothesis is true. Because power increases with sample size, estimates of power are often used in research planning to anticipate the amount of data that should be collected. Closely related to power is replication probability, which is the probability that a result found to be significant in one experiment will also be found to be significant in a replication of the experiment. Replication probability can be used to assess the reliability of a finding. To estimate power and replication probability, the point estimate from a first experiment is used as the alternative hypothesis to contrast with the null hypothesis. Unfortunately, a point estimate yields little information about other alternative hypotheses that are reasonably consistent with the initial data. The other alternative hypotheses can span a very wide range, with each one yielding very different estimates of power and replication probability. Therefore, the replication probability has been determined to be 'virtually unknowable' [14]. Thus, NHST in combination with point estimation leaves the scientist with unclear estimates of power and replication probability, and hence provides a very weak basis for assessing the reliability of an outcome.

Computational constraints

As outlined above, NHST provides a paucity of dubious information. To obtain this, the analyst is also subject to many computational constraints. For example, in analysis of variance (ANOVA), computations are much easier to conduct and interpret if all conditions have the same number of data points (i.e. so-called balanced designs) [Ref. 13, pp. 320–343]. Standard ANOVA also demands homogeneity of variances across the conditions. In multiple regression, computations (and interpretation of results) can go haywire if the predictors are strongly correlated. In χ^2 analyses of contingency tables, the expected values should be approximately 5 or greater for approximated p values to be adequate. Various analyses suffer when data points are missing. There are numerous other computational strictures when pursuing point estimates and NHST.

A feeble foundation for empirical science

In summary, when a researcher conducts NHST, the analysis typically begins with many restrictive computational assumptions and the result is a point estimate of parameters, with no clear range of other plausible parameter values, little information about how the parameter values trade off against each other, estimates of power and replication probability that can be 'virtually unknowable' [14] and a declaration of significance based on a p value that depends on the experimenter's intention for stopping data collection and the analyst's inquisitiveness about other conditions.

Bayesian data analysis

In Bayesian data analysis, the researcher uses a descriptive model that is easily customizable to the specific situation without the computational restrictions in conventional NHST models. Before considering any newly collected data, the analyst specifies the current uncertainty for parameter values, called a prior distribution, that is acceptable to a skeptical scientific audience. Then Bayesian inference yields a complete posterior distribution over the conjoint parameter space, which indicates the relative credibility of every possible combination of parameter values. In particular, the posterior distribution reveals complete information about correlations of credible parameter values. Bayesian analysis also facilitates straightforward methods for computing power and replication probability. There are no p values and no corrections for multiple comparisons, and there is no need to determine whether the experimenter intended to stop when $n = 47$ or ran a clandestine significance test when $n = 32$. Moreover, Bayesian analysis can implement cumulative scientific progress by incorporating previous knowledge into the specification of the prior uncertainty, as deemed appropriate by peer review. This section briefly explains each of these points, with an accompanying example.

Model flexibility and appropriateness

In Bayesian data analysis, the descriptive model can be easily customized to the type and design of the data. For example, when the dependent variable is dichotomous (e.g. correct or wrong) instead of metric (e.g. response time) and when there are several different treatment groups, then a model analogous to ANOVA can be built that directly models the dichotomous data without assuming any approximations to normality. Box 1, with its accompanying Figures 1 and 2, provides a detailed example. The Bayesian model also has no need to assume homogeneity of variance, unlike NHST ANOVA.

Bayesian inference is also computationally robust. There is no difficulty with unequal numbers of data points in different groups of an experiment (unlike standard methods for NHST ANOVA). There is no computational problem when predictors in a multiple regression are strongly correlated (unlike least-squares point estimation). There is no need for expected values in a contingency table to exceed 5 (unlike NHST χ^2 tests). Bayesian methods do not suffer from these problems because Bayesian inference effectively applies to one datum at a time. When each datum is considered, the posterior distribution is updated to reflect that datum.

Box 1. Example of a Bayesian model for data analysis

Consider an experiment that investigated the difficulty of learning a new category structure after learning a previous structure [38]. Figure 1 shows a schematic of the four types of structure. All participants learned the top structure and then each person was trained on one of the four structures in the bottom of the diagram. In the initially learned structure, only the two horizontal dimensions are relevant and the vertical dimension is irrelevant. In the reversal shift, all category assignments are reversed. In the relevant shift, one of the initially relevant dimensions remains relevant. In the irrelevant shift, the initially irrelevant dimension becomes relevant. In the compound shift, a different compound of dimensions is relevant. Different theories of learning make different predictions for the relative difficulty of these shifts. Variations of this relevance-shift design have been used by subsequent researchers [39,40] because it decouples cue relevance from outcome correlation and it compares reversal and relevance shifts without introducing novel stimulus values.

A Bayesian model for this design is shown in Figure 2. The essential ideas are simple: the accuracy observed for each individual in the shift phase is assumed to reflect the underlying true accuracy for that individual, and individual accuracy values are assumed to come from a distribution determined by the shift difficulty. (Other distributions for individual differences could be used if desired [e.g. 41,42].) The primary goal of the analysis is to estimate the parameters of the group distributions. These parameters include the mean μ_c for the cth condition and the certainty κ_c, which can be thought of as the reciprocal of standard deviation: high certainty implies a narrow spread for accuracy for that condition. The group estimates mutually inform each other via the global-level distributions, which provide shrinkage of the group estimates. The prior constants were chosen as only mildly informed.

Although this article highlights the use of conventional descriptive models, Bayesian methods are also advantageous for estimating parameters in other models, such as multidimensional scaling [15], categorization [16], signal detection theory [17], process dissociation [18] and domain-specific models such as children's number knowl-

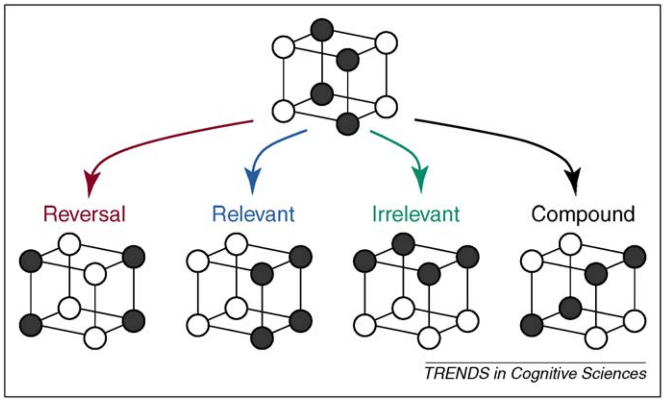

Figure 1. Four types of shift trained in a learning experiment [38]. The stimuli had three binary-valued dimensions, indicated by the cube edges, and each stimulus had an experimenter-specified binary category label, indicated by the color of the disk at each corner. More information is in Box 1. [Adapted with permission from Ref. 38].

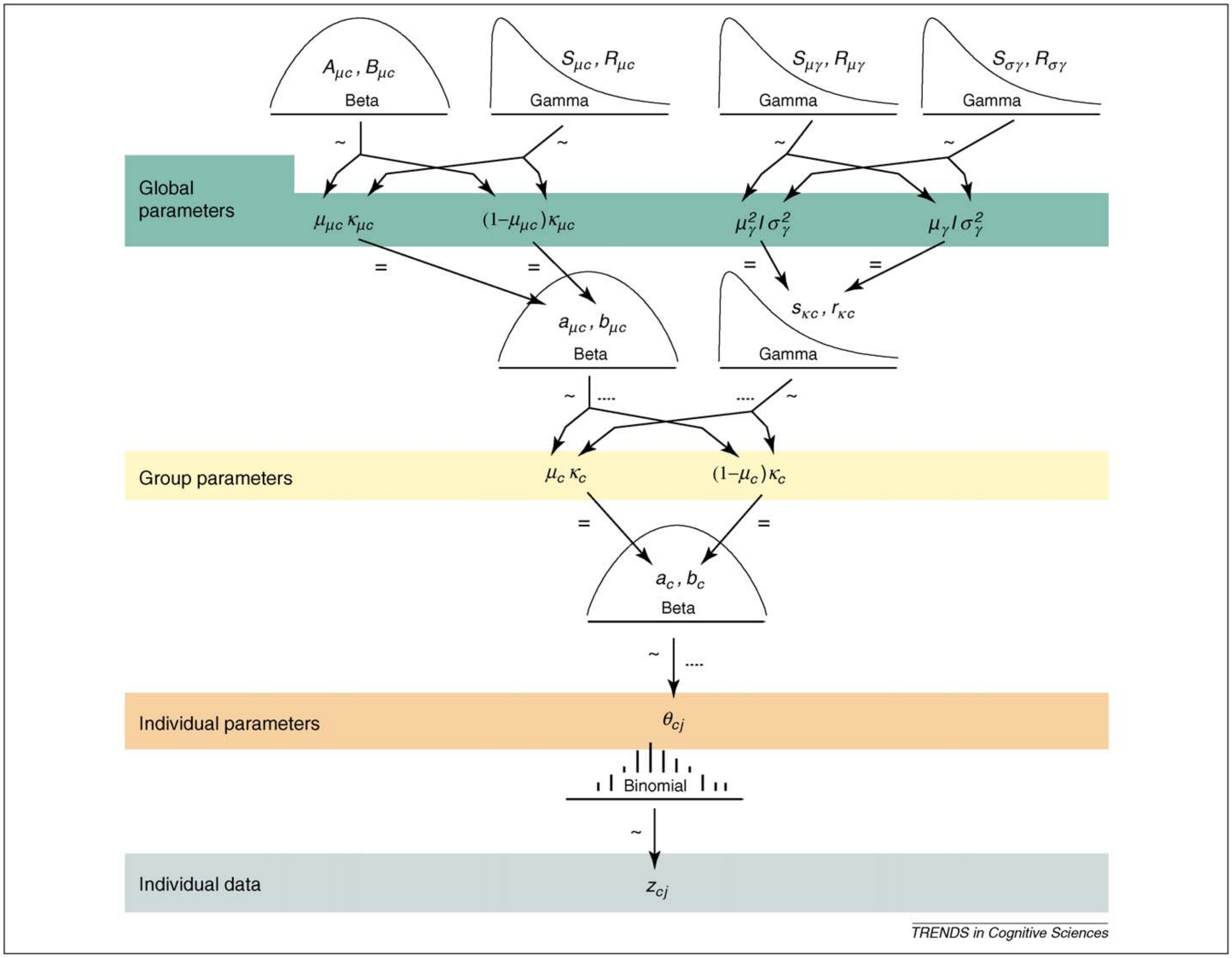

Figure 2. Diagram of hierarchical model for experiment in Figure 1. The distributions are represented by iconic caricatures, which are not meant to indicate the actual shape of the prior or posterior distributions. The model uses 12 group and global parameters, as well as 240 individual parameters. At the bottom of the diagram, the number correctly identified by individual *j* for condition *c* is denoted z_{cj}. This number represents the underlying accuracy θ_{cj} for that individual. The different individual accuracy value come from a beta distribution for that condition, which has mean μ_c and certainty κ_c. These group-level estimates of μ_c and κ_c are the parameters of primary interest. The group-level parameters come from global-level distributions, whereby data from one group can influence estimates for other groups. Shrinkage for the group means is estimated by the global certainty $\kappa_{\mu c}$ and shrinkage for the group certainties is estimated by the global standard deviation σ_{γ}. The top-level prior distributions are vague and only mildly informed. The top-level γ distributions could be replaced by uniform or folded-*t* distributions [43].

edge [19], among many others. Parameters in conventional models have a generic descriptive interpretation, such as a slope parameter in multiple linear regression indicating how much the predicted value increases when the predictor value increases. Parameters in other models have more domain-specific interpretation, such as a position parameter in multidimensional scaling indicating relative location on latent dimensions in psychological space.

Richly informative inferences

Bayesian analysis delivers a posterior distribution that indicates the relative credibility of every combination of parameter values. This posterior distribution can be examined in any way deemed meaningful by the analyst. In particular, any number of comparisons across group parameters can be made without penalty because the posterior distribution does not change when it is examined from different perspectives. Box 2, with Figures 3 and 4, provides an example.

The posterior distribution inherently provides credible intervals for the parameter estimates (without reference to sampling distributions, unlike confidence intervals in NHST). For example, in contingency table analysis, credible intervals for cell probabilities are inherent in the analysis, unlike χ^2 tests. The posterior distribution inherently shows covariation among parameters, which can be very important for interpreting descriptive models in applications such as multiple regression.

Coherent power analysis and replication probability

In traditional power analysis, the researcher considers a single alternative value for the parameter and determines the probability that the null hypothesis would be rejected. This type of analysis does not take into account uncertainty for the alternative value. The replication probability, which is related to power, is the probability that rejection of the null hypothesis would be achieved (or not) if data collection were conducted a second time. Because point

Box 2. The posterior distribution and multiple comparisons

The full posterior distribution for the model in Figure 2 is a joint distribution over the 252-dimensional parameter space. Full details of how to set up and execute a Bayesian analysis are provided in [20]. The group-level parameters are shown in Figure 3. Each point is a representative value sampled from the posterior distribution. Note that credible parameter values can be correlated; for example, within group 2 there is a positive correlation ($r = 0.29$) between μ_2 and κ_2.

In the actual data, the median accuracy is 0.531 for condition 4 and 0.906 for condition 1, a difference of 0.375. But in the posterior distribution, the median μ is 0.598 for condition 4 and 0.886 for condition 1, a difference of only 0.288. This compression of the estimates relative to the data is one manifestation of shrinkage, which reflects the prior knowledge that all the groups come from the same global population.

The prior distribution at the top level of the model was only mildly informed by previous knowledge of human performance in this type of experiment. The posterior distribution changes negligibly if the prior information is changed within any reasonable range, because the data set is moderately large and the hierarchical structure of the model means that low-level parameters can be mutually constrained via higher-level estimates.

The posterior distribution can be examined from arbitrarily many perspectives to extract useful comparative information. For example, to assess the magnitude of difference between the means of the reversal and relevant conditions, the difference $\mu_{Rev} - \mu_{Rel}$ is computed at every representative point and the distribution of those differences is inspected, as shown in the upper-left histogram of Figure 4. It is evident that the modal difference is 0.097 and the entire posterior distribution falls far above zero. Therefore, it is highly credible that reversal shift is easier than relevant shift. A crucial idea is that the posterior distribution does not change when additional comparisons are made and there is no need for corrections for multiple comparisons. The prior structure already provided shrinkage, as informed by the data, which mitigates false alarms.

We can also examine differences in precision across groups. The first histogram in the third row of Figure 4 indicates that the precision for the reversal group is credibly higher than that for the relevant group. This difference in within-group variability would not be revealed by traditional ANOVA because it presumes equal variances for all groups.

estimation in NHST yields no posterior distribution over credible parameter values, the replication probability is 'virtually unknowable' [14].

By contrast, Bayesian analysis provides coherent methods for computing the power and replication probability. Bayesian power analysis uses the posterior distribution to sample many different plausible parameter values, and for each parameter value generates plausible data that simulate a repetition of the experiment. The simulated data can then be assayed regarding any research goal, such as excluding a null value or attaining a desired degree of accuracy for the parameter estimate. Box 3 and Figure 5 provide an example and some details [see Ref. 20, for an extended tutorial]. The Bayesian framework for computing power and replication probability is the best we can do, because the posterior distribution over the parameter values is our best representation of the world based on the information we currently have. The Bayesian framework naturally incorporates our uncertainty into the estimates of power and replication probability. Other research goals involving the highest density interval (HDI; defined and exemplified in Figure 4) can be used to define power [21–26], as can goals involving model comparison [22,27].

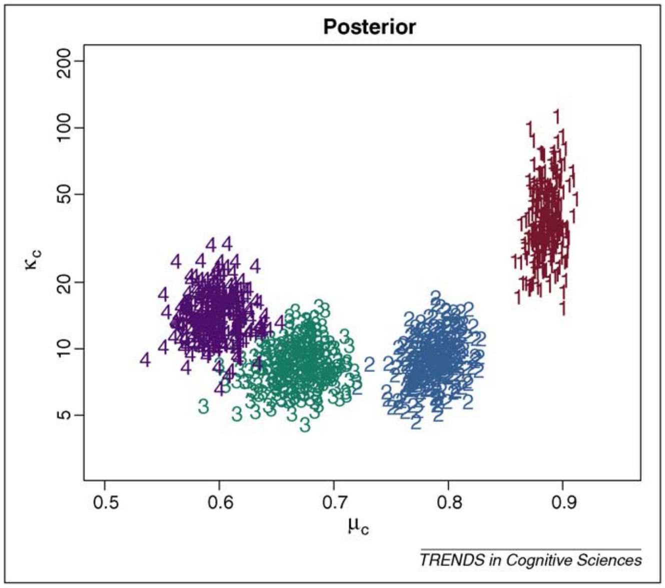

Figure 3. Credible posterior values for the group-level means μ_c and certainties κ_c for each shift condition in Figure 1. Conditions: 1 = reversal, 2 = relevant, 3 = irrelevant, 4 = compound. The points are representative values from the continuous posterior distribution. The program for generating the posterior distribution in Figures 3 and 4 was written in the R language [44] using the BRugs interface (BRugs user manual (the R interface to BUGS), http://mathstat.helsinki.fi/openbugs/data/Docu/BRugs%20Manual.html) to the OpenBUGS version [45] of BUGS [46].

Appropriateness of the prior distribution

Bayesian analysis begins with a prior distribution over the parameter values. The prior distribution is a model of uncertainty and expresses the relative credibility of the parameter values in the absence of new data. Bayesian analysis is the mathematically normative way to reallocate credibility across the parameter values when new data are considered. The resulting posterior distribution is always a compromise between the prior credibility of the parameter values and the likelihood of the parameter values for the data. In most applications, however, the prior distribution is vague and only mildly informed, and therefore has little influence on the posterior distribution (Figures 3 and 4).

Prior distributions are not covertly manipulated to predetermine a desired posterior. Instead, prior distributions are explicitly specified and must be acceptable to the audience of the analysis. For scientific publications, the audience consists of skeptical peer reviewers, editors and scientific colleagues. Moreover, the robustness of conclusions gleaned from a posterior distribution can be checked by running the Bayesian analysis with other plausible prior distributions that might better accord with different audience members.

Prior distributions should not be thought of as an innocuous nuisance. On the contrary, consensually informed prior distributions permit cumulative scientific knowledge to rationally affect conclusions drawn from new observations. As a simple example of the importance of prior distributions, consider a situation in which painstaking survey work has previously established that in the general population only 1% of subjects abuse a certain dangerous drug. Suppose that a person is randomly selected from this population for a drug test and the test yields a positive result. Suppose that the test has a 99% hit rate and a 5% false alarm rate. If we ignore the prior knowledge, we

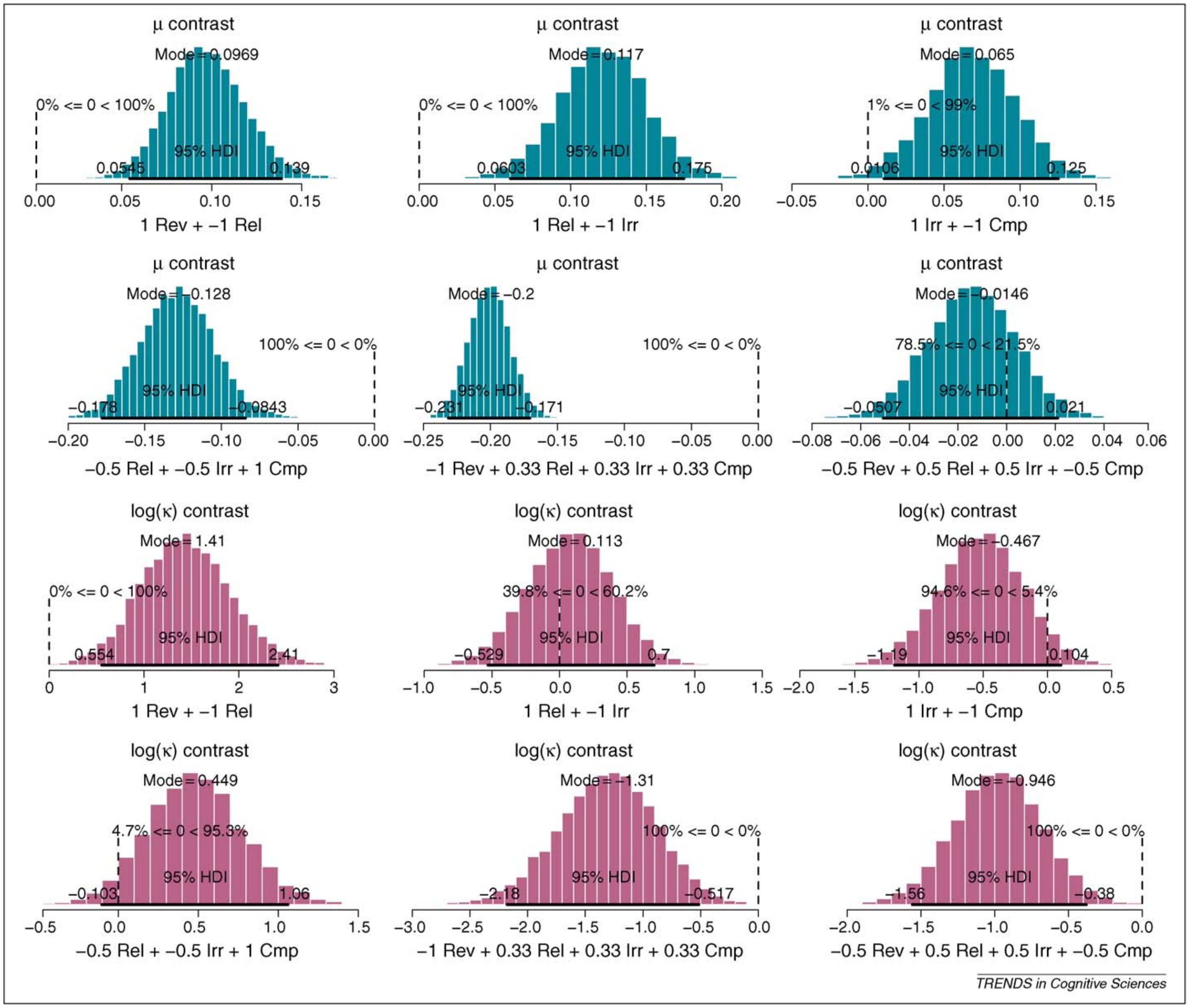

Figure 4. Posterior distributions of selected comparisons of the parameter values. The highest density interval is denoted 95% HDI, such that all values within the interval have higher credibility than values outside the interval, which spans 95% of the distribution. Various contrasts among μ and κ values are shown, including complex comparisons as motivated by different theories. For example, if difficulty in learning the shift is based only on the number of exemplars with a change in category, then reversal should be more difficult than the other three shifts, which should be equally easy. The difference between the average for non-reversal and reversal shifts is shown in the middle column of the second row, from which it is evident that the difference is credibly in the opposite direction. If there is no influence of the first phase on the second, then the relevant and irrelevant structures, which have only one relevant dimension, should be easier to learn than the reversal and compound structures, which have two relevant dimensions. The corresponding contrast is shown in the right-hand column of the second row, which indicates that zero is among the credible differences.

would conclude that there is at least a 95% chance that the tested person abuses the drug. But if we take into account the strong prior knowledge, then we conclude that there is only a 17% chance that the person abuses the drug.

Some Bayesian analysts attempt to avoid mildly informed or consensually informed prior distributions and opt instead to use so-called objective prior distributions that satisfy certain mathematical properties [e.g., 28]. Unfortunately, there is no unique definition of objectivity. Perhaps more seriously, in applications that involve model comparison, the posterior credibility of the models can depend dramatically on the choice of objective prior distribution for each model [e.g., 29]. These model-comparison situations demand prior distributions for the models that are equivalently informed by prior knowledge.

In summary, incorporation of prior knowledge into Bayesian analysis is crucial (recall the drug test example) and consensual (as in peer review). Moreover, this can be cumulative. As more research is conducted in a domain, consensual prior knowledge can become stronger, reflecting genuine progress in science.

Models of cognition and models of data

The posterior distribution of a Bayesian analysis only tells us which parameter values are relatively more or less credible within the realm of models that the analyst cares to consider. Bayesian analysis does not tell us what models to consider in the first place. For typical data analysis, descriptive models are established by convention: most empirical researchers are familiar with cases of the gener-

Box 3. Retrospective power and replication probability

Although the posterior distributions in Figures 3 and 4 show credibly non-zero differences between the conditions, we can ask what was the probability of achieving that result. To conduct this retrospective power analysis, we proceed as indicated in Figure 5. Because our best current description of the world is the actual posterior distribution, we use it to generate simulated data sets, each of which is subjected to a Bayesian analysis like that for the original data.

In 200 simulated experiments, using the same number of subjects per group as in the original experiment, for 100% the 95% HDI for $\mu_{Rev} - \mu_{Rel}$ fell above zero, for 88% the HDI for $\mu_{Rel} - \mu_{Irr}$ fell above zero, for 46% the HDI for $\mu_{Irr} - \mu_{Cmp}$ fell above zero, and for 20% the HDI for $\kappa_{Rev} - \kappa_{Rel}$ fell above zero. Thus, depending on the goal, the experiment might have been over- or underpowered. Because our primary interest is in the μ_c parameters, follow-up experiments should reallocate subjects from the relevant condition to the compound condition. Unlike traditional ANOVA [13, pp. 320–343], Bayesian computations are not challenged by so-called unbalanced designs that have unequal subjects per condition.

We might also be interested in the probability of reaching the research goal if we were to replicate the experiment (i.e. run the experiment a second time). In 200 simulated experiments using the actual posterior distribution as the generator for simulated data, and using the actual posterior distribution as the prior distribution for the Bayesian analysis (for detailed methods see [20]), for 100% the 95% HDI for $\mu_{Rev} - \mu_{Rel}$ fell above zero, for 99% the HDI for $\mu_{Rel} - \mu_{Irr}$- fell above zero, for 85% the HDI for $\mu_{Irr} - \mu_{Cmp}$ fell above zero and for 96% the HDI for $\kappa_{Rev} - \kappa_{Rel}$ fell above zero. Thus, the replication probability is computed naturally in a Bayesian framework. In NHST, the replication probability is difficult to compute and interpret [14] because there is no posterior distribution to use as a data simulator.

alized linear model [30,31] such as linear regression, logistic regression and ANOVA. Given this space of models, the rational approach to infer what parameter values are most credible is Bayesian. Therefore, cognitive scientists should use Bayesian methods for analysis of their empirical data. Advanced introductions can be found in [32–36] and an accessible introduction including power analyses can be found in [20]. Bayesian methods for data analysis are available now and will be the preferred method for the 21st century [37], regardless of whether or not cognitive scientists invent sustainable Bayesian models of mind.

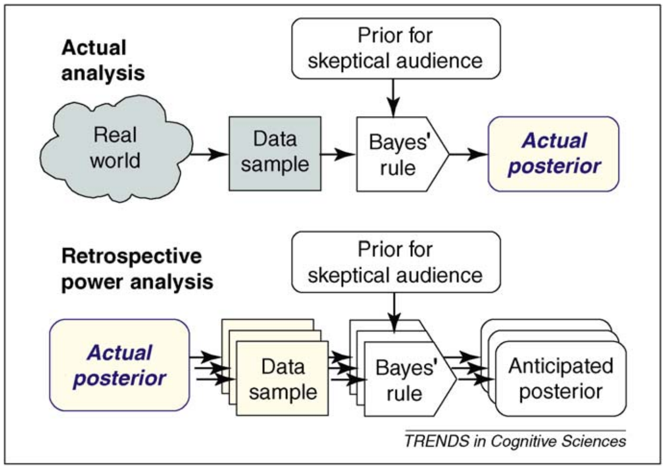

Figure 5. Bayesian power analysis. The upper panel shows an actual analysis in which the real world generates one sample of data and Bayesian analysis uses a prior distribution acceptable to a skeptical audience, thereby generating the actual posterior distribution. The lower panel shows a retrospective power analysis. The parameter values of the actual posterior distribution are used to generate multiple simulated data sets, each of which is subjected to Bayesian analysis. When computing the replication probability, the prior distribution is replaced by the actual posterior distribution because the replication builds on the previous experiment (process not shown) [20].

References

1 Chater, N. *et al.* (2006) Probabilistic models of cognition. *Trends Cogn. Sci.* 10, 287–344
2 Anderson, J.R. (1990) *The Adaptive Character of Thought*, Erlbaum
3 Chater, N. and Oaksford, M., eds (2008) *The Probabilistic Mind*, Oxford University Press
4 Doya, K. *et al.*, eds (2007) *Bayesian Brain: Probabilistic Approaches to Neural Coding*, MIT Press
5 Griffiths, T.L. *et al.* (2008) Bayesian models of cognition. In *The Cambridge Handbook of Computational Psychology* (Sun, R., ed.), pp. 59–100, Cambridge University Press
6 Kruschke, J.K. (2008) Bayesian approaches to associative learning: from passive to active learning. *Learn. Behav.* 36, 210–226
7 Edwards, W. *et al.* (1963) Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242
8 Lee, M.D. and Wagenmakers, E.J. (2005) Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychol. Rev.* 112, 662–668
9 Wagenmakers, E.J. (2007) A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* 14, 779–804
10 Kruschke, J.K. (2010) Bayesian data analysis. *Wiley Interdisciplin. Rev. Cogn. Sci.* DOI: 10.1002/wcs.72.
11 Berger, J.O. and Berry, D.A. (1988) Statistical analysis and the illusion of objectivity. *Am. Sci.* 76, 159–165
12 Lindley, D.V. and Phillips, L.D. (1976) Inference for a Bernoulli process (a Bayesian view). *Am. Stat.* 30, 112–119
13 Maxwell, S.E. and Delaney, H.D. (2004) *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (2nd edn), Erlbaum.
14 Miller, J. (2009) What is the probability of replicating a statistically significant effect? *Psychon. Bull. Rev.* 16, 617–640
15 Lee, M.D. (2008) Three case studies in the Bayesian analysis of cognitive models. *Psychon. Bull. Rev.* 15, 1–15
16 Vanpaemel, W. (2009) BayesGCM: software for Bayesian inference with the generalized context model. *Behav. Res. Methods* 41, 1111–1120
17 Lee, M.D. (2008) BayesSDT: software for Bayesian inference with signal detection theory. *Behav. Res. Methods* 40, 450
18 Rouder, J.N. *et al.* (2008) A hierarchical process-dissociation model. *J. Exp. Psychol.* 137, 370–389
19 Lee, M.D. and Sarnecka, B.W. (2010) A model of knower-level behavior in number concept development. *Cogn. Sci.* 34, 51–67
20 Kruschke, J.K. (2010) *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*, Academic Press/Elsevier Science
21 Adcock, C.J. (1997) Sample size determination: a review. *Statistician* 46, 261–283
22 De Santis, F. (2004) Statistical evidence and sample size determination for Bayesian hypothesis testing. *J. Stat. Plan. Infer.* 124, 121–144
23 De Santis, F. (2007) Using historical data for Bayesian sample size determination. *J. R. Stat. Soc. Ser. A* 170, 95–113
24 Joseph, L. *et al.* (1995) Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician* 44, 143–154
25 Joseph, L. et al. Some comments on Bayesian sample size determination. Statistician 44, 167–171.
26 Wang, F. and Gelfand, A.E. (2002) A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Stat. Sci.* 17, 193–208
27 Weiss, R. (1997) Bayesian sample size calculations for hypothesis testing. *Statistician* 46, 185–191
28 Berger, J. (2006) The case for objective Bayesian analysis. *Bayes. Anal.* 1, 385–402
29 Liu, C.C. and Aitkin, M. (2008) Bayes factors: prior sensitivity and model generalizability. *J. Math. Psychol.* 52, 362–375
30 McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models* (2nd edn), Chapman and Hall/CRC.
31 Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *J. R. Stat. Soc. Ser. A* 135, 370–384

32 Carlin, B.P. and Louis, T.A. (2009) *Bayesian Methods for Data Analysis* (3rd edn), CRC Press.

33 Gelman, A. et al. (2004) *Bayesian Data Analysis* (2nd edn), CRC Press.

34 Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilievel/Hierarchical Models*, Cambridge University Press

35 Jackman, S. (2009) *Bayesian Analysis for the Social Sciences*, Wiley

36 Ntzoufras, I. (2009) *Bayesian Modeling Using WinBUGS*, Wiley

37 Lindley, D.V. (1975) The future of statistics: a Bayesian 21st century. *Adv. Appl. Prob.* 7, 106–115

38 Kruschke, J.K. (1996) Dimensional relevance shifts in category learning. *Connect. Sci.* 8, 201–223

39 George, D.N. and Pearce, J.M. (1999) Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *J. Exp. Psychol. Anim. Behav. Process.* 25, 363–373

40 Oswald, C.J.P. *et al.* (2001) Involvement of the entorhinal cortex in a process of attentional modulation: evidence from a novel variant of an IDS/EDS procedure. *Behav. Neurosci.* 115, 841–849

41 Lee, M.D. and Webb, M.R. (2005) Modeling individual differences in cognition. *Psychon. Bull. Rev.* 12, 605–621

42 Navarro, D.J. *et al.* (2006) Modeling individual differences using Dirichlet processes. *J. Math. Psychol.* 50, 101–122

43 Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayes. Anal.* 1, 515–533

44 R Development Core Team (2009) R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*.

45 Thomas, A. *et al.* (2006) Making BUGS open. *R News* 6, 12–17

46 Gilks, W.R. *et al.* (1994) A language and program for complex Bayesian modelling. *Statistician* 43, 169–177