Journal of Parapsychology, 78(2), 170-182, 2014

BAYESIAN AND CLASSICAL HYPOTHESIS TESTING: PRACTICAL DIFFERENCES FOR A CONTROVERSIAL AREA OF RESEARCH

By J. E. Kennedy

ABSTRACT: The use of Bayesian analysis and debates involving Bayesian analysis are increasing for controversial areas of research such as parapsychology. This paper conceptually describes the philosophical and modeling differences between Bayesian and classical analyses, and the practical implications of these differences. Widely accepted statistical conventions have not yet been established for Bayesian analysis in scientific research. The recommendations from the FDA guidance on using Bayesian methods are appropriate for confirmatory experiments. This guidance recommends that the study design and protocol include (a) specification of the prior probabilities and models that will be used, (b) specification of the criteria that will be considered acceptable evidence, (c) operating characteristics for the probability of Type I error and power of the analysis, and (d) an estimate of the relative roles of prior probability versus the data from the current experiment in producing the final results. Both classical and Bayesian methods are valid when properly applied with confirmatory methodology that includes prespecification of statistical methods, and prospective evaluations of inferential errors and power. Evaluations of inferential errors and power measure the validity of a planned hypothesis test, including Bayesian analysis. Unfortunately, the use of confirmatory methodology has been rare in psychology and parapsychology.

Keywords: Bayesian analysis, classical analysis, inferential errors, confirmatory research, subjective probability

The use of Bayesian analysis has been rapidly increasing in science and is becoming conspicuous in scientific controversies. For example, Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) argued that classical analyses supporting parapsychological effects are evidence that classical methods are faulty and should be replaced with Bayesian methods. Bem, Utts, and Johnson (2011) responded that certain aspects of this analysis were flawed, but agreed that Bayesian methods have advantages that will be increasingly utilized in scientific research. Debates like this typically focus on specialized technical points without presenting the fundamental assumptions and models that provide the crucial context for understanding and evaluating the arguments.

The present article is intended to describe conceptually the philosophical assumptions, models, and practical aspects that differ between Bayesian and classical hypothesis testing. This discussion should allow a person to conceptually understand the descriptions of methodology and the findings for experimental research that uses Bayesian analyses, and to follow debates about conflicting conclusions from research data. In addition, some potentially controversial claims and practices with Bayesian methods are described, as well as recommendations for methodology for confirmatory experiments. References are not provided for concepts that are commonly described in writings on Bayesian methods.

The discussion here focuses on evaluating the evidence for an ESP or psi experimental effect using a binomial analysis, as is common in parapsychology. Bayesian methods can also be used for other types of analyses. The basic principles discussed here also apply for other analyses.

When discussing current limitations, uncertainties, or debates about a statistical topic, I sometimes offer my opinion about the optimal strategy for handling the matter. Some of these opinions are prefaced with qualifiers such as “in my opinion” or “my perspective is.” These qualifiers are intended to indicate that a detailed technical discussion of the topic is beyond the purposes of the present article, and that others may have differing opinions.

Is Probability a Condition of the Physical World or a Condition of a Human Mind?

Bayesian and classical analyses are based on different philosophical perspectives about the nature of probability. Consider the case of a colleague who goes into a separate room and flips a coin. After the coin has been flipped, the colleague knows the outcome, but a person in the other room does not.

Objective Probability

One perspective is that after the coin has been flipped there is no uncertainty about the outcome. It is what it is. If the coin came up heads, the probability that it is heads is one, and the probability that it is tails is zero. The fact that a person in another room does not know the state of the coin is irrelevant. Probability in this case is objectively based on the state of the physical world. It is not an accurate representation to describe the state of the coin as being uncertain after the state has been physically determined.

Classical hypothesis testing is based on this philosophy of probability. A scientific hypothesis such as "do some people have psychic abilities" is a question about the existing state of the world. The world is what it is, and the fact that a particular person is uncertain about the truth of a hypothesis does not affect the existing state of the world. Variation and probability pertain to the outcomes of future experiments and observations, not to the properties of an existing state of the world. This perspective on probability is also called the *frequentist* interpretation because it assumes probability is based on the frequency of occurrence of an outcome when the random event or observation is repeated many times.

The logic for statistical analysis is to determine the probability for the outcome of an experiment given that a certain state of the world exists. The statistical models treat the parameters for the state of the world as constant and the outcome of an experiment as variable.

Subjective Probability

An alternative philosophical perspective is that probability is based on the beliefs in a human mind and therefore is subjective. The fact that the state of the coin has been determined does not resolve the uncertainty for a person who does not know the outcome. Uncertainty and probability exist for that person. The probability for a person in the room with the coin is completely different than for a person in another room.

Bayesian statistics are based on subjective probability and prescribe how a person's beliefs should change as new data are obtained. A mathematical model determines the optimal beliefs given the initial beliefs and the new data. This strategy assumes that the uncertainty in a person's mind can be quantitatively modeled and that a person's beliefs should rationally follow the mathematical laws of probability theory.

A person's initial beliefs and uncertainty are mathematically represented with *prior probability distributions*. These represent the person's beliefs prior to collecting data for the current study. Ideally, everything that a person believes about a topic is quantitatively incorporated into the prior probabilities. For example, any concerns about misconduct or biased methodology in previous studies must be incorporated quantitatively into the prior probability values.

After the data have been collected for the current study, the analysis combines or updates the prior probabilities with the evidence from the new data to produce the *posterior probability*. This mathematically represents what the person should rationally believe given the prior beliefs and the data from the current study.

In Bayesian models, parameters representing the existing state of the world are treated as variable and the observed outcome of an experiment is treated as constant. The variability in the parameters for the existing state of the world represents the uncertainty in a person's mind, not variation or fluctuations in the actual state of the world. Cases with variations in the state of the world, such as in a random effects analysis, are a different aspect of variability in the model.

Different Uses of Probability

Both philosophical perspectives on probability appear to me to be valid. They focus on different manifestations of probability. Objective probability attempts to directly model uncertainty in the physical world whereas

subjective probability attempts to directly model uncertainty in a human mind. Both approaches involve algorithms for drawing inferences about the world. Both quantify uncertainty using mathematical probability distributions of hypothetical possibilities for the value of terms in mathematical models. Both assume that the mathematical models can be verified and improved by making observations.

The key question is how useful the two approaches are in practice. I tend to favor one or the other, depending on the context for the use of probability.

In situations such as gambling games in casinos, the probabilities are precisely known. All possible outcomes of a series of random events can be fully enumerated. The probability that a certain outcome will occur on a series of trials is clear, and the concept of probability based on many repetitions seems natural. Many parapsychology experiments also have these well-defined properties.

The other extreme would be situations such as commodity markets and other investment decisions. In these cases, the probabilities are not precisely known and are not constant over time.

Another important factor is the type of question being asked. A question such as "do some people have psychic abilities" is about an existing state of the world. On the other hand, a question such as "should I invest my retirement savings in the commodities market based on the predictions of a psychic" is a personal decision about future actions more than a scientific question about the state of the world. These latter situations usually involve potential risks and rewards, and are more difficult to conceptualize in terms of repeated observations.

For me, when a question focuses on an existing state of the physical world and can be evaluated with repeated observations using clearly applicable probability models, the methods of objective probability are a natural fit. When a question focuses on a personal decision that involves risks and rewards or poorly defined probabilities, subjective probabilities are a natural fit. Note that the operation of a market is based on the assumption that people have different subjective probabilities about the outcomes of future events. If everyone had the same beliefs, commodity markets and stock markets would not be possible because there would be only buyers or only sellers.

Scientific Research

Scientific researchers have traditionally taken great pride in being objective. They have modeled the basic properties of the world as being independent of the human mind. The philosophy of objective probability emerged from and is consistent with that worldview.

Subjective probability brings the diversity of a market environment to scientific research, and complicates analyses by including models of the personal beliefs in a human mind as well as models of the external world. Advocates of Bayesian methods, of course, argue that pure objectivity does not occur and that subjective probability is more realistic of what actually happens in science. However, another perspective is that a prominent injection of subjectivity into scientific methods will unnecessarily further degrade the admittedly imperfect objectivity of science and hinder the development of consensus.

These debates have no clear resolution at present. Both approaches have assumptions about how human beliefs should ideally be influenced by evidence. From my perspective, the claims that one approach is better than the other need to be evaluated empirically—and that remains to be done. Most scientific research, and particularly experiments, can be reasonably evaluated with either approach.

A more pragmatic question is what are the differences between these two approaches in practice? It appears to me that both approaches are logically valid and should eventually reach the same conclusions for scientific hypotheses about an existing state of the world. Ease of use and efficiency in reaching those conclusions may differ. Of course, classical methods currently have advantages from much more widely available software and more extensive practical experience with the methods and software. In addition, classical methods have widely accepted conventions for statistical methodology and simpler mathematical methods.

Differences in Describing Results

Classical hypothesis tests evaluate an experiment by comparing the observed outcome to the distribution of other outcomes that could have occurred if the results were produced by chance fluctuations. If the probability

or *p value* of the observed outcome under this null hypothesis is less than a prespecified criterion, the outcome is interpreted as a significant result that provides evidence for the alternative or experimental hypothesis.

A Bayesian analysis typically compares the probability that the alternative or experimental hypothesis is true with the probability that the null hypothesis is true, given the prior probabilities and the experimental data. The null hypothesis is that only chance is operating. The comparison is made by forming the ratio of the two probabilities. This ratio is the *odds* that the alternative hypothesis is true. Larger values of the odds are favorable for the alternative hypothesis.

For an estimate of the effect size with 95% confidence, a classical frequentist analysis describes the *confidence interval* as having a .95 probability that the range contains the true value. A Bayesian analysis describes the *credible interval* as having a .95 probability that the true value lies within this range. Obviously, few statistical users will consider the theoretical distinction between these descriptions to be important in practice.

Differences in Calculating and Interpreting Probability

Classical methods are based on hypothetical repetitions of the experiment independent of a person's beliefs, whereas Bayesian methods are contingent on certain prior beliefs and give different results for different prior beliefs. The results of a Bayesian analysis cannot be assumed to apply to persons who have different prior beliefs. Classical methods have the assumption that objective scientific evidence can be developed without regard for a person's prior beliefs.

The difference between Bayesian and classical perspectives can be seen easily when evaluating data with the binomial probability model. The binomial model has two key parameters: *P*, which is mean chance expectation for a particular outcome occurring in an event or trial; and *X*, the number of times that the outcome actually occurs in a group of trials. The probability model for a classical hypothesis test assumes that *P* is constant and *X* is a random variable. Data analysis is based on a probability distribution for *X* that represents all possible outcomes that could have occurred. Bayesian analysis switches this and takes *X* as constant and *P* as a random variable. Data analysis is based on a probability distribution for *P* that represents the beliefs and uncertainty in a person's mind about the true value of *P*.

Classical Hypothesis Testing

The usual convention is that the null hypothesis is rejected and the alternative hypothesis is accepted if the *p* value for the experiment is .05 or less. The *p* value is the probability of obtaining an experimental outcome as extreme or more extreme than the observed outcome if the null hypothesis is true. An incorrect conclusion or inferential error can occur for a hypothesis test. Accepting the alternative hypothesis when the null hypothesis is true is a Type I error. When the null hypothesis is true and the usual convention is applied, 5% of experiments can be expected to have a Type I error.

The technically justifiable conclusion for a classical hypothesis test is the simple binary outcome of whether or not the null hypothesis is rejected. Note that the *p* value is used to infer whether the null hypothesis is true, but it is not a direct measure of the probability that the null hypothesis is true. The philosophical assumptions and mathematical derivations for classical analyses provide the probability of obtaining the experimental outcome given that the null hypothesis is true. This is conceptually different than the probability that the null hypothesis is true given the experimental outcome—which is evaluated with Bayesian analysis. Greater confidence occurs for outcomes with smaller *p* values; however, this greater confidence is basically qualitative.

For a hypothetical example, consider an experiment attempting to detect a PK influence on an electronic random event generator (REG). If the outcome is 5,100 hits in 10,000 trials where the probability of a hit is .5 by chance, the *p* value for this outcome under the null hypothesis is $p = .046$ two-sided. This result is less than .05 and would be considered significant evidence for a PK effect.

The statistical power of an experiment is an important limiting factor that is often overlooked when interpreting the results. An experiment with a small sample size can fail to support the alternative hypothesis when that hypothesis is actually true. This is a Type II inferential error. The statistical power is the expected proportion

of experimental outcomes that will support the alternative hypothesis when that hypothesis is true. Power is usually determined based on estimated effect sizes from previous studies. The usual recommendation for experimental design is that the power should be at least .80, and preferably higher. Unfortunately, many experiments in parapsychology and psychology have been designed without regard for power and have had much lower power (Kennedy, 2013a).

When experiments with low power produce nonsignificant outcomes, the interpretation is ambiguous. The results could be due to the alternative hypothesis being false or due to the small sample size. Most experiments with low power can be expected to produce nonsignificant outcomes and to contribute little to scientific knowledge beyond providing estimates for designing more powerful studies.

Power analysis evaluates the statistical validity of a hypothesis test. Although a classical hypothesis test does not provide a direct probability that a hypothesis is true, power analysis provides probabilities that correct inferences will be made about hypotheses, and it can justify confidence in the experimental conclusions.

Bayesian Hypothesis Testing

A typical Bayesian analysis for a simple experiment uses methods for comparing which of two probability models is correct. One model is for the alternative hypothesis that an experimental effect is occurring and the other model is for the null hypothesis that only chance is operating. The posterior probability is calculated for each model. As usual, the posterior probability is calculated by updating the prior probability with the evidence from the data in the current study. The ratio of the posterior probabilities for the two models is the odds that the alternative hypothesis is true given the prior probabilities and the experimental data. These odds are a convenient way to compare two probabilities. The odds provide a direct measure of the probability that the alternative hypothesis is true, rather than an indirect inference as with classical analysis.

Widely accepted conventions have not been established for the magnitude of odds that is considered adequate evidence. Discussions of this topic usually reference Jeffreys (1961, p. 432), which (with minor rounding) describes odds of 3 to 10 as "substantial," 10 to 32 as "strong," 32 to 100 as "very strong" and greater than 100 as "decisive." Odds of 1 to 3 are "not worth more than a bare mention."

Odds of less than 1 can be inverted to provide the odds that the null hypothesis is true. This ability to provide direct quantitative evidence supporting the null hypothesis is an important feature of Bayesian analysis.

Jeffreys (1961, p. 435) said that he used an odds of 3 the way classical analysts use $p = .05$, and an odds of 10 the way classical analysts use $p = .01$. He also noted that inferential errors will sometimes occur with these criteria.

A Bayesian analysis of an experiment can have three possible outcomes: the final odds can (a) exceed the criterion supporting the alternative hypothesis, (b) exceed the criterion supporting the null model, or (c) fall into the intermediate zone that does not convincingly support either model. An experiment with a small sample size will likely have the latter result.

Odds near one indicate that the sample size is not adequate to evaluate whether the null hypothesis or alternative hypothesis is true. Probabilities for Type I error and power are based on classical ideas about hypothetical repetitions of an experiment. Inoue, Berry, and Parmigiani (2005) note that Bayesian methods for determining sample size have not become standardized and widely used. They also note that many Bayesian analysts use the concepts of Type I error and power in practice, and that this mixing of approaches is useful. Kruschke (2011) provides a useful discussion of power in the context of Bayesian analysis.

Prior probability for the alternative (psi) hypothesis. The starting point for a Bayesian hypothesis test is the prior probability that the hypothesis of interest is true. For a properly conducted experiment, this prior probability will be specified at the design stage prior to collecting data for the experiment. This prior probability is typically expressed as prior odds, which are the prior probability that the alternative or psi hypothesis is true divided by the prior probability that the null hypothesis is true. These prior odds will be adjusted or updated based on the evidence from the data in the current experiment to produce the final (posterior) odds that the alternative hypothesis is true.

In theory, the optimal strategy is to set the prior probabilities based on information from previous research. These are called *informative* priors. In an ideal world, the posterior probabilities from the first study would become the prior probabilities for the second study, and this would continue with each subsequent study. This strategy is optimal if the prior information is accurate. However, this strategy can also propagate bias if the previous studies have methodological errors, misconduct, selective reporting, or other biases—as sometimes occur in the real world.

If methodological bias is suspected in previous studies, the posterior probabilities from those studies need to be adjusted for the possible bias when developing the prior probabilities for the next study. This situation demonstrates the general principle that posterior probabilities may need to be modified based on subjective opinions about factors that are outside the mathematical models used for data analysis.

Informative priors are problematic for a controversial area like parapsychology because subjective beliefs about previous research vary dramatically. The points of dispute typically focus on the methodology in previous research. Skeptic David Marks (2000, pp. 306-307) states that when he began investigating the claims for remote viewing and ganzfeld experiments, his subjective prior probability that these effects can occur was one-tenth. After delving into the methodology and findings for these experiments, his subjective probability for ganzfeld ESP ability was one-millionth and for remote viewing was one-billionth. Wagenmakers et. al. (2011) argued the prior probability for the psi hypothesis should be very close to zero and gave 10^{-20} as an example. Proponents of psi who find the existing methodology and evidence compelling have equally strong subjective beliefs. For example, Radin (2006, p. 276) presented the "odds against chance" for ganzfeld research to be $3x10^{19}$ to 1 and the overall evidence for psi as $1.3x10^{104}$ to 1. Although Radin appears to have mistakenly interpreted classical *p* values as the probability that the null hypothesis is true, these odds can be accepted as his personal beliefs. These strong subjective beliefs, pro or con, will dominate the evidence from almost any reasonable experiment.

An alternative strategy is to use a *noninformative* prior probability that biases neither for nor against the alternative hypothesis. One obvious option is to assign equal prior probabilities to the alternative hypothesis and the null hypothesis. With this strategy, the prior odds are 1 and the experimental conclusion will be based on the evidence from the data in the current experiment. Noninformative priors emulate the classical assumption that scientific evidence from a study can and should be developed without regard for a person's prior beliefs.

Bayes factor. The Bayes factor is a measure of the evidence from the data in the current experiment. It is the ratio for the probability of obtaining the experimental outcome under the alternative hypothesis divided by the probability of obtaining the experimental outcome under the null hypothesis. For a parapsychological experiment, this ratio is the odds that the experimental outcome is due to psi rather than to chance fluctuations.

The final (posterior) odds that the alternative hypothesis is true are derived by multiplying the prior odds that the alternative hypothesis is true by the Bayes factor. If a noninformative prior is used, the prior odds are 1 and the Bayes factor gives the final odds for the analysis. The Bayes factor is sometimes discussed as a likelihood ratio, or more precisely, the ratio of *marginal likelihoods*.

Although the Bayes factor is not influenced by the prior probability that the alternative hypothesis is true, the calculation of the Bayes factor incorporates a different prior probability in a very fundamental way—as discussed in the next section.

Prior probability distribution for effect size. Calculation of the Bayes factor requires that the probability for the experimental outcome be estimated under the alternative hypothesis. This calculation requires assumptions about the effect size for the alternative hypothesis (*P* in the binomial model). The assumptions about effect size for this calculation are derived from a prior probability distribution that represents the beliefs and uncertainty in a person's mind about the magnitude of the effect. For properly conducted research, this prior probability distribution for effect size should be specified before data collection begins. The strategy for selecting a prior probability for effect size can be informative or noninformative.

Not surprisingly, selecting a prior probability distribution for effect size based on previous research is subject to widely differing opinions for an area such as parapsychology. Different prior probability distributions for psi effect size were a major factor in the debates between Wagenmakers et. al. (2011) and Bem et al. (2011), and between Jefferys (1990) and Dobyns (1992). The arguments may have no clear resolution given the subjective nature of probability in Bayesian analysis.

Unfortunately, attempts to find unbiased, noninformative prior probability distributions for effect size have also been problematic. Biases in favor of either the experimental or null model appear to be virtually inevitable—and can be counterintuitive. For example, a diffuse prior probability distribution that allows a wide range for effect size appears on the surface to represent a very open-minded prior belief. However, a diffuse prior makes the Bayes factor favor the null hypothesis.

For example, a uniform distribution is frequently recommended as a noninformative prior distribution for binomial effect sizes. The uniform prior gives equal probability to any effect size (*P* in the binomial model) between

0 and 1. Using the online binomial Bayes factor calculator provided by Rouder (2012), 5,100 hits in 10,000 trials with a uniform prior of beta(1,1) for the calculator gives a two-sided Bayes factor of 10.8 supporting the null hypothesis. This is considered strong evidence in favor of the null hypothesis. As noted above, the classical binomial analysis gives $p < .05$ supporting the alternative or psi hypothesis.

The fact that Bayesian analyses of small effect sizes tend to support the null hypothesis when classical analyses support the alternative hypothesis is well known among statisticians. Some proponents of Bayesian analysis argue that this shows that classical analyses are flawed (e.g., Jefferys, 1990). However, applying Bayesian analyses to simulated data indicates that these discrepancies can reflect low power and inferential errors in Bayesian hypothesis testing, particularly with diffuse prior probabilities (Kennedy, in press).

Much work remains to be done to understand the consequences of different prior probability distributions for effect size and to develop unbiased methods. Debates about appropriate prior probabilities for effect size can be expected given the current limited understanding and high potential for unrecognized bias. Most books on Bayesian analyses recommend model evaluations and sensitivity analyses that evaluate how different prior probabilities affect the final conclusions from an analysis.

Objective Bayesian analysis. Methods for objective Bayesian analysis attempt to minimize the subjectivity and potential biases from prior probabilities. One common practice is to base conclusions on the Bayes factor. However, the prior probability distribution for effect size is a fundamental part of the Bayes factor and currently can be expected to be a source of potential bias and controversy.

Like classical analyses, objective Bayesian methods strive to identify what a person should conclude given the experimental data and without biases from prior beliefs. As the use of Bayesian analysis for scientific research matures, conventions for objective analysis will likely become established. At present, widely accepted statistical conventions have not yet been established for generating objective scientific evidence from the intrinsically subjective nature of Bayesian analysis.

Cumulative conclusions for each experiment. An ideal Bayesian analysis with informative priors analyzes an experiment as a research synthesis that incorporates the findings from previous studies. The prior probabilities incorporate all previous information and are combined with the data from the current study to provide a cumulative conclusion. A separate research synthesis would be redundant.

That is a very different strategy than the classical concept of independent replications that are evaluated in a separate research synthesis. Independent replications have an important role in establishing the validity of a scientific finding because they counteract potential biases such as methodological errors, misconduct, and selective reporting. However, independent replication can be compromised by informative priors in Bayesian analyses. With informative priors, the results of previous studies are directly incorporated into the statistical analysis of an experiment, which makes the experiment not independent of the previous studies.

Objective Bayesian methods minimize the dependence between studies and enhance the validity of scientific findings. Both Bayesian and classical approaches assume that replications tend to be unbiased and will eventually counteract methodological biases. This assumption is discussed in a later section.

Other Differences

Bayesian methods require more information and more complex models than classical methods. Proponents argue that this is beneficial because it makes the subjective aspects of an analysis explicit and conspicuous. An alternative view is that the complexity and subjectivity significantly reduce the efficiency of reaching conclusions about scientific hypotheses. My opinion is that these speculations need to be supported with quantitative evidence. An obvious first step is to compare the statistical power or the amount of data needed to establish evidence for an effect using classical analyses versus the more complex Bayesian models (e.g., Kennedy, in press). Another important comparison is the amount of data needed to raise concerns about and counteract a case of scientific misconduct or methodological bias.

Adjustments for sequential analysis, optional stopping, and multiple analyses are very important in classical hypothesis testing. Many Bayesian analysts believe that these need no adjustments with Bayesian hypothesis tests. However, there are differing opinions on this (e.g., Berry & Hochberg, 1999). Here too, these discussions have been primarily based on theory with little practical empirical evaluation of inferential error rates.

In the 1990s, computer approximation methods were developed for Bayesian analyses. The writings prior to that time often discussed the great computational difficulty in working with Bayesian methods. Those discussions are no longer applicable. In fact, the general consensus now is that complex models such as hierarchical or multi-level random effects are often easier to work with using Bayesian methods. However, the methods use approximations, which raise concerns about identifying cases for which the approximations do not work well.

Advocates of Bayesian methods frequently criticize the use of *p* values. Without getting into the myriad of technical details, my perspective is that classical hypothesis testing is useful and reliable for confirmatory experiments with prospectively set power and probability of a Type I error. I have not seen criticisms that invalidate the acceptance/rejection of hypotheses in this situation. A nonsignificant result in a well-powered experiment is evidence supporting the null hypothesis.

Both Bayesian and classical methods have certain assumptions that must be met for the results to be valid. Most classical analyses assume that the error terms (and thus the observations within a group or treatment) are independent and identically distributed. Bayesian methods assume the observations in a group or treatment are *exchangeable*, which means the observations are from the same distribution and any permutation of the observations is equally likely. Unrecognized confounding factors compromise both Bayesian and classical assumptions. In general, the applicability of these assumptions is most clear for studies that properly utilize randomness.

Bayesian Analyses for Confirmatory Experiments

The prior probabilities with Bayesian analyses apparently can alter the conclusions from almost any reasonable experiment. The influence can be in either direction—from conservative analyses to exaggerated effects. My impression from the past debates in parapsychology is that skeptics can relatively easily find and justify prior probabilities that produce unfavorable results, and proponents can as easily find and justify prior probabilities that are favorable. Dobyns (1992) shows this quantitatively for one example of psi data.

Those presenting conservative Bayesian analyses of psi experiments have generally assumed that the Bayesian results were correct and that they demonstrate that classical statistical methods are flawed (Berger & Delampady, 1987; Jefferys, 1990; Wagenmakers et. al., 2011). The possibility that the selected Bayesian methods had low power or were biased was not seriously considered in these writings. However, initial investigations with simulated data suggest that these discrepancies may reflect low power and inferential errors for the Bayesian methods (Kennedy, in press).

Inferential error rates and power are useful measures for evaluating statistical methods and are important factors in designing confirmatory research. A statistical hypothesis test requires several methodological decisions that affect how the test performs. These decisions include the selection of the prior probabilities, specific statistical models, acceptance criteria, and sample size. Evaluations of inferential errors and power indicate how well these factors work together to provide an effective decision-making process. These evaluations indicate how reliably the Bayesian models of the human mind correctly detect conditions in the external world.

Bayesian analysts often argue that prespecifying acceptance criteria and sample size are not needed with Bayesian analyses. That argument implicitly assumes that inferential errors can be ignored with Bayesian analysis. That may be applicable for exploratory research; however, when the purpose of an experiment is confirmation of a controversial effect, more formal methods and quantitative evaluation of expected error rates are needed.

In theory, Bayesian hypothesis tests start with prespecified prior probabilities, which are then updated from the data in the current experiment to produce posterior probabilities and conclusions. In practice, the current state of Bayesian analysis has potential for researchers to start with the experimental data and then to adapt the prior probabilities to produce the desired conclusions. These adaptations can be applied in the context of the sensitivity analyses and model evaluations that are needed given the lack of established conventions and limited understanding of the implications of different prior probabilities. The potential for maneuvering to produce the desired results is enhanced by any ambiguity about the criteria that are considered acceptable evidence.

One important distinction between confirmatory and exploratory research is that decisions about analysis methodology should be made prospectively for confirmatory research. Methodological decisions for exploratory research can be made during data analysis.

FDA Recommendations

The U.S. Food and Drug Administration (FDA, 2010) developed guidance for using Bayesian statistics that offers a pragmatic and balanced perspective. The great majority of writings on Bayesian analyses focus on the exploratory stage of research. The FDA guidance is a rare exception that discusses methodology appropriate for confirmatory research. The guidance is for using Bayesian methods when seeking approval of medical devices. It was developed after public review and comment, and it would be appropriate in any situation in which the experimental results will be challenged, including parapsychology.

The FDA guidance takes the position that Bayesian analyses can be advantageous in some situations, but supporting analyses are usually needed to provide convincing evidence. The guidance recommends that the study design and protocol include (a) specification of the prior probabilities and models that will be used, (b) specification of the criteria that will be considered acceptable evidence, (c) operating characteristics for Type I errors and power of the analysis, (d) an estimate of the relative roles of prior probability versus the data from the current experiment in producing the final results, and (e) sensitivity analyses that determine how different prior probabilities and models affect these evaluations. These analyses require a substantial amount of effort that will typically be done prospectively with simulations. This information is also appropriate for Bayesian reanalyses that challenge findings from previous experiments.

Note that the FDA guidance does not indicate that a fixed sample size be specified. The guidance does not discourage Bayesian sequential analyses, but it does recommend that the decision criteria be specified in advance and that simulations be run to evaluate the effective probability of a Type I error and the power of the test.

In addition, the FDA recommendations are an effective way to address the uncertainties, differing opinions, and potential controversies with Bayesian analyses. Possible biases for small effects and diffuse prior probabilities need to be evaluated. As another example, two-sided tests in Bayesian analyses tend to be sensitive to prior probabilities and to favor either the alternative or null hypothesis. The evaluations recommended by FDA should show the net practical effects of any potential analysis issues that may or may not be recognized by the investigators.

My impression is that the theoretical developments in Bayesian analysis are significantly ahead of the practical experience in using the methods. I expect that significant pitfalls and reality checks will be found as more experience is gained with these methods. For example, with classical regression analyses, considerable practical experience was required before analysts began to usefully understand the implications of correlations among predictor variables (Hocking, 1983). At present, the writings on Bayesian analyses focus on theory with few simulation studies or empirical investigations of the applicability of the theory in practice. The supportive analyses recommended by the FDA guidance provide valuable assurances that the main analyses actually perform as advertised.

FDA takes the position that both Bayesian and classical approaches are valid if properly applied. The FDA guidance notes that "[w]hile the Bayesian approach can often be favorable to the investigator with good prior information, the approach can also be more conservative" (p. 10). The guidance also notes that "[t]he flexibility of Bayesian models and the complexity of the computational techniques for Bayesian analyses create greater possibility for errors and misunderstandings" (p. 10).

I would add that the flexibility, complexity, and subjectivity of Bayesian analyses make potential biases very difficult for a typical scientist to understand. Reliance on retrospective sensitivity analyses and model evaluations can introduce significant potential for biases. Without the supplemental analyses at the planning stage, an experiment cannot be expected to provide convincing evidence if challenged. Of course, more casual methods can be used for exploratory research.

Exploration, Confirmation, Application

Both classical and Bayesian analyses assume that experimental research is self-correcting and will eventually produce valid, compelling conclusions. Biased results from an experiment will be rectified by other experiments that are unbiased.

This idealistic philosophical hope does not consider the evidence that sustained levels of methodological noise and bias can occur in academic research—particularly in psychology and parapsychology (Ioannidis, 2012;

Kennedy, 2013a, 2014; Pashler & Wagenmakers, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievet, 2012). The typical research practices in recent years in psychology and parapsychology have been exploratory and provide many opportunities for biases and misconduct. The amount of methodological bias and misconduct that actually occurs cannot be reasonably estimated. Undetected cases are likely. These exploratory research practices cannot be expected to resolve a controversy such as the occurrence of psi—particularly when the findings consistently differ among experimenters (Kennedy, 2014).

Confirmation

Well-designed confirmatory experiments are required to make experimental research self-correcting and to provide convincing, valid conclusions (Kennedy, 2013a, 2014; Pashler & Wagenmakers, 2012, Wagenmakers et al., 2012). Confirmatory methodology is well established for regulated medical research, but it has not been part of the research culture for psychology and parapsychology. Key components of confirmatory methodology include prespecification of statistical methods and acceptance criteria, sample size based on power analysis, public prospective registration of experiments, experimental procedures that make intentional or unintentional data alterations by one person difficult, documented formal validation of software, and sharing data for analyses by others (Kennedy, 2013a, 2013b, 2014; KPU Registry, 2014). Meta-analysis of exploratory studies does not eliminate the need for well-designed confirmatory research (Kennedy, 2013a).

With the advent of the KPU study registry (KPU Registry, 2012), confirmatory research methodology can be implemented for parapsychological research. Preregistration of statistical methods, including prior probabilities for Bayesian analyses, is fundamental for convincing confirmatory research. The standard for study registration in medical research and with the KPU registry is *public, prospective* registration, with requirements for the content of the registration (International Committee of Medical Journal Editors, 2005; KPU Registry, 2012). For comparison, study registration as currently offered by Open Science Framework (https://osf.io/) is basically self-registration that allows an experimenter to determine the content of the registration and to wait until the experimental results have been obtained before deciding whether to make the study and registration known to others. These options undermine much of the methodological value of study registration and allow an experimenter to withhold the experiment and/or registration if the results do not come out as hoped. However, if the results are favorable, the experimenter can misleadingly (and retrospectively) present the study as preregistered.

Skeptics have suggested that psi experiments are actually a control group that provides empirical evidence for the magnitude of methodological bias that occurs with current exploratory practices. That hypothesis is applicable to most experimental research in psychology as well, and remains plausible until confirmatory methodology is implemented—or until there is convincing evidence that does not rely on experiments.

Application

The most convincing evidence for psi would come from the development of practical applications. Psi clearly has great potential for practical applications that would substantially and conspicuously alter daily life. Sufficient effort and money have been devoted to developing applications of psi that tangible progress would be expected if the psi effects were even minimally reliable (Kennedy, 2003). Although there have been a few sporadic episodes of impressive results, the prevailing experience has been that efforts to develop applications were dropped when the results did not sustain the interest of those supporting and/or conducting the research. From this perspective, the absence of useful applications is evidence for the absence of convincing psi effects.

Given the lack of well-designed confirmatory research, my impression is that many scientists implicitly look to application rather than experiments when evaluating the evidence for psi. However, the rationale for that strategy is rarely discussed openly because it raises doubts about the integrity of scientific research far beyond parapsychology.

Conclusions

Both classical and Bayesian hypothesis tests are valid when properly applied. Well-designed confirmatory experiments with prespecified analyses and prospective evaluations of inferential errors and power are essential

for convincing evidence with either approach. However, the research culture for psychology and parapsychology has focused on exploration with very little attention to convincing confirmation. This situation undermines both classical analyses and Bayesian analyses. Unless confirmatory methodology is more widely implemented, strong evidence for psi will need to be based on the development of practical applications rather than on experimental research.

Evaluations of inferential errors and power measure the validity of a planned hypothesis test and are needed for both Bayesian and classical analyses. Classical analyses focus on modeling the uncertainty in the physical world—which tends to keep the models grounded in reality. Bayesian analyses focus on modeling the uncertainty in a human mind—which adds a layer of abstraction that provides greater flexibility and higher potential for the models to become unrealistic. Evaluations of inferential errors and power verify that all the components of a hypothesis test combine to provide useful inferences about reality. In addition, an estimate of the relative roles of prior probability versus the data from the current experiment is needed when planning confirmatory Bayesian analysis. Given the current lack of widely accepted conventions for Bayesian methods, prospective sensitivity analyses that evaluate different prior probabilities are also useful.

For most experiments, classical hypothesis tests require significantly less effort at present. One useful option would be to use classical methods as the primary analysis, plus a Bayesian analysis to explore and compare the methods. The effort required for Bayesian analyses will likely be reduced as applicable software evolves and the implications of different strategies for determining prior probability become more clearly understood through experience.

With classical analyses, 80% or more of properly designed confirmatory experiments are expected to provide independent evidence for an effect, with the great majority of experimenters obtaining successful replications. Corresponding expectations will likely be developed for confirmatory experiments with Bayesian analyses, including objective methods that minimize the propagation of bias from mistakes and misconduct in previous research. Of course, practical applications can be expected with this degree of reliability.

References

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*, 716-719. Retrieved from http://dl.dropboxusercontent.com/u/8290411/ResponsetoWagenmakers.pdf

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317-352. Retrieved from http://www.stat.duke.edu/~berger/papers/p-values.pdf

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*, 215-227. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378375899000440

Dobyns, Y. H. (1992). On the Bayesian analysis of REG data. *Journal of Scientific Exploration*, *6*, 23-45. Retrieved from http://www.scientificexploration.org/journal/jse_06_1_dobyns.pdf

Food and Drug Administration (2010). *Guidance on the use of Bayesian statistics in medical device clinical trials*. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

Hocking, R .R. (1983). Developments in linear regression methodology: 1959-1982. *Technometrics*, *25*, 219-230.

Inoue, L. Y. T., Berry, D. A., & Parmigiani, G. (2005). The relationship between Bayesian and frequentist sample size determination. *American Statistician*, *59*, 79-87. Retrieved from http://www.ime.unicamp.br/~lramos/mi667/ref/06berry05.pdf

International Committee of Medical Journal Editors (2005). *Clinical trial registration*. Retrieved from http://icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

Ioannidis, J. P. A. (2012).Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645-654. Retrieved from http://pps.sagepub.com/content/7/6/645.full

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration*, *4*, 153-169. Retrieved from http://www.scientificexploration.org/journal/jse_04_2_jefferys.pdf

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Clarendon Press.

Kennedy, J. E. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *Journal of Parapsychology*, *67*, 53-74. Retrieved from http://jeksite.org/psi/jp03.pdf

Kennedy, J. E. (2013a). Can parapsychology move beyond the controversies of retrospective meta-analysis? *Journal of Para-*

psychology, *77*, 21-35. Retrieved from http://jeksite.org/psi/jp13a.pdf
Kennedy, J. E. (2013b). [Letter to the editor]. *Journal of Parapsychology*, *77*, 304-306. Retrieved from http://jeksite.org/psi/method_letter.pdf
Kennedy, J. E. (2014). *Experimenter misconduct in parapsychology: Analysis manipulation and fraud*. Retrieved from http://jeksite.org/psi/misconduct.pdf.
Kennedy, J. E. (in press). Beware of inferential errors and low power with Bayesian analyses: Power analysis is needed for confirmatory research. *Journal of Parapsychology*. Retrieved from http://jeksite.org/psi/confirm_bayes.pdf
KPU Registry (2012). Koestler Parapsychology Unit Registry for Parapsychological Experiments. Retrieved from http://www.koestler-parapsychology.psy.ed.ac.uk/TrialRegistry.html
KPU Registry (2014). Exploratory and confirmatory analyses. Retrieved from http://www.koestler-parapsychology.psy.ed.ac.uk/Documents/explore_confirm.pdf
Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.
Marks, D. (2000). *The psychology of the psychic* (2nd ed.). Amherst, NY: Prometheus Books.
Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528-530. Retrieved from http://pps.sagepub.com/content/7/6.toc
Radin, D. (2006). *Entangled minds*. New York: Paraview.
Rouder, J. (2012). Bayes factor for a binomially distributed observation. Online calculator retrieved from http://pcl.missouri.edu/bf-binomial
Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426-432. Retrieved from http://www.ejwagenmakers.com/2011/WagenmakersEtAl2011_JPSP.pdf
Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. L. J, & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632-638. Retrieved from http://pps.sagepub.com/content/7/6/632.full.pdf+html

Broomfield, CO, USA
jek@jeksite.org

Abstracts in Other Languages

Spanish

PRUEBAS DE HIPÓTESIS BAYESIANA Y CLÁSICA:
DIFERENCIAS PRÁCTICAS PARA UN ÁREA DE INVESTIGACIÓN POL´EMICA

RESUMEN: El uso del análisis Bayesiano y los debates relacionados con tal análisis están aumentando en áreas controvertidas de investigación como la parapsicología. Este artículo describe conceptualmente las diferencias filosóficas y de modelamiento entre los análisis Bayesiano y clásico, y las implicaciones prácticas de estas diferencias. No se han establecido convenciones estadísticas ampliamente aceptadas aún para el análisis Bayesiano en la investigación científica. Las recomendaciones de la FDA sobre el uso de métodos Bayesianos son apropiadas para los experimentos confirmatorios. Esta guía recomienda que el diseño del estudio y protocolo incluyan: (a) especificación de las probabilidades previas y modelos que se van a utilizar, (b) especificación de los criterios que serán considerados como evidencia aceptable, (c) características de funcionamiento para la probabilidad de error Tipo I error y poder del análisis, y (d) una estimación de la importancia relativa de la probabilidad a priori frente a los datos del experimento actual en la producción de los resultados finales. Tanto los métodos clásicos como los Bayesianos son válidos cuando se aplican correctamente en la metodología de confirmación que incluye especificación previa de métodos estadísticos y evaluaciones prospectivas de los errores inferenciales y el poder. Las evaluaciones de los errores inferenciales y el poder miden la validez de una prueba de hipótesis planeada, incluyendo el análisis bayesiano. Desafortunadamente, el uso de la metodología de confirmación ha sido poco frecuente en la psicología y la parapsicología.

French

LE TEST D'HYPOTHESES BAYESIEN ET CLASSIQUE :
DIFFERENCES PRATIQUES DANS UN CHAMP DE RECHERCHE CONTROVERSE

RESUME : L'utilisation de l'analyse bayésienne et les débats impliquant l'analyse bayésienne prennent une place de plus en plus importante dans des champs controversés tels que celui de la parapsychologie. Cet article décrit conceptuellement les différences philosophiques et de modèle entre les analyses bayésiennes et classiques, ainsi que les implications pratiques de ces différences. Des conventions statistiques largement acceptées n'ont pas encore été établies pour l'analyse bayésienne dans la recherche scientifique. Les recommandations de la FDA pour l'emploi des méthodes bayésiennes sont appropriées pour des expériences confirmatoires. Il est conseillé que la conception de l'étude et du protocole incluent (a) une spécification des probabilités a priori et des modèles qui seront utilisés, (b) une spécification des critères qui détermineront ce qui serait une preuve acceptable, (c) les caractéristiques opérantes pour la probabilité de l'erreur de Type 1 et la puissance de l'analyse, et (d) une estimation des rôles rela- tifs des probabilités a priori versus des données de l'expérience en question dans la production des résultats finaux. Tant les méthodes classiques que les méthodes bayésiennes sont valides lorsqu'elles sont correctement appliquées dans une méthodologie confirmatoire qui inclue la préspécification des méthodes statistiques et des évaluations prospectives des erreurs inférentielles et de la puissance. Les évaluations des erreurs inférentielles et de la puissance mesurent la validité d'un test d'hypothèse planifié, y compris pour l'analyse bayésienne. Malheureusement, l'em- ploi d'une méthodologie confirmatoire est rare tant en psychologie qu'en parapsychologie.

German

BAYESSCHE UND KLASSISCHE HYPOTHESENPRÜFUNG:
PRAKTISCHE UNTERSCHIEDE FÜR EIN KONTROVERSES FORSCHUNGSGEBIET

ZUSAMMENFASSUNG: Die Verwendung der Bayesschen Analyse und Diskussionen unter Einschluss der Bayesschen Analyse haben für kontroverse Forschungsgebiete wie die Parapsychologie an Bedeutung gewonnen. Dieser Artikel beschreibt konzeptuell die philosophischen und Modellierungsdifferenzen zwischen Bayesschen und klassischen Analysen und die praktischen Implikationen dieser Unterschiede. Bisher konnten sich statistische Konventionen in Bezug auf die Bayessche Analyse in der wissenschaftlichen Forschung noch nicht allgemein durchsetzen. Die Empfehlungen nach Vorgabe der FDA über die Verwendung der Bayesschen Methoden sind für Bestätigungsexperimente angemessen. Diese Vorgabe empfiehlt, dass die Studienplanung und das Protokoll folgendes einschließen: (a) die Spezifizierung der a priori-Wahrscheinlichkeiten und der verwendeten Modelle, (b) die Spezifizierung der Kriterien, die als akzeptable Evidenz eingestuft werden, (c) die wirksamen Kennzeichen für die Wahrscheinlichkeit des Fehlers 1. Art und der Analyse der Teststärke und (d) eine Abschätzung über die relative Rolle der a priori-Wahrscheinlichkeit versus der Daten im vorliegenden Experiment beim Zustandekommen der endgültigen Ergebnisse. Sowohl die klassische wie auch die Bayessche Methode sind zulässig, wenn sie korrekt bei der Bestätigungsmethodologie angewendet werden, was die vorherige Spezifizierung der statistischen Methoden und prospektive Einschätzungen der Fehler beim Schlussfolgern und der Teststärke einschließt. Einschätzungen der Fehler beim Schlussfolgern misst die Validität einer geplanten Hypothesenprüfung, einschließlich der Bayesschen Analyse. Bedauerlicherweise wird die Bestätigungsmethodologie in Psychologie und Parapsychologie selten angewandt.

Other Methodology Articles