Journal of Parapsychology, 79(1), 53–64

BEWARE OF INFERENTIAL ERRORS AND LOW POWER WITH BAYESIAN ANALYSES: POWER ANALYSIS IS NEEDED FOR CONFIRMATORY RESEARCH

By J. E. Kennedy

ABSTRACT. Errors in inference can occur with any hypothesis testing method, including Bayesian analysis. The evaluation of expected rates of inferential errors is important when planning confirmatory research, but inferential errors have rarely been addressed in writings on Bayesian hypothesis testing. The present investigation applied classical and Bayesian hypothesis testing methods to binomial data with certain effects and to data simulating the null hypothesis. The Bayesian analyses generally had substantially lower power (probability of correctly detecting an effect), particularly for small effect sizes. For data with a small effect size and power of .80 for a classical analysis, the probability that the Bayes factor with a uniform prior correctly reached 3 or higher supporting the alternative model was only .173. The probability that the Bayes factor was 3 or higher incorrectly supporting the null model was .619. These findings verify that quantitative evaluation of expected inferential error rates is essential when designing confirmatory studies that use Bayesian analyses. The argument that biases in favor of the null model are appropriate for small effect sizes because of potential methodological problems is based on exploratory research and is not appropriate for well-designed confirmatory research that focuses on a pre-established effect size.

Keywords: Bayesian analysis, hypothesis test, inferential errors, statistical power, confirmatory research

Confirmatory research is the foundation for valid scientific findings. Exploratory research is usually the creative step that is the starting point for a line of research. However, exploratory research is prone to various questionable methodological practices (Ioannidis, 2012; Kennedy, 2014b; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kevit, 2012). Confirmatory research provides the convincing evidence that makes science valid and self-correcting. Exploration and confirmation are both essential for science. In the past few decades, research in the social sciences has developed an unhealthy emphasis on exploratory research without adequate consideration of confirmation (Ioannidis, 2012; Wagenmakers et al., 2012). Fortunately, a more balanced perspective has been rapidly emerging (e.g., Open Science Collaboration, 2012), although some psychological researchers currently continue to advocate statistical methods that blur the distinction between exploration and confirmation (see Kennedy, 2015). Meta-analysis of exploratory studies does not eliminate the need for well-designed confirmatory research (Kennedy, 2013a).

Confirmatory methodology for experiments includes certain key practices: prespecification of the statistical methods and the criteria for acceptable evidence, sample sizes 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, 2014b; KPU Registry, 2014). These practices are based on established confirmatory methodology in regulated medical research and on widely recognized principles for good research methodology. The Koestler Parapsychology Unit study registry now provides public prospective registration for parapsychological experiments (KPU Registry, 2012; Watt & Kennedy, 2015).

The quantitative evaluation of potential inferential errors is a fundamental factor for planning confirmatory research. Errors in inference and power analysis have long been standard topics in statistical textbooks for psychologists (e.g., Hays, 1963; Keppel, 1973). Exploratory research usually focuses on either *p*

values or effect sizes, with little or no consideration of sample size, power, and inferential errors. Properly designed confirmatory research incorporates an explicit, balanced recognition of the interactions among effect size, sample size, statistical significance, and inferential errors.

Evaluation of inferential errors and power quantifies the statistical validity of a planned hypothesis test. These evaluations determine the rates of correct and incorrect inferences if the true effect size is a certain value, and the corresponding rates if the null hypothesis is true. The evaluations can be done for different effect sizes to form a curve that covers the range of effects of interest. This curve represents the operating characteristics for the hypothesis test. For confirmatory research, a certain minimum effect size is often of interest and is the focus of the evaluation.

In well-designed confirmatory research, all analysis decisions that could affect the experimental results are made prior to data collection. These decisions include the specific statistical methods, criteria for acceptable evidence, specification of any data adjustments, and criteria for excluding any data from the analyses. If this information cannot be prespecified, the study is more exploratory than confirmatory. These methodological decisions should be publicly registered before data collection begins.

The use of Bayesian analysis is rapidly increasing in scientific research. The basic philosophy, assumptions, and models for Bayesian analyses were described conceptually in a previous paper (Kennedy, 2014a). That paper also pointed out the need for direct comparisons of Bayesian and classical methods for confirmatory research.

Kruschke (2011, p. 321) defines statistical power in a Bayesian context as "[t]he probability of achieving the goal [of the study], given the (suspected) true state of the world." He describes the value of using simulated data to evaluate power when planning research, but he does not address all types of inferential errors.

Most writings on Bayesian methods have focused on exploratory research without considering inferential errors or prespecifying the criteria for acceptable evidence. Few writings address confirmatory practices. One notable exception is the document by the U.S. Food and Drug Administration (2010) that has recommendations for the use of Bayesian methods when seeking approval of medical devices. The document recommends that the experimental design and protocol include the prior probabilities and statistical models that will be used, the criteria for acceptable evidence, and the operating characteristics for type I errors and power.

The present paper quantitatively compares the power and rates of inferential errors for classical and Bayesian analyses for examples of confirmatory experiments that use binomial analysis. The primary purpose is to investigate and to verify the need to evaluate statistical power and inferential error rates for Bayesian hypothesis tests.

Classical Power Analysis

The decision process for a classical hypothesis test for confirmatory research is to accept or to reject the null hypothesis. If the null hypothesis that the results are due to chance is rejected, the experiment is interpreted as providing evidence for the alternative or experimental hypothesis. Thus, the experiment has two possible outcomes—the null hypothesis is either accepted or rejected. Note that this decision process applies to well-designed confirmatory research but not to the more common exploratory analyses that are typically underpowered. In underpowered studies, one cannot accept or support the null hypothesis because nonsignificant results could be due to low power rather than to the null hypothesis being true.

Classical power analysis determines the sample size needed for an experiment to have a high probability of reaching the correct conclusion. The power analysis must consider the probability of rejecting the null hypothesis when the alternative hypothesis is true and the probability of accepting the null hypothesis when the null is true. Standard power calculations determine the sample size from four factors: (a) the desired power of the experiment, which is the probability that the null hypothesis will be rejected if the alternative or experimental hypothesis is true, (b) the effect size for the alternative hypothesis, (c) the alpha or significance level for the analysis, which is the probability of making a type I error that incorrectly rejects the null hypothesis when it is true, and (d) whether the statistical test is one- or two-sided.

For confirmatory research, the effect size for a power calculation is typically based on previous research and/or on an effect that would be meaningfully convincing or useful. The usual recommendations are that the power be at least .8 for confirmatory research and preferably higher, such as .95. The alpha or significance level is typically set at .05. However, the convention for alpha = .05 has become overly rigid and does not consider the nature of the phenomenon being investigated or the distinction between exploratory and confirmatory research. For confirmatory research on controversial effects, an alpha of .01 may be more appropriate. On the other hand, an alpha larger than .05 is often appropriate for small exploratory studies. The fixation on alpha = .05 has promoted ambiguity about whether research is exploratory or confirmatory.

When the effect size for the alternative hypothesis is estimated from previous data, the estimate may not be accurate. If the true effect size is greater than the estimate, the experiment will have greater power than the calculated power. If the true effect size is less than the estimate, the experiment will have lower power. Good practice is to consider the confidence interval for the effect size estimate and to use an effect size for the power calculation that is from the lower part of the confidence interval.

Power calculators are available online and as free programs—notably the G*Power program (Faul et al., 2007; Faul, Erdfelder, Lang, & Buchner, 2012). For binomial power calculations, some calculators use a normal approximation that is not accurate for small samples sizes. Exact calculations, such as performed by the G*Power program, are preferable for small samples sizes.

Background for Bayesian Analysis

Bayesian analysis is based on the philosophical position that probability reflects uncertainty in the human mind rather than uncertainty in the physical world (Kennedy, 2014a). Classical analysis is based on models of uncertainty in the physical world. Bayesian analysis requires models for the human mind in addition to models of processes in the physical world. The analysis starts with *prior probability distributions* that model the beliefs and uncertainties in a human mind prior to the experiment. These prior probabilities are updated based on the results of the experiment to produce *posterior probability distributions* that represent what a person should believe after the experiment. Obviously, a Bayesian analysis produces different results for different prior probabilities or beliefs. *Objective Bayesian methods* attempt to minimize the subjective aspects and potential biases of prior probability distributions.

The *Bayes factor* is a measure of the evidence from the current study and has a key role in Bayesian hypothesis testing. It is the ratio of the posterior probability for the experimental outcome if the outcome was produced by the alternative model divided by the posterior probability for the experimental outcome if the outcome was produced by the null model. The calculation of posterior probability for the alternative model requires a prior probability distribution for the effect size for the alternative model. The debates about Bayesian results often hinge on differing opinions about the prior probability distribution for an effect size. Objective Bayesian hypothesis tests typically focus on the Bayes factor.

Widely accepted conventions have not been established for the magnitude of the Bayes factor (or odds) that is considered acceptable evidence. Discussions of this topic usually refer to Jeffreys (1961), who said he used an odds of 3 the way a classical analyst uses $p = .05$ and an odds of 10 the way a classical analyst used $p = .01$. Most applications of Bayesian hypothesis tests have been for exploratory research and have not specified a criterion for acceptable evidence.

An important feature of Bayesian analysis is that the Bayes factor can be inverted to give the odds that the results were produced by chance, as assumed for the null model. The criterion for acceptable evidence can be applied to the null model as well as to the alternative model. This can provide evidence directly supporting the null hypothesis.

A Bayesian hypothesis test can have three possible outcomes. The Bayes factor can exceed the criterion supporting the alternative model, or it can exceed the criterion supporting the null model, or it can 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.

Proponents of Bayesian analysis have sometimes argued that type I error and power are classical concepts that should not be considered with Bayesian analysis. That argument is basically a philosophical

stance that virtually eliminates any discussion of inferential errors or accountability for an analyst. Fortunately, the majority of analysts appear to recognize that inferential errors can occur with Bayesian analyses and should be addressed in scientific research, particularly confirmatory research. Kruschke (2011) discusses statistical power from a Bayesian perspective.

Evaluations of inferential error rates and power determine whether the methodological decisions about prior probabilities, specific statistical models, acceptance criteria, and sample size combine to form an effective process for making scientific inferences about a phenomenon. Researchers planning a confirmatory study must make decisions for each of these factors, and those choices affect the probability of making an incorrect inference. These decisions should be based on an understanding of how the different possible choices affect the operation of the hypothesis test. The effectiveness of a confirmatory hypothesis test needs to be quantitatively evaluated with all of the decisions for these factors operating together.

Investigation Plan

A basic strategy for evaluating statistical decision-making processes is to apply the processes to data with known properties, as in the common practice of using simulations to evaluate and compare statistical methods. This strategy can be used to evaluate any decision-making process, including Bayesian hypothesis tests. For purposes of evaluation, a hypothesis test can be conceptualized as a black box that has an input of data and outputs a decision. The black box can be any type of hypothesis test and is evaluated by observing the output when the input data have known properties. The same input data can be used to evaluate different hypothesis testing methods.

The first step is to develop a model for generating data that simulates an effect that the experiment is intended to detect, and a model that generates data for the null hypothesis. The planned statistical analysis is applied to data from each model, and the rates of correct and incorrect inferences determined. This pragmatic approach underlies traditional power analysis and indicates the statistical validity of the hypothesis test. An appropriate effect size or range of effect sizes for these evaluations is usually obvious for confirmatory research.

These evaluations estimate the probabilities of making correct and incorrect inferences for an experiment if the effect has a specific effect size. For binomial data, the effect is the value of the parameter P in a binomial model. Most binomial hypothesis tests assume that the effect being investigated has a fixed but unknown value of P. As discussed in the previous paper (Kennedy, 2014a), Bayesian prior probability distributions represent the uncertainties in the beliefs in a human mind, not variability or random effects in the phenomenon being investigated. The prior probabilities are part of the hypothesis test, not part of the models for generating data to evaluate the hypothesis test. The data for evaluating the hypothesis-testing methods simulate conditions in the external world. Thus, a fixed value of P is used to simulate a phenomenon for a typical binomial hypothesis test.

In the comparisons described below, the sample sizes for an experiment were determined using classical power analysis, and then the corresponding probabilities of correct and incorrect inferences for a Bayesian analysis were determined for that sample size and effect size. A Bayes factor of 3 was set as the criterion for significant or acceptable evidence for the alternative model or for the null model. The number of hits for obtaining a Bayes factor of 3 was found, and the probabilities of reaching that number were determined for both the null model and the modeled effect.

As noted above, a Bayesian hypothesis test has three possible outcomes. The evaluations here provide the probability of each possible outcome if the null hypothesis is true and if the modeled alternative hypothesis is true.

The cases described below are for studies with $P = .5$ for the null hypothesis, as is typical for parapsychological experiments with random event generators (REGs). Two different effect sizes were investigated as alternative models, $P = .53$ and $P = .503$. Both are within the range of effects reported as evidence in parapsychological experiments. For each effect size, the sample size was determined for three different sets of parameters for the classical power analysis: (a) alpha = .05 and power = .80, (b) alpha = .05 and power = .95, (c) alpha = .01 and power = .95. These options cover the range of power analysis that would

reasonably be used in designing a confirmatory experiment. The classical sample sizes were determined using the G*power program (Faul et al., 2012).

The comparisons were based on two-sided tests. This was done because two-sided tests are often recommended in general and are specifically appropriate for parapsychology given the established occurrence of psi missing. Psi missing has been prominent in studies with REGs. Also, two-sided analyses are the only option for the online binomial Bayes factor calculator that was used.

The Bayes factor analysis used the online binomial Bayes factor calculator provided by Rouder (2012). As is common for Bayesian binomial analysis, this calculator has a beta distribution for the prior probability and requires that the two beta parameters be specified. The parameters beta(1,1) were used for one analysis. These provide a uniform prior distribution that is frequently recommended as an objective Bayesian prior and is based on the assumption that any hit rate is equally possible between 0% and 100%. However, this "objective" prior distribution does not reasonably represent the beliefs of either proponents or skeptics of psi. A second analysis was done with parameters beta(22,22). These parameters represent a belief that effects between .4 and .6 are most likely. This distribution is symmetric and has 82% of the distribution between .4 and .6. The beta(22,22) prior is more representative of beliefs based on previous research than is the beta(1,1) prior. However, the selection of this prior distribution was somewhat arbitrary, and arguments can be made that a different prior distribution is preferable. Differing opinions about prior probability is the nature of Bayesian analysis. The purpose here is to investigate the value of power analysis and inferential error evaluations for Bayesian hypothesis tests. Identification and justification of an optimal prior distribution for a particular situation is a different task, and it may be affected by the findings from the present investigation. Online graphs and tables of the beta distribution (e.g., Casio Computer Company, 2014) can be used to display and explore the different options.

Once the number of hits for a Bayes factor of 3 was determined for the different sample sizes and beta parameters, the Stat Trek (2014) online binomial calculator was used to calculate the probabilities for the different possible outcomes under the null and alternative models. The binomial calculator gives cumulative probabilities that are equivalent to those obtained from running an extremely large number of simulations. The Appendix provides more detailed information about this process and gives the cutoff values that were used. Because the distributions investigated here are symmetric, the evaluations for a positive deviation give the same results as those for a negative deviation. For example, the evaluation of $P = .53$ gives the same inferential error rates and power as an evaluation of $P = .47$.

The Jeffreys prior that uses beta(.5,.5) is sometimes recommended as an objective prior and was also initially examined. This is a U-shaped distribution that is higher on the tails near 0 and 1 and thus is even less realistic than the uniform distribution. When it became apparent that this prior gives lower power than the uniform distribution, further consideration of it was rejected.

Results

The results are shown in Tables 1 and 2 for the two different effect sizes. The null model is designated as H_0 in the tables and the alternative model as H_1. The column for "Probability $BF(H_1) \geq 3$" under "If H_1 is True" gives the basic power for the Bayesian analysis, that is, the probability of obtaining a Bayes factor significantly (≥ 3) supporting the alternative model if the alternative model is true. The column for "Probability $BF(H_1) \geq 3$" under "If H_0 is True" gives the alpha level or probability of type I error for the Bayesian analysis, that is, the probability of obtaining a Bayes factor significantly in favor of the alternative model if the null model is true. The column for "Probability $BF(H_0) \geq 3$" under "If H_1 is True" is important because it gives the probability of obtaining a Bayes factor significantly in favor of the null model if the alternative model is actually true.

The Bayesian analyses generally have substantially lower power than the classical analyses, particularly for the uniform prior distribution. As shown in Table 1 for the hit rate of 53%, when the classical analysis has a power of .80, the power for the Bayes factor with a uniform prior is .393, and the probability of obtaining a Bayes factor that incorrectly supports the null model is .294. The probability of making a type I error with the uniform prior is .002 in all three cases and, thus, is much lower than the classical alpha level.

The power increases and error rates decrease for the beta(22,22) prior and as sample size increases. For the case with classical power of .95 and alpha of .01, the Bayesian power and probability of type I error for the beta(22,22) prior are .944 and .008 respectively, which are very close to the classical values.

The lower power and higher associated errors are dramatic for the smaller hit rate of 50.30%. As shown in Table 2, when the classical analysis has a power of .80, the power for the Bayes factor with a uniform prior is only .173, and the probability of obtaining a Bayes factor that incorrectly supports the null model is .619. With the beta(22,22) prior, the probability of the Bayes factor incorrectly supporting the null hypothesis is still greater than the probability of it correctly supporting the alternative model (.386 versus .318). For all the cases in Table 2, the Bayesian analysis has substantially less power, and in most of them it is substantially biased in favor of the null model. The probability of making a type I error is essentially zero.

Table 1

Inferential Errors and Power for Classical and Bayesian Binomial Analyses for H_1 Effect Size = 53.00% and H_0 (Chance) = 50.00%

Classical Power Analysis			Analysis for Bayes Factor (using *N* from the classical power analysis)						
Alpha	Power	*N*	Prior Beta Parameters	If H_1 is True (*P* = .53)			If H_0 is True (*P* = .50)		
				Probability $BF(H_1) \geq 3$	Probability $BF(H_0) \geq 3$	Probability $BF < 3$	Probability $BF(H_1) \geq 3$	Probability $BF(H_0) \geq 3$	Probability $BF < 3$
.05	.80	2,189	1,1	.393	.294	.313	.002	.976	.022
			22,22	.629	.069	.302	.014	.814	.172
.05	.95	3,613	1,1	.672	.106	.222	.002	.982	.016
			22,22	.844	.017	.139	.010	.866	.124
.01	.95	4,963	1,1	.846	.035	.119	.002	.984	.014
			22,22	.944	.004	.052	.008	.888	.104

$BF(H_1) \geq 3$ indicates that the Bayes factor is 3 or greater, supporting the alternative model.
$BF(H_0) \geq 3$ indicates that the Bayes factor is 3 or greater, supporting the null model.
$BF < 3$ indicates that the Bayes factor is less than 3 for both the alternative and null models.

Table 2

Inferential Errors and Power for Classical and Bayesian Binomial Analyses for H_1 Effect Size = 50.30% and H_0 (Chance) = 50.00%

Classical Power Analysis			Analysis for Bayes Factor (using *N* from the classical power analysis)						
Alpha	Power	*N*	Prior Beta Parameters	If H_1 is True (*P* = .503)			If H_0 is True (*P* = .500)		
				Probability $BF(H_1) \geq 3$	Probability $BF(H_0) \geq 3$	Probability $BF < 3$	Probability $BF(H_1) \geq 3$	Probability $BF(H_0) \geq 3$	Probability $BF < 3$
.05	.80	218,187	1,1	.173	.619	.208	.000	.998	.002
			22,22	.318	.386	.296	.002	.988	.010
.05	.95	361,059	1,1	.417	.337	.246	.000	.998	.002
			22,22	.601	.160	.239	.000	.992	.008
.01	.95	494,843	1,1	.643	.162	.195	.000	.998	.002
			22,22	.795	.061	.144	.000	.992	.008

$BF(H_1) \geq 3$ indicates that the Bayes factor is 3 or greater, supporting the alternative model.
$BF(H_0) \geq 3$ indicates that the Bayes factor is 3 or greater, supporting the null model.
$BF < 3$ indicates that the Bayes factor is less than 3 for both the alternative and null models.

Discussion and Conclusions

These findings verify that quantitative evaluation of inferential errors is needed for confirmatory research with Bayesian analysis, as well as for confirmatory research with classical analysis. The investigations found that the Bayesian analyses of the binomial data tended to have lower power and higher rates of associated inferential error than the classical hypothesis tests. For small effect sizes, the Bayesian analyses were strongly biased toward supporting the null model when the alternative model was true.

The analyses here should not be taken as evidence that classical hypothesis tests are generally superior to Bayesian methods, or as recommendations about prior probabilities or the magnitude that a Bayes factor must reach to provide acceptable evidence. Rather, my expectation is that useful prior probabilities and criteria for acceptable evidence can be found for Bayesian analyses if inferential errors and power are evaluated. The primary conclusion here is that these evaluations are essential for confirmatory research.

Small Effects

The differences in conclusions between Bayesian and classical analyses for small effects are well known among statisticians and are a manifestation of the *Jeffreys-Lindley paradox*. Diffuse prior probability distributions, such as a uniform distribution, enhance this paradox. Some proponents of Bayesian methods have argued that these differences between classical and Bayesian results indicate that classical methods are flawed (e.g., Jefferys, 1990; Wagenmakers & Grünwald, 2005). This argument appears to be based on the assumption that Bayesian methods are theoretically optimal and should be the standard for evaluating other statistical methods. However, a more common and less disputable strategy for evaluating statistical methods is to use data that simulate effects of interest. These more empirical evaluations are less susceptible to overly optimistic theoretical ideas. The evaluations here show that Bayesian analyses can be strongly biased in favor of the null model when the alternative model is actually true. Similar conclusions have been reached by others (e.g., Bem, Utts, & Johnson, 2011; Dobyns, 1992).

Some defenders of Bayesian methods attempt to put a positive spin on the bias against small effects by claiming that small effects are usually due to methodological artifacts or to slight meaningless effects for the phenomenon being investigated (e.g., Jefferys, 1990). That argument may have some plausibility for exploratory research, but it is not appropriate for well-designed confirmatory research that is attempting to verify previous findings. The argument is basically a speculation that is impervious to empirical data. Larger sample sizes cannot overcome these a priori assumptions that the effects are meaningless. The result is that small effects could be largely excluded from scientific findings with Bayesian analysis—a position that is inconsistent with the basic principles of empirical scientific research.

A more appropriate strategy for reaching valid scientific conclusions is to conduct well-designed confirmatory research rather than use biased statistical methods. Parapsychological research is just beginning to implement this type of confirmatory methodology.

Counterintuitive, Arbitrary Priors

The prior probabilities in Bayesian hypothesis testing can have counterintuitive consequences. For example, a uniform prior probability for effect size appears on the surface to represent a very open-minded prior belief that is typically described as "objective." However, a diffuse prior like this makes any specific effect size have a relatively low probability—including the specific effect size that is the experimental outcome. A low prior probability for the experimental outcome makes the Bayes factor favor the null hypothesis. In effect, the power of the hypothesis test is spread over an unrealistic "objective" range that results in the analysis favoring the null hypothesis. On the other hand, a narrow prior probability distribution can have a much higher or much lower probability for a particular effect size, depending on the details of the distribution and the specific effect.

The evaluation of inferential errors using simulated data for an effect of interest is very useful for revealing counterintuitive biases for a planned Bayesian hypothesis test. The Jeffreys-Lindley paradox is another manifestation of counterintuitive effects that can bias a hypothesis test.

In practice, the selection of a prior probability distribution is often substantially arbitrary. A wide range of prior probability distributions from diffuse to narrow usually can be justified as reasonably plausible. The beta(22,22) prior distribution was somewhat arbitrarily selected for the present investigation and will likely be considered inappropriate by some researchers. I cannot defend it as optimal or as specifically representing my personal beliefs. I can only say that it appears to me to be within the wide range of plausible priors.

Inferential errors and power can be useful factors in selecting prior probabilities for confirmatory research. Most writings on Bayesian analysis focus on the exploratory stage of research and recommend retrospective sensitivity analysis of prior probabilities. However, for confirmatory research, the sensitivity analyses need to be done at the planning stage and the selected priors included in the preregistered study information.

If an analysis has a high probability of inferential errors for an effect size that is of primary interest to the experimenter, the design and/or analysis need to be modified. For confirmatory research, this situation needs to be discovered at the planning stage rather than through retrospective sensitivity analysis. An efficient strategy may be to find a prior probability distribution that gives useful rates of inferential errors, and then consider whether the selected prior is within the range of plausible priors.

The potential for inferential errors also affects the interpretation of the final experimental results. For a classical analysis with low power, a nonsignificant result is ambiguous because the result could be due to low power or to the experimental hypothesis being false. A more fundamental point is that the absence of a careful power analysis typically indicates exploratory methodology that is prone to various questionable practices that can be difficult to detect from the final report. This is true whether or not the experimenter describes the research as exploratory. Similarly, for a Bayesian analysis that has a high probability of incorrectly supporting the null model when the alternative model is true, an experimental outcome supporting the null model is ambiguous, and the experiment is likely exploratory and prone to other questionable methodological practices that could be difficult to detect from the final report.

Some Bayesian analysts (e.g., Kruschke, 2011) consider the Bayes factor to be an undesirable hypothesis-testing method because of the high potential for bias. These analysts propose alternative Bayesian methods. If the proposed alternative methods are used for confirmatory research, prespecification of the criteria for acceptable evidence and evaluation of inferential errors and power are needed.

Limitations of Posterior Probabilities

An often underappreciated limitation of Bayesian analysis is that the mathematical functions representing the posterior probabilities do not represent the multifactorial contingencies of actual scientific beliefs. The validity and meaning of the Bayes factor (or *p* value) for an experiment depend on the methodology that was used. The statistical evidence from an experiment is contingent upon good methodology, but key methodological factors (e.g., preregistration of the planned analysis, software validation, and measures to prevent fraud) are not considered in the statistical models. This larger context that is not represented in the mathematical models can be decisive when evaluating the evidence from an experiment. Attempts to quantitatively adjust posterior probabilities and subsequent prior probabilities for these methodological factors are inevitably subjective and imprecise.

Parapsychological research with electronic REGs is a clear example of the importance of the larger context for research. Meta-analyses of experiments with REGs have consistently found that smaller studies have larger effects (Kennedy, 2013a). This pattern is a recognized symptom of methodological bias, but could also be a property of psi (Kennedy, 2013a). Either way, the data are not consistent with a straightforward statistical analysis. The evidence from parapsychological experiments with REGs depends more on a person's opinion about this property of the results than on the specific *p* values or Bayes factors that are produced in an analysis.

Reasonable expectations for convincing experimental evidence are: (a) confirmatory methodology as described in the introduction, and (b) reliable confirmatory results that have properties consistent with the assumptions for the statistical analyses. Parapsychological research and most psychological research have not yet met these standards (Kennedy, 2013a, 2014b; Wagenmakers et al., 2012).

Overall Summary

The key points in this paper can be summarized as:

1. Quantitative evaluation of expected inferential errors and power is essential for planning statistically-based confirmatory research, including research with Bayesian analyses. These evaluations can be done by determining the probability of correct and incorrect inferences for the planned analysis when applied to data with and without the effect that the study is designed to detect. These evaluations are relatively easy for binomial analyses.
2. Much work remains to be done to develop reasonably unbiased Bayesian methods. The results of Bayesian hypothesis tests currently represent explorations of the properties of poorly understood and somewhat arbitrarily selected mathematical functions more than beliefs that a person has or should have. Better understanding of errors in inference will contribute significantly to the development of Bayesian methods.

The first conclusion above is consistent with the U.S. Food and Drug Administration (2010) recommendations on the use of Bayesian methods when seeking approval of medical devices. These recommendations recognize the current uncertainties and limitations of Bayesian analyses and are appropriate for confirmatory research that is expected to receive critical scrutiny (Kennedy, 2014a).

Exploratory analyses present different challenges than those discussed here. One problem is that researchers often report exploratory analyses only if the outcomes are suggestive of an effect. This introduces a fundamental bias for false-positive errors and shows the need for confirmatory research. Attempts to provide convincing evidence from exploratory research by altering the type I error rates do not eliminate the bias from selective reporting and do not eliminate the need for confirmatory research. The debates about Bayesian versus classical statistics have often implicitly focused on attempts to develop convincing results from exploratory research. Psychologists and parapsychologists appear to be reaching the same conclusion as has previously been reached for regulated medical research: well-designed confirmatory research is required for convincing scientific evidence.

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

Casio Computer Company. (2014). Beta distribution (chart) calculator. Retrieved from http://keisan.casio.com/exec/system/1180573226

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2012). G*Power 3 [Computer software]. Retrieved from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3.

Hays, W. L. (1963). *Statistics*. New York, NY: Holt, Rinehart, and Winston.

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, England: Clarendon Press.
Kennedy, J. E. (2013a). Can parapsychology move beyond the controversies of retrospective meta-analysis? *Journal of Parapsychology*, *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/jp13let.pdf
Kennedy, J. E. (2014a). Bayesian and classical hypothesis testing: Practical differences for a controversial area of research. *Journal of Parapsychology*, *78*, 170–182. Retrieved from http://jeksite.org/psi/jp14.pdf
Kennedy, J. E. (2014b). Experimenter misconduct in parapsychology: Analysis manipulation and fraud. Retrieved from http://jeksite.org/psi/misconduct.pdf
Kennedy, J.E. (2015). Critique of Cumming's "new statistics" for psychological research: A perspective from outside psychology. Retrieved from http://jeksite.org/psi/critique_new_stat.pdf
Keppel, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.
KPU Registry. (2012). Registry for parapsychological experiments. Retrieved from https://koestlerunit.wordpress.com/study-registry/
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.
Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660. Retrieved from http://pps.sagepub.com/content/7/6/657.full
Rouder, J. (2012). Bayes factor for a binomially distributed observation. Retrieved from http://pcl.missouri.edu/bf-binomial
Stat Trek. (2014). Binomial calculator: Online table. Retrieved from http://stattrek.com/online-calculator/binomial.aspx
U.S. 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
Wagenmakers , E., & Grünwald, P. (2005). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science*, *17*, 641–642.
Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. J., & Kevit, R. (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
Watt, C., & Kennedy, J. E. (2015). Lessons from the first two years of operating a study registry. *Frontiers in Psychology*, *6* (article 173), 1–4. DOI: 10.3389/fpsyg.2015.00173

Broomfield, CO, USA
jek@jeksite.org

Appendix

The online binomial Bayes factor calculator provided by Jeff Rouder (2012) can be used for two-sided binomial analyses. This calculator provides the Bayes factor that has the posterior probability of the experimental outcome with the null model divided by the posterior probability of the experimental outcome with the alternative model. For $P = .5$, this Bayes factor as a function of the number of hits is a symmetric bell-shaped curve. If a Bayes factor of 3 is considered acceptable evidence, outcomes in the middle of the curve with Bayes factors greater than 3 are evidence supporting the null hypothesis. The points on each side where the Bayes factor just reaches 3 are the cutoff values or critical values for supporting the null hypothesis. Bayes factors of .333 or less on the tails support the alternative hypothesis. If the Bayes factor is inverted to give the posterior probability with the alternative model divided by the posterior probability with the null model, these tail values are 3 or greater supporting the alternative hypothesis. The points on each side where the Bayes factor just reaches .333 or less are the cutoff values or criteria for supporting the alternative hypothesis. Outcomes with a Bayes factor that falls between 3 and .333 do not clearly support either the alternative or null hypothesis. For a given number of trials, prior beta parameters, and P for the null model, the cutoff values can easily be found by trying different numbers of hits until the Bayes factor

just reaches 3 or .333.

Once these cutoff points for the Bayes factor have been determined, the power and probability of errors in inference can be found using standard binomial calculators. The Stat Trek (2014) website provides a useful online binomial calculator that uses exact methods and normal approximations as appropriate.

Classical power can be calculated with a binomial calculator as the probability of reaching the cutoff for significance when *P* in the binomial model is the hit rate (effect size) assumed for the alternative hypothesis in the power evaluation. Likewise, the alpha level or probability of type I error can be calculated with a binomial calculator as the probability of reaching the cutoff for significance when *P* is the hit rate for the null hypothesis.

These same types of calculations can be done using the cutoff criteria for the Bayes factor. The use of the binomial calculator is straightforward once the Bayes factor cutoff values have been found. For the cutoff for the null hypothesis (Rouder's Bayes factor = 3), the relevant cumulative probability is less than or equal to the cutoff. For the cutoff for the alternative hypothesis (Rouder's Bayes factor = .333), the relevant cumulative probability is greater than or equal to the cutoff. The only complication is in handling both sides or tails. For the examples discussed here, the probability of outcomes on the negative side was negligible when the alternative model was applied. Both sides need to be considered when the null model is applied.

The Bayes factor cutoff values for the analyses in Tables 1 and 2 are listed below.

N	Prior Beta Parameters	*N* Hits for $BF(H_1) = 3$	*N* Hits for $BF(H_0) = 3$
2,189	1,1	1,167	1,147
	22,22	1,153	1,125
3,613	1,1	1,902	1,877
	22,22	1,885	1,851
4,963	1,1	2,595	2,566
	22,22	2,575	2,537
218,187	1,1	109,969	109,818
	22,22	109,859	109,680
361,059	1,1	181,676	181,486
	22,22	181,536	181,314
494,843	1,1	248,778	248,559
	22,22	248,617	248,361

Abstracts in Other Languages

French

ATTENTION AUX ERREURS INFÉRENTIELLES ET AUX FAIBLES PUISSANCES DANS LES ANALYSES BAYESIENNES : L'ANALYSE DE PUISSANCE EST NÉCESSAIRE POUR LA RECHERCHE CONFIRMATOIRE

RÉSUMÉ : Les erreurs dans les inférences peuvent survenir avec n'importe quelle méthode pour tester des hypothèses, dont l'analyse bayésienne. L'évaluation des taux attendus d'erreurs inférentielles est importante lorsque l'on prévoit des recherches confirmatoires, mais les erreurs inférentielles ont rarement été discutées dans la littérature sur l'analyse bayésienne. La présente étude applique les méthodes de test classique et bayésienne à des données binomiales avec certains effets et à des données simulant l'hypothèse nulle. Les analyses bayésiennes ont généralement une puissance substantiellement plus faible (probabilité de détecter

correctement un effet), particulièrement avec des petites tailles d'effet. Pour des données avec une petite taille d'effet et une puissance de 0,80 pour une analyse classique, la probabilité que le facteur de Bayes avec une probabilité a priori uniforme atteignant 3 ou plus soit correctement en faveur du modèle alternatif fut seulement de 0,173. La probabilité que le facteur de Bayes fut de 3 ou plus et soutienne incorrectement le modèle nul était de 0,619. Ces données vérifient que l'évaluation quantitative des taux d'erreur inférentielle attendus est essentielle lorsque l'on conçoit des études confirmatoires qui utilisent des analyses bayésiennes. L'argument selon lequel les biais en faveur du modèle nul sont appropriés pour des petites tailles d'effet du fait des potentiels problèmes méthodologiques est basé sur la recherche exploratoire et n'est pas appropriée pour des recherches confirmatoires bien conçues qui se concentrent sur une taille d'effet préétablie.

German

VORSICHT VOR INFERENZFEHLERN UND GERINGER TESTSTÄRKE BEI BAYESSCHEN ANALYSEN: EINE ANALYSE DER TESTSTÄRKE WIRD BEI BESTÄTIGUNGSFORSCHUNG BENÖTIGT

ZUSAMMENFASSUNG: Fehler bei Schlussfolgerungen können bei jeder hypothesenüberprüfenden Methode auftreten, auch bei der Bayesschen Analyse. Die Abschätzung der zu erwartenden Anzahl von Inferenzfehlern ist wichtig, wenn eine Bestätigungsforschung geplant wird, aber Inferenzfehler wurden nur selten in der Literatur über Bayessche Hypothesenüberprüfung diskutiert. In der vorliegenden Untersuchung wurden die klassische und die Bayessche Methode zur Hypothesenüberprüfung auf binomiale Daten bei bestimmten Effekten und auf Daten zur Simulation der Nullhypothese angewandt. Die Bayesschen Analysen weisen im allgemeinen eine deutlich geringere Teststärke auf (die Wahrscheinlichkeit, einen wahren Effekt zu erkennen), besonders bei kleineren Teststärken. Bei auf klassische Weise analysierten Daten mit einer geringen Effektstärke und einer Teststärke von .80 beträgt die Wahrscheinlichkeit, dass der Bayes-Faktor mit einem gleichförmigen Prior von 3 oder höher das alternative Modell zu Recht bestätigt, nur .173. Die Wahrscheinlichkeit, dass der Bayes-Faktor von 3 oder höher das Nullmodell fälschlicherweise bestätigt, beträgt .619. Diese Ergebnisse belegen, wie wichtig die quantitative Abschätzung erwarteter Inferenzfehler ist, wenn man Bestätigungsstudien unter Verwendung von Bayesschen Analysen plant. Das Argument, dass Abweichungen zugunsten des Nullmodells für geringe Effektgrössen angemessen sind wegen potentieller methodologischer Probleme, beruht auf exploratorischer Forschung und ist für eine wohlüberlegte Bestätigungsforschung unangemessen, die auf eine vorher festgelegte Effektstärke abzielt.

Spanish

CUIDADO CON LOS ERRORES INFERENCIALES Y BAJO PODER EN ANÁLISIS BAYESIANOS: SE NECESITA UN ANÁLISIS DE PODER PARA LA INVESTIGACIÓN CONFIRMATORIA

RESUMEN: Los errores en la inferencia pueden ocurrir con cualquier método de prueba de hipótesis, incluyendo al análisis Bayesiano. La evaluación de las tasas esperadas de errores inferenciales es importante en la planificación de la investigación confirmatoria, pero rara vez se han abordado los errores inferenciales en los escritos sobre la prueba de hipótesis Bayesiana. Apliqué métodos de prueba de hipótesis clásicos y Bayesianos a datos binomiales con efectos conocido y datos que simulaban la hipótesis nula. El análisis Bayesiano generalmente tuvo un poder sustancialmente menor (la probabilidad de detectar correctamente un efecto), en particular para los efectos pequeños. Para los datos con un efecto pequeño y poder de 0.80 en un análisis clásico, la probabilidad de que el factor de Bayes con un cálculo previo uniforme de 3 o más apoyando al modelo alternativo fue de sólo 0.173. La probabilidad de que el factor de Bayes fuera 3 o más apoyando incorrectamente al modelo nulo fue 0.619. Estos resultados verifican que la evaluación cuantitativa de los índices de error inferencial esperados es esencial en el diseño de estudios de confirmación que utilizan análisis Bayesianos. El argumento de que los sesgos a favor del modelo nulo son apropiados para los efectos pequeños por posibles problemas metodológicos se basa en la investigación exploratoria y no es apropiado para la investigación confirmatoria bien diseñada que se centra en un efecto de tamaño preestablecido.