# Bayesian Analysis

Ruud Wetzels[1], Don van Ravenzwaaij[2], Eric-Jan Wagenmakers[3]
[1]The European Foundation for the Improvement of Living and Working Conditions
[2]University of New South Wales
[3]University of Amsterdam

**Abstract**

A Bayesian analysis concerns the quantification and updating of knowledge according to the laws of probability theory. A Bayesian analysis allows researchers to estimate parameters (e.g., how effective is a new treatment?) or to conduct hypothesis testing and model selection (e.g., is the new treatment more effective than the standard treatment?). The Bayesian paradigm offers concrete advantages for the practical research worker; one of these is the ability to attach probabilities to hypotheses and parameters, and another is the ability to collect data until the evidence conclusively supports the null hypothesis or the alternative hypothesis.

**Keywords:** Statistical inference, Bayes factor, Sequential testing.
Word Count: 7492

Statistics is of central importance to all empirical sciences, including clinical psychology. In a world without statistics it remains unclear, for instance, whether a new treatment against depression is superior to the standard treatment; whether the standard treatment is superior to placebo; or whether the placebo treatment is superior to no treatment whatsoever. Only by using statistics can we draw conclusions from data that are inherently noisy. A world without statistics is reigned by chaos, anarchy, and arbitrariness. It is true that some findings are clear enough to pass what Berkson called the *interocular traumatic test* – when the data are so compelling that conclusion hits you straight between the eyes. However, "...the enthusiast's interocular trauma may be the skeptic's random error. A little arithmetic to verify the extent of the trauma can yield great peace of mind for little cost." (Edwards, Lindman, & Savage, 1963, p. 217).

Because of its pivotal role in separating the empirical wheat from the chaff, clinical psychologists are confronted with statistics as soon as they enter university. There students are implicitly or explicitly taught that statistics comprises a toolbox of objective methods, and that the main challenge is to select the tool –ANOVA, regression, correlation– that is appropriate for the situation at hand. Following selection of the right tool, the remaining work is highly automatized. Data are analyzed in a software package such as SPSS, and the final result is captured in a single number: the $p$ value. If $p < .05$, the null hypothesis can be rejected and one is allowed to claim that, for instance, the new treatment is better

than the standard treatment. These teachings carry over to later research practices, and consequently most articles in clinical psychology rely exclusively on the $p$ value to back up their statistical claims.

This state of affairs is unfortunate, for at least three reasons. First, it ignores the fact that the statistics taught (i.e., classical, orthodox, or frequentist statistics) comes with deep conceptual and practical problems. Many students, for instance, remain unaware that $p$ values depend on the intention with which the data were collected, that they overestimate the evidence against the null hypothesis, and that one may not continue data collection until the $p$ value falls below .05. Only by teaching the limitations of a particular method can it be properly understood. Second, the current state of affairs falsely suggest that statistics speaks with one voice, and that for a particular problem there exists one single correct answer. Perhaps this is comforting to the student, but in reality statisticians do not agree on how to analyze even the simplest problems. This disagreement reflects fundamentally different opinions about the nature of probability and the goal of scientific inference. The lines between competing statistical camps can be drawn a number of ways, but the most often-used distinction, one elaborated on below, is that between *frequentists* and *Bayesians*. Third, the current state of affairs is old-fashioned as it reflects the statistical state-of-the-art from half a century ago. In those times, the field of statistics was dominated by frequentists, and Bayesian statisticians constituted a small but stubborn minority. Times have changed, however, and a Bayesian revolution has transformed statistics.

## Bayesian Statistics

In deductive reasoning, general laws allow one to make statements about an individual case with absolute certainty: if all humans are mortal, and if Socrates is a human, then it follows that Socrates must be mortal. However, the empirical sciences usually proceed by inductive reasoning, where observation of individual cases alter the plausibility of a general law. The observation of every new white swan makes it more plausible that *all swans are white*; however, inductive reasoning is inherently probabilistic and does not offer the kind of absolute certainty that is associated with deduction: indeed, Australia is home to swans that are black. Hence, inductive reasoning reflects the knowledge that we have gleaned from experience. Such knowledge can be trivial (as in expecting the sun to rise in the morning) or honed through years of training (as for expert Go players) – the bottom line is that inductive knowledge is generated from experience or experiments, not from mathematical proof.

Bayesian statistics can be viewed as a formalization of inductive reasoning; Bayesian statistics provides a normative account of how rational agents update their beliefs as they are confronted with new information. In order to update beliefs in a normative fashion, these beliefs need to be formalized and changed within the framework of probability calculus. As observed by the first real Bayesian statistician, Pierre-Simon Laplace (1829), "(...) the theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it."

The axiomatic foundations of Bayesian inference have been outlined for instance in Cox (1946), de Finetti (1974), and Jeffreys (1961). These and other works have shown how the Bayesian framework rests on a few core statements or axioms. For instance, Jeffreys'

second axiom says that "The theory must be self-consistent; that is, it must not be possible to derive contradictory conclusions from the postulates and any given set of observational data." Indeed, one of the philosophical attractions of Bayesian statistics is that it avoids, by its very construction, internal inconsistencies or incoherencies.

In order to update knowledge it first needs to be quantified. Bayesian statisticians quantify knowledge (or uncertainty, or degree of belief) by means of probability distributions. Consider a one-euro coin that I just drew from my pocket, and ask yourself what the probability if of it landing heads on the next toss. Perhaps your best guess is $1/2$; however, you might entertain a margin of uncertainty around the $1/2$ mark to account for the fact that the coin may not be perfectly balanced due to design or due to wear and tear. If you are relatively certain that the coin is balanced then the probability distribution that reflects your knowledge is highly peaked around $1/2$ (i.e., a small interval around $1/2$ contains the bulk of the probability distribution); If you are less certain then the probability distribution will be less peaked and more broadly spread out (i.e., a large interval around $1/2$ is needed to contain the bulk of the probability distribution). Note that you may use all kinds of methods to quantify or enhance your knowledge: in particular, you may examine other one-euro coins and see how often they land heads rather than tails. At any rate, by quantifying knowledge through probability distributions the Bayesian statistician does not single out one particular value as unique or special; instead a range of different values is considered, and all of these are subsequently taken into account.

After quantifying knowledge by probability distributions, the data come in and the distributions are updated. For instance, assume you were relatively certain that my one-euro coin had a probability of $1/2$ to fall heads. Now you observe the following sequence of tosses: $\{H, H, H, H, H, T, H\}$. The information from these tosses is used to update your knowledge about the propensity of the coin. The Bayesian framework ensures that this updating is done in a manner that is rational and coherent. For instance, the ultimate result does not depend on the order in which the data came in, nor does it depend on whether the data came in all at once or one at a time.

In sum, Bayesian statistics has provided humankind with a formal theory of induction: it is common sense expressed in numbers. Bayesian statements are coherent, meaning that they are internally consistent. Uncertainty is quantified through probability distributions. These probability distributions are updated by new information, using the laws of probability calculus. The next two sections provide concrete illustrations of the general principle.

*Bayesian parameter estimation*

At first consider a single statistical model or hypothesis, say $\mathcal{M}_1$. In the Bayesian paradigm, the parameters of this model, $\theta$, are assigned a *prior distribution* that reflect our uncertainty or degree of belief about $\theta$, denoted $p(\theta \mid \mathcal{M}_1)$. This prior distribution for $\theta$ is updated after encountering data $y$ to yield a *posterior distribution* $p(\theta \mid y, \mathcal{M}_1)$. The posterior distribution represents all we know about $\theta$ (under the assumption that the statistical model is correct).

Consider an example on the suicidality of adolescent sexual minority members (SMMs; Plöderl et al., in press). Many studies have reported increased rates of suicide attempts among SMMs; however, the picture is less clear for completed suicides. A methodologically

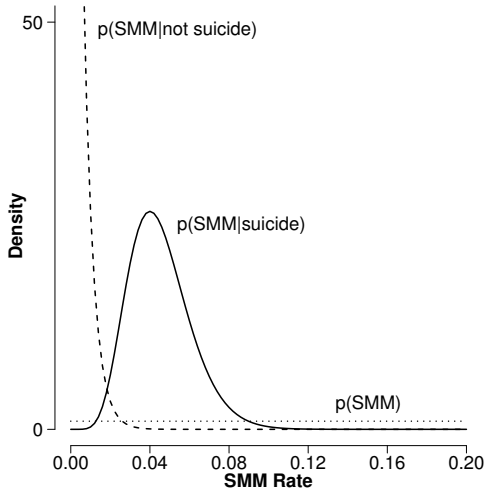


*Figure 1.* Prior and posterior distributions for the reported proportion of sexual minority members among living adolescents and adolescents who committed suicide. Data combined from Shaffer et al. (1995) and Renaud et al. (2010).

strong procedure to study completed suicides is the autopsy method. In the autopsy method, sexual orientation is determined by informants, both for a group of individuals who died by suicide and for a matched control group of individuals who did not commit suicide. The autopsy method therefore yields two proportions: the proportion of SMMs in the group whose members died by suicide and the proportion of SMMs in the matched control group. Collapsing the two autopsy studies performed so far (i.e., Shaffer, Fisher, Hicks, Parides, & Gould, 1995; Renaud, Berlim, Begolli, McGirr, & Turecki, 2010) the data show that out of the 175 adolescents who committed suicide, 7 were classified as SMM; out of the 202 matched controls, none was classified as SMM. Hence, the crucial comparison is between the SMM rate of 7/175 or 4% in the suicide group and 0/202 or 0% in the matched control group.

Figure 1 shows the result of one possible Bayesian analysis (see Plöderl et al., in press for details). The dotted, flat distribution is the prior distribution for $\theta$, where $\theta$ is the SMM rate in both the suicide group and the matched control group. This prior distribution quantifies our knowledge about $\theta$ *before* seeing the actual data. We chose a prior that is uninformative, meaning that it assigns equal prior weight to all values from 0 to 1. This prior may strike one as implausible, because the proportion of SMMs in the population is generally believed to be smaller than 15%. One alternative therefore is to propose a uniform prior distribution for $\theta$ that ranges from 0 to, say, .15. But because this alternative prior distribution is also flat the main result is not affected.

Figure 1 also shows two posterior distributions; the first posterior distribution (solid line) reflects what we know about the rate of SMMs in the suicide group (i.e., 7/175 or 4%). This distribution assigns the highest plausibility to the value of .04 or 4%, but other values

are likely as well; values exceeding 10%, however, are highly unlikely. For this posterior distribution, a 95% *credible interval* extends from .02 to .08, meaning that it is 95% certain that the true value of $\theta$ lies in this interval – in other words, 2.5% of the posterior mass for $\theta$ is lower than .02, and another 2.5% is higher than .08. The probability that $\theta$ is in between .05 and .10 equals the area of the posterior distribution from .05 to .10; here, this probability is about 1/3. In addition, the posterior distribution shows that the value of $\theta = .04$ is about 4 times more likely than the value of $\theta = .02$, because the posterior density is about 4 times as high at $\theta = .04$ as it is at $\theta = .02$.

The second posterior distribution (dashed line) reflects what we know about the rate of SMMs in the matched control group (i.e., 0/202 or 0%). Note that even though the control group does not feature a single reported SMM among 202 adolescents, the posterior distribution still has considerable mass away from zero, and the 95% credible interval extends from 0 to .02. It is of course also possible to plot the posterior distribution for the difference between the SMM rates, and to test whether or not this difference includes zero (Plöderl et al., in press; see also the next section).

The updating from prior to posterior distribution proceeds via Bayes' rule. This rule states that the posterior distribution $p(\theta \mid y, \mathcal{M}_1)$ is proportional to the product of the prior $p(\theta \mid \mathcal{M}_1)$ and the likelihood $f(y \mid \theta, \mathcal{M}_1)$:

$$p(\theta \mid y, \mathcal{M}_1) = \frac{p(\theta \mid \mathcal{M}_1) f(y \mid \theta, \mathcal{M}_1)}{m(y \mid \mathcal{M}_1)} \propto p(\theta \mid \mathcal{M}_1) f(y \mid \theta, \mathcal{M}_1). \tag{1}$$

In this equation, $m(y \mid \mathcal{M}_1)$ is the marginal probability of the data, a normalizing constant that does not involve $\theta$. The $\propto$ symbol stands for "is proportional to". The posterior distribution $p(\theta \mid y, \mathcal{M}_1)$ is a rational compromise between prior knowledge $p(\theta \mid \mathcal{M}_1)$ and the information coming from the data, $f(y \mid \theta, \mathcal{M}_1)$. Hence, the posterior distribution contains all that we know about $\theta$ (under model $\mathcal{M}_1$) after observing the data $y$. Note that the posterior distribution is conditional on the data $y$ that have been observed; data that could have been observed, but were not, do not affect Bayesian inference.

*Bayesian model selection*

Bayesian parameter estimation focuses on determining the size of an effect, taking its presence for granted. In contrast, Bayesian model selection focuses on testing for the presence of an effect, regardless of its size. Consider for example the choice between models $\mathcal{M}_1$ and $\mathcal{M}_2$. Bayes' rule dictates how the prior probability of $\mathcal{M}_1$, $p(\mathcal{M}_1)$, is updated through the data to give the posterior probability of $\mathcal{M}_1$, $p(\mathcal{M}_1 \mid y)$:

$$p(\mathcal{M}_1 \mid y) = \frac{p(\mathcal{M}_1) m(y \mid \mathcal{M}_1)}{p(\mathcal{M}_1) m(y \mid \mathcal{M}_1) + p(\mathcal{M}_2) m(y \mid \mathcal{M}_2)}. \tag{2}$$

In the same way, one can calculate the posterior probability of $\mathcal{M}_2$, $p(\mathcal{M}_2 \mid y)$. The ratio of these posterior probabilities is given by

$$\frac{p(\mathcal{M}_1 \mid y)}{p(\mathcal{M}_2 \mid y)} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \frac{m(y \mid \mathcal{M}_1)}{m(y \mid \mathcal{M}_2)}, \tag{3}$$

which shows that the change from prior odds $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ to posterior odds $p(\mathcal{M}_1 \mid y)/p(\mathcal{M}_2 \mid y)$ is given by the ratio of marginal probabilities $m(y \mid \mathcal{M}_1)/m(y \mid \mathcal{M}_2)$, a

| Bayes factor $BF_{12}$ | | | Interpretation |
|---|---|---|---|
| | $>$ | 100 | Extreme evidence for $\mathcal{M}_1$ |
| 30 | – | 100 | Very Strong evidence for $\mathcal{M}_1$ |
| 10 | – | 30 | Strong evidence for $\mathcal{M}_1$ |
| 3 | – | 10 | Moderate evidence for $\mathcal{M}_1$ |
| 1 | – | 3 | Anecdotal evidence for $\mathcal{M}_1$ |
| | 1 | | No evidence |
| 1/3 | – | 1 | Anecdotal evidence for $\mathcal{M}_2$ |
| 1/10 | – | 1/3 | Moderate evidence for $\mathcal{M}_2$ |
| 1/30 | – | 1/10 | Strong evidence for $\mathcal{M}_2$ |
| 1/100 | – | 1/30 | Very Strong evidence for $\mathcal{M}_2$ |
| | $<$ | 1/100 | Extreme evidence for $\mathcal{M}_2$ |

Table 1: Evidence categories for the Bayes factor $BF_{12}$ (after Jeffreys, 1961).

quantity known as the *Bayes factor* (Jeffreys, 1961). The log of the Bayes factor is often interpreted as the weight of evidence provided by the data.

Thus, when the Bayes factor $BF_{12} = m(y \mid \mathcal{M}_1)/m(y \mid \mathcal{M}_2)$ equals 5, this indicates that the observed data $y$ are 5 times more likely to occur under $\mathcal{M}_1$ than under $\mathcal{M}_2$; when $BF_{12}$ equals 0.1, this indicates that the observed data are 10 times more likely under $\mathcal{M}_2$ than under $\mathcal{M}_1$. Even though the Bayes factor has an unambiguous and continuous scale, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 1. We replaced the labels "worth no more than a bare mention" with "anecdotal", "decisive" with "extreme", and "substantial" with "moderate". These labels facilitate scientific communication but should be considered only as an approximate descriptive articulation of different standards of evidence.

Bayes factors represent "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648) and "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378). Nevertheless, Bayes factors come with two important challenges. The first challenge is computational: for many models, the marginal likelihoods $m(y)$ from Equation 3 are difficult to calculate. This happens because models are generally *composite*, meaning that they have one or more free parameters, and each set of parameter values yields a different likelihood. In order to combine these different likelihoods into a single number, the marginal likelihood, one needs to consider the likelihood for all parameter values separately, and then compute their weighted average. Formally, the marginal likelihood $m(y)$ can be expressed as $\int_\Theta f(y \mid \theta)p(\theta)\,\mathrm{d}\theta$: it is a weighted average across the parameter space, with the prior distribution providing the averaging weights. The computational challenge grows with the dimensionality of the parameter space. Nevertheless, modern computational methods coupled with the ever-increasing power of desktop computers are currently overcoming the computational challenge.

The second challenge is conceptual. For parameter estimation, the prior on the pa-

rameters is generally overwhelmed by the data rather quickly. Intuitively, rational agents may have different prior beliefs, but the influx of data drives them inexorably towards a common opinion. Hence, the philosophical discussion on the subjectivity of the prior distribution is not particularly worrisome to a Bayesian; for most data sets, the precise shape of the prior distribution is practically irrelevant. For model selection, however, the prior on the parameters does exert a lasting influence, particularly for parameters that are unique to the models under consideration. This happens because the prior on the parameters is part and parcel of the model specification. A model with a fixed value of $\theta$ (e.g., $\theta \sim N(0, \sigma \to 0)$) is simpler than a model for which $\theta$ is free to vary (e.g., $\theta \sim N(0, \sigma = 1)$), just as a model in which $\theta$ is allowed to vary only a little (e.g., $\theta \sim N(0, \sigma = 0.1)$) is simpler than a model in which $\theta$ is allowed to vary a lot (e.g., $\theta \sim N(0, \sigma = 10)$).

In Bayesian model selection, the marginal likelihood is determined by the proportion of the parameter space that yields a good fit to the observed data. Because the marginal likelihood is an average across the entire parameter space, complex models with high-dimensional parameter spaces are not necessarily desirable — large regions of the parameter space may yield a fit to the data that is very poor, dragging down the average. Hence, a good model is a parsimonious model that uses only those parts of the parameter space that are required to provide an adequate account of the data. In this sense the Bayes factor can be viewed as an automatic Ockham's razor, also known as Ptolemy's principle of parsimony.

The conceptual challenge can be met in several ways. Subjective Bayesians may insist that the prior distribution for all parameters needs to be determined by a thorough process of "prior elicitation" that extracts knowledge from subject-matter experts. On the other hand, objective Bayesians seek to establish a set of prior distributions with good properties for testing that can be used regardless of the subject-matter (e.g., Jeffreys, 1961). For instance, one useful default is the unit-information prior, where the prior precision is determined by the amount of information in a single observation — under this prior, an approximation to the Bayes factor is known as BIC, the Bayesian Information Criterion. The next example shows the Bayes factor with default priors in action.

Consider the possibility of a linear relation between scores on the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990) and the Anhedonic Depression scale of the Mood And Anxiety Symptom Questionnaire (MASQ.AD; Watson & Clark, 1991). PSWQ and MASQ.AD scores were obtained from a group of 40 first-year psychology students at the University of Amsterdam (data courtesy of Maurice Topper). Figure 2 suggests that there is a positive correlation, $r = .55$, between scores on PSWQ and scores on MASQ.AD. But how confident can we be that the correlation is in fact present? Wetzels and Wagenmakers (2012) describe a default test for the presence of a correlation that is based on the linear regression framework outlined in Liang, Paulo, Molina, Clyde, and Berger (2008). This framework allows us to obtain peace of mind by doing a little arithmetic to verify the extent of the interocular trauma conveyed by Figure 2.

In fact, little if any arithmetic is needed here. The default prior distributions have been specified as detailed in Liang et al. (2008), and an accompanying `R` function requires only that we enter the observed correlation (i.e., $r = .55$) and the number of participants (i.e., $n = 40$). The result is $BF_{10} \approx 107$, indicating that the data are over one hundred times more likely to occur under the alternative hypothesis $\mathcal{H}_1$ than under the null hypothesis $\mathcal{H}_0$;

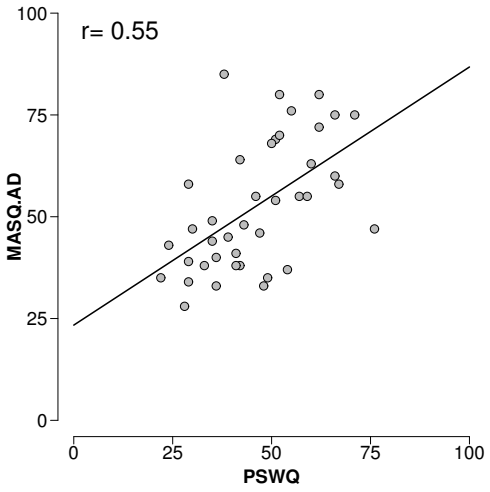


*Figure 2.* Scores on the Penn State Worry Questionnaire and the Anhedonic Depression scale of the Mood And Anxiety Symptom Questionnaire for a group of 40 first-year psychology students at the University of Amsterdam. Data collected in 2012, courtesy of Maurice Topper.

this constitutes "extreme evidence" in favor of the presence of a correlation (cf. Table 1).

This example can be extended to highlight one of the practical advantages of Bayesian analysis for clinical psychology. The concept that underlies this advantage is the *stopping rule principle* which states that "(...) the reason for stopping experimentation (the *stopping rule*) should be irrelevant to evidentiary conclusions about $\theta$." (Berger & Wolpert, 1988, p. 74). This means that evidence, quantified by the Bayes factor, may be monitored as the data accumulate – data collection may stop whenever the evidence is conclusive, be it in favor of $\mathcal{H}_0$ or in favor of $\mathcal{H}_1$. As pointed out by Edwards et al. (1963, p. 193), "(...) the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience."

Hence the stopping rule principle allows us to monitor the Bayes factor for the presence of a correlation between scores on PSWQ and scores on MASQ.AD. Figure 3 shows the result. Note that after the first 11 participants, the Bayes factor provides modest support in favor of the null hypothesis; also note that after 30 participants the evidence is still only anecdotal.

*Markov chain Monte Carlo*

For a long time, researchers could only proceed with Bayesian inference when the posterior distribution was available in closed form. As a result, practitioners interested in models of realistic complexity did not much use Bayesian inference. This situation changed dramatically with the advent of computer-driven sampling methodology generally known as *Markov chain Monte Carlo* (i.e., MCMC). Using MCMC techniques such as

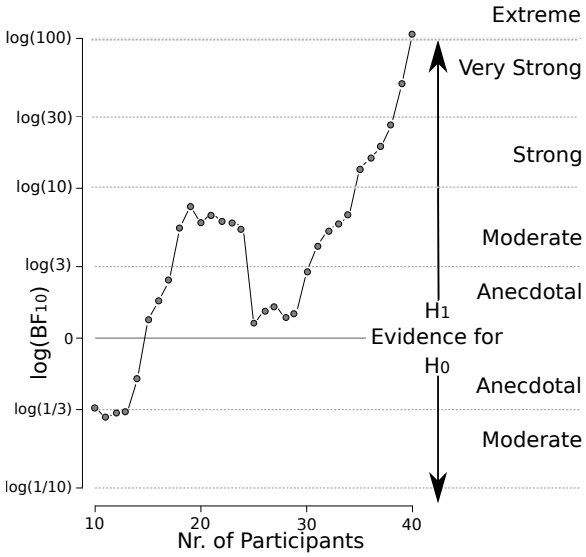


*Figure 3.* A Bayesian sequential hypothesis test for the presence of a correlation between PSWQ and MASQ.AD scores. The Bayes factor $BF_{10}$ compares two hypotheses: the null hypothesis $\mathcal{H}_0$ postulates the absence of a correlation, and the alternative hypothesis $\mathcal{H}_1$ postulates its presence. Note that the evidence may be monitored as the data accumulate, and that it is possible to quantify evidence in favor of $\mathcal{H}_0$.

Gibbs sampling or the Metropolis-Hastings algorithm, researchers can now directly sample sequences of values from the posterior distribution of interest, foregoing the need for closed form analytic solutions. The current adage is that Bayesian models are limited only by the user's imagination.

In order to visualize the increased popularity of Bayesian inference, Figure 4 plots the proportion of articles published in the *Journal of the American Statistical Association* that contain the words "Bayes" or "Bayesian". The JASA time line in Figure 4 confirms that Bayesian methods have become increasingly mainstream. Also shown is the time line for one of the flagship journals in clinical psychology, the *Journal of Consulting and Clinical Psychology*. The JCCP time line suggests that clinical psychology has yet to take advantage of the recent developments in Bayesian statistics.

## Comparison to Frequentist Statistics

Frequentist inference is based on the idea that probability is a limiting frequency. This means that frequentists feel comfortable assigning probability to a repeatable event in which the uncertainty is due to randomness, such as getting a full house in poker (i.e., aleatory uncertainty, O'Hagan, 2004). But a frequentist must refuse to assign probability to an event where uncertainty is also due to lack of knowledge, such as the event of Andy Murray ever winning Wimbledon (i.e., epistemic uncertainty, O'Hagan, 2004).

Because uncertainty about parameters is epistemic, frequentist inference does not allow probability statements about the parameters of a statistical process. For instance, the fact that a frequentist 95% confidence interval for the normal mean $\mu$ is [-0.5, 1.0] does

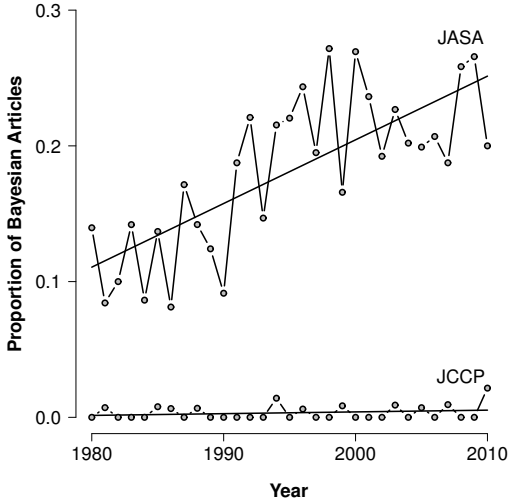


*Figure 4.* Change over time in the proportion of articles containing the words "Bayes" or "Bayesian", separately for the *Journal of the American Statistical Association* and the *Journal of Consulting and Clinical Psychology*.

not mean that there is a 95% probability that $\mu$ is in [-0.5, 1.0]. Instead, what it means is that if the same procedure to construct confidence intervals was repeated very many times, for all kinds of different data sets, then in 95% of the cases would the true $\mu$ lie in the 95% confidence interval.

Discussion of frequentist inference is complicated by the fact that current practice has become an unacknowledged amalgamation of the $p$ value approach advocated by Fisher (Fisher, 1958) and the $\alpha$-level approach advocated by Neyman and Pearson (Neyman & Pearson, 1933). Hubbard and Bayarri (Hubbard & Bayarri, 2003) summarize and contrast the paradigms as follows:

> "The level of significance shown by a $p$ value in a Fisherian significance test refers to the probability of observing data this extreme (or more so) under a null hypothesis. This data-dependent $p$ value plays an epistemic role by providing a measure of inductive evidence against $H_0$ in single experiments. This is very different from the significance level denoted by $\alpha$ in a Neyman-Pearson hypothesis test. With Neyman-Pearson, the focus is on minimizing Type II, or $\beta$, errors (i.e., false acceptance of a null hypothesis) subject to a bound on Type I, or $\alpha$, errors (i.e., false rejections of a null hypothesis). Moreover, this error minimization applies only to long-run repeated sampling situations, not to individual experiments, and is a prescription for behaviors, not a means of collecting evidence." (Hubbard & Bayarri, 2003, p. 176)

Clearly then, Fisher's approach differs from that of Neyman and Pearson, and only the latter approach is truly frequentist. Here we perpetuate the confusion and refer to both

the Fisherian and the Neyman-Pearson procedure as "frequentist".

Thus, the most fundamental differences between Bayesian and frequentist methods are the following:

**1.** For Bayesians, probability reflects their uncertainty or degree of belief; for frequentists, probability is the frequency of occurrence when the number of replications grows infinitely large.

**2.** For Bayesians, inference is conducted over the parameter space, conditional on the observed data; for frequentists, inference is conducted over the sample space of alternative outcomes, whereas the model parameters are considered fixed but unknown.

**3.** For Bayesians, all inference is conditional on pre-experimental information about the model parameters (i.e., the prior distribution); for frequentists, all inference is conditional on the sampling plan (i.e., the intention with which the data were collected, without which it is impossible to define the sample space). Both elements have a distinct subjective flavor.

## Frequentist Statistics, Bayesian Conclusions

Many authors have written about the limitations of Bayesian statistics, and many more have discussed the shortcomings of frequentist statistics. A different perspective is motivated by Gigerenzer's Freudian analogy of an applied researcher's statistical psyche (Gigerenzer, 1993). In this analogy, the applied researcher has a Superego that wants to follow the Neyman-Pearson tradition; it seeks to contrast two well-defined hypotheses (i.e., the null hypothesis and an alternative hypothesis), it operates using concepts of $\alpha$-level and power, and it is generally concerned with procedures that will work well in the long run. In contrast, the applied researcher's Ego follows the Fisherian tradition; it does not posit a specific alternative hypothesis, it ignores power, and it computes a $p$ value that is supposed to indicate the statistical evidence against the null hypothesis. Finally, the applied researcher's Id is *Bayesian*, and it desperately wants to attach probabilities to hypotheses. However, this wish is suppressed by the Superego and Ego. In its continual struggle to obtain what it desires, the Id—although unable to change the statistical analysis procedures that are used—wields its influence to change and distort the interpretations that these analysis procedures afford. In other words, many applied researchers use frequentist methods to draw Bayesian conclusions. Two examples follow.

*Case 1: From a significant p value it is concluded that the null hypothesis is false*

Consider the following hypotheses from the field of experimental psychology:

**1.** Stereotypic movements activate the corresponding stereotype. Hence, "participants who were unobtrusively induced to move in the portly way that is associated with the overweight stereotype ascribed more stereotypic characteristics to the target than did control participants, $t(18) = 2.1, p < .05$" (Mussweiler, 2006, p.18).

**2.** Increased conception risk is positively associated with several measures of race bias, particularly for those women who are vulnerable to sexual coercion. The critical two-way interaction yielded $p = .04$ (Navarrete, Fessler, Fleischman, & Geyer, 2009).

**3.** Concepts and word meanings are based partly on implicit simulations of our own actions—hence, the reading of action verbs (e.g., to throw) should result in the preferential activation of left premotor cortex in right-handers, and right premotor cortex in left-handers. For the critical three-way interaction $p = .04$, and for the critical two-way interaction $p = .02$ (Willems, Hagoort, & Casasanto, in press).

No doubt, these hypotheses are interesting and daring—daring in part because many people will find them surprising or even counterintuitive. Given the $p$ values reported above, it is customary to conclude that evidence has been collected in favor of the specified hypotheses, and that the associated null hypotheses can be rejected.

This reasoning does not hold, however, for two reasons. First, the $p$ value is not a posterior probability of the null hypothesis (with equal prior probability on $\mathcal{H}_0$ and $\mathcal{H}_1$). In fact, many researchers have shown that $p$ values (when misinterpreted as a posterior probability) tend to overestimate the evidence against the null hypothesis, sometimes dramatically so (e.g., (Berger & Sellke, 1987)). Second, extraordinary claims require extraordinary evidence. Thus, a Bayesian might well argue that an implausible hypothesis $\mathcal{H}_1$ should receive relatively little prior weight. Consequently, the data (i.e., the Bayes factor) would have to be particularly compelling in order for $\mathcal{H}_1$ to overcome the prior odds that are stacked against it (see Equation 3).

*Case 2: From a non-significant p value it is concluded that the null hypothesis is true*

In the Fisherian paradigm, $p$ values can only be used to reject the null hypothesis. The APA task force on statistical inference stressed this point by issuing the warning "Never use the unfortunate expression 'accept the null-hypothesis'." (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599). But applied researchers are often interested in demonstrating the absence of an effect, making it easy to draw premature conclusions. Consider the following examples, taken again from the field of experimental psychology:

**1.** "Participants in the memory condition were as likely to retrospectively judge a short sequence to be random as they were to retrospectively judge a long sequence to be random [$\chi^2 = 0.07, p > .79$]" (Olivola & Oppenheimer, 2008, p. 995).

**2.** "(...) there was no difference in hit rates [.782 vs. .771; $t(35) = 0.740, p = .464, p_{rep} = .57$]" (Gomez, Shutter, & Rouder, 2008).

**3.** "(...) the emotional significance of the stimuli did not result in a bias effect (...) Neither the main effect of foil valence nor the interaction between target and foil valence reached significance, both $Fs < 1$" (Zeelenberg, Wagenmakers, & Rotteveel, 2006, p.289).

In the above cases, it is entirely possible that there was an effect, but that the power of the experiment was too low for it to be detected. Again, the crucial mistake is to interpret the $p$ value as a posterior probability of the null hypothesis. Therefore, the claims above are not warranted, and the only way to quantify the evidence in favor of the null hypothesis rigorously is by means of a Bayesian hypothesis test.

## Concluding Comments

Bayesian analysis presents a coherent theory of inductive reasoning and thereby allows researchers to attach probabilities to hypotheses and parameters. Although Bayesian analysis has greatly increased in popularity in the field of statistics proper, its use in clinical psychology is still limited. Indeed, in clinical psychology (and other empirical sciences) the standard statistical analyses are frequentist. However, the desired conclusions are Bayesian, and, as a result, frequentist findings are often interpreted as if they were Bayesian. Applied researchers should be aware of the philosophical and practical differences between Bayesian and frequentist inference, and take special care to use a method of inference that is in line with the conclusions that they wish to draw.

SEE ALSO: Bayes Theorem; Clinical Mathematical Psychology; Null hypothesis significance testing debate; Theories of Truth

## Further Reading

Berger, J. O. & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician, 49*, 293-337.

Lee, M. D. & Wagenmakers (in press). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

## References

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112–139.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.).* Hayward (CA): Institute of Mathematical Statistics.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *The American Journal of Physics*, *14*, 1–13.

de Finetti, B. (1974). *Theory of probability, vol. 1 and 2.* New York: John Wiley & Sons.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Fisher, R. A. (1958). *Statistical methods for research workers (13th ed.).* New York: Hafner.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale (NJ): Erlbaum.

Gomez, P., Shutter, J., & Rouder, J. N. (2008). Memory for objects in canonical and noncanonical viewpoints. *Psychonomic Bulletin & Review*, *15*, 940-944.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence ($p$'s) versus errors ($\alpha$'s) in classical statistical testing. *The American Statistician*, *57*, 171–182.

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

Laplace, P.-S. (1829). *Essai philosophique sur les probabilités.* Brussels: H. Remy.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of $g$ priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*, 487–495.

Mussweiler, T. (2006). Doing is for thinking! Stereotype activation by stereotypic movements. *Psychological Science*, *17*, 17–21.

Navarrete, C. D., Fessler, D. M. T., Fleischman, D. S., & Geyer, J. (2009). Race bias tracks conception risk across the menstrual cycle. *Psychological Science*, *20*, 661-665.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, *231*, 289–337.

O'Hagan, A. (2004). Dicing with the unknown. *Significance*, *1*, 132–133.

Olivola, C. Y., & Oppenheimer, D. M. (2008). Randomness in retrospect: Exploring the interactions between memory and randomness cognition. *Psychonomic Bulletin & Review*, *15*, 991–996.

Plöderl, M., Wagenmakers, E.-J., Tremblay, P., Ramsay, R., Kralovec, K., Fartacek, C., & Fartacek, R. (in press). Suicide risk and sexual orientation: A critical review. *Archives of Sexual Behavior.*

Renaud, J., Berlim, M. T., Begolli, M., McGirr, A., & Turecki, G. (2010). Sexual orientation and gender identity in youth suicide victims: An exploratory study. *Canadian Journal of Psychiatry*, *55*, 29–34.

Shaffer, D., Fisher, P., Hicks, R. H., Parides, M., & Gould, M. (1995). Sexual orientation in adolescents who commit suicide. *Suicide and Life–Threatening Behavior*, *25*, 64–71.

Watson, D., & Clark, L. A. (1991). *The mood and anxiety symptom questionnaire.* (University of Iowa, Department of Psychology, Iowa City)

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Willems, R. W., Hagoort, P., & Casasanto, D. (in press). Body–specific representations of action verbs: Neural evidence from right– and left–handers. *Psychological Science.*

Zeelenberg, R., Wagenmakers, E.-J., & Rotteveel, M. (2006). The impact of emotion on perception: Bias or enhanced processing? *Psychological Science*, *17*, 287–291.