Title

Statistical Inference in Behavioral Research: Traditional and Bayesian Approaches

Authors

Etz, Alexander
Goodman, Steven
Vandekerckhove, Joachim

Publication Date

2021-04-02

Peer reviewed

Statistical inference in behavioral research: traditional and Bayesian approaches

Alexander Etz[a], **Steven N. Goodman**[b], **and Joachim Vandekerckhove**[a,*]

This is the author final version of a chapter published in *Research Integrity in the Behavioral Sciences*.

Null hypothesis significance testing (NHST) has long been a mainstay of scientific research, more in some scientific fields than others. It persists despite numerous calls across multiple scientific disciplines to abandon or at least modify the practice. In 2016, the American Statistical Association issued a statement decrying the use of the "bright line" $p < 0.05$ criterion as leading to a "considerable distortion of the scientific process." There are a number of alternatives to NHST that don't share its logical and practical deficiencies. First among them is Bayesian inference, which can be viewed as both a calculus of evidence and of belief. The Bayesian definition of "evidence" differs profoundly from what the p-value represents. In this chapter, we review deficiencies in NHST and provide an introduction to Bayesian reasoning, with particular attention to its relationship to the truth of scientific claims.

statistics | inference | Bayesian | frequentist | JASP | Bayes factor | p-value

Science starts and ends in uncertainty. As such, it should not be surprising that the properties and indeed integrity of any scientific method depends on how it represents uncertainty. The most basic measure we have for representing uncertainty is probability. Many scientists are surprised to learn that probability is an extraordinarily difficult and complex measure, philosophically and scientifically. We will begin this discussion of methods in statistical inference by outlining how different conceptions of probability lead to different approaches to inference. Many controversies about how to make proper conclusions based on quantitative data have their roots in controversies about the meaning of probability itself.

The original meaning of probability derived from the same root as "approbation," which related to the degree to which an opinion or action was supported by evidence, such as when deciding on the guilt or innocence of a suspect. This kind of probability was "epistemic" in nature; it related to one's degree of belief, or a logical relationship between the opinion and the strength of underlying evidence. This notion was completely distinct from the notion of "chance," as exemplified in games of chance, or gambling. This kind of probability came to be designated "aleatoric", related to games, or "stochastic," derived from the Greek term meaning "to aim at" or "to guess." The development of the probability concept engaged leading thinkers throughout the centuries, many of whom became known for other contributions in philosophy, mathematics, statistics, economics, medicine and physics – just a few of the many fields in which probabilities arise. A partial list includes Fermat, Pascal, Huygens, Bernoulli, deMoivre, LaPlace, Gauss, Quetelet, Boole, Venn, Markov, Kolmogoroff, Von Mises, Carnap, Keynes, Ramsey, Jeffreys, Reichenbach, de Finetti, Savage and Lindley.

This rich intellectual history evolved into a number of conceptions of probability upon which methods of statistical inference are based. These inferential methods share the strengths, limitations and intellectual motivations of their probabilistic foundations. The dominant approaches to probability can be divided into the "frequentist" and "epistemic" schools, the latter subdivided into "logical" and "subjective" conceptions of probability.

The frequentist approach to probability is actually the more recent of the two, having only been formalized in the early 20th century. It arose along with the Logical Positivist movement in philosophy that attempted to put science on a firm grounding by building it up from elements based on direct observation. The frequentist approach represented an attempt to make probability as objective and measurable a scientific quantity as physical measurements like height, weight, and mass. This was achieved by defining the probability of event A as being equal to its proportion in a pre-specified, in principle observable, "collective" of repeatable random events – equivalent to a "long-run frequency." The idea was that if we could observe this proportion, this probability would be objective, uniquely specified and observable. This probability was deductive in nature, in that once the collective was specified, the probability of outcomes within it would be set, or as von Mises famously declared, "First the collective, then the probability."

Two features of this definition pose operational and logical problems for systems of inference that use it. First, it did not apply to individual events, but rather to the collective itself (i.e., the "long run"). Thus, if an experiment is generating a single outcome to which we want to assign a probability, this definition says that we cannot apply a probability to that individual experiment, only to the "long run" of repetitions. This is why virtually every traditional statistical measure, from p-values to confidence intervals to Type I and Type II error rates, have definitions starting with "if this experiment is repeated." The resulting numbers are technically not intended to apply to the result in hand, but rather to the long-run of results within which that experiment sits.

The second problematic feature is that the long-run is constructed through a thought experiment. One can indeed "imagine" what would happen if the experiment were repeated many times, but this differs from having the multiple repetitions in hand. The issue is that there may not be a consensus on what experiment was done and therefore which "long run" is relevant; a given result can be a legitimate member of several different long runs (Wrinch & Jeffreys, 1919). To complicate this further, outcomes within a given long run may not be equiprobable; for instance, the border

[a]University of California, Irvine; [b]Stanford School of Medicine
All authors contributed to the final draft.
[*]To whom correspondence should be addressed.

of a "tail area" used to calculated a p-value is almost always the most probable outcome within that tail, and grouping them together sometimes violates inferential intuition. So the conditions for using frequentist probability as a foundation for inference comes at a price; the resulting numbers cannot be used to apply to an individual experiment, observers must agree on what the hypothetical "long run" will look like, and demonstrably different elements of a long run may be treated similarly. These properties generate the requirement for rigid pre-specification of all experimental procedures, including outcome measures and stopping rules, and cautions that 95% confidence intervals don't mean we have 95% confidence in any individual interval.

In contrast, epistemic probability, whether logical or subjective, can apply to individual events or to propositions that are not repeatable events. It is a "degree of justified belief", with the justification arising from underlying evidence. The "logical" subtype requires that the correspondence between belief and the evidence be unique, based on logical relationship, which can be difficult to establish. The subjective subtype allows for variation among individuals, but raises the question of whether this inter-subjective variability renders it illegitimate as a scientific tool.

What makes probability "scientific?". The question of what makes methods based on these types of probability "scientific", or correspond with the truth, is a central issue for science. For the frequentist, a "scientific" or objective probability is based on correspondence with an imagined empirical reality. This reality is not usually observable, and is typically based on statistical models, calculations or simulations. Even if these models are correct, or agreed upon, they apply only to the data, not to hypotheses, so a frequentist has no language or measure of uncertainty about hypotheses giving rise to the data.

The "scientific property" of epistemic probability is that probabilities are (a) consistent (i.e., one would never believe simultaneously that $P(A) > P(B)$ and $P(A) < P(B)$), and (b) coherent, in that one would never act based on these beliefs in ways that guaranteed one would be worse off. This leads us directly to Bayes' theorem; one can only satisfy these conditions if one's epistemic probabilities are modified by empirical data using Bayes' Theorem. Bayes' Theorem also guarantees that with accumulating data, intersubjective differences will eventually disappear and probability estimates will conform to observable reality (under weak conditions; Diaconis & Freedman, 1986). So epistemic probability in some sense ends where frequentist probability tries to start; a correspondence with observed reality. As Kendall (1949) stated, "Neither party can avoid using the ideas of the other to set up and justify a comprehensive theory."

Bayes' theorem. It was the need to answer fundamental questions about the behavior of games and rational betting strategies that led to developments in the calculus of probabilities. It was the theorem of an amateur scientist, the Rev. Thomas Bayes, that has reverberations today. He set out to answer the question of how much one should bet on one player versus another player in an interrupted game of chance. In solving this, he came up with an equation that today is uncontroversial as a mathematical expression, and indeed elementary. It was that the probability of two events occurring together, A and B, could be decomposed into different ways: the probability of A given B, times the probability of B, or the probability of B given A, times the probability of A. This can be written:

$$\begin{aligned} P(A \wedge B) &= P(A|B) \times P(B) \\ &= P(B|A) \times P(A). \end{aligned}$$

Equating the two expressions on the right, this can be re-arranged to yield Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}.$$

This is merely the algebra of conditional probability, subject to no more controversy than $1 + 1 = 2$. The difficulty begins when we assign meanings to A or B such that their probabilities cannot be observed. If A is a scientific hypothesis and B is data, Bayes' theorem becomes:

$$P(\text{Hypothesis}|\text{Data}) = \frac{P(\text{Data}|\text{Hypothesis}) \times P(\text{Hypothesis})}{P(\text{Data})}$$

This equation requires us to define a measure that corresponds to the probability of a hypothesis being true (e.g., $P(\text{Hypothesis})$), with or without data. This kind of probability falls into the "epistemic" category; logically justified, perhaps, but not necessarily empirically confirmable.

So Bayes' theorem, when applied to the process of inference—drawing conclusions about nature based on observed data—requires an epistemic probability of a hypothesis. This was historically known as "inverse" probability because it allows us to "invert" $P(\text{Data}|\text{Hypothesis})$ to $P(\text{Hypothesis}|\text{Data})$, but it is now more commonly called "Bayesian" probability (Fienberg, 2006). The acceptance or rejection of this foundational concept is at the core of the controversy about the use of Frequentist and Bayesian methods in statistics.

Statistical inference. Statistical inference is a subset of the broader subject of scientific inference. Inference in general examines how truth value transfers from one proposition to another. One such proposition can be a statement about observed data, and the other about a hypothesis that generates the data. An inference from the general (hypothesis) to the specific (data) is called "deductive" and is *truth preserving*, in that the conclusions are true if the premises are true. This is what makes it an attractive foundation for empirical science; it guarantees—as in pure mathematics—that all statements deriving from the premises are valid if the premises are true. But it comes at the price of not expanding our knowledge beyond what is already in the premises. Making a statement about the truth of a hypothesis based on observed data is a form of inductive inference, also called "ampliative" inference, in that the conclusion (about a hypothesis) has more explanatory power than the premises (the data). So inductive logic "amplifies" our knowledge, but at the price of not knowing if our conclusions about the hypotheses are correct.

In statistical inference, the hypotheses are probabilistic statements about nature, i.e. statistical. Examples of statistical hypotheses are "a response rate is 10%," or "the success rates of two interventions are equal." Under such hypotheses, one can predict the distribution of observations one would expect under specified experimental conditions. A prediction based on probabilistic formulae, of how often various outcomes will arise under a specified statistical hypothesis is deductive, and the attendant probabilities,"frequentist." Any probabilistic statements about the underlying truth are by definition epistemic, or "Bayesian."

Origins of frequentist inference. The central challenge of statistical inference is how to make statements not about observable data, but about the hypotheses that give rise to them, the essence of inductive reasoning. Until the early 20th century, a widely accepted methodology for how to use data to ascertain the truth of underlying hypotheses did not exist. Scientists and statisticians were familiar with the mathematics of probability, but how to use those mathematical properties to draw conclusions about nature from data was far more unsettled. The problem lay in the measure of probability itself; there was a well known formula that could guide inference about probabilities—Bayes' theorem—but its use required the acceptance of epistemic probability that many rejected as a foundation for sound science. The challenge was whether a method could be constructed, based purely on frequentist probability, that could provide a measure of uncertainty about underlying hypotheses without the machinery of Bayes or attendant Bayesian probabilities.

Frequentist inference as we know it today was really born in the 1920s and 1930s, as a reaction to the Bayesian model. The pioneers in this frequentist revolution were Ronald Fisher, Jerzy Neyman, and Egon Pearson. Fisher was a mathematician, geneticist and active experimentalist, the latter in the field of agriculture. Neyman and Pearson were mathematical statisticians. The driving motivation was to develop a new framework of inference that was "objective," in the sense that it was not based on epistemic uncertainty. Fisher believed that "The theory of [Bayesian inference] is founded upon an error, and must be wholly rejected" (p. 10 Fisher, 1925). What was this fatal error? He objected to the use of Bayes theorem when there was no basis to estimate the prior probability of a hypothesis. He rejected the widespread practice of assigning the same probability to all possible explanations, also known as a "uniform prior" (see also Aldrich et al., 2008; Zabell, 1989). Fisher's noted that representing ignorance with a uniform prior on one scale (e.g., x) can correspond to a (possibly highly) informative prior on another scale (e.g., x^2). According to Fisher, this error merely created the illusion of the Bayesian paradigm being objective and thereby "scientific."

In the 1920s Fisher constructed his own view of how inference could be conducted, without needing to specify prior distributions (uniform or otherwise). Fisher developed new approaches to both testing and estimation. He took the idea of a tail-area probability, used by Karl Pearson, and made it his central tool for statistical testing, calling it the "p-value", short for "probability" value, or "associated probability." The use of the p-value side-stepped the topic of prior probabilities by only considering what data might be observed under the hull hypothesis. It was originally intended to be used as a measure of evidence against the null hypothesis to be combined with other sources of evidence, and not as an "error rate" associated with a decision. He suggested that the ".05 level" might be a useful benchmark, but not for determining whether the null hypothesis was likely to be false, but for deciding whether an experiment was worth repeating (some more history about the origins of the .05 level is given in Cowles & Davis, 1982). He stated that the .05 threshold represented weak evidence, and that should one consider the null hypothesis to be false only if, upon repeated experimentation, "a properly designed experiment rarely fails to give this level of significance" (Fisher, 1926). So the "one and done" modern practice of declaring theoretical confirmation based on a single significant experiment is antithetical to the practice suggested by its originator.

Fisher's influence expanded immeasurably with the publication in 1925 of his landmark statistical textbook, *Statistical methods for research workers* (Fisher, 1925). This textbook was the first of its kind, aimed at practicing scientists, filled with practical examples showing how to analyze common experimental designs, and served to popularize the use of the p-value. This book, revised 14 times, was a scientific best seller from the time of its publication until after Fisher's death in 1962.

Fisher's approach to inference had both formal and informal components. Two of Fisher's contemporaries, Jerzy Neyman and Egon Pearson, began to try to reframe Fisher's ideas in a more formal mathematical framework. In 1928 they proposed a modification to the original testing procedure of Fisher (Neyman & Pearson, 1928). Their idea was to introduce an alternative hypothesis to contrast to the null hypothesis of Fisher, and to propose formal decision rules for accepting and rejecting these hypotheses. They formally introduced the now familiar notions of Type I and II errors and power. They proposed that statistical properties of various decision rules should be studied, and in 1933 they derived the properties of optimal statistical tests (Neyman & Pearson, 1933). Neyman would later turn his sights onto the topic of optimal estimation as well, proposing the now ubiquitous confidence interval procedure (Neyman, 1937).

The innovations by Fisher, Neyman, and Pearson in the late 1920s and early 1930s served as the foundation of modern mathematical statistics. In the following sections we outline the differences between them, the ways in which frequentist testing and estimation is done today, and summarize a number of common criticisms levied at them.

The basics of frequentist testing. In a statistical testing context we are concerned with deciding which of a set of competing hypotheses are true. A statistical hypothesis is a statement that refers to either values of population parameters or the functional forms of statistical models. Tests of these statistical hypotheses are used as a stand-in for tests of our actual scientific hypotheses. For instance, the statement that "the average height of men in the population is not equal to that of women" is a hypothesis about the difference between the averages of populations of men and women. If we let the parameter δ represent the difference in height between the populations of men and women, then a translation of our hypothesis into statistical language would be that $\delta \neq 0$.

In the Neyman and Pearson theory of hypothesis testing there are typically two competing hypothesis: the *null* hypothesis and the *alternative* hypothesis. The alternative hypothesis is chosen so that it corresponds with our hypothesis of interest; the null hypothesis is (typically) constructed such that it represents the complement of the alternative hypothesis. In our heights example, the hypothesis that $\delta \neq 0$ would be our alternative hypothesis and thus its complement $\delta = 0$ would be our null hypothesis.

In the Neyman and Pearson framework, the outcome of a hypothesis test is a binary decision: either reject the null hypothesis and accept the alternative, or accept (do not reject) the null hypothesis. This leads to the possibility of making two types of errors: rejecting the null hypothesis when it is actually true (false positive, α or Type I error), and not rejecting the null hypothesis when it is actually false (false negative, β or Type II error). The set of observations that would lead to rejection are called the *rejection region* of the test. If the observed data are in the rejection region (e.g., $Z > 1.96$) then one is supposed to "reject" the null hypothesis. In this framework, one chooses the rejection region such that there is no more than α% chance of making a Type I error, while at the same time keeping the chance of making a Type II error to a

minimum. In practice, the use of .05 for the Type I error and less than .20 for the Type II error has become standard.

Fisher's significance testing shares many features with that of Neyman and Pearson, with a few key differences. In Fisher's approach, one only considers which data might be observed if the null hypothesis were true. Then, once the data are in hand one computes the p-value, which is the probability of the observation plus the probability of any other observations that show a larger discrepancy with the null hypothesis. For example, if one observes a Z-value of 2.5, then the p-value is the probability of observing a Z-value greater than or equal to 2.5. Fisher felt that the exact level of the p-value was informative, whereas in a hypothesis test the only information to be used was whether the result fell in the rejection region, and a p-value was not calculated at all. If the p-value is small then one has evidence to suggest the null hypothesis is not true, with smaller p-values providing stronger evidence. On the topic of whether to regard a given p-value is considered "significant," in this approach it is "open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result" (p. 13 Fisher, 1971). This represented the kind of informality that the Neyman-Pearson hypothesis test was designed to eliminate. As noted previously, Fisher also put a particular premium on replication of significance, an idea revived in the 2000's under the rubric of research reproducibility (e.g., Goodman, Fanelli, & Ioannidis, 2016).

Fisher in particular very strongly rejected hypothesis tests as being too algorithmic and thereby anti-scientific, ironically a criticism that is today aimed at his innovation, the use of p-values. But today, the two approaches are typically taught and practiced as a unifed set of methods. For instance, a researcher might set up a null and alternative hypothesis and choose a rejection region such that they have an α level of .05, and then once the data are observed they report: a) whether the data fall into the rejection region and the null hypothesis was rejected, as well as b) the strength of the evidence against the null hypothesis in the form of a p-value. More historical detail on that controversy and the rise of frequentist inference can be found in Hald (2008), Goodman (1993), and Gigerenzer (1993, 2004).

Criticisms of frequentist testing. A number of criticisms have been levied toward the frequentist approaches to hypothesis testing. First, researchers have a tendency to misinterpret p-values and to use them draw improper inferences from their data (Greenland et al., 2016). A survey by Oakes (1986) (replicated by Haller & Krauss, 2002) illustrates these misconceptions. Oakes quizzed psychology researchers' interpretations of frequentist hypothesis tests by presenting an experiment that results in $t(18) = 2.7$ and $p = .01$. In response, 35% of the researchers marked as true the statement, "The probability of the null hypothesis has been found;" 85% endorsed the statement, "The probability that the decision taken is wrong is known;" 60% endorsed the statement "A replication has a 0.99 probability of being significant." None of these statements follow from the result of either a Fisher significance test or a Neyman and Pearson hypothesis test.

Other critiques have focused on the statistical properties of the null-hypothesis significance testing procedure. The p-value is defined as the probability, if the null hypothesis were true, of results as or more extreme than those observed in the experiment. That is, a p-value takes into account not only the results that were actually observed in the experiment, but also those that could have potentially been observed but were not. This dependence on unobserved data has been seen as an inherent weakness of the procedure (e.g., Jeffreys, 1961), and many take issue with the ambiguity in the definition of which outcomes are "more extreme" than those observed because this depends critically on the sampling plan (Lindley, 1993; Goodman, 1999a) — and the sampling plan is often arbitrarily chosen (in many research labs) or unknown (in the case of naturally occurring data, meta-analyses, etc.).

Another challenge associated with frequentist testing procedures is that they are not always logically consistent. Schervish (1996) and Royall (1997) demonstrate a number of general cases where both the process of using p-values as measures of evidence, as well as the process of strict reject/accept hypothesis tests, can lead to paradoxical inferences. Consider two researchers, Pat and Oliver, who want to test whether men and women have different heights. Both specify a point null hypothesis that the average difference is zero (i.e., $\delta = 0$), but Oliver is only interested in whether men are taller, so decides to use a one-sided test. They both agree to use $\alpha = .05$ to determine their rejection region, meaning Oliver rejects his null hypothesis if $Z > 1.64$ and Pat rejects her null hypothesis if $|Z| > 1.96$. If the experiment results in $1.64 < Z < 1.96$, then Oliver rejects his one-sided null hypothesis, $p < 0.05$, and asserts that men are taller on average than women. At the same time, Pat cannot reject her null hypothesis, because her calculated p-value is greater than 0.05, and hence cannot assert the weaker logical claim that men are either taller *or* shorter than women. Thus we are licensed to conclude that men are taller than women, but, paradoxically, have to withhold judgment about whether they are taller or shorter. Examples like these also challenge the notion that frequentist testing is completely objective; we have the same data, the same null hypothesis, yet we cannot know the p-value (and the decision) without knowing what is in the scientists' minds.

A last issue with p-values as measures of evidence is that they incorporate no information about effect magnitude. A large effect in a small study can have the same p-value as a small effect in a large study. This violates basic scientific intuition that if two observed effects have the same "statistical distance" from the null effect (e.g. the same number of standard errors), the one further from the null contradicts it more strongly. The p-value does not have that property because it is calculated only in relation to one hypothesis. Some statisticians and philosophers object on logical grounds to calling the p-value "evidence," saying that "evidence" must explicitly compare hypotheses, as the purpose of evidence is to modify belief; "evidence" is a construct that elevates data from being a neutral observation to something inferentially relevant. This is the framework for the Bayes Factor, an alternative measure to the p-value, discussed later.

Frequentist estimation. Recently the field of psychology has seen efforts to replace the apparently problematic practice of hypothesis testing with a focus on estimation (Cumming, 2014), a practice long advocated (Gardner & Altman, 1986) and now standard in biomedical research (Altman, Machin, Bryant, & Gardner, 2013). Estimation has a different goal than hypothesis testing. In estimation we begin by assuming a model for how the data were generated. A unique model is determined by its parameters, and in an estimation setting we wish to estimate which parameter values of the model are consistent with the observed data. Compared to hypothesis testing, where we start with specific hypothesis statements about the parameter values, the outcome of an estimation procedure is more open ended.

Frequentist estimation problems consists of two components: finding a point estimate of the parameter and computing an interval around it, which are called "confidence intervals." The point estimates represent the best guess for the parameter values. The informal goal of providing a confidence interval is to provide a designated level of uncertainty about that "best guess" consistent with the data. While this seems straightforward, the formal definition of a confidence interval is rather convoluted, because of its foundation in frequentist probability. Namely, "An X% confidence interval for a parameter θ is an interval (L, U) generated by a procedure that in repeated sampling has an X% probability of containing the true value of θ, for all possible values of θ" (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

It is important to appreciate the subtle implications of this definition. The frequentist paradigm allows us to make statements about *the procedure that was used to generate the estimates*, and this is different from making statements about *the estimates themselves*. To illustrate the distinction, take the following example (due to D. Basu; Ghosh, 1988). To determine a 95% confidence interval, let us ignore any data we have collected and instead generate a random number between 0 and 1 from a uniform distribution, deciding as follows:

$$\begin{cases} CI = \emptyset & \text{if number is 0.05 or smaller} \\ CI = (-\infty, +\infty) & \text{otherwise} \end{cases}$$

Note that this odd procedure has the same property: with a probability of one in twenty (5%), it will generate the empty interval $\emptyset$, which does not contain the true parameter, and with a complementary probability 95%, it will generate the infinite interval that does contain the true parameter. Hence, according to the definition, any interval generated by this procedure is a valid 95% confidence interval. It is hopefully clear that these intervals are useless for scientific inference.[1]

This example is artificial, but the same principle applies in situations where we believe strongly in the null hypothesis (e.g., the existence of ESP), or any range of hypotheses. We do an experiment that generates a 95% CI on the ESP effect size of, say, {0.3 to 0.6}. We would recognize this as very unusual, and would not accord it a 95% probability of including the truth because that would mean that after this one result we would have a 95% or greater belief in the existence of ESP. So by merely looking at the interval (i.e., the data), we accord different probabilities of it actually including the truth depending on the strength of prior evidence or beliefs. This shows that we do not accord every observed interval the same 95% chance of including the truth. If we strongly believe in the null, we will accord every observed interval including the null much higher than 95% chance of including the truth, and every CI not including the null a much lower than 95% chance of including the truth. If our prior evidence/belief is justified, this would be confirmed empirically.

From these thought experiments, we recognize that the properties of a confidence interval generating procedure do not necessarily transfer to the confidence intervals themselves, and that we need another inferential approach to know what credibility to apply to any particular observed interval. This is not a new insight. The subtle distinction between *properties of an interval* and *properties of the process that generated an interval* is why Neyman used the neologism "confidence," instead of "probability" to describe the interval, as he was aware that the confidence level did not accord with the frequentist notion of probability.

Criticisms of frequentist estimation. Most confidence intervals used in practice can be seen as inversions of one hypothesis test or another, in that the parameter values inside a $1-\alpha$% confidence interval are precisely those which would not be rejected by a level α hypothesis test. Thus, these confidence intervals necessarily inherit the statistical criticisms of hypothesis tests mentioned above.

Like hypothesis tests, confidence intervals are often misinterpreted. Hoekstra, Morey, Rouder, and Wagenmakers (2014) provided researchers and students with a survey about confidence intervals, analogous to the survey conducted by Oakes (1986) about hypothesis testing. This survey presented the result of an experiment with a 95% confidence interval for the mean ranging from 0.1 to 0.4. In response, 86% of the researchers from this sample marked as true the statement, "The 'null hypothesis' that the true mean equals 0 is likely to be incorrect"; 59% endorsed the statement "There is a 95% probability that the true mean lies between 0.1 and 0.4"; 47% endorsed the statement "The probability that the true mean equals 0 is smaller than 5%"; 58% endorsed the statement "If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4." Just as with the previous survey, none of these statements follow from the result of a confidence interval.

Bayesian methods. In the previous section, we have focused largely on null hypotheses and the associated concepts of the p-value and the confidence interval. Both of these common statistical entities are based on the calculation of the direct probability or *likelihood function* $P(\text{Data}|\text{Hypothesis})$ (Etz, 2018; Royall, 1997; Goodman & Royall, 1988). That is, they involve the probability (or probability density) of observing a particular data set, or one that is more extreme, if the null hypothesis model were true.

Likelihood functions have a number of critical useful properties. Perhaps their most useful property is that they are often easy to compute – the expected distribution of the data under various hypotheses is simply the probability density under a hypothesis, and can be calculated using a standard probability density function. For example,the probability of seeing 20 heads and 10 tails in a sequence of 30 flips of a fair coin is readily computed with the binomial formula (it is .028).

A second critical property of the likelihood is that the probability of the data under the model provides a complete quantitative description of the model's predictions. If $P(\text{Data}|\text{Hypothesis})$ is zero, then those data would falsify the hypothesis. If $P(\text{Data}|\text{Hypothesis})$ is unknown or undefined, then the hypothesis makes no known statement about those data. If $P(\text{Data}|\text{Hypothesis})$ is large, then the model strongly predicts those data. While verbal descriptions of hypotheses (and of statistical models more generally) can be useful for our understanding, statistical models are ultimately defined by what they predict and any evaluation of a model must make use of this probability one way or another.

As discussed in the previous section, classical estimation methods often involve finding those model parameters that maximize the likelihood – that is, we find for our model those parameters the make the data that we have observed the most probable. The likelihood function does not tell us which parameter values are more likely than others—the inverse probability—it tells us only that *if these parameters were true, then these data would be expected*.

[1] Note that an exactly analogous procedure can be conceived for null hypothesis significance testing: Reject the null hypothesis if and only if the 20-sided die comes up 1. Such a procedure guarantees that, in the long run, we will falsely reject approximately 5% of all true null hypotheses. The procedure is nevertheless entirely useless.

The inverse probability is the quantity more relevant to inference about the truth than the p-value!

The way in which probabilities are expressed in natural language is so prone to misunderstanding that there is a name for the particular error. The *confusion of the inverse* is the cognitive illusion that the distinct probabilities $P(A|B)$ and $P(B|A)$ are similar in magnitude. In reality, they are not – a common illustration is the case where A = "Jane is a US citizen," and B = "Jane is a member of the US Senate." In that case, $P(A|B)$ is 1, whereas $P(B|A)$ is close to 0. Similarly, if we flip a standard coin 5 times and get 5 heads (i.e., $P(\text{Data}|\text{fair coin})$), that probability is 1/32, but the probability that the coin is fair given the 5 heads (i.e., $P(\text{fair coin}|\text{Data})$) would still be 1 because—outside of statistics textbooks—biased coins do not exist (Gelman & Nolan, 2002).

Recognition that the p-value value was both widely misused in a "bright line" fashion and misinterpreted as an inverse probability led the American Statistical Association to issue a remarkable statement about p-values in 2016 (Wasserstein & Lazar, 2016), the first such statement in its 125-year history. The most important two points of the statement were that $p \leq .05$ does not mean that the probability of the null hypothesis is less than .05, and that the use of $p \leq .05$ as an indicator of the truth or falsity of a scientific claim represented poor scientific practice.

Bayesian updating. The inverse probability $P(\text{Hypothesis}|\text{Data})$—the probability that the hypothesis is true given the data—is often called a "posterior" probability since it is the probability of the hypothesis after considering the data, contrasted with the "prior" probability of the hypothesis before seeing the data. Compared to the likelihood, the posterior is a little more difficult to calculate. The calculation involves Bayes' theorem, which we will rewrite slightly differently from before:

$$\underbrace{P(\text{Hypothesis}|\text{Data})}_{\text{Posterior probability}} = \underbrace{P(\text{Hypothesis})}_{\text{Prior probability}} \times \underbrace{\frac{P(\text{Data}|\text{Hypothesis})}{P(\text{Data})}}_{\text{Updating factor}}$$

In that equation, $P(\text{Hypothesis})$ is the prior probability of the hypothesis, and $P(\text{Data}|\text{Hypothesis})/P(\text{Data})$ is an updating factor that captures how much more likely the hypothesis becomes once the data are factored in (Rouder & Morey, in press; Keynes, 1921; Carnap, 1950). This updating factor is a measure of the strength of evidence supporting the hypothesis (Royall, 1997; Berger & Wolpert, 1988; Edwards, Lindman, & Savage, 1963).

Bayesian methods have a number of attractive properties for use in science. Because they derive directly from epistemic probability theory, they are guaranteed to be internally consistent, and because they are built on a formal system, they do not rely on shortcuts, heuristics, or leaps of logic. Most importantly, because Bayesian methods allow us to calculate the probability that an hypothesis is true, their use is particularly attractive for behavioral scientists (Etz & Vandekerckhove, 2018; Vandekerckhove, Rouder, & Kruschke, 2018; Edwards et al., 1963).

Bayesian testing. The power of Bayesian methods comes with certain requirements. Because the system of inference is formal, the researcher is required to be similarly precise in the specification of their statistical assumptions – as in any formal system, the conclusions are only as good as the assumptions. It is important for the analyst to make only assumptions that are reasonable, defensible, or otherwise tenable (e.g., because it can be demonstrated that the conclusions are invariant under multiple sets of assumptions). This can be challenging to researchers accustomed to statistical analyses that work out-of-the-box and do not appear to demand such efforts, but classical methods are equally or more assumptive, just not transparently so.

How to represent the prior probability of the hypothesis, $P(\text{Hypothesis})$, is often the most contentious in model comparison exercises. Scientists conduct their research to determine the probability that an hypothesis is true, so quantifying that probability before they start is sometimes difficult, particularly if it is an unfamiliar exercise. Of course, there is nothing illogical about factoring prior information into our ultimate evaluation of an hypothesis, but sometimes that information is difficult to quantify. In such cases, it might be desirable to instead limit the scope of the calculation and *assess first only how much is learned from the data at hand*. This is where a new quantity, the *Bayes factor*, becomes useful.

Consider the case where there are two competing hypotheses, A and B, and where we have some relevant data D. For this case, there will exist two posterior probabilities: $P(A|D)$ and $P(B|D)$. A handy way of expressing which hypothesis is more likely than the other is the posterior odds $P(A|D)/P(B|D)$. Using Bayes' theorem, we know that

$$P(A|D) = \frac{P(D|A) \times P(A)}{P(D)}$$

and accordingly that

$$\underbrace{\frac{P(A|D)}{P(B|D)}}_{\text{Posterior odds}} = \underbrace{\frac{P(A)}{P(B)}}_{\text{Prior odds}} \times \underbrace{\frac{P(D|A)}{P(D|B)}}_{\text{Bayes factor}}.$$

We can read this expression in an intuitive way: the relative *posterior* probability of two hypotheses is their relative *prior* probability multiplied by the relative evidence provided by the data. This relative evidence is the predictive success of the two hypotheses and is called the *Bayes factor* (or often the likelihood ratio if A and B have no additional parameters). In other words, the relative predictive success of two hypotheses is what determines how truth value is reallocated between them.

An example of Bayesian testing Consider the case of bistable perception. In this phenomenon, a single perceptual stimulus can be seen or heard two different ways. A figure might look like a vase one moment, but look like two faces the next moment; or a drawing might look like a duck one second, a rabbit the next.

Some ambiguous percepts differ between individuals: a dress in a photograph might appear blue and black to some people, gold and white to others; or a sound clip might sound like "YANNY" to some and like "LAUREL" to others. Suppose that a researcher claims that teenagers are more likely to hear YANNY than LAUREL – three times more likely, in fact (i.e., 75% chance of YANNY). Another researcher claims there is no preference (i.e., 50% chance of YANNY). Because both of these claims are quite specific, let us call the latter claim the "point null hypothesis" and the former claim the "competing-point hypothesis." Further suppose that the researchers collected data from 30 teenagers and found that $Y = 20$ heard YANNY while $L = 10$ heard LAUREL.

We can now calculate how strongly either hypothesis had predicted this outcome. In both cases, the probabilities are easily obtained with the binomial formula. For the point null hypothesis,

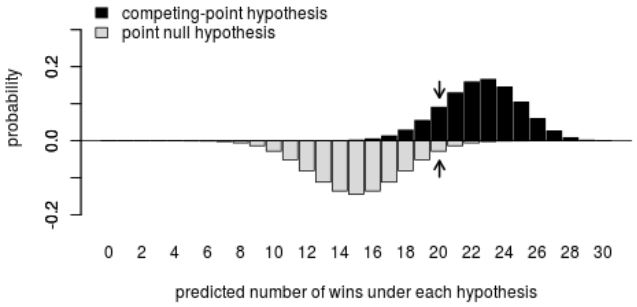


Fig. 1. An illustration of the computation of the Bayes factor using visualizations from Etz et al. (2018). Shown are the predictive distributions of two competing hypotheses: one 'point null hypothesis' under which some event happens with a 50% probability (lighter bars, upside-down) and one 'competing-point hypothesis' under which it happens with a 75% probability (darker bars, upright). The predictions are for an experiment with 30 trials in which the event can either occur or not. The arrows indicate the case where the event happened 20 times out of 30. This outcome is predicted with a probability of .0280 under the point null hypothesis and with a probability of .0909 under the competing-point hypothesis. The Bayes factor between these two models is simply the ratio of these probabilities, here 3.247.

the probability is

$$\begin{aligned} P(Y = 20, L = 10|\text{point null}) &= \binom{Y+L}{Y} \times 0.50^{Y} \times 0.50^{L} \\ &= \binom{30}{20} \times 0.50^{20} \times 0.50^{10} \\ &= 0.0280, \end{aligned}$$

while for the competing-point hypothesis, the probability is

$$\begin{aligned} P(Y = 20, L = 10|\text{competing}) &= \binom{30}{20} \times 0.75^{20} \times 0.25^{10} \\ &= 0.0909. \end{aligned}$$

Hence, the competing-point hypothesis is supported more strongly by these observations by a factor of $.0909/.0280 = 3.247$. Figure 1 illustrates this comparison graphically.

A Bayes factor of 3.247 is generally considered to be only weak evidence (Wagenmakers, Marsman, et al., 2018; Kass & Raftery, 1995; Goodman, 1999b). Combined with a perfectly ambivalent prior (50% on either claim, or a prior ratio of 1), this Bayes factor brings us to a posterior probability of only about 76% in favor of the competing-point hypothesis – not a very high probability.

Bayesian estimation. In the Bayesian framework, the distinction between testing and estimation is less clear-cut than it is in the frequentist framework. It is useful to think of the two practices as the ends of a continuum. The continuum captures how many possible states of the world are being considered. If we are interested in the probability θ that a coin comes up heads, we might limit our possible hypotheses to $A : \theta = 0.50$ and $B : \theta = 1.0$. This has all the bearing of a testing scenario. Alternatively, we might consider $A : \theta = 0.0$, $B : \theta = 0.5$, $C : \theta = 1.0$, which still has the appearance of a testing context. However, if we permit that θ might be any of $(0.01, 0.02, \ldots, 0.99, 1.00)$, then it is less clear if we are estimating a parameter or selecting between 101 models. If we allow θ to be anywhere from 0.5 to 1.0, or from 0.0 to 1.0, then we are more obviously dealing with an estimation scenario. Hence, the Bayesian estimation task can be seen an extension of Bayesian hypothesis testing, in which truth value is reallocated among many possible parameter values.

An example of Bayesian estimation Behavioral researchers rarely have strong quantitative theories that permit statements such as "the probability that a teenager will hear YANNY is 75%." Instead, much behavioral research is conducted in a context of discovery: we seek to quantify effect sizes or estimate parameters (Cumming, 2014), rather than to discriminate between a set of competing accounts. The simplest way to illustrate Bayesian estimation is to use *conjugate* families of distributions. A prior distribution and likelihood for the data are said to be conjugate when the resulting posterior distribution is in the same class as the prior distribution. For example, updating a normal prior distribution with normally distributed data results in a normal posterior distribution with a new mean and standard deviation. In what follows we will illustrate conjugate updating and estimation using the conjugate family that includes beta prior distributions and binomial likelihood functions.

Continuing with the bistable perception phenomenon, we might consider two researchers interested in estimating the fraction ϕ of teenagers who hear YANNY versus LAUREL. The two researchers, independently from one another, retrieve the data from the previous example (a sample of 30 teenagers, 20 of whom hear YANNY) and use it to estimate ϕ. However, the two researchers differ in their prior conceptions of this proportion. Researcher 1 believes that teenagers are relatively homogeneous, and will to a large extend either all hear LAUREL or all hear YANNY (i.e., ϕ will be close to 0% or 100%). Researcher 2 believes that the population is more likely to be split, and ϕ is most likely close to 50%. This difference in prior beliefs is displayed with the dashed lines in both panels of Figure 2. Researcher 1's prior is well captured by a $beta(0.5, 0.5)$ distribution while Researcher 2's prior is best described with a $beta(2, 2)$ distribution.

The $beta(a, b)$ prior is defined as

$$P(\phi|a, b) = \phi^{a-1}(1 - \phi)^{b-1} \times C_{\text{prior}}$$

where C_{prior} is a ϕ-independent scaling parameter that ensures the function describes a proper distribution (i.e., one whose mass totals 1). The interpretation of the parameters a and b is revealed when we compute the posterior distribution.

To obtain the posterior distribution of ϕ, we multiply the prior with the binomial likelihood of the data, which is:

$$P(Y, L|\phi) = \phi^{Y} \times (1 - \phi)^{L} \times C_{\text{likelihood}},$$

where we again capture all factors that do not contain ϕ into a single scaling parameter. The posterior is then, by Bayes' theorem:

$$\begin{aligned} P(\phi|Y, L) &= P(Y, L|\phi) \times P(\phi|a, b) \times C_{\text{post}} \\ &= \left[\phi^{Y} \times (1 - \phi)^{L}\right] \times \left[\phi^{a-1}(1 - \phi)^{b-1}\right] \times C_{\text{post}}. \end{aligned}$$

Some algebraic rearrangement yields

$$\begin{aligned} P(\phi|Y, L) &= \phi^{a-1+Y}(1 - \phi)^{b-1+L} \times C_{\text{post}} \\ &= \phi^{a'-1}(1 - \phi)^{b'-1} \times C_{\text{post}}, \end{aligned}$$

where we have first collected all scaling factors into a new factor C_{post} and then introduced the updated parameters $a' = a + Y$ and $b' = b + L$. This rearrangement illustrates the conjugacy of the beta prior and the binomial likelihood: the posterior distribution of ϕ again follows a $beta$ distribution. Due to the conjugacy, adding

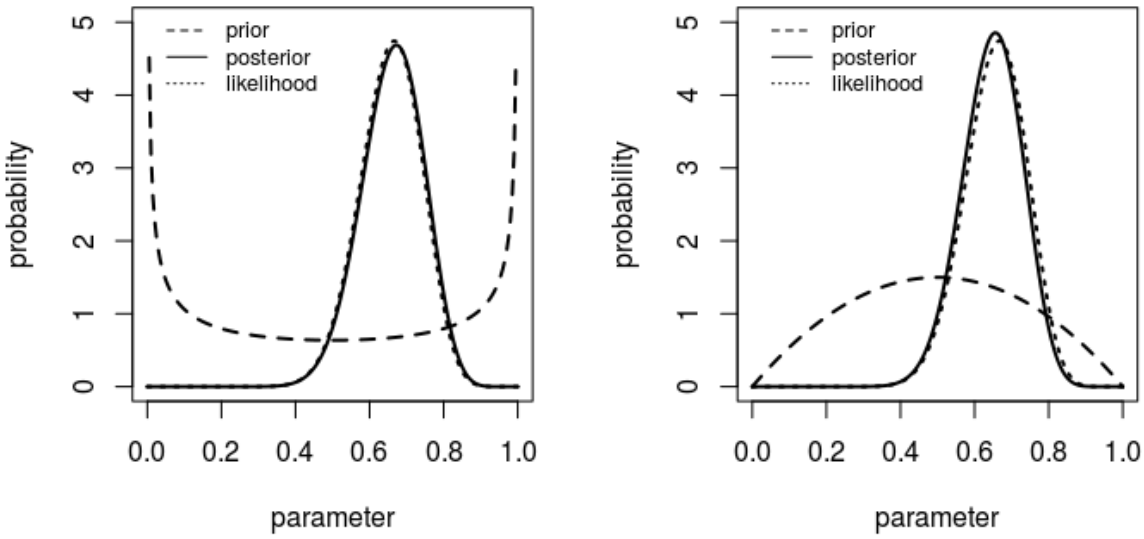


Fig. 2. Prior and posterior distributions of two independent researchers seeing the same data. Researcher 1 (left) expects the ratio parameter to be extreme; either close to 0 or close to 1. Researcher 2 (right) expects the parameter to be closer to .5. Both researchers observe the same data: a sample of 30 yields 20 observations in one category and 10 in the other. When these data are factored in, the differences between the two researchers dissipate – even the dramatic difference in priors is easily overwhelmed by a modest amount of data.

further observations is easy: simply increment a' with the number of new YANNY observations and increment b' with the number of new LAUREL observations.

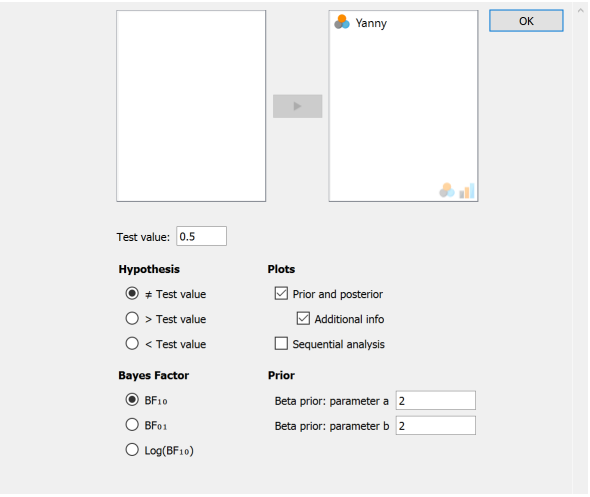
The way in which a and b absorb the number of observations of each type also reveals an interesting interpretation of these parameters: the value of a and b *prior to seeing the data* can be interpreted as the number of times the researcher had already observed YANNY and LAUREL occurrences (respectively). The "effective prior sample size" $a + b$ expresses the strength of the available prior information. Note that this remains true if we were to add a second batch of observations: the effective sample size after the first batch is $a' + b'$, and upon observing new data (Y_2, L_2) the new parameters would be $a'' = a' + Y_2$ and $b'' = b' + L_2$. Updating the probability density with new data is a matter of incrementing the parameters of the distribution and (in this case) does not require complex mathematical exercises.

It is also easy to see how the data will quickly overwhelm the prior: Researcher 1 has the equivalent prior information of one observation and Researcher 2 has the equivalent of four observations. These quickly pale when incremented by 30 observations.

With the parameters of the posterior distribution $P(\phi|Y, L)$ in hand, we can now compute a number of interesting quantities, such as the most likely value of ϕ (the posterior mode): $\frac{a'-1}{a'+b'-2}$, which is $.6724$ for Researcher 1 and $.6563$ for Researcher 2. We could also compute the posterior probability that $\phi > .5$, which is 0.9670 for Researcher 1 and and 0.9599 for Researcher 2. Here, again, the dramatic difference in the prior distribution makes little difference in the ultimate quantities of interest. Both researchers conclude that ϕ is close to 2/3rd and is very likely greater than one half.

Software. Simple Bayesian methods require no more computational effort than standard approaches. To illustrate the ease of Bayesian computation we will recreate the estimation analysis above using the software JASP (JASP Team, 2018). JASP is a statistical program with a graphical user interface, meaning no knowledge of scripting or coding is necessary to perform a Bayesian analysis. We have created a data file containing 20 YANNY and 10 LAUREL responses, available for download at `https://osf.io/ksvdp/`. If we open this file in JASP and select "Bayesian binomial test" from the frequencies drop-down button we are brought to an options menu. In this menu we can specify that the success counts are in the YANNY column of the data, and also specify that we wish to use a beta(2,2) prior distribution for the success parameter (JASP uses θ for our ϕ). JASP will then generate the results of the Bayesian analysis in the right-most panel, in the form of a plot of the prior and posterior distributions of the probability of success, which we present in Figure 3 (right panel). Note that this posterior distribution exactly matches that from the right panel of Figure 2. JASP provides a 95% credible interval for the parameter, which in this case ranges from .482 to .796.

Whereas Bayesian analysis of common designs is possible in software such as JASP, Bayesian calculation for complex modeling is often computationally intensive. Fortunately, user-friendly tools for Bayesian analysis have emerged in recent years and have incorporated into virtually all standard statistical software (e.g. Stata, SPSS, SAS, and R), as well as in new specialized software specifically for Bayesian analyses (e.g. JASP). Introductions to the use of general-purpose Bayesian software can be found in Matzke, Boehm, and Vandekerckhove (in press); van Ravenzwaaij, Cassey, and Brown (in press) and Wagenmakers, Love, et al. (2018), but more tutorials appear on a regular basis.



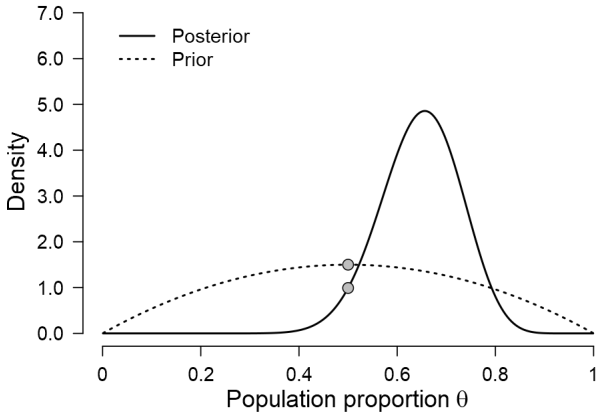


Fig. 3. A Bayesian binomial test in JASP. **Left:** The interface for the Bayesian binomial test. **Right:** Default output from the test applied to the YANNY/LAUREL data set.

Statistics and research integrity

The recent crisis of confidence in psychological science (Pashler & Wagenmakers, 2012) has led to repeated calls for better statistical methods, better study designs, and better social incentive structures. The link of the analytic and inferential issues above to research integrity is fairly direct, in that research integrity, writ large, is the fidelity of the scientific process and resulting conclusions to the truth. If it can be shown that the manner in which studies are done, data is analyzed, or conclusions drawn is likely to systematically deviate from the truth, then by definition we have a challenge to research integrity. There has been an evolution of professional and scientific norms within the behavioral sciences that have exaggerated and reified some of the most unfortunate misconceptions and misuses to which frequentist methods are subject; in particular that statistical significance is an arbiter of truth, that the credibility of a claim can be assessed without considering the prior probability of the claim being true, and that non-significant studies are uninformative. These misconceptions and misuses include but are not limited to:

- ✗ The false belief that a single experiment is sufficient to prove a theory;
- ✗ Resistance to replication of important experiments;
- ✗ The difficulty of publishing research replications;
- ✗ Strong selection pressure at journals for significant findings (i.e., publication bias);
- ✗ Professional advancement dependent on publications in highly cited journals, with acceptance influenced by statistical significance;
- ✗ Widespread use of "p-hacking";
- ✗ Failure to use sample sizes that correspond to a priori plausible or scientifically important effect sizes;
- ✗ Failure to share data;
- ✗ Failure to prepare, share or publish research protocols.

A number of the practices above, particularly those related to replication and publication, can be related directly to extreme versions of frequentist philosophy. But they also mean, because so many of these practices are deeply entrenched and indeed, institutionalized, that moving to Bayesian methods cannot solve all of the challenges to scientific integrity in the social and behavioral sciences. Nevertheless, they represent a critical piece of a multifaceted strategy that the behavioral sciences must adopt if its findings and claims are to be regarded as reliable.

It is not critical for the entire analytic approach of the behavioral sciences to move to Bayesianism for many of these changes to be made. it is interesting to look at the methodologic evolution within biomedicine, which has not given up frequentist methods, but has avoided some of the particularly egregious practices seen in the behavioral sciences. Most importantly, in clinical biomedicine there is a culture of disbelief in single studies, particularly small ones, and knowledge is not regarded as established until a sufficiently large collection of studies generates convincing evidence, as shown in systematic reviews and meta-analyses. Underpowered studies are strongly disfavored at the major journals, and strong emphasis is put on estimation together with testing, particularly for nonsignificant studies with equivocal findings. Study protocols are now routinely requested by the major journals, and the law requires that randomized trials must be preregistered at clinicaltrials.gov within 21 days from when the first patient is enrolled, and the RCT results reported in clinicaltrials.gov regardless of outcome, with government penalties for noncompliance.

This is not to suggest that the field of clinical research has solved or avoided all of the issues of research integrity that are now plaguing behavioral and social science, but it went down similar paths of awareness and reform starting 3-4 decades ago and adopted a number of practices that have blunted some of the worst potential effects of improper understanding or implementation of frequentist philosophy and methods. That said, problems remain. Interestingly, many of the innovations currently being suggested for the behavioral sciences are now being adopted within the biomedical sciences as ways to accelerate progress there, particularly the move to open science.

Changes in the practices of an entire discipline require far more than a change in analytic philosophy; these must be accompanied by changes in education, professional norms and expectations, funding, other rewards, and publication. But understanding how the analytic philosophy contributed to some of these practices is critical to make the right changes and thereby improve the trust that those inside or outside the behavioral sciences put in the field and its findings.

Acknowledgements

This work was supported by National Science Foundation grants #1534472 and #1658303 to JV and National Science Foundation Graduate Research Fellowship Program #DGE-1321846 to AE.

References

Aldrich, J., et al. (2008). R. A. Fisher on Bayes and Bayes' theorem. *Bayesian Analysis*, *3*(1), 161–170.

Altman, D., Machin, D., Bryant, T., & Gardner, M. (2013). *Statistics with confidence: confidence intervals and statistical guidelines*. John Wiley & Sons.

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

Carnap, R. (1950). *Logical foundations of probability*. Chicago: The University of Chicago Press.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*(5), 553.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-?29.

Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, 1–26.

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

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, 2515245918773087.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian analysis*, *1*(1), 1–40.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, *33*, 503–513.

Fisher, R. A. (1971). *The design of experiments* (7th ed.). New York: Hafner.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p-values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, *292*(6522), 746–750.

Gelman, A., & Nolan, D. (2002). You can load a die, but you can't bias a coin. *The American Statistician*, *56*(4), 308–311.

Ghosh, J. K. (Ed.). (1988). *Statistical information and likelihood. A collection of critical essays by Dr. D. Basu*. Springer New York. doi:

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.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio–Economics*, *33*, 587–606.

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, *137*, 485–496.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine*, *130*, 995–1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, *130*, 1005–1013.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341), 341ps12–341ps12.

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, *78*(12), 1568–1574.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337–350.

Hald, A. (2008). *A history of parametric statistical inference from Bernoulli to Fisher, 1713-1935*. Springer Science & Business Media.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*, 1–20.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164.

JASP Team. (2018). *JASP (Version 0.9)[Computer software]*. Retrieved from `https://jasp-stats.org/`

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan & Co.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Matzke, D., Boehm, U., & Vandekerckhove, J. (in press). Bayesian inference for psychology. Part III: Parameter estimation in nonstandard models. *Psychonomic Bulletin & Review*.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, *236*, 333–380.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 175–240.

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.

Oakes, M. W. (1986). *Statistical inference*. Epidemiology Resources.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Rouder, J. N., & Morey, R. D. (in press). Teaching Bayes' theorem:

Strength of evidence as predictive accuracy. *The American Statistician*.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

Schervish, M. J. (1996). P values: What they are and what they are not. *The American Statistician*, *50*, 203–206.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (in press). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p–values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, *38*, 715–731.

Zabell, S. (1989). R. A. Fisher on the history of inverse probability. *Statistical Science*, 247–256.