Journal of Education and Learning; Vol. 6, No. 4; 2017
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Beyond P values and Hypothesis Testing: Using the Minimum Bayes Factor to Teach Statistical Inference in Undergraduate Introductory Statistics Courses

Robert Page[1] & Eiki Satake[2]

[1] Department of Mathematics, Framingham State University, Framingham, Massachusetts, USA

[2] Institute for Liberal Arts and Interdisciplinary Studies, Emerson College, Boston, Massachusetts, USA

Correspondence: Robert Page, Department of Mathematics, Framingham State University, 100 State St, Framingham, Massachusetts, 01701, USA. Tel: 1-508-626-4773. E-mail: rpage@framingham.edu.

Received: July 6, 2017 Accepted: July 19, 2017 Online Published: August 2, 2017

doi:10.5539/jel.v6n4p254 URL: http://doi.org/10.5539/jel.v6n4p254

Abstract

While interest in Bayesian statistics has been growing in statistics education, the treatment of the topic is still inadequate in both textbooks and the classroom. Because so many fields of study lead to careers that involve a decision-making process requiring an understanding of Bayesian methods, it is becoming increasingly clear that Bayesian methods should be included in classes that cover the P value and Hypothesis Testing. We discuss several fallacies associated with the P value and Hypothesis Testing, including why Fisher's P value and Neyman-Pearson's Hypothesis Tests are incompatible with each other and cannot be combined to answer the question "What is the probability of the truth of one's belief based on the evidence?" We go on to explain how the Minimum Bayes Factor can be used as an alternative to frequentist methods, and why the Bayesian approach results in more accurate, credible, and relevant test results when measuring the strength of the evidence. We conclude that educators must realize the importance of teaching the correct interpretation of Fisher's P value and its alternative, the Bayesian approach, to students in an introductory statistics course.

Keywords: Bayesian statistics, conditional probability, hypothesis testing, minimum Bayes factor, P value, statistical inference

1. Introduction

In recent years, Bayesian statistics has gone from being a controversial theory on the fringe of mainstream statistics to being widely accepted as a valuable alternative to more common classical approaches. Indeed, Bayesian methods have become increasingly common in a range of fields, including marketing, economics, school assessment, nuclear waste disposal, medicine, and law. They have, for example, permeated all of the major areas of medical research, from clinical trials to survival modeling and decision-making about the use of new technologies (Ashby, 2006).

Interest in Bayesian statistics has also been growing in statistics education. Increasingly, though not commonly, elementary statistics texts today are introducing Bayesian methods using Bayes' Rule within a section on conditional probability (see for example, De Veaux, Velleman, & Bock, 2008; Sullivan, 2007; Triola, 2007; Larson & Farber, 2006; Bluman, 2004). This is a significant step forward from what was observed more than a decade ago when Bayes' rule and Bayesian methods were rarely covered or treated as optional topic at best (Satake, Gilligan, & Amato, 1995). A few texts, such as Berry (1996), Bluman (2004), and Larson and Farber (2006), discuss the theorem in great detail with examples and exercise problems. Unfortunately, for the most part, our informal survey reveals that the treatment of the topic in textbooks and the classroom is often sparse and inadequate. Since students in many fields today are likely to enter careers that will include a decision-making theory that requires a good working knowledge of Bayesian methods and ideas, the importance and practical necessity of integrating the Bayesian approach in classes involving P value and Hypothesis Testing approaches is becoming increasingly clear (Satake & Amato, 2008).

In this paper, we will discuss the following:

- Several fallacies of frequentist statistical methods, specifically P value and Hypothesis Testing.

- The use of the Minimum Bayes' Factor (MBF) for statistical inference as an alternative to the frequentist. Methods.
- The reason why the Bayesian approach provides more accurate, credible, and relevant test results than the frequentist statistical methods when we measure the strength of the evidence.
- The reason why Bayesian methods should be included in an introductory statistics course.

Typically, in an introductory statistics course, the primary procedure of statistical inference presented is the *combined method* of Fisher's P value approach and Neyman-Pearson's Hypothesis Test approach. In fact, the method is the most widely used, even beyond undergraduate statistics classes. Furthermore, the combined method is considered to be a so-called *best compromised* method by many researchers and practitioners in the various fields, and it has become the standard method in classroom instruction and scientific journal article writings. Unfortunately, the P value and Hypothesis Test approaches are conceptually different and incompatible with each other (Goodman, 1999, 2005). Therefore, it is illogical to combine the two methods and use them as a single coherent approach to statistical inference. In fact, the founders of each method, Fisher, Neyman and Pearson, greatly disagreed with the others' approach to statistical inference and disliked one another personally (Goodman, 1999). Based on an informal survey of widely used textbooks (Sullivan, 2014; Triola, 2014; Brase C. H. & Brase C., 2015) almost none cover the history of the debate. Indeed, most treat these two incompatible approaches as a single, coherent method.

The main objective of a statistical inference is to best estimate the characteristics of the parameter of a target population, based on what was actually observed in a sample drawn from the population. In other words, a statistical inference seeks to make an inference about a general outcome (population) from a specific outcome (sample). This process is called *inductive reasoning*. Therefore, what one eventually must answer is, "how credible is one's hypothesis, based on what was actually observed?" Contrast this with the question, "how credible is one's sample data, based on the assumption that the null hypothesis is true?" Answering the latter question is known as *deductive reasoning*. Using P values, Hypothesis Tests, or the two combined to answer the main question of statistical inference is in many cases irrelevant and meaningless, because one is using deductive reasoning to answer an inductive-natured question. In other words, one is not interested in measuring the credibility of the data, but rather in measuring the truth of the hypotheses based on the data. The deductive quantitative tools such as P values and Hypothesis Tests do not serve well in answering inductive-natured statistical inference questions. One needs an inductive tool to answer questions directly related to statistical inference. That is one of the main reasons why students and even instructors in a statistics course often misinterpret the results of their statistical analyses (Gigerenzer et al., 2004).

Regardless of the type, statistical methods were originally intended to measure the strength of *evidence* for drawing an accurate and reliable conclusion about the population parameters of interest, based on a sample. Unfortunately, in most cases the term "evidence" is not correctly interpreted by students or even correctly defined by instructors. A statistical definition of evidence, according to Goodman and Royall (1998), is given as "a property of data that makes us alter our beliefs about how the world around us is working". In short, what one means by "measuring the strength of evidence" must lead to a determination of the truth of one's belief (or hypothesis) in the end, based on what is actually observed. It is a simple fact that neither of the frequentist methods, even when combined, can accomplish this task. In the next two sections of this paper we will discuss three frequentists approaches to statistical inference, followed by an introduction to an alternative inductive-natured statistical method called *Bayesian methods*.

2. Fisher's P Value

> ***Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.***
>
> **–R. A. Fisher (1937)**

This approach consists of two components, namely (1) a null hypothesis (H_0), and (2) a test statistic (derived from an observed value under the null hypothesis distribution). Graphically, in a right-sided test, the P value is the area under the null hypothesis distribution (H_0 distribution) curve from the test statistic (denoted by x) and beyond (See Figure 1). For a left-sided or two-sided test, the concept is similar.

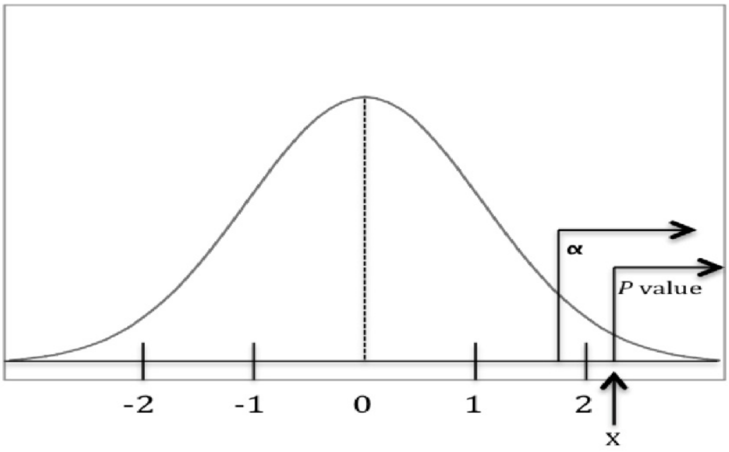

Figure 1. The bell-shaped curve represents the probability of every possible outcome under the null hypothesis. Both α (the type 1 error rate) and the *P value* are "tail areas" under this curve. The tail area for α is set before the experiment, and a result can fall anywhere within it. The *P value* tail area is known only after a result is observed, and, by definition, the result will always lie on the border of that area

As seen in Figure 1, the correct interpretation of the P value is the probability of obtaining a test statistic equal to, *or more extreme than*, the observed value under the assumption that the null hypothesis is true. The basic rule for decision-making is that if any sample data has a corresponding P value less than an ordinary benchmark of 0.05 (a typical choice in the behavioral and social science research, as Fisher stated), the result is considered to be *statistically significant*.

Although there are many articles and texts involving P value fallacies in the literature (e.g., Goodman, 1999, 2005; Matthew, 2001; Maxwell & Satake, 2010; Satake & Murray, 2014; Winkler, 2001; etc.), the following two fallacies are the most serious in terms of why the P value is neither suitable for measuring the strength of evidence nor for drawing an accurate conclusion about the population of interest. First, the correct definition of the P value is made difficult for interpretation of the results because it is not part of any formal procedure of statistical inference. Specifically, by including the probability of data that is "more extreme" than actual observed data, the conclusion is made imprecise. Later, we will show just how much the P value can overstate the amount of evidence against H_0. Second, the P value does not calculate what the investigator wants to know—namely, the credibility of the research hypothesis (H_a, also known as the alternative hypothesis) in light of the observed data. As we mentioned before, the inability to calculate such a probability makes the P value an inappropriate tool when conducting statistical inference. Symbolically, Fisher's P value is calculated as follows:

$$\text{P value} = P\,[(x \geq \theta) \mid (H_0 \text{ is true})] \tag{1}$$

where x = the test statistic for an arbitrary sample and θ is the observed value.

Of course, the probability that an investigator ultimately wishes to find is:

$$P\,[(H_0 \text{ is true}) \mid (x = \theta)]. \tag{2}$$

The common error made by investigators is to interpret the Fisher's P value as the latter probability, *which it is not*.

3. Neyman-Pearson's Hypothesis Tests

> ***…no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.***
>
> ***But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate (conflicting) hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong.***
>
> **–J. Neyman and E. Pearson (1933)**

Historically, the two main approaches have been Fisher's P value, as discussed previously, and Neyman-Pearson's hypothesis test. The major conceptual differences between them are listed below.

- Fisher's P value was originally intended for a flexible inferential measure in a single experiment, whereas Neyman-Pearson's Hypothesis Test was a rule of *behavior*, not inference, in the hypothetical long run experiment (Goodman, 1999).
- Fisher's P value indicates a tail area starting from a point calculated by the sample data, whereas Neyman-Pearson's Hypothesis Test uses the pre-trial fixed "error rates", such as Type I (α) and Type II (β) errors, and the region for rejecting H_0 before the experiment begins.
- Fisher's P value measures the amount of evidence against only a null hypothesis, whereas Neyman-Pearson's Hypothesis Test states two conflicting hypotheses and measures the strength of evidence based on the pre-trial error rates and the critical region. The decision to reject H_0 is contingent upon whether or not the test statistic falls into the critical region.

Philosophically, Neyman and Pearson held a strong position on the use of deductive reasoning (from *general to specific*), whereas Fisher rejected their deductive and mechanistic approaches to statistical inference. Although both the P value and pre-trial Type 1 error rate (α) are *tail area* probabilities under the null hypothesis distribution, a P value is a post-experiment probability calculated from the actual sample data, while α is a fixed pre-experiment error rate selected prior to sample data analysis. Therefore, while it is natural that statistical novices fail to distinguish between the two, they are conceptually different probability values.

Importantly, Neyman-Pearson's Hypothesis Test lacks the ability to accomplish the following tasks:

1) Measure the strength of evidence accurately.

2) Assess truth of a research hypothesis from a single experiment (Goodman, 1999).

In other words, Neyman-Pearson's Hypothesis Test only gives conventions for dichotomous results such as "Reject H_0" or "Do not reject H_0" based on preset long-term error rates. One cannot conclude how justified a particular decision is after an experiment, nor how credible a hypothesis is prior to the experiment. More importantly, one cannot assign a probability value to how much the sample data supports a given hypothesis, which is the essential meaning of *measuring the strength of evidence*. One is only able to conclude, after a single experiment, that the final decision was made by a standard procedure that controls error frequencies in the long run (Dienes, 2011).

Fisher's P value and Neyman-Pearson's Hypothesis Test are fundamentally incompatible. However, the two approaches were blended into a single approach, which has mistakenly become a standard approach to teaching statistical inference (Maxwell & Satake, 2010).

4. The Combined Method

In the previous section, we noted that neither Fisher's P value Test nor Neyman-Pearson's Hypothesis Test can accurately measure the strength of evidence to calculate the probability of the truth of one's belief. Specifically, Fisher's P value calculates the rarity of the observed value and "more extreme" values, given that the null hypothesis is true. Neyman-Pearson's Hypothesis Test also prevents one from calculating the probability of the truth of one's belief. Thus one can only obtain the dichotomous decision, namely "Reject H_0" or "Do not reject H_0", without any specific numerical probability indicating how strongly the evidence is for or against H_0. Therefore, each method needs the other to compensate for its respective limitations. Naturally, researchers were tempted to combine the two methods. Goodman (1999) commented on this issue as follows:

> The hypothesis tests approach offered scientists a Faustian bargain—a seemingly automatic way to limit the number of mistaken conclusions in the long run, but only by abandoning the ability to measure evidence and assess truth from a single experiment. It is doubtful that hypothesis tests would have achieved their current degree of acceptance if something had not been added that let scientists mistakenly think they could avoid that trade-off. That something turned out to be Fisher's "P value", much to the dismay of Fisher, Neyman, Pearson, and many experts on statistical inference who followed (p. 999).

The combined method operates as follows. First, and before the experiment, an investigator sets the Type 1 error of his choice (α, usually 5%) and power (β, usually a value exceeding 80%). Then the investigator calculates a P value from a sample and rejects the null hypothesis if the P value is less than the preset Type 1 error rate, i.e., reject H_0 if $p < \alpha$. This combined method procedure appears, completely deductively, to associate a probability with the null hypothesis, like Fisher's P value method, within the context of a method that controls the chance of errors. On this issue, Goodman (1999) states:

> The key word here is "probability", because a probability has an absoluteness that overwhelms caveats that it is not a probability of truth or that it should not be used mechanically (p. 1000).

As a result, the combined method is currently considered by many scientists, educators, practitioners, and others as a "more scientific and objective" method for measures of the strength of evidence in statistical inference (Satake, 2014). Contrary to what many believe, the combined method still cannot answer the question "What is the probability of the truth of one's belief based on the evidence?" If the main goal of statistical inference is to answer that question, then the combined method is not sufficient. Fortunately, there is an alternative approach, unknown to many statistical novices as well as educators, that *can* answer the question. We refer to such an approach as the Bayesian method.

5. Prelude to Bayesian Methods: Several Key Points on Frequentists Methods

Goodman (2008) pointed out that the correct interpretation of a P value is made difficult, even among skilled researchers, because it is not part of any formal system of statistical inference. Therefore, the P value's inferential meaning is widely, and often wildly, misunderstood. For instance, Oaks (1986) asked 70 academic psychologists to interpret $p < 0.01$ as a probability, and only 8 out of 70 participants interpreted it correctly. The results of other surveys of professionals in the social and clinical sciences indicate a similar, apparently widespread, misunderstanding of the P value and Hypothesis Test approaches (Mittag & Thompson, 2000; Nelson, Rosenthal, & Rosnow, 1986; Dar, Serlin, & Omer, 1994). Given that many academics and researchers have spent countless hours studying and teaching the P value and Hypothesis Test approaches in statistics courses at all levels, the natural question arises: Why are we not successful in teaching these concepts?

Kline (2004) warrants two factors as the reason. First, the P value and Hypothesis Test approaches lack transparency as inference systems. Pollard and Richardson (1987) noted that it is difficult to explain the logic of these methods and dispel confusion about them. Second, it is a general human weakness in reasoning with conditional probabilities, especially those best viewed from a relative frequency perspective (Anderson, 1998).

Furthermore, many people mistakenly believe that the incorporation of the P value into the Hypothesis Test framework is the right way to measure the strength of statistical evidence. Several misconceptions have been created in statistics education as a result. Before we proceed to the main theme of the paper, we will itemize some important points regarding the confusion about the frequentist methods.

Facts about the P value, Hypothesis Test, and Combined approaches that are most commonly misunderstood are summarized below (Goodman, 1999).

Fisher's P value approach is intended to measure the strength of the evidence using a flexible inferential measure. Neyman-Pearson's Hypothesis Tests are rules of behavior, not inference. Neyman and Pearson held that Fisher's inductive reasoning was an illusion and that the only meaningful parameters of importance in an experiment were constraints on the number of statistical "errors" the researcher would make, which should defined before an experiment (α). Fisher rejected mechanistic approaches to inference, believing in a more flexible, inductive approach to science, such as a mathematical likelihood.

The two methods are philosophically and conceptually different. That is to say, it is incorrect to conclude: "Reject H_0 if $p < \alpha$" or "Do not reject H_0 if $p \geq \alpha$". Because of Fisher's P value's resemblance to the pretrial error rate (α), it was absorbed into the Hypothesis Test framework. This created two illusions: first, that an "error rate" could be measured after an experiment, and second, that this post-trial "error rate" could be regarded as a measure of inductive evidence. However, α is a pre-specified error rate (before studying a sample) while a P value is the conditional probability of finding a sample with data equal to or more extreme than the actual observed sample data, *given that H_0 is true*. Furthermore, α is the probability of a set of future outcomes, represented by the "tail area" of the null distribution. Implicit in the concept is that one doesn't know which of those outcomes will occur. The tail area represented by Fisher's P value is quite different; we know the outcome, and by definition it lies exactly on the border of the tail area. The "error rate" concept requires that a result can be anywhere within the tail area, which is not the case with Fisher's P value. An error rate interpretation of the P value implies partial ignorance about the results, but if we had such ignorance, we could not calculate the P value. Another factor that makes the error rate interpretation of P values problematic is that they are always calculated conditionally (given that H_0 is true).

Fisher's P value is calculated based on a single experiment, whereas the α value is set before an experiment and designed for the long run.

As a practicing scientist, Fisher had an abiding interest in creating an objective, quantitative method to aid the process of inductive inference, i.e., drawing conclusions from observations. He did not believe that the use of Bayes' formula to convert *prior probabilities* (the probability, before an experiment, that H_0 is true) to *posterior probabilities* (the probability, after an experiment, that H_0 is true) was justified in scientific research, where prior probabilities are usually uncertain. He ultimately proposed the inferential methods that did not require prior probabilities of hypotheses. Fisher's P value approach is one of those inferential methods. The statement "One can never accept H_0, only fail to reject it" was a feature, not of Neyman-Pearson's Hypothesis Tests, but of Fisher's Significance (P value) Tests. It is important to note that Fisher's P value approach had no H_a. On the other hand, there was no measure of evidence in the Neyman-Pearson Hypothesis Tests, though some continue to interpret it as if there were.

Neyman and Pearson held that the best one can do with deductive probability theory is to use a rule for statistically dictated behavior, which they claimed would serve well in the long run. They also stated that there does not exist reasoning called "Inductive". If it were to be used properly, they believed the term should be called "Inductive behavior".

There is no definitive measure tool to calculate the strength of the evidence in the Hypothesis Tests, although many people used Fisher's P value as such a tool. The results are limited to either "Reject H_0 under a given Type 1error" or "Fail to reject H_0 under the given Type 1 error".

Even Fisher thought that null hypothesis testing was the most primitive type in a hierarchy of statistical analyses and should be used only for problems about which we have very little knowledge or none at all (Gigerenzer et al., 1989). Interestingly, Fisher also thought that using a 5% level of significance indicated a lack of statistical thinking; he believed that the level of significance should depend on the context of the investigation.

6. Bayesian Perspectives: How to Measure the Strength of the Evidence

Many of the fallacies about the P value and Hypothesis Test approaches are the result of researchers, teachers, and students wishing to determine posterior probabilities without using Bayesian techniques. Unfortunately, frequentist statistical methods only calculate the probability of occurrence of values equal to or more extreme than the observed data value under the assumption of H_0 being true. Additionally, even if the P value were correctly used, frequentists still neglect to include the "more extreme" values when they draw a conclusion. What is desired is to calculate the probability of truth for a hypothesis based on data that were actually observed in a sample, while excluding the "more extreme" hypothetical outcomes that may be observed in the long run. Such a probability can only be obtained using Bayesian methods.

In Bayesian methods, there are three main components namely *prior odds of the null hypothesis* (a subjective component stated by an investigator before seeing data), *Bayes' factor* (an index that measures the strength of actual evidence), and *posterior odds of the null hypothesis* (a more objective component derived through a combination of prior odds and the Bayes' factor). The posterior odds reflect the probability of truth of the null hypothesis (Maxwell & Satake, 2010). A simplified process of the three components is illustrated below (See Figure 2).

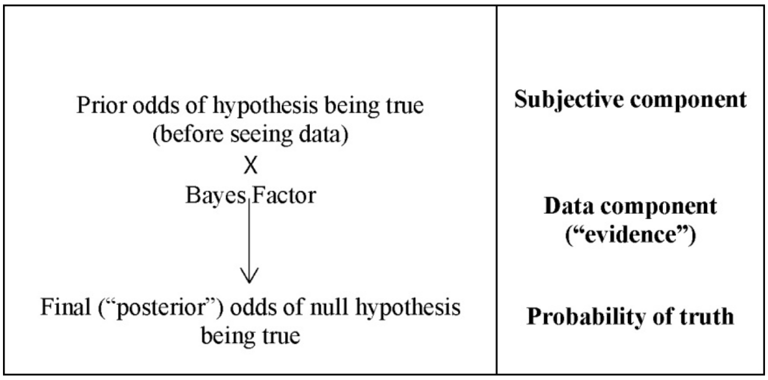
Prior odds of hypothesis being true
(before seeing data)
X
Bayes Factor
Final ("posterior") odds of null hypothesis being true
Subjective component
Data component ("evidence")
Probability of truth

Figure 2. Bayes Theorem, in words

Bayesian methods provide a proper way to measure and to combine evidence. They are symbolically written as follows:

$$\frac{P(H_0|\text{Data})}{P(H_a|\text{Data})}=\frac{P(H_0)}{P(H_a)}\bullet\frac{P(\text{Data}|H_0)}{P(\text{Data}|H_a)} \tag{3}$$

In words, the left-hand side of the equation above represents the posterior odds ratio, and the right-hand side is the product of the prior odds ratio and Bayes' factor (aka the likelihood ratio). Because Bayes' factor is a factor of the posterior odds ratio, from the Bayesian standpoint, the better the hypothesis predicts the data, the stronger the evidential support for that hypothesis.

There are two main differences between Fisher's P value and Bayes' factor. First, the calculation of Fisher's P value involves both observed data and "more extreme" hypothetical outcomes, whereas one calculates Bayes' factor using only observed data. Second, Fisher's P value is calculated in relation to only the null hypothesis, whereas Bayes' factor is calculated in relation to both the null and alternative hypotheses. In other words, Bayes' factor is a comparison of how well two competing hypotheses (H_0 and H_a) predict the data. The hypothesis that predicts the actual observed data better is the one that is said to have more evidence supporting it (Goodman, 1999).

Furthermore, Bayes' factor can be interpreted as the mathematical ratio of the credibility of evidence under two conflicting conditions (the null hypothesis versus the alternative hypothesis). For instance, if the Bayes' factor is 1/3, the result can be interpreted in three ways:

1) Objective probability: the actual observed result is three times as probable under H_a as it is under H_0.

2) Inductive inference: the evidence supports H_a three times as strongly as it does H_0.

3) Subjective probability: assuming that we start with $P(H_0) = P(H_a)$, the odds of H_0 to H_a after we observed data are one-third what they were before seeing the data.

Therefore, Bayes' factor modifies one's initial belief about an event of interest. After observing how much Bayes' factors of certain sizes change various initial beliefs, one is able to determine the strength of the evidence. In other words, one can precisely distinguish between "strong evidence" and "weak evidence" using Bayes' factor.

7. Minimum Bayes' Factor

Goodman (1999) noted that statistical investigations get more difficult when the alternative hypothesis is stated as it is usually posed. For example, consider the alternative hypothesis "the true difference is not zero". This type of alternative hypothesis is referred to as a *composite hypothesis*, because it is composed of multiple simple hypotheses ("The true difference is 1%, 2%, 3%, …"). Katki (2008) noted that the superiority of Bayes' factor over Fisher's P value for measuring the strength of evidence is a consequence of considering alternative values of the composite hypothesis mentioned above. At the same time, he also stated that the difficulty of using Bayes' factor is related to the selection of a specific alternative value to be compared with the null value. Bayes' factor is sensitive to the choice of alternative values, and it may lead to unreasonable test results if they are not properly chosen. Therefore, an investigator must choose the alternative value that best represents all possible distinct alternative values. One way to measure the evidence for a composite hypothesis requires averaging all possible distinct alternative values (infinitely many alternative values exist, for instance, when we say "the mean is different from zero"). Choosing just that hypothesis among all components of the composite hypothesis is like looking across all patient subgroups in clinical decision-making situation. This process can be extremely complex, impractical, and time consuming.

Fortunately, there are simpler ways to select an alternative value. One can select the alternative value with the largest effect against the null hypothesis, and cite this value as the summary of evidence across all subgroups of all possible distinct values contained in a composite alternative hypothesis. This particular technique was suggested by Goodman (1999) and Katki (2008), and is called the *Minimum Bayes' Factor* (MBF). The MBF reflects the largest amount of evidence against the null hypothesis.

Therefore, whatever effect is being investigated, the best-supported alternative hypothesis is always that the unknown true effect is equal to the actual observed effect. Symbolically it is written as $X = \mu$ where X is the observed value and μ is the hypothesized mean value in H_a. Thus, the MBF uses the best-supported alternative hypothesis that has the largest amount of evidence against H_0. It is the worst-case scenario for H_0 among all possible distinct Bayes' factors, because no alternative value has a larger amount of evidence against H_0 than the MBF does. Hence, when we compare the P value with a Bayes' factor to measure the strength of the evidence against H_0, we must calculate the Bayes' factor for the hypothesis that corresponds to the actual observed difference. This is an excellent benchmark against Fisher's P value when comparing frequentists' and Bayesian methods.

Furthermore, Goodman (1999) addressed the objectivity of the MBF as follows:

> The minimum Bayes' factor is a unique function of the data that is at least as objective as the P value. In fact, it is more objective because it is unaffected by the hypothetical long-run results that can make the P value uncertain (p. 1010).

The process for translating between Fisher's P value and the MBF is summarized below:

Step 1: Since the P value is always calculated based on the observed value, one must formulate MBF in exactly the same way. Symbolically, it is composed of two probabilities written as:

$$P(\text{Evidence}|H_0) \text{ and } P(\text{Evidence}|H_a). \tag{4}$$

Step 2: Given the fact that a lesser P value means less support for H_0, one must again formulate the MBF in the same manner. This means, when calculate the MBF, one must place $P(\text{Evidence}|H_0)$ in the numerator, while $P(\text{Evidence}|H_A)$ is placed in the denominator. Therefore,

$$MBF = \frac{P(\text{Evidence}|H_0)}{P(\text{Evidence}|H_a)}. \tag{5}$$

Step 3: Convert the MBF into a Gaussian approximation form (aka, Standard Normal Distribution form) for calculation purposes. The formula is

$$MBF = \frac{P(\text{Evidence}|H_0)}{P(\text{Evidence}|H_a)} = e^{-\frac{z^2}{2}}. \tag{6}$$

Where z represents the corresponding z-value from the null effect. For example, if P=0.05 (two-sided test) is given, z is obtained as $\pm 1.96.$ Hence,

$$MBF = e^{\frac{(\pm 1.96)^2}{2}} = 0.15 \text{ or } \frac{1}{6.8}. \tag{7}$$

As a result, the observed value supports H_a 6.8 times as strongly as it does H_0 (Maxwell & Satake, 2010).

Table 1. Final (Posterior) probability of the null hypothesis after observing various Bayes factors, as a function of the prior probability of the null hypothesis

Strength of Evidence	Bayes Factor	Decrease in Probability of the Null Hypothesis	
		From	To No Less Than
Weak	1/5	0.9	0.64*
		0.5	0.17
		0.25	0.06
Moderate	1/10	0.9	0.47
		0.5	0.09
		0.25	0.03
Moderate to Strong	1/20	0.9	0.31
		0.5	0.05
		0.25	0.02
Strong to Very Strong	1/100	0.9	0.08
		0.5	0.01
		0.25	0.003

*Calculations were performed as follows: A probability (Prob) of 90% is equivalent to an odds of 9, calculated as Prob÷(1-Prob). Posterior odds = Bayes factor ● prior odds; thus, (1/5) ● 9 = 1.8. Probability odds÷(1+odds); thus, 1.8÷2.8 = 0.64.

Calculating the MBF shows that a P value greatly exaggerates the strength of evidence against the null hypothesis—one of the main reasons that the observed effects derived from the various clinical studies often do not predict true effects well (Goodman, 1999). Since a P value of 0.05 has a maximum evidential strength of the MBF of about 6.8 (shown above), it falls in the category of, at most, moderate strength against the null hypothesis (see Table 1 above). If one starts at a prior probability of even odds on the alternative hypothesis, then after collecting the evidence the probability of the null hypothesis being true drops to 13% and the probability of H_a rises to 87%. Put another way, an experiment beginning with $P(H_0) = P(H_a) = 0.5$ still has at least a 13% chance of the alternative hypothesis being wrong after observing a P value of 0.05.

Another advantage of the Bayesian approach over the frequentist methods is that the conclusion is not dichotomous. For instance, under hypothesis tests, "Reject H_0" leads to "a treatment is effective", and "do not reject H_0" leads to "a treatment is not effective" in a clinical setting. Similarly, in Fisher's P value approach "$p < 0.05$" leads to "the treatment is found to be statistically significant" whereas "$p \geq 0.05$" indicates "the treatment is found to be not statistically significant". In Bayesian methods, unlike Fisher's P value approach, a specific prior probability value is assigned in advance to both H_0 and H_a to calculate the posterior odds after one examines the evidence. Therefore, an investigator can obtain both "statistical significance" and "practical significance" (the magnitude of the change between prior and posterior probabilities) of an experiment's data. This feature should be appealing to researchers, especially in the medical and clinical sciences, because the method allows them to calculate the effectiveness of a treatment directly from the data. It also fulfills the mission of so-called evidence-based practice (Maxwell & Satake, 2010; Satake, 2014).

8. Illustrated Example: Calculations of Fisher's P Value and the MBF

In this section, we discuss the following issues:

- How are Fisher's P value and the MBF calculated under the H_0 distribution?
- Why does the MBF approach produce a more accurate and reliable result than Fisher's P value approach when measuring the strength of the evidence against H_0?

For illustration purposes, we provide a relevant example that would be typically taught in an introductory statistics course.

Example (Adapted from the book **General Statistics**, 4th edition, by W. Chase and F. Bown):

To confirm her belief that abused children will show elevated levels of depression, a psychologist gave a test called the Profile of Mood States (POMS) to a sample of 49 abused children. The results showed a mean depression score of 16.6 and a standard deviation of 5.6. Can she conclude that abused children in general have a mean depression level that is different from the national norm of 15?

Solution

Calculating an observed value based on a sample, assuming that a two-sided test is conducted:

Step 1: State H_0 and H_a.

$$H_0: \mu = 15 \tag{8}$$

$$H_a: \mu \neq 15 \tag{9}$$

Step 2: Calculate an observed value.

$$\hat{\theta} = \frac{16.6-15}{5.6/\sqrt{49}} = \frac{1.6}{0.8} = 2 \tag{10}$$

Step 3: Calculate Fisher's P value and the MBF.

Fisher's P value = $0.0228 \times 2 = 0.0456$ (since it is a two-sided test). (11)

Symbolically, the result is

Fisher's P value = $P\left(\left|\hat{\theta}\right| \geq 2 \middle| H_0 \text{ is true}\right) = 0.0456.$ (12)

$$MBF = e^{-\frac{(2)^2}{2}} = e^{-2} = 0.1353. \tag{13}$$

Since $0.1353 = \frac{1}{7.4}$, the result is that the alternative hypothesis is 7.4 times as likely as the null hypothesis, based on the observed evidence. Symbolically,

$$\text{MBF} = P\left(\hat{\theta} = 2 \middle| H_0 \text{ is true}\right) = 0.1353. \tag{14}$$

After noting that $\frac{0.1353}{0.0456} = 2.97$, it can be seen that Fisher's P value states the amount of evidence against H$_0$ as 2.97 times as much as the MBF does. This means that Fisher's P value exaggerates *statistical significance* almost 3 times as much as the MBF. Therefore, the conclusion derived from Fisher's P value is less accurate as a measure of the strength of evidence against H$_0$.

9. Conclusion

While Fisher's P value approach remains the most widely used type of P value taught in introductory and intermediate statistics courses, the disturbing fact is that Fisher's P value often leads to an inaccurate test result because it overstates the amount of evidence against the null hypothesis. In other words, one is more likely to conclude the result as "statistically significant" than when one uses MBF. As noted earlier, even Fisher did not clearly explain the importance of including the probability of more extreme data than the actual observed data (Goodman, 1999). Therefore, there is no sufficient evidence to argue for or against Fisher's inclusion of the more extreme data. The probability of a "more extreme" event is hypothetical; hence, in the authors' view, it should be excluded if one is to continue to use the P value as the chief measure of the strength of *exact* evidence. The strength of evidence against the null hypothesis must be measured based on only what is actually observed, not including unobserved and hypothetical "more extreme" data that may occur over the long run.

Pedagogically, we firmly believe that the following key points must be emphasized in statistics classrooms. First, before the MBF is introduced, instructors must stress the conceptual meaning and correct interpretation of Fisher's P value, rather than mere knowledge of the procedures for its calculation. Generally speaking, in the field of statistics, there is little value in knowing a set of procedures if students do not understand the underlying concepts. If a student understands the concepts well, then particular procedures will be straightforward and easy to learn. Second, students must have a good working knowledge of basic probability theory, particularly conditional probability, before the topics of P value and Hypothesis Tests are presented. If the topic of probability theory is thoroughly covered as a prerequisite, students are more likely to understand the meaning of Fisher's P value and less likely to misinterpret the results. This is contrary to what many introductory statistics instructors believe: that probability is the least essential topic in their courses, and that statistical analysis can be taught with only a superficial exposure to probability.

Then what can statistics educators do to help students fully understand the conceptual and mathematical differences among Fisher's P value, Neyman-Pearson's Hypothesis Tests, the combined method, and the MBF approach? In general, a theory is better understood when contrasted with another. Therefore, students will develop a better insight and deeper understanding of the fundamentals of statistical inference if they are taught the MBF approach in parallel with one or more of the other three (e.g., Fisher's P Value approach). Although there are several articles pertaining to Fisher's P value fallacies (Goodman, 1999; etc.), not enough efforts to make statistical novices aware of such fallacies are made in statistics classrooms. One reason is that instructors tend to focus on, and invest their lecture time in, explaining the mathematical derivations and mechanical procedures of the Fisher's P value for statistical inference, to the detriment of the correct interpretation of the results. As a consequence, many students at the introductory and intermediate levels are unable to understand, interpret, and apply the results correctly; they still believe that Fisher's P value with the combined method is the ultimate quantitative tool for measuring the truth of the hypothesis given evidence (Satake & Murray, 2014). To remedy this misconception, we, as statistics educators, must emphasize the limitations of traditional methods in a statistics classroom and explore Bayesian approaches, such as the MBF, to promote students' understanding of the topic of statistical inference at a deeper level.

Secondly, it is also essential to implement more intellectually stimulating and challenging content and courses into the current statistics curriculum to further develop students' quantitative skills and reasoning. The National Council of Teachers of Mathematics recommends that statistics courses should be intriguing with a level of challenge that invites speculation, that the intensity of courses increase, and that teachers implement the style described in "Integrating new knowledge gained with pre-existing intellectual constructs" (Briner, 1999; NCTM Report, 2000). So, by introducing the MBF, students are able to experience a more mathematically challenging and intellectually stimulating learning environment.

From a pedagogical standpoint, the incorrect interpretation and illusions about the meaning of the P value are truly embarrassing to our teaching profession. Furthermore, they will severely damage our students' quantitative reasoning skills and scientific literacy. Gigerenzer et al. (2004) echoed the authors' idea by stating that one quick remedy to eliminate such illusions and misunderstanding about the P value among students, professionals, and statistics instructors is to introduce Bayes' rule. The importance and practical necessity of integrating Bayes' rule within the teaching of conditional probability is becoming increasingly clear (Satake & Amato, 2008). For instance, several researchers and educators have advocated the inclusion of Bayesian perspectives into undergraduate and graduate classroom teaching. Greenland (2008) stated, "Thus, even if frequentist results remain the norm for presentation, the inclusion of Bayesian perspectives in teaching and analysis is strongly recommended". Moore (2001) stated, "although the discipline of statistics is healthy, its place in academe is not". In relation to Moore's comment on undergraduate statistics curriculum, Bolstad (2002) suggested "introducing Bayesian methods may be the key to revitalizing our undergraduate programs in statistics". Kline (2004) also recommended that statistics education be reformed. Specifically, he stated that the role of P values and hypothesis tests should be deemphasized so that more time can be spent showing students how to determine whether a result has substantive significance and how to replicate it using Bayesian methods. Yet, in reality, many introductory and intermediate statistics textbooks still do not emphasize methods beyond traditional statistical methods and tests (Capraro & Capraro, 2002).

Educators must realize the importance of teaching the correct interpretations of Fisher's P value and its alternative, the MBF approach, to students. An increased effort to do so will result in students with greater knowledge of statistical inference as well as improved scientific literacy. It is time for us to move forward to a more meaningful, evidence-based statistics.

References

Amato, P. P., & Satake, E. (2008). An Alternative Version of Conditional Probabilities and Bayes' Rule: An Application of Probability Logic. *The AMATYC Review*, *29*(2), 41-50.

Amato, P. P., Gilligan, W., & Satake, E. (1995). Using an N X M Contingency Table to Determine Bayesian Probabilities: An Alternative Strategy. *The AMATYC Review*, *16*, 34-43.

Anderson, J. L. (1998). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology [online]*, *2*(1), 2. https://doi.org/10.5751/ES-00043-020102

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine*, *25*, 3589-3631. https://doi.org/10.1002/sim.2672

Berry, D. (1996). *Statistics: A Bayesian Perspective*. Belmont, CA: Wadsworth.

Bluman, A. (2004). *Elementary Statistics: A step by step approach* (5th ed.). McGrill-Hill.

Bolstad, W. M. (2002). *Teaching Bayesian Statistics to Undergraduates: Who, What, Where, When, Why and How*. ICOTS6 International Conference on Teaching Statistics, Capetown, South Africa.

Bown, F., & Chase, W. (1999). *General Statistics* (4th ed.). Wiley.

Brase, C. H., & Brase, C. (2015). *Understandable Statistics: Concepts and Methods* (11th ed.). Cengage Learning.

Briner, M. (1999). *Learning Theories*. Denver: University of Colorado. Retrieved February 6, 2003, from http://curriculum.calstatela.edu/faculty/psparks/theorists/501learn.htm

Capraro, R. M., & Capraro, M. M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement*, *62*(5), 771-782. https://doi.org/10.1177/001316402236877

Dar, R., Omer, H., & Serlin, R. C. (1994). Misuse of statistical tests in three decades of Psychotherapy research. *Journal of Consulting and Clinical Psychology*, *6*, 75-82. https://doi.org/10.1037/0022-006X.62.1.75

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2008). *Stats Data and Models* (2nd ed.). Boston, MA: Pearson-Addison Wesley.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Journal of the Association for Psychological Science*, *6*, 274-290. https://doi.org/10.1177/1745691611406920

Fisher, R. A. (1937). Professor Karl Pearson and the Method of Moments. *Annals of Eugenics*, *7*, 303-318. https://doi.org/10.1111/j.1469-1809.1937.tb02149.x

Gigerenzer, G. (2004). Mindless statistics. *Journal of Soc Economics*, *33*, 587-606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., & Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*, *102*(4), 684-704. https://doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., Oliver V., & Stefan, K. (2004). *The null ritual: What you always wanted to know about significance testing but were afraid to ask*. SAGE. https://doi.org/10.4135/9781412986311.n21

Goodman S. N. (1988). Royall R. Evidence and scientific research. *Am J Public Health*, *78*, 1568-1574. https://doi.org/10.2105/AJPH.78.12.1568

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, *30*(12), 995-1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Goodman, S. N. (1999b). Toward evidence-based medical statistic. 2: The Bayes factor. *Annals of Internal Medicine*, *130*(12), 1005-1021. https://doi.org/10.7326/0003-4819-130-12-199906150-00019

Goodman, S. N. (2005). Introduction to Bayesian methods I: Measuring the strength of evidence. *Clin Trials*, *2*, 282-290. https://doi.org/10.1191/1740774505cn098oa

Goodman, S. N. (2008). Dirty dozen: Twelve p value misconceptions. *Elsevier*, *45*, 135-140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greenland, S., & Poole, C. (2013). Living with P values: Resurrecting a Baytesian Perspective on Frequentist Statistics. *Epidemiology*, *24*(1), 62-68. https://doi.org/10.1097/EDE.0b013e3182785741

Katki, H. A. (2008). Invited Commentary: Evidence-based evaluation of P values and Bayes factors. *American Journal of Epidemiology*, *168*(4), 384-388. https://doi.org/10.1093/aje/kwn148

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association. https://doi.org/10.1037/10693-000

Larson, R., & Farber, B. (2006). *Elementary Statistics: Picturing the world* (5th ed.). Pearson.

Lee, J. J. (2011). Demystify statistical significance—Time to move on from the P value to Bayesian analysis. *Editorials, JNCI*, *103*(1), 2-3. https://doi.org/10.1093/jnci/djq493

Matthews, R. A. J. (2001). Why should clinicians care about Bayesian methods? *Journal of Statistical Planning and Inference*, *94*(1), 43-58. https://doi.org/10.1016/S0378-3758(00)00232-9

Maxwell, L. D., & Satake, E. (2010). *Scientific Literacy and Ethical Pratice: Time for a Check-Up*. Annual American Speech-Language-Hearing Association Convention.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, *29*(4), 14-20. https://doi.org/10.2307/1176454

Moore, D. S. (2001). Undergraduate programs and the future of academic statistics. *The American Statistician*, *55*, 1-6. https://doi.org/10.1198/000313001300339860

Murray, A. V., & Satake, E. (2014). Teaching an Application of Bayes' Rule for Legal Decision-Making: Measuring the Strength of Evidence. *Journal of Statistics Education*, *22*(1), n1.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, *41*, 1299-1301. https://doi.org/10.1037/0003-066X.41.11.1299

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A Mathematical, Physical and Engineering Sciences*, *231*, 289-337. https://doi.org/10.1098/rsta.1933.0009

Oaks, M. (1986). *Statistical inference: Commentary for the social and behavioural sciences* (1st ed.). Chichester: Wiley.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making type 1 errors. *Psychological Bulletin*, *2013*, 156-163. https://doi.org/10.1037/0033-2909.102.1.159

Satake, E. (2014). *Evidence-based statistics for clinical professionals: What really prevents us from moving forward*. Keynote presentation at the annual symposium of LSU-School of ALLIED Health, New Orleans, LA.

Sullivan, M. (2007). *Informed Decisions Using Data* (2nd ed.). Pearson.

Sullivan, M. (2014). *Fundamentals of Statistics* (4th ed.). Pearson.

Triola, M. F. (2007). *Elementary Statistics* (10th ed.). Addison-Wesley.

Triola, M. F. (2014). *Elementary Statistics* (4th ed.). Pearson.

Winkler, R. L. (2001). Why Bayesian Analysis Hasn't Caught on in Healthcare Decision-Making. *International Journal of Technology Assessment in Health Care*, *17*, 56-66. https://doi.org/10.1017/S026646230110406X

Copyrights