Thailand Statistician
July 2019; 17(2): i-x
http://statassoc.or.th
Editorial

# How to Test without P-Values?

**Abstract**

In view of the p-value crisis in statistics in particular, and in sciences in general, as reported in the news and in the literature during over few decades, this paper aims simply at reminding statisticians of this serious problem (at the heart of statistical inference), as well as providing information for adjusting their teaching and research "culture". The main message is this. While the notion of P-values does provide some useful information from the observed data, it is not enough to use it alone to make decisions (in the context of statistical testing of hypotheses). As such, new sound inference procedures are needed for decision-making.

**Keywords:** Bayes factor, Bayesian tests, calibration of p-values, minimum Bayes factor, Neymann-Pearson tests, P-values, significance tests.

## 1. Introduction

Upfront, it should be clear by now (and definitely about time) that the use of the notion of p-values, as a tool in carrying out statistical hypothesis testing (i.e., p-value-based hypothesis testing), must be abandoned. However, it is not easy just do so if we do not really understand why our "p-value culture" (in standard frequentist statistics) is wrong! Thus, to be somewhat comprehensive, we will clarify this point as well.

This paper is viewed as a follow-up on Nguyen (2016) in which the news of a ban on the use of p-values in testing (by a journal) was elaborated without discussing alternative inference procedures.

The ASA's statement about p-values Wasserstein (2016) should be, finally, a wake-up call for all statisticians. We will elaborate on it a bit in subsequent sections. But while the "wrongdoing" of p-values in testing could be easy to explain (see Section 2 below), the "urgent" and "reasonable" question of statisticians seems to be "how to test without using p-values?", in other words, what are the possible alternatives? While there are several alternatives based on information theory, e.g., Anderson (2008), Burnham (2002), Cumming (2012), we will focus on the one which seems more familiar, namely Bayesian testing (e.g. Kock (2007)), especially on the proposed "Minimum Bayes Factor" as in Page (2017). A detailed introduction to Bayesian statistics constitues an alternative to p-values in testing.

As our paper is somewhat devoted to teachers of introductory statistics courses at a time where software might not be adjusted yet, and new text books have not yet been written, we recommend also Page (2017) as a guideline for teachers to write their own lecture notes.

This paper is organized as follows. In Section 2, we will elaborate upon why p-values (alone) are not suitable as an inference procedure to reach decisions in (frequentist) hypothesis testing. In Secion 3, we elaborate on general statistical inference as a basis for knowledge discovery from empirical sciences. In Section 4, we suggest a tentative "solution" to teaching testing, in the same spirit of Page (2017).

## 2. A Closer Look at P-values

In frequentist statistics, we used to use p-values to make decisions in testing. The typical situation is the so-called "tests of significance", also known as null hypothesis significance testing (NHST). In recent news, a study claims that giving solid food to babies at an earlier age make them sleep longer.

This conclusion is based on the following familiar statistical inference procedure (see standard texts such as Casella (2002), Freedman (2007), Freedman (2005), Kutner (2004)). Let $\mu, \nu$ denote the averages of sleeping time (per day) of babies when they are consuming standard baby food and solid food, respectively. To find out whether there is an "effect" (when changing food regime) on sleeping times, statisticians considered the null hypothesis $H_o : \nu - \mu = 0$. If $H_o$ is rejected, the test is declared "statistically significant". The decision to reject or accept $H_o$ is of course based on available data (evidence). Let $Y$ be the difference between two sample means of two random samples (of sleeping times) in two baby populations. Intuitively, large values of $Y$ provide an evidence against $H_o$ (i.e., we suspect that $H_o$ might be false). $Y$ is our test statistic. Suppose we are in "standard assumptions" (used in classrooms) so that the distribution of $Y$ can be identified under $H_o$, allowing us to compute our p-values of the test, when observing $Y = y$, i.e., $p(y) = P(Y \geq y|H_o)$ which is the probability of seeing the data (and possibly more extreme than it) if $H_o$ is "true". If $p(y)$ is very small, say, less than $0.01$ (or $0.001$), we reject $H_o$, otherwise, do not reject it. The justification for doing so is this. If $H_o$ were true, what we saw is a very rare event. A rare event is an event with very small chance to happen. Hence, e.g., as spelling out in Freedman (2007), p. 480-481, "It is an argument by contradiction" noting that "p-value is not the chance of $H_o$ being right".

*Remark.* While the authors of Freedman (2007) realized that the use of p-values as an inference procedure is based on an "argument by contradiction", they failed to realize that the "argument by contradiction" is invalid (to be elaborated shortly). However, they did express their feeling about the whole business of testing of significance "correctly" (Freedman, 2007, p. 562-563):

"Nowadays, tests of significance are extremely popular. One reason is that the tests are part of an impressive and well-developed mathematical theory. This sounds so impressive, and there is so much mathematical machinery clanking in the background, that tests seem truly scientific-even when they are complete nonsense. St Exupery understood this kind of problem very well: When a mystery is too overpowering, one dare not disobey (The Little Prince)"

When you toss a coin, say, $8$ times, and got $8$ heads (just like in the story of "the lady tasting tea" of R. Fisher), of course you suspect that the coin is not fair. But is "suspicion" alone enough to jump to a conclusion? Remember a "doubt" (from evidence) could be an indication, say, of guilt, but usually, it is not enough to conclude! This is precisely the number $(6)$ in the ASA's statement Wasserstein (2016):

$(6)$ "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis"

From this statement, while a p-value does provide some evidence, say, against the null hypothesis (as stated as number (1) in Wasserstein (2016)), we like to know "what else do we need to make a logical decision?". Well, it could be a "calibration of p-values", i.e., a (non linear) transformation of them to another more meaningful scale (to be elaborated in Section 4).

Now, before discussing, on a firm logical ground, why p-values created a "statistical crisis in science" Gelman (2014), let's look at how "results" from testing using p-values could be reliable?

Unlike observational data, say, in economics, where we could have only one time series (not replicable, although, under a given specific model, we could use simulations to produce "similar" data), experiments, such as in clinial trials, can surely be repeated (replicated) to check whether claims in published research are correct (or to be trusted), just like in physics. It can be said that the actual crisis of p-values is emphasized with the shocking paper "Why most published research findings are false?" (Inoannidis (2005)). Also, the reliability of results of testing in regression analysis (the Bread-and-Butter tool in, say, econometrics), has been criticized, e.g., Wheelan (2013).

In physics, when a proposed "law" (of nature) is violated by some experiments, all physicists will agree to change it. It is not so "nice" in statistics! Of course, physical laws are validated by predictions, and in Newtonian mechanics, there is no uncertainty involved (except measurement errors). And, unlike probability which is mathematical, statistics is designed as a "science" based on "inference" (ways to make decisions in the face of uncertainty). An inference procedure is a way to arrive at conclusions. It is not a (mathematical) theorem! It is based on a reasoning process which,

in turn, is based on some type of logic (needed to support the rationale of the procedure). That is why statisticians can "argue" (not prove) in favor of an inference procedure of their choice (e.g., frequentist inference vs Bayesian inference). But reliability of results in real-world applications could be used to judge the merit of a chosen inference procedure. And it is precisely the case here with p-values. As we will see shortly, a closer look into the notion of p-values will reveal "theoretical" reasons to abandon it.

First, the p-value of a test statistic $Y$ in a NHST for a null hypothesis $H_o$, when $Y = y$, is precisely $p(y) = P(Y \geq y|H_o)$, i.e., the probability of observing the observed value of the test or more extreme values than it (but not "yet" observed!), if $H_o$ is true (i.e., if the observed data is generated by the model $H_o$). Roughly speaking, it is the probability of observing the observed data given the null hypothesis. It is not the probability of $H_o$ being "true" given the data (evidence). This is reemphasized, as statement number (2) in the ASA's statement Wasserstein (2016), namely

(2) "p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone"

Thus, semantically, a p-value is not qualified as a quantity to form a sound inference procedure.

*Remark.* Of course, in the context of frequentist statistics, $H_o$ (when identified as a subset of the parameter space), is not a random event, and hence we cannot even talk about "probability of $H_o$ being true" Perhaps, in such a case, the notion of "likelihood" could be used instead, where the likelihood of a subset of the parameter space can be obtained by some generalization of the well-known concept of "likelihood of a parameter value given the data", just like in "likelihood ratio test".

Secondly, what is the logic underlying the use of p-values in testing? Well, using small values of a p-value (as a statistic) to infer a conclusion is clearly a form of "proof by contradiction" or "modus tollens" in logical deduction. Specifically, it is based on a reasoning process of the form: If $A \Longrightarrow B$ ($A$ implies $B$) is true, then $B^c \Longrightarrow A^c$ (not $B$ implies not$A$) is true. It is so, since in two-valued (binary: true or false) logic, e.g., in mathematics, $(A \Longrightarrow B) = A^c \cup B$ (not $A$ or $B$), logically, so that $(B^c \Longrightarrow A^c) = B \cup A^c = A^c \cup B = (A \Longrightarrow B)$.

*Remark.* In binary logic (where $1, 0$ denote true and false, respectively), the truth table of the material conditional $A \Longrightarrow B$ is given by

$$\begin{bmatrix} A & B & A \Longrightarrow B \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

which is the same as the truth table of $A^c \cup B$. That is why $(A \Longrightarrow B)$ is logically equivalent to $A^c \cup B$.

But this logical equivalence, and hence its modus tollens, is only valid for binary logic! In statistical reasoning, we are in fact dealing with a multivalued logic, where the truth of an assertion is expressed as a probability value in $[0, 1]$. To show that modus tollens does not hold any more in such a multivalued logic, it suffices to give some counter examples! Not only counter examples abound, they are well-known. The best well-known is the so-called "penguin triangle" in Artificial Intelligence (typical rules that have exceptions): Since almost all birds fly, and penguins do not fly, hence penguins are not birds! For more details, see Nguyen (2016). Another example of using the wrong syllogictic reasoning is this. If $H_o$ is true then what we observed should not be a rare event. Seeing that the observed data is a rare event (i.e., an event with small probability), we conclude that $H_o$ is not true.

Finally, one more thing to note on the inference procedure based on p-values, namely "small values of p-values lead to the rejection of the null hypothesis". Before observing the data $Y = y$, the p-value is a statistic (a random variable which is observable) $p(Y)$ with values in $[0, 1]$. Often, $p(Y)$ is uniformly distributed on $[0, 1]$ (more details in the next section), and as such, as pointed out in Briggs (2016), p.178-179, there is nothing special about small values of $p(Y)$! In other words, why then

small values of $p(Y)$ would suggest that the null hypothesis is untrue? Of course, you are free to "argue" that "why not?". But if you explain to people who hire you to conduct the experiment how you reach your conclusions, do you think they will trust your results? noting that real-world applications of statistical inference affect the everyday life of many people, see e.g. Wheelan (2013). Rather than "argueing", why don't we read the well-known Lindley's paradox (see next section) which provides a situation where small p-values do not necessarily imply that $H_o$ has a low probability to be untrue.

*Remark.* Despite number (3) in Wasserstein (2016), namely

(3) "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold"

there is, recently an attempt "in defense of p-values" by changing the $\alpha$ threshold Benjamin (2017). But it is still a threshold! But, immediately, such "defense" is countered by Trafimow (2018) in the spirit that "the problem with p-values is not about the magnitude (level) of the $\alpha$, but about the logic underlying their uses".

In summary, the use of p-values (alone) for carrying out tests should be put to rest once for all. The situation is not similar to a debate (in computer science) some time ago about "how to quantify uncertainty? as an additive or as a non additive measure?" where there was things such as "In defense of probability"! There should be no "In defense of p-values". Rather, as in physics, we should look for how to live without p-values, even this culture has been around with us for over a century, and, as a quote, "we have taught our students the wrong thing too well". And that will require another closer look at "what is a sound statistical inference?" in order to figure out possible replacements of p-values in testing.

## 3. What is a Sound Statistical Inference?

In NHST of Fisher, the inference procedure (i.e., the way to use data to arrive at conclusions) is based on p-values and hence flaw since it is based on a wrong logic. In this section, we look at two other familiar inference procedures: the Neyman-Pearson (NP) and the Bayes procedures.

In accusing Fisher's NHST framework as "worse than useless", Neyman and Pearson (NP) embarked in shaping Fisher's testing setting into a decision one, as you used to learn in text books.

*Remark.* Although NHST is a decision problem after all, its main goal is discovery of knowledge which is the spirit of "science" (not about decisions). Neyman-Pearson theory of testing is formulated in a decision theory framework, suitable for social sciences.

Let's dig into the well-known Neyman-Pearson approach to hypothesis testing to find out whether their theory would collapse together with Fisher's NHST? This is important! Most statisticians will say: we are using NP and not NHST so that we might not face the p-value problem. In the past, there were discussions about "Are NHST and NP theories the same or different?", perhaps for the sake of "philosophy"! But, now, in view of the p-value crisis, this question becomes more important, and urgent! Specifically, don't you want to know whether NP theory is affected by p-value crisis? Of course, it is if NP theory is p-value-based!

For this purpose, regardless of whether we talk about "inference" or "decision-making" in Fisher and NP theories, respectively, we just focus on "how they reach their conclusions ?", i.e., what are the inference procedures they used?

Consider, typically, a NP testing problem in linear regression (e.g. Kutner (2004)) $H_o : \theta = 0$ vs $H_a : \theta \neq 0$, with type I error $\alpha$ (the type II error $\beta$ is in the "background"!). Under "standard assumptions", you use a $t$-test $T$. You compare the observed value $T = t^*$ with the $(1-\alpha)$ quantile of the $t$-distribution (under $H_o$) which is the same as comparing its p-value with $\alpha$. While the semantic of type I error is different, $\alpha$ plays the role of a threshold on p-values for making decisions (a sort of "defuzzification" of the term "small"). As such, the main engine in NP is p-value! In other words, while NP testing is formulated in a different framework and with another purpose than Fisherian testing, its inference procedure is the same, i.e., using p-values: small p-values (those which are less than the type I error viewed as a number in $[0, 1]$) lead to rejection of $H_o$. This is why in using NP testing format, you also compute p-values (in fact, using MINITAB), as taught in text books.

Let's take a closer look at p-values in NP testing. In a standard text like Casella (2002), after all fancy mathematics in NP theory, the notion of p-values appears (p. 397) with the definition : "A p-value $p(X)$ is a test statistic satisfying $0 \leq p(x) \leq 1$ for every sample point $x$. *Small values of* $p(X)$ *give evidence that* $H_a$ *is true.* A p-value is *valid*, if for every $\theta \in \Theta_o$ (corresponding to $H_o$) and every $0 \leq \alpha \leq 1$, $P_\theta(p(X) \leq \alpha) = \alpha$".

Note that, since $P_\theta(p(X) \leq \alpha)$ is nothing else than the (cumulative) distribution function (evaluated at $\alpha$) of $p(X)$, under $\theta \in \Theta_o$, a "valid" p-value statistic $p(X)$ is a random variable, uniformly distributed on the unit interval $[0, 1]$ , under $H_o$.

As detailed in Casella (2002), p 397-399, a valid p-value statistic $p(X)$ can be used to construct a NP test which rejects $H_o$, at level $\alpha$, when $p(X) \leq \alpha$, i.e., the rejection of $H_o$ in NP testing is determined by small values of p-value statistic, where "small" is defined as "less than $\alpha$". On page 397 of Casella (2002), you see the statement "The smaller the p-value, the stronger the evidence for rejecting $H_o$"! Why it is so when $p(X)$ is uniformly distributed on $[0, 1]$ ?

Thus, the inference procedure of NP testing is also p-value-based, and hence suffers the same drawback as NHST.

We turn now to another inference procedure which is not p-value-based.

Again, consider a simple setting in hypothesis testing. The (observable) random variable of interest is $X$ which has, say, a probability density of the form $f(x|\theta)$ where the true (parameter) $\theta_o$ is unknown, but known to be in a known parameter space $\Theta$. Given, say, a random sample $\mathbf{x} = (x_1, x_2, ..., x_n)$ from the population $X$, we wish to figure out whether $\theta_o \in \Theta_o$ or not, where $\Theta_o \subseteq \Theta$. Let's call $H_o : \theta \in \Theta_o$ the "null" hypothesis, and $H_a : \theta \in \Theta_a$ the alternative hypothesis (e.g., $\Theta_a = \Theta \backslash \Theta_o$, the set complement of $\Theta_o$ in $\Theta$).

In NHST, we seek a statistic $T(\mathbf{X})$ whose large values make us suspect that $H_o$ is unlikely. The p-value of this test statistic is $p(\mathbf{x}) = P(T(\mathbf{X}) \geq T(x)|H_o)$ which is roughly $P(D|H_o)$, the probability of seeing the data $D$ (in fact, the data $\mathbf{x}$ or more extreme than it), if $H_o$ is true. Now, $P(D|H_o) \neq P(H_o|D)$ which is the probability of $H_o$ being true given the data (of course, if we can "formulate" it), is it obvious that $P(H_o|D)$ should be used, instead, to shed light on whether the data reveals the possible truth of $H_o$? When a testing problem (like in NP framework) is viewed as a model selection problem, the comparison of $P(H_o|D)$ with $P(H_a|D)$ to decide which model (hypothesis) is more likely to be true, is common sense reasoning. An inference procedure based on common sense reasoning could be a sound inference procedure (a procedure which is at least not illogical).

To carry out this inference, we need to be able to "formulate" the informal quantity $P(H_o|D)$. Such a formulation is precisely at the heart of *Bayesian statistics.* The key step is to treat $H_o$ (in fact, the subset $\Theta_o \subseteq \Theta$) as a random event, so that $P(H_o|D)$ becomes a bona fide conditional probability value.

Given $X$ distributed as $f(x|\theta)$, $\theta \in \Theta$, prior to seeing data, we could have some information (even subjectively) on the location of the true $\theta_o$ on $\Theta$. The epistemic uncertainty about $\theta_o$ could be quantified as a probability distribution, just like for stochastic uncertainty (remember: "probability theory is nothing but common sense reduced to calculation" (Laplace)). That is the Bayesian *approach* to statistical inference. A (subjective) probability distribution $\pi(.)$ on $\Theta$, representing the statistician's prior knowledge about $\theta_o$, is called a prior distribution. When equiped the population parameter $\theta$ with a prior probability density $\pi(.)$, we treat $\theta$ as a (unobservable) random variable, so that we have now a random vector $(X, \theta)$ whose joint density is $f(x, \theta) = f(x|\theta)\pi(\theta)$. After seeing the data $\mathbf{X} = \mathbf{x}$, we obtain the updated/ posterior distribution of $\theta$, as

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f(\theta, \mathbf{x})d\theta} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x}|\theta)\pi(\theta)d\theta}.$$

All inferences about $\theta$ are then based upon this posterior distribution $\theta \to f(\theta|\mathbf{x})$, see e.g., Kock (2007). The posterior distribution $f(.|\mathbf{x})$ allows us to "formalize" $P(\Theta_o|\mathbf{x})$ as

$$P(\Theta_o|\mathbf{x}) = \int_{\Theta_o} f(\theta|\mathbf{x})d\theta.$$

Let's specify this to the situation of testing. The prior $\pi(.)$ over $\Theta$ could be built from priors on hypotheses and within hypotheses as follows. Let's our prior on $\{H_o, H_a\}$ (i.e., on $\{\Theta_o, \Theta_a\}$ be $\lambda(\Theta_o) = 1 - \lambda(\Theta_a)$; Let $\pi_o(.), \pi_a(.)$ be prior densities on $\Theta_o, \Theta_a$, respectively. Then, our prior $\pi(.)$ on $\Theta = \Theta_o \cup \Theta_a$ is taken as

$$\pi(\theta) = \begin{cases} \lambda(\Theta_o)\pi_o(\theta) & \text{if} \quad \theta \in \Theta_o \\ \lambda(\Theta_a)\pi_a(\theta) & \text{if} \quad \theta \in \Theta_a. \end{cases}$$

From it, it follows that

$$P(\Theta_o|\mathbf{x}) = \int_{\Theta_o} f(\theta|\mathbf{x})d\theta = \int_{\Theta_o} f(\mathbf{x}|\theta)\pi(\theta)d\theta/f(\mathbf{x}) =$$

$$\lambda(\Theta_o)\int_{\Theta_o} f(\mathbf{x}|\theta)\pi_o(\theta)d\theta/f(\mathbf{x}).$$

Similarly,

$$P(\Theta_a|\mathbf{x}) = \int_{\Theta_a} f(\theta|\mathbf{x})d\theta = \int_{\Theta_a} f(\mathbf{x}|\theta)\pi(\theta)d\theta/f(\mathbf{x}) =$$

$$\lambda(\Theta_a)\int_{\Theta_o} f(\mathbf{x}|\theta)\pi_a(\theta)d\theta/f(\mathbf{x}),$$

so that

$$\frac{P(\Theta_o|\mathbf{x})}{P(\Theta_a|\mathbf{x})} = \frac{\lambda(\Theta_o)}{\lambda(\Theta_a)} \times \frac{\int_{\Theta_o} f(\mathbf{x}|\theta)\pi_o(\theta)d\theta}{\int_{\Theta_o} f(\mathbf{x}|\theta)\pi_a(\theta)d\theta}.$$

The ratio $B^o_a(\mathbf{x}) = \frac{\int_{\Theta_o} f(\mathbf{x}|\theta)\pi_o(\theta)d\theta}{\int_{\Theta_o} f(\mathbf{x}|\theta)\pi_a(\theta)d\theta}$, representing the change from the prior odds $\frac{\lambda(\Theta_o)}{\lambda(\Theta_a)}$ to the posterior odds $\frac{P(\Theta_o|\mathbf{x})}{P(\Theta_a|\mathbf{x})}$, after seeing the data, is called the *Bayesian factor* of $H_o$ with respect to $H_a$. We could let $P(\mathbf{x}|H_i) = \int_{\Theta_i} f(\mathbf{x}|\theta)\pi_i(\theta)d\theta$, for $i = o, a$, as a "likelihood" of $\Theta_i$, given the data.

The Bayesian inference is based on $\frac{P(\Theta_o|\mathbf{x})}{P(\Theta_a|\mathbf{x})}$ via the calculation of $B^o_a(\mathbf{x}) : H_o$ is prefered to $H_a$ if $\frac{P(\Theta_o|\mathbf{x})}{P(\Theta_a|\mathbf{x})} > 1$.

Here is a well-known example, in the spirit of Lindley's paradox. In tossing a coin $n = 104,900,000$ times, we got $x = 52,263,000$ heads. Question: Is the coin fair or not?

To answer this question, we set up our familiar testing problem. Let $X$ be the number of heads in $n$ tosses, and let $\theta \in \Theta = [0, 1]$ be the probability of getting heads in a single toss. Then the distribution of $X$ is

$$f(x|\theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}.$$

Let $\Theta_o = \{\frac{1}{2}\}$, and $\Theta_a = [0, 1]\backslash\{\frac{1}{2}\}$. We are going to test $H_o : \theta = \frac{1}{2}$ vs $H_a : \theta \neq \frac{1}{2}$.

In NP testing, with, say, at $\alpha = 0.01$, we compute the p-value $P(X \geq 52,263,000\,|\theta = \frac{1}{2}) = 0.0003$, and reject $|H_o$.

In Bayesian testing, consider the most "neutral" situation where $\lambda(\Theta_o) = \lambda(\Theta_a) = \frac{1}{2}$ (so that $B^o_a(x) = \frac{P(H_o|x)}{P(H_a|x)}$), and $\pi_o(\frac{1}{2}) = 1$, $\pi_a(\theta) = 1$ for $\theta \neq \frac{1}{2}$ (uniform on $[0, 1]\backslash\{\frac{1}{2}\}$).

We have

$$B_a^o(x) = \frac{f(x|\frac{1}{2})}{\int_0^1 f(x|\theta)d\theta} = 15.4$$

clearly indicating that $P(H_o|x) > P(H_a|x)$, i.e., accepting (strongly) $H_o$. In fact the probability of $H_o$ being true is

$$P(H_o|x) = \frac{f(x|\theta = \frac{1}{2})\lambda(H_o)}{f(x|\theta = \frac{1}{2})\lambda(H_o) + \lambda(H_a)\int_0^1 f(x|\theta)\pi_a(\theta)d\theta} = 0.94.$$

## 4. An Alternative to P-values for Hypothesis Testing

In view of the analysis in the previous section, it should appear clear to us that if we seek a sound inference procedure for testing, Bayesian testing is an alternative to p-values.

While the "orthodox" critiscism of Bayesian statistics was that "bringing subjective information into a statistical analysis is not scientific: let the data speak for itself", we should re-examine it at least on two counts. First, as it is well-known, there are natural sciences (e.g., physics) and social sciences (e.g., microeconomics) where the latter involves humans and focuses mainly on decision-making. Thus, "scientific" or not depends upon which sciences we are talking about. Secondly, if we have prior (subjective or not) information we should use it, since more information should help us to make better decisions, as testified in machine learning.

On the logical ground alone, statistical inferences based on p-values alone are not sound inferences. As such, as we might continue to teach frequentist statistics, we have no choice but to teach Bayesian testing in it, at the traditional place dominated by p-value culture (of course, after explaining clearly to students why p-values are not "appropriate").

While there are other possible replacements (or even abandoning entirely hypothesis testing) , e.g., Anderson (2008), Burnham (2002), Cumming (2012), we suggest here, as in Page (2017), Bayesian testing because Bayesian statistics is a well-established theory, a "partially" well-known approach to statistical inference in academics, and more importantly, it serves as a sort of "modification" of frequentist testing to repair its "inverse logic".

Unlike frequentist inference where a toolkit is available to use before seeing the data, Bayesian statistics only provides the users with guidelines for how to proceed. Of course, how to pick priors seems to be a problem, as well as computational aspects!

For computational problems, a text like the following is useful:

*Bayesian Essentials With R*
J. M. Marin and C. P. Robert
Springer, 2014.

As for the priors, there are situations (often considered in frequentist statistics) where a *calibration of p-values* is possible, i.e., transforming p-values into another (non linear) scale compatible with the Bayesian testing spirit (and hence not illogical), called the *minimum Bayes factor* (MBF), as recommended in Page (2017). Here is an example of a calibration of a p-value to a MBF, noting that the p-value statistic is uniformly distributed on the unit interval $[0, 1]$.

Let our observable variable (population) $X$ taking values in $\mathcal{X} = (0, 1]$, which has a probability density function $f(x|\theta) = \theta x^{\theta-1}$, a Beta$(\theta, 1)$ on $(0, 1]$, where the unknown parameter $\theta \in \Theta = (0, 1]$. We wish to find out whether in fact $X$ is uniformly distributed on $(0, 1]$ (i.e., $\theta = 1$), or it follows a Beta distribution with $0 < \theta < 1$.

First, the "null" hypothesis $H_o : \theta = 1$ (a simple hypothesis identified as $\Theta_o = \{1\} \subseteq \Theta = (0, 1]$), and the alternative hypothesis $H_a : 0 < \theta < 1$ (a composite hypothesis identified as $\Theta_a = (0, 1) \subseteq \Theta = (0, 1]$).

If we are going to carry out this test from the Bayesian point-of-view, then let's focus on the computation of the Bayes factor $B_{0a} = \frac{P(x|H_0)}{P(x|H_a)}$ (noting that $P(x|H_0)$ is a density value, interpreted as the "likelihood" of $H_o$ given the data $x$, it is not quite the familiar P-value!).

Since $H_o$ here is a simple hypothesis, the distribution of $X$ inder it is well specified, namely

$$P(x|H_o) = P(x|\theta = 1) = f(x|\theta = 1) = \theta x^{\theta-1} = 1,$$

which is the "likelihood" of $H_o$, not its "probability"!.

How to compute $P(x|H_a)$ ?

Since $H_a$ is a composite hypothesis with the parameter $\theta \in (0, 1)$, to compute $P(x|\theta \in (0, 1))$, we need a prior distribution $\pi(.)$ of $\theta$ on $(0, 1)$, so that

$$P_\pi(x|H_a) = \int_{\Theta_a} f(x|\theta)\pi(\theta)d\theta = \int_0^1 f(x|\theta)\pi(\theta)d\theta.$$

Clearly, that depends on our prior $\pi(\cdot)$.

The approach consists of replacing, in this example, $\int_0^1 f(x|\theta)\pi(\theta)d\theta$ (which is the mean of $\theta \to f(x|\theta)$ with respect to $\pi(\theta)$), by its maximum, over all possible priors $\pi(\cdot)$, i.e., independent of them. Since $\int_0^1 \pi(\theta)d\theta = 1$, for any probability density function $\pi(\cdot)$, we see that (the mode is greater than the mean):

$$\int_0^1 f(x|\theta)\pi(\theta)d\theta \leq \max_{\theta\in\Theta_a} f(x|\theta) = \max_{0<\theta<1} f(x|\theta).$$

From which,

$$B_{oa} = \frac{P(x|H_0)}{P(x|H_a)} = \frac{1}{\int_0^1 f(x|\theta)\pi(\theta)d\theta} \geq \frac{1}{\max_{0<\theta<1} f(x|\theta)} = MBF_{oa}.$$

Let's compute $\max_{0<\theta<1} f(x|\theta)$ for $f(x|\theta) = \theta x^{\theta-1}$, for this example.

We have

$$\frac{d}{d\theta}[\theta x^{\theta-1}] = x^{\theta-1} + \theta\frac{d}{d\theta}(x^{\theta-1}).$$

If we let $y(\theta) = x^{\theta-1}$, then $\log y = (\theta - 1)\log x$, so that $\frac{y'}{y} = \log x \Longrightarrow y' = y\log x = x^{\theta-1}\log x$, hence

$$\frac{d}{d\theta}[\theta x^{\theta-1}] = x^{\theta-1} + \theta\frac{d}{d\theta}(x^{\theta-1}) = x^{\theta-1} + \theta[x^{\theta-1}\log x] = x^{\theta-1}[1 + \theta\log x]$$

so that $\frac{d}{d\theta}[\theta x^{\theta-1}] = 0$ when $\theta = -\frac{1}{\log x} \in (0, 1)$ if $x < \frac{1}{e}$,
from which

$$\max_{0<\theta<1} f(x|\theta) = f(x|-\frac{1}{\log x}) = (-\frac{1}{\log x})x^{-\frac{1}{\log x}-1} = -\frac{1}{x\log x}x^{-\frac{1}{\log x}} = -\frac{1}{ex\log x}$$

(since if we let $y = x^{-\frac{1}{\log x}}$, then $\log y = -1 \Longrightarrow y = e^{-1} = \frac{1}{e}$).
Thus,

$$\frac{1}{\max_{0<\theta<1} f(x|\theta)} = MBF_{oa} = \begin{cases} -ex\log x & \text{if } x < \frac{1}{e} \\ 1 & \text{otherwise,} \end{cases}$$

which is independent of any prior of $\theta$ on $H_a$. For example, if $x = 0.05$, then $MBF_{oa} = 0.41$.

*Remark.* Suppose $X$ is our p-value statistic when we use frequentist statistics, which is uniform on the unit interval (this is called a "valid p-value statistic", which corresponds to a N-P testing framework) Then for $x = 0.05$, we have $MBF_{oa} = 0.41$. The transfomation $p \to -ep\log p$ (for example) is called a *calibration* (but it is *not* just a change of scale). By considering MBF, we are in a different conceptual framework.

*Final remarks.* Recently, the ASA has released a special issue "Moving to a World Beyond $p < 0.05$" of The American Statistician (73), March 21, 2019, which called for abandoning the use of "statistical significant" and offered much to replace it. It is the final say on the crisis of p-value. As such, until new text books on introductory statistics appear, the simplest "teaching strategy" for teachers of statistics could be this. When it comes to the chapter on (frequentist) testing of hypotheses, explain and discuss the actual "problem" of using p-values to carry out tests. Then, introducing Bayesian statistics with emphasis on Bayesian testing. It should be noted that, in view of the "p-value problem", teaching both frequentist statistics and Bayesian statistics is beneficial to students. When it comes to the topic of (linear) regression analysis (for prediction), first, a discussion about the use of p-values in the 'traditional culture" should be engaged with students. If possible, especially for multivariate linear regression models, mention or teach a bit of LASSO (Least Absolute Shrinkage and Selection Operator), as an update estimation method to estimate regression parameters, an alternative for OLS (Ordinary Least Squares) or ridge regression, which is free of p-values! This is so since, like other "algorithmic modeling" procedures in machine learning, p-values and tests are absent.

## References


Anderson DA. Model based inference in the life sciences. Springer; 2008.

Benjamin DJ, Berger JO, Johannesson M, et al (2017) Redefine statistical significance. Nat. Hum. Behav. 2017; 2: 6-10.

Briggs W. Uncertainty: The soul of modeling, probability & statistics. Springer; 2016.

Burnham KP. and Anderson DR. Model selection and Multimodel Inference: A Practical Information - Theoretic Approach. Springer; 2002.

Casella G. and Berger RL. Statistical Inference. Duxbury; 2002.

Cumming G. Understanding The New Statistics. Routledge; 2012.

Freedman D. Pisani R. and Purves R. Statistics (Fourth Edition), W.W. Norton; 2007.

Freedman DA. Statistical Models: Theory and Practice. Cambridge University Press; 2005

Gelman A. and Robert CP. The statistical crises in science. Am. Sci. 2014; 102(6): 460-465.

Inoannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2(8): e124.

Kock KR. Introduction to Bayesian Statistics. Springer; 2007.

Kutner MH. Nachtsheim CJ. and Neter J. Applied linear regression models. McGraw-Hill/Irwin; 2004.

Nguyen HT. On evidential measures of support for reasoning with integrated uncertainty: A lesson from the ban of P-values in statistical inference. In: Huynh VN, Inuiguchi M, Le B, Le B, Denoeux T. editors. Integrated Uncertainty in Knowledge Modelling and Decision Making. IUKM 2016. Lecture Notes in Computer Science. 2016; 9978: 3-15.

Nguyen HT. Editorial: Why P-values are banned? Thai. Stat. 2016; 14(2): i-iv.

Page R. and Satake E. Beyond P-values and hypothesis testing: Using the minimum Bayes factor to teach statistical inference in undergraduate introductory statistics courses. JEL. 2017; 6(4): 254-266.

Trafimow D, Amrhein V, Areshenkoff CN, et al. (2018) Manipulating the alpha level cannot cure significance testing. Front Psychol. 9:699. doi: 10.3389/fpsyg.2018.00699.

Wasserstein RL. and Lazar NA. The ASA's statement on p-values: Context, process, and purpose. Am. Stat. 2016; 70(2): 129-133.

Wheelan C. Naked statistics. W.W. Norton; 2013.

Hung T. Nguyen
Associate Editor