A critical evaluation of the current "*p*-value controversy"

Stefan Wellek*[1,2]

[1] Department of Biostatistics, CIMH Mannheim, Mannheim Medical School of the University of Heidelberg, D–68159 Mannheim, J5, Germany
[2] Department of Medical Biostatistics, Epidemiology and Informatics, University of Mainz, D–55101 Mainz, Germany

Received 8 January 2017; revised 29 March 2017; accepted 7 April 2017

This article has been triggered by the initiative launched in March 2016 by the Board of Directors of the American Statistical Association (ASA) to counteract the current *p*-value focus of statistical research practices that allegedly "have contributed to a reproducibility crisis in science." It is pointed out that in the very wide field of statistics applied to medicine, many of the problems raised in the ASA statement are not as severe as in the areas the authors may have primarily in mind, although several of them are well-known experts in biostatistics and epidemiology. This is mainly due to the fact that a large proportion of medical research falls under the realm of a well developed body of regulatory rules banning the most frequently occurring misuses of *p*-values. Furthermore, it is argued that reducing the statistical hypotheses tests nowadays available to the class of procedures based on *p*-values calculated under a traditional one-point null hypothesis amounts to ignoring important developments having taken place and going on within the statistical sciences. Although hypotheses testing is still an indispensable part of the statistical methodology required in medical and other areas of empirical research, there is a large repertoire of methods based on different paradigms of inference that provide ample options for supplementing and enhancing the methods of data analysis blamed in the ASA statement for causing a crisis.

Keywords: Bayesian inference; Data mining; Measures of evidence; Multiplicity correction; Prediction; Reproducibility of experiments.

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Review of key concepts and their history

In its simplest form, the concept of "*p*-value" can be traced back to the year 1900 where Karl Pearson published a paper (Pearson, 1900) describing the basic steps making up his χ^2-test for goodness of fit of a fully specified k-nomial distribution (with k denoted $n+1$ in the original paper where n stands for the number of degrees of freedom of the null distribution of the test statistic) to an unknown distribution of the same type from which a single random sample has been taken. As is nowadays taught already in one of the first lessons of a course in applied statistics, the *p*-value of this test has to be obtained by computing the probability that a centrally χ^2-distributed random variable with $k-1$ degrees of freedom becomes at least as large as the value computed for the χ^2-test statistic with the data under analysis. More generally and precisely speaking, the *p*-value associated with a test of a simple null hypothesis H_0 specifying a single distribution for a real-valued function $t(\boldsymbol{X})$ of the collection $\boldsymbol{X}$ of random variables observed in generating the data, is defined to be $P[t(\boldsymbol{X}) \geq t(\boldsymbol{x})|H_0]$

*Corresponding author: e-mail: stefan.wellek@zi-mannheim.de, Phone: +49-621-1703-6001

with $\boldsymbol{x}$ denoting the "realized" value of $\boldsymbol{X}$ taken in the currently analyzed sample. In order to make this definition mathematically exact, it must additionally be supposed that $t(\cdot)$ is a measurable function on the sample space of $\boldsymbol{X}$ that can even be of nonfinite dimension. In many (if not the majority of) applications, H_0 corresponds instead of a single distribution to a whole, maybe even uncountable family $\left(P_\theta\right)_{\theta\in\Theta_0}$ of distributions on the sample space of $\boldsymbol{X}$ that causes technical complications for the proper use of the concept of a p-value. The usual way of extending its definition to the case of composite (nonsimple) null hypotheses is to set $P[t(\boldsymbol{X}) \geq t(\boldsymbol{x})|H_0] := \sup_{\theta\in\Theta_0} P_\theta[t(\boldsymbol{X}) \geq t(\boldsymbol{x})]$. In modern theory of frequentist inference, it is quite often the case that a p-value $P[t(\boldsymbol{X}) \geq t(\boldsymbol{x})|H_0]$ has to be considered as a random variable itself, replacing the realized value $\boldsymbol{x}$ with some random variable $\tilde{\boldsymbol{X}}$ taking values in the same sample space as $\boldsymbol{X}$ but following a possibly different distribution.

During the history of modern statistics, the view that every beginner has to know the definition of a p-value and the way it has to be handled in practical work has been by no means universally accepted at all times. Actually, in its current form the concept was a focus of the Fisherian school of frequentist inference whereas in the work of Neyman and E. S. Pearson, it played virtually no role at all. From the Fisherian viewpoint as coherently explained in his still very influential book "Statistical Methods for Research Workers" (SMRW) first published in 1925, calculation of p-values is a crucial step to be performed in statistical testing in order to *quantify the evidence* against H_0 in a scientifically meaningful and easy to interpret manner. In contrast, for Neyman and Pearson, the primary objective of statistical testing was *decision making* between H_0 and some prespecified alternative hypothesis H_1 which often, but not necessarily, states the logical complement of the null hypothesis H_0. In Neyman and Pearson's approach, every nonrandomized test is uniquely and exhaustively defined by a "critical region" $\mathcal{C}$ consisting of all points in the sample space of the data $\boldsymbol{X}$ for which a decision in favor of H_1 has to be taken. In contrast to other variants of statistical decision making, hypotheses testing is done under the restriction that the probability of obtaining data that fall in the critical region (also called the rejection region) must not exceed some upper bound α whenever H_0 is in fact true. Formally, this means that $\mathcal{C}$ must satisfy the condition that

$$\sup_{\theta\in\Theta_0} P_\theta[\boldsymbol{X} \in \mathcal{C}] \leq \alpha. \tag{1}$$

In the Neyman–Pearson framework, the upper bound α (usually called the "significance level" of the test) to the probability of committing a type-I error plays the role of a constant which, in principle, can be assigned any value that the experimenter considers reasonable. The recommendation to use $\alpha = 0.05$ as a default (as long as no multiplicity issues are involved) had become customary already in R.A. Fisher's earlier years and was explicitly formulated in the first edition of his SMRW (Fisher, 1925, § 20). The actual starting point of Neyman and Pearson's (NP) theory of hypotheses testing is the fairly obvious fact that, given anything else, the level condition (1) is far from determine the critical region C uniquely. In most cases, there are infinitely many tests maintaining some given significance level α and among these tests, there will typically be large differences in terms of the type-II error risk or, in other words, of power. The key achievements of the NP theory are methods for the construction of testing procedures that are optimal in the sense of maximizing the power among large classes of competing solutions to the same testing problem. A thorough and systematic exposition of these results has been available since 1959 when the first edition of E.L. Lehmann's monograph "Testing Statistical Hypotheses" appeared (two substantially extended editions came out in subsequent decades, with the most recent one authored jointly with J. P. Romano). A fact whose relevance for considerations about an allegedly inappropriate use of p-values can hardly be overestimated is that the role played by this concept in the NP theory is marginal at best. On the one hand, p-values often provide a convenient tool for defining critical regions. On the other hand, it can well happen that the construction of an optimal solution to some testing problem leads to a critical region that does not admit of a representation in terms of a p-value (noticeable examples are optimal tests for two-sided equivalence problems with nonsymmetrically specified equivalence margins—see Wellek, 2010).

The very brief overview of the history and basic concepts of statistical testing given in this section should suffice for making clear that an assessment of the role to be assigned to the methodology of hypotheses tests in potentially causing a crisis in empirical and experimental research must address two different questions:

(I) Is hypotheses testing per se insufficient as a methodological backbone of the empirical, inductive sciences?
(II) Given the answer to question (I) is "no" (so that the dissatisfaction with current practice of statistical inference mainly reflects inadequacies in the handling of hypotheses tests), what are the major sources of these flaws and how can they be avoided?

We address the second of these basic questions first since the ASA statement (for the full text see Wasserstein and Lazar, 2016) focuses upon *p*-values rather than the hypotheses testing paradigm as such, and *p*-values are most liable to mis- and overinterpretation among the basic elements of statistical testing.

2 Frequently encountered patterns of logically flawed inferences from *p*-values

2.1 *p*-values are not type-I error risks of tests

Even in our era, the medical literature is full of examples of statements according to which a *p*-value of, say 0.015 means that the decision in favor of the respective alternative hypothesis could be taken with a type-I error risk of 1.5% at most. As has been repeatedly pointed out also in the medical literature (see in particular Goodman, 1999), that kind of interpretation is grossly misleading and suggests a degree of certainty for the decision made in the study, which is not supported by the properties of the testing procedure. In particular, the significance level holds in the "long run" of infinitely many applications of the test and changing it upon the results of a single application is inadmissible. Bearing this elementary fact in mind substantially contributes to avoiding exaggerated expectations regarding the amount of evidence (in the widest, unfortunately rather vague sense of the term), in favor of the working hypothesis of a given trial provided by a significant result.

2.2 *p*-values versus measures of effect size

A small *p*-value is often mistaken as a result indicating that the effect under study is large. Even in elementary courses on statistical methods for nonstatisticians, it is easy to explain why *p*-values provide no suitable basis for quantifying sizes of effects. For example, let $\bar{X}$ denote the mean of a sample of n independent observations from a normal distribution with unit variance and unknown expected value to be tested for exceedance of zero. The *p*-value of the optimal test is then given by $\Phi\big(-n^{1/2}\bar{x}\big)$ (with $\Phi(\cdot)$ denoting the standard normal cumulative distribution function and $\bar{x}$ the observed value of $\bar{X}$), which for $\bar{x} > 0$ obviously decreases with n. Since in this setting the only reasonable measure of effect size is clearly given by $\bar{x}$, the *p*-value can thus take arbitrarily extreme (i.e. small) values even when the true effect size almost vanishes.

The undesirable consequences of the fallacy of misinterpreting *p*-values as sizes of effects are totally obvious for statisticians well-grounded in basic theory of inference and were also recognized in practical guidelines for research workers decades ago. The most frequently cited source of the latter provenance is Cohen (1962), a review article that appeared in a psychological journal. Although a general definition of the notion of effect size in precise statistical terminology is not given in that paper, the basic idea is readily explained: In order to measure the size of the effect tested by means of a given testing procedure, one uses a point estimator of the parameter in terms of which the hypotheses are formulated. It is

important to note that even in standard settings, the question of how to choose the measure of effect size does not always produce a unique answer. A particularly noticeable example of that kind of ambiguity is the t-test for differences between two homoscedastic Gaussian distributions from which independent samples are available. Despite the simplicity of this setting, there are three different options for choosing "the parameter of interest," all of which can be convincingly justified: (i) the nonstandardized difference $\mu_1 - \mu_2$ of population means; (ii) $(\mu_1 - \mu_2)/\sigma$; (iii) $(\mu_1 - \mu_2)/\sigma^2$. By far the most popular of these choices is (ii), and since Cohen (1962) it has become customary to call $(\mu_1 - \mu_2)/\sigma$ *the* effect size for the two-sample setting. The reasons in favor of this choice are twofold: firstly, the t-statistic is well known to be the building block of an uniformly most powerful invariant (UMPI) test of the null hypothesis $(\mu_1 - \mu_2)/\sigma \leq \theta_0$ for any fixed $\theta_0 \in \mathbb{R}$ (for negative θ_0, the corresponding alternative hypothesis specifies noninferiority of population 1). Secondly, $(\mu_1 - \mu_2)/\sigma$ is independent of the physical dimension of the underlying measurement scale and that was viewed by Cohen as a sufficient reason per se for its preference. Option (i) can be considered the better choice whenever the measurement scale is fixed and has a clearcut biological or medical meaning (e.g. the reduction of blood pressure under anti-hypertensive medication). Option (iii), although rarely considered in practice, can be justified by the fact that in the two-sample problem with data $(X_1, \ldots, X_m) \sim \mathcal{N}(\mu_1, \sigma^2)$ and $(Y_1, \ldots, Y_n) \sim \mathcal{N}(\mu_2, \sigma^2)$, the joint distribution of all observations constitute a 3-parameter exponential family with main parameter $\theta^* = (\mu_1 - \mu_2)/\sigma^2$ (see Lehmann and Romano, 2005, § 5.3). All three measures of effect size are readily estimated by (i) $\bar{X} - \bar{Y}$, (ii) $(\bar{X} - \bar{Y})/S$ and (iii) $(\bar{X} - \bar{Y})/S^2$, with $S^2 = \left(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2\right)/(m + n - 2)$.

An important aspect that is often ignored when effect sizes are reported supplementing p-values and significance statements, is that confidence limits (which likewise are standard components of data analyses leading beyond the traditional "p-value culture") should be calculated in accordance with the choice of the effect size measure. In the two-sample t-test setting, this rule is clearly violated when standard confidence limits are combined with Cohen's effect size measure $(\mu_1 - \mu_2)/\sigma$. Confidence limits for the latter can and should also be calculated exactly, but this requires application of an extra confidence procedure yielding limits that are markedly different from $\bar{X} - \bar{Y} \pm t_{m+n-2,1-\alpha/2}\, S\sqrt{(1/m + 1/n)}$. Although the theoretical results enabling calculation of exact optimal confidence limits for standardized means of normal distributions have been known for a very long time (see Lehmann, 1959, §§ 3.5, 6.14.53), tools for using these methods in practice were made available only fairly recently (for an overview see Kelley, 2007).

2.3 p-values must not be confused with probabilities of hypotheses

Perhaps the most common and difficult to counteract misinterpretation of p-values is that they are probabilities with which the respective null hypothesis holds true. As a statistician, one is inclined to think that it should be easy to explain even to applied researchers without any statistical training why in the frequentist framework, that interpretation is fundamentally wrong: hypotheses are statements about parameters and thus nonrandom quantities. Hence, each hypothesis is either true or false, irrespective of whether or not the value of the parameter is known. Accordingly, when viewed as an "event," H_0 is realized or not realized with absolute certainty, and it makes no sense to assign a probability to it. The only way to change this fundamental fact is to treat parameters as unobservable random quantities that opens up the possibilities of Bayesian modeling.

2.4 p-values provide no usable information about chances of reproducibility

Quantitative assessments of reproducibility of experiments are an area with respect to which even "hard-core" frequentists will hardly be able to dispute that satisfactory answers to scientifically very natural and important questions can only be found having recourse to the Bayesian paradigm of inference. As has been convincingly argued by Goodman (1992), an adequate formalization of the

concept is to define the chance of reproducibility as the Bayesian predictive probability that the outcome of an independent replicate of the current study falls in the rejection region of the same test as carried out before at the prespecified level of significance. A special case in which the idea underlying this approach (as well as the algorithm to be used for its implementation) can easily be illustrated, is the two-sample setting with binary data and the difference between the responder rates as the parameter of interest.

Let the sufficient statistics for the data from the current trial be given by independent variables $X \sim \mathcal{B}(m, \pi_1)$ and $Y \sim \mathcal{B}(n, \pi_2)$ with $\mathcal{B}(k, p)$ denoting the binomial distribution with parameters $k \in \mathbb{N}$, $p \in (0, 1)$. A sensible and customary choice of a prior distribution is obtained through applying Jeffreys' (1939) rule (for a brief description of the rationale behind it see below, § 4.5), which in this case leads one to view (p_1, p_2) as independent, identically beta-distributed with parameters $(1/2, 1/2)$. The posterior distribution corresponding to this "objective" (also called "noninformative") reference prior is well known (see e.g. Wellek, 2005, § 5) to be likewise the product of two beta-distributions with parameters $(a, b) = (x + 1/2, m - x + 1/2)$ and $(a, b) = (y + 1/2, n - y + 1/2)$, respectively. If in a replication study, samples of sizes m' and n' are drawn obtaining responder counts X' and Y' respectively, then routine calculations show that the joint posterior distribution (also called the "predictive distribution") of (X', Y') has probability mass function

$$\mathrm{pred}(x', y') = \mathrm{pred}_1(x')\mathrm{pred}_2(y') \tag{2}$$

where

$$\mathrm{pred}_1(x') = \frac{\Gamma(m'+1)\Gamma(m+1)\Gamma\left(x'+x+\frac{1}{2}\right)\Gamma\left(m'+m-x'-x+\frac{1}{2}\right)}{\Gamma(x'+1)\Gamma(m'-x'+1)\Gamma\left(x+\frac{1}{2}\right)\Gamma\left(m-x+\frac{1}{2}\right)\Gamma(m'+m+1)}, \tag{3}$$

$$\mathrm{pred}_2(y') = \frac{\Gamma(n'+1)\Gamma(n+1)\Gamma\left(y'+y+\frac{1}{2}\right)\Gamma\left(n'+n-y'-y+\frac{1}{2}\right)}{\Gamma(y'+1)\Gamma(n'-y'+1)\Gamma\left(y+\frac{1}{2}\right)\Gamma\left(n-y+\frac{1}{2}\right)\Gamma(n'+n+1)}, \tag{4}$$

and $\Gamma(t) := \int_0^\infty u^{t-1}e^{-u}\,du\ \forall\, t > 0$. In the literature, distributions of the form (3) and (4) are known under different names. In Bayesian statistics, the most customary terminology is to call a distribution with probability mass function (3) and (4) beta-binomial with parameters $\left(m', x+\frac{1}{2}, m-x+\frac{1}{2}\right)$ and $\left(n', y+\frac{1}{2}, n-y+\frac{1}{2}\right)$, respectively (see, e.g. Gelman et al., 2013, Table A.2.)

In the binomial two-sample setting, the problem of assessing the chances for reproducing a significant result arises whenever the selected test statistic T (being a real-valued function defined on the set of all pairs of nonnegative numbers), with the observations (x, y) from the current study exceeded the critical value $c_{m,n}(\alpha)$, say, of the corresponding test at nominal level α. A promising approach to assessing the reproducibility of this result consists of determining the posterior predictive probability of the event $\{(x', y') \big| T(x', y') > c_{m',n'}(\alpha)\}$. For illustration, let us suppose that in the primary study, the sample sizes were $m = n = 100$, with $x = 70$, $y = 50$ as the observed counts of responders. With these data, the score test of $H_0 : \pi_1 - \pi_2 \leq 0.075$ versus $H_1 : \pi_1 - \pi_2 > 0.075$ [← relevant superiority of the treatment given to patients in arm 1 of the trial] at nominal level $\alpha = 0.05$ rejects since we have

$$\begin{aligned} T(x, y) &= \frac{x/m - y/n - 0.075}{\sqrt{\frac{1}{m}\left(\hat{\pi}_2\left(\delta_0\right)+\delta_0\right)\left(1-\hat{\pi}_2\left(\delta_0\right)-\delta_0\right)+\frac{1}{n}\hat{\pi}_2\left(\delta_0\right)\left(1-\hat{\pi}_2\left(\delta_0\right)\right)}} = \\ &= \frac{(x-y)/100 - 0.075}{\sqrt{\frac{1}{100}(0.56451+0.075)(1-0.56451-0.075)+\frac{1}{100}0.56451(1-0.56451)}} = \\ &= 1.81107 > 1.645, \end{aligned}$$

with $\hat{\pi}_2(\delta_0)$ denoting the maximum likelihood estimate of π_2 under the restriction $\pi_1 - \pi_2 = \delta_0 = 0.075$ (cf. Farrington and Manning, 1990, Eq. (12)). The asymptotic p-value corresponding to the observed value of the test statistic is $1 - \Phi(1.81107) = 0.0351$.

The posterior predictive probability of obtaining a significant result with the data from the new study, given the responder counts (x, y) observed in the current study, can now be written

$$P_{pred}\left[\text{Rejection of } H_0 \text{ in the new study}\middle| X = x, Y = y\right] = \sum_{(x',y')\in C'_\alpha} \text{pred}_1(x')\ \text{pred}_2(y') \qquad (5)$$

where C'_α denotes the set of all pairs (x', y') of nonnegative integers with $x' \le m'$, $y' \le n'$, such that $\dfrac{x'/m' - y'/n' - 0.075}{\sqrt{\frac{1}{m'}(\hat{\pi}'_2(\delta_0) + \delta_0)(1 - \hat{\pi}'_2(\delta_0) - \delta_0) + \frac{1}{n'}\hat{\pi}'_2(\delta_0)(1 - \hat{\pi}'_2(\delta_0))}} > z_{1-\alpha}$. In this case, in the definition of the "new" critical region C'_α, $\hat{\pi}'_2(\delta_0)$ stands for the ML estimate of the reference responder rate π_2 computed from (m', x'), (n', y') under the restriction that $\pi_1 - \pi_2 = \delta_0$. Evaluating (5) with $m = 100 = n$, $x = 70$, $y = 50$, $\delta_0 = 0.075$, $\alpha = 0.05$, and $m' = 100 = n'$ yields $P_{pred}\left[\text{Rejection of } H_0 \text{ in the new study}\middle| X = x, Y = y\right] = 0.55018$. Thus, although the p-value obtained from the original study was fairly small, the chances of getting a significant result in a replication study would be only slightly larger than 50%. Further calculations show that the chance of reproducibility as measured in terms of the Bayesian predictive probability (5), when considered as a function of the p-value obtained from the current study, is not even monotonically decreasing. For instance, if the sample sizes available for the current study were both 500 rather than 100 and the absolute frequencies of responders were observed to be $x = 400$ and $y = 330$, the score test for the problem $H_0 : \pi_1 - \pi_2 \le 0.075$ versus $H_1 : \pi_1 - \pi_2 > 0.075$ would yield a p-value of less than 1%. Nevertheless, recalculating the posterior predictive probability (5) of getting a significant result in the same test in a replication study with $m' = 100 = n'$, one obtains an even more disappointing result, namely 29%. [SAS/IML code to reproduce these results is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/10.1002/bimj.201700001/suppinfo).]

3 Banning significance testing from making up the backbone of statistical inference – a real option for biostatistical practice?

3.1 Basic criteria of good practice of hypotheses testing

(i) Whenever a test based on a p-value is performed, the latter must be used and interpreted in the correct manner. The list of possible and frequently occurring misinterpretations of p-values is actually much longer than that considered in the previous section. In fact, one of the papers supplementing the ASA statement (Greenland et al., 2016) contains a compilation of no less than 18 different misinterpretations of that concept.

(ii) In most study reports and papers presenting the results from analyzing experimental or empirical data, several or even large numbers of p-values are presented. It must be made unequivocally clear which of these p-values ares used as a basis for making significance statements, and which were computed only as descriptive summary measures of parts of the data.

(iii) If the confirmatory part of an analysis comprises several or even a multitude of tests, one of the well-established multiple testing procedures must be applied rather than carrying out each elementary test at the same nominal significance level, which would be appropriate in a simpler study for which the confirmatory part of the statistical analysis reduces to taking a single significance decision. Generally, a multiple testing procedure which by construction provides control over the familywise error rate (FWER) has to be preferred where FWER is defined to be the

risk of erroneously rejecting at least one of the null hypotheses under consideration. Regarding the development of procedures with this property, impressive progress has been made during the last five decades. The repertoire of multiple testing procedures available nowadays for routine applications (for an overview focusing on biopharmaceutical applications see Dimitrienko et al., 2009) comprises both efficient nonparametric solutions making no assumptions about the dependencies between elementary tests, and procedures tailored for fully specified dependence structures. In extreme cases where, like in genomewide association studies (GWAS), the number of elementary tests is typically huge, it might be acceptable to weaken the requirement on the level of protection against the type-I error risk using a procedure tailored for controlling the so-called false discovery rate (FDR), in the sense of Benjamini and Hochberg (1995) rather than the FWER. An example of a dataset for which more null hypotheses are rejected by the Benjamini–Hochberg as compared with Holm's FWER-controlling procedure is presented by Victor et al. (2010).

(iv) Whenever tests of different power are available for the same setting and given all relevant specifications, the most powerful of these procedures should be applied. Likewise, if out of two tests at the same level being commensurate in power against the specific alternative of interest, one relies on asymptotics whereas the other one is an exact procedure, the latter should be preferred.

3.2 Logical limitations and paradoxes

The logic behind statistical hypotheses testing is not free of elements deeming fairly artificial and, in line with this, very difficult to understand to most nonstatisticians. Perhaps, the most conspicuous of these features is the intrinsic asymmetry of the roles played by the two hypotheses making up a testing problem: The possibility of being rejected subject to a known bound to the risk of taking a wrong decision and hence of confirming its logical counterpart (typically its complement) exists only for H_0, whereas the null hypothesis itself can never be confirmed in the same sense using the same test. If H_0 states that the effects of two treatments on some given outcome variable are the same, then the logical asymmetry between the hypotheses implies that absence of a significant difference in efficacy does not allow one to infer significant equi-efficacy. By common sense, this is a paradox since it seems to run counter the basic logical law of double negation.

Another logical subtlety which researchers making use of hypotheses testing methods must keep in mind is that the degrees of certainty expressed in terms of significance level and power do not relate in a straightforward manner to the respective statements made about the data under analysis, but are features exhibited "in the long run" by the inferential procedure used to establish these statements. In other words, the adjective "confirmatory" taken in the strict sense only indicates a quality of methods, and between results obtained by means of a confirmatory procedure and those which merit being called "confirmed" there remains a nonignorable gap.

A true paradox on that cannot be resolved without relativizing the principle giving rise to it, is connected with the concept of FWER and the sensible handling of procedures enabling its control. A basic question that automatically comes up in that context is how to discriminate between tests belonging to the same versus separate "families." As regards this question, nothing has changed in terms of the truth of the bon mot formulated by Rupert G. Miller (1966) in the introduction to the first edition of his reference work on simultaneous statistical inference: "A nonmultiple comparisonist regards each separate statistical statement as a family .... At the other extreme is the ultraconservative statistician who has just a single family consisting of every statistical statement he might make during his lifetime. ... There are few statisticians who would adhere to the first principle, but the author has never met one of the latter variety. ... Most statisticians fall somewhere in between ... but have no well-formulated principles on family size or constitution."

3.3 Underused flexibility regarding the formulation of hypotheses

One of the standard criticisms raised against the basic paradigm behind hypotheses testing is that a null hypothesis stating a null effect of the treatment under evaluation, is in most cases known a priori not to hold true in the strict sense (cf. Berger, 1985, § 4.3.3.II). Furthermore, when the corresponding test allows one to reject the assumption of identical efficacy of the treatments with sufficiently high "confidence," the researcher knows almost nothing that may be of real interest (see Berger, loc. cit., p. 20, Example 8). This criticism can easily be invalidated by recalling the fact that the scope of hypotheses formulations leading to testing problems, for which fully satisfactory or even optimal solutions are available, is much wider than underlies the classical scheme of testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, with θ measuring the population treatment effect. Actually, it was already in the early 1950s that tests were constructed for the null hypothesis $\xi_1 \leq \xi \leq \xi_2$ versus $\xi < \xi_1$ or $\xi > \xi_2$ with ξ denoting the population mean of a Gaussian distribution of unknown variance σ^2, from which a single random sample has been taken, and $[\xi_1, \xi_2]$ as an arbitrary nonempty interval which, in applications, will typically enclose zero (Hodges and Lehmann, 1954). In contrast to this problem for which an unbiased test is cumbersome to compute and has a critical region with a fairly strange shape, for the analogous problem $\theta_1 \leq \xi/\sigma \leq \theta_2$ versus $\xi/\sigma < \theta_1$ or $\xi/\sigma > \theta_2$ referring to the effect size in the sense of Cohen (1962), an optimal test exists that is very easy to implement and even allows exact computation of power and sample size. Although the same holds true for many other problems with null hypotheses specifying that the parameter of interest lies in some prespecified interval whose endpoints are called relevance limits in the jargon of modern biostatistics, the number of researchers making use of tests of that form is still amazingly low. The advantages of basing the confirmatory analysis of a study on such a test (for a brief overview of optimal tests for relevant differences for frequently occurring settings see Wellek, 2010, Ch. 11) are clear enough: A significant result is unlikely to be obtained unless the effect size has an amount considered large enough to justify the costs (in the widest sense of the word) entailed with the implementation of the treatment under evaluation.

Another important option for adapting the hypothesis formulation to the question of primary interest for a researcher is obtained by treating the assumption of relevant differences as the null hypothesis to be tested. Retaining the notation introduced in the previous paragraph so that θ stands for the selected measure of population effect size, a testing problem of this third kind reads $H_0 : \theta \leq -\varepsilon_1$ or $\theta \geq \varepsilon_2$ versus $H_1 : -\varepsilon_1 < \theta < \varepsilon_2$ with prespecified $\varepsilon_1, \varepsilon_2 > 0$. In current biostatistical terminology, a testing problem of this form is called an equivalence problem with equivalence margins ε_1, ε_2. Roughly speaking, an equivalence testing procedure, which is valid in terms of the significance level allows a researcher to "prove" the null hypothesis of the traditional two-sided problem relating to the same target parameter. Of course, the significance level is now an upper bound to the probability of declaring the treatment effects to be essentially the same although they actually differ by more than an amount considered practically irrelevant or negligible. The construction, implementation, and comparative evaluation of equivalence testing procedures has been an area of active biostatistical research since the early 1970s (for a systematic exposition of equivalence testing methods see Wellek, 2010). Originally, the interest in equivalence testing has been triggered through the introduction of an abridged regulatory procedure for the approval of so-called generic drugs to the market requiring to show equivalence between the new and the primary manufacturer's formulation of the drug with respect to so-called bioavailability measures being defined as pharmacokinetic characteristics of the time×concentration profile recorded upon administration of a drug in healthy volunteers. Studies performed for that purpose are called comparative bioavailability studies and the confirmatory analysis of the data obtained from such a trial is still the best known and most frequently referenced field of application for statistical equivalence testing procedures (monographs dealing exclusively with bioequivalence assessment were published by Chow and Liu, 2009, and Hauschke et al., 2007). However, the range of research questions that cannot be adequately addressed without recourse to equivalence testing methods is much broader. Actually, a little bit of thought reveals that there is a good number of problems for which the classical approach via testing a conventional one-point null hypothesis fails to

provide an answer to the question that really matters. A whole class of problems for which this holds true can be subsumed under the heading of model validation by means of goodness-of-fit (GoF) tests: In the interest of a consistent terminology, all tests of that group should be addressed as tests of lack of fit, due to the fact that the null hypothesis to which they are tailored states that the distance between the true model underlying the data and the model assumed by theory is zero. In order to merit their name, GoF tests should be constructed as equivalence tests with the alternative hypothesis specifying that the distance between true and theoretical model falls below some prespecified equivalence margin (for a selection of solutions to problems of testing for goodness rather than lack of fit see Wellek, 2010, Ch. 9).

A third option for a hypotheses formulation allowing one to address research questions, which cannot be answered by means of traditional two-sided tests, consists of considering generalized one-sided problems of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, with θ_0 denoting an arbitrarily prespecified point in the space of the parameter θ of interest. Assuming without loss of generality that the null value of θ indicating identical efficacy of some experimental as compared with a reference treatment is given by zero itself, there are two different objectives to be pursued by carrying out a test for a problem of that form: If θ_0 is chosen to be some negative number $-\varepsilon$ with $\varepsilon > 0$ and the corresponding alternative hypothesis is aimed to be confirmed by means of an appropriate testing procedure, this means that a researcher considers it sufficient to rule out that the experimental treatment falls short in efficacy from the reference treatment by more than a clinically acceptable margin. Studies being designed according to this basic scheme are called noninferiority (or one-sided equivalence) trials and have played a steadily increasing role in medical research during the last two decades. The increasing importance of noninferiority trials is reflected by the fact that two monographs of their own (Rothmann et al., 2011; Ng, 2014) have been published on the statistical methodology for trials of that type. The choice $\theta_0 = +\varepsilon$ makes a test of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ a test for relevant superiority of the experimental over the reference treatment and thus a one-sided version of a test for relevant differences. The fact that the construction of such tests is under shift models a straightforward exercise and their application in the confirmatory analysis of clinical trials is highly desirable, was pointed out several decades ago by Victor (1987).

In conclusion of the facts reviewed in this subsection, one can say that criticizing the methodology of statistical testing for an undue lack of flexibility with respect to its coverage of different types of questions raised in scientific research, is unwarranted.

3.4 Reasons why binary decision making is indispensible in medicine and related fields

Provided that all options for avoiding misuses of p-values and failures to adequately translate prototypical scientific research questions into statistical hypotheses are made use of in practice, the discussion of the problems raised in the ASA statement has to focus on the following question: Was the overwhelmingly broad acceptance that hypotheses testing has found in the scientific community as a basis of inference for many decades more of a handicap for true scientific progress than a real success? Answering this crucial question with "yes" would entail that one had to abandon statistical decision making as the gold-standard yardstick for discriminating between empirical results whose validity is restricted to the sample(s) actually analyzed and those admitting generalization to the underlying population(s). Adopting the phrase put in the title of the comment on the ASA statement contributed by Mayo (2016), one cannot but state that taking this action would amount to "throw out the error control baby with the bad statistics bathwater." Actually, there are unrefutable reasons why at least in the medical area of application of statistical methodology, binary decision making cannot be dispensed with.

(a) The ultimate goal of research in the medical sciences is to provide patients, physicians, and other professionals involved in the diagnosis and treatment of pathological health conditions, with a rational basis for deciding on promising measures for improving that status in a given

individual case. The vast majority of these decisions are clearly binary in nature. For example it must be decided whether or not the evidence obtained from a recent trial for the efficacy of some drug is sufficiently strong for justifying to expose the patient with the risk of a severe adverse advent having occurred upon administration of the same drug under comparable circumstances. Similarly, one has to instantaneously take a stand on whether or not the chance of early detection of some tumor entity by means of a radiation-intensive procedure balances out the health risk entailed in administrating the necessary x-ray dose to a patient of the given configuration of known predictors for that risk, and so on. In all these cases, the alternatives between which a choice has to be made, clearly make up a simple dichotomy.

(b) In epidemiology, taking binary decisions is likewise an issue of crucial importance. Typically, a case-control study about nutritional or otherwise environmental risk factors for selected diseases must eventually answer for each exposure variable under investigation the question of whether or not a substantial risk increase has to be attributed to it. This must be done without unnecessary temporal delay and as unambiguously as possible, since in the positive case, actions have to be taken for removing the respective source of exposure.

(c) No regulatory authority having to review drugs and other healthcare products can perform its main function without taking binary decisions almost every day. In particular, all decisions about the market approval of drugs taken by the FDA and the EMA are necessarily options between just two plain alternative actions, namely rejection or acceptance.

4 Options and needs for supplementing statistical analyses based on hypotheses tests

4.1 Confidence interval estimation

The methodology of confidence interval construction for population parameters was developed not much later than that of hypotheses testing. However, the origin of the concept is more difficult to identify. According to Lehmann (1959, p. 120), the first author who explained its meaning in the correct manner was E.B. Wilson in a JASA paper that appeared in 1927. The systematic theory of confidence bounds and intervals was developed and published in a couple of papers by Jerzy Neyman in the 1930s. At the core of his work is the well-known duality theorem, whose importance for practice is hard to overestimate. Roughly speaking, the theorem states that to each family of tests there corresponds a unique confidence set. The resulting confidence level is the complement of the significance level α at which each of the tests making up that family is carried out, and given the data $\boldsymbol{x}$ observed in the experiment or study under analysis, the confidence set consists of all parameter values θ_0 such that $\boldsymbol{x}$ falls in the acceptance region of the test of the null hypothesis $\theta = \theta_0$, with θ denoting the population parameter of interest.

Although it was before World War II that the theory of confidence estimation appeared as a topic on the table of contents of every good textbook on mathematical statistics, it took until the 1980s at least until the respective procedures were incorporated in the toolbox of practicing biostatisticians. Presumably, the major reason for this slow adaptation of one of the most useful tools of frequentist inference is closely related to the fact that the concept of a confidence level is still harder to understand for nonstatisticians than the distinction between p-value and significance level. For instance, only very few clinicians or medical researchers will refrain from verbalizing the result that the 95% confidence interval for the difference δ of two responder rates estimated from the data of a two-arm study was computed to be $(0.032, 0.285)$, say, by stating that the true value of δ lies between 3.2% and 28.5% with probability 95%. Being told that this is fundamentally wrong will typically leave them in a state of irritation about the artificiality of the statistical way of thinking. Admittedly, it is one of the advantages of the Bayesian paradigm that it offers an approach to interval estimation leading to results for which the interpretation in question is the correct one (see § 4.4 below).

Nowadays, there is a broad consensus among biostatisticians that the computation of confidence intervals for all parameters estimated with the data obtained from a study is an indispensable part of each state-of-the-art statistical analysis as is the incorporation of full details of the results of interval estimation in the study report. A crucial role in the path toward incorporating interval estimation in biostatistical practice was played by a group of statisticians who published a series of articles from 1986 through 1991 in the British Medical Journal (compiled in a book edited by Altman et al., 2000), each on confidence interval based inference for contributors to medical journals.

Like *p*-values, confidence limits can be used both as part of the confirmatory and the exploratory section of the analysis of a study. If the first option is chosen, the same problems of multiplicity arise as have to be coped with in hypothesis testing. The risk of obtaining at least one confidence interval failing to include the true value of the respective parameter increases with the number of confidence intervals constructed at some prespecified fixed level, analogously to the type-I error risk inflation entailed in checking several *p*-values for nonexceedance of some fixed nominal upper bound. Counteracting this undesirable effect requires adjustment of the confidence levels to be used in calculating the limit(s) of the individual intervals. The simplest option available for that purpose is applying Bonferroni's correction that amounts to using $1 - \alpha/k$ rather than $1 - \alpha$ as the nominal confidence level in each of the k steps. Converting by means of the duality principle multiple testing procedures using, like the very popular method of Holm (1979), nonconstant nominal levels depending on *p*-values obtained from other elementary tests, into confidence sets is a much more difficult problem. An easy-to-implement solution for settings where the elementary tests one starts from are one-sided tests based on asymptotically normal statistics, was developed by Strassburger and Bretz (2008) and Guilbaud (2008).

The one-to-one correspondence between confidence intervals and families of tests stated by Neyman's duality theorem is not the only relationship of far-reaching importance that can be shown to hold between interval estimation and hypotheses testing. Another link with a completely different logical basis is the key to a simple though suboptimal generic solution to problems of equivalence testing: Let $\big(\underline{\theta}(\cdot\,;\alpha), \bar{\theta}(\cdot\,;\alpha)\big)$ denote a pair of real-valued functions on the sample space $\boldsymbol{\mathcal{X}}$ of the random vector $\boldsymbol{X}$ describing the dataset under analysis, such that $\underline{\theta}(\mathbf{X};\alpha)$ and $\bar{\theta}(\mathbf{X};\alpha)$ are lower and upper confidence bounds, respectively, for θ at the *same one-sided* confidence level $1 - \alpha$. Then, a valid test for the equivalence problem $H_0 : \theta \leq -\varepsilon_1$ or $\theta \geq \varepsilon_2$ versus $H_1 : -\varepsilon_1 < \theta < \varepsilon_2$ can be carried out simply through checking whether or not the realized confidence bounds satisfy the relation $\big(\underline{\theta}(\mathbf{x};\alpha), \bar{\theta}(\mathbf{x};\alpha)\big) \subset (-\varepsilon_1, \varepsilon_2)$. The mathematical proof of this fact, which is called the "principle of interval inclusion" in the pertinent literature, is almost trivial (see, e.g., Wellek, 2010, § 3.1). The principle was first formulated and introduced as a tool for the confirmatory analysis of bioequivalence studies by W.J. Westlake (1972). As applied with the data obtained from a 2×2-crossover design with lognormally distributed measurements, the interval inclusion rule continues to be the standard procedure recommended by the regulatory guidelines for assessing the so-called average bioequivalence of two different formulations of the same drug. Equivalence tests carried out by way of checking for confidence interval inclusion are applied in many areas beyond the field of bioequivalence assessment, which is another demonstration of the potential of methods of interval estimation for extending the arsenal of confirmatory methods rooted in the frequentist paradigm of inference.

The possibilities of making use of confidence intervals in the exploratory part of an analysis are manifold, and exploiting the methodology of interval estimation is doubtlessly one of the most promising options for preventing frequentist inference from the reproach that virtually all scientific problems are eventually broken down into questions being answered by a simple yes or no. As was explained in some more detail in § 2.3, a major prerequisite for making appropriate use of this tool is that interval estimation is done for the effect size measure of major interest rather than a parameter for which computation of confidence limits is particularly simple (for an important example recall § 2.3). In many cases, interval estimation for the effect size measure of choice will admit no exact solution, but has to be based on large sample approximation. Typically, many approximate formulas are available for some given parametric function, and it is of course desirable to select the one providing the highest

numerical precision. A highly recommendable source of guidance for computing the approximate standard errors required for interval estimation is contained in the statistical algorithms chapter of the manual to the Review Manager 5 of The Cochrane Collaboration (see Deeks and Higgins, 2010).

4.2 Variable selection and model fitting in high-dimensional settings

A fast growing field of application for statistical methods is the analysis of so-called high-dimensional data. The meaning of this term, which has become fairly fashionable in recent years, is hard to define precisely. Basically, it is used for a class of datasets in which the number n of observational units is much smaller than the dimension k of the vector of measurements taken from each unit. The question, which is left to answer in order to make this concept precise, is simply how large k has to be relative to n for qualifying a dataset to fall within the realm of methods for high-dimensional data. Usually, such data are considered from a regression perspective and within that framework, the dimension is high as soon as the number of degrees of freedom left under a suitable model for the residual variance becomes negative. By this and other, content-related reasons, confirmatory assessment of the amount of association between single or sets of covariables is typically not an issue, so that any kind of problems entailed by the potential mis- or over-use of p-values definitely need not be taken in consideration. Instead, the regression models to be fitted are used for purposes of statistical learning (in the sense of the influential monograph by Hastie et al., 2011), that is prediction or classification depending on whether the output variable is continuous or categorical.

Generally, a major objective of such an analysis is to identify a tendentially small subset of the full set of covariables which, when used to build up a predictor or classifier, yields favorable values of the respective error criterion. Of the classical techniques of covariable selection, stepwise forward selection is the only option for high-dimensional settings since both best-subset and backward selection require that the design matrix C is of full rank. Due to the discreteness of the process of variable selection, the results obtained by means of stepwise forward selection tend to be unstable, which also affects the prediction error. The standard approach to determining combined predictors of reduced variability is via shrinkage of regression coefficients. The shrinking method with the longest tradition in statistics is that introduced by Hoerl (1962) and Hoerl and Kennard (1970) under the label "ridge regression." Its algebraic basis is a result given by Tikhonov (1963), according to which regularization (i.e. removing singularity) of the matrix $\mathrm{C}'\mathrm{C}$ can be done through adding a nonnegative "tuning parameter" λ to its diagonal elements. The corresponding shrunken regression coefficients are the components of the vector $(\mathrm{C}'\mathrm{C}+\lambda\mathrm{I})^{-1}\mathrm{C}'\boldsymbol{y}$, with I and $\boldsymbol{y}$ denoting the $k\times k$ identity matrix and the $n\times 1$ vector of observed values of the output variable, respectively. Following Hoerl (1962, 1970), in statistical applications the ridge regression coefficients are usually represented as solutions to the problem of minimizing the penalized sum of squares $(\boldsymbol{y}-\mathrm{C}\boldsymbol{\beta})'(\boldsymbol{y}-\mathrm{C}\boldsymbol{\beta})+\lambda\sum_{j=1}^{k}\beta_j^2$. Although both definitions lead to different algorithms, they are mathematically equivalent. A more recent variant of penalized regression was introduced by Tibshirani (1996) under the acronym LASSO (Least Absolute Shrinkage and Selection Operator). It differs from ridge regression by the norm used for measuring the length of the vector of regression coefficients in the penalization term: In the LASSO, the L_1-norm replaces squared Euclidean distance from the origin. Precisely, the penalized sum of squares to be minimized now reads $(\boldsymbol{y}-\mathrm{C}\boldsymbol{\beta})'(\boldsymbol{y}-\mathrm{C}\boldsymbol{\beta})+2\lambda\sum_{j=1}^{k}|\beta|_j$. Despite the apparent similarity between both versions of penalized regression, there are clear-cut differences in the results obtained by them when applied to the same dataset: Ridge regression shrinks all coefficients obtained from OLS regression by the factor $1/(1+\lambda)$ so that none of the individual predictors are completely discarded. In contrast, LASSO shrinkage consists of shifting all OLS estimates in absolute value by λ toward zero, so that a number of covariables will be fully eliminated from the linear combination used as the combined predictor. In other words ridge regression does not lead to selection whereas the LASSO does exactly this. An alternative to LASSO regression, which likewise leads both to variable selection and shrinkage of

coefficients, is L_2-boosting with componentwise least squares as base procedure (for details see, e.g. Bühlmann and Hothorn, 2007).

All three approaches to covariable selection and shrinkage of regression coefficients are not only available for the classical linear model, but have been adapted both for logistic regression and the proportional hazards model for censored survival data. For the logistic model, ridge-type estimation of regression coefficients was considered by Duffy and Santner (1989) and le Cessie and van Houwelingen (1992). An analogous proposal for the Cox model was made by Verweij and van Houwelingen (1994). Adaptations of the LASSO for the logistic and proportional hazards model were described by Tibshirani (1996, § 8) and Tibshirani (1997), respectively. Not surprisingly, the boosting algorithm could likewise be modified in a way allowing its application both to logistic regression models (see Bühlmann and Hothorn, 2007, § 7.1) and proportional hazards regression (Ridgeway, 1999). All of these proposals have in common that they view the mean squared error, which is the traditional measure of prediction (in)accuracy for the linear model with Gaussian output variables, as a special case of a negative log-likelihood to be replaced by its analogue for the other models. A crucial question to be raised in this context is whether sufficiently convincing reasons can be given for considering the negative of any log-likelihood function exhibiting some regularity conditions as providing a suitable measure of the prediction error entailed in the models to be built up.

4.3 Tree-based methods

Algorithms relying on so-called decision trees as building blocks are widely and successfully used as conceptually simple tools both for predicting a quantitative outcome variable and classification of sampling units with respect to an outcome criterion of the categorical type. A decision tree consists of a collection of nested dichotomies, which partition the total sample in an increasing number of subgroups. Each subgroup that is not split any further in the course of this process is called a leaf (or terminal node) of the tree under generation. If the dependent variable is a quantity that shall be predicted, its average over each leaf is used as the predicted value for all sampling units contained within it. In a classification problem, each element of a terminal node is assigned to the class that the majority of its elements belong to. Along each branch of the tree, the consecutive splits are conducted via comparing, in each unit belonging to the respective node, the same single covariable selected from the components of the whole k-dimensional vector of regressors with the same cutoff. The major challenge in constructing the tree is to optimize the choice of splitting variables and associated cutoffs in terms of the prediction and classification error, respectively. In principle, a decision tree can be grown using an arbitrarily large number k of covariables where, given some stopping criterion for the process, the number of terminal nodes will increase with k. If exceeding some limit, this increase will entail the risk of overfitting, which is avoided through "pruning" the initially grown tree afterwards by collapsing branches originating from the same node. Optimal pruning strategies were described and analyzed by Breiman et al. (1984) and Ripley (1966).

The accuracy of tree-based predictions and classifications can be substantially improved through combining the outputs of a multitude of trees fitted consecutively to the same dataset. Approaches that serve that purpose were developed in the literature on machine learning under the term "tree boosting algorithms." The manner the individual trees have to be constructed, and what "combining" means precisely, depends on the type of output variable. In the case of a regression problem, that is a setting with continuous output variable, the predictions provided by the individual trees have to be summed up, and the tree to be added at the next step is constructed through fitting a tree of the same structure to the residuals obtained for the current model. When the output variable is binary, so that the trees are used for classification, the combined classification is the majority vote over the individual trees, and the most popular boosting algorithm (AdaBoost.M1 - see Freund and Schapire, 1997) proceeds by growing the next tree with observations reweighted according to the misclassification rate produced in the previous step.

A popular resampling-based alternative to tree boosting is known under the keyword "random forests" (Breiman, 2001). By definition, a random forest consists of decision trees grown for a sufficiently large number of bootstrap samples from the observed dataset $\{(y_1, x_{i1}, \ldots, x_{ik})|i = 1, \ldots, N\}$. As a second step guided by external randomization, random forest formation entails selection of a fixed number $\tilde{k} \ll k$ of covariables at each node, to be searched through to find the optimal split for that node. This nodewise random selection of covariables provides a high chance of selecting disjoint sets of covariables in growing different trees, which reduces the correlation between the terms one has to average at the end. Random forest construction methods are even applicable when the output variable is the possibly right-censored waiting time until some failure event (cf. Ishwaran et al., 2008, 2014; Wright et al., 2017).

4.4 Inference based on the Bayesian paradigm

At the beginning of any, however cursory, overview of Bayesian methods it is of crucial importance to recall the fundamental philosophical differences between the Bayesian and the frequentist approach to statistical inference. The fact that the latter is usually called the "classical" approach is at odds with the history of the statistical sciences: The birth year of the Bayesian school of statistical inference is commonly quoted to be 1763 where Thomas Bayes' "Essay towards solving a problem in the doctrine of chances" was first published. In this treatise, one finds not only the Bayesian formula but, perhaps still more important, a formulation of the nowadays so called Bayesian postulate, according to which in the "state of ignorance," all possible values of the distributional parameter of interest have to be equally weighted a priori. Irrespective of how the prior distribution is specified, it is the distinctive feature of any Bayesian analysis that all inferences are derived from probability distributions assigned to the unknown parameters on which the distributions underlying the data do depend. Treating population parameters as random variables that can be manipulated using the full spectrum of rules of the calculus of probability, is of course in sharp contrast to the frequentist view: From the frequentist viewpoint, population parameters are unobservable constants about which no meaningful probability statements can be made.

The possibility of applying the whole machinery of the calculus of probability in establishing statements about population parameters leads to much more elaborate inferences than can be obtained by following the lines of a frequentist analysis of the same dataset. Inferential settings for which this enrichment becomes particularly obvious are hypotheses testing and interval estimation: In the Bayesian framework, hypotheses are events in the parameter space (viewed as a sample space sui generis), and the check of the p-value for nonexceedance of the prespecified significance level is replaced by a statement about the posterior probability with which the null hypothesis holds true. Similarly, the Bayesian approach to interval estimation leads to regions (called credible regions in Bayesian language) about which one can say that they contain the true parameter value with known (posterior) probability. As was pointed out in § 4.1 (in accordance with Burton, 1994), this is exactly the form of a confidence statement which most clinicians and experimental researchers expect to be the appropriate one, although it is inadmissable within the framework of frequentist inference.

It cannot, and should not, be overlooked that the advantage of providing direct answers to questions considered natural for the majority of users of statistical methods is bought at the price that, except for fairly special cases of Bayesian analyses based on objective priors (cf. Bayarri and Berger, 2004), Bayesian inference does not allow one to control for risks of drawing erroneous conclusions from the data. Still more can be said: concepts like type-I error risk (of tests of hypotheses) and risk of noncoverage (of interval estimates) are in principle non-Bayesian in nature and have no role to play in fully Bayesian analyses potentially involving informative priors. This fact is certainly one of the major reasons, why for the longest time, results of Bayesian analyses were not accepted by the regulatory authorities as sufficient for a positive decision on an application for market approval of a medical product. The second serious obstacle, which had hindered until fairly recently the adoption

of Bayesian methods for statistical practice, is the complexity of the computational problems to be solved in calculating posterior probabilities for models involving parameters of higher dimension. Due to the advances achieved during the last three decades in the field of Markov chain Monte Carlo (MCMC) algorithms for generating samples from posterior distributions inaccessible to analytical methods, this is no longer an issue. The MCMC methodology was not only refined and optimized in numerous technical papers (an excellent overview of that methodology can be found in the monograph by Gelman et al., 2013) but serves also as the backbone of a widely used software package for the routine implementation of Bayesian inference made available under the name WinBugs by Spiegelhalter et al. (2003). These technical developments played an important role in paving the way for some slow realignment of the policy of the regulatory authorities toward abandoning reluctance of taking evidence from Bayesian analyses in consideration (cf. CDRH, 2010).

Despite the availability of powerful tools for the precise estimation of posterior probabilities, and even when one is willing to refrain from error risk control in the frequentist sense, there remains as a major problem in Bayesian inference that the results are potentially sensitive to changes of the prior distribution specified for the model parameters. A frequently recommended though not uncontroversial proposal for standardizing Bayesian inference in that respect is to use a uniform reference prior throughout (being fully "neutral" in the sense of assigning each of the possible values of the unknown parameter(s) the same prior weight). The controversy about this proposal has a long history focusing on the question of how to decide what transform of the parameters the Bayesian postulate should refer to. The still most convincing solution to that problem is to apply a rule that was established by Jeffreys (1939) and followed in § 2.4 in the special case of the two-sample setting with binary data. The idea behind this rule is to construct a distribution on the parameter space of $\boldsymbol{\theta}$ (which might be multi-dimensional), which remains invariant under arbitrary differentiable one-to-one transformations of $\boldsymbol{\theta}$. What Jeffreys showed is that a density that satisfies this invariance condition has to be proportional to $|\mathbf{I}(\boldsymbol{\theta})|^{1/2}$ with $\mathbf{I}(\boldsymbol{\theta})$ denoting the Fisher information contained in a single observation from the model assumed for the data. Although the approach via the invariance principle is not free of logically questionable aspects (mainly due to the fact that it may lead, even for standard models, to nonintegrable improper priors), it has been widely and successfully used in the practice of Bayesian inference. As evidence supporting this statement, it might suffice to recall that one of the still most frequently referenced textbooks on Bayesian statistics (Box and Tiao, 1973) presents almost exclusively analyses based on noninformatory priors satisfying Jeffreys' invariance principle. A heuristic justification that is often invoked (see, e.g., Spiegelhalter et al., 2004, § 5.5) for this choice stems from the well-known fact (cf. Breslow, 1990) that there are many instances of Bayesian analyses based on noninformative priors that lead to similar results as frequentist methods.

5 Discussion

The review presented in the body of this paper of major problems of statistical inference and possible approaches to their solution is necessarily cursory, and the coverage of topics it provides is by no means complete. Nevertheless, it suffices as a basis for critically addressing the six principles formulated in the ASA statement for improving "the conduct or interpretation of quantitative science."

ad **1.** ***p*-values can indicate how incompatible the data are with a specified statistical model.** This statement is correct and fully in line with the basic definitions given above in § 1.

ad **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.** This is again just a recapitulation of a defining feature of a *p*-value rather than a point of controversy.

ad **3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.** This depends on whether or not additional sources of

information are available as a basis for decision making. In some cases it may be that the results of just a single study or trial are available at the point in time when the decision has to be taken. Then, relying on the *p*-value as the decisive criterion might be the most reasonable thing one can do.

ad **4.** **Proper inference requires full reporting and transparency.** Completeness and transparency of reporting are basic requirements of any kind of activity which merits being classed as scientific, and that should be kept in mind irrespective of what school of statistics one adheres to.

ad **5.** **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.** Mistaking *p*-values as measures of effect size is one of the distinguishing characteristics of a flawed statistical data analysis. The fact that this mistake is still encountered very often in practice and even in papers published in high-ranking journals cannot be ascribed as a logical flaw to the badly understood concept.

ad **6.** **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.** This statement must be criticized in itself: as long as there is no consensus about the precise meaning of the concept of evidence, distinguishing between good and poor measures of evidence does not shed much light on the scientific value of the concept under discussion.

All in all, the six principles formulated in the ASA statement provide much in terms of a guidance for good practice of handling and interpreting *p*-values and hypothesis tests, and little (if anything) of a convincing plaedoyer for abandoning the classical paradigm of statistical inference as a basis of scientific data analysis and evaluation. This should not be misinterpreted as a lack of determination in taking appropriate steps to overcome recognized deficits of the "traditional" methodology of statistical analysis based on the frequentist paradigm. Rather, it reflects the fact that a radical rejection of the classical principles of statistical inference as was recently proclaimed by the editors of a specific psychological journal banning both *p*-values and confidence limits (Trafimow, 2014; Trafimow and Marks, 2015), is of virtually no help as long as no conclusively substantiated alternative can be offered. What has to be affirmed without any reservation is that binary decision making by means of hypothesis tests, although indispensible in a substantial number of research areas, and in particular in the realm of regulatory affairs, requires enhancement and supplementation by having recourse to methods that have been developed without the intention to provide control over error risks of any kind. The brief review given in the core sections of this paper on the increasing number of options of that kind made available in the recent literature will hopefully demonstrate that both major branches of our science deserve any effort we are able to make in developing improved and innovative methods.

Acknowlegments The author is particularly grateful to Dankmar Böhning as the Editor for inviting him to write this paper. The Co-Editor is gratefully acknowledged for providing a list of typos, and an anonymous reviewer for his expert and thoughtful report on the paper. Last but not least, in terms of English language polishing, the paper considerably benefited from the expert proofreading done by Susan Martin.

Conflict of interest
The author declares that he has no competing interests.

References

Altman, D., Machin, D., Bryant, T. and Gardner, M., Eds. (2000). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines* (2nd edn.). BMJ Books, London, UK.

Bayarri, M. J. and Berger, J. O. (2004) The interplay of Bayesian and frequentist analysis. *Statistical Science* **19**, 58–60.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edn.). Springer, New York, NY.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.
Box, G. P. E. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.
Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.
Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.
Breslow, N. (1990). Biostatistics and Bayes. *Statistical Science* **5**, 269–298.
Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science* **22**, 477–505.
Burton, P. R. (1994). Helping doctors to draw appropriate inferences from the analysis of medical studies. *Statistics in Medicine* **13**, 1699–1713.
CDRH (2010). *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health, Rockville MD. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm.
le Cessie, S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics* **41**, 191–201.
Chow, S. C. and Liu, J. P. (2009). *Design and Analysis of Bioavailability and Bioequivalence Studies* (3rd edn.). Chapman & Hall/CRC, Boca Raton, FL.
Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology* **65**, 145–153.
Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London, UK.
Deeks, J. J. and Higgins, J. P. T. (2010). Statistical Algorithms in Review Manager 5. http://tech.cochrane.org/revman.
Dmitrienko, A., Tamhane, A. C. and Bretz, F., Eds. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC, Boca Raton, FL.
Duffy, D. E. and Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics: Theory and Methods* **18**, 959–980.
Farrington, C. P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unit relative risk. *Statistics in Medicine* **9**, 1447–1454.
Fisher, R. J. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd Ltd., Edinburgh, UK.
Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation* **121**, 256–285.
Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufman, San Francisco CA, pp. 148–156.
Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences* **55**, 119–139.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd edn.). Chapman & Hall/CRC, Boca Raton, FL.
Goodman, S. N. (1992). A comment on replication, *p*-values and evidence. *Statistics in Medicine* **11**, 875–879.
Goodman, S. N. (1999). Towards evidence-based medical statistics. 1: the *p* value fallacy. *Annals of Internal Medicine* **130**, 995–1004.
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. and Altman, D. G. (2016). Statistical tests, p-values, confidence intervals and power: a guide to misinterpretations. *Online discussion of the ASA Statement on Statistical Significance and p-Values.* http://www.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/utas_a_1154108_sm5368.pdf.
Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. *Biometrical Journal* **50**, 678–692.
Hastie, T., Tibshirani, R. and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd edn.). Springer, New York, NY.
Hauschke, D., Steinijans, V. W. and Pigeot, I. (2007). *Bioequivalence Studies in Drug Development: Methods and Applications*. Wiley & Sons, Chichester, UK.
Hodges, J. L. and Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B* **16**, 262–268.
Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress Symposium Series* **58**, 54–59.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 5–67.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics* **2**, 841–860.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J. and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics* **15**, 757–773.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: theory, application and implementation. *Journal of Statistical Software* **20**, 1–24.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley & Sons, New York, NY.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd edn.). Springer, New York, NY.

Mayo, D. G. (2016) Don't throw out the error control baby with the bad statistics bathwater: a commentary. *Online discussion of the ASA Statement on Statistical Significance and p-Values.* http://www.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/utas_a_1154108_sm4621.pdf.

Miller, R. G. (1966). *Simultaneous Statistical Inference*. McGraw-Hill Book Company, New York, NY.

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

Neyman, J. (1938). L'estimation statistique traitée comme un problème classique de probabilité. *Actualités Scientifiques et Industrielles* **739**, 25–57.

Ng, T.-H. (2014). *Noninferiority Testing in Clinical Trials: Issues and Challenges*. Chapman & Hall/CRC, Boca Raton, FL.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* **50**, 157–175.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics* **31**, 172–181.

Ripley, B. D. (1966). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.

Rothmann, M. D., Wiens, B. L. and Chan, I. F. S. (2011). *Design and Analysis of Non-Inferiority Trials*. Chapman & Hall/CRC, Boca Raton, FL.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS User Manual Version 1.4* MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley & Sons, Chichester, UK.

Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine* **27**, 4914–4927.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics* **4**, 1035–1038.

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology* **36**, 1–2.

Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology* **37**, 1–2.

Verweij, J. M. and van Houwelingen, J. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine* **13**, 2427–2436.

Victor, A., Elsäßer, A., Hommel, G. and Blettner, M. (2010). Judging a plethora of *p*-values: how to contend with the problem of multiple testing. Part 10 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International* **107**, 50–56.

Victor, N. (1987). On clinically relevant differences and shifted nullhypotheses. *Methods of Information in Medicine* **26**, 155–162.

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

Wellek, S. (2005). Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes. *Biometrical Journal* **47**, 48–61.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd edn.). Chapman & Hall/CRC, Boca Raton, FL.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmacological Sciences* **61**, 1340–1341.

Wilson, E. B. (1927). Probable inference, the law of succession and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

Wright, M. N., Dankowski, T. and Ziegler, A. (2017). Unbiasedsplit variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine* **36**, 1272–1284.

