Final Draft Post-Refereeing:
Taroni F., Biedermann A., Bozza S. 2016, Statistical hypothesis testing and common misinterpretations: Should we abandon p-value in forensic science applications?, Forensic Science International, 259, e32-e36.
DOI: http://dx.doi.org/10.1016/j.forsciint.2015.11.013

Statistical hypothesis testing and common misinterpretations: should we abandon *p-value* in forensic science applications?

Abstract

Many people regard the concept of hypothesis testing as fundamental to inferential statistics. Various schools of thought, in particular frequentist and Bayesian, have promoted radically different solutions for taking a decision about the plausibility of competing hypotheses.

Comprehensive philosophical comparisons about their advantages and drawbacks are widely available and continue to span over large debates in the literature. More recently, controversial discussion was initiated by an editorial decision of a scientific journal [1] to refuse any paper submitted for publication containing null hypothesis testing procedures. Since the large majority of papers published in forensic journals propose the evaluation of statistical evidence based on the so called *p-values*, it is of interest to expose the discussion of this journal's decision within the forensic science community. This paper aims to provide forensic science researchers with a primer on the main concepts and their implications for making informed methodological choices.

Keywords: Frequentist approach, Bayesian methodology, Bayes' theorem, p-value, Degrees of belief, Hypothesis testing

1. Introduction

Misunderstandings regularly arise in papers published in forensic science and medical journals when scientists report on statistical analyses conducted as part of casework or research. It is commonplace that statistical data analysis in a forensic context refers to presume the so-called frequentist perspective, associated with the idea that statistical conclusions can be entirely objective with known error rates. Consider, for instance, the problem of hypothesis testing, where a key component to assist scientists in drawing conclusions about competing propositions is often represented by the observed significance level, generally known as the *p-value*. Examples of such applications abound in literature. A large majority of the published papers in contemporary forensic journals propose statistical treatments based on this quantity. Use of this methodology is common in other fields also, such as psychology [2]. Recently, the editors of the journal *Basic and Applied Social Psycology* [1] published an editorial expressing the intention to ban from publication in their journal any paper containing null hypothesis testing procedures. This announcement has echoed widely, from general weekly science journals [e.g., 3] to specialist groups such as the *International Society for Bayesian Analysis* [4].

This paper aims at exposing the discussion of this decision and its motivation within the forensic science community. Our main concern is not the correctness or usefulness of frequentist statistical procedures, but rather misinterpretations surrounding the use of such procedures. There is a need in forensic science to emphasize what the various approaches allow scientists to say, and what they do not entitle them to say.

[5] noted that the temptation to use *p-values* and the frequentist approach must be resisted for several different reasons. First and foremost, the fundamental mistaken notion is to believe that if an observation is rare under a given hypothesis, then it can be regarded as evidence against that hypothesis. This logical misinterpretation is one source of confusion about what a *p-value* really means [6]. Often such a value does *not* address the questions that scientific research requires [7], in the sense that it fails to provide the probabilities of the competing hypotheses conditional on the observation, which is what the researcher is mainly interested in.

This paper is structured as follows. Section 2 offers an overview of the statistical inferential problem, from both the frequentist and the Bayesian points of view. Section 3 focuses on the topic of hypothesis testing, following again the two schools of thought, presenting the computation of a *p-value* and a posterior probability of competing hypotheses, respectively. Relevant limitations and common misinterpretations related to the implementation of null hypothesis significance testing procedures in forensic science applications are discussed in Section 4. Conclusions pointing out why we recommend a Bayesian approach to hypothesis testing are given in Section 5.

2. The inferential problem: frequentist and Bayesian schools of thought

Applications of statistics occur in virtually all fields of endeavor, business, social sciences, physical science, biological sciences, health science, forensic science, and so on. Although the specific details differ somewhat in the different fields, problems related to data analysis can be treated with a common general theory of statistics.

Consider the simple situation of observing a random variable X which is constructed as having probability density $f(x \mid \theta)$, with θ being an unknown parameter summarizing some relevant characteristics of the population that is under study (say, a population mean). To approach inferential statistics about θ, that is to infer a value for the population parameter θ, the two major schools of thought – *frequentist* and *Bayesian* – differ on their interpretations of the notion of *probability*. The frequentist scientist thinks of probability in terms of population frequencies: a probability is defined as the long-run relative frequency with which events occur under repeated observations. Consequently, it is difficult to think of a probability that is related to a single event. Under the Bayesian paradigm, probability represents a degree of belief in the assertion that an event will occur. So, it is possible to think of a probability as a property attached to a singular event. Extended critical discussions on probability definitions and their application can be found, for example, in [8], [9] and [10].

The consequence is that, for a frequentist, parameter θ is a fixed but unknown number and no probability statement can be made about it. For a Bayesian scientists, though θ may be fixed, its value is typically unknown and is subject to many peoples' uncertainties. The fundamental element of the Bayesian paradigm states that all uncertainties characterizing a problem must be described by probability distributions. Probabilities are interpreted as a conditional measure of uncertainty associated with the occurrence of a particular event given the available information, the observed data and the accepted assumptions about the mechanism which has generated the data. They provide a measure of personal degrees of belief in the occurrence of an event in these conditions. Statistical inference about a quantity of interest is described as the modification of the uncertainty about its true value in the light of acquired data, and Bayes' theorem specifies how this should be done. Hence, under the Bayesian paradigm, the uncertainty about a population parameter θ is modeled by a probability distribution $\pi(\theta)$, called the *prior distribution*, that summarizes the knowledge that is available on the values of θ before observations are obtained. Notice that parameter θ is treated as a random variable in order to describe personal uncertainty about its true value, not its variability (typically, parameters are fixed quantities). Specifically, a Bayesian analysis treats parameters as random, whereas observed data x are treated as fixed (in fact, these are the only data You[1] dispose of). In summary, the result of a Bayesian analysis is a posterior probability statement, posterior in the literal sense, because the posterior distribution describes personal beliefs about θ after looking at the available data x. Bayes' theorem allows the initial information on the parameter (θ) to be updated by incorporating the information contained in the observations (x). Inference is then based on the posterior distribution, $\pi(\theta \mid x)$, the distribution on θ conditional on x:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{f(x)},$$

where $f(x \mid \theta)$ is the likelihood function, and $f(x)$ is the marginal distribution of x.

Frequentist scientists will proceed in the opposite way. They treat data as random even after observation. In fact, if you repeated the study, the data might come out differently. Data, thus, are seen as a repeatable random sample, so that any given experiment is an instance in an infinite sequence of possible repetitions of the same experiment that is conducted under identical conditions. Vice versa among frequentists, the parameter is considered as a fixed unknown constant to which no probability distribution can be assigned [11], because it actually exists as a numerical characteristic of the data generating process. As such, one of its possible values has probability 1 while all other possible values have probability 0. See the discussion by [11]. In summary, instead of conditioning on the (observed) data to update personal beliefs on the parameter θ of interest, a frequentist scientist conditions on the parameter θ to assess the plausibility of the data one observes.

3. Opposite views to hypothesis testing

The concept of hypothesis testing is widely regarded as fundamental for inferential purposes. The statistical testing of a hypothesis involves a decision about its plausibility based on observed data. In forensic science, as well as in other disciplines such as law, medicine or physics, many scientists are attracted by the so-called falsificationist scheme. Therein, the aim is to reason − based on data − about the acceptability of a theory or a

[1] The capital letter in the 'You' is used here to emphasize the personalistic and subjectivist perspective to probability.

hypothesis, either by confirming or disconfirming it. This scheme of reasoning implies that a hypothesis should yield verifiable predictions, on the basis of which it can be tested to be true or false. If the empirical consequence of a hypothesis is experimentally shown to be false, then the hypothesis is refuted. As noted by [12], falsificationism is nothing but an extension of the proof by contradiction to the experimental method: "[t]he proof by contradiction of standard dialectics and mathematics consists in assuming true a hypothesis and in looking for (at least) one of its logical consequences that is manifestly false. If a false consequence exists, then the hypothesis under test is considered false and its opposite true" (at p. 5). This means that, as we should expect, once a hypothesis is refuted, no further evidence can ever confirm it, unless the refuting evidence or some portion of the background assumptions (knowledge) is revoked [13].

Application in practice of the falsificationist reasoning scheme faces several problems that conflict with a probabilistic view of the world [14]. Thus it is tempting for a scientist to test a hypothesis that claims the opposite of what one actually believes, that is, a hypothesis one believes a priori to be false, and then state – through appropriate experimental data – that this hypothesis is false. The statistical formalization of this procedure is due to the statistician R.A. Fisher who proposed a test of significance, the implementation of which involves the rejection of a null hypothesis if an appropriate test statistics exceeds some particular value, based on a pre-assigned significance level. The underlying ideas stipulate that a practitioner computes the *p-value* (or observed significance level) which represents the probability, assuming that the null hypothesis is true, that one would obtain a value of the test statistic that is as extreme as, or more extreme, than that obtained from the data. Following this line of reasoning, one rejects the null hypothesis if the *p-value* is less than some pre-specified significance level. Otherwise, whenever the *p-value* is greater than a pre-assigned significance level, no conclusion can be drawn until additional evidence becomes available. Note that the significance level is often considered to be arbitrary, but in fact its choice ought to be based on the relative utilities accessible from the two possible correct choices and the possible erroneous choices [15]. Nonetheless in practice, a 5% level (sometimes 1%) is commonly used as the threshold for the assessment of the significance of departures from a null hypothesis.

In this scheme, hypotheses are assumed to be fixed and it is the data which vary. That is, one looks at the probability of the data if the null hypothesis is true. However, for inferential purposes in forensic science applications the opposite is far more interesting: the probability of the null or alternative hypothesis given the observed data. This probability can be obtained by applying Bayes' rule. One should be careful, thus, not to perform the tempting but erroneously, logical inversion of probability terms [16]. As noted in the previous section regarding inference about a population parameter θ, which may take one of a range of possible values, the Bayesian approach similarly applies to hypothesis testing where it allows one to definite a prior probability (i.e., a probability that is assigned prior to observations of data) for each hypothesis. This is an expression of the personal degree of uncertainty about a hypothesis' truthfulness, and is the basis on which posterior probabilities of the competing hypotheses can be obtained. Note that the rejection or acceptance of the null hypothesis is not meant as an assertion of its truth or falsity, only that it is less or more probable than the alternative, as described by [17, at p. 34]:

> A set of observations may be logically consistent with several different hypotheses, even though some of the hypotheses are inherently less plausible than others and even if the observations are more reasonably accounted for by some hypotheses than by others. We all characteristically draw conclusions in such situations and these conclusions guide further thinking and research and influence our behavior. The conclusions drawn are uncertain, however, so that it is reasonable to seek some quantitatively consistent way of characterizing this uncertainty. Bayes' theorem is important because it provides an explication for this process of consistent choice between hypotheses on the basis of observations and for quantitative characterization of their respective uncertainties.

Clearly, in the Bayesian perspective, one's interest is directed towards the more probable hypothesis[2], and so no pre-assigned significance level is required. Relevant limitations of testing of significance, common misinterpretations and advantages of the Bayesian approach will be considered in the next section.

4. Relevant limitations of *p-value*

The Bayesian method allows one to overcome several difficulties that arise with null hypothesis significance testing procedures, even though it should be acknowledged that conclusions based on a *p-value* may not *necessarily* be misleading. However, and despite the fact that the *p-value* is perceived as a relatively simple concept to

[2]We do not cover here discussion of more advanced decision-theoretic approaches that conceptualise the choice between competing hypotheses not only in terms of the (posterior) probabilities of the propositions, but also in terms of utility or loss functions that capture one's preferences regarding the consequences of one's decisions (e.g., the false acceptance or rejection of a hypothesis).

implement and to interpret, there is reason to suspect that non-adequately trained users may not fully understand the procedure's real meaning and appropriate conclusions.

In particular, users must be aware that when performing testing of significance competing hypotheses are not equivalent. There is in fact an asymmetry associated with them: one collects data (or evidence) against the null hypothesis before it is rejected, while the acceptance of the null hypothesis does not imply an assertion about its truthfulness, rather that there is little evidence against it.

Therefore, conclusions such as the following typical example:

> There was no significant difference ($t = 0.027$, $p = 0.978$)[3] between rewarded and unrewarded participants individual responses, with respect to their opinion profiles on each individual signature, nor was there any significant difference between individuals across the signature types ($t = 0.05$, $p = 0.996$) [18][p. 30],

ought *not* be interpreted as an assertion of almost truthfulness of the null hypothesis (stating that there are no differences between individual responses from different kind of participants), but rather as a recognition that the data provide little evidence against it.

One of the major misunderstanding that accompanies the reporting on significance testing consists of interpreting the *p-value* as the probability that the null hypothesis H_0 is true. A small observed significance level is often thought to justify a statement of the sort 'having looked at these data, I believe that hypothesis H_0 is quite implausible'. That is, when $p = 0.05$, it is tempting to state that there is only a 5% probability that the null hypothesis is true. And vice versa, if we consider the above example, a *p-value* equal to 0.978 may be wrongly perceived as a large probability of the null hypothesis being true. This is not what a *p-value* means. The *p-value* is a statement about the plausibility of the data and other, more extreme, hypothetical and unobserved data sets, conditional on a null hypothesis. Since the *p-value* is calculated by assuming the null hypothesis being correct, it cannot portray the chance that this hypothesis is true.

This misreading is also known as the fallacy of the 'transposed conditional' [19], which arises when the use of a *p-value* introduces confusion about (a) a probability about some aspects of the data, assuming the null hypothesis to be true, and (b) the probability that the hypothesis is true, conditioning on the observed data. Confusion may also arise since we may encounter situations where the two approaches yield similar or even identical numerical results, such as in one-tailed hypothesis testing[4] [20], but the reported probabilities have a radically different interpretation. Conversely, examples can be found where the two approaches produce irreconciliable results, such as in two-sided hypothesis testing [21].

We can provide a very simple example to show that when performing a test of significance, a frequentist decision on the basis of a *p-value* suggesting null hypothesis rejection may seem appropriate even if data are more likely under the null hypothesis rather than the alternative. Suppose that it is of interest to examine the accuracy of a laboratory scale. Measurements X are assumed to be Normally distributed with known variance $\sigma^2 = 0.01^2$, $X \sim \mathcal{N}(\theta, 0.01^2)$. For this purpose, a weight standard of 1000mg is used. Thus, it is of interest to test whether the standard weight will be found to be equal to 1000mg (H_0) or whether it will be found either smaller or larger (H_1). Thirty measurements are taken with a sample mean $\bar{x} = 999.995$. A significance test of level 0.05 would allow one to reject the null hypothesis since it is easy to show that the test statistic falls into the rejection region and that the *p-value* is lower than the significance level ($p = 0.006$). However, if we perform a Bayesian test by introducing a prior density under the alternative hypothesis (the standard weight is different from 1000mg), $\mathcal{N}(1000, 0.1^2)$, and assuming the null hypothesis a priori as likely as the alternative, standard calculations will provide a posterior probability of the null hypothesis that is equal to 0.56. So, from one side (frequentist) one would reject the null hypothesis, from the other side (Bayesian) we have a posterior probability equal to 0.56 (see Appendix). Of course the computation of a posterior probability also depends on the prior probability elicitation, however the reader can refer to [21] which, for a large class of prior distributions, provides lower limits of the posterior probability of the null hypothesis that are still larger than the *p-value*. Thus, one can face situations where one would reject a hypothesis that has a non-negligible probability.

A further argument against the widespread practice of significance testing is given by the lack of reproducibility of results. Probabilities in the frequentist approach must be based on repetition under identical conditions. Thus, if one were to repeat analyses many times, then on only 5% of those occasions one would falsely reject the null hypothesis. Using [22]'s words, the perspectives can be summarized as follows:

[3] Note that t is the observed value of a given test statistic (see the Appendix for an illustrative example) and p is the *p-value*.

[4] A test is one-tailed when deviations of the population parameter from some benchmark value are considered possible in only one direction; vice versa, whenever deviations in either direction are considered theoretically possible, the test is said to be two-tailed.

> Indeed, one might argue that those types of posterior probability statement are exactly what one wants from a data analysis, letting us make statement of the sort 'how plausible is hypothesis H_0 in light of these data?' A frequentist *p-value* answer a different question 'how frequently would I observe a result at least as extreme as the one obtained if H_0 were true?', which is a statement about the plausibility of the data given the hypothesis. Turning this assessment about the hypothesis requires another step in the frequentist chain of reasoning (e.g. conclude H_0 is false if the *p-value* falls below some preset level) (at p.33).

Beyond general research questions regarding experimentally repeatable trials under controlled situations, does operational forensic science offer realistic possibilities to plan experiments? The answer appears to be no. It suffices to consider a criminal trial where there is uncertainty about the defendant's guilt, and the evidence is represented by material found at the crime scene and witnesses' statements. The point is that probability statements about hypotheses lie within the province of the Bayesian approach. Under the frequentist school of thought, hypotheses do not have probabilities associated with them: hypotheses about θ are either true or false. The truth or falsity of a hypothesis does not only depend on the data observed, the correct interpretation of the *p-value* thus is much more laborious. As insisted previously, the *p-value* cannot measure the probability of the truth of the null hypothesis because its calculation assumes the null hypothesis to be true. There were long debates about these aspects. Consider [23, at p. 284], for instance: "one can't have a measure assuming something to be true while simultaneously measuring how likely the same thing is to be false". Similarly [24, at p. 42]: "to interpret a p-value as the probability that the null hypothesis is true is not only wrong but also dangerously wrong. The danger arises because this interpretation ignores how plausible the hypothesis might have been in the first place". Note that the underlying debate is still going on [4].

Examples and discussion on these misleading aspects, related to legal affairs, are also presented in the *Reference manual on statistics* [25] and an extended discussion on the fallacy of the transposed the conditional – in this context also called *p-value* fallacy – is given in [26].

More technical aspects limit the use of *p-values*. First, identical *p-values* for two different experiments can lead to very different conclusions depending on the sample size. Examples are given in [27]. Second, the definition of the *p-value* that takes into account the probability to obtain results at least as extreme as the observed data is in contradiction with the likelihood principle according to which once data are observed, no other values matter. Hypotheses should be compared by how well they explain the data. Therefore, the hypothetical extreme values that might have been observed are irrelevant since they do not describe how well the null hypothesis explains the available data [28]. On the contrary, Bayesian inference obeys this principle because it uses the fixed x seen [19].

5. Conclusion

The traditional null hypothesis significance testing based on the *p-value* as a method of statistical inference has provoked numerous counter arguments. In view of the above, the answer to the question *'Should we abandon p-values in forensic science applications?'* rejoins the position of [19, at pag 359]:

> My own view is that significance tests, violating the likelihood principle and using transpose conditionals, are incoherent and not sensible.

We have highlighted and discussed some of them, although they are not new (see, for example, [29]), because there is a tendency among forensic scientists to support their conclusions in published research by means of this summary.

The Bayesian method would provide a logically defensible answer both in experimental studies, where there is room for planned experiments, as well as in observational studies. In forensic science, it is feasible to formulate hypotheses about the state of the world and examine them by conditioning on available new information. Bayes' theorem is a basic corollary of the theorem of compound probabilities [30] and is uncontroversial [31]. Bayes' rule is the logical framework for updating degrees of beliefs.

Many forensic scientists, however, remain reluctant to consider Bayesian methods. The perceived simplicity of the *p-value* and its apparent neutrality, often conflict with the Bayesian paradigm. Its spread is constrained on two main grounds: (i) computational issues (e.g., the derivation of the posterior distribution may be perceived as a tedious task), and (ii) the myth of objective probabilities [32], which is part of the myth of objective knowledge [33]. Computational challenges do not represent a convincing argument since, over decades, there have been and continue to be ever increasing improvements in computational facilities. As far as the presumed 'objectivity' of the frequentist approach is concerned, it must be pointed out that no statistical approach can be entirely neutral. Both

the prior distribution $\pi(\theta)$ and the probability density $f(x \mid \theta)$ describe personal beliefs: one about the parameter, and one about the data. However, while the specification of a prior distribution on θ is often controversial, the parametric model proposed by the scientist is uncritically accepted [34].

Appendix

The example given in Section 4 focuses on testing the null hypothesis $H_0 : \theta = 1000$ against the alternative hypothesis $H_1 : \theta \neq 1000$. Measurements are assumed to be Normally distributed with known variance $\sigma^2 = 0.01^2$. Assume the null hypothesis is believed, a priori, to be as likely as the alternative, that is $Pr(H_0) = Pr(H_1) = 0.5$, and that the prior density under the alternative hypothesis is Normal with prior mean $\mu = 1000$ and prior variance $\tau^2 = 0.1^2$. Thirty measurements were taken and a sample mean $\bar{x} = 999.995$ was observed.

Bayesian hypothesis testing involves the computation of a Bayes factor that measures the change produced by the evidence when going from the prior to the posterior distribution in favour of one hypothesis as opposed to another. In the specific case, it can be shown that the Bayes factor reduces to [35]

$$\begin{aligned} BF &= \left(1 + \frac{n\tau^2}{\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2/n} - \frac{1}{\tau^2 + \sigma^2/n}\right)(\bar{x} - \mu)^2\right) \\ &= \left(1 + \frac{30 \cdot 0.1^2}{0.01^2}\right)^{1/2} \exp\left(-\frac{1}{2}\left(\frac{1}{0.01^2/30} - \frac{1}{0.01^2/30 + 0.1^2}\right)(-0.004)^2\right) \\ &= 1.29 \end{aligned}$$

The posterior probability of hypothesis H_0 can then be computed according to

$$Pr(H_0 \mid x) = \left(1 + \frac{1}{BF}\right)^{-1} = \left(1 + \frac{1}{1.29}\right)^{-1} = 0.56.$$

Vice versa, a null hypothesis significance test with significance level α, would have a rejection region:

$$|T_n| = \left|\frac{\bar{X}_n - 1000}{\sigma/\sqrt{n}}\right| \geq \left|z_{\alpha/2}\right|$$

In the specific case, the observed value of the test statistics T_n takes value

$$\frac{\bar{x} - 1000}{\sigma/\sqrt{n}} = \frac{999.995 - 1000}{0.01/\sqrt{30}} = -2.73,$$

and falls into the rejection region. The null hypothesis is therefore rejected, with an observed significance level (i.e., *p-value*) smaller than 0.01,

$$P\left[|T_n| > 2.73 \mid H_0\right] = 0.006.$$

References

[1] D. Trafimow, M. Marks, Editorial, Basic and Applied Social Psychology 37 (2015) 1–2.
[2] J. Kruschke, H. Aguinis, H. Joo, The time has come: Bayesian methods for data analysis in the organizational sciences, Organizational Research Methods 15 (2009) 722–752.
[3] J. Leek, R. Peng, *p* values are just the tip of the iceberg, Nature 520 (2015) 612.
[4] A. Schmidt, J. Berger, P. David, J. Kadane, T. O'Hagan, L. Pericchi, Banning null hypothesis significance testing, The ISBA Bulletin 22 (2015) 5–9.
[5] I. Good, The interface between statistics and philosophy of science, Statistical Science 3 (1988) 386–412.
[6] R. Nuzzo, Statistical errors, Nature 506 (2014) 150–152.
[7] B. Lecoutre, Training students and researchers in Bayesian methods, Journal of Data Science 4 (2006) 207–232.
[8] D. Lindley, Probability, in: C. Aitken, D. Stoney (Eds.), The use of statistics in forensic science, Ellis Horwood, New York, 1991, pp. 27–50.
[9] T. Aven, G. Reniers, How to definee and interpret a probability in a risk and safety setting, Safety Science 51 (2013) 223–231.
[10] F. Lad, Operational Subjective Statistical Methods : a Mathematical, Philosophical, and Historical Introduction, John Wiley & Sons, New York, 1996.
[11] J. Kadane, Prime time for Bayes, Controlled Clinical Trials 16 (1995) 313–318.
[12] G. D'Agostini, From observations to hypotheses - Probabilistic reasoning versus falsificationism and its statistical variations, in: Vulcano Workshop on Frontier Objects in Astrophysics and Particle Physics, Vulcano, Italy, 2004.
[13] C. Howson, P. Urbach, Scientific Reasoning - The Bayesian Approach, 2nd Edition, Open Court, Chicago and La Salle, 1996.

[14] P. Dixon, The p-value fallacy and how to avoid it, Canadian Journal of Experimental Psychology 57 (2003) 189–202.
[15] E. Lehmann, J. Romano, Testing Statistical Hypotheses, 3rd Edition, Springer, New York, 2005.
[16] D. Wijeysundera, P. Austin, J. Hux, W. Beattie, A. Laupacis, Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials, Journal of Clinical Epidemiology 62 (2003) 13–21.
[17] J. Cornfield, Bayes theorem, Review of the International Statistical Institute 35 (1967) 34–49.
[18] T. Dewhurst, B. Found, K. Ballantyne, D. Rogers, The effects of extrinsic motivation on signature authorship opinions in forensic signature blind trials, Forensic Science International 236 (2014) 127–132.
[19] D. Lindley, Understanding Uncertainty, revised edition Edition, John Wiley & Sons, Hoboken, 2014.
[20] G. Casella, R. Berger, Reconciling bayesian and frequentist evidence in the one-sided testing problem, Journal of the American Statistical Association 82 (1987) 106–111.
[21] J. Berger, T. Sellke, Testing a point null hypothesis: the irreconcilability of p values and evidence, Journal of the American Statistical Association 82 (1987) 112–122.
[22] S. Jackman, Bayesian Analysis for Social Sciences, John Wiley & Sons, Chichester, 2009.
[23] S. Goodman, Introduction to bayesian methods i: measuring the strength of evidence, Clinical Trials 2 (1995) 282–290.
[24] A. O'Hagan, The Bayesian statistics: principles and benefits, in: M. van Boekel, A. Stein, A. van Bruggen (Eds.), Bayesian statistics and quality modelling in agro-food production chain (Wageningen UR Frontis Series), Kluwer Academic Publishers, Dordrecht, 2004, pp. 31–45.
[25] D. Kaye, D. Freedman, Reference Manual on Scientific Evidence, in: National Research Council, Reference Manual on Scientific Evidence, The National Academic Press, Washington D.C., 2011, pp. 211–302.
[26] S. Goodman, Toward evidence-based medical statistics. 1: The p-value fallacy, Annals of Internal Medicine 130 (1999) 995–1004.
[27] D. Spiegelhalter, K. Abrams, J. Myles, Bayesian approaches to clinical trials and health-care evaluations, John Wiley & Sons, Chichester, 1966.
[28] M. Lavine, What is bayesian statistics and why everything else is wrong, The Journal of Undergraduate Mathematics and Its Applications 20 (1999) 165–174, www.math.umass.edu/ lavine/whatisbayes.pdf.
[29] W. Rozeboom, The fallacy of the null-hypothesis significance test, Psychological Bulletin 57 (1960) 416–428.
[30] B. de Finetti, Theory of Probability, A Critical Introductory Treatment, Vol. 1, John Wiley & Sons, London, 1974.
[31] W. Salmon, The Foundations of Scientific Inference, University of Pittsburgh Press, Pittsburgh, 1966.
[32] B. de Finetti, Sul significato soggettivo della probabilità, Fundamenta Mathematicae XVII (1931) 298–329.
[33] S. Senn, Bayesian, likelihood, and Frequentist approaches to statisticst, Applied Clinical Trials August (2003) 35–38.
[34] D. Lindley, Is our view of bayesian statistics too narrow?, in: J. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), Bayesian Statistics 4, Oxford University Press, 1992, pp. 1–15.
[35] F. Taroni, S. Bozza, A. Biedermann, P. Garbolino, C. Aitken, Data analysis in forensic science: a Bayesian decision perspective, John Wiley & Sons, Chichester, 2010.

Acknowledgments

We are greatly indebted to Dr. Frank Lad of the University of Canterbury, Department of Mathematics and Statistics, for the constructive comments that greatly helped to improve the quality of this paper. Alex Biedermann gratefully acknowledges the support of the Swiss National Science Foundation through grant No. BSSGI0_155809.