Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 1, 127–144
DOI: 10.1214/14-BJPS264

Assigning probabilities to hypotheses in the context of a binomial distribution

Casper J. Albers[a], Otto J. W. F. Kardaun[b] and Willem Schaafsma[a]
[a]*University of Groningen*
[b]*Max Planck Institute for Plasmaphysics*

Abstract. Given is the outcome s of $S \sim \mathrm{B}(n, p)$ (n known, p fully unknown) and two numbers $0 < a \le b < 1$. Required are probabilities $\alpha_<(s)$, $\alpha_{a,b}(s)$, and $\alpha_>(s)$ of the hypotheses $\mathrm{H}_<$: $p < a$, $\mathrm{H}_{a,b}$: $a \le p \le b$, and $\mathrm{H}_>$: $p > b$, such that their sum is equal to 1. The degenerate case $a = b (= c)$ is of special interest. A method, optimal with respect to a class of functions, is derived under Neyman–Pearsonian restrictions, and applied to a case from medicine.

1 Introduction

In this paper, we propose a method to assign probabilities to hypotheses concerning p in the context of binomial hypothesis testing. Given is the result $S = s$ of n successive i.i.d. Bernoulli(p) trials, with no prior information whatsoever about the location of p in $\Theta = [0, 1]$. We assign a number $\alpha_<(s)$ to the hypothesis $\mathrm{H}_<$: $p < a$, as a principle-dependent *estimate of the truth value* $\mathbf{1}_{[0,a)}(p)$ of $\mathrm{H}_<$: $p < a$ and extend this approach to assigning numbers $\alpha_<(s)$, $\alpha_{a,b}(s)$, and $\alpha_>(s)$ to hypotheses $\mathrm{H}_<$: $p < a$, $\mathrm{H}_{a,b}$: $a \le p \le b$, and $\mathrm{H}_>$: $p > b$. A method of inference, under the restriction that the sum of these probabilities is equal to one, is given in Section 5.

Our approach comprises of the following steps. We estimate the indicator function of the hypothesis of interest under the so-called restriction of weak unbiasedness. Finding the optimal estimate is done through minimising a proper integrated risk function, in this paper the quadratic one. As such, our approach synthesises the testing and estimation contexts. Our method is compared with other methods currently en vogue.

Historical context

Jacob Bernoulli (1713) had established that the possibility of $a \le \frac{s}{n} \le b$ is, in his words, "morally certain" if n is sufficiently large and p is such that $a \le p - 0.01$ and $b \ge p + 0.01$. He proved that the probability of the event $\{a \le n^{-1} S \le b\}$ can

Key words and phrases. Hypothesis testing, p-values, q-values, squared error loss, weak unbiasedness.

Received January 2013; accepted September 2014.

be made arbitrarily close to 1 by choosing n sufficiently large. To apply his result, p has to be known. Thomas Bayes (1763) thought about the inverse problem where p is unknown but the outcome s is known and the probability of $a \le p \le b$ has to be specified. He would have regarded the hypothesis $\mathrm{H}_{a,b}$ as "morally certain" if its "probability" is larger than 0.999, say. How to specify such kind of "probability," that is the question. Considering the degenerate case $a = b (= c)$ of a null hypothesis, Karl Pearson (1900) raised the idea that $\mathrm{P} = \mathrm{P}(\chi_1^2 \ge (s - nc)^2/(nc(1-c)))$ is a fairly reasonable criterion for the probability that H_0: $p = c$ is true.

Bayesians emphasise that the probabilities $\alpha_<(s)$, $\alpha_{a,b}(s)$ and $\alpha_>(s)$ are "assessments of probability which should correspond to posterior probabilities". In the mid-eighties, Berger and coworkers (see, e.g., Berger (2003)) emphasised that p-values and posterior probabilities tend to be considerably different, an issue known since quite some time. This motivated the last author to initiate the idea of estimating truth values with respect to (w.r.t.) squared-error loss (Schaafsma (1989)). This idea gained popularity after a series of publications by, mainly, Hwang and Casella (such as Hwang et al. (1992), Hwang and Yang (2001)).

2 Principles behind the estimation of truth values

Given some hypothesis H about p, for example, the hypothesis $\mathrm{H}_{a,b}$: $a \le p \le b$, which estimator α: $\{0, 1, \dots, n\} \to [0, 1]$ should we use for estimating $\mathbf{1}_\mathrm{H}(p)$? Let $\Theta = [0, 1]$ denote the set of theoretical possibilities θ for the true value p and let Θ_H denote the set of all θ for which H is true and let Θ_A denote that where H is false (i.e., its logical negation A is true). Let $X_\theta \sim \mathrm{B}(n, \theta)$ denote any random variable (having the same distribution as $S \sim \mathrm{B}(n, p)$ only if $p = \theta$). Classical statisticians will start out by studying the bias function

$$\mathrm{B}(\theta, \alpha) = \mathbf{E}\alpha(X_\theta) - \mathbf{1}_\mathrm{H}(\theta)$$

given an estimator α, using notation $\mathbf{1}_\mathrm{H} = \mathbf{1}_{\Theta_\mathrm{H}}$. The requirement of unbiasedness, that is, $\mathrm{B}(\theta, \alpha) = 0$ for all θ, cannot be satisfied because $\mathbf{E}\alpha(X_\theta)$ is a continuous function of θ whereas $\mathbf{1}_\mathrm{H}$ is not. As replacement, we advocate to require *weak unbiasedness*, that is,

$$\begin{cases} \mathbf{E}\alpha(X_\theta) \ge \frac{1}{2} & \text{if } \theta \in \Theta_\mathrm{H}, \\ \mathbf{E}\alpha(X_\theta) \le \frac{1}{2} & \text{if } \theta \notin \Theta_\mathrm{H}, \end{cases} \tag{2.1}$$

which implies weak similarity in the sense that $\mathbf{E}\alpha(X_a) = \mathbf{E}\alpha(X_b) = \frac{1}{2}$ if $\mathrm{H} = \mathrm{H}_{a,b}$ is considered. In the degenerate case $a = b (= c)$, we use the restriction $\mathbf{E}\alpha(X_c) = \frac{1}{2}$ of weak similarity without worrying about weak unbiasedness. (If $c = \frac{1}{2}$ then this will automatically be satisfied but otherwise a complication appears, which may perhaps be removed by requiring $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{E}\alpha(X_\theta)|_{\theta=c} = 0$.)

Subsequently, thinking in terms of estimation errors, the classical statistician may choose some loss function. We suggest $\mathrm{L}(\theta, a) = (a - \mathbf{1}_{\mathrm{H}}(\theta))^2$ such that the risk function corresponds to the mean squared error

$$\mathrm{R}(\theta, \alpha) = \big(\mathrm{B}(\theta, \alpha)\big)^2 + \mathrm{Var}\big(\alpha(X_\theta)\big).$$

This paper is based on squared-error loss because of its simplicity and properness. Some results, however, can easily be generalised to proper loss functions in general (see below). It is here that, in our view, classical statisticians should accept arguments from their Bayesian colleagues implying that, for example, absolute-error loss is "improper" which is a reason to avoid its use in estimating truth values.

Remark on properness

Savage (1951) advocated thinking in terms of a utility function $\mathrm{U}(\theta, a)$ instead of a loss function $\mathrm{L}(\theta, a)$. He also noted that, given such utility function, a loss function is obtained by specifying $\mathrm{L}(\theta, a) = (\sup_a \mathrm{U}(\theta, a)) - \mathrm{U}(\theta, a)$ as the regret w.r.t. the utility of the most profitable action. To be acceptable in the elicitation of probabilities, that is, estimation of truth values, the utility function should be *proper* in the sense that, in our special context of estimating $\mathbf{1}_{\mathrm{H}}(t)$, the expected utility $\mathbf{E}\mathrm{U}(T, a)$ is maximal as a function of a iff $a = \mathrm{P}(T \in \Theta_{\mathrm{H}})$, no matter the choice of the distribution of random variable T. The class of (strictly) proper utility functions is extremely large, much larger than the ones of the form

$$\mathrm{U}(\theta, a) = f_{\mathrm{H}}(a)\mathbf{1}_{\mathrm{H}}(\theta) + f_{\mathrm{A}}(a)\mathbf{1}_{\mathrm{A}}(\theta) \tag{2.2}$$

which will be discussed here. By definition, the properness requirement is satisfied iff $pf_{\mathrm{H}}(a) + (1-p)f_{\mathrm{A}}(a)$ is maximal as a function of a iff $a = p$ and that this holds for every $p \in [0, 1]$.

If $\mathrm{U}(\theta, a)$ is proper then the corresponding regret-loss $\mathrm{L}(\theta, a)$ is proper as well (in the sense that the expected loss $\mathbf{E}\mathrm{L}(T, a)$ is minimal if $a = \mathrm{P}(T \in \Theta)$). The utility function $\mathrm{U}(\theta, a) = (2a - a^2)\mathbf{1}_{\mathrm{H}}(\theta) + (1 - a^2)\mathbf{1}_{\mathrm{A}}(\theta)$ is most attractive in our opinion because (i) it is proper (because $(2a - a^2)p + (1 - a^2)(1 - p) = 1 - p(1 - p) - (a - p)^2$ is maximal if $a = p$) while (ii) $0 \le \mathrm{U}(\theta, a) \le 1$ is such that $\sup_a \mathrm{U}(\theta, a) = 1$ (for every fixed value of θ). Note that the regret $1 - \mathrm{U}(\theta, a) = (a - \mathbf{1}_{\mathrm{H}}(\theta))^2$ is the squared-error loss function. Estimators of truth values are called q-values if this quadratic loss function is used in their derivation.

Another proper loss function is the logarithmic one $L(\theta, a) = -\log a$ if $\theta \in \Theta_{\mathrm{H}}$, $L(\theta, a) = -\log(1 - a)$ if $\theta \in \Theta_{\mathrm{A}}$. Quadratic loss and logarithmic loss are mathematically attractive, for example, because $L(\theta, a)$ is convex as a function of a and constant as a function of θ on Θ_{H} and Θ_{A}, respectively. Moreover, they are "fair" in the sense that nothing changes if Θ_{H} is replaced by Θ_{A} and a by $1 - a$. Our experiences based on quadratic loss are not much different from those based on logarithmic loss, when the probabilities and their assignments stay away from being close to 0 or 1 (Kardaun and Schaafsma (2015)).

Given any ("regular") proper loss function, the *unrestricted* minimisation of $\int_0^1 \mathrm{R}(\theta,\alpha)\,\mathrm{d}\theta$ is achieved if $\alpha = \alpha_{\text{Bayes}}$, where

$$\alpha_{\text{Bayes}}(x) = \frac{1}{\beta(x+1, n-x+1)} \int_{\{\theta;\mathbf{1}_{\mathrm{H}}(\theta)=1\}} \theta^x (1-\theta)^{n-x}\,\mathrm{d}\theta. \tag{2.3}$$

The Bayesian will regard $\int_0^1 \mathrm{R}(\theta,\alpha)\,\mathrm{d}\theta$ as the Bayes risk w.r.t. Bayes' prior; the non-Bayesian considers this integral as the area under the risk function.

In the degenerate case $a = b = c$ of testing H_0: $p = c$ the value $\alpha_{\text{Bayes}}(x)$ assigned to H_0 is equal to 0 for all x. The consequence $\mathbf{E}\alpha_{\text{Bayes}}(X_c) = 0$ is in obvious conflict with the weak-unbiasedness requirement. In many situations, nobody will believe that p is *exactly* equal to some predetermined value c. In these cases, some dogmatism is involved in the requirement of weak unbiasedness. In other practical situations, it is very well possible that p is *exactly* equal to, for example, $c = \frac{1}{2}$. The Bayesian will assign a positive prior probability ρ to this possibility and modify α_{Bayes} accordingly. The choice $\rho = \frac{1}{2}$ of the "unbiased Bayesian" is somewhat dogmatic as well. Some Bayesians suggest to leave the choice of ρ to the problem-owner who, hence, can compute "his own" posterior probability. We prefer to choose ρ such that the corresponding Bayes estimator is weakly similar (see Section 6).

3 Testing H_0: $p = c$ by assigning a q-value

Pearson (1900) claimed (for situations much more general than the present one) that the p-value

$$\alpha_{\text{Pearson}}(s) = \mathrm{P}\left(\chi_1^2 \geq \frac{(s-nc)^2}{nc(1-c)}\right) \tag{3.1}$$

provides a "fairly reasonable" criterion for the probability that the data can be considered as having arisen from random sampling. Fisher suggested to use the "exact" p-value

$$\alpha_{\text{Fisher}}(s) = \sum_{\{x:b_{n,c}(x) \leq b_{n,c}(s)\}} b_{n,c}(x), \tag{3.2}$$

where, here and in the sequel,

$$b_{n,c}(x) = \binom{n}{x} c^x (1-c)^{n-x}. \tag{3.3}$$

In the Neyman–Pearson interpretation, the p-value is defined as the *smallest* value of the nominal level of significance for which H_0 is rejected. Modern statisticians, Bayesians (e.g., Berger (2003)) as well as non-Bayesians (e.g., Lehmann (1959)), reject the idea to regard p-values as posterior probabilities.

It is in line with Section 2 to construct an estimator α: $\{0, 1, \dots, n\} \to [0, 1]$ of the true value $\mathbf{1}_{\{c\}}(p)$ of the indicator function $\mathbf{1}_{\{c\}} : \Theta \to \{0, 1\}$ of H_0 such that α is (in some sense) optimal with respect to squared-error loss, that is,

$$\mathrm{R}(\theta, \alpha) = \begin{cases} \mathbf{E}\big(\alpha(X_\theta) - 1\big)^2 & \text{if } \theta = c, \\ \mathbf{E}\big(\alpha(X_\theta)\big)^2 & \text{if } \theta \neq c. \end{cases} \tag{3.4}$$

The requirement of weak unbiasedness is satisfied if and only if $\mathbf{E}\alpha(X_c) \geq \frac{1}{2}$ and $\mathbf{E}\alpha(X_\theta) \leq \frac{1}{2}$ if $\theta \neq c$, where $X_c \sim \mathrm{B}(n, c)$. A necessary condition for weak unbiasedness is that of weak similarity, that is, $\mathbf{E}\alpha(X_c) = \frac{1}{2}$ or, equivalently, that of continuity of the risk function.

The requirement of weak unbiasedness is in exact agreement with Lehmann's (1959) decision-theoretic concept of unbiasedness if squared-error loss is replaced by absolute-error loss. We are somewhat critical w.r.t. the mathematical niceties involved. Continuity of the mean squared error is quite restrictive. For a well-established null-hypothesis H_0, for instance, the probability c of selecting a random digit from the SETUN computer (which used ternary logic, cf. Rescher (1969)) being $\frac{1}{3}$, one would like to suppress $R(c, \alpha)$ by allowing $\mathbf{E}\alpha(X_c)$ to be somewhat larger than $\frac{1}{2}$. This is at the cost of an increase of $R_{\theta,\alpha}$ for $\theta \neq c$. This may be acceptable as many of these are comparatively small. In other situations, when H_0 is more arbitrary, it may be preferable to increase $R(c, \alpha)$ in order to reduce $R(\theta, \alpha)$ for $\theta \neq c$.

Theorem 1. *The problem of minimising*

$$\int_0^1 \mathrm{R}(\theta, \alpha)\, \mathrm{d}\theta \tag{3.5}$$

under the restrictions

$$\begin{gathered} 0 \leq \alpha(x) \leq 1 \ (x = 0, 1, \dots, n), \\ \mathbf{E}\alpha(X_c) = \frac{1}{2} \end{gathered} \tag{3.6}$$

of weak similarity has as solution

$$\alpha^*(s) = \frac{b_{n,c}(s)}{2\sum_{x=0}^{n}(b_{n,c}(x))^2}. \tag{3.7}$$

Proof. Note that

$$\int_0^1 \mathrm{R}(\theta, \alpha)\, \mathrm{d}\theta = \int_0^1 \sum_{x=0}^{n} \alpha^2(x) b_{n,\theta}(x)\, \mathrm{d}\theta = \frac{1}{n+1} \sum_{x=0}^{n} \alpha^2(x). \tag{3.8}$$

Using notations $b_n = (b_{n,c}(0), \dots, b_{n,c}(n))$ and $\alpha = (\alpha(0), \dots, \alpha(n))$, note that $\|\alpha\|^2$ is minimal under the restriction that the inner product $\langle \alpha, b_n \rangle$ is equal to $\frac{1}{2}$ if

$\alpha = \alpha^* := \frac{1}{2} b_n / \|b_n\|^2$. Solution α^* is non-negative. (Partial proofs of $\alpha^* \le 1$ were obtained and numerical verifications up to $n = 10^6$ were made, but a mathematical proof in general, as far as we know, is not available.) □

Remark 1. Note that the approximation

$$\alpha^*(x) \approx \big(\pi c(1-c)n\big)^{1/2} b_{n,c}(x) \tag{3.9}$$

is a consequence of $\sum_{x=0}^{n} (b_{n,c}(x))^2 = \mathrm{P}(X_c - X_c' = 0) \approx (2\pi 2c(1-c)n)^{-1/2}$, where X_c, $X_c' \sim \mathrm{B}(n, c)$ are independent. (Apply the local form of the CLT.) Theorem 1 can easily be generalised to the situation where the restriction $\mathbf{E}\alpha(X_c) = \frac{1}{2}$ is replaced by $\mathbf{E}\alpha(X_c) = \psi$. The normal approximation suggests that truncation is necessary if $\psi > 2^{-1/2} = 0.71$.

Remark 2. In the preceding text, we took $\Theta = [0, 1]$. When, on basis of theoretical arguments, it is impossible that, for instance, $p < c$, one-sided testing of H_0: $p = c$ against A: $p > c$ is indicated. Now, the natural parameter space is $\Theta = [c, 1]$ and we have to minimise

$$\int_c^1 \sum_{x=0}^{n} \alpha^2(x) b_{n,\theta}(x)\, \mathrm{d}\theta = (n+1)^{-1} \sum_{x=0}^{n} \alpha^2(x) \sum_{s=0}^{x} b_{n+1,c}(s) \tag{3.10}$$

under the restriction $\sum_{x=0}^{n} \alpha(x) b_{n,c}(x) = \frac{1}{2}$. This provides

$$\alpha(s) = \min\left(1, \lambda b_{n,c}(s) \Big/ \sum_{x=0}^{s} b_{n+1,c}(x)\right), \tag{3.11}$$

where λ is such that the equality constraint is satisfied.

Figure 1 presents graphs of the two-sided q-value of α^* and of the one-sided q-value α ($\lambda = 1.0345$, following from a numerical analysis employing the bisection method) both for testing H_0: $p = c$ in case $c = \frac{1}{2}$ and $n = 12$. These graphs specify the entire "Bayes-optimal weakly-similar" method of inference. For example, $s = 10$ the exact value

$$\alpha^*(10) = \frac{2^{12}}{2} \binom{12}{10} \Big/ \sum_{x=0}^{12} \binom{12}{x}^2 = 2^{11} \binom{12}{10} \Big/ \binom{24}{12} = \frac{33{,}792}{676{,}039} \tag{3.12}$$

which can be approximated via

$$\pi^{1/2} 3^{1/2} \binom{12}{10} 2^{-12} = 0.0495. \tag{3.13}$$

Note that we have used the identity $\sum_{x=0}^{n} \binom{n}{x}^2 = \binom{2n}{n}$ the validity of which is

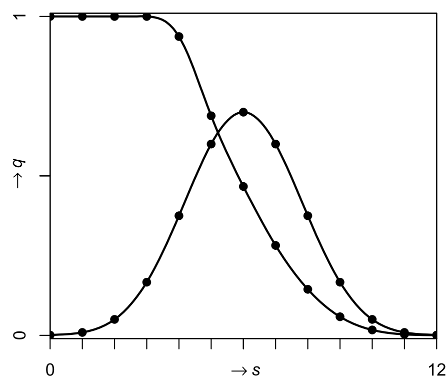

Figure 1 *The optimal estimator α^* of Theorem 1 in the case $n = 12$, $c = \frac{1}{2}$ (symmetric) and the optimal estimator α of Remark 2 in the same case if $p < \frac{1}{2}$ is considered impossible (decreasing).*

directly derived by calculating the coefficient of x^0 of $(x+1/x)^{2n} = ((x+1/x)^n)^2$, see, for example, Greene and Knuth (1982).

The one-sided q-value

$$\begin{aligned} \alpha(10) &= 1.0345 \binom{12}{10} 2^{-12} \Big/ \left(1 - \left(\binom{13}{11} + \binom{13}{12} + \binom{13}{13}\right) 2^{-13}\right) \\ &= 0.0169 \end{aligned} \tag{3.14}$$

indicates that the value $\alpha^*(10) = 0.05$ is considerably decreased if the possibility that $p < \frac{1}{2}$ is deemed impossible. These assessments $\alpha^*(10) = 0.050$ and $\alpha(10) = 0.017$ have, in our view, a better foundation than the corresponding (exact) p-values

$$2\left(\binom{12}{10} + \binom{12}{11} + \binom{12}{12}\right) 2^{-12} = 0.039 \tag{3.15}$$

and than half of it, that is, 0.019, respectively.

Discussion

The differences between optimal q-values and the corresponding p-values are not very alarming in this example.

4 Testing $\mathbf{H}_<$: $p < c$ by assigning a q-value

It is often practically impossible that p is exactly equal to c (see Section 6 for a counter example). In such situations, we may try to estimate $\mathbf{1}_{[0,c)}(p)$ or $\mathbf{1}_{[0,c]}(p)$ w.r.t. squared-error loss. First, we note that the unrestricted minimisation of $\int_0^1 \mathrm{R}(\theta, \alpha)\, \mathrm{d}\theta$ provides, using the combinatorial identity, from Pascal's triangle, $\binom{n+1}{x} = \binom{n}{x} + \binom{n}{x-1}$,

$$\begin{aligned}\alpha_{\text{Bayes}}(s) &= \int_0^c \theta^s (1-\theta)^{n-s}\, \mathrm{d}\theta / \beta(s+1, n-s+1) \\ &= \sum_{x=s}^{n+1} b_{n+1,c}(x) \\ &= (1-c) \sum_{x=s}^{n} b_{n,c}(x) + c \sum_{x-1=s}^{n} b_{n,c}(x) \\ &= (1-c) b_{n,c}(s) + \sum_{x=s+1}^{n} b_{n,c}(x).\end{aligned} \tag{4.1}$$

Under the restriction $\mathbf{E}\alpha(X_c) = \frac{1}{2}$ of weak similarity (or continuity of the risk function) one obtains

$$\alpha_{\text{ws}}(c, s) = \frac{1}{2} b_{n,c}(s) + \sum_{x=s+1}^{n} b_{n,c}(x) \tag{4.2}$$

which differs from $\alpha_{\text{Bayes}}(s)$ if $c \neq \frac{1}{2}$. As α_{ws} is a decreasing function of s, it is not only weakly similar but also weakly unbiased. It minimises the Bayes risk under the restriction of weak unbiasedness (the proof goes along the lines of Section 5) and should, therefore, be acceptable for the unbiased Bayesian. At present, there is a considerable number of papers (cf. Hwang et al. (1992), Hwang and Yang (2001), Wells (2010)) considering α_{ws} and several of its generalisations.

Figure 2 provides a graph of $\alpha_{\text{Bayes}} = \alpha_{\text{ws}}$ for the case $(n, c) = (12, \frac{1}{2})$. It is interesting to compare this graph (black line) with that of the one-sided q-value

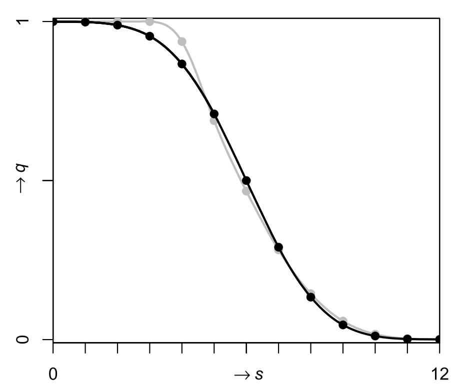

Figure 2 *The one-sided q-value $\alpha_< = \alpha_{\text{Bayes}} = \alpha_{\text{ws}}$ estimating $\mathbf{1}_{[0,1/2]}$, as given in, for example, (4.2), (grey) and the one-sided q-value α estimating $\mathbf{1}_{\{1/2\}}$ from Figure 1 (black).*

α (decreasing) of Figure 1. Both are such that $\mathbf{E}\alpha(X_{1/2}) = \mathbf{E}\alpha_<(X_{1/2}) = \frac{1}{2}$. Note that $\alpha_{\text{Bayes}}(10) = 0.017$, whereas $\alpha_{\text{ws}}(10) = 0.011$.

5 Settling a crucial issue

Whether one uses p-values or optimal q-values, the estimates $\alpha_<(s)$, $\alpha_0(s)$ and $\alpha_>(s)$ of $\mathbf{1}_{[0,c)}(p)$, $\mathbf{1}_{\{c\}}(p)$ and $\mathbf{1}_{(c,1]}(p)$ derived in Sections 3 and 4 are not probabilistically coherent because their sum $1 + \alpha_0(s)$ is larger than 1. This lack of probabilistic coherency is not something to worry about too much if $\alpha_0(s)$ is very small, as it will be in the case of Section 6. Nor is it of interest if utilisations are discussed based on choosing a decision d such that some expectation $c_<(d)\alpha_<(s) + c_0(d)\alpha_0(s) + c_>(d)\alpha_>(s)$ is optimal. Nevertheless, it is inconvenient that q-values and p-values display a lack of probabilistic coherency. This difficulty does not occur if, for example, a Bayesian approach is chosen with prior $\rho\varepsilon_{\{c\}} + (1-\rho)\mu$ where $\mu \sim \text{U}(0,1)$ is Bayes' prior. The choice of ρ (and μ) is a hot issue. In this section, we present an alternative theory for constructing estimates $\alpha_<(s)$, $\alpha_{a,b}(s)$, $\alpha_>(s)$ under the additional restriction of probabilistic coherency: $\alpha_<(s) + \alpha_{a,b}(s) + \alpha_>(s) = 1$ (see the end of Section 6 for a comparison with the Bayesian approach).

To arrive at a specific solution, we concentrate attention on an integrated risk, for example, $\int_0^1 \text{R}(\theta,\alpha)\,\text{d}\theta$ where $\alpha = (\alpha_<, \alpha_{a,b}, \alpha_>)$. Bayesians may emphasize that, given any proper loss function (such that $\inf_\alpha \int_0^1 \text{R}(\theta,\alpha)\,\text{d}\theta < \infty$), the integrated risk $\int_0^1 \text{R}(\theta,\alpha)\,\text{d}\theta$ is minimal if α contains the probabilities prescribed by the Beta$(s+1, n-s+1)$ distribution. A drawback of this approach is that $\alpha_{a,b}(s)$ will be tiny (i.e., "incredibly" small) if $b-a$ is very small (or even 0, if $a = b(=c)$). This shows that the Bayesian has to replace Bayes' prior by something else depending on a and b (see the beginning of this section). We concentrate attention on the proper loss function

$$\begin{aligned} \text{L}_w(\theta,\alpha(x)) &= \big(\mathbf{1}_{[0,a)}(\theta) - \alpha_<(x)\big)^2 + w\big(\mathbf{1}_{(b,1])}(\theta) - \alpha_>(x)\big)^2, \\ w &> 0 \end{aligned} \tag{5.1}$$

(which has its origin in Epstein (1969)) and restrict attention to the class of Lehmann-unbiased rules $\alpha = (\alpha_<, \alpha_>)$. This condition is satisfied if and only if $\alpha_<$ is a weakly-unbiased estimator of $\mathbf{1}_{[0,a)}(p)$ and $\alpha_>$ is a weakly-unbiased estimator of $\mathbf{1}_{(b,1]}(p)$. It can be established, for any $w > 0$ that $\int_0^1 \text{R}_w(\theta,\alpha)\,\text{d}\theta$ is minimal under the restriction of Lehmann-unbiasedness if $\alpha_<(s)$, $\alpha_{a,b}(s) = 1 - \alpha_<(s) - \alpha_>(s)$, and $\alpha_>(s)$ are the probabilities prescribed by the $\frac{1}{2}$ Beta$(s, n-s+1) + \frac{1}{2}$ Beta$(s+1, n-s)$ distribution.

(Proofs are not presented because it is unsatisfactory that $\alpha_{a,b}(s)$ is tiny if $b-a$ is very small. The underlying distributional inference $Q(s) = \frac{1}{2}$ Beta$(s, n-s+$

$1) + \frac{1}{2}\,\mathrm{Beta}(s+1, n-s)$ has its origin in Kroese (1994), Kroese et al. (1995) and Salomé (1994).) The loss function L_w allows, under certain conditions, that *the task of constructing $\alpha_<$ can be separated from the task of constructing $\alpha_>$* which is certainly convenient. Strict properness of L_w follows from the fact that, if T is some real-valued random variable with $\mathrm{P}(T < a) = p_<$, $\mathrm{P}(a \le T \le b) = p_{a,b}$, and $\mathrm{P}(T > b) = p_>$, then

$$\begin{aligned}
&\mathbf{E}\mathrm{L}_w(T, a) \\
&\quad = (1-a_<)^2 p_< + a_<^2(1-p_<) + w\big((1-a_>)^2 p_> + a_>^2(1-p_>)\big) \qquad (5.2) \\
&\quad = p_<(1-p_<) + (a_< - p_<)^2 + w p_>(1-p_>) + w(a_> - p_>)^2
\end{aligned}$$

is minimal as a function of $(a_<, a_>)$ iff $a_< = p_<$ and $a_> = p_>$. To avoid the difficulty that $\alpha_{a,b}(s)$ will be tiny if $b - a$ is very small, the Neyman–Pearsonian will be attracted by the idea to impose some restriction on the class of estimators $\alpha = (\alpha_<, \alpha_{a,b}, \alpha_>)$, for example, that to the class $\mathscr{D}_{1,\psi_2}$ of rules satisfying

$$\mathbf{E}\alpha_<(X_a) = \psi_1, \qquad \mathbf{E}\alpha_>(X_b) = \psi_2, \tag{5.3}$$

where $\psi_1, \psi_2 \in [0, 1]$ should satisfy $\psi_1 + \psi_2 \le 1$ (at least if $a = b = c$).

Theorem 2. *The Bayes risk $\int_0^1 \mathrm{R}_w(\theta, \alpha)\,\mathrm{d}\theta$ is minimal under the restriction $\alpha \in \mathscr{D}_{\psi_1\psi_2}$ if and only if $\alpha_<$ minimises*

$$\int_0^1 \mathbf{E}\big(\mathbf{1}_{[0,a)} - \alpha_<(X_\theta)\big)^2\,\mathrm{d}\theta \tag{5.4}$$

under the restriction $\mathbf{E}\alpha_<(X_a) = \psi_1$ and $\alpha_>$ minimises

$$\int_0^1 \mathbf{E}\big(\mathbf{1}_{(b,1]} - \alpha_>(X_\theta)\big)^2\,\mathrm{d}\theta \tag{5.5}$$

under the restriction $\mathbf{E}\alpha_>(X_b) = \psi_2$, provided that the q-values $\alpha_<^*$ and $\alpha_>^*$ thus obtained satisfy $\alpha_<^* + \alpha_>^* \le 1$.

Proof. Trivial. □

Lemma 1. *For $\psi_1, \psi_2 \le \frac{1}{2}$, the optimal q-values $\alpha_<^*$ and $\alpha_>^*$ are given by*

$$\alpha_<^*(s) = \max\left(\sum_{x=s+1}^{n} \binom{n}{x} a^x (1-a)^{n-x} + \lambda_1 \binom{n}{s} a^s (1-a)^{n-s}, 0\right), \tag{5.6}$$

where λ_1 is such that $\mathbf{E}\alpha_<^*(X_a) = \psi_1$, and

$$\alpha_>^*(s) = \max\left(\sum_{x=0}^{s-1} \binom{n}{x} b^x (1-b)^{n-x} + \lambda_2 \binom{n}{s} b^s (1-b)^{n-s}, 0\right), \tag{5.7}$$

where λ_2 is such that $\mathbf{E}\alpha_>^*(X_b) = \psi_2$.

Proof. See the Appendix. □

If either ψ_1 or ψ_2 is larger than $\frac{1}{2}$ (a theoretical possibility which we consider to be "morally rejectable" as it implies that the risk at a boundary point is smaller than that in its neighbourhood), then in (5.6) and (5.7) $\max(\dots, 0)$ should be replaced by $\min(\dots, 1)$. To apply Theorem 2, we have to verify that $\alpha^*_< + \alpha^*_> \leq 1$. If $\psi_1 = \psi_2 = \frac{1}{2}$ then truncation is not necessary, $\lambda_1 = \lambda_2 = \frac{1}{2}$ and one obtains the solution with $p_<(x) = \{Q(x)\}([0, c))$ and $Q(x) = \frac{1}{2}\,\mathrm{Beta}(x, n - x + 1) + \frac{1}{2}\,\mathrm{Beta}(x + 1, n - x)$ discussed before. If ψ_1 and ψ_2 are smaller than $\frac{1}{2}$, then it follows from the above that $\alpha^*_< + \alpha^*_> < 1$.

As an example, for $(n, c) = (12, \frac{1}{2})$ numerical analyses were made for various values of $\psi_1 = \psi_2 (= \psi)$. Results for $\psi = \frac{1}{2}$ were already presented in Figure 2. Figure 3 presents results for $\psi = \frac{1}{4}$. Note that, in this case $\mathbf{E}\alpha^*_0(X_c) = \frac{1}{2}$.

6 Practical example and discussion

In 2008, the Dutch Acute Pancreatitis Study Group (Besselink et al. (2008)) was astounded when they analysed the results of their carefully designed double-blind comparison of "placebo" and "treatment" based on a about 150 patients from each of the two categories. They had assumed that the beneficial effects would outweigh the possible risks of complications for these—severely ill—patients. The interim analysis (at half-way) did not point to any significant difference. In the final analysis they observed $s = 9$ deaths in the placebo group and $t = 24$ deaths in the treatment group. A Poisson approximation followed by conditioning w.r.t. $n = s + t = 33$ (a more precise analysis is discussed in Kardaun and Schaafsma (2015)) leads us to the problem of discussing truth or falsity of hypotheses like

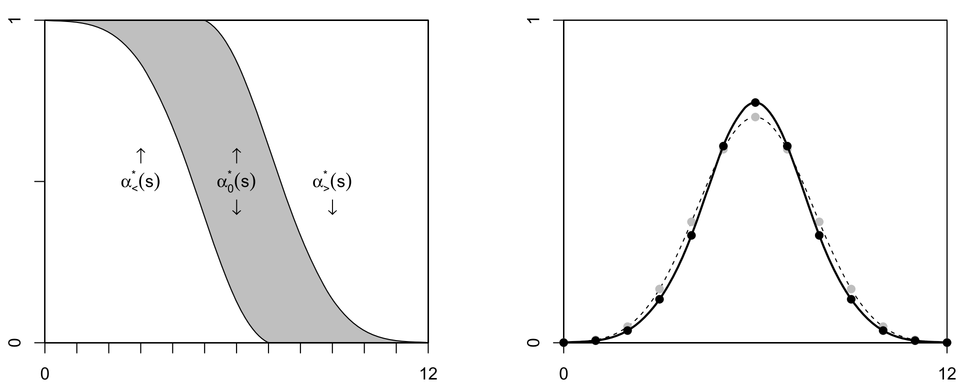

Figure 3 *Left: a display of α^* for $\psi = \frac{1}{4}$ for $(n, c) = (12, \frac{1}{2})$. Right: a display of the optimal two-sided q-value α^* derived in Section 3 (see Figure 1) (dashed), and the q-value α^*_0 derived as part of the coherent system displayed on the left.*

$\mathrm{H}_<$: $p < \frac{1}{2}$, H_0: $p = \frac{1}{2}$ and $\mathrm{H}_>$: $p > \frac{1}{2}$, given the outcome $s = 9$ of $S \sim \mathrm{B}(33, p)$, p denoting the probability that a case of death is from the placebo group. The research workers had taken for granted, a priori, that treatment would be beneficial, that is fewer cases of death in the treatment group. Their intention was to ignore the possibility that $p < \frac{1}{2}$ and, hence, to test H_0: $p = \frac{1}{2}$ versus the one-sided alternative $\mathrm{H}_>$: $p > \frac{1}{2}$. The experimental results, however, were in the opposite direction. Thus, they replaced their intention to test H_0 versus $\mathrm{H}_>$: $p > \frac{1}{2}$ by a two-sided testing approach. (This case study provides a clear indication that one-sided testing is only appropriate if one is not only "morally" certain but "absolutely" certain that the inequality constraint imposed, here $p \geq \frac{1}{2}$, is valid.)

Discussion

This problem is of particular interest because it is relevant to discuss whether or not p is *exactly* equal to $\frac{1}{2}$ (see also the "concluding remarks"). The two-sided p-value

$$\alpha_{\mathrm{Fisher}}(9) = \left(\sum_{x=0}^{9} + \sum_{x=24}^{33}\right) \binom{33}{x} \left(\frac{1}{2}\right)^{33} = 0.0135 \tag{6.1}$$

is such that H_0 is rejected at the 5% level. (The UMPU size 5% and UMPI size 5% Neyman–Pearson tests are such that H_0 would even have been rejected if $s = 10$ and, with some type of randomisation probability, even if $s = 11$.) To express statistical uncertainty in such accept—reject statement, the neo-Bayesian claim that p-values tend to be misleadingly small whenever they are small. If one computes the posterior probability

$$\alpha_{\mathrm{Bayes}}(9) = \frac{\binom{33}{9} 2^{33}}{\binom{33}{9} 2^{33} + 1/34} = 0.132 \tag{6.2}$$

based on $\rho = \frac{1}{2}$ and $\mu \sim \mathrm{U}(0, 1)$, then the controversy is obvious.

Classical statisticians reject the idea that (p, s) is the outcome of a pair $(\tilde{T}, \tilde{X})$ of random variables with joint distribution such that $\mathcal{L}(\tilde{T}) = \rho\varepsilon_c + (1-\rho)\mu$ and $\mathcal{L}(\tilde{X}|\tilde{T} = \theta) = \mathrm{B}(n, \theta)$.

In this paper, we are fascinated by the idea that $\mathbf{1}_c(p)$ should be estimated w.r.t. squared-error loss. The Bayes-optimal weakly similar q-value $\alpha^*(9)$ of Section 3 provides

$$\alpha^*_{\mathrm{ws}}(9) = \frac{\binom{33}{9} 2^{33}}{2\sum_{x=0}^{33} \binom{33}{x}^2} = \frac{(1/2)\binom{33}{9}}{\binom{66}{33}} = 0.0229. \tag{6.3}$$

(Kardaun and Schaafsma (2015) contains several alternative estimators, as well as generalisations and some asymptotic theory. They suggest that the estimate $\alpha^*_{\mathrm{ws}}(9)$

just obtained is, perhaps, a bit too small because weak similarity is bit too conservative.)

A subject-matter remark is that a particularly high mortality rate was found in the subgroup of patients with bowel ischemia, who, all of them, received probiotics. (As described in the paper, this diagnosis was made within two weeks of admission to the hospitals.) From nine of these patients, only one survived. Hence, subtracting the eight patients of this subgroup who did pass away, one gets for the mortality in the non-bowel-ischemia group: $(\text{probiotics}, \text{placebo}|\text{total}) = (16, 9|25)$ instead of $(24, 9|33)$ for the entire group. Therefore, in the absence of ischemia, the statistical significance is considerably less, the hypothesis $p = \frac{1}{2}$ is not traditionally rejected at the two-sided 5% level, even based on a simple normal approximation with continuity correction (which can be calculated by hand): $Z = |16 - 12.5 - \frac{1}{2}|/\sqrt{25/4} \simeq 1.2$. As an approximation to $\mathrm{P}(\{X \leq 9\} \cup \{X \geq 16\})$ this leads to the two-sided p-value $\mathrm{P}\{|Z| > 1.2\} = 23\%$ ("exact": 22.95%) instead of 1.5% ("exact" p-value: 1.35%) for the entire group. The weakly-similar q-value (see Theorem 1) gives in this case $\tilde{\alpha}_{25}(9) = \binom{25}{9} 2^{24} / \binom{50}{25} = 27\%$. Without any continuity correction, to obtain an approximation of $\frac{1}{2}(\mathrm{P}(\{X \leq 9\} \cup \{X \geq 16\}) + (\mathrm{P}\{X < 9\} \cup \{X > 16\}))$, one gets $Z \simeq 1.4$ for the $(16, 9|25)$ sub-group and $Z \simeq 2.6$ for the entire group of patients, with corresponding two-sided mid p-values of 16% and 0.9%, respectively. For some further discussion on continuity correction and mid p-values, see Hirji et al. (1991) and Lancaster (1961).

Though several subgroup analyses were performed and the two-week survival rates of the patients that did not develop bowel ischaemie was given, the longer term survival ratio was not precisely put under the "candle light" (from Rao (1999)) in the original article in *The Lancet*.

Concluding remarks

It is an inconvenient truth that we, mathematical statisticians, advertise our findings using words like optimal, rational, coherent, admissible, unbiased, etc., made specific in mathematical form and, hence, suggesting that anything else is sub-optimal, irrational, etc., while there are many reasons to doubt. In this paper, we derived an optimal method to coherently specify estimators $\alpha_<(s)$, $\alpha_0(s)$, $\alpha_>(s)$ on basis of a Neyman–Pearsonian choice of two numbers ψ_1 and ψ_2. Choosing $\psi_1 = \psi_2 = \frac{1}{2}$ may be interesting if one accepts $\alpha_0 \equiv 0$. Dealing with the problem of the present section, the choice $\psi_1 = \psi_2 = \frac{1}{4}$ leads to the Bayes-optimal triple $\alpha^* = (\alpha^*_<, \alpha^*_0, \alpha^*_>)$ with

$$\alpha^*_{33;<}(9) = 0.9832, \qquad \alpha^*_{33;0}(9) = 0.0168, \qquad \alpha^*_{33;>}(9) = 0.0000. \tag{6.4}$$

The choice $\psi = \frac{1}{4}$ leads to similar results as the requirement of weak similarity $\mathbf{E}\alpha_0(X) = \frac{1}{2}$. Preoccupation with H_0: $p = \frac{1}{2}$ implies that this requirement is not completely satisfactory because one would like to have that the average estimate of $\mathbf{1}_{\{1/2\}}(p)$ is closer to 1, if H_0: $p = \frac{1}{2}$ is true, than $\frac{1}{2}$. This implies that there are also reasons to choose ψ somewhat smaller than $\frac{1}{4}$. For instance, when $\psi = 0.2$, $\alpha_0(9) = 0.0206$. Finally, it is mentioned that the probiotica case study from 2008 also illustrates the importance of investigating the variability of binomial probabilities associated with subgroup selection.

Appendix: Proof of Lemma 1

A reference to the fundamental lemma in the q-value approach in Schaafsma et al. (1989) might do. This reference, however, did not provide a full proof (because three different approaches were available). We present an explicit proof for the special situation of this paper. Note that Fubini's theorem provides

$$\int_0^1 \mathbf{E}\big(\mathbf{1}_{[0,a)} - \alpha_<(X_\theta)\big)^2 \,\mathrm{d}\theta = \sum_{s=0}^{n} \binom{n}{s} \bigg(\big(1-\alpha_<(s)\big)^2 \int_0^a \theta^s (1-\theta)^{n-s}\,\mathrm{d}\theta + \big(\alpha_<(s)\big)^2 \int_a^1 \theta^s (1-\theta)^{n-s}\,\mathrm{d}\theta \bigg). \tag{A.1}$$

Using

$$\int_0^a \frac{\theta^s(1-\theta)^{n-s}}{\beta(s+1, n-s+1)}\,\mathrm{d}\theta = \sum_{x=s+1}^{n+1} b_{n+1,a}(x) \tag{A.2}$$

we have to solve the optimisation problem

$$\text{Minimise} \sum_{s=0}^{n} \bigg(\big(1-\alpha_<(s)\big)^2 \sum_{x=s+1}^{n+1} b_{n+1,a}(x) + \big(\alpha_<(s)\big)^2 \sum_{x=0}^{s} b_{n+1,a}(x) \bigg) \quad \big(\stackrel{\text{def.}}{=} M_{\alpha_<}(s)\big) \tag{A.3}$$

subject to

$$0 \le \alpha_<(s) \le 1 \ (s = 0, \ldots, n) \tag{A.4}$$

$$\sum_{s=0}^{n} \alpha_<(s) b_{n,a}(s) = \psi_1. \tag{A.5}$$

Ignoring the (trivial) inequality constraints, we derive, by writing down a Lagrangian and differentiation w.r.t. $\alpha_<(s)$ ($s = 0, 1, \ldots, n$), that the minimum will be achieved if

$$-2\big(1-\alpha_<(s)\big) \sum_{x=s+1}^{n+1} b_{n+1,a}(x) + 2\alpha_<(s) \sum_{x=0}^{s} b_{n+1}(x) + \lambda b_{n,a}(s) = 0 \tag{A.6}$$

which leads to

$$\begin{aligned}\alpha_<(s) &= \sum_{x=s+1}^{n+1} b_{n+1,a}(x) - \frac{1}{2}\lambda b_{n,a}(s)\\ &= \sum_{x=s+1}^{n} b_{n,a}(x) + b_{n,a}(s)\Big(a - \frac{1}{2}\lambda\Big),\end{aligned} \tag{A.7}$$

where λ has to be determined such that

$$\begin{aligned}\sum_{s=0}^{n} \alpha_<(s) b_{n,a}(s) &= \mathrm{P}\big(B_{n,a} > B'_{n,a}\big)\\ &\quad + \Big(a - \frac{1}{2}\lambda\Big)\mathrm{P}\big(B_{n,a} = B'_{n,a}\big) = \psi_1.\end{aligned} \tag{A.8}$$

Here $B_{n,a}$, $B'_{n,a} \sim \mathrm{B}(n,a)$ are taken to be independent. If $\psi_1 = \frac{1}{2}$, then the equality is satisfied if $(a - \lambda/2) = \frac{1}{2}$. One obtains α_{sim} where truncation is not necessary. If, however, ψ_1 is less than $\frac{1}{2}$, then $a - \lambda/2$ will have to be smaller than $\frac{1}{2}$ and, usually, smaller than 0 with as a consequence that the inequality constraint $\alpha_<(s) \geq 0$ may be violated. Note that $\alpha^*_<$ is then well-defined. If $\psi_1 > \frac{1}{2}$, then $\alpha - \frac{1}{2}\lambda$ has to be larger than $\frac{1}{2}$ and a different type of truncation is needed as indicated in the discussion after Lemma 1. To establish that truncation according to $\alpha^*_<(s)$ (with λ replacing $a - \lambda/2$ constant) provides the solution, suppose that $\alpha_<(s) = \alpha^*_<(s) + \varepsilon(s)$ is another vector satisfying the constraints. Note that $\alpha^*_<(s) = 0$ implies $\varepsilon(s) \geq 0$ and that $\sum_{s=0}^{n} \varepsilon(s) b_{n,a}(s) = 0$. To establish that $M_{\alpha^*_<}(s) < M_{\alpha_<}(s)$, note that $M_{\alpha_<}(s) - M_{\alpha^*_<}(s)$ is equal to

$$\begin{aligned}&\sum_{s=0}^{n} \big(\varepsilon(s)\big)^2 - 2\sum_{s=0}^{n} \varepsilon(s)\big(1 - \alpha^*_<(s)\big) \sum_{x=s+1}^{n+1} b_{n+1,a}(x)\\ &\quad + 2\sum_{s=0}^{n} \varepsilon(s)\alpha^*_<(s) \sum_{x=0}^{s} b_{n+1,a}(x).\end{aligned} \tag{A.9}$$

It suffices to prove

$$\sum_{s=0}^{n} \varepsilon(s)\Bigg(\alpha^*_<(s)\sum_{x=0}^{s} b_{n+1,a}(x) - \big(1 - \alpha^*_<(s)\big)\sum_{x=s+1}^{n+1} b_{n+1,a}(x)\Bigg) \geq 0, \tag{A.10}$$

or, equivalently that

$$\sum_{s=0}^{n} \varepsilon(s)\alpha^*_<(s) \geq \sum_{s=0}^{n} \varepsilon(s) \sum_{x=s+1}^{n+1} b_{n+1,a}(x). \tag{A.11}$$

Splitting the summation in the l.h.s. into cases with $\alpha^*_<(s) > 0$ and cases with $\alpha^*_<(s) = 0$, we have that the first sum is equal to

$$\begin{aligned}
&\sum_{\{s;\alpha^*_<(s)>0\}} \varepsilon(s)\left(\sum_{x=s+1}^{n} b_{n,a}(x) + \lambda b_{n,a}(s)\right) \\
&\quad = \sum_{\{s;\alpha^*_<(s)>0\}} \varepsilon(s)\left(\sum_{x=s+1}^{n+1} b_{n+1,a}(x) + (\lambda - a) b_{n,a}(s)\right)
\end{aligned} \tag{A.12}$$

while the second sum satisfies

$$\sum_{\{s;\alpha^*_<(s)=0\}} \varepsilon(s)\alpha^*_<(s) \geq \sum_{\{s;\alpha^*_<(s)=0\}} \varepsilon(s)\left(\sum_{x=s+1}^{n} b_{n,a}(x) + \lambda b_{n,a}(s)\right), \tag{A.13}$$

because $\alpha^*_<(s) = 0$ implies $\varepsilon(s) \geq 0$ while $0 = \alpha^*_<(s) \geq \sum_{x=s+1}^{n} b_{n,a}(x) + \lambda b_{n,a}(s)$. Combining these sums provides

$$\begin{aligned}
\sum_{s=0}^{n} \varepsilon(s)\alpha^*_<(s) &\geq \sum_{\{s;\alpha^*_<(s)=0\}} \varepsilon(s)\left(\sum_{x=s+1}^{n} b_{n,a}(s) + \lambda b_{n,a}(s)\right) \\
&\qquad + (\lambda - a) \sum_{\{s;\alpha^*_<(s)>0\}} \varepsilon(s) b_{n,a}(s) \\
&= \sum_{\{s;\alpha^*_<(s)=0\}} \varepsilon(s)\left(\sum_{x=s+1}^{n} b_{n,a}(s) + a b_{n,a}(s)\right) \\
&\qquad + (\lambda - a) \sum_{s=0}^{n} \varepsilon(s) b_{n,a}(s).
\end{aligned} \tag{A.14}$$

The first term is nonnegative because $\varepsilon(s) \geq 0$ and the second term is equal to 0.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect those of their affiliations. The authors sincerely thank an anonymous referee for his or her helpful comments and suggestions.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* **53**, 370–418. Reprinted in *Biometrika* **45** (1958), 293–315. (With an introduction by G. A. Barnard.)

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? (with discussion). *Statist. Sci.* **18**, 1–32. MR1997064

Bernoulli, J. (1713). *Ars Conjectandi. Accedit tractatus de seriebus infinitis, et epistola Gallice scripta de ludo pilæ reticularis*. Basel: Impensis Thurnisiorum. Translation into German as "*Wahrscheinlichkeitsrechtnung (Ars Conjectandi)*" and reprinted by R. Haussner (1899) (162 pages). Leipzig: Verlag von Wilhelm Engelmann.

Besselink, M. G. H., et al. for the Dutch Acute Pancreatitis Study Group (2008). Probiotic prophylaxis in predicted severe acute pancreatitis: A randomised, double-blind, placebo-controlled trial (with discussion). *The Lancet* **371**, 651–659.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* **8**, 985–987.

Greene, D. H. and Knuth, D. E. (1982). *Mathematics for the Analysis of Algorithms*, Chapter 1, 2nd ed. Basel: Birkhäuser.

Hirji, K. F., Tan, S.-J. and Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Stat. Med.* **10**, 1137–1153.

Hwang, J. T., Casella, G., Robert, C., Wells, M. and Farrell, R. (1992). Estimation of accuracy of testing. *Ann. Statist.* **20**, 490–509. MR1150356

Hwang, J. T. and Yang, M. (2001). An optimality theory for mid p-values in 2×2 contingency tables. *Statist. Sinica* **11**, 807–826. MR1863164

Kardaun, O. J. W. F. and Schaafsma, W. (2015). *Distributional Inference*. In preparation.

Kroese, A. H. (1994). Distributional inference: A loss function approach. Ph.D. thesis, Univ. Groningen.

Kroese, A. H., Van der Meulen, E. A., Poortema, K. and Schaafsma, W. (1995). Distributional inference. *Stat. Neerl.* **49**, 63–82. MR1333179

Lancaster, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56**, 223–234. MR0124107

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*, 1st ed. New York: Wiley. 3rd ed. with J. P. Romano (1997). Heidelberg: Springer.

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* **50**, 157–175. Reprinted in *Breakthroughs in Statistics, Vol. II* (N. L. Johnson and K. Kotz, eds.) 1–11. (With an introduction by G. A. Barnard.)

Rao, C. R. (1999). *Statistics and Truth: Putting Chance to Work*, 2nd ed. River Edge, NJ: World Scientific Publishing Co. MR1474730

Rescher, N. (1969). *Many-Valued Logic*. New York: McGraw Hill.

Salomé, D. (1994). Statistical inference via fiducial methods. Ph.D. thesis, Univ. Groningen.

Savage, L. J. (1951). The theory of statistical decision. *J. Amer. Statist. Assoc.* **46**, 55–67.

Schaafsma, W. (1989). Discussing the truth or falsity of a statistical hypothesis H and its negation A. In *Proceedings of the International Workshop on Theory and Practice in Data Analysis*. *Rep. Math.* **89**, 150–166. Berlin: Akademie der Wissenschaften der DDR. MR1015754

Schaafsma, W., Tolboom, J. and Van der Meulen, E. A. (1989). Discussing truth of falsity by computing a q-value. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.) 85–100. Amsterdam: North-Holland. MR1089626

Wells, M. T. (2010). Optimality results for mid p-values. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*. *IMS Collections* **6**, 184–198. Beachwood, OH: IMS. MR2798519

C. Albers
Department of Psychometrics and Statistics
University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
The Netherlands
E-mail: c.j.albers@rug.nl
URL: www.casperalbers.nl

O. Kardaun
Max Planck Institute for Plasmaphysics
Boltzmannstrasse 2
85748 Garching
Germany
E-mail: otto.kardaun@ipp.mpg.de

W. Schaafsma
Department of Mathematics
University of Groningen
P.O. Box 800
9700 AC Groningen
The Netherlands