Foundational issues in statistical inference

C.J. Albers
Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen,
P.O. Box 800, NL-9700 AZ Groningen, The Netherlands.
Corresponding author, c.j.albers@rug.nl

O.J.W.F. Kardaun
Max-Planck Institut für Plasmaphysik, Boltzmannstraße 2, D-85748 Garching, Germany

W. Schaafsma
Institute of Mathematics and Computing Science, University of Groningen

A.G.M. Steerneman
Department of Econometrics, University of Groningen

A. Stein
Wageningen University, P.O. Box 16, NL-6700AA Wageningen, The Netherlands, and
ITC, P.O. Box 6, NL-7500AA Enschede, The Netherlands

Abstract

Statistical inference is about using statistical data (x) to formulate an opinion about something that is defined well, but unknown (y). Testing a hypothesis H about y is one of the possibilities, the estimation or prediction of y is another one. We concentrate the attention on estimation or prediction in the sense that an opinion is required in the form of a probability distribution $Q = Q(x)$ on the space $\mathscr{Y}$ of all theoretical possibilities.

The data x being statistical, it is natural to incorporate probabilistic arguments in the context to let x speak about y. Assuming that (x, y) is the outcome of a pair (X, Y) of random variables (in the sense of probability theory), the 'true' distribution P of (X, Y) exists. It may be exactly known in simulations and in thought experiments, but it is only partially known in real-world investigations. That is why the context to let x speak about y will involve at least some specification of a family $\mathscr{P} = \{P_\theta;\ \theta \in \Theta\}$ of theoretically possible P's. We assume that the probabilistic aspects of the situation are sufficiently convincing to aim at a probabilistic form of the opinion about y, given nature's message x and 'the context'.

If a probability statement is needed about some hypothesis H with respect to y, then we construct an estimator or predictor α of the truth value of H and, if the estimator seems reasonable, we use $\alpha(x)$ as the (epistemic) probability of H.

If a distributional inference is needed about a *real-valued* unknown y then, apart from using the

Bayesian approach, we can construct an inference by defining its distribution function G_x such that, for any $z \in \mathbb{R}$, $G_x(z)$ is equal to $\alpha_z(x)$ where α_z is some estimator of the truth value of $\mathrm{H}_z : y \leq z$. We claim that it is appropriate, in this respect, to use the p-value to estimate the truth value of H_z (if y is the true value of a parametric function). The imposed probabilistic coherency should not be taken too seriously because the underlying restriction of strong similarity, the Fisherian requirement that $G_X(Y)$ follows a $\mathcal{U}(0,1)$ distribution, may be 'very' reasonable if a distributional inference about y is required, but it is often not more than 'fairly' reasonable if a probability statement about H_z is needed.

Key words: foundations of statistical inference, philosophy of statistics, hypothesis testing, distributional inference, strong similarity, applied statistics

MSC: 62A01

1 Introduction

To clarify our intentions, we start from the following quotation by Venn [89], which we found in (R.T.) Cox [16]:

> 'In every case in which we extend our inferences by Induction or Analogy, or depend on the witness of others, or trust to our own memory of the past, or come to a conclusion through conflicting arguments, or even make a long and complicated deduction by mathematics or logic, we have a result of which we can scarcely feel as certain as of the premises from which it was obtained. In all these cases then we are conscious of varying quantities of belief, but are the laws according to which the belief is produced and varied the same? If they cannot be reduced to one harmonious scheme, if in fact they can at best be brought to nothing but a number of different schemes, each with its own body of laws and rules, then it is vain to endeavor to force them into one science.'

Cox observes that 'Venn himself belonged to the school of authors who define probability in statistical terms and restrict its meaning to examples in which it can be so defined'. Cox is more ambitious. He discusses principles 'valid over the whole field of probable inference'. We are much less ambitious (see Appendix B.7 for further discussion).

Restricting the attention to statistical inference, i.e. induction on the basis of a set x of statistical data, and largely ignoring the possibility of extending inferences 'by Analogy or by depending on the witness of others, etc.', we still shall encounter difficulties of a fundamental nature which preclude the existence of compelling conclusions. The theory of statistical inference and decision-making is, indeed, 'nothing but a number of different schemes, each with its own body of laws and rules'. Some unification will and should be pursued but this should not be done too dogmatically.

Section 2 will be about the relative-frequency definition of probability. In Section 3 a mathematical generalization is discussed. Unfortunately, most of the probabilities discussed in Sections 2 and 3 are theoretically well-defined but practically not exactly known. As concrete numerical specifications are necessary, we need a different kind of 'probabilities'. A general perspective is presented in Section 4. A leading idea is that (epistemic) probabilities are 'most' reasonable degrees of belief which should be constructed as estimates of truth values with *squared-error loss* in our mind (or some other *proper* loss function). A different leading idea is discussed and illustrated in Section 5. Our Fisher-Neyman-Pearson-Wald perspective will be confronted with two alternatives in Section 6. The last two sections were motivated by the title 'Statistics: Reflections from the Past and Visions for the Future' of the San Antonio conference in honor of C.R. Rao on the occasion of his 80th birthday.

In general, the reader may be disappointed by the lack of compelling results and of firm recommendations. This, however, is not unnatural. Just like Fisher and Rao we like to emphasize that, after a 'scrutiny of the data', one should 'let the data speak' (i.e. adopt an empiricist attitude) and 'express statistical uncertainty'. In Appendix C we shall have to admit that

the developed theory is only of limited interest for actual practice which, of course, is our real purpose.

2 The relative-frequency concept as a statistical basis for the understanding of probability

Suppose that a discourse is about a finite population Ω of $m = \#(\Omega)$ distinguishable elements denoted by symbols like ω and w. Here ω denotes an arbitrary element and w is used if a particular element is under investigation. It is a matter of naive set theory to understand the meaning of the relations $\omega \in A$ and $A \subset B$, and of the operations $A \cup B$, $A \cap B$, $\bar{A}$, etc. 'Axioms' like $\overline{A \cup B} = \bar{A} \cap \bar{B}$ are then evident. A $1:1$ correspondence exists between the subsets of Ω and their indicator functions. The relations and operations between sets have their analogues in respectively $\mathbf{1}_A(\omega) = 1$, $\mathbf{1}_A \leq \mathbf{1}_B$, $\mathbf{1}_A \vee \mathbf{1}_B$, $\mathbf{1}_A \wedge \mathbf{1}_B$, and $1 - \mathbf{1}_A$. If we consider an arbitrary element ω of Ω, then we can discuss the truth or falsity of the proposition $\omega \in A$. Such propositions may have alternative formulations but if we restrict the attention to those which are well defined with respect to Ω, then we know for every $\omega \in \Omega$ whether or not such a proposition is true and, as a consequence, that any well-defined proposition can be identified with a proposition of the form $\omega \in A$. We use π_A as a notation for this proposition, and observe that, in classical logic, the relations and operations discussed have their analogues in, respectively, π_A is true, $\pi_A \Rightarrow \pi_B$, $\pi_A \vee \pi_B$, $\pi_A \wedge \pi_B$, and $\neg\pi_A$.

Counting. Using $\#(A)$ to denote the number of elements in some subset A of Ω, the set function $\#$ thus defined satisfies

$$A \cap B = \varnothing \Rightarrow \#(A \cup B) = \#(A) + \#(B).$$

The relative frequency $\mathrm{R}(A) = \#(A)/\#(\Omega)$ satisfies the same additivity property, together with $\mathrm{R}(\Omega) = 1$. In the next paragraph we assume that some 'statistician' has done the counting and, hence, that $\#(A)$ and $\mathrm{R}(A)$ are known to us.

Interpretation. Suppose that an individual w is taken at random from Ω and that we have to discuss the hypothesis H: $w \in A$. The ideal degree of belief in H is, of course, the truth value $\mathbf{1}_A(w)$. This, however, is unknown to us a priori because w can be anywhere in Ω.

Why is it 'compelling' to use the relative frequency $\mathrm{R}(A)$ as 'the mean' between 0 and 1, to 'approximate' $\mathbf{1}_A(w)$ and to 'express our opinion' about H? The randomness assumption implies that, a priori, all $\omega \in \Omega$ are equally likely. There are $\#(A)$ possibilities in favour of H and, hence, it is very reasonable to use $\mathrm{R}(A)$ as the chance that H: $w \in A$ holds true. This definition of chance or (degree of) probability is not applicable in other, more general, situations. That is why a different definition is needed. One possibility arises from the mathematical fact that $\mathrm{R}(A)$ is optimal with respect to squared error loss. If one minimizes

$$\sum_{\omega \in \Omega} \left(\mathbf{1}_A(\omega) - r\right)^2 = \sum_{\omega \in \Omega} \left(\mathbf{1}_A(\omega) - \mathrm{R}(A)\right)^2 + m\left(\mathrm{R}(A) - r\right)^2$$

as a function of the real number r, then one obtains $\mathrm{R}(A)$. The argument is fairly subtle because minimization of

$$\sum_{\omega\in\Omega} |\mathbf{1}_A(\omega) - r|^c = \#(A)|1-r|^c + (m - \#(A))r^c$$

leads to a solution satisfying $\mathrm{logit}(r) = (c-1)^{-1}\mathrm{logit}(R(A))$ for $c \in [0,\infty)$, where $\mathrm{logit}(r) = \log(r/(1-r))$. Obviously, $r = \mathrm{R}(A)$ only for $c = 2$. For $c = 1$, the minimum is achieved if $r = \mathbf{1}_{(1/2,1]}(\mathrm{R}(A)) + \frac{1}{2}\mathbf{1}_{\{1/2\}}(\mathrm{R}(A))$. The quadratic loss function, however, is not the only one providing $\mathrm{R}(A)$ as the optimal value. If one uses $-\log|\mathbf{1}_{\bar{A}}(\omega) - r|^c$ for c in $(0,\infty)$ instead of $(\mathbf{1}_A(\omega) - r)^2$ then one also arrives at $\mathrm{R}(A)$ because

$$\#(A)\log r + (m - \#(A))\log(1-r)$$

is maximum as a function of r if $r = \mathrm{R}(A)$. In this paper we restrict ourselves to the quadratic loss function.

Providing partial information. The considerations above were completely a priori. If the experiment has already been performed, then some information about w may be communicated to us. It is very important to know exactly how the *source of information* proceeds. Suppose that it provides us with the information that $w \in B$. It now seems reasonable to use the conditional relative frequency

$$\mathrm{R}_B(A) = \frac{\#(A\cap B)}{\#(B)} = \frac{\mathrm{R}(A\cap B)}{\mathrm{R}(B)}$$

as the (posterior) probability of H_A: $w \in A$, given the truth of π_B: $w \in B$. (Note that $\#(B) > 0$.) Unfortunately there is a snake in the grass. The interpretation of $\mathrm{R}_B(A)$ as a probability in the degree-of-belief sense may be inappropriate if it cannot be assumed that the source produces '$\omega \in B$' if and only if $\omega \in B$ (here for arbitrary element ω). An example will illustrate this.

Example 1. Suppose that $\Omega = \{1,2,3\}$ and that w is taken at random from Ω such that, a priori, the three probabilities concerned are equal to $\frac{1}{3}$. Suppose also that the truth or falsity of H_A: $w = 1$ is of interest while the source provides us with the information that w is odd or, equivalently, that H_B: $w \in \{1,3\}$ is true. One would like to jump to the conclusion that $\mathrm{R}_B(A) = \frac{1}{2}$ is the most reasonable degree of belief in H_A, given the oddity of w. This conclusion is, indeed, appropriate if the source is known to provide the information 'ω is odd' if and only if $\omega \in B = \{1,3\}$. If the source is allowed to make statements different from 'odd' or 'even' or, e.g., is allowed to remain silent (like in some games of chance) then $\mathrm{R}_B(A) = \frac{1}{2}$ will usually not be the most reasonable degree of belief in H_A and a probability statement should, perhaps, not be made (see also Albers (*et al.*) [1,2]). If, e.g., the source is uncertain about 'the oddity of unity' and keeps silent if $w = 1$ is observed, then the information $w \in B = \{1,3\}$ implies that $w = 3$ and, hence, that the 'probability' of H_A

should be assessed as 0. Note that this invalidates the weak syllogism

$$\frac{\begin{array}{l} w \in A \text{ implies } w \in B \\ w \in B \end{array}}{w \in A \text{ becomes more plausible}}$$

defended and motivated in Jaynes and Bretthorst [38].

3 The mathematical theory of probability

The relative frequency R is not the only set function satisfying the axioms

(1) $\mathrm{P}(\Omega) = 1$
(2) $A \cap B = \varnothing \Rightarrow \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

of a probability measure. Given any function p: $\Omega \to [0, 1]$ satisfying $\sum_\omega p(\omega) = 1$, the axioms are satisfied by the set function P which is defined by $\mathrm{P}(A) = \sum_{\omega \in A} p(\omega)$.

Largely ignoring the concrete numerical specification of functions like p and P, mathematicians have developed a fascinating theory of probability, with Kolmogorov [51] as a first culmination point. In his statistical theory, Ω is a very general, possibly even infinite-dimensional 'underlying' outcome space, while P is a set function satisfying axioms like 1 and 2. The numerical specification of values $\mathrm{P}(A)$ of interest is usually left to the statistician (or his client). In his practice, there is only one set function P which deserves to be called 'the' probability distribution or, perhaps, there is none. We restrict the attention to situations where the *existence* of an underlying true distribution P is not too questionable. What we have to worry about is the precise numerical specification of $\mathrm{P}(A)$ for the hypotheses H: $w \in A$ of interest. The numerical values $\alpha(x)$ we shall produce are only estimates or approximations of the true values $\mathrm{P}(A)$. They are based on the information we have about the actual outcome w of the 'random' experiment (now with 'random' in the sense of chance dependent). This information is made available to us in the form of a set x of statistical data.

Example 1, continued. If we assume that there is an 'experimental' probability t of remaining silent in the case of unity, then we can incorporate this information about the source of information by extending Ω to obtain $\Omega = \{\omega_0, \omega_1, \omega_2, \omega_3\}$ where ω_0 symbolizes that the outcome is 1 but the source remains silent, ω_1 that the outcome is 1 and 'w is odd' is reported, ω_2 that the outcome is 2 and 'w is even' is reported, and ω_3 that the outcome is 3 and 'w is odd' is reported. Now, for $0 \leq t \leq 1$, P is determined by $p(\omega_0) = \frac{1}{3}t$, $p(\omega_1) = \frac{1}{3}(1-t)$, $p(\omega_2) = p(\omega_3) = \frac{1}{3}$, while $A = \{\omega_0, \omega_1\}$ and $B = \{\omega_1, \omega_3\}$ are such that, according to standard probability theory,

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{1-t}{2-t}$$

which is less than $\mathrm{P}(A) = \frac{1}{3}$ if t is larger than $\frac{1}{2}$ (see the end of Section 2).

The significance of this result depends to a large extent on the availability of reliable evidence with respect to t. If such information is not sufficiently accurate, then one should not make a probability statement. If we really would know the exact value of t, then the 'procedure' which assigns $(1-t)/(2-t)$ if B occurs, and 0 if B does not occur, is optimal, in the sense of minimum mean squared error.

We now turn to an example from sampling theory where the specification of P is beyond doubt (if sampling is really at random) but, nevertheless, a vexed issue appears. The example is a bit complicated and the reader might prefer to skip to Section 4.

Example 2, the Problem of the Reference Class (see Reichenbach [67]). In Section 2 we became fascinated by the idea that, under certain conditions with respect to the source of information, $\mathrm{R}_B(A)$ is the most reasonable degree of belief in H: $w \in A$, given the information $w \in B$. The numerical evaluation of this conditional relative frequency requires knowledge of $\#(A \cap B)$ and of $\#(B)$. While such knowledge may be available in some situations from games of chance, it is not at our disposal in many situations from practice. A plethora of examples from practice exists, e.g. that of Giannone *et al.* [31]:

A: 'H-mode is reached in a plasma discharge';

not(A): 'L-mode is reached';

B: $P_{\mathrm{L}}/P_{\mathrm{thres}} > 2.0$;

not(B): $P_{\mathrm{L}}/P_{\mathrm{thres}} < 2.0$;

where P_{L} is the heating power and P_{thres} is the minimum heating power required to achieve H-mode according to some preciously established empirical scaling (see also C.5).

Suppose now, for example, that the population size m is known and that sampling has been such that the individual w of interest is the $(n+1)$st in a sequence $w_1, \ldots, w_{n+1}$ taken successively at random and (independently) without replacement from the population. We assume that all necessary precautions have been taken. For $w_1, \ldots, w_n$, full information is available about the truth or falsity of $w_i \in A$ and of $w_i \in B$. As a notation we use

$$n_{uv} = \#\{i = 1, \ldots, n;\ \mathbf{1}_A(w_i) = u,\ \mathbf{1}_B(w_i) = v\}$$

where $u, v \in \{0, 1\}$, which can be tabulated as follows.

	A	not(A)	
B	n_{11}	n_{01}	n_{+1}
not(B)	n_{10}	n_{00}	n_{+0}
	n_{1+}	n_{0+}	n

Similarly, m_{uv} is used if the population is considered. Note that $\sum n_{uv} = n = n_{++}$, $\sum m_{uv} = m$. For w_{n+1} we know that $\mathbf{1}_B(w_{n+1}) = 1$, i.e. that event B has occurred, and we are interested in the truth or falsity of H: $w_{n+1} \in A$. In the absence of any further (relevant) information we shall try to use the information available to specify a 'most' or, at least, a 'very' reasonable degree of belief $\alpha \in [0,1]$ in the truth of H: $w_{n+1} \in A$. It is intuitively clear that this number should be somewhere between the relative frequency n_{1+}/n_{++} of A occurrences in the sample and the conditional relative frequency n_{11}/n_{+1} of A occurrences among the B occurrences in the sample (if $n_{+1} \geq 1$). To analyze the situation mathematically, introduce $\Omega = \{(\omega_1, \ldots, \omega_{n+1});\ \omega_i \text{ is the } i\text{-th drawing}\}$ and specify P by defining

$$p(\omega) = m^{-1}(m-1)^{-1} \cdots (m-n)^{-1}.$$

Note that the outcome $x = X(w)$ of the random vector $X = (X_{11}, X_{12}, X_{21}, \ldots, X_{n1}, X_{n2}, X_{n+1,2})$, with $X_{i1}(\omega) = \mathbf{1}_A(\omega_i)$ and $X_{i2}(\omega) = \mathbf{1}_B(\omega_i)$, has to be used to make a probability statement about the outcome $y = Y(w)$ of the random variable Y which is defined by $Y(\omega) = \mathbf{1}_A(\omega_{n+1})$. Within this context it is natural to consider the physical probabilities $\mathrm{P}(Y = 1) = m_{1+}/m$ and $\mathrm{P}(Y = 1|X = x) = (m_{11} - n_{11})/(m_{+1} - n_{+1})$. As our knowledge about m_{1+}, m_{11} and m_{+1} is only partial, the epistemological status of these results is that these probabilities are 'ideals' which we do not know but can try to approximate. This may be useful but we should not forget that the real issue is to approximate the truth value $y = \mathbf{1}_A(w_{n+1})$ of the hypothesis H: $y = 1$. It is intuitively clear that such an approximation should not be made if the sample size n is very small (making the n_{uv} insufficiently informative). In principle, however, we try to construct a procedure α: $\mathfrak{X} = \{0,1\}^{2n+1} \to [0,1]$ such that the mean squared error

$$\mathbf{E}(Y - \alpha(X))^2$$

is 'as small as possible'. The outcome $\alpha(x)$ is then used as the assessment required. This task is difficult to elaborate upon, because this risk depends on the true but unknown values m_{uv} $(u, v = 0, 1)$. We return to this problem at the end of Section 6.

4 A general perspective

After a scrutiny of the data and of the source(s) by which they are produced, the statistician may be involved in a situation of the following kind.

Problem. Given the data x and the definition of the unknown y of interest, a statistical inference about y is required. To specify the relation between x and y, and the meaning of the adjective 'statistical', we take it for granted that (x, y) is the outcome $(X(w), Y(w))$ of a pair (X, Y): $\Omega \to \mathfrak{X} \times \mathscr{Y}$ of random variables, possibly with Y: $\Omega \to \mathscr{Y}$ degenerate in the sense that $\mathrm{P}(Y = y) = 1$ for the true value y (see Section 5 for an example).

The underlying probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is usually not specified as precisely as in Section 3. In principle, however, the statistician should have such a specification in mind, either with

P known (see Section 3, Example 2) or with P partially known (see Example 1 and Section 5). In practice the discourse will then be about the distribution $P = \mathrm{P} \circ (X, Y)^{-1}$ induced on the space $\mathfrak{X} \times \mathscr{Y}$ of 'all' theoretical possibilities $(\xi, \eta) = (X(\omega), Y(\omega))$ $(\omega \in \Omega)$ for the true value $(x, y) = (X(w), Y(w))$ of interest. In some situations not all points in $\mathfrak{X} \times \mathscr{Y}$ are theoretically possible.

In statistical practice, 'much' of P may be known, but 'not everything'. Let $\mathscr{P}$ denote the class of all theoretical possibilities for the true distribution P of (X, Y). In practice some elements of $\mathscr{P}$ are much more likely (a priori) than other ones. Such information should not be ignored if it is 'sufficiently reliable'. Here are two examples of how such information can be incorporated:

(1) Make an initial guess $P^{(0)} \in \mathscr{P}$ of P and proceed by indicating 'something' about the uncertainty a priori with respect to the specification of $P^{(0)}$. An example from the theory of nonparametric density estimation can be found in Albers *et al.* [3] where an initial guess ψ of the unknown density function f is to be made a priori and, next, the data are used to provide an estimate of the density by fine-tuning this initial guess. To specify the fine-tuning mechanism, some information about the distance $||\psi - f||_1$ is needed.

(2) Impose a 'parametric' model $\mathscr{P}_0 = \{P_\theta; \theta \in \Theta\} \subset \mathscr{P}$ and make the assumption $P \in \mathscr{P}_0$. Under this assumption, the true value t of θ is defined by $P = P_t$. The identifiability of θ is tacitly assumed. It is not necessary that Θ is finitely dimensional; in Section 5 the part of θ is played by the density f. In addition, some further knowledge may be available a priori about the 'likelihood' of the elements of Θ, e.g. that $\theta^{(0)}, \ldots, \theta^{(m)}$ are 'typical' values for t.

If the probabilistic constructions are sufficiently reliable then a probabilistic form for the inference about y is indicated. Two cases are of interest:

(1) *Hypothesis Testing.* If the truth or falsity of a hypothesis H: $y \in \mathscr{Y}_{\mathrm{H}}$ has to be discussed, then we shall pursue the construction of a method of inference α: $\mathfrak{X} \to [0, 1]$ such that the outcome $\alpha(x)$ is a 'very' reasonable degree of belief in H; it is regarded as an estimate of the truth value $\mathbf{1}_{\mathscr{Y}_{\mathrm{H}}}(y)$ of H (with respect to quadratic loss , see Section 2, α will then be called a q-value).

(2) *Distributional Inference.* If y has to be estimated (or predicted) in the sense that an 'opinion' about y is needed (or if a synthesis is needed of assessments of probability $\alpha(x)$, as derived under (1), for varying $\mathscr{Y}_{\mathrm{H}}$), then we pursue the construction of a method of inference Q: $\mathfrak{X} \to \mathscr{Y}^*$ where $\mathscr{Y}^*$ denotes the class of 'all' probability measures on $\mathscr{Y}$, and $Q(x)$ specifies our opinion about y.

Ad (1). Note that hypothesis testing in the sense of (1) is a special case of Distributional Inference in the sense of (2): take $\mathbf{1}_{\mathscr{Y}_{\mathrm{H}}}(y)$ as the unknown of interest and, for arbitrary $\xi \in \mathfrak{X}$, identify $Q(\xi) = (1 - \alpha(\xi))\epsilon_0 + \alpha(\xi)\epsilon_1$ with the 'probability' or degree of belief, $\alpha(\xi)$, in H (for details, see Salomé [71, pp. 39 and 60]). Aiming at the minimization of the mean squared

error

$$\mathbf{E}\left(\mathbf{1}_{\mathscr{Y}_{\mathrm{H}}} - \alpha(X)\right)^2 = \begin{cases} \mathbf{E}(1-\alpha(X))^2 & \text{if H is true} \\ \mathbf{E}(\alpha(X))^2 & \text{if H is false,} \end{cases}$$

one might feel attracted by the idea to require that $\alpha\colon \mathfrak{X} \to [0,1]$ is such that the unbiasedness requirement

$$\mathbf{E}(\mathbf{1}_{\mathscr{Y}_{\mathrm{H}}} - \alpha(X)) = 0$$

is satisfied. In many estimation problems, however, this requirement is too strong. That is why it will not be considered. It seems very reasonable to require that the q-value $\alpha\colon \mathfrak{X} \to [0,1]$ is *weakly unbiased*, in the sense that

$$\mathbf{E}(1-\alpha(X))^2 \leq \mathbf{E}(\alpha(X))^2 \text{ if H is true}$$

$$\mathbf{E}(\alpha(X))^2 \leq \mathbf{E}(1-\alpha(X))^2 \text{ if H is false,}$$

or, equivalently, that

$$\mathbf{E}\alpha(X) \geq \tfrac{1}{2} \text{ if H is true}$$

$$\mathbf{E}\alpha(X) \leq \tfrac{1}{2} \text{ if H is false}$$

(see Schaafsma [75])

The requirement of *strong unbiasedness* we shall impose in Section 7.2 is less easily understood and appreciated. It requires that $\alpha(X)$ is stochastically larger, respectively smaller, than a random variable following the $\mathcal{U}(0,1)$ distribution if H is true, respectively false. Motivation via the related concept of strong similarity can be found below.

Ad (2). Though distributional inference can be pursued in greater generality, the interesting case of a *real-valued* unknown y is of paramount interest. Restricting ourselves to this case, note that the procedure $Q : \mathfrak{X} \to \mathbb{R}^*$ is now uniquely determined by the family $\mathcal{G} = \{\mathcal{G}_\xi;\ \xi \in \mathfrak{X}\}$ of distribution functions associated to it. We have $\mathcal{G}_\xi(z) = \{Q(\xi)\}((-\infty, z])$. The notation used here expresses the difference between the data x actually available and the values ξ theoretically possible for x, a priori. See the distinction between ω and w made earlier. In the sequel, when it will be clear from the context whether x refers to the actual data or to the theoretical possible values, we will not make a notational distinction between x and ξ.

Definition. The procedure $Q\colon \mathfrak{X} \to \mathbb{R}^*$ is said to be *strongly similar* if

$$\mathcal{L}\, \mathcal{G}_X(Y) = \mathcal{U}(0,1)$$

holds for the true distribution P of (X, Y) (where $\mathcal{L}$ denotes 'the distribution of').

Motivation 1. If the ideal distributional inference $Q^*(\xi) = \mathcal{L}\ (Y \mid X = \xi)$ has a continuous distribution function $\mathcal{G}^*_\xi$, no matter the a priori possible outcome ξ of X, then $\mathcal{L}\ (\mathcal{G}^*_\xi(Y) \mid X =$

$\xi) = \mathcal{U}(0,1)$ for every $\xi \in \mathfrak{X}$ and, hence, (1) $\mathcal{G}_X^*(Y)$ and X are independent, (2) $\mathcal{L}\,\mathcal{G}_X^*(Y) = \mathcal{U}(0,1)$. The last property of the ideal solution is made imperative.

A second motivation is needed because the above argument requires a context where Y is not degenerate. If, e.g., one estimates some population characteristic $y = f(t)$ then the ideal inference $Q^*(\xi) = \epsilon_y$ is degenerate. We exploit Neyman's concept of (a family of) confidence intervals:

Motivation 2. If one uses a strongly similar procedure Q: $\mathfrak{X} \to \mathbb{R}^*$ to generate the intervals

$$\left\{\left[\mathcal{G}_\xi^{-1}(\tfrac{1}{2}\alpha),\ \mathcal{G}_\xi^{-1}(1-\tfrac{1}{2}\alpha)\right];\ \xi \in \mathfrak{X}\right\}$$

then, automatically,

$$\mathrm{P}\left(\mathcal{G}_X^{-1}(\tfrac{1}{2}\alpha) \le Y \le \mathcal{G}_X^{-1}(1-\tfrac{1}{2}\alpha)\right) = 1-\alpha$$

if the $\mathcal{G}_\xi$ are strictly increasing and continuous, i.e. the confidence intervals are 'exact'.

In practice one will often have to content oneself with approximate (asymptotic) results. An example can be found at the end of the following section.

5 Credibility anchored in probability

Let the outcomes $x_{[1]} < \cdots < x_{[n]}$ of an independent random sample $X_1 < \cdots < X_n$ be available from some (largely) unknown distribution on $\mathbb{R}$ with distribution function F and strictly positive density $f = F'$. Being interested in making a distributional inference about the quantile $\xi_p = F^{-1}(p)$, we note that

$$\mathrm{P}\left(\xi_p \le X_{[i]}\right) = \mathrm{P}\left(p \le F(X_{[i]})\right) = \sum_{s=0}^{i-1} \binom{n}{s} p^s (1-p)^{n-s}$$

because the event of interest occurs if and only if at most $i-1$ of the independent $\mathcal{U}(0,1)$-distributed random variables $U_j = F(X_j)$ fall below p (see Thompson [85]).

The distribution function $G_{x_{[1]},\ldots,x_{[n]}}$ of the distributional inference $Q = Q(x_{[1]}, \ldots, x_{[n]})$ about $y = \xi_p$ may be chosen such that $G_{x_{[1]},\ldots,x_{[n]}}(z)$ indicates the degree of belief in H: $\xi_p \le z$. Next, equating $Q(\xi_p \le x_{[i]})$ to $\mathrm{P}(\xi_p \le X_{[i]})$ we obtain that

$$G_{x_{[1]},\ldots,x_{[n]}}(z) = \sum_{s=0}^{i-1} \binom{n}{s} p^s (1-p)^{n-s}$$

if $z = x_{[i]}$ $(i = 1, \ldots, n)$.

In practice, one will use a continuous distribution function such that the epistemic probability assigned to $[x_{[i]}, x_{[i+1]}]$ is $\binom{n}{i} p^i (1-p)^{n-i}$. It is a matter of mathematics to establish that the procedure described is *asymptotically strongly similar*. The essence of the proof is that for any $c \in (0,1)$,

$$\lim_{n\to\infty} \mathrm{P}\left(G_{X_{[1]},\ldots,X_{[n]}}(\xi_p) \le c\right) = c.$$

For a complete proof, see Albers [1] or Kardaun and Schaafsma [46]. Note that the bootstrap distribution for ξ_p (see Efron [26]) is asymptotically equivalent to the procedure just described. De Bruin *et al.* [12] adapted the ideas presented here to obtain a nonparametric estimate $\hat{f}_n$ of f which, for $n \to \infty$, did not converge as quickly as usual kernel estimates. An improvement $\hat{f}_n^{(m)}$ of $\hat{f}_n$ was developed in Albers *et al.* [3]. It is suggested that this improved version is better than 'usual' kernel estimates if the initial guess ψ of f is not too unreliable. In practice, this initial guess may be slightly data dependent, it can e.g. follow the $\mathcal{N}(\bar{x}, s^2)$ distribution, or be based on fitting a S_B curve to the first four moments (or the first two moments and the minimum and maximum) of the empirical distribution, see Johnson and Kitchen [39]. In comparisons of $\hat{f}_n^{(m)}$ and kernel estimates based on ψ as well, the performance of both types of estimates is approximately equally good.

6 Science or Technology? Rationalist or Empiricist?

Our attitude is that of the Fisher-Neyman-Pearson-Wald school. The attention is concentrated on discussions about the construction of methods of inference, *functions* prescribing inferences (or decisions). It has to be admitted that a considerable amount of 'surrealism' is involved in the shift of attention from the data x at hand to 'arbitrary' data $\xi \in \mathfrak{X}$. This shift of attention is chosen to accommodate statistical assumptions of randomness, independence, etc. The 'surrealism' involved may easily lead to peculiar conflicts and incoherencies. See, e.g., Appendix B.8.

With respect to attitudes in general, something useful can be found in the conversation with Tukey, reported in Fernholz and Morgenthaler [28]: 'A good statistician must be a schizophrenic, because he has to deal with uncertainty and the measurement of uncertainty - that's his main task - and to do this using the most certain tool we have which is mathematics. He has to bridge the gap here'. Somewhat later, referring to the question whether statistics is a science, see also the quotation to Venn [89] in Section 1, Tukey said: 'I would tend to think it would have been more accurate to say science and technology' and he referred to 'Data Analysis, including Statistics' as a pure technology. With these preparations in mind, two attitudes, competing with that of the Fisher-Neyman-Pearson-Wald school, will next be discussed. Both provide a more direct approach to the problem of evaluating the data x at hand. The 'surrealism' involved is of a different kind.

Approach 1. 'Data analysts' concentrate the attention on the use of well defined numerical principles (e.g. least squares). They avoid 'statistical fuss' because the randomness, independence, homogeneity, and other assumptions underlying a statistical analysis, which are often not satisfied, are ignored. It is, indeed, true that many data sets in psychometry, meteorology, biology, etc., are evaluated ignoring a mathematical-statistical foundation. Corresponding results are expected to be less reliable than results based on a mathematical-statistical foundation, at least if underlying assumptions are realistic.

Approach 2. 'Bayesians' wholeheartedly accept the probabilistic constructions of Section 4 and, in fact, provide an extension such that an approximation $\tilde{P} = \mathcal{L}(\tilde{X}, \tilde{Y})$ of P is obtained and $Q(x) = \mathcal{L}\ (\tilde{Y}|\ \tilde{X} = x)$ can be used as an approximation to the ideal inference $Q^*(x) = \mathcal{L}\ (Y|\ X = x)$. More precisely, they proceed as follows. Firstly they use the additional information available (see Section 4) to postulate a parametric model $\mathscr{P} = \{P_\theta;\ \theta \in \Theta\}$ and a prior distribution $\tau \in \Theta^*$. Next they discuss an artificial triple $(T, \tilde{X}, \tilde{Y})$ of random variables such that (1) $\mathcal{L}\ (T) = \tau$, (2) $\mathcal{L}\ ((\tilde{X}, \tilde{Y})|\ T = \theta) = P_\theta$. Next they concentrate the attention on the marginal distribution $\tilde{P} = \int P_\theta\ \mathrm{d}\tau(\theta)$ and continue as above. Axiomatic theory of De Finetti or of R.T. Cox or, more recently, of Grünwald and Philip-David [32] can be invoked to support this approach. These theories will be examined in Appendix B.

Returning to Example 2 of Section 3. (The reader that skipped the example earlier, can now skip to Section 7.) The following argument has its origin in proposals made in pattern recognition. Note that the relative frequency of B-occurrences, among the n_{0+} $\bar{A}$-observations is n_{01}/n_{0+}. Similarly that among the A-observations is n_{11}/n_{1+}. The 'dissimilarity' between $\mathbf{1}_B(w_{n+1}) = 1$ and the first, resp. the second relative frequency is n_{00}/n_{0+}, resp. n_{10}/n_{1+}. Defining the 'affinity' as the reciprocal value and normalizing, we obtain the number

$$\alpha_1(x) = \frac{n_{1+}/n_{10}}{(n_{1+}/n_{10}) + (n_{0+}/n_{00})}$$

which one might interpret as a 'posterior probability' of $w_{n+1} \in A$, given $w_{n+1} \in B$. We are not in favour of this type of argument but have to accept that it can be made. It provides an example of Approach 1. Next we illustrate Approach 2. Note that the difficulty in Section 3 was that the ideal solution

$$\mathrm{P}(Y = 1|\ X = x) = \frac{m_{11} - n_{11}}{m_{+1} - n_{+1}}$$

cannot be used because the population frequencies m_{11} and m_{+1} are not known to us. With respect to the general perspective of Section 4, the vector $(m_{00}, m_{01}, m_{10}, m_{11})$ plays the part of the true value t of the underlying parameter θ. It is not unreasonable to suggest that $t = (m_{00}, m_{01}, m_{10}, m_{11})$ itself is the outcome of some random vector $T = (M_{00}, \ldots, M_{11})$. It is mathematically attractive (but statistically questionable because of the underlying i.i.d. assumptions) to postulate that $\mathcal{L}\ (T)$ is the multinomial $M(m; p_{00}, \ldots, p_{11})$ distribution, because this implies that

$$\begin{pmatrix} \tilde{X}_{11} \\ \tilde{X}_{12} \end{pmatrix}, \ldots, \begin{pmatrix} \tilde{X}_{n1} \\ \tilde{X}_{n2} \end{pmatrix}, \begin{pmatrix} \tilde{Y} \\ X_{n+1,2} \end{pmatrix}$$

are independent and identically distributed with $\tilde{\mathrm{P}}(\tilde{X}_{i1} = u,\ \tilde{X}_{i2} = v) = p_{uv}$. If we accept this model then we obtain

$$\tilde{\mathrm{P}}\left(\tilde{Y} = 1|\ \tilde{X}_{11} = x_{11}, \ldots, \tilde{X}_{n+1,2} = 1\right) = \frac{p_{11}}{p_{+1}}$$

and this would settle the issue if p_{11} and p_{+1} were known. Unfortunately, practice is usually such that even the existence of these probabilities (as relative frequencies in a superpopulation) is questionable. Classically it can of course be estimated by n_{11}/n_{+1} and, in case the features AS and B are mutually uncorrelated, also by n_{+1}/n. Nevertheless, one can continue this Bayesian approach by postulating a prior distribution for the vector $(p_{00}, \ldots, p_{11})$, e.g. a Dirichlet distribution. We, however, return to the Fisher-Neyman-Pearson-Wald approach.

Which *method of inference* is most appropriate to settle the Problem of the Reference Class? At the end of Section 3 it became clear that we need a procedure α: $\{0,1\}^{2n+1} \to [0,1]$ such that $\mathbf{E}\,(Y - \alpha(X))^2$ is 'as small as possible'. Two proposals are as follows:

(1) Reichenbach [67] suggested to take information of the kind $w_{n+1} \in B$ into account if such information is 'relevant' and the statistics involved are 'sufficiently reliable'. A quantification is as follows. Firstly, to discuss the relevance of the information $w_{n+1} \in B$ we consider the outcome s of the standardized test statistic

$$S = \frac{N_{11} - N_{1+}N_{+1}/n}{\sqrt{\frac{N_{0+}N_{1+}N_{+0}N_{+1}}{n^2(n-1)}}}$$

of Fisher's exact test for the 2×2 table with the n_{uv} as outcomes. Next, referring to the Bayesian setting with the p_{uv}, we regard (see Section 7.2)

$$\alpha(s) = \exp\left(-\tfrac{1}{2}s^2\right)/\sqrt{2}$$

as degree of belief in favour of the hypothesis H_0: $p_{00}p_{11} = p_{10}p_{01}$ that the 'features' A and B are independent and the information $w_{n+1} \in B$ is completely irrelevant. Finally we take the convex combination

$$\alpha_2(x) = \frac{\alpha(s)n_{1+}}{n} + \frac{(1-\alpha(s))n_{11}}{n_{+1}}$$

as the degree of belief in H: $y = 1$ because n_{1+}/n is the sampling analogue of the relative frequency $\mathrm{R}(A) = m_{1+}/m$ which is optimal (see Section 2) in the case of complete irrelevance of the information $w_{n+1} \in B$. The sampling analogue n_{11}/n_{+1} of m_{11}/m_{+1} is chosen in the case of relevance. (If $n_{11} = n_{+1} = 0$, we can add $\frac{1}{2}$ to both n_{11} and n_{01}, corresponding to Jeffreys's prior, see e.g. Salomé [71, p. 110].)

(2) In [34], Hinkley, in the situation of Example 2, requires that we derive the conditional probability

$$p(x,y) = \mathrm{P}\left(X = x,\ Y = y \,\middle|\, \begin{pmatrix} \tilde{N}_{00} & \tilde{N}_{01} \\ \tilde{N}_{10} & \tilde{N}_{11} \end{pmatrix} = \begin{pmatrix} n_{00} & n_{01} \\ n_{10}+1-y & n_{11}+y \end{pmatrix}\right),$$

where $\begin{pmatrix} \tilde{N}_{00} & \tilde{N}_{01} \\ \tilde{N}_{10} & \tilde{N}_{11} \end{pmatrix}$ is the complete-data sufficient statistic. This function $p(x,y)$ is then re-

garded as the likelihood of y, given the data x. A normalization provides the degree of belief

$$\alpha_3(x) = \frac{p(x,1)}{p(x,0)+p(x,1)} = \frac{n_{11}+1}{n_{1+}+2}$$

in favor of H: $y = 1$, which is the procedure based on the traditional Bayes-Laplace prior, see Kardaun *et al.*[44, Question 13].

A comparative analysis. For various $\theta = \begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix}$ and n, we studied the behaviour of $R(\theta, \alpha_j) = \mathbf{E}(Y - \alpha_j(X))^2$ for $j = 1, 2, 3$ as follows. For $M_1 = 20000$ experimental settings, where in each setting the m_{ij} were chosen uniformly random (as multiples of 5, between 5 and 150), and n was chosen uniformly random (as a multiple of 5 between 20 and m), we performed $M_2 = 2000$ experiments within each setting. For each experiment we computed the three α_i's and the corresponding losses $(\mathbf{1}_{w_{n+1}} - \alpha_i(w_1, \ldots, w_n))^2$, and averaged these for the M_2 replications, after which we compared the M_1 settings.

From these simulations, we concluded that, in general, α_2 performs best: it has the lowest average risk (on average, $R(\theta, \alpha_1) = 1.12R(\theta, \alpha_2)$, and $R(\theta, \alpha_3) = 1.17R(\theta, \alpha_2)$), and in 60.6% of the experimental settings its risk was lower than that of α_1 and α_3, as is shown in this table:

	Rank		
Procedure	Best	Second	Third
α_1	21.4%	46.4%	32.2%
α_2	60.6%	27.3%	12.1%
α_3	18.0%	26.4%	55.7%

A detailed study analyzing under which settings of (θ, n) the procedures performed best and worst, showed that α_3 performs poorest when $m_{01} \leq 25$ and/or $\det(\theta) < 0$; α_2 performs best, by far, if $m_{10} \leq 25$. The differences between the three procedures vanish quite rapidly if n grows (for $n > 400$, the three risks differ by less than one percent).

7 Reflections from the past

7.1 A general background of some inductive inferences

It is nice to argue that the main task of the statistician is to make predictive statements about the real world. Such statements may be directly 'verifiable' or 'falsifiable'. In practice, however, the attention is usually restricted to inductive inferences about 'true' values of unknown parameters. Such inferences are difficult to verify or falsify (except for artificial

situations with simulated data). Restricting the attention to this area, a very powerful 'technology' or 'methodology' is as follows. Suppose that a parametric family $\mathscr{P} = \{P_\theta;\ \theta \in \Theta\}$ has been decided upon to characterize 'all' possible values for the distribution P of X and suppose that one is interested in the 'true' value $y = f(t)$ of a given function $f\colon \Theta \to \mathscr{Y}$. We assume that $\Theta = \mathbb{R}^p$ with p finite and that f is not too complicated, the cases (1) $\mathscr{Y} = \Theta$, $f(\theta) = \theta$ and (2) $\mathscr{Y} \subset \mathbb{R}$ being of particular interest. Note that t denotes the true value of θ, based on the assumption $P \in \mathscr{P}$, and is defined by $P_t = P$. To derive a distributional inference about y, it seems reasonable to start from a distributional inference $Q(x) \in \Theta^*$ about t and to use the induced measure $\{Q(x)\} \circ f^{-1}$ as the inference about y. (Later we shall mention some drawbacks of this approach.)

To derive a distributional inference $Q(x)$ about t we assume that densities p_θ exist (in the generalized sense of Radon-Nikodym derivatives with respect to some σ-finite measure μ). Next we consider the likelihood function $l_x(\theta) = p_\theta(x)$ and introduce some weight function $w(\theta)$. Finally, if the actual situation allows this construction, we consider

$$q_x(\theta) = \frac{l_x(\theta)w(\theta)}{\int_\Theta l_x(\theta)w(\theta)\ \mathrm{d}\lambda(\theta)}$$

as the density of our distributional inference $Q(x)$ about t (here λ denotes Lebesgue measure on $\mathbb{R}^p$, restricted to Θ). To facilitate computation and interpretation, one can make a normal approximation $\mathcal{N}_p(\hat\theta, \hat\Sigma)$, where $\hat\theta$ is defined by $q_x(\hat\theta) = \max_\theta q_x(\theta)$ and the Taylor expansion

$$\log q_x(\theta) = \log q_x(\hat\theta) + \tfrac{1}{2}(\theta - \hat\theta)'\Delta_x(\theta - \hat\theta) + \dots$$

is used with Δ_x as notation for the Hessian, the matrix of second order derivatives of $\log(q_x)$, evaluated at $\hat\theta$. Next, ignoring the θ-dependence of the remainder term, one obtains

$$q_x(\theta) \approx c_x \exp\left(-\tfrac{1}{2}(\theta - \hat\theta)'\hat\Sigma^{-1}(\theta - \hat\theta)\right)$$

the right-hand side of which is the density of the multivariate normal $\mathcal{N}_p(\hat\theta, \hat\Sigma)$ distribution with $\hat\Sigma = -(\Delta_x)^{-1}$; c_x being a standardization constant. If f is a linear function, then the distributional inference about $y = f(t)$, derived as the induced measure $\{Q(x)\} \circ f^{-1}$, is normal as well, at least if one uses $Q(x) = \mathcal{N}_p(\hat\theta, \hat\Sigma)$. If f is not linear, then a linearization, based on the delta method, may be applied, or a more complicated error propagation based on quadratic approximation of the function f and including the propagation of skewness and kurtosis, may be used (see Kardaun [45, Section 1.8]).

Many mathematical statisticians have contributed to the scientific consolidation of this powerful technology. Some work of this kind is necessary because there are situations of a vexed nature (see, e.g., the beginning of Subsection 7.2) which cannot be dealt with satisfactorily by using $Q(x) = \mathcal{N}_p(\hat\theta, \hat\Sigma)$. A beautiful report of this work can be found in Reid [68]. Its title 'Asymptotics and the Theory of Inference' summarizes that 'asymptotic arguments are an inevitable consequence of a frequency-based theory of probability'. With respect to the technology just mentioned there are some complications which we consider crucial and

which are not dealt with satisfactorily by asymptotic theory. These complications are: (1) should the choice of the weight function w depend, to some extent, on the function f defining the unknown $y = f(t)$ of interest? (2) which parametrization should be chosen and which weight function w and, above all, (3) how can one resolve the basic incongruity that consists in the fact that, for any monotonic function $u(\theta)$ of θ, the likelihood function used in classical, frequentist statistics transforms as $l_{x;u}(u(\theta)) = l_{\{x;\theta\}}(\theta)$, whereas $q_x(\theta)$, at least if interpreted as a posterior probability in a (formal or subjective) Bayesian approach, transforms as $q_{x;u}(u(\theta)) = q_{x;\theta}(\theta)|u'(\theta)|$ with $u'(\theta) = \frac{d}{d\theta}u(\theta)$, see Kardaun *et al.* [44, Question 13]. (The issues involved are interrelated.) Asymptotic theory is of interest but ignores the choices to be made in practice where sample sizes are fixed (and not too large). The following two subsections provide an alternative to the formal-Bayes theory just mentioned.

7.2 Hypothesis Testing

If one is interested in the testing of a null hypothesis, e.g. H_0: $f(t) = 0$, then the approach of Section 7.1 provides that the induced measure $\{Q(x)\} \circ f^{-1}$ on $\mathscr{Y}$ will assign measure 0 to $\{0\}$. As the testing of such null hypotheses is crucial in model building, something had to be done about this difficulty. Pearson [65] started the 'modern' theory of statistics (see Section 8) by stating that the probability

$$\alpha_{\mathrm{Pearson}}(x) = \mathrm{P}_0\left(\chi^2_{k-1} \geq g(x)\right)$$

of exceedance is a 'fairly reasonable' criterion for the probability that the null hypothesis in his 'goodness-of-fit' situation is true. Here $g(x) = \sum(o_i - e_i)^2/e_i$ is the outcome of his test statistic (o_i denoting the observed and e_i the expected frequency of cell i ($i = 1, \ldots, k$)).

Pearson's formulation indicates that he was aware of 'the nearness of an abyss' [66]. Note that the probability distribution P_0 refers to the theoretical situation that the distribution of $g(X)$ is considered if the null hypothesis is true. By calling the 'p-value' $\alpha_{\mathrm{Pearson}}(x)$ a 'fairly reasonable' criterion, Pearson indicated that, perhaps, some 'more reasonable' criterion exists. Note that, in principle, the 'probability that H_0 is true' is equal to 1 if, indeed, H_0 is true. It is equal to 0 if H_0 is false.

Somewhat later in his goodness-of-fit paper, Pearson was using his p-value to suggest the odds ratio $\alpha(x)/(1-\alpha(x))$ for betting purposes. We now know that this suggestion is misleading, since some overconfidence is displayed. Small values of $\alpha(x)$ (and hence of $\alpha(x)/(1-\alpha(x))$) are too small to be reliable for betting purposes (and large values are too large). In the Neyman-Pearson theory, the p-value $\alpha(x)$ is defined as the *smallest* nominal level α of significance for which H_0 is rejected on the basis of the data x. The selection involved implies that, indeed, small values are too small to be reliable. That large values are too large is seen by considering the case with $o_i \approx e_i$ where the p-value is almost 1 whereas such observations are also in line with theoretical possibilities θ close to, but slightly different from, those specified by H_0. In

Bayesian theory, it is also argued (see Berger [7] for a recent survey) that overconfidence is displayed. We are not very satisfied by the alternative procedures these authors proposed. In our opinion, a 'q-value' $\alpha(x)$ is needed, α: $\mathfrak{X} \to [0,1]$ being an estimator of the truth value of H_0, derived on the basis of the quadratic loss function (see Section 2). This concept of q-value has its origin in Schaafsma [75] and Schaafsma *et al.* [76]. See also Huang *et al.* [35]. It is different from the 'Bayesian' concept of q-value introduced in Storey [83].

In the special case $k = 2$ of Pearson's problem, the outcome

$$\begin{aligned} g(x) &= (o_1 - e_1)^2\{e_1^{-1} + (n - e_1)^{-1}\} \\ &= \frac{(o_1 - np_1^{(0)})^2}{np_1^{(0)}(1 - p_1^{(0)})} \end{aligned}$$

of his test statistic is equal to the square of the test statistic which, with some abuse of notation, is now denoted as

$$X = \frac{O_1 - np_1^{(0)}}{\sqrt{np_1^{(0)}(1 - p_1^{(0)})}}$$

such that the following situation is 'approximately' relevant. (An exact approach to Pearson's problem with $k = 2$ will be provided in Section 8.)

Problem. Given the outcome x of $X \sim \mathcal{N}(t, 1)$ where the true value t of $\mathbf{E}X$ is completely unknown a priori. The interest is in testing H_0: $t = 0$, with 'testing' in the sense that a 'very' reasonable degree of belief $\alpha(x)$ in H_0 is required. As the testing of such 'null' hypothesis (versus A: $t \neq 0$) is a vexed issue, we shall also discuss the testing of H: $t \leq 0$ versus A: $t > 0$.

The q-value underlying Pearson's p-value is now given by

$$\alpha_{\text{Pearson}}(x) = 2\,(1 - \Phi(|x|))\,.$$

Fisher [29] emphasized that such p-values are uniformly distributed on $[0,1]$ if H_0 is true. This suggests the following definitions (see Section 4).

Definition 1. The q-value α: $\mathbb{R} \to [0,1]$ is said to be *Fisher-unbiased* (or strongly unbiased) if $\alpha(X)$ is stochastically larger, respectively smaller, than a random variable $U \sim \mathcal{U}(0,1)$ if H_0 (or, rather, H) is true, respectively false.

Definition 2. The q-value α: $\mathbb{R} \to [0,1]$ is said to be *Fisher-similar* (or strongly similar) if $\alpha(X)$ has the uniform distribution if H_0 is true (or, more generally, if the true value t of θ is in the common boundary of Θ_H and Θ_A).

It is a matter of mathematical analysis to establish that the p-value just defined is *uniformly best strongly similar* for testing H_0: $t = 0$, for a very large variety of loss functions, if the attention is restricted (by 'invariance') to estimators α: $\mathbb{R} \to [0,1]$ of $\mathbf{1}_{\{0\}}(t)$ depending on x via $|x|$. There is a considerable amount of Fisherian dogmatism involved in the requirements of strong similarity and strong unbiasedness. The uniformization of testing null hypotheses these restrictions provide can also be achieved by not aiming at a uniformly distributed

random variable U but by concentrating the attention on another random variable V with a fixed distribution on $[0,1]$, e.g. Beta$(2,2)$. Theory in Van der Meulen *et al.* [58] provides some motivation for using the uniform distribution $\mathcal{U}(0,1)$ (in terms of risk continuity for 'all' proper loss functions). The best motivation, however, is the one provided at the end of Section 4. Unfortunately, this motivation does not refer to hypotheses of the form H_0: $t = 0$ but to hypotheses of the form H: $t \leq 0$.

In our opinion it is less 'surrealistic' (see Section 8) to concentrate the attention, primarily, on the estimation of the indicator function $\mathbf{1}_{\{0\}}(t)$ of H_0 with squared error loss. The requirement of *weak unbiasedness*

$$\mathbf{E}\alpha(X) \geq \tfrac{1}{2} \quad \text{if and only if } \mathrm{H}_0 \text{ (or H) is true}$$

implies that of *weak similarity*

$$\mathbf{E}\alpha(X) = \tfrac{1}{2} \quad \text{if } \mathrm{H}_0 \text{ is true,}$$

which, in situations like the present one, is equivalent to the requirement of continuity of the mean squared error.

There are many weakly unbiased estimators of $\mathbf{1}_{\{0\}}(t)$. The p-value has a position between the extremes $\alpha \equiv \frac{1}{2}$ and $\alpha(x) = \mathbf{1}_{[0,\Phi^{-1}(3/4)]}(|x|)$. Let us elaborate on the idea of Neyman and Pearson to consider the likelihood ratio

$$\lambda(x) = \frac{\max_{\theta\in\Theta_H} p_\theta(x)}{\max_{\theta\in\Theta} p_\theta(x)}$$

which, in the present situation, provides $\lambda(x) = \exp(-\frac{1}{2}x^2)$. The estimator $\lambda\colon \mathbb{R} \to [0,1]$ thus defined can be regarded as an estimator of $\mathbf{1}_{\{0\}}(t)$, but $\mathbf{E}\lambda(X_0)$ is different from $\frac{1}{2}$ (it is $\frac{1}{\sqrt{2}}$) and that is why some correction will be needed if it has to be made weakly unbiased. We use X_θ as a notation for a fictitious random variable having the same distribution as X if t happens to be θ. One possibility is to take

$$\alpha_1(x) = \exp(-\tfrac{1}{2}x^2)/\sqrt{2}.$$

The q-value thus defined minimizes the area under the mean-squared-error curve, among all weakly similar q-values, because it is the solution of the following optimization problem

$$\begin{array}{rl} \text{Minimize} & \displaystyle\int_{-\infty}^{\infty} \mathbf{E}\left(\alpha(X_\theta)\right)^2 \,\mathrm{d}\theta \\ \text{Restrictions} & \mathbf{E}\left(\alpha(X_0)\right) = \frac{1}{2} \\ & 0 \leq \alpha \leq 1. \end{array}$$

The proof is easy (see Schaafsma [75] for an outline and Schaafsma [76] for an application). Another possibility is to determine the constant c such that $\mathbf{E}(\lambda(X_0))^c = \frac{1}{2}$, this provides

$$\alpha_2(x) = \exp\left(-\tfrac{3}{2}x^2\right).$$

Neyman and Pearson themselves accepted some form of 'Fisherian dogmatism' by computing the p-value

$$\begin{aligned}\alpha_3(x) &= \mathrm{P}(\lambda(X_0) \leq \lambda(x)) \\ &= 2\,(1 - \Phi(|x|))\,.\end{aligned}$$

Alternative suggestions have been made, and elaborated upon, in, e.g., Schaafsma (*et al.*) [75,76], Van der Meulen (*et al.*) [57,58], and Berger [7].

The comparison of the underlying methods of inference $\alpha\colon \mathfrak{X} \to [0,1]$ is a matter of concern. We are attracted by the idea that the mean-squared error

$$\mathrm{R}(\theta, \alpha) = \begin{cases} \mathbf{E}(1 - \alpha(X_0))^2 & \text{if } \theta = 0 \\ \mathbf{E}(\alpha(X_\theta))^2 & \text{if } \theta \neq 0 \end{cases}$$

should be considered as characteristic of performance and that the requirement $\mathbf{E}\alpha(X_0) = \frac{1}{2}$ of weak similarity should be respected (partly because it is equivalent to continuity of the risk function). In this respect it is an attractive property of α_1 that it minimizes the area under the risk function. On the other hand, it has to be admitted that the 'very fact' that $\mathrm{H}_0\colon t = 0$ has to be tested implies that alternatives not too far from 0 deserve additional attention. If this is incorporated in the objective function then one will see that some overconfidence is displayed by using α_1. We are attracted by the way out of these difficulties which is presented by ideas about most stringent tests, minimax regret (SMR) procedures, etc. Such 'objectivistic' optimum properties have the advantage that they do not depend on the parametrization. For elaborations we refer to Van der Meulen (*et al.*) [57,58]. These elaborations are not presented here because Section 8 will provide the essence of an 'exact' approach to the problem of testing $\mathrm{H}_0\colon t = p_1^{(0)}$ on the basis of the outcome s of $S \sim \mathrm{Bin}(n, t)$.

With respect to the general goodness-of-fit problem with more than two cells, Fisher [29] suggested that Pearson's p-value had to be replaced by

$$\alpha_{\mathrm{Fisher}}(x) = \mathrm{P}\left(\chi^2_{k-1-c} \geq g(x)\right)$$

because the distribution of $g(X)$ under H_0 is like that of χ^2_{k-1-c} (asymptotically). Here c denotes the number of parameters to be estimated efficiently under H_0. In the case of a simple null hypothesis, Fisher and Pearson agreed. Pearson was unable to appreciate the modification needed if $c \geq 1$. Perhaps he already felt that his own p-values were too small if they were small, and now Fisher made them even smaller. Note that α_{Fisher} is (asymptotically) strongly similar.

Neyman and Pearson [63] used their likelihood-ratio criterion to derive test statistics. They claimed that Pearson's statistic $g(x)$ should be replaced by some function of the likelihood ratio $\lambda(x)$. Wilks [95] suggested the form $-2\log\lambda(x)$ and derived, under certain conditions, that the distribution of $-2\log\lambda(X_0)$ is approximately that of χ^2_r (where $r = \dim(\Theta) -$

$\dim(\Theta_0)$). The course of the history of statistics would have been different if these leaders would have been more critical with respect to the Pearson-Fisher idea to use the p-value $\alpha(x) = \mathrm{P}(\chi^2_r \geq -2\log\lambda(x))$ as 'the' degree of belief in H_0. Neyman was interested in the creation of a non-dogmatic theory of statistics, but he himself was fairly dogmatic about the likelihood-ratio principle and he did not criticize the idea that the p-value should be used. The p-value is, of course, a very convenient by-product of the Neyman-Pearson theory. By restricting the attention to the Neyman-Pearson formulation, Lehmann [54] was able to achieve his goal which was to *derive tests as solutions of clearly stated optimum problems*. Our purpose is similar though we are interested in deriving optimal q-values rather than optimal Neyman-Pearson level-α tests. A crucial situation is as follows, see also Stein [80].

Problem. Let $x \in \mathbb{R}^r$ be the outcome of $X \sim \mathcal{N}_p(t, I_r)$ and suppose that the testing of H_0: $t = 0_r$ is of interest. Which q-value α: $\mathbb{R}^r \to [0,1]$ should we use to estimate $\mathbf{1}_{\{0\}}(t)$ with squared error loss?

For the case $r = 1$ we obtained $\alpha_1(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2}$ as the q-value which minimizes the area under the mean squared error graph, among all weakly-similar procedures. We regard this result as 'reasonable and fairly satisfactory' though still some overconfidence is involved. For the general case $r > 1$ there are two very easy methods to satisfy the requirement of weak similarity. First, there is the p-value $\alpha_1(x) = \mathrm{P}(\chi^2_r \geq ||x||^2)$. Secondly, we can consider the likelihood ratio $\lambda(x) = \exp(-\frac{1}{2}||x||^2)$ and determine the constant c such that $\mathbf{E}(\lambda(X_0))^c = \frac{1}{2}$. It follows from

$$\mathbf{E}\exp\left(-\tfrac{1}{2}c\chi^2_r\right) = (c+1)^{-r/2}$$

that weak similarity is achieved if $c = 2^{2/r} - 1$. This provides

$$\alpha_2(x) = \exp\left(-\tfrac{1}{2}(2^{2/r}-1)||x||^2\right).$$

For $r = 1$ we obtain the result $\alpha_2(x) = \exp(-\frac{3}{2}|x|^2)$ discussed before as a modification of the p-value (into the wrong direction). For the case $r = 2$ (see also Salomé *et al.* [70]), we obtain

$$\alpha_2(x) = \exp\left(-\tfrac{1}{2}||x||^2\right)$$

which coincides with the p-value because $\alpha_2(x) = \lambda(x)$ in this case is such that $-2\log\lambda(X_0) = ||X_0||^2$ has the χ^2_2 distribution and $\mathrm{P}(\chi^2_2 \geq ||x||^2) = \exp(-\frac{1}{2}||x||^2)$. For $r \geq 3$ we obtain that the modification of α_1 provided by α_2 is in the right direction in the sense that small p-values are replaced by larger ones.

We are not satisfied by these results based on taking $\alpha(x) = \lambda(x)^c$ for an appropriate c. For $r \leq 2$, overconfidence will be displayed and for r large, 'diffidence' may appear (the values of $\alpha(x)$ being too close to $\frac{1}{2}$). We have arrived at the feeling that expressing a degree of belief in H_0 when testing this null-hypothesis $t = 0_r$ versus A: $t \neq 0_r$ itself is considerably questionable if r is 'not very small'. The reason is that one is formulating a null hypothesis which is extremely unlikely while the alternative is left unrestricted. See the end of Section 8 and the end of Appendix B.1, for further discussion.

The methodology of Subsection 7.1 is a useful tool with many Bayesian and non-Bayesian extensions, e.g. with respect to its implementation. Nevertheless one should not be dogmatic about this approach. Some complications were mentioned at the end of Subsection 7.1, and Subsection 7.2 dealt with situations where the methodology of Subsection 7.1 does not work. For the special case of a *real-valued function* f, these complications can (sometimes) be avoided by using the following approach which we regard as a proper interpretation of Fisher's 'fiducial argument'.

Suppose that a distributional inference $Q(x)$ is required about the 'true' value $y = f(t)$ of a given, nice and smooth, real-valued function f: $\Theta \to \mathbb{R}$, e.g. the function $f(\theta) = ||\theta||^2$ with respect to the problem studied at the end of Section 7.2. A procedure Q: $\mathfrak{X} \to \mathbb{R}^*$ for making such distributional inferences is then uniquely determined by the corresponding family $\mathcal{G} = \{G_x;\ x \in \mathfrak{X}\}$ of distribution functions where $G_x(z) = (Q(x))((-\infty, z])$, and it may be useful to make the requirement of strong similarity $\mathcal{L}\, G_X(y) = \mathcal{U}(0,1)$ introduced in Section 4 under Ad (2). Note that $y = f(t)$ is a deterministic constant: we are dealing with estimation, not prediction.

The idea is now to require that $G_x(z) = \alpha_z(x)$ is the 'most appropriate' p-value for testing H_z: $y \leq z$ against A_z: $y > z$. Such p-values are (asymptotically) uniformly distributed on $[0,1]$ if $y = z$ and, hence, the strong similarity of the procedure Q appears as a consequence.

We can now exploit theory from Lehmann's book [54] to construct such p-values as the smallest value of α for which the 'most powerful similar size-α test' (if it exists) rejects H_0. Along these lines, *uniformly best* strongly-similar procedures can sometimes be constructed. Details can be found in Salomé [70] and Kardaun and Schaafsma [46].

Example. If in the situation of the problem at the end of Subsection 7.2 a distributional inference is required about the true value $y = f(t) = ||t||^2$, then it is 'natural' to restrict the attention to the outcome $||x||^2$ of $||X||^2 \sim \chi'^2_{r;||t||^2}$ and to construct $Q(x)$ by defining $G_x(z) = \mathrm{P}(\chi'^2_{r;z} \geq ||x||^2)$ if $z \geq 0$. For $z < 0$ we, of course, use $G_x(z) = 0$. The jump $G_x(0) = \mathrm{P}(\chi^2_r \geq ||x||^2)$ at 0 corresponds to the p-value for testing H_0: $t = 0_r$ (because $||t|| = 0 \Leftrightarrow t = 0_r$).

We are fascinated by this neo-Fisherian way of generating distributional inferences and, hence, confidence intervals. There are, of course, some technical difficulties but more important than these mathematical details is the discussion about the question whether or not this identification of $G_x(z)$ with a p-value is appropriate. For the hypothesis testing problem H_z: $y \leq z$ versus A_z: $y > z$ the p-value may be 'fairly' reasonable but it is certainly not 'most' reasonable. Does the existence of alternative 'q-values' affect the reasonableness of the strongly similar procedure for making distributional inferences about y? The motivations for the concept of strong similarity, given in Section 4, were more convincing in the case of making a distributional inference than when testing a hypothesis. We believe that requiring

strong similarity (exactly or asymptotically) is very appealing in the case of distributional inference and is only moderately appealing in the case of hypothesis testing.

Example. Suppose that the outcome x of $X \sim \mathcal{N}(t, 1)$ has to be used to (1) make a distributional inference $Q(x)$ about the true value $t = \mathbf{E}X$ of the parameter $\theta = \mathbf{E}X_\theta$, where t, a priori, can be anywhere on the real line $\mathbb{R}$, and to (2) make a probability statement about H: $t \leq 0$ in the form of a 'very reasonable' degree of belief $\alpha(x)$.

Ad (1). Note that $Q(x) = \mathcal{N}(x, 1)$ is extremely appealing, firstly because it is anchored in physical probability see Section 5 and identify, for arbitrary z', the credibility $G_x(z' + x)$ of $t \leq z' + x$ with the probability $\mathrm{P}(t \leq z' + X) = \mathrm{P}(X - t \geq -z') = \Phi(z')$ to obtain $G_x(z) = \Phi(z - x)$ by substituting $z' = z - x$, and, secondly, because the procedure derived is optimal from a large number of perspectives (see Kardaun and Schaafsma [46]).

Ad (2). Note that the p-value $\alpha(x) = 1 - \Phi(x)$ associated to the 'optimal' distributional inference $Q(x) = \mathcal{N}(x, 1)$ just mentioned, is uniformly best strongly similar for testing H: $t \leq 0$ against A: $t > 0$, for a large variety of proper loss functions. If the attention is concentrated on squared error loss then the p-value is having the optimum property of minimizing the integrated mean squared error among all estimators α: $\mathfrak{X} \to [0, 1]$ of $\mathbf{1}_{(-\infty,0]}(t)$ and, hence, also among all weakly unbiased estimators (and among all equivariant estimators). Optimality of the p-values remains valid if the loss function is replaced by another proper one. Nevertheless, we are of the opinion that $\alpha(x) = 1 - \Phi(x)$ is only 'fairly' reasonable. Estimators exist which are more reasonable because 'they pay more attention to alternatives which are not too far from 0'. To derive such estimators, it is natural to restrict the attention to the estimators α: $\mathfrak{X} \to [0, 1]$ satisfying the equivariance property $\alpha(x) + \alpha(-x) = 1$ (such q-values are automatically weakly similar). The envelope risk (in θ) of these equivariant estimators is easily obtained by considering the posterior probability

$$\alpha_\theta(x) = \frac{\exp(-\frac{1}{2}(x + \theta)^2)}{\exp(-\frac{1}{2}(x + \theta)^2) + \exp(-\frac{1}{2}(x - \theta)^2)}$$

with respect to the prior τ which assigns probability $\frac{1}{2}$ to $-\theta$ as well as to θ ($\theta \geq 0$), this is best equivariant in θ. Van der Meulen [57] determined the value of θ such that the maximum regret is minimum among these 'somewhere minimum risk equivariant' procedures. Such statistical optimum properties were introduced in Schaafsma [73] by adapting ideas of Wald [91] about most stringent Neyman–Pearson tests. Van der Meulen [57] found $\theta \approx \frac{1}{2}$ and established, numerically, that the q-value

$$\alpha(x) = \frac{\exp(-\frac{1}{2}x)}{\exp(-\frac{1}{2}x) + \exp(\frac{1}{2}x)}$$

thus obtained is not much different from the equivariant estimator which has minimax regret among all equivariant estimators.

Discussion. The comparison between the (procedures behind the) p-value $1 - \Phi(x)$ and the q-value $1/(1 + \mathrm{e}^x)$ (and other q-values) is not straightforward. Much depends on the purpose

one has in mind. If this purpose is that of estimating $\mathbf{1}_{(-\infty,0]}(t)$ with squared-error loss then it is obvious from the preceding analysis that the area below the regret function of the p-value is smaller than that of any other q-value whereas the maximum of the regret function (assumed at 0) is larger than that of the q-value $1/(1+\mathrm{e}^x)$. Visual inspection of the graphs of the regret function, constructed by Van der Meulen [57], provides the impression that it is nicer to minimize the area under the regret function than to minimize the maximum. This impression changes if one takes into account that alternatives θ close to 0 are of much more interest than alternatives very far away from 0.

If the purpose of testing H: $t \leq 0$ against A: $t > 0$ is to choose one of the following decisions

$$a_0 : \text{maintain H: } t \leq 0$$

$$a_1 : \text{reject H and accept A: } t > 0$$

on the basis of an asymmetric loss function, where an error of the second kind costs c times as much as an error of the first kind ($0 < c < 1$), then the 'principle of minimum expected posterior loss' provides the solution that decision a_1 is made if and only if the degree of belief $\alpha(x)$ in H is less than c times the degree of belief in $1-\alpha(x)$ in A, i.e., if and only if $\alpha(x) \leq c/(1+c)$. It can be established easily that the use of the p-value $\alpha(x) = 1-\Phi(x)$ is 'optimal' in the following respect: the decision procedure

$$d(x) = \begin{cases} a_1 \text{ if } \alpha(x) < c/(1+c) \\ a_0 \text{ if } \alpha(x) > c/(1+c) \end{cases}$$

defined by it has uniformly minimum risk among all minimax risk procedures (or, equivalently, among all procedures which are 'Lehmann-unbiased', see Schaafsma [73,74]). For this 'special purpose' the p-value $1-\Phi(x)$ is certainly more appropriate than the q-value $1/(1+\mathrm{e}^x)$. Note that using the p-value is also optimal in the sense that the area under the risk function is minimum.

The situation changes somewhat if one considers the choice between the *three* decisions

$$a_{\mathrm{H}} : \text{accept H: } t < 0$$

$$a_0 : \text{ remain undecided}$$

$$a_{\mathrm{A}} : \text{accept A: } t > 0$$

and uses a loss function where an error of the second kind (decide upon a_0 whereas $t \neq 0$) costs c times as much as an error of the first kind (either deciding upon a_{H} if $t > 0$, or deciding upon a_{A} if $t < 0$). Using the p-value, of course, corresponds to the minimization of the area under the risk (or regret) curve. The minimax risk procedure provides a_0 for all $x \in \mathbb{R}$. To compromise between these principles, the Bayesian might consider some $\mathcal{N}(0, \tau^2)$

prior distribution. He will then have some difficulty in specifying τ^2. In Schaafsma [73,74] a non-Bayesian compromise is chosen based on the restriction by Lehmann-unbiasedness. He considers it 'optimal' to use the 'minimax-regret Lehmann-unbiased' procedure. For $c = 1/9$ and $c = 1/19$ the 'doubtful regions' $\{x;\ d(x) = a_0\}$, for different procedures, are as follows:

Procedure	$c = 1/9$	$c = 1/19$
Using the p-value	$[-1.65, +1.65]$	$[-1.96, +1.96]$
Minimax-regret Lehmann-unbiased	$[-1.73, +1.73]$	$[-2.02, +2.02]$
Using the q-value $1/(1 + \mathrm{e}^x)$	$[-2.20, +2.20]$	$[-2.94, +2.94]$
Minimax risk	$[-\infty, +\infty]$	$[-\infty, +\infty]$

Another development of the three decision problem in a Neyman-Pearson-Wald type formulation, where the loss ratios for the three types of error have to be settled by the end user of the test, has been elaborated in Kardaun [40, Section 2.3].

Conclusion. The issue of using the outcome x of $X \sim \mathcal{N}(t, 1)$ for testing H: $t \leq 0$ versus A: $t > 0$ by providing a degree of belief $\alpha(x)$ in favour of H, in the absence of further information, is not settled easily, since different considerations lead to different decision procedures. (A small consolidation is that for a number, albeit not all, practical problems, the three criteria above lead to the same decision. In case of conflicting decisions, i.e. doubtful regions $[-a, +a]$ with $2 < a < 3$, the situation is such that model assumptions and sample-size considerations should be (re-)scrutinized.) Lehmann (personal communication, 1969) criticized the constancy of the loss functions suggested in the two- and three-decision problem formulated. The actual loss resulting from an error of the first kind will be an increasing function of the distance between t and the hypothesis. If one specifies such loss function $\mathrm{L}(\theta, a)$ then application of the 'principle of minimum expected posterior loss' requires the availability of a distributional inference $Q(x)$ about t. We suggested under Ad (1) that the construction of such distributional inference is a less controversial affair than the making of a probability statement about H: $t \leq 0$.

8 Trying to develop a vision for the future

The history of statistics is part of the development of science at large. In this respect one might reread the quotation from Venn [89] cited in Section 1 and note that, in this paper, we restricted ourselves to induction in the sense of statistical inference, i.e. the evaluation of data such that randomness (and other) assumptions are acceptable. Unfortunately, completely compelling results did not appear. Some surrealistic reference to 'analogous situations' is often necessary. This, of course, is natural because a sample cannot supply full information about the population.

There is an analogy between the history of statistics and that of science at large. Our perspective is as follows. During the first half of the 20th century, science was dominated by the

modernists who had positive opinions about the possibility of mankind to arrive at consensus and to structure its discourse by being explicit about terminology, axioms, procedures, etc. In Austria there was the Wiener Kreis, in England the Unity-of-Science Movement, and in the Netherlands the Significa Movement. Their *constructivist* attitude is represented by the works of Hilbert, K. Pearson, R.A. Fisher, A.N. Whitehead, De Finetti, and many others. There were also critical comments, e.g. by Russell, Brouwer, etc.

During the second half of the 20th century, the mood was less optimistic. The *post-modernists* were very skeptical about the possibility of arriving at consensus by structuring the discourse. Their analyses of 'texts' made them very much aware of the hidden agenda's, the selection of arguments, the misuse of statistics, etc. Their *deconstructivist* attitude is represented by the work of Gödel in logic and of C. Stein in statistics. The Lehmann-Wald complete class theorem provides an interesting end-product of the idea of Neyman that a non-dogmatic theory of statistics has to be established. In the Netherlands Van Dantzig worried about the 'crisis of uncertainties'.

It is natural to think about 'what next?' From the philosophical side, there are two tendencies: (1) that of the *surrealists* who enjoy the 'ironic debate' about everything which can go wrong, and (2) that of the *reconstructivists* who accept the post-modernistic criticism but try to recover that what was 'good' in the ideas of the modernists. We hope that the reader will appreciate our attempts along this second line of thought.

In Subsection 7.2 we considered the special case $k = 2$ of Pearson's problem which, essentially, is Bayes's problem where the outcome x of $X \sim \mathrm{Bin}(n, p)$ has to be used to test H_0: $p = p^{(0)}$. We discussed various approaches, mostly based on asymptotic theory, to assign a degree of belief to this hypothesis . An exact method will now be presented by using q-values. The estimate $\alpha(x)$ is called a q-value if the underlying estimator α: $\{0, 1, \ldots, n\} \to [0, 1]$ of the true value $\mathbf{1}_{\{p^{(0)}\}}(\theta)$ of the indicator function of H_0 is optimal with respect to mean squared-error, or risk,

$$\mathrm{R}(\theta, \alpha) = \begin{cases} \mathbf{E}(\alpha(X_\theta) - 1)^2 \text{ if } \theta = p^{(0)} \\ \mathbf{E}(\alpha(X_\theta))^2 \qquad \text{if } \theta \neq p^{(0)}. \end{cases}$$

To enforce a 'solution', we introduce the requirement of weak unbiasedness, which is satisfied if and only if $\mathbf{E}\alpha(X_\theta) \geq \frac{1}{2}$ if $\theta = p^{(0)}$ and $\mathbf{E}\alpha(X_\theta) \leq \frac{1}{2}$ if $\theta \neq p^{(0)}$. A condition, necessary for weak unbiasedness, is that of *weak similarity*: $\mathbf{E}\alpha(X_{p^{(0)}}) = \frac{1}{2}$. In Albers *al.* [5] it is shown that the procedure $\alpha^* : \{0, 1, \ldots, n\} \to [0, 1]$ defined by

$$\alpha^*(x) = \frac{\binom{n}{x}\left(p^{(0)}\right)^x\left(1 - p^{(0)}\right)^{n-x}}{2\sum_{k=0}^{n}\left(\binom{n}{k}\left(p^{(0)}\right)^k\left(1 - p^{(0)}\right)^{n-k}\right)^2}$$

is the unique solution in the class of weakly similar q-values which minimizes the area $\int_0^1 \mathrm{R}(\theta, \alpha)\ \mathrm{d}\theta$ under the risk function. The q-value $\alpha^*(x)$ seems 'more reasonable' than the

p-value of Pearson and its 'exact' analogue of Fisher. Nevertheless, the danger of overconfidence still exists. Note that the procedure $\alpha^*(x)$ is weakly unbiased only in the case $p^{(0)} = \frac{1}{2}$. The procedure α^* just derived is 'very reasonable', but not 'most reasonable'. Such a 'most reasonable' solution does not exist.

Conclusion. Fisher, Rao, and many others tried to let the data speak, after a scrutiny of these data. The messages statisticians produce are 'context dependent'. Not only the data but also the 'context' has to be scrutinized. Compelling results will not be obtained but some results are 'almost compelling', many procedures are 'fairly' reasonable or, even, 'very' reasonable. Some skepticism, however, should be maintained. In Kardaun *et al.* [44] a remark was made about the making of probability statements on hypotheses of the form H: $y \in \mathscr{Y}_{\mathrm{H}}$. They argued that 'probabilities' are used to refine the making of yes-or-no statements. Unfortunately these probabilities in the degree-of-belief sense are often considerably unreliable in the range, say between 0.1 and 0.9, for which they are designed. Perhaps one should defer opinion in such situations. We will elaborate further upon this in Appendices B and C.

Appendices

We did not want to burden the main text with too much controversy, too many details, etc. Nevertheless it will, hopefully, be clear from the preceding that we are interested in reviving the essence of Fisher's ideas about fiducial inference and that we try to accomplish this goal by using a Neyman-Pearson-Wald approach based on loss functions, restrictions to classes of nice procedures, in particular those which are *strongly similar*. This, obviously, implies that we are somewhat critical with respect to the Bayesian approach. To clarify our position, Appendix A contains a discussion supporting the usefulness of the restriction by strong similarity while Appendix B provides a critical examination of various tendencies in statistics. We have to accept that Induction cannot be accomplished without incorporating something 'irrational', an argument by 'analogy' or some other 'surrealistic' construction of the mind. In Appendix C a variety of applications is discussed. Is the theory of strongly similar procedures (for making distributional inferences) of practical interest? That is the question which we shall try to answer with respect to these applications. The answer is, of course, that it is of some, limited, interest and that alternative Bayesian and non-Bayesian approaches should not be belittled.

A An example of the usefulness of the requirement of strong similarity

In Section 1 we referred to statistical science at large by quoting Venn [89]. Next, we restricted ourselves to induction in the sense of statistical inference. In this restricted area, it seems reasonable to use probabilistic terminology to express uncertainty. In the other areas indicated by Venn, it is usually impossible, misleading, and counter-productive to use such terminology (at least in the end report of an investigation). In some applications of mathematics, e.g. of dynamical systems, one is sometimes referring to examples from the real world to motivate some purely mathematical elaboration. The 'inferences' about the real world are then not, primarily, obtained by induction but by analogy. In mathematical statistics the issue is Induction, but to settle the issue we are often, more or less explicitly, referring to analogy. Anyway, in statistical inference, there are statistical and there are systematic uncertainties. Expressing uncertainty may have a negative utility in the discourse, especially if this discourse has the form of a debate. It may have a positive utility if a person has a debate with himself or if it can be assumed that all participants in a discourse are expressing 'their' uncertainties 'scientifically'. Even then it may be impossible to combine their opinions with the same scientific care (see Genest and Zidek [30]). A positive exception is as follows.

Example. Combining independent trustworthy opinions, in distributional form, about a 'deterministic' (physical) constant. Suppose that k research workers (or laboratories) have been involved in the determination of the true value t of a physical constant. Ignoring the possibility of systematic errors and of sources of information common to them, suppose that their investigations were performed *independently* and have resulted in distributional inferences

$Q_1, \ldots, Q_k$ with distribution functions $G_1, \ldots, G_k$ such that the underlying procedures are *strongly similar*. In Kardaun and Schaafsma [46] several methods are discussed to combine these opinions such that the combination procedure is strongly similar as well. One possibility is to take

$$G(z) = \Phi\left(k^{-1/2} \sum_{i=1}^{k} \Phi^{-1}(G_i(z))\right).$$

It is easy to see that the independence of the underlying data-generating random variables $X_1, \ldots, X_k$ implies that the $G_i(t) = G_{X_i}(t)$ are independent as well, all having the uniform distribution because of the strong similarity. Hence, $\Phi^{-1}(G_i(t)) \sim \mathcal{N}(0,1)$ and the independence implies that $k^{-1/2} \sum_{i=1}^{k} \Phi^{-1} G_i(t) \sim \mathcal{N}(0,1)$, with $G(t) \sim \mathcal{U}(0,1)$ as a consequence.

Conclusion of Appendix A. The requirement of strong similarity was introduced at the end of Section 4. Although some surrealism is involved also here, the example is one of the many instances where it is useful.

B Surrealistic tendencies in Statistics

In Section 8 we made it clear that we are interested in 'reconstructing' that what was good in the work of the modernists like Fisher, De Finetti, R.T. Cox, Neyman, and many others. We are not optimistic about the possibilities of mankind to organize 'scientific' discourses satisfactorily, in general. That is why we restrict ourselves to statistical inference and decision analysis where we hope to have a 'high-quality' scientific discourse leading to some consensus among the participants, see Appendix A. Post-modernists have emphasized the possibility of context-dependency, lying and cheating, withholding information, and of being overconfident. Example 1 of Section 2 is paradigmatic in this respect. When is a discourse of high scientific quality? As logic, mathematics, statistics, computation, inter-personal relationships, etc., may be involved, it is a delicate task to arrive at a 'concerted opinion'. A prerequisite, for some form of agreement is that a *scrutiny of the data* provides consensus about the set x of statistical data available for analysis. To *let the data speak* and *express statistical uncertainties*, methods of inference are needed. Which methods? That is the question which generations of statisticians have discussed. The scrutiny should, of course, be extended from the data to the *contexts* one invokes and to the *principles* behind the choice of method. Lehmann's book [54] was paradigmatic in this respect: he wanted to construct tests as solutions to well-defined optimum problems. In his treatment of the Neyman-Pearson theory, p-values appear as convenient tools: by reporting the smallest value of the nominal level of significance α at which H is rejected, one has a device to report the result of a Neyman-Pearson test for an arbitrary value of α. Lehmann was very critical with respect to the use of such 'credibility' as a 'probability'. We share his concerns and, in this respect, agree with the Bayesians who, for other reasons, rejected the use of p-values as degrees of belief (see numerous papers, especially in Berger [7]).

As indicated in Section 8, we are somewhat critical about many 'surrealistic tendencies'. It

is however unavoidable that the discussion involves some surrealistic element (cf. Mumford [62]). To let the data speak, some mathematical argument will be needed to extrapolate from what is given to what is unknown. Whether one uses words like rational, mathematical, optimal, coherent, unbiased, natural, etc., one should never forget that some fiction, some surrealistic argument is involved. Statistical inferences and decisions are mixtures of facts and fictions. If the factual input is overwhelming then one should not worry, but in statistical practice the attention is often concentrated on situations where the factual evidence is not overwhelming. Which rationalizations are appropriate? Some 'lines of thought' are as follows.

B.1 Emphasis on parametric models

By assuming that $P \in \mathscr{P} = \{P_\theta;\ \theta \in \Theta\}$, the statistician is doing something he is not really believing in, at least not if assumptions of 'normality', 'exponentiality', or 'homogeneity' are involved. In Section 5 we went beyond such a parametric context and in Section 8 we considered Bayes' problem where the assumption $P \in \mathscr{P}$ is completely acceptable (though the assumption that t can be anywhere in $\Theta = [0, 1]$ is somewhat peculiar). Parametric models are often specified as the result of some preliminary analysis of the data. The assumption $P \in \mathscr{P}$ will not be made if the null hypothesis (that P is indeed in $\mathscr{P}$) is rejected at some nominal level of significance, say $\alpha = 0.05$. It is in this respect that the following discussion is of some interest.

Modern statistics started with Pearson's goodness-of-fit test, see Stigler [82], Kotz [52], and Hald [33]. In Section 7.2 we worried about the question whether it is reasonable to assign a probability (in the degree-of-belief sense) to H_0: $t = 0_r$ if the outcome x of $X \sim \mathcal{N}_r(t, I_r)$ is available. For $r = 1$ we arrived at something reasonable, not the p-value but some q-value. For r large, say $r > 3$, we believe that the testing of such null hypotheses should be avoided as much as possible, because agreement about procedures is difficult to achieve. A discussion with the problem-owner may, in some situations, provide a relevant simplification of the problem.

Example. One of us (Schaafsma) was involved in a controversy among physical anthropologists (Culotta [19], Van Vark aand Bilsborough [87,88], Wolpoff [96]). Wolpoff had expressed the opinion that a sample of Levantine Neanderthal skulls was not more dispersed than the modern world population. Van Vark and Bilsborough arrived at the opinion, after analyzing the data available, that the Levantine skulls were more dispersed because the 'appropriate' χ^2 test (on a very large number of degrees of freedom) provided an extremely small p-value. A scrutiny of the computational results, however, indicated that for some variables (principal components) the Levantine skulls were significantly *less* variable whereas for other variables (the majority) they were significantly *more* variable. It was clear that the χ^2 methodology was inappropriate. In our opinion, the research worker should specify a function f: $\Theta \to \mathbb{R}$ such that $f(t) > 0$ indicates that the Levantine population is less variable and that $f(t) < 0$ indicates that it is more variable than the present-day world population (taking the general-

ized variance provides a natural possibility). This supports our view that many applications of χ^2 tests (based on a large number of degrees of freedom) are not very appropriate because null hypotheses of the form H_0: $t = 0_r$ should not be tested by using the χ^2 statistic 'blindly'.

B.2 Emphasis on methods of inference

If one needs an inference or decision on the basis of a particular set x of data then it is considerably peculiar to shift the attention to methods or procedures applicable to arbitrary sets ξ. In some pattern recognition problems this approach is natural because the set x given is only one out of the many to follow. But in other situations there is an alienation from the actual issue which is to evaluate the data x at hand. The next step is even more problematic. It may be natural to discuss whether, for instance, an estimator is unbiased in the factual sense that $\mathbf{E}d(X) = g(t)$ for the true value t of θ, and to discuss its mean squared error $\mathbf{E}(d(X) - g(t))^2$, but it is much less natural and, in fact, extremely surrealistic to require that $\mathbf{E}d(X_\theta) = g(\theta)$ holds for *all* theoretically possible values θ of t and to study the risk function $\mathbf{E}(d(X_\theta) - g(\theta))^2$ for $\theta \in \Theta$. We use the notation X_θ to denote any random variable having the distribution X would have had if t would have been equal to θ. In our notation $X = X_t$ is the random variable X: $(\Omega, \mathcal{F}, \mathrm{P}) \to \mathfrak{X}$ actually observed. The usual notation does not emphasize this difference. In logic, paradoxes arise from the comparison of various 'worlds', one or none of these worlds being the real one, see Albers *et al.* [4]. With respect to distributional inference one should be aware of the surrealism involved in requiring that $\mathcal{L}\ G_{X_\theta}(g(\theta)) = \mathcal{U}(0,1)$ holds for all θ. It is a miracle that, in some situations of practical interest, such rationalizations provide something useful.

B.3 Wald's minimax risk and related minimax regret suggestion

In his review of Von Neumann-Morgenstern [59], Wald [93] expressed his view that the problem of constructing a statistical decision function can be regarded in game-theoretic terms, Nature playing the part of Player I and the statistician being Player II. By using this perspective, minimax risk procedures can sometimes be derived. The actual situation is different because Nature is not particularly interested in 'choosing some prior to maximize Nature's expected profit'. From a practical point of view, minimax risk procedures are often useless since they concentrate too much on the 'worst case'. If, e.g., one wants to construct a minimax risk q-value for testing H: $p = p^{(0)}$, then one arrives at the trivial procedure $\alpha \equiv \frac{1}{2}$. Here 'least favourable priors' concentrate in the close neighborhood of $p^{(0)}$. In Schaafsma [74] a theory of multiple decision problems was developed by generalizing ideas of Lehmann. Unfortunately, the situation was degenerated: relevant results based on Wald's minimax risk requirement and on Lehmann's requirement of decision-theoretic unbiasedness are much less abundant than adherents of the Neyman-Pearson-Wald school have hoped.

B.4 The Lehmann-Wald Complete Class theorem

A beautiful result established in Wald [94] and motivated by earlier results of Lehmann, is that procedures which are not (extended) Bayes have risk function which can be made smaller everywhere by choosing a convenient (extended) Bayes procedure. This result provides some support for the Bayesian approach. There are, however, alternative properties, like those of unbiasedness or similarity, which may be lost by applying such improvement. The discussion about the choice of a procedure remains difficult and, perhaps, some simple 'unbiased' procedure, not too far from the class of admissible ones, is 'most recommendable'.

B.5 De Finetti's Representation Theorem

This beautiful result establishes *axiomatically* that, in the situation of Bayes's problem, any probability measure Q on the Kolmogorov σ-field $\mathcal{F}$ in $\{0,1\}^\infty$ is exchangeable (i.e. invariant under permutations) if and only if there exists a prior distribution τ on $[0,1]$ such that

$$Q(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\, \mathrm{d}\tau(\theta).$$

The theorem is used by personalist Bayesians to support their idea that the making of inferences or decisions, on the basis of statistical data, cannot be performed 'rationally' or 'probabilistically coherently', unless one adopts the Bayesian approach and accepts full responsibility for the choice of some prior distribution τ. In response to the discussion by J.M. Bernardo in Kardaun *et al.* [44] it has been argued that some type of Munchhausianism is involved in this interpretation: one starts from the existence of some numerically specified probability measure Q, different from the true but unknown one P, on the Kolmogorov σ-field in $\{0,1\}^\infty$, and derives the existence of a distribution τ on $[0,1]$ such that the probability measure Q is of the form indicated. However, such probability measure τ neither exists in the real world nor in the mind of the person who is involved in the inferential business: before making any observation he knows nothing exactly. To impose the existence of a probability measure Q on $\mathcal{F}$, governing all epistemic probabilities, is a matter of wishful thinking overemphasizing some vague analogy between epistemic and physical probabilities.

B.6 Using equivariance principles

In the objectivist-Bayesian literature and Neyman-Pearson-Wald literature one likes to refer to situations where a 'natural' solution can be obtained by requiring the invariance or equivariance of some procedure. Any mathematical mind will appreciate the beauty and usefulness of such result. The snake in the grass is that a surrealistic argument is involved, another 'Munchhausianism' as discussed in Kardaun *et al.* [44]. This is as follows. One starts from the belief in the existence of a 'natural' procedure. Next one explores the group invariance

of the situation and makes the suggestion that in some transformed situation, isomorphic to the original one, the same procedure should be applied because it was 'natural'. Next one makes interpretations of the requirement of equivariance (or invariance) thus motivated. If this results in a 'uniformly best equivariant procedure' then it looks as if the premise of everything, the existence of a 'natural' procedure, has been established.

B.7 Motivations for the Bayesian approach using the axiomatics of R.T. Cox and Jaynes

Here, again, some surrealistic fallacy appears. Cox [16] argues as follows: "every conjecture is based on some hypothesis, which may consist wholly of actual evidence or may include assumptions made for the argument's sake. Let h denote an hypothesis and i a proposition reasonably entitled to partial assent as an inference. The probability is a measure of this assent, determined more or less precisely, by the two propositions, i and h. It is therefore a numerical function of propositions [...] Let us denote the probability of the inference i on the hypothesis h by the symbol $i|h$. The probability on the hypothesis h of the inference formed by conjoining the two inferences i and j is represented in the notation just given by $i \cdot j|h$. By the axiom (1.ii) this probability is a function of the two probabilities: $i|h$, the probability of the first inference on the original hypothesis, and $j|h \cdot i$, the probability of the second inference on the hypothesis formed by conjoining the original hypothesis with the first inference."

The snake in the grass is that Cox forgets the provision he formulates so wisely in the beginning, by using the worlds 'reasonably entitled to partial assent as an inference'. By continuing on the basis of the existence of a numerical function $i|h$ which can be applied to arbitrary (i, h), e.g. to $(i \cdot j, h)$ and to $(j, h \cdot i)$, he ignores that many propositions i may *not* be 'reasonably' entitled to partial assent (given h). In statistics some wishful thinking will be needed but we should restrict ourselves to situations and procedures such that the inferences we make are 'sufficiently trustworthy'.

B.8 Other issues related to the limited usefulness of probabilistic coherency

The purpose of making probability statements and distributional inferences is, of course, that they can be used as if they were factually true. They, of course, are not. Fisher was criticized because he manipulated epistemic probabilities (with respect to the Behrens-Fisher problem) as if they were Kolmogovorian. Distributional inferences are defined by requiring that epistemic probabilities fit within the framework of a probability measure. How much probabilistic coherency should we impose, or tolerate? If a Neyman-Pearson-Wald statistician has proposed some procedure, e.g. a test, a confidence interval or a classification procedure, then this procedure can be applied to the data x at hand. The underlying pre-data characteristics of the procedure (probabilities of error of various kinds, e.g. the probability a priori that the true value $y = \eta(t)$ will not be contained by the confidence interval prescribed) are obvious. They may depend on the true value t of the underlying parameter. If they, indeed, depend on

t then they are unknown and a numerical specification will require their estimation. This will involve complications especially because the pre-data characteristics of a procedure lose much of their appeal after the data x are made available. Several authors, e.g. Blyth and Staudte [8,9], are interested in an assessment of data-dependent probabilities of the two types of error after a Neyman-Pearson test has been performed. Restricting ourselves to the simple case of testing H: $t \leq 0$ versus A: $t > 0$ on the basis of the outcome x of $X \sim \mathcal{N}(t, 1)$, a Type I error is committed is $x \geq \Phi^{-1}(1-\alpha)$. Given such outcome x, what is a reasonable assessment of the degree of belief one should have in H: $t < 0$? Elsewhere in this paper we suggest that, for arbitrary $\xi \in \mathfrak{X}$, it is reasonable to use $(1 + \mathrm{e}^{\xi})^{-1}$ as degree of belief in favor of H. Whether this assessment should be affected by the knowledge that $x \geq \Phi^{-1}(1-\alpha)$ is a vexed issue which we shall not discuss here, see Kardaun and Schaafsma [46] for details.

Remark. It is also possible that a pre-data characteristic does not depend on t and one nevertheless would like to make an assessment of the 'underlying concept' a posteriori. The original motivation for making assessments of probability by estimating indicator functions with squared-error loss came, partly, from Kiefer [50] who was interested in the posterior probability that the true value $y = \eta(t)$ is in some Neymanian confidence interval $[\underline{d}(x), \overline{d}(x)]$. The prior (or pre-data) probability is equal to the confidence coefficient $1-\alpha$ but the posterior probability will depend on the outcome x observed. Kiefer tried to estimate or, rather, predict the indicator function $\mathbf{1}_{[\underline{d}(x),\overline{d}(x)]}(y)$. Here, again, it is a vexed issue whether or not one should use the estimate $Q(x)([\underline{d}(x), \overline{d}(x)])$ suggested by any 'optimal' procedure Q: $\mathfrak{X} \to \mathbb{R}^*$ for making distributional inferences about y. Our answer is that some adaptation is necessary, at least in general, but also that it is difficult to decide upon such adaptation.

Conclusion of Appendix B

In practice, statistical inferences and decisions cannot be dispensed with, cf. Draper and Smith [24]. If the factual evidence is weak then, perhaps, one should defer the formulation of opinions and the making of decisions. As the evidence accumulates, however, one will comply with the need to express opinions quantitatively by making probability statements, distributional inferences, etc., as part of the basis of subsequent decision making. Underlying methods, however, are not easily agreed upon. The suggestions made, discussed and criticized in Sections B.1, ..., B.8, may all be useful, at least to a certain extent. In the non-dogmatic approach we suggest, it is a matter of 'muddling through', since a universally acceptable approach does not exist. This conclusion is not unlike that which Berger indicated in his conversation with Wolpert [97].

C Strongly similar distributional inferences in practice

Statistical inferences are needed to quantify the scientific discourse. Are the suggestions made in this paper, those with respect to the requirement of strong similarity in particular, of

practical interest? This question will be discussed from a variety of perspectives, all based upon the professional experience of (some of) the authors.

It is not difficult to anticipate that the conclusion will be similar to that made by Savage [72] about Fisher's Fiducial Inference. There is only a 'handful of applications' if *exact* results based on a *uniformly best* strongly similar procedure are required. But there are many more applications if one regards the requirement of strong similarity as something which, indeed, should be satisfied *exactly*, the question of optimality or uniqueness of recommendation being only of secondary interest. The number of applications is without bound if the requirement of strong similarity is only regarded as an 'ideal' to be *approximated*. Much asymptotic theory can be interpreted in this way. It leaves the mathematical statistician unsatisfied: one would like to define and establish the 'asymptotic optimality' of sequences of procedures proposed. This, however, is extremely complicated (see Wald [92], Snijders [78], etc.).

C.1 Medical statistics

Several of the authors are interested in the areas of (medical) diagnosis and prognosis. In principle this is a matter of making technological extensions of Example 2 of Section 3 (the problem of the Reference Class, see also the end of Section 6). Theories and computer programs have been developed to construct CONfidence intervals of POSterior probabilities and confidence bands of survival functions. There is a relationship with the GRECON project to be discussed in Section C.2. Indeed, the emphasis in the programs POSCON (used in Engelmoer and Roselaar [27], see also Kardaun and Schaafsma [46]) CONCERN (see Kardaun (and Schaafsma) [40,46]) is on making simplifying model assumptions such that not too many parameters have to be estimated. Anyway, the statistical and systematic uncertainties involved in this business are such that, in our opinion, probability statements may be reliable if the probabilities are close to 0 or 1, i.e. in situations where almost no uncertainty exists. But they may be very unreliable in the interval between, say, 5% and 95% for which they are designed. This suggests that the task of expressing uncertainty adequately is a formidable one. D.R. Cox *et al.* [15] formulated this as follows: 'in clinical practice, it is quite common for information on 100 or more variables to be routinely collected in each patient, giving the statistician the unenviable task of summarizing the joint effect of these variables on survival'. By computing standard errors of posterior probabilities, we can display statistical uncertainties and by repeating the analysis for a different model, we get an impression of the systematic uncertainties. These latter uncertainties should also be expressed but we do not have a general prescription how to proceed in practice, cf. Appendix A and the reference to Genest *et al.* [30] made there.

C.2 Expressing uncertainties in Econometrics

Econometrics has part of its roots in The Netherlands, because of the important contributions by Nobel prize winner Jan Tinbergen (1903-1994) and Henri Theil (1924-2000). Especially Tinbergen can be recognized as one of the founding fathers of econometrics in the 1930's. Tinbergen wrote his PhD Thesis in physics under the supervision of Ehrenfest; an appendix gives applications to economics. He worked as a statistician with Statistics Netherlands (Central Bureau of Statistics, CBS) from 1929 to 1945. For the Dutch economy he constructed the first complete macro-econometric model, which was rather simple, but it could be used for economic policy making. He became the first director of the CPB, the Netherlands Bureau for Economic Policy Analysis. CPB was formally founded by law in 1947 and still plays an important role in the preparation of economic policy. Nowadays macro-econometric models are indispensable for developing economic policy. In principle the basic data are in the form of a multivariate time series. If various countries are considered simultaneously, or regions within a country, then a relationship with the subject of C.3 will appear.

In 1977 the GRECON-group, a group of Groningen econometricians, originally consisting of B. Bos, R.H. Ketellapper, M.A. Kooyman, and W. Voorhoeve from the Econometric Institute of the University of Groningen, The Netherlands, built a rather small model for the Dutch economy. The models used by CPB were large, the number of predetermined variables being larger than the length of the sample period. The two-stage least squares (2SLS) estimation procedure (cf. [20]) for simultaneous econometric models could therefore not be applied. Moreover, predictions from large models are not very accurate. This should be expressed by making prediction intervals. The GRECON-group constructed a relatively simple model that could predict the most important economic target variables, such as the rate of unemployment, the price level of private consumption, and total production. In Ketellapper *et al.* [49] this model is presented. Its estimation (on the basis of 2SLS methodology) was outlined. Moreover, the uncertainties in the predictions were described by means of estimated standard deviations. It is not unreasonable to inform those involved in discussions about macro-econometric policy that statistical and other uncertainties are involved. The GRECON predictions were of the form $\hat{y} \pm 2s$, where $\hat{y}$ denotes the prediction and s the standard deviation of the predictor. The choice of these intervals was based on asymptotic theory using linearity and normality assumptions. The prediction error was considered to be approximately normal. This shows that, in principle, predictions were made in the form of $\mathcal{N}(\hat{y}, s^2)$ distributional inferences. At those days, the predictions by the CPB were in the form of point predictions. The statistical uncertainties had to be ignored because the models used were so large that 2SLS methodology could not be applied. In Dehling *et al.* [21] a comparison was made of the CPB point predictions and the GRECON predictions in the form of $\mathcal{N}(\hat{y}, s^2)$ distributional inferences, based on a loss function $\mathrm{L}(y, Q)$ which they regard as the distributional analogue of squared-error loss (more can be found in Kardaun and Schaafsma [46]). The CPB point predictions are identified as distributional inferences $\mathbf{1}_{\{\hat{y}\}}$ with point mass 1 in the estimate (/prediction) $\hat{y}$.

The loss function $\mathrm{L}(y, Q)$ introduced is such that, interestingly,

$$\mathrm{L}(y, \epsilon_{\hat{y}}) = |y - \hat{y}|$$

corresponds to the absolute error, not its square. The loss function $\mathrm{L}(y, Q)$ is such that it pays to express uncertainty if this is done adequately. That is part of the reason why the losses of the predictive inferences made by GRECON were somewhat smaller, on the average, than those of the predictions made by CPB.

At present complicated macro-ecomic models are widely used. In principle they determine the joint distribution $\mathcal{L}\,(Y_1, Y_2, \ldots)$ of a vector Y of variables at time points $t = 1, 2, \ldots$ where observations are made. The model contains some unknown parameters (regression coefficients, variances, etc.) which have to be estimated from the data – the outcome $x = (x_1, \ldots, x_n)$ of the vectors $Y_1, \ldots, Y_n$ before time point $n+1$ (see also Section C.3). In practice the situation is complicated by the fact that the data available at some point in time are not exactly known: it takes some time before the actual value of some macro-economic figure becomes available. Anyway, computational facilities make it possible to simulate predictions from a model. A point-prediction is determined by substitution of the values of the coefficients and the values of the predetermined variables. On their turn these values are generated from certain predetermined probability distributions. In this way the output distribution of a predictor can be simulated depending on the uncertainties with regard to the coefficients and the uncertainties with regard to the values of the predetermined variables. By plotting its density function using smoothing techniques, a graphical illustration of the relative likelihoods of the possible outcomes of the predictor is obtained. This is, of course, a pragmatic form of distributional inference which may turn out to be skewed or heavy tailed. Since the generation of graphical representations has become relatively easy, many researchers construct, quite naturally, distributional inferences (without using this terminology, however). There are, of course, many issues involved. We discuss two of these, one with respect to the evaluation and one with respect to the precise specification. The precise specification of such distributional inference depends on whether the density simulated is directly regarded as an approximation to the ideal inference $\mathcal{L}\,(Y_{n+m} \mid Y_1 = x_1, \ldots, Y_n = x_n)$ or as the basis of an approximation to the distribution of the prediction error. In the latter case one will obtain an estimate of the density of the prediction error by shifting the simulated density over a distance $\hat{y}$ to the left. If one has a point estimate $\hat{y}$ of y and an estimate $\mathcal{L}\,(Z)$ of the distribution of the prediction error then it is *not* most natural to use the density of the distribution simulated originally (it corresponds to $\mathcal{L}\,(\hat{y} + Z)$ but it is more natural (if y refers to location rather than scale) to report $Q(x) = \mathcal{L}\,(\hat{y} - Z)$ as the distributional inference about the true value $y = y_{n+m}$. If normality is imposed, or symmetry around 0, then $\mathcal{L}\,(\hat{y} + Z) = \mathcal{L}\,(\hat{y} - Z)$ and there is no difference.

The evaluation of predictive distributions is a matter of concern. On the one hand one can use the loss function $\mathrm{L}(y, Q)$ just discussed. On the other hand, one can refer to the value $G(y)$ of the distribution function of the predictive distribution Q in the true value y. Such $G(y)$

should behave like drawings from $\mathcal{U}(0,1)$, at least if one wants to respect the requirement of strong similarity.

C.3 Spatial statistics

Whereas time-dependence was essential in C.2, space-dependence is essential in spatial statistics (Cressie [17]). In principle one is interested in variations of a vector $Y(s,t)$ of variables in space and time, but in our introduction the attention is restricted to only one variable $Y(s)$ at location s in some 2-dimensional space, such as a farmer's field, a forest, the soil under contaminated water, etc., but also 1-dimensional lines and 3-dimensional volumes may be considered. (For the 1-dimensional case there is a close correspondence with the time-series models studied in C.2.) As variable we may consider the soil-pH, the concentration of lead in an environmental study, or the amount of rainfall. The question then is, given a finite set x of realizations $x_1 = y(s_1), \ldots, x_n = y(s_n)$ of $Y(s_1), \ldots, Y(s_n)$, firstly to estimate (or predict) the value $y = y(s_0)$ of $Y(s_0)$ at an unvisited location $s_0 = (s_{01}, s_{02}) \in \mathbb{R}^2$ (with estimation in the sense of point estimation), secondly to indicate the statistical uncertainty in this point estimate, e.g. by studying the variance of the distribution of prediction error, thirdly to obtain a confidence interval for the true value $y = y(s_0)$ to be predicted or, almost equivalently, to obtain a distributional inference Q about this true value y. Note that it is often not possible to obtain multiple samples at a single location, and that, for inferential purposes, it is unrealistic to shuffle the spatially collected data.

To be able to make any inference, an assumption of stationarity is to be made (Matheron [55]). Several of these occur in the literature, the most common ones being, in an order of decreasing strength, as follows.

Strong stationarity: translation invariance is required, i.e.

$$\mathrm{P}(Y(s_1) > y_1, \ldots, Y(s_n) > y_n) = \mathrm{P}(Y(s_1 + h) > y_1, \ldots, Y(s_n + h) > y_n)$$

for any translation vector $h = (h_1, h_2)$; i.e. the joint probability distribution of any finite set of variables is independent of translation.

Second order stationarity: $\mathbf{E}Y(s) = \mu$, independent of the location s, and

$$\mathrm{Cov}(Y(s), Y(s+h)) = C(h)$$

for some positive definite function $C(h)$, often called the covariance function, only depending on $h = (h_1, h_2)$ and not on location $s = (s_1, s_2)$.

Intrinsic stationarity: $\mathbf{E}Y(s) = \mu$ independent of location s and $\mathrm{Var}(Y(s) - Y(s+h)) = \frac{1}{2}\gamma(||h||)$ for some negative-definite function $\gamma(||h||)$, called the variogram, only depending upon $||h||$ and not on location $s = (s_1, s_2)$.

To obtain a prediction $\hat{y}$ of the true outcome y of $Y(s_0)$, ordinary 'kriging' offers a linear combination of the observations under the condition that the variance of the prediction error

is minimized. Ordinary kriging also provides us with the kriging variance, being equivalent to the prediction error variance. Cokriging includes data from two or more variables (Stein and Corsten [81]). It is not too difficult to extend this ordinary kriging estimate towards blocks, i.e. average values for parcels. Note that some estimate of the variogram is needed and that statistical and systematic uncertainties involved in this estimation are ignored (Cui *et al.* [18]). Having an estimate $\hat{y}$ and a prediction-error variance s^2, one can proceed as in C.2 by making a distributional inference $\mathcal{N}(\hat{y}, s^2)$ about y or by constructing an approximately 95% confidence interval $\hat{y} \pm 2s$. If y is known to be in $\mathbb{N}^0$ then the inference $\mathcal{N}(\hat{y}, s^2)$ will be improved if the corresponding probabilities of $(-\infty, \frac{1}{2})$, $(\frac{1}{2}, \frac{3}{2})$, etc., are assigned to the possibilities 0, 1, etc. (For a refinement of this statement, see Kardaun and Schaafsma [46].)

There are various reasons to question the 'probabilities' assigned by $\mathcal{N}(\hat{y}, s^2)$ (or by its discrete analogue) to some interval. That is one of the reasons why several alternative procedures have been constructed to estimate the probability that a value at an unvisited location occurs within an interval. The most simple procedure, called indicator kriging, considers a threshold value y_c. Then the probability that a value less than y_c occurs is obtained by considering the random variables $\mathbf{1}_{\{Y(s)<y_c\}}$. One of the possibilities to arrive at an assessment of the probability of $Y(s_0) < y_c$ required is to use a linear prediction of $\mathbf{1}_{\{Y(s_0)<y_c\}}$ on the basis of the outcomes $x_1, \ldots, x_n$ of $Y(s_1), \ldots, Y(s_n)$ observed. As an alternative to the assessment of $\mathrm{P}(Y(s_0) < y_c)$, one may consider disjunctive kriging (Matheron [56]), which transforms the observed distribution by means of, for example, Hermite polynomials to an approximation of the Gaussian distribution, leading to a prediction for which the distribution can be derived from the assumed distribution of the data, followed by a back-transformation. Model-based geostatistical procedures, including Bayesian forms of kriging (Diggle *et al.* [22]), allow a more general approach using the underlying spatial models. They extend the procedures described so far to also include count data, for example. Clearly, by being able to estimate $\mathrm{P}(Y(s_0) < y_c)$, one is also able to estimate $\mathrm{P}(a < Y(s_0) < b)$ for any $a < b$.

In several studies, the use of such a probability estimating method is avoided, because it is sometimes cumbersome to apply and provides only limited information if a full distribution is required at an unvisited point. Such a distribution may be Gaussian, but may also be unspecified, skewed, finite, bimodal, or may have another anomaly. If that is the case, spatial statistics often relies on spatial simulations. With the advent of modern computer systems, increasingly many simulation approaches are developed. Common procedures are matrix decomposition, the turning bands method and sequential procedures. The state of the art is similar to that of Appendix C.2. A nice overview is provided in Chilès and Delfiner [13].

C.4 Statistical inference in microarray analysis

The technology of so-called microarrays is a young and challenging field in genomics and bioinformatics. Microarrays are high-throughput mechanisms designed to measure the quantity of ribonucleic acids (RNA) in different samples. One of the most common types of microarrays

are so-called cDNA microarrays, where the abundance of 'single-stranded DNA' is measured for two biological samples (e.g. skin tissues). This level of abundance yields insight in the properties of the genes of the samples studied. Common medical research focusses on, e.g., cancer research where a healthy cell tissue is compared with a diseased one in order to learn which genes play a role in the development of the disease. Common plant research investigates issues like yield-improvement and immunity for certain deceases. Many other biological and medical applications exist. See, e.g., Sebastiani *et al.* [77] for further introduction. The data obtained are in the order of thousands of measured gene expressions for both samples, and these data arise as drawings from convolutions of various different, and unknown, statistical distributions. A statistical formulation is as follows. The data given are the realizations of $X_1, \ldots, X_n$, arising from an unknown distribution F. Required is an estimate r of ρ, which is defined as some function of F. An example from medical studies is that where the measurement for patient i is $X_i = (X_{1,i}, X_{2,i})$ where $X_{1,i}$ is, e.g., a p-dimensional vector (with, e.g., $p = 2500$) of measured gene expressions (usually, each gene is measured a couple of times to reduce technical variation) and $X_{2,i}$ some quantitative measure of the severity of the disease. To study the relationship between the gene expressions and the severity of the disease, one may try to derive a point prediction of the outcome y of X_2 or make a distributional inference or, simply, study the conditional distribution of X_2 given X_1. Another possibility (see C.3) is to define a threshold value c for the severity of the disease and to replace X_2 by a binary variable $Y = \mathbf{1}_{[c,\infty)}(X_2)$.

If, e.g., the attention is concentrated on the prediction of such binary variable Y, then one may try to design a classification statistic $\rho(X)$, e.g. a linear one of the form $a + b'X_1$, discriminating between the cases with $y = 0$ (using $a + b'x_1 < 0$ as indicator) and those with $y = 1$ (where $a + b'x_1 > 0$ is needed). A feature of this technology, in common with the previous appendices, but more dominant, is that thousands of explanatory variables are measured whereas the sample sizes available for estimation (or prediction) purposes are quite small. This problem is known as the 'small n, large p'-problem. In microarray analysis, it is not uncommon that experiments are carried out with only about 40 patients(/plants/test animals/etc.), while it is known that, due to technical and biological variation, the variance of the measurements is quite large, even when taking some replicates.

Dudoit *et al.* [25] describe a three-step approach to analyze such problems in a statistically sound manner, which is, in general, also applicable to other fields of applied statistics, but described here for the field of microarray analysis:

(i) an intensive and thorough search of the parameter space to generate good candidate estimators. To do this within satisfactory computer-time, it is often necessary to reduce the dimensionality of the problem under investigation. Reduction on the basis of known biological/medical reasons is preferable, since this reduction can take place before the collection of data. It is not uncommon to also reduce dimensionality technically by classifying the genes studies in a (much lower) number of classes, based on their measured properties. This is done by (adapted versions of) multivariate techniques such as cluster analysis, discriminant anal-

ysis and principal component analysis. See Speed *et al.* [79] and Amaratunga and Cabrera [6] for an overview of the usage of these techniques in microarray analysis.

(ii) an approach for selecting an optimal estimator among these candidates. One approach to do this, in line with the thoughts of this article, is to specify some parametric loss function, and to minimize some characteristic of the corresponding risk function. Among the most commonly used techniques in microarray analysis are multinomial logistic (or probit) regression classifiers and 'naive Bayes classifiers'. 'Although these model-based approaches provide a quantification of the uncertainty of the predictive model [..] model-free approaches are currently the most popular' (Sebastiani *et al.* [77]). These model-free methods include Fisher linear discriminant analysis, nearest neighbor classification trees and support vector machines (see, e.g., Vapnik [86]).

(iii) a method for reliably assessing the performance of the resulting estimator. To assure that the found estimates describe real biological processes and are not just due to the abundance of the data, these estimates have to be validated. Usually this is done by permutation-like methods as the bootstrap (cf. Kerr and Churchill [48]), cross-validation (cf. Dudoit *et al.* [25]) or similar techniques.

C.5 Scientific discourse on the design of a future tokamak

The tokamak (a Russian acronym of 'тороидальная камера с магнитными катушками', i.e., a toroidal chamber with magnetic field coil, a name coined by Yavlinski (see Dolgov–Saveliev [23], Mukhovatov [60], Braams and Stott [11]) is a promising type of machine used in magnetic fusion research. The goal is to confine a toroidal plasma (i.e. an ionized gas, in which electromagnetic forces play a predominant role) in a strong magnetic field such that, by elevating the density (the number of ions and electrons per unit volume) and by creating (through additional heating) a sufficiently elevated temperature, an appreciable amount of energy per unit time is generated from nuclear fusion reactions. In principle, the working of such a (prospective) fusion reactor is as follows. The presently considered fuel components are deuterium and tritium. By the process of nuclear fusion ($\mathrm{Deuterium}_2^1 + \mathrm{Tritium}_3^1 \rightarrow \mathrm{Helium}_4^2 + \mathrm{Neutron}_1^0$), the two hydrogen isotopes produce energetic α particles (helium) and neutrons, which heat the plasma as well as the material of the surrounding wall and the so-called the blanket which absorbs the neutrons. In a prospective reactor, the fluid which cools the blanket is to drive, outside the reactor, conventional turbines that generate electricity. The planned Next-Step device ITER is intermediate between present-day larger scale devices such as ASDEX Upgrade, DIII-D, JET, JT-60U, Tore Supra and a technico-environmentally as well as commercially viable fusion reactor. It is designed to study plasmas for which the internal heating by alpha particles exceeds the external (auxiliary) heating, and in addition to investigate a number of technological issues related to the blanket and wall material, see Lackner *et al.* [53]. To give an impression of the size: the major radius of the JET torus is 3.0 m, of ITER FEAT 6.2 m and of a future reactor some 8–10 m. In all cases, from a technological point of

view, the best 'aspect ratio', i.e. the ratio between major and minor torus radius is between 3 and 4. By influencing the plasma shape (i.e., also by deviating from an elliptical cross-section) one can modify to some extent, presumably through the plasma pedestal parameters [36], also the confinement time. For further information, the reader is referred to [37,11,47,45].

Of course, there are many aspects to be considered in designing a major Next-Step tokamak device. We describe here three statistical aspects of the physics involved. (i) The problem of plasma size and shape: the device should have a sufficiently large size (i.e., favourable ratio between volume and surface area) to reduce the energy loss through heat transport in the plasma (by conduction, convection, etc.). The thermal isolation of the plasma can be expressed by the (effective) heat diffusivity χ or, equivalently, the confinement time $\tau_E = a^2/\chi$, where a is the plasma minor radius. The conventional theory of plasma transport does not well predict the actually observed heat losses, and empirical global energy confinement scalings have been derived by applying regression analysis based on internationally assembled data sets, see e.g. [37]. A number of statistical aspects are: (1) finding an appropriate functional form of the dependence of τ_E on plasma size, plasma shape and physical plasma parameters such as the ('poloidal') magnetic field, plasma density and heating power; (2) finding a representative data set describing a particular plasma regime (such as L–mode, (quiescent) H–mode, see [90,98,14,69,31,45], etc.); (3) the propagation of measurement errors and the ensuing regression with errors in (the regression) variables; (4) the search for the influence of additional satellite variables (called 'hidden variables') on the confinement, and (5) the relation to plasma physical transport theories.

(ii) Prediction of the 'operational space' where the plasma is in a particular state, for instance in the L-mode (where the confinement time is about a factor 1.5–2.0 lower than in the H-mode), or H-mode with a particular type of ELMs (Edge-Localized Modes) that prevent too strong heat loads on the wall material. A useful statistical approach here is discriminant analysis. (In this context, of particular interest is its relation with regression analysis, see [41,42,45].) Another aspect is the need to avoid the density limit (see [37] and Borrass *et al.* [10]) as well as the avoidance (or at least the mitigation) of plasma disruptions, see Pautasso *et al.* [64]. In the latter case, neural networks have been applied as a kind of universal method to approximate multidimensional surfaces, which constitutes an alternative to (non-parametric) discriminant analysis.

(iii) A measure of plasma fusion performance is the ratio $Q = P_{fus}/P_{aux}$, where P_{fus} is the plasma fusion power and P_{aux} the 'auxiliary' heating power, externally injected into the plasma. As described in Mukhovatov *et al.* [61], it is interesting to predict Q in future fusion devices based on (a) empirical confinement scalings and (b) a semi-empirical approach, where the plasma pedestal temperature is estimated by empirical regression models (based on an international multi-tokamak pedestal database), and where the heat transport of the plasma core is calculated by theory-based plasma transport models, such as the Multi-Mode Model (MMM), and linear and non-linear versions of the Gyro Landau Fluid (GLF) model, which are implemented in numerical computer codes. In integrated transport modelling, also the

sawtooth region (in the central part of the plasma) and the ELM region (in the outer torus annulus) have to be taken into account, see e.g. [84].

A particular aspect of predicting the pedestal temperature in Next-Step devices, preferably by means of a distributional inference, is the need of incorporating in the analysis the errors in the regression variables, while simplifying a more complicated error propagation in the chain leading to the actual temperature and density measurements. Of particular interest are also methods of model selection and the analysis of causal models in such situations. For non-linear regression models, a specific statistical question is which 'distance' of the data and the regression surface is to be minimised: (a) the sum of the residuals, i.e., of the distances in the metric of the errors, between the observed data and the regression surface (which is a generalization of least squares), or (b) a softer 'penalty function' (which is a generalization of maximum likelihood), to avoid for flexible regression surface modelling overfitting of the data. The state of the art is that estimates are made and provided with standard errors suggesting normal distributions as distributional inferences.

In the specific case of confinement time prediction for ITER, in [43,61], five different characterisations ('definitions') of a 95% interval estimate were considered and the width of a corresponding confidence distribution (using terminology adopted in [44], by Schweder's suggestion) for the confinement time has been assessed to be four times (for the ITER-1998 design) or three times (for ITER FEAT) the statistical standard deviation from ordinary least squares regression based on $N_{eff} = N/4$ 'effective' observations.

Conclusion of Appendix C

If real-world problems have to be solved like the ones indicated in this appendix, then it is almost always a mixture of mathematics, science, technology, and art which guides our behaviour. Statisticians like to emphasize the *scientific* aspect that it are the data which should speak (after a scrutiny) and that statistical and systematic *uncertainties should be expressed*. Mathematical aspects should, of course, not be ignored. It is the mathematician's task to prove theorems and to establish rigorously that certain propositions are true or false. An example is the statement made in C.3 that $\mathcal{N}(\hat{y}, s^2)$ (and other continuous) distributional inferences can be improved if y is known to be in $\mathbb{N}^0$. The device mentioned there may lead to an improvement 'in general' but it does not 'always' have to do so. To obtain a universal improvement, a slightly different modification has to be made (see Kardaun and Schaafsma [46]).

Such mathematical subtleties are, of course, ignorable from the technological point of view that 'issues have to be settled' rather than 'problems have to be solved'. In this respect it is our view that the making of distributional inferences is an important tool to express opinions about unknown true values, given the numerical evidence available. This tool should, however, not be used if methodological uncertainties are too large. There are many situations where

the making of point estimates (or predictions) is too misleading to be recommended and that also their replacement by distributional inferences is not trustworthy because *another statistician, using the same observational material, may arrive at considerably different inferences.* Whether a Bayesian approach is chosen or one with the requirement of strong similarity in mind, it is always some kind of an art to arrive at a statistical inference. This aspect of being somewhat *artificial* is not ignorable if complicated issues have to be settled as indicated in C.1, ..., C.4 and, especially, in C.5. For the theoretical statistician or epistemologist, it is interesting to think about the applied statistician's unenviable task of making distributional inferences summarizing what the data have to say about crucial scientific issues.

Acknowledgements
The second author wishes to acknowledge F. Engelmann for his careful reading of Appendix C.5.

References

[1] C.J. Albers. *Distributional Inference: The Limits of Reason.* PhD Thesis, University of Groningen, 2003. Also available on internet, http://www.ub.rug.nl/eldoc/dis/science/c.j.albers/

[2] C.J. Albers and W. Schaafsma. *How to assign probabilities, if you must.* Statistica Neerlandica, **55**(3): 346–357, 2001

[3] C.J. Albers and W. Schaafsma. *Estimating a density by adapting an initial guess.* Computational Statistics and Data Analysis, **43**(1/2): 27–36, 2002

[4] C.J. Albers, B.P. Kooi and W. Schaafsma. *Trying to resolve the two-envelope problem.* Synthese, 2004 (to appear, preprint available via http://gbic.biol.rug.nl/∼calbers/)

[5] C.J. Albers and W. Schaafsma. *Optimal weakly similar procedures for solving Bayes's problem.* 2005 (to be published)

[6] D. Amaratunga and J. Cabrera. *Exploration and Analysis of DNA Microarray and Protein Array Data.* Wiley, Hobroken (NJ), 2004

[7] J.O. Berger. *Could Fisher, Jeffreys and Neyman have agreed on testing? (With discussion).* Statistical Science, **18**(1): 1–32, 2003

[8] C.R. Blyth and R.G. Staudte. *Estimating statistical hypotheses.* Statistics & Probability Letters, **23**: 45–52, 1995

[9] C.R. Blyth and R.G. Staudte. *Hypothesis estimates and acceptability profiles for* 2×2 *contingency tables.* Journal of the American Statistical Association; Theory and Methods, **92**(438): 694–699, 1997

[10] K. Borrass, A. Loarte, C.F. Maggi, V. Mertens, P. Monier, R. Monk, J. Ongena, J. Rapp, G. Saibene, R. Saartori, J. Schweinzer, J. Stober, W. Suttrop and EFDA-JET Workprogramme collaborators. *Recent H-mode density limit studies at JET.* Nuclear Fusion, **44**: 752–760, 2004

[11] C.M. Braams and P. Stott. *Nuclear Fusion: Half A Century Of Magnetic Confinement Fusion Research*. Institute of Physics Publishing, 2002

[12] R. de Bruin, D. Salomé, W. Schaafsma. *A semi-Bayesian method for nonparametric density estimation*. Computational Statistics & Data Analysis, **30**(1): 19–30, 1999

[13] J.P. Chilés and P. Delfiner. *Geostatistics: Modelling Spatial Uncertainty*. Wiley, Chichester, 1999

[14] J.P. Christiansen, J. DeBoo, O.J.W.F. Kardaun, S.M. Kaye, Y. Miura, *et al. Global Energy Confinement Database for ITER (special topic)*. Nuclear Fusion, **32**: 291–338, 1992

[15] D.R. Cox and D. Oakes. *Analysis Of Survival Data*. Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1984

[16] R.T. Cox. *The Algebra of Probable Inference.* John Hopkins University Press, Baltimore, 1961

[17] N.A.C. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1991

[18] H. Cui, A. Stein, D.E. Myers. *Extension of spatial information, Bayesian kriging and updating of prior variogram parameters.* Environmetrics 6: 373–384, 1995

[19] E. Culotta. *Pulling Neanderthals back into our family tree*. Science **252**(5004): 376

[20] R. Davidson and J.G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1993

[21] H.G. Dehling, T.K. Dijkstra, H.J. Guichelaar, W. Schaafsma, A.G.M. Steerneman, T.J. Wansbeek and J.T. van der Zee. *Structuring the inferential contest*. In: D.A. Berry, K.M. Chaloner and J.K. Geweke (eds). *Bayesian Analysis in Statistics and Econometrics*, Chapter 46: 539–547, John Wiley & Sons, New York, 1996

[22] P.A. Diggle, J.A. Tawn and R.A. Moyeed. *Model based geostatistics* (*with discussion*). Journal of the Royal Statistical Society, Series C, **47**: 299–350, 1998

[23] G.G. Dolgov-Saveliev, V.S. Mukhovatov, V.S. Strelkov, M.N Shepelev and N.A. Yavlinksi. *Investigation of a toroidal discharge in a strong magnetic field.* In: N.R. Nilson (ed.). *Proceedings of the Fourth International Conference on Ionization Phenomena in Gases, August 17–21, 1959*, volume II, Part IV, 947–953. North-Holland, Amsterdam, 1960

[24] N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 1966

[25] S. Dudoit, M.J. van der Laan, S. Keleş, A.M. Molinaro, S.E. Sinisi and S.L. Teng. *Loss-based estimation with cross-validation: applications to microarray data analysis and motif finding.* ACM SIGKDD Explorations, **5**(2): 37–49, 2003

[26] B. Efron. *Bootstrap methods: another look at the jackknife*. Annals of Statistics, **7**(1): 1–26, 1979

[27] M. Engelmoer and C.S. Roselaar. *Geographical Variation in Waders.* Kluwer, Dordrecht, 1998

[28] L.T. Fernholz and S. Morgenthaler. *Remembering J.W. Tukey*. Statistical Science, **18**: 346–356, 2003

[29] R.A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930

[30] C. Genest and J.V. Zidek. *Combining probability distributions: a critique and an annotated bibliography*. Statistical Science, **1**(1): 114–148, 1986

[31] L. Giannone, A.C.C. Sips, O.J.W.F. Kardaun, F. Spreitler, W Suttrop. *Regime identification in ASDEX Upgrade*. Plasma Physics and Controlled Fusion, **46**: 835–856, 2004

[32] P.D. Grünwald and A. Philip David. *Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory*. Annals of Statistics, **32**(4): 1367–1433, 2004

[33] A. Hald. *A History Of Mathematical Statistics From 1750 To 1930*. Wiley, New York, 1998

[34] D.V. Hinkley. *Predictive likelihood*. Annals of Statistics, **7**(4): 718–728, 1979

[35] J.T. Huang, G. Casella, C. Robert, M.T. Wells, and R.H. Farrell. *Estimation of accuracy in testing*. Annals of Statistics, **20**: 409–509, 1992

[36] L.D. Horton *et al. Dependence of H–mode pedestal parameters on plasma magnetic geometry*. Plasma Physics and Controlled Fusion, **44**: A273–A278, 2002

[37] ITER Physics Basis Document. Nuclear Fusion, **39**: 2173–2664, 1999

[38] E.T. Jaynes, and G.L. Bretthorst (ed.). *Probability Theory: The Logic Of Science*. Cambridge University Press, 2003

[39] N.L. Johnson and J.O. Kitchen. *Some notes on tables to facilitate fitting S_B curves*. Biomerika, **58**: 223–227 and 657–668, 1971

[40] O.J.W.F. Kardaun. *On Statistical Survival Analysis*. PhD Thesis, University of Groningen, 1986

[41] O.J.W.F. Kardaun, J.W.P.F. Kardaun, S.-I. Itoh, and K. Itoh. *Discriminant analysis of plasma fusion data*. In: Y. Dodge, J. Whittaker (eds.), Computational Statistics X, Volume 1, Physica-Verlag, Heidelberg, 163–170, 1993

[42] O.J.W.F. Kardaun, *et al. Generalising regression and discriminant analysis: catastrophe models for plasma confinement and threshold data*. In: A. Prat (ed.), Computational Statistics XII, Volume 1, Physica-Verlag, Heidelberg, 313–318, 1996

[43] O.J.W.F. Kardaun. *Interval estimation of global H-mode energy confinement in ITER*. Plasma Physics and Controlled Fusion, **41**: 429–469, 1999

[44] O.J.W.F. Kardaun, D. Salomé, W. Schaafsma, A.G.M. Steerneman, J.C. Willems, and D.R. Cox. *Reflections on fourteen cryptic issues concerning the nature of statistical inference*. International Statistical Review, **71**(2): 277–318, 2003. Chinese translation by Yu Zhu in: Statistics and Information Tribune, (5,6), 2004 and (1), 2005

[45] O.J.W.F. Kardaun. *Classical Methods of Statistics with Applications in fusion-oriented Plasma Physics*. Springer–Verlag, Heidelberg, 2005 (forthcoming)

[46] O.J.W.F. Kardaun and W. Schaafsma. *Distributional Inference, Towards a Bayes-Fisher-Neyman Compromise*. 2003 (preprint, available on request)

[47] M. Kaufmann. *Plasmaphysik und Fusionsforschung: Eine Einführung.* Teubner–Verlag, Wiesbaden, 2003

[48] M.K. Kerr and G.A. Churchill. *Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments.* Proceedings of the National Academy of Sciences of the USA, **98**: 8961-8965, 2001

[49] R.H. Ketellapper, B. Bos, M.A. Kooyman and W. Voorhoeve. *A simultaneous econometric model for the Dutch economy.* Statistica Neerlandica, **31**: 141–159, 1977

[50] J. Kiefer, H.P. Wynn. *Optimum and minimax exact treatment designs for one-dimensional autoregressive error processes.* Annals of Statistics, **12**(2): 431–450, 1984

[51] A.N. Kolmogorov. *Grundbegriffe der Warscheinlichkeitsrechnung.* Springer–Verlag, Heidelberg, 1933

[52] S. Kotz and N.L. Johnson (eds.). *Breakthroughs in Statistics. Volume II: Methodology and Distribution.* Springer Series in Statistics: Perspectives in Statistics. Springer-Verlag, New York, 1992

[53] K. Lackner, R. Andreani, D. Campbell, M. Gasparotto, D. Maisonnier, and M.A. Pick. *Long-term fusion strategy in Europe.* Journal of Nuclear Materials, **307–311**:10–20, 2002

[54] E.L. Lehmann. *Testing Statistical Hypotheses* (second reprinted edition). Springer-Verlag, New York, 1997

[55] G. Matheron. *The intrinsic random functions and their applications.* Advances in Applied Probability, **5**: 439–468, 1973

[56] G. Matheron. *A simple substitute for conditional expectation: the disjunctive kriging.* In: M. Guarascio and M. Dacid and C. Huijbrechts (eds.), *Advanced geostatistics in the mining industry*, Reidel, Dordrecht, 221–236, 1976

[57] E.A. van der Meulen. *Assessing Weights of Evidence for Discussing Classical Statistical Hypotheses.* PhD.-thesis, University of Groningen, 1992

[58] E.A. van der Meulen and W. Schaafsma. *Assessing weights of evidence for discussing classical statistical hypotheses.* Statistics & Decisions, **11**(3): 201–220, 1993

[59] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour.* Princeton University Press, 1944

[60] V.S. Mukhovatov. Итоги науки и техники, сериа физика Плазмы, Chapter Токамаки. Академиа наук СССР, Москва, Book 1, Part 1, 1980

[61] V.S. Mukhovatov, *et al. Comparison of ITER Performance Predicted by Semi-Empirical and Theory-Based Transport Models.* Nuclear Fusion, **43**: 942–948, 2003

[62] D. Mumford, *The dawning of the age of stochasticity.* In: V. Arnold, M. Atiyah, P. Lax and B. Mazur (eds.), *Mathematical frontiers & perspectives*, American Mathematical Society, 1999

[63] J.R. Neyman and K. Pearson. *On the use and interpretation of certain test criteria for purposes of statistical inference.* Biometrika **A20**:175–240/264–294, 1928

[64] G. Pautasso, C. Tichmann, S. Egorov, T. Zehetbauer, *et al. On-line prediction and mitigation of disruptions in ASDEX Upgrade.* Nuclear Fusion, **42**: 100-108, 2002

[65] K. Pearson. *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, 1900

[66] K. Pearson. *The fundamental problem of practical statistics.* Biometrika, **13**: 1–16, 1920

[67] H. Reichenbach. *Elements of Symbolic Logic.* Macmillan, London, 1948

[68] N. Reid. *2000 Wald Memorial Lecture: Asymptotics and the Theory of Inference.* Annals of Statistics, **31**(6): 1695–1731, 2003

[69] F. Ryter for the Threshold Database Working Group, *H–mode Power Threshold Database for ITER (special topic).* Nuclear Fusion, **36**: 1217–1264, 1996

[70] D. Salomé and R. de Bruin and W. Schaafsma. *Q-values for χ^2 problems.* Statistics & Decisions, **17**(3): 285–291, 1999

[71] D. Salomé. *Statistical Inference via Fiducial Methods.* PhD Thesis, University of Groningen, 1998. Also available on internet, http://www.ub.rug.nl/eldoc/dis/science/d.salome/

[72] L.J. Savage. *The Theory of Statistical Decision.* Journal of the American Statistical Association, **46**: 55-67, 1951

[73] W. Schaafsma. *Hypothesis Testing Problems with the Alternative Restricted by a Number of Parameters.*, PhD Thesis, University of Groningen, 1966. Also: Noordhoff, Groningen, 1966

[74] W. Schaafsma. *Minimax risk and unbiasedness for multiple decision problems of type I.* Annals of Statistics, **5**: 1684–1720, 1969

[75] W. Schaafsma. *Discussing the truth or falsity of a statistical hypothesis H and its negation A.* International Workshop on Theory and Practice in Data Analysis (Berlin, 1988), Akademie der Wissenschaften der DDR, 150–166, 1989

[76] W. Schaafsma, J. Tolboom and E.A. van der Meulen. *Discussing truth or falsity by computing a Q-value.* In: Y. Dodge (ed.). *Statistical data analysis and inference*, North-Holland, Amsterdam, 85–100, 1989

[77] P. Sebastiani, E. Gussoni, I.S. Kohane and M.F. Ramoni. *Statistical challenges in functional genomics.* Statistical Science, **18**(1): 33-59, 2003

[78] T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* PhD thesis, University of Groningen, 1979

[79] T. Speed (ed.). *Statistical Analysis of Gene Expression Microarray Data.* Interdiciplinary Statistics Series, Chapman and Hall, New York, 2003

[80] C. Stein. *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution.* Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Volume I, . University of California Press, 197–206, 1956

[81] A. Stein, L.C.A. Corsten. *Universal kriging and cokriging as a regression procedure.* Biometrics 47: 575–588, 1991

[82] S.M. Stigler. *The History of Statistics. The Measurement of Uncertainty Before 1900.* Harvard University Press, Cambridge, 1986

[83] J.D. Storey. *The positive false discovery rate: a Bayesian interpretation and the q-value.* Annals of Statistics, **31**(6): 2013–2035, 2003

[84] W. Suttrop, O. Gehre, J.C. Fuchs, H. Reimerdes, W. Schneider, J. Schweinzer and the ASDEX Upgrade Team. *Effects of type-I Edge-Localized Modes on Transport in ASDEX Upgrade.* Plasma Physics and Controlled Fusion, **40**: 771–774, 1998

[85] W.R. Thompson. *On confidence ranges for the median and other expectation distributions for populations of unknown distribution form.* Annals of Mathematical Statistics, **7**: 122–128, 1936

[86] V. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998

[87] G.N. van Vark and A. Bilsborough. *Shaking the family tree.* Science, **253**(5022): 834–834, 1991

[88] G.N. van Vark and A. Bilsborough. *Human cranial variability, past and present.* American Journal of Physical Anthropology, **95**(1): 89–91, 1994

[89] J. Venn. *The Logic of Chance: an Essay on the Foundations and Province of the Theory of Probability*, 3rd edition. McMillan, London and New York, 1888

[90] F. Wagner for the ASDEX Team. *Regime of Improved Confinement and High Beta in Neutral-Beam-Heated Divertor Discharges of the ASDEX Tokamak.* Physical Review Letter, **49**: 1408–1412, 1982

[91] A. Wald. *Tests of statistical hypotheses concerning several parameters when the number of observations is large.* Transactions of the American Mathematical Society, **54**: 426–482, 1943

[92] A. Wald. *On a statistical problem arising in the classification of an individual into one or two groups.* Annals of Mathematical Statistics, **15**: 145–162, 1944

[93] A. Wald. *Book review of Von Neumann-Morgenstern [59].* The Review of Economic Statistics, **24**(1): 47–52, 1947

[94] A. Wald. *Statistical Decision Functions.* Wiley, New York, 1950

[95] S.S. Wilks. *The large-sample distribution of the likelihood ratio for testing composite hypotheses.* Annals of Mathematical Statistics, **9**: 60–62, 1938

[96] M.H. Wolpoff. *Levantines and Londoners.* Science, **255**(5041): 142, 1992

[97] R.L. Wolpert. *A conversation with James O. Berger.* Statistical Science, **19**(1): 205–218, 2004

[98] P.N. Yushmanov, T. Takizuka, K. Riedel, O.J.W.F. Kardaun, J.G. Cordey, S.M. Kaye and D.E. Post. *Scalings for Tokamak Energy Confinement.* Nuclear Fusion, **30**: 1999–2008, 1990