Distributional inference

A. H. Kroese

*Department of Mathematics, University of Groningen, P.O. Box 800,
9700 AV Groningen, The Netherlands*

E. A. van der Meulen

*Department of Medical Statistics, Erasmus University Rotterdam,
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands*

K. Poortema

*Faculty of Applied Mathematics, University of Twente, P.O. Box 217,
7500 AE Enschede, The Netherlands*

W. Schaafsma

*Department of Mathematics, University of Groningen, P.O. Box 800,
9700 AV Groningen, The Netherlands*

The making of statistical inferences in distributional form is conceptually complicated because the epistemic 'probabilities' assigned are mixtures of fact and fiction. In this respect they are essentially different from 'physical' or 'frequency-theoretic' probabilities. The distributional form is so attractive and useful, however, that it should be pursued. Our approach is in line with Wald's theory of statistical decision functions and with Lehmann's books about hypothesis testing and point estimation: loss functions are defined, risk functions are studied, unbiasedness and equivariance restrictions are made, etc. A central theme is that the loss function should be 'proper'. This fundamental concept has been explored by meteorologists, psychometrists, Bayesian statisticians, and others. The paper should be regarded as an attempt to reconcile various schools of statisticians. By accepting what we regard as good and useful in the various approaches we are trying to develop a nondogmatic approach.

Key Words & Phrases: decision theory, fiducial inference, foundations, invariance, unbiasedness.

1 Introduction

The making of statistical inferences is an uncertain affair. That is why different schools have emerged and the statistical controversy has appeared. We worry about this controversy, because the lack of consensus revealed by it impedes the respectability of our science. It is true that the debate between the various schools is less heated than it used to be. The statistical community has started to realize that every school has something useful to say and that the making of applications requires an eclectic attitude. Some situations are dealt with most appropriately by using a data-analytic

approach in which statistical modelling is kept at a minimum. Other situations require a Neyman–Pearson test or a p -value. Sometimes a prior distribution can be defined and a Bayesian approach is indicated. It is easy to agree with the following statement in (EFRON, 1970):

Both the objective and the subjective schools have attempted to broaden the philosophical basis of their respective models. These attempts at greater applicability range from the ingenuous (in his lifetime the statistician would err 5% of the time) to the heroic (subjective probability). Nevertheless it is a fact of life that in many practical situations neither frame is satisfactory. In such situations the difficulty in choosing an appropriate loss function is dwarfed by the fundamental problem of choosing an applicable framework in which to view the experiment and its results.

Yet, in the present paper, we focus on the problem how, *given* a probabilistic model that describes the experiment, inferences should be made. The reason is that this, and not the problem how to select an appropriate model, has received most attention in papers about the statistical controversy.

Suppose that the data x have to be used to discuss the unknown y of interest. We assume that (x, y) is the outcome of (X, Y) , where X and Y are random variables taking values in spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. The joint distribution P is assumed to belong to some parametrized family $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$. Inference about $g(\theta)$, where g is some deterministic function, is a special case of this formulation (identify y and $g(\theta)$ and let Y have the distribution which assigns point mass 1 to y).

The form of the inference is crucial. We focus on the form of a probability distribution on $\mathcal{Y}$. A well-known way to construct such distributional inferences is via a Bayesian analysis. If a prior distribution is chosen for θ , the joint distribution of θ , X , and Y can be constructed. Next, by the laws of probability, the (marginal) conditional distribution of Y given $X = x$ is obtained. This is the Bayesian solution w.r.t. the prior τ . ‘Fiducial inference’ and ‘structural inference’ are also well-known. An introduction to fiducial inference can be found in FISHER (1959) and BUEHLER (1980). Its controversial history is described in SEIDENFELD (1992) and ZABELL (1992). The structural approach to distributional inference can be found in FRASER (1968). The distributional form has been rejected as too ambitious and controversial by the ‘frequentists’ after they had seen that Fisher made some misinterpretations, for example with respect to the Behrens-Fisher problem. It is not difficult, however, to realize that alternative forms of inference, e.g. estimates equipped with standard errors, are used in practice as if they are distributional inferences. The distributional form is attractive as it provides an intrinsic way to express the uncertainties the statistician has about y .

Why are distributional inferences so controversial? The main reason is that the ‘probability’ distributions assigned are distributions of ‘epistemic’ probabilities which describe our ‘knowledge’ about the unknown of interest (epistemology = theory of

knowledge). The distributional inference summarizes the information one has about y with no particular application in mind. These epistemic probabilities should not be mixed up with 'physical' or 'frequency-theoretic' probabilities. Epistemic probabilities are mixtures of fact and fiction. We use the word 'fact' to denote those real things whose existence and definition one might agree about: the data x , the definition of the unknown y about which an inference has to be made, the true, usually unknown, distribution which determines the physical probabilities. The word 'fiction' is used to denote the additional inputs the statistician needs to perform his tasks. These inputs are not necessarily purely fictive. There may be a factual core. Examples are the 'principles', 'loss functions', 'prior distributions', etc. which the statistician exploits.

Of course, if a distributional inference is given about y , other probability statements can be derived by the calculus of probability. An induced distributional inference can immediately be obtained about $f(y)$ where f is some mapping. One of the major paradigms in Bayesian statistics is that inferential probability statements should 'cohere' in the sense that for all f such an induced probability distribution is 'inferentially valid'. Lindley writes 'My thesis is that sampling-theorists have failed to consider coherence and, in consequence, have produced unsatisfactory results', see LINDLEY (1990). In our theory it may happen that the inferences obtained for y and $f(y)$ do not cohere in the sense indicated. As distributional inferences and the 'probabilities' they define are mixtures of fact and fiction there is no compelling reason why these probabilities should cohere in the same way physical probabilities or relative frequencies do. In fact there is a very good reason why the epistemic probabilities should *not* be subjected to such restrictive framework. This is explained in Section 8 and has the consequence that our approach to distributional inference is less 'dogmatic' than most other approaches. An interesting paper, expressing a similar point of view as this paragraph, is WILKINSON (1977).

The purpose of this paper is to give an introduction to the loss function approach to distributional inference. Hence we will focus on procedures for making distributional inferences and we will compare these on the basis of their risk functions. Of course the choice of a loss function is a matter of concern. The loss function approach to distributional inference is not new. For example, in EATON (1982) the problem of evaluating one method of distributional inferences (i.e. using improper prior distributions) is treated using a loss function formulation. Philosophy, problem formulation and proper loss functions (see the next section) are discussed in the introduction to that paper.

2 Properness

Our theory of distributional inferences is in line with Wald's theory of statistical decision functions. The set of all probability measures on $\mathcal{Y}$ is denoted by $\mathcal{Y}^*$. A procedure for making distributional inferences is a function $Q: \mathcal{X} \rightarrow \mathcal{Y}^*$. A loss function $L: \mathcal{Y} \times \mathcal{Y}^* \rightarrow \mathbb{R}$ is needed to express the loss incurred if the inference $v \in \mathcal{Y}^*$

is made whereas the true value is equal to y . It is natural to choose the loss function such that $L(y, v) \geq 0$ with equality if and only if $v = \varepsilon_y$. (Dirac measure ε_y assigns point mass 1 to y .) There is yet another extremely important property which L should satisfy. That is the property of *properness*.

DEFINITION 1. *The loss function L is said to be proper (or Bayes-fair) if $\mathcal{L}Y = v_1 \in \mathcal{Y}^*$ implies*

$$EL(Y, v_2) \geq EL(Y, v_1) \quad \text{for all } v_2 \in \mathcal{Y}^*$$

The underlying idea is obvious: if one would know that y is from distribution v_1 then it is this distribution which should appear if the expected loss is minimized. In other words, the evaluation of the distributional inference should be such that the statistician produces the true distribution if this is known. There is an abundance of proper loss functions. An example for $\mathcal{Y} = \mathcal{R}$ is

$$L(y, v) = \int_{\mathcal{R}} (1_{[y, \infty)}(z) - G(z))^2 dz \quad (1)$$

where v is a probability measure on $\mathbb{R}$ and $G(\cdot)$ the corresponding c.d.f. Note that $1_{[y, \infty)}(z)$ corresponds with the c.d.f. of ε_y , and hence $L(y, v)$ is the squared L_2 -distance between the c.d.f.'s of v and ε_y . Properness of this loss function follows from the observation that

$$\begin{aligned} EL(Y, v_2) &= \int G_2^2(z)(1 - G_1(z)) + (1 - G_2(z))^2 G_1(z) dz \\ &= \int (G_1(z) - G_2(z))^2 dz + \int G_1(z)(1 - G_1(z)) dz \end{aligned}$$

Here $Y \sim v_1$ and $G_i(\cdot)$ is the c.d.f. of v_i , $i = 1, 2$. It can be shown that loss function (1) is a so-called 'quadratic loss function': under the condition that the first moment of v is finite, $L(y, v)$ is equal to $-1/2 \iint |u - v|(v - \varepsilon_y)(du)(v - \varepsilon_y)(dv)$. The proof of this assertion can be found in KROESE (1994). The class of quadratic loss functions was introduced for finite $\mathcal{Y}$ in STAËL VON HOLSTEIN et al. (1978) and for arbitrary $\mathcal{Y}$ in EATON (1982). We think that loss function (1) is attractive, also because $L(y, \varepsilon_z) = |y - z|$ for all y and z . More examples of proper loss functions can be found in HENDRICKSON et al. (1971), SAVAGE (1971), SCHERVISH (1989), etc.

The procedures we will construct will be chosen (partly) on the basis of their risk (= expected loss) functions. Such approaches are sometimes called decision-theoretic. We do not use this formulation because we insist upon a clear separation between statistical inference and decision making. The statistician provides a distributional inference about y based on the data x . This distributional inference describes the uncertainty about y after having observed $X = x$. The statistician carries the main responsibility for making this inference and should be evaluated by a proper loss function. Next it will usually be other people who worry about subsequent discussions and decisions.

The risk function R of a procedure Q is defined as $R(\theta, Q) = E_{\theta} L(Y, Q(X))$. In general no procedure exists of which the risk is uniformly smaller than that of any other procedure. One way to solve this problem is to require the risk to be small in some overall sense. For example the minimax rule can be chosen or the Bayes rule w.r.t. some probability measure τ . It is a simple fact, see e.g. BERNARDO (1979), that properness of the loss implies that the Bayes risk $r(\tau, Q) = \int R(\theta, Q) d\tau(\theta)$ is minimum if Q corresponds, at least formally, with the Bayesian solution (as described in the previous section) w.r.t. the prior τ . Hence, using a proper loss function has the consequence that if a prior distribution is available, based on solid information, the Bayesian answer is obtained. If such a prior does not exist, considering the integrated risk is less appealing and so is the Bayesian approach. Another possibility is to restrict the class of procedures by some requirement. Classical examples are the reductions by unbiasedness and equivariance. These will be considered in the following sections.

3 Inference about a success probability

Suppose that s successes have been observed in a sequence of n independent Bernoulli experiments, all with the same unknown success probability p . So s is the outcome of $S \sim B(n, p)$. The problem to make a distributional inference about p in this situation has received considerable attention in the literature. In his famous posthumous essay, Thomas Bayes used an argument that resembles the concept of using the $U[0, 1]$ distribution as prior such that the posterior Beta $(s + 1, n - s + 1)$ is obtained as distributional inference. Other authors who obtained the same solution are Laplace, Pearson, and many others. Some Bayesians advocated the Beta $(\frac{1}{2}, \frac{1}{2})$ prior distribution. For a summary of the Bayesian approach to the binomial problem, see GEISSER (1984). Fisher had problems in extending his fiducial argument to discrete distributions and only provided an asymptotically valid result, see FISHER (1959). The same remark applies to Fraser's structural inference, see FRASER (1968).

Although we do not criticize the solutions above, we like to present an alternative approach. In practice the differences will be small. Any procedure $Q: \{0, \dots, n\} \rightarrow [0, 1]^*$ for making distributional inferences about p is characterized by the distribution functions $G_s(\cdot)$, $s = 0, \dots, n$. An appealing property to require of such a procedure is that 'on the average' equal mass is put to the left and the right of the true value. The operationalization of this definition requires that one considers the random variable $X \sim B(n, \theta)$ for arbitrary $\theta \in \Theta = (0, 1)$. The notation emphasizes the difference between the random variable S actually realized and the theoretical construct X .

DEFINITION 2. *The procedure Q is said to be unbiased if $E_{\theta} G_X(\theta) = 1/2$ holds for all $\theta \in (0, 1)$.*

In LEHMANN (1951) a concept of unbiasedness was introduced that depends on the loss function of interest. In a decision theoretic setting with loss function L , a

procedure d is said to be unbiased if $E_\theta L(\theta, d(X)) \leq E_\theta L(\theta_1, d(X))$ for all $\theta, \theta_1 \in \Theta$. It can be shown that, except for some complications if θ is 0 or 1, Definition 2 corresponds with Lehmann's concept of unbiasedness if loss function (1) is considered. The proof is not too interesting and is skipped here. The Bayes–Laplace procedure $Q^{(BL)}$ defined by $Q^{(BL)}(s) = \text{Beta}(s+1, n-s+1)$ with corresponding c.d.f.'s $G_i^{(BL)}(\cdot)$, is not unbiased as it satisfies

$$E_\theta G_X^{(BL)}(\theta) = \frac{1}{2} + (\theta - \frac{1}{2})P(X = X') \quad (2)$$

where X and X' are independent and $B(n, \theta)$ distributed. Formula (2) can be proved using the fact that the order statistics of independent uniformly distributed r.v.'s on $[0, 1]$ are Beta distributed.

We use some modification of Fisher's fiducial argument to generate a particular unbiased procedure. The idea is that $\alpha_z(s) = G_z(z)$ expresses the 'weight of the evidence' carried by the observation s in favour of $H_z: p \leq z$, the complementary weight being assigned to the alternative hypothesis $A_z: p > z$. One usually uses the p -value $P_z(X \geq s)$ or $P_z(X > s)$ as the weight in favour of H_z . Taking, for $z \in [0, 1]$, the arithmetic average of these two, we obtain

$$G_z^{(0)}(z) = P_z(X > s) + \frac{1}{2}P_z(X = s) \quad (3)$$

The corresponding procedure $Q^{(0)}$ is given by

$$Q^{(0)}(s) = \frac{1}{2}\text{Beta}(s, n-s+1) + \frac{1}{2}\text{Beta}(s+1, n-s) \quad (4)$$

where, by definition, $\text{Beta}(0, g) = \varepsilon_0$ and $\text{Beta}(f, 0) = \varepsilon_1$ for all $f, g > 0$. In the derivation of (4) a similar argument is used as the verification of (2). The rule $Q^{(0)}$ has an interesting optimality property: consider as loss function the following obvious modification of (1).

$$L(p, v) = \int_0^1 (1_{[p, \infty)}(z) - G(z))^2 dz \quad (5)$$

THEOREM 1. *The procedure $Q^{(0)}$ is unbiased. It is Bayes-best unbiased in the sense that it minimizes the Bayes risk with respect to the uniform prior among all unbiased procedures, for loss function (5).*

PROOF: See Appendix A.

Of course the Bayes risk of $Q^{(0)}$ is larger than that of $Q^{(BL)}$, the risk of $Q^{(BL)}$ not being uniformly smaller than that of $Q^{(0)}$. In KROESE (1994) it is proved that $Q^{(0)}$ is inadmissible and simple procedures are described explicitly that dominate $Q^{(0)}$ (for sample sizes 1 and 2). The existence of such procedures is no compelling reason for rejecting $Q^{(0)}$. These other rules are not unbiased. The unbiasedness results in some loss of performance. This is natural if unbiasedness (or equivariance) restrictions are made.

REMARK. There are other definitions of unbiasedness that can be considered. If a distributional inference $Q(s)$, with distribution function $G_s(\cdot)$, has to be used to make an estimate of p , then it is natural to consider the expectation of $Q(s)$ as the estimate of p . The procedure Q is said to be expectation-unbiased if the so defined point estimator is unbiased, i.e.

$$\mathbf{E}_\theta \left(\int_0^1 z dG_X(z) \right) = \theta \quad \forall \theta \in [0, 1]$$

Note that neither $Q^{(BL)}$ nor $Q^{(0)}$ is expectation-unbiased. The procedure Q defined by $Q(s) = \text{Beta}(s, n - s)$ (with obvious interpretation of the degeneracies), however, is. Also for the concept of expectation-unbiasedness results can be derived like that of Theorem 1. We refer to KROESE (1994).

The concept of expectation-unbiasedness has a drawback in comparison with the concept of unbiasedness discussed earlier. The drawback is that it depends on the parametrization. If $\log\{\theta/(1 - \theta)\}$ is used instead of θ then no expectation-unbiased procedure Q exists as this function of θ does not allow an unbiased estimate. The concept of expectation-unbiasedness is of considerable intuitive interest, however, especially in those, possibly complicated, situations where one has already constructed a satisfactory unbiased estimator and wants to express the involved certainty by replacing the estimator by some procedure for making distributional inferences. It is then natural to require that this procedure is expectation-unbiased.

4 Inference about a location parameter

Another problem that has received considerable attention in the literature is the 'measurement problem', see for example FRASER (1968). Stripped to its essentials, the problem is as follows. Let the error $e = x - \mu$ in the measurement x of some physical constant $\mu \in \mathbb{R}$ be regarded as the outcome of a r.v. E having some known continuous error distribution with c.d.f. $F(\cdot)$. Hence x is regarded as the outcome of a random variable X with distribution function $F_\mu(\cdot)$, where $F_\mu(z) = F(z - \mu)$. If this model is accepted, the random variable X and its probability distribution $\mathcal{L}X$ have a clear physical and frequency-theoretic interpretation in terms of repeated measurements of μ .

Suppose a distributional inference is required about μ on the basis of the data x . We assume nothing is known in advance about μ . By writing $\mu = x - e$ it seems natural to take $Q^{(1)}(x) = \mathcal{L}(x - E) = \mathcal{L}(x + \mu - X)$ as distributional inference about μ . 'As nothing is known about μ prior to the experiment, the probability distribution of E does not change by the observation of $X = x$ '. Of course $Q^{(1)}(x)$ has no interpretation in terms of repeated experiments because μ is a deterministic constant.

It is easy to see that the existence of a density $F' = f$ implies that $Q^{(1)}(x)$ has the likelihood function $l_x(\mu) = f(x - \mu)$ as its density. So $Q^{(1)}$ corresponds with using Lebesgue measure as formal prior distribution for μ . $Q^{(1)}$ also corresponds with the fiducial argument: let $G_x^{(1)}(\cdot)$ be the c.d.f. corresponding to $Q^{(1)}(x)$. The continuity of

F suffices to show that $G_x^{(1)}(\mu)$ is identical to the p -value $\alpha_\mu(x) = P_\mu(X \geq x)$ (the fiducial argument, compare the previous section).

The so defined procedure $Q^{(1)}$ is also supported by our loss function approach: $Q^{(1)}$ is MRE (Minimum Risk Equivariant) w.r.t. translations if the loss function is proper and invariant. A procedure is (translation) equivariant if a shift in the data leads to a corresponding shift of the distributional inference. The invariance of the loss function means that the loss does not change if the true value y and the distributional inference are shifted over the same distance. The precise choice of the loss function is not important in the present context because $Q^{(1)}$ is MRE for *every* proper and invariant loss function (provided its risk is finite). An abundance of such loss functions exists, an example is (1).

The proof that $Q^{(1)}$ is MRE is as follows: the *equivariance* of $Q: \mathbb{R} \rightarrow \mathbb{R}^*$ implies that $Q(x)$ is obtained from $Q(0)$ by shifting $Q(0)$ over a distance x . The *invariance* of L implies that the risk of an equivariant procedure can be written as

$$E_\mu L(\mu, Q(X)) = EL(\mu - X, Q(0))$$

The *properness* of L implies that the r.h.s. is minimum if $Q(0) = \mathcal{L}(\mu - X)$. Hence $Q(x) = \mathcal{L}(x + \mu - X) = Q^{(1)}(x)$. In EATON (1992) the admissibility of $Q^{(1)}$ is established (under some conditions) for a special class of loss functions.

5 An example of predictive inference

It is often argued that prediction should be central in statistics, because predictive statements are verifiable. In this respect it is interesting to consider the following extension of Section 4. Suppose the measurement x has to be used to make a distributional inference about some future measurement y . Here x and y are regarded as outcomes of independent random variables X and Y , having distribution functions $F_1(\cdot - \mu)$ and $F_2(\cdot - \mu)$, respectively. F_1 and F_2 are regarded as known whereas $\mu \in \mathbb{R}$ is unknown. A review of approaches is as follows:

(I) As the distribution function $F_2(\cdot - \mu)$ of Y has a solid physical and frequency-theoretic interpretation, it is natural to require that the distributional inference $Q(x)$ about y has *the same form*. This implies that the corresponding distribution function $G_x(\cdot)$ should satisfy $G_x(z) = F_2(z - \hat{\mu}(x))$ where $\hat{\mu}(x)$ is some estimate of μ . This approach is often followed in practice. It is misleading, though, as $Q(x)$ should be more dispersed than the distribution corresponding with $F_2(\cdot)$. The statistical uncertainties, arising from the fact that μ is unknown, should be 'added' to the physical uncertainties about the future measurement. If μ is replaced by an estimate and the resulting distribution is used as distributional inference about y the uncertainties are underestimated.

(II) As the distribution function $G_x(\cdot)$ is some kind of *estimate* of $F_2(\cdot - \mu)$, the classical *unbiasedness* requirement $E_\mu G_x(z) = F_2(z - \mu)$, for all z and μ , might be made. If a random variable Z can be constructed such that (a) Z and

X are independent, for all μ , and (b) $\mathcal{L}_\mu(X + Z) = \mathcal{L}_\mu Y$, then $G_x(z) = P(Z \leq z - x)$ defines the estimate required, because $E_\mu G_x(z) = E_\mu P((Z \leq z - X)|X) = P_\mu(Z + X \leq z) = P_\mu(Y \leq z) = F_2(z - \mu)$. This approach to distributional inference fails if Y is not sufficiently more dispersed than X and consequently such a random variable Z does not exist. If $F_1 = F_2$ then the unbiased estimator of the distribution function of Y corresponds to the Dirac measure ε_x with mass 1 at the observation X . The statistical uncertainties involved in the prediction of y are suppressed rather than expressed if this approach is chosen.

(III) In Section 4 we saw that many approaches lead to $\mathcal{L}(x + \mu - X)$ as distributional inference about μ on the basis of the observation x . A predictive distribution for y can, of course, be obtained by integrating out the conditional distribution of Y given μ . This will lead to the same inference as the following argument: it is natural to write $y = x + (\mu - x) + (y - \mu)$ and to suggest the distributional inference $Q^{(1)}(x) = \mathcal{L}(x + (\mu - X) + (Y - \mu)) = \mathcal{L}(x + (Y - X))$. Note that the distribution of $Y - X$ does not depend on μ .

(IV) The solution obtained under (III) can also be obtained directly in our loss function approach. Assuming in Section 4 that the loss function is proper and invariant under translations, $Q^{(1)}$ is MRE, provided that the risk function is finite. Note that the risk

$$EL(Y, Q(X)) = EL(Y - X, Q(0))$$

is minimum if $Q(0) = \mathcal{L}(Y - X)$ and, hence, if $Q(x) = \mathcal{L}(x + Y - X)$. Under some conditions, the procedure $Q^{(1)}$ is admissible for a special class of loss functions as is demonstrated in EATON (1993).

6 General MRE procedures

There are no difficulties in generalizing most of the results of the preceding two sections to multidimensional translation problems. In fact, the MRE results can be generalized to (parametric) models, invariant under groups other than the translation group. If there exists a 'sufficiently large' group that leaves the model invariant and, moreover, a proper and invariant loss function exists, then in general the MRE procedure can be found. It does not depend on which particular loss function is chosen. We will not make this more concrete here; the details can be found in KROESE (1994). It is interesting that the MRE procedure, if it exists, corresponds to Fraser's structural inference or, equivalently, to the formal Bayes rule if the right Haar measure is used to construct a prior. So at least in the case where there is sufficient invariance, various approaches lead to the same distributional inference. This inference therefore comes close to 'inductive logic'. The MRE rule may be inadmissible unfortunately. An easy example is as follows:

EXAMPLE 1. Suppose the outcome x of $X \sim N_p(\mu, I_p)$ has to be used to make

a distributional inference $Q(x)$ about μ , where $p \geq 3$. Consider the loss function

$$L(\theta, v) = -\log f(\theta) \quad (6)$$

where $f = dv/d\lambda$ is the density of v with respect to Lebesgue measure on $\mathbb{R}^p$. Loss function (6) is proper, see BERNARDO (1979), and translation invariant. Hence, by an argument like that in Section 4, the procedure Q , defined by $Q(x) = N_p(x, I_p)$, is MRE w.r.t. this loss function. In STEIN (1955) it was shown that estimators $d: \mathbb{R}^p \rightarrow \mathbb{R}^p$ exist such that

$$E_\mu \|d(X) - \mu\|_2^2 < E_\mu \|X - \mu\|_2^2$$

holds for all $\mu \in \mathbb{R}^p$. Now we compare the risk of Q with that of Q' defined by $Q'(x) = N_p(d(x), I_p)$. The risk of Q' is equal to

$$\begin{aligned} E_\mu L(\mu, Q'(X)) &= E_\mu [-\log((2\pi)^{-n/2} \exp(-\frac{1}{2} \|\mu - d(X)\|_2^2))] \\ &= \frac{1}{2} n \log 2\pi + \frac{1}{2} E_\mu \|d(X) - \mu\|_2^2 \end{aligned}$$

and it is immediately clear that this is uniformly less than the (constant) risk of the MRE rule Q which, as a consequence, is inadmissible if logarithmic loss is considered.

Hence, exactly as in point estimation, an MRE procedure is not necessarily admissible. A similar result, for other loss functions, can be found in EATON (1982). A second complication is that a proper and invariant loss function does not always exist. An example is given in the next section.

7 Student's problem

Section 4 deals with inference about μ in the case of a prescribed error distribution. A more difficult and possibly more realistic context is that where the error distribution is known up to a scale parameter. We will focus on the case where normality assumptions are made. It is not difficult to extend the argument to a general location-scale family, but the normal case is of crucial interest.

Suppose that the outcome $(x_1, \dots, x_n)$ of an independent random sample from $N(\mu, \sigma^2)$ has to be used to make a distributional inference about μ , while σ is unknown. The problem is invariant under affine transformations, as $(ax_i + b, \dots, ax_n + b)$ can be regarded as a random sample from $N(a\mu + b, a^2\sigma^2)$ for all $(a, b) \in \mathbb{R}^+ \times \mathbb{R}$. It is natural to choose a loss function $L: \mathbb{R} \times \mathbb{R}^* \rightarrow \mathbb{R}$ that is proper and invariant under affine transformations. That is, to choose the loss function such that the loss does not change if the true value μ and the distributional inference are shifted and rescaled in the same way. It can be proved, however, that no 'satisfactory' such loss functions exist, see KROESE (1994). Here 'satisfactory' means that the risk of the best equivariant rule is finite and that the loss function is not constant. An attractive approach now is to focus on the (translation invariant) loss function (1). If the attention is restricted to the affine equivariant procedures,

the procedure that uniformly minimizes the risk function can be found. In Appendix B it is proved that this is procedure $Q^{(0)}$ defined by

$$Q^{(0)}(x_1, \dots, x_n) = \mathcal{L}\bar{x} + \frac{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}}{\sqrt{n}} T_n$$

where T_n is a random variable with the Student t -distribution with n degrees of freedom. The 'standard' procedure $Q^{(1)}$, defined by

$$Q^{(1)}(x_1, \dots, x_n) = \mathcal{L}\bar{x} + \frac{\left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}}{\sqrt{n}} T_{n-1}$$

is also affine equivariant. Therefore standard Student inference is not admissible w.r.t. loss function (1). It is of some interest to allow the loss function to depend on σ . If for example the loss function $L(\{\mu, \sigma\}, \nu) = -\log f(\mu) - \log(\sigma)$ is chosen where f is the density of ν , then $Q^{(1)}$ is best affine equivariant and $Q^{(0)}$ thus not admissible (see Appendix B). We prefer the first approach, because we do not like the loss function to depend on σ . The evaluation of a distributional inference will in practice be performed on the basis of μ and Q only.

8 Coherency?

Facing uncertainty, it is natural not to know exactly what to do. The statistician can follow some precise procedure but he will always worry about whether this is reasonable. We quote from SAVAGE (1977):

A more serious difficulty with our exploitation of personal probability is this. Perhaps we are too far, in our capacities, from the ideally coherent man to benefit always by emulating him. Unlike him, we are the victims of vagueness and of relative inability to count and calculate. Indeed, were it not for this human weakness, the theory of personal probability would have no human use. Yet, since we are imperfect, we might conceivably be led systematically astray in trying to emulate perfection.

The 'ideally coherent man' or 'rational decision maker' some Bayesians and philosophers discuss is an idealized person who is coherent in the very narrow *probabilistic* sense that the probabilities assigned by him comply with each other: if a distributional inference $Q(x)$ is provided about an unknown y then for him this defines a distributional inference about $z = f(y)$ for all mappings f .

Though we like to apply probabilistic coherency as a 'principle' (see Section 9), we reject it as a 'dogma' because it is an example of what Whitehead has called 'a fallacy of misplaced concreteness'. Epistemic probabilities are mixtures of fact and fiction and there is a good scientific reason why distributional inference should not be required to cohere in the restrictive sense described above. The fictions (e.g. prior

distributions and loss functions) most appropriate for one situation are not necessarily in line with the fictions needed in another situation. It is too restrictive to require that always the same fiction should do. Note that this is also exemplified in the Bayesian literature. If one is interested in testing a point null hypothesis $H: \theta = \theta_0$, a prior is used that assigns positive mass to θ_0 , see for example BERGER (1985). If a different testing problem is considered a different prior is chosen. In our loss function formulation it is possible that a procedure Q for making distributional inferences about y has some optimality property (for example it can be MRE, Bayes-best unbiased, minimax), while the induced procedure for $f(y)$ does not have this optimality property. Examples will be given in the following sections.

In WILKINSON (1977) 'empirical evidence' is given for noncoherence as an intrinsic property of observational data. For example, suppose that the outcome x of a random variable $X \sim N_p(\theta, I_p)$ has been observed. Here $N_p(\theta, I_p)$ denotes the p -dimensional normal distribution with mean θ and the identity matrix as covariance matrix. Many will agree that $Q(x) = N_p(x, I_p)$ is the 'correct' distributional inference about θ . The induced distributional inference about $\|\theta\|^2$, however, is very unattractive, see Section 3.5 in Wilkinson's paper. Note that the expectation of this induced distributional inference is $p + \|x\|^2$, while $\|X\|^2$ has a noncentral χ^2 distribution with mean $p + \|\theta\|^2$.

Giving up the principle of probabilistic coherence leads to ad hocery which is unpleasant but not unrealistic. One starts to understand what Neyman meant by his 'Life is complicated but not uninteresting'. He wanted to develop a nondogmatic approach to statistics and that is what we want too. In the following sections the question of coherency will be central.

9 Multiple hypothesis testing

Suppose that an acceptable, e.g. MRE, rule $Q^{(1)}$ is available for making distributional inferences about y , whereas the genuine interest is not in y itself but in 'its group membership' or, equivalently, in the truth or falsity of corresponding hypotheses H_i ($i = 1, \dots, m$). To be precise, suppose that $\mathcal{Y} = \mathcal{Y}_1 + \dots + \mathcal{Y}_m$ is some partitioning of the space $\mathcal{Y}$ of all theoretically possible y values. We are interested in the hypotheses $H_i: y \in \mathcal{Y}_i$ ($i = 1, \dots, m$). One of these must be true, but which one?

Instead of giving a multiple-decision formulation directly, we can make a distributional inference about $i = i(y)$, where $i: \mathcal{Y} \rightarrow \{1, \dots, m\}$ is the indicator of the hypothesis which is true. So we are looking for a reasonable procedure $Q: \mathcal{X} \rightarrow \{1, \dots, m\}^*$. Such a procedure Q can be obtained by 'projecting' $Q^{(1)}$: define $\hat{Q}^{(1)}$ by

$$\hat{Q}^{(1)}(x) = (Q^{(1)}(x)(\mathcal{Y}_1), \dots, Q^{(1)}(x)(\mathcal{Y}_m)) \quad (7)$$

Note that $\hat{Q}^{(1)}(x) = Q^{(1)}(x) \circ i^{-1}$. One should not be misled by these formal niceties. The problem of making a distributional inference about i requires the formulation of a loss function $L(i, \nu)$, where ν is now a probability measure on $\{1, \dots, m\}$, and of a couple of principles which are not necessarily in line with those behind the construction of $Q^{(1)}$. Much will depend on the specific application in mind but the

properness of L is crucial. In the references of Section 2 an abundance of proper loss functions for this problem can be found. Very well-known is the 'Brier score' introduced in BRIER (1950):

$$L(i, v) = \|e_i - q\|_2^2 \quad (8)$$

where e_i is the i th unit vector and q is the vector of probabilities assigned by v . If the classes are ordered, the 'Ranked Probability Score' introduced in EPSTEIN (1969) can be considered. Loss function (1) is a generalization of this Ranked Probability Score.

Whether or not the procedure $\hat{Q}^{(1)}$ defined in (7) is most attractive, it certainly is a convenient and natural rule. Having specified an epistemic probability distribution $Q(x)$, for example $\hat{Q}^{(1)}(x)$, on $\{1, \dots, m\}$, immediate applications can be made to multiple-decision problems: suppose the space $\mathcal{A} = \{a_1, \dots, a_n\}$ consists of a finite number of actions, usually statements about the truth or falsity of various combinations of $H_1, \dots, H_m$. The loss function L , given by

$$L(y, a_j) = w_{ij}$$

if $y \in \mathcal{Y}_i$ ($i = 1, \dots, m; j = 1, \dots, n$), is constant on the various subsets of $\mathcal{Y}$ or, equivalently, the loss depends on y via the group-membership variable $i = i(y)$.

In SCHAAFSMA (1969) such formulations were considered and it was established that Wald's minimax risk principle leads to degenerate results in the sense that, in the classical setting where $y = t$ (the true value of a parameter), least favourable distributions tend to concentrate on 'the common boundary'. A natural way out of these difficulties is to abandon the minimax principle and to proceed along the lines of the present section by taking the action a_j which minimizes the expected loss

$$\sum_{i=1}^m w_{ij} Q(x)(\{i\})$$

with respect to the epistemic probabilities suggested. If $\hat{Q}^{(1)}$ is chosen, probabilistically coherent results are obtained. This should, however, not obscure the fact that these special-purpose interpretations of $Q^{(1)}$ are not necessarily optimal. For example, in many probability models the MRE rule $Q^{(1)}$ is minimax, see KIEFER (1957). The corresponding multiple-decision rule, however, will usually neither be MRE nor minimax.

A referee noticed the following: suppose a distributional inference $Q^{(v)}(x)$ is obtained about y by using the Bayes rule w.r.t. a probability measure τ and a proper loss function. A distributional inference about $i(y)$ can be obtained by projecting $Q^{(v)}(x)$. If the loss function for the problem of making a distributional inference about $i(y)$ is proper, for example if loss function (8) is used, then this induced distributional inference corresponds to the Bayes rule w.r.t. τ in this problem. Hence considering Bayes rules w.r.t. a fixed prior distribution leads to probabilistically coherent inferences, in the sense that induced/projected distributional inferences are inferentially valid. A similar observation was made in EATON (1982). It is not clear, however, that the same prior distribution is appealing if we are interested in different

aspects of y . In fact it is obvious that it isn't, at least not if the aspects are sufficiently different. We think that the prior should depend to some extent on the aspect of y of interest. In that case the probabilistic coherence of the inferential statements is lost, but increased relevance is gained. An extreme example is that where we are interested in testing a point null hypothesis. This will be considered in the next section.

10 Classical hypothesis testing

It is interesting to specialize Section 9 to the case where $m = 2$ and $y = t$ is the true value of the underlying parameter. We shall present an example where the suggestions of Section 9 work out nicely and another one where the principle of probabilistic coherency has to be abandoned because it leads to something completely unacceptable. The two hypotheses H_1 and H_2 involved will now be denoted as 'the' hypothesis H ($=H_1$) and its negation A ($=H_2$).

To be precise, let x be the outcome of a random variable X whose distribution $P = \mathcal{L}(X)$ on the sample space $\mathcal{X}$ is assumed to be in the family $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ such that the true value t of θ can be defined by $P = P_t$. We consider the dichotomy $\Theta = \Theta_H + \Theta_A$ and we are interested in the truth or falsity of the hypotheses $H: t \in \Theta_H$ and $A: t \in \Theta_A$. Suppose we have to make a choice from

a_0 : no sufficient evidence for deciding between H and A ,

a_H : H is accepted and A is rejected.

a_A : A is accepted and H is rejected.

in the light of the loss function L_c defined by

$L_c(\theta, a)$	a_0	a_H	a_A
$\theta \in \Theta_H$	c	0	1
$\theta \in \Theta_A$	c	1	0

where c is (much) smaller than $1/2$.

EXAMPLE 2. Consider the context of Section 4 and suppose that one is interested in the question whether $H: \mu(=t) < 0$ or $A: \mu(=t) > 0$, the possibility of $\mu(=t) = 0$ being ignorable. Section 4 and Section 9 suggest to proceed as follows: first, $Q^{(1)}$ is defined by $Q^{(1)}(x) = \mathcal{L}(x + \mu - X)$ with corresponding c.d.f. $G_x^{(1)}(\cdot)$ defined by $G_x^{(1)}(z) = 1 - F(x - z)$. Secondly, $\hat{Q}^{(1)}$ is determined by the epistemic probabilities $\hat{Q}^{(1)}(x) =$

$$(\hat{Q}^{(1)}(x)(\Theta_H), \hat{Q}^{(1)}(x)(\Theta_A)) = (G_x^{(1)}(0), 1 - G_x^{(1)}(0)) = (1 - F(x), F(x))$$

Thirdly the action that has to be chosen is determined by minimizing the (epistem-

ically) expected loss, which equals c if a_0 , $F(x)$ if a_H , and $1 - F(x)$ if a_A is chosen. In this way we obtain the three-decision rule

$$d(x) = \begin{cases} a_H & \text{if } x < F^{-1}(c) \\ a_0 & \text{if } F^{-1}(c) \leq x \leq F^{-1}(1-c) \\ a_A & \text{if } F^{-1}(1-c) < x \end{cases}$$

These procedures $Q^{(1)}$, $\hat{Q}^{(1)}$, and d are very satisfactory in the sense that they are nicely probabilistically coherent. $Q^{(1)}$ is MRE, $\hat{Q}^{(1)}$ has minimum integrated risk, if finite, w.r.t. Lebesgue measure for proper loss functions, and d minimizes the integrated risk, if finite, w.r.t. Lebesgue measure for loss function L_c .

EXAMPLE 3. Consider the context of Section 4 and suppose that one is interested in the question whether the (point null) hypothesis $H: \mu = 0$ is true or its negation $A: \mu \neq 0$. Now the requirement of probabilistic coherency leads us nowhere because $Q^{(1)}(x)(\Theta_H) = 0$ and $d(x) = a_A$ for all $x \in \mathcal{X}$. As was said before, in some Bayesian literature the prior is adapted to the null hypothesis of interest, see for example BERGER (1985).

The standard frequentist approach to this problem is to compute the 'exceedance probability' which, in the present context, can be identified as $\alpha(x) = 2 \min \{1 - F(x), F(x)\}$. A procedure Q can be defined by $Q(x) = (\alpha(x), 1 - \alpha(x))$. In the recent literature (HWANG et al., 1992, VAN DER MEULEN, 1992, VAN DER MEULEN et al., 1993, SCHAAFSMA, 1989, SCHAAFSMA et al., 1989) attempts are made to embed such ' p -values' $\alpha(x)$ in a theory which focuses on 'the intrinsic meaning' of the procedure $\alpha: \mathcal{X} \rightarrow [0, 1]$, which is that it should 'approximate the ideal value' which is 1 if H is true and 0 if H is false. The idea is that the procedure $\alpha: \mathcal{X} \rightarrow [0, 1]$ should be regarded as an estimator of the true value $1_{\{0\}}(\mu)$ of the indicator function corresponding to the hypothesis H . Compliance with the Bayesian approach requires that this procedure should be evaluated on the basis of a *proper* loss function, for example squared error loss. If one studies the risk function w.r.t. squared error loss, i.e.

$$R(\mu, \alpha) = E_\mu(1_{\{0\}}(\mu) - \alpha(X))^2$$

as a function of μ , then a discontinuity appears at 0 unless the estimator $\alpha: \mathcal{X} \rightarrow [0, 1]$ satisfies

$$E_0 \alpha(X) = \frac{1}{2}$$

This is obvious because $R(\mu, \alpha)$ can be decomposed into the variance of $\alpha(X)$ which is a continuous function of μ and the squared bias which is equal to $(1 - E_0 \alpha(X))^2$ if $\mu = 0$ and $E_\mu \alpha^2(X)$ if $\mu \neq 0$. Theory has been developed (see for instance, VAN DER MEULEN et al., 1993) for minimizing certain expressions, e.g. integrated risks, under the restriction that the estimator $\alpha: \mathcal{X} \rightarrow [0, 1]$ is 'similar', i.e. that $E_0 \alpha(X) = 1/2$. Further motivation for this and other concepts of similarity can be found in the paper cited above. If the theory is applied to the present example where P is the standard

normal distribution, and the integrated mean squared error $\int R(\mu, \alpha) d\mu$ is minimized among all procedures satisfying $E_0\alpha(X) = 1/2$, then the estimator

$$\alpha(x) = \frac{e^{-x^2/2}}{\sqrt{2}}$$

is obtained, see SCHAAFSMA (1989). Note that this 'weight of evidence' can immediately be applied to the three-decision problem formulated. It is natural to expect that the 'similarity' or 'continuous-risk' requirement involved in the construction of α makes it inadmissible. The conjecture is true. An alternative estimator α' has explicitly been constructed by J. Tolboom (personal communication, 1988) such that $R(\mu, \alpha') < R(\mu, \alpha)$ holds for all μ .

11 Discussion

There are many papers and books about the statistical controversy. An interesting survey is WILKINSON (1977). The summary of that paper contains the following statement:

One key to resolving the conflict lies in recognizing that inferential probability derived from observational data is inherently *noncoherent* in the sense that their inferential implications cannot be represented by a single probability distribution on the parameter space...

Noncoherence is mathematically inherent in confidence and fiducial theory, and provides a basis for reconciling the Fisherian and Neyman–Pearson viewpoints. A unified theory of confidence-based inferential probability is presented, and the fundamental incompatibility of this with Subjective Bayesian theory is discussed.

Our conclusions are only slightly different. The loss function approach we recommend, does not exclude any type of approach. The statistical controversy will never be resolved unless we agree that there are many ways to argue from the observation x to a distributional inference $Q(x)$ about the true value y of the unknown concerned. The different approaches can be compared on the basis of a (proper) loss function. Sometimes 'optimum' procedures for making distributional inferences can be found. Replacing the criterion by another one, though, will usually lead to another procedure being 'optimal'.

Also the statistical model will always be doubtful, not only because the true distribution P may be outside the class $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, but especially because the model chosen is only one from a family of $\mathcal{P}$'s satisfying $P \in \mathcal{P}$. There is considerable fiction involved in everything and that is why nobody can be sure. Except for this, many models do not allow the construction of an 'optimum' procedure. In that case some 'reasonable' procedure has to be found.

All this does not imply that the making of distributional inferences is useless or 'too ambitious'. It simply means that difficulties and responsibilities are involved in

making such inferences. We, statisticians, have to train ourselves in resolving these difficulties.

Appendix A (Proof of the theorem in Section 3)

It is easy to see that $Q^{(0)}$ is unbiased: let X and X' be independent $B(n, \theta)$ distributed random variables. Formula (3) implies that

$$\mathbf{E}_\theta G_X^{(0)}(\theta) = P_\theta(X > X') + \frac{1}{2} P_\theta(X = X') = \frac{1}{2}$$

for all $\theta \in (0, 1)$. It remains to be proved that $Q^{(0)}$ minimizes the integrated risk w.r.t. the uniform prior and loss function (5) among all unbiased procedures.

Let as before $G_s^{(\text{BL})}(\cdot)$ denote the distribution function of the Beta $(s+1, n-s+1)$ distribution. The latter is the posterior distribution of θ if $U[0, 1]$ is used as prior and s is observed. Let $m(s)$ denote the corresponding marginal probability that $X = s$. Hence $m(s) \equiv 1/(n+1)$. For an arbitrary unbiased procedure Q , that generates inferences $Q(s)$ with corresponding c.d.f.'s $G_s(\cdot)$, we can write

$$\begin{aligned} \int_0^1 R(\theta, Q) d\theta &= \int_0^1 \left[\int_0^1 \mathbf{E}_\theta (1_{[\theta, \infty)}(z) - G_X(z))^2 d\theta \right] dz \\ &= \int_0^1 \left[\sum_{s=0}^n \int_0^1 (1_{[\theta, \infty)}(z) - G_s(z))^2 dG_s^{(\text{BL})}(\theta) \right] m(s) dz \\ &= \int_0^1 \left[\sum_{s=0}^n [G_s^{(\text{BL})}(z)(1 - G_s(z))^2 + (1 - G_s^{(\text{BL})}(z))G_s^2(z)] m(s) \right] dz \\ &= \int_0^1 \left[\sum_{s=0}^n \left[\left(G_s^{(\text{BL})}(z) + \frac{c(z)P_z(X=s)}{m(s)} \right) (1 - G_s(z))^2 \right. \right. \\ &\quad \left. \left. + \left(1 - G_s^{(\text{BL})}(z) - \frac{c(z)P_z(X=s)}{m(s)} \right) G_s^2(z) \right] m(s) \right] dz \end{aligned}$$

where $c: [0, 1] \rightarrow \mathbb{R}$ is arbitrary. The last equality follows from

$$\sum_{s=0}^n P_z(X=s)(1 - G_s(z))^2 - P_z(X=s)G_s^2(z) = (1 - 2\mathbf{E}_z G_X(z)) = 0$$

for all $z \in (0, 1)$. Here we make use of the assumption that Q is unbiased. If a function c can be found such that

$$G_s^{(0)}(z) = G_s^{(\text{BL})}(z) + \frac{c(z)P_z(X=s)}{m(s)} \quad (9)$$

for all s and z , we have finished the proof. This follows from the fact that for all

$p \in [0, 1]$ the expression $p(1-q)^2 + (1-p)q^2$ is minimum, as a function of q , for $q = p$. Now observe that for all s and z we have

$$\begin{aligned} G_s^{(\text{BL})}(z) + \frac{c(z)P_z(X=s)}{m(s)} &= \text{Beta}_{s+1, n-s+1}(z) + c(z)(n+1)P_z(X=s) \\ &= P_z(X > s) + (z + c(z)(n+1))P_z(X=s) \end{aligned}$$

Here $\text{Beta}_{f,g}(\cdot)$ is the c.d.f. of a Beta distribution with parameters f and g . If we choose $c(z) = (1/2 - z)/(n+1)$, for all z , then (9) holds and thus the theorem is proved.

Appendix B (Proof of the assertions of Section 7)

Suppose $(x_1, \dots, x_n)$ is the outcome of an independent random sample from $N(\mu, \sigma^2)$. A distributional inference is required about μ while σ^2 is unknown. By sufficiency, this inference should only depend on $(\bar{x}, s)$, where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Now define the loss function

$$L_0(\{\mu, \sigma\}, \nu) = \frac{1}{\sigma} \int_{\mathbb{R}} (1_{[\mu, \infty)}(z) - G(z))^2 dz \quad (10)$$

Note that the natural ordering of procedures is the same for loss functions (1) and (10). The problem with loss function (10) is invariant under affine transformations, i.e. under the group $\mathcal{G} = \mathbb{R}^+ \times \mathbb{R}$ with actions

$$(g_1, g_2) \circ (\bar{x}, s) = (g_1 \bar{x} + g_2, g_1 s)$$

$$(g_1, g_2) \circ (\mu, \sigma) = (g_1 \mu + g_2, g_1 \sigma)$$

$$(g_1, g_2) \circ \nu = g_2 + g_1 \nu$$

Here $\circ$ denotes group action. The invariance of L_0 is easily verified. Note that L_0 is not proper, properness meaning that the expected loss, w.r.t. some joint distribution for μ and σ , is minimum for the corresponding marginal distribution of μ . A procedure Q , only depending on $(\bar{x}, s)$, is equivariant iff it is of the form $Q(\bar{x}, s) = \bar{x} + s\nu_0$, where ν_0 is some probability distribution on $\mathbb{R}$ not depending on $(\bar{x}, s)$. Let Q be the equivariant procedure that corresponds with ν_0 in this way and let $G(\cdot)$ denote the c.d.f. of ν_0 . The risk of Q can be written as

$$\begin{aligned} E_{\mu, \sigma^2} L_0(\{\mu, \sigma\}, Q(\bar{X}, S)) &= E_{\mu, \sigma^2} L_0\left(\left\{\frac{\mu - \bar{X}}{S}, \frac{\sigma}{S}\right\}, \nu_0\right) \\ &= E_{\mu, \sigma^2} \int_{\mathbb{R}} \left(\frac{S}{\sigma} G^2(z) 1_{(-\infty, (\mu - \bar{X})/S)}(z) \right. \\ &\quad \left. + \frac{S}{\sigma} (1 - G(z))^2 1_{((\mu - \bar{X})/S, \infty)}(z)\right) dz \end{aligned} \quad (11)$$

$$= \int_{\mathbb{R}} \left(\mathbb{E}_{\mu, \sigma^2} \frac{S}{\sigma} \right) G^2(z) - 2 \left(\mathbb{E}_{\mu, \sigma^2} \frac{S}{\sigma} 1_{[(\mu - \bar{X})/S, \infty)}(z) \right) G(z) \\ + \left(\mathbb{E}_{\mu, \sigma^2} \frac{S}{\sigma} 1_{[(\mu - \bar{X})/S, \infty)}(z) \right) dz$$

We see that the risk is minimum for $G^{(0)}(\cdot)$ defined by

$$G^{(0)}(z) = \frac{\mathbb{E}_{\mu, \sigma^2} \frac{S}{\sigma} 1_{[(\mu - \bar{X})/S, \infty)}(z)}{\mathbb{E}_{\mu, \sigma^2} \frac{S}{\sigma}}$$

A routine calculation shows that $G^{(0)}(\cdot)$ is the cumulative distribution function belonging to the distribution of $\sqrt{(n-1)/n} T_n$, where T_n has the Student t -distribution with n degrees of freedom. Hence the best equivariant procedure $Q^{(0)}$ is given by

$$Q^{(0)}(x_1, \dots, x_n) = \mathcal{L}\bar{x} + \frac{s\sqrt{n-1}}{n} T_n = \mathcal{L}\bar{x} + \frac{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}}{\sqrt{n}} T_n$$

It follows that $Q^{(1)}$ is inadmissible w.r.t. loss function (10) and thus w.r.t. loss function (1).

The best equivariant rule can also be calculated w.r.t. loss function L given by

$$L(\{\mu, \sigma\}, v) = -\log f(\mu) - \log(\sigma) \quad (12)$$

This loss function is proper, in the above described sense that the expected loss w.r.t. a joint distribution for μ and σ is minimum for the corresponding marginal distribution of μ . Loss function (12) is also invariant under $\mathcal{G}$. Now (11) immediately implies that the optimal v_0 is equal to $\mathcal{L}_{\mu, \sigma}(\mu - \bar{X})/S$, which is equal to $\mathcal{L}T_{n-1}/\sqrt{n}$. This implies that $Q^{(1)}$ is the best equivariant rule w.r.t. loss function (12) and hence $Q^{(0)}$ not admissible.

Acknowledgement

We thank the referees for their remarks, which led to several improvements.

References

- BERGER, J. O. (1985), *Statistical decision theory and Bayesian analysis*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo.
- BERNARDO, J. M. (1979), Expected information as expected utility, *Annals of Statistics* 7, 686–690.
- BRIER, G. W. (1950), Verification of forecasts expressed in terms of probability, *Monthly Weather Review* 78, 1–3.
- BUEHLER, R. (1980), Fiducial inference, in: *R. A. Fisher: An appreciation*, Springer Lecture Notes in Statistics, New York, 109–118.
- EATON, M. L. (1982), A method for evaluating improper prior distributions, *Statistical Decision Theory and Related Topics III* 1, 329–352.

- EATON, M. L. (1992), A statistical diptych: admissible inferences-recurrence of symmetric Markov chains, *Annals of Statistics* 20, 1147–1179.
- EATON, M. L. (1993), On the inadmissibility of some best invariant predictive distributions: The translation parameter case, in: *Statistics and probability: A Raghu Raj Bahadur Festschrift*, Wiley Eastern Limited, New Delhi, 191–202.
- EFRON, B. (1970), Comment on “On the inference and decision models of statistics” by Blyth, C. R., *Annals of Mathematical Statistics* 41, 1049–1054.
- EPSTEIN, E. S. (1969), A scoring system for probability forecasts of ranked categories, *Journal of Applied Meteorology* 8, 985–987.
- FISHER, R. A. (1959), *Statistical methods and scientific inference*, (2nd ed.), Oliver and Boyd, Edinburgh.
- FRASER, D. A. S. (1968), *The structure of inference*, Wiley, New York.
- GEISSER, S. (1984), On prior distributions for binary trials, *American Statistician* 38, 244–251.
- HENDRICKSON, A. D. and R. J. BUEHLER (1971), Proper scores for probability forecasters, *Annals of Mathematical Statistics* 42, 1916–1921.
- HWANG, J. T., G. CASELLA, C. ROBERT, M. WELLS and R. H. FARELL (1992), Estimation of accuracy in testing, *Annals of Statistics* 20, 490–509.
- KIEFER, J. C. (1957), Invariance, minimax sequential estimation and continuous time processes, *Annals of Mathematical Statistics* 28, 573–601.
- KROESE, A. H. (1994), forthcoming Ph.D. thesis, University of Groningen.
- LINDLEY, D. V. (1990), The 1988 Wald Memorial Lectures: the present position in Bayesian statistics, *Statistical Science* 5, 44–89.
- MEULEN, E. A. VAN DER (1992), *Assessing weights of evidence for discussing classical statistical hypotheses*, Ph.D. thesis, University of Groningen.
- MEULEN, E. A. VAN DER and W. SCHAAFSMA (1993), Assessing weights of evidence for discussing classical statistical hypotheses, *Statistics & Decisions* 11, 201–220.
- SAVAGE, L. J. (1971), Elicitation of personal probabilities and expectations, *Journal of the American Statistical Association* 66, 783–801.
- SAVAGE, L. J. (1977), The shifting foundations of statistics, in: *Logic, laws, and life: some philosophical complications*, U. Pittsburgh, 3–18.
- SCHAAFSMA, W. (1969), Minimax risk and unbiasedness for multiple decision problems of type I, *Annals of Mathematical Statistics* 40, 1684–1720.
- SCHAAFSMA, W. (1989), Discussing the truth or falsity of a statistical hypothesis H and its negation A, in: *International workshop on theory and practice in data analysis, Proceedings*, Berlin GDR. REPORT R-MATH-01/89, 150–166.
- SCHAAFSMA, W., J. TOLBOOM and E. A. VAN DER MEULEN (1989), Discussing truth or falsity by computing a Q value, in: *Statistical data analysis and inference*, North Holland, Amsterdam, 85–100.
- SCHERVISH, M. J. (1989), A general method for comparing probability assessors, *Annals of Statistics* 17, 1856–1879.
- SEIDENFELD, T. (1992), R. A. Fisher's fiducial argument and Bayes' theorem, *Statistical Science* 7, 358–368.
- STAËL VON HOLSTEIN, C.-A. S. and A. H. MURPHY (1978), The family of quadratic scoring rules, *Monthly Weather Review* 106, 917–924.
- STEIN, C. (1955), Inadmissibility of the usual estimator of the mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1, 197–206.
- WILKINSON, G. N. (1977), On resolving the controversy in statistical inference, *Journal of the Royal Statistical Society Series B* 39, 119–171.
- ZABELL, S. L. (1992), R. A. Fisher and the fiducial argument, *Statistical Science* 7, 369–387.

Received: June 1993. Revised: April 1994.