Augustin, Pöhlmann:

On Robust Sequential Analysis - Kiefer-Weiss Optimal Testing under Interval Probability

Sonderforschungsbereich 386, Paper 261 (2001)

Online unter: http://epub.ub.uni-muenchen.de/

On Robust Sequential Analysis – Kiefer-Weiss Optimal Testing under Interval Probability

Thomas Augustin
University of Munich
Department of Statistics
Akademiestr. 1
D-80799 München
Germany

Sigrid Pöhlmann
University of Dortmund
Department of Statistics
D-44221 Dortmund
Germany

September 28, 2001

Abstract

Usual sequential testing procedures often are very sensitive against even small deviations from the 'ideal model' underlying the hypotheses. This makes robust procedures highly desirable. To rely on a clearly defined optimality criterion, we incorporate robustness aspects directly into the formulation of the hypotheses considering the problem of sequentially testing between two interval probabilities (imprecise probabilities). We derive the basic form of the Kiefer-Weiss optimal testing procedure and show how it can be calculated by an easy-to-handle optimization problem. These results are based on the reinterpretation of our testing problem as the task to test between nonparametric composite hypotheses, which allows to adopt the framework of Pavlov (1991). From this we obtain a general result applicable to any interval probability field on a finite sample space, making the approach powerful far beyond robustness considerations, for instance for applications in artificial intelligence dealing with imprecise expert knowledge.

Keywords: Interval probability, imprecise probabilities, sequential testing, robustness, Kiefer-Weiss optimality, total-variation neighbourhood models, least favorable pairs, composite hypotheses
AMS classification: Primary: 60A05, 62A01, 62L10; secondary: 62F05, 62F35, 62G10, 62G20, 62G35

1 Introduction

Sequential and group sequential procedures help ceteris paribus to reduce the sample size. So they have become the standard way of analysis especially in areas, where sampling cost of each unit is high, like in quality management and in many types of clinical trials (e.g. Jennison and Turnbull (1999)).
In contrast to the fixed sample case, the problem of robustness has rarely been addressed in sequential analysis, neglecting the fact that many of the standard procedures must be suspected to be highly sensitive to even small deviations from the 'ideal model' specifying a certain parametric distribution. But in many situations the distributional assumptions may be satisfied only approximately: for instance the measurements may be imprecise or outliers may occur. Furthermore, sometimes it is even impossible to formulate an 'ideal model' precisely. This is especially true in applications in artificial intelligence, where the models stem from, naturally rather imprecise, expert judgements.
One approach to take robustness into account (e.g. Christmann (1999)) will be called *ex post robustifcation* in this paper: procedures which are optimal for the 'ideal model' are robustified by passing over to robust versions of the statistic they are based on. (For instance, in the simplest case, using the median instead of the mean.) To find such robustifications, one tries to transfer experience from the case of a fixed sample size to the sequential case hoping that "what's good for fixed sample size can not be bad in the sequential case". The performance of such robustified versions then is evaluated with respect to certain measures of performance (for instance the breakdown point) or is justified by appropriate behaviour in simulation studies.
This paper would like to bring up a conceptually different approach for discussion. We propose to incorporate robustness considerations directly into the formulation of the hypotheses and *then* to search for optimal procedures in this extended setting. This so-to-say *ex ante robustification* has the appealing property that the whole development stands under a certain, precisely

defined optimality criterion (in our case a Kiefer-Weiss-type criterion). Therefore, the solutions gained are eo ipso justified to be optimal for the setting considered.
To formulate such hypotheses prepared for robustness, the natural framework is the notion of interval probability, also known as imprecise probability. This concept provides a superstructure upon the models commonly used in robust statistics to describe small deviations from an 'ideal model' as well as outliers (see e.g. Huber (1981, Chapter 10) or the review (and the extensions) in Augustin (2001)). Additionally interval probability is the tool per se to express uncertain knowledge in form of expert opinions probabilistically (e.g. Shafer (1976), Weichselberger and Pöhlmann (1990), Yager, Fedrizzi and Kacprzyk (1994)). A general survey on imprecise probabilities and a comprehensive bibliography can be found on the "imprecise probability page" (de Cooman and Walley (2001)); recent developments are discussed, for instance, in Bernard (2001).
To make this paper self-contained, in Section 2 we briefly collect some basics from the theory of interval probability. Section 3 turns to sequential testing and states the optimality criterion under consideration. Our main result describing the basic form of the optimal procedure is formulated and proven in Section 4. There we also discuss this procedure as well as aspects of its practical calculation, and finally illustrate it with a didactic example.

2 Interval probability

In the whole paper we will confine ourselves to a finite sample space $\mathcal{Y} = \{y_1, ..., y_n\}$ with n elements y_i and consider, w.o.l.g., $\mathcal{A} = \mathcal{P}(\mathcal{Y})$ as the σ-field on $\mathcal{Y}$ containing arbitrary events A. Singletons will separately be denoted by $E_j = \{y_j\}$, $j = 1, \ldots, n$.
Interval-valued assignments are symbolized by capital letters $P(\cdot)$ and are called **interval probabilities**; the lower interval limit is denoted by $L(\cdot)$, the upper one by $U(\cdot)$. As the name interval probability suggests, the probability of every event A is described by an interval $[L(A), U(A)] \subseteq [0; 1]$ instead of a single real number $p(A)$. To distinguish in notation and terminology, we call every probability in the usual sense, i.e. every set function satisfying Kolmogoroff's axioms, **classical probability** and denote it by small letters $p(\cdot)$.
The concept of interval probability allows to express the quality of infor-

mation or the degree of uncertainty in the probability statement itself. By this also robustness aspects can be taken into account properly. If there are doubts about the underlying model or if many outliers have to be expected, neighborhood models can be formulated leading to wider intervals. On the opposite, small intervals reflect probabilistic information with high reliability. Several axiomatizations for interval probabilities have been suggested in literature, which materially coincide in the case of a finite sample space. According to them interval-valued set functions

$$\begin{array}{rcccl} P & : & \mathcal{A} & \to & \{[L,U] : 0 \leq L \leq U \leq 1\} \\ & & A & \mapsto & [L(A), U(A)] \end{array}$$

can be distinguished with respect to the relation between the non-additive set functions $L(\cdot)$ and $U(\cdot)$ and the set

$$\mathcal{M} := \{p(\cdot) : L(A) \leq p(A) \leq U(A) \quad \forall A \in \mathcal{A}\}$$

of all classical probabilities $p(\cdot)$ being in accordance with them.
If at least $\mathcal{M} \neq \emptyset$, which is understood as a minimum requirement, the assignment can be interpreted as not contradictory to the concept of probability. In this paper we join Weichselberger's terminology, calling $P(\cdot)$ **R-probability** and $\mathcal{M}$ its **structure** (cf. Weichselberger and Pöhlmann (1990), Weichselberger (2001, Chapter 2)). [1]
If there is additionally an one-to-one correspondence between interval limits and the structure such that

$$\inf_{p \in \mathcal{M}} p(A) = L(A) \ , \quad \forall A \in \mathcal{A},$$

$$\sup_{p \in \mathcal{M}} p(A) = U(A) \ , \quad \forall A \in \mathcal{A},$$

an R-probability $P(\cdot)$ is called **F-probability** (cf. Weichselberger and Pöhlmann (1990) and Weichselberger (2001)).
Since there are well-defined ways to proceed from R- to F-probabilities (Weichselberger (2001), Chapter 2.5 and 2.6), we confine ourselves in the following to F-probability.

[1] In the frequentist theory of interval probability (e.g. Papamarcou and Fine (1986)) the set function $L(.)$ is called "dominated". Walley (1991) gives a behaviorial characterization of such assignments as "avoiding sure loss".

Note that in this situation necessarily $L(\cdot)$ and $U(\cdot)$ are conjugated:

$$U(A) = 1 - L(\neg A),\ A \in \mathcal{A}. \tag{1}$$

Therefore, one of the two set-functions $L(\cdot)$ or $U(\cdot)$ is sufficient to describe $P(\cdot)$.
In the way it was defined above, interval probability is characterized by assigning probability components to all events of the σ-field $\mathcal{A}$. It is a huge advantage of interval probability that it is possible to construct interval probability from any assignment on arbitrary subsets $\mathcal{A}_L$, $\mathcal{A}_U$ of $\mathcal{A}$. For this, consider partial assignments $\tilde{L}(\cdot)$ on $\mathcal{A}_L$ and $\tilde{U}(\cdot)$ on $\mathcal{A}_U$ such that

$$\begin{aligned}\tilde{\mathcal{M}} := \{p(\cdot) \in \mathcal{M} :\ & p(A) \geq \tilde{L}(A), \quad \forall A \in \mathcal{A}_L, \\ & p(A) \leq \tilde{U}(A), \quad \forall A \in \mathcal{A}_U\} \neq \emptyset.\end{aligned}$$

Then it can be shown that $P(\cdot) = [L(\cdot), U(\cdot)]$ with

$$L(A) := \inf_{p \in \tilde{\mathcal{M}}} p(A), \quad \forall A \in \mathcal{A},$$

$$U(A) := \sup_{p \in \tilde{\mathcal{M}}} p(A), \quad \forall A \in \mathcal{A},$$

is an F-probability with structure $\tilde{\mathcal{M}}$. That is, it is reflecting exactly the information contained in $\tilde{L}(\cdot)$ and $\tilde{U}(\cdot)$.
An important special case for applications is the situation where $\mathcal{A}_L$ and $\mathcal{A}_U$ are consisting of all singletons $E_j, j = 1, \ldots, n$. Then one is led to the theory of **probability intervals (PRI)** as described in Weichselberger and Pöhlmann (1990). In this case the limits $L(\cdot)$ and $U(\cdot)$ will be summarized in the following way:

$$\begin{bmatrix} L(E_1) & U(E_1) \\ \vdots & \vdots \\ L(E_n) & U(E_n) \end{bmatrix}.$$

3 Sequential testing

To prepare the study of sequential tests between interval probabilities and to introduce the notation used throughout this paper, let us briefly review some basics of sequential analysis (e.g. Ghosh (1970), Irle (1990)).

3.1 Classical theory

Consider two hypotheses H_0 and H_1, specifying two sets $\mathcal{W}_0$, $\mathcal{W}_1$ of probability distributions, with $\mathcal{W}_0 \cap \mathcal{W}_1 \neq \emptyset$, on the same measurable space. In sequential analysis one solves the task of deciding between H_0 and H_1 by considering successively repeated observations. Given bounds α_i, $i \in \{0,1\}$, on the overall probabilities of falsely rejecting hypothesis H_i, one has to decide at every time point whether one is ready to accept H_0, or to accept H_1, or whether a further observation has to be drawn. This leads to

Definition 1 *:*
Consider a finite space $\mathcal{Y}$, a sequence $X_1, X_2, \ldots$ of independent random elements mapping from a measurable space $(\Omega, \mathcal{G})$ into $(\mathcal{Y}, \mathcal{P}(\mathcal{Y}))$ with common probability law $p(\cdot)$, and the filtration $\mathcal{A}_1, \mathcal{A}_2, \ldots$ adapted to $X_1, X_2, \ldots$.

*a) A **sequential test** for testing $H_0 : p(\cdot) \in \mathcal{W}_0$ versus $H_1 : p(\cdot) \in \mathcal{W}_1$ is a pair (N, D) where N is a stopping time with respect to the sequence $\mathcal{A}_1, \mathcal{A}_2, \ldots$ and D is an $\mathcal{A}_N$-measurable decision rule specifying which hypothesis is to be accepted once sampling has stopped.*

*b) For every sequential test (N, D) denote by $\gamma_i(N, D, p)$, $i \in \{0,1\}$, the overall probability of deciding in favour of $H_{i'}$, $i' \in \{0,1\}$, $i' \neq i$, if $p(\cdot) \in \mathcal{W}_0 \cup \mathcal{W}_1$ is true. Then, given two bounds α_0 and α_1, let $\boldsymbol{\mathcal{K}_{\alpha_0,\alpha_1}}$ be the **set of all sequential tests** (N, D) with $\gamma_i(N, D, p) \leq \alpha_i, \forall p(\cdot) \in \mathcal{W}_i$, $i \in \{0,1\}$.* □

Most work on sequential analysis considers the case of two simple hypotheses of the form $H_0 : p(\cdot) = p_0(\cdot)$ versus $H_1 : p(\cdot) = p_1(\cdot)$ where $p_0(\cdot)$ and $p_1(\cdot)$ are classical probabilities. Typically $p(\cdot)$ is described by a real-valued parameter θ being an element of a parameter space Θ, so that one tests de facto:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1 \,,$$

where, without loss of generality, $\theta_0 < \theta_1$ can be assumed. In this case two criteria have been suggested to distinguish one element of $\mathcal{K}_{\alpha_0,\alpha_1}$ as optimal: Wald and Wolfowitz (1948) proposed to define a test as optimal if it minimizes both $I\!E_{\theta_0} N$ and $I\!E_{\theta_1} N$ among all tests $(N, D) \in \mathcal{K}_{\alpha_0,\alpha_1}$, which also contains level-α_0 tests based on fixed sample sizes. This problem possesses a general solution, namely the sequential probability ratio test (SPRT) between θ_0 and θ_1, which firstly was introduced by Wald (1947).

The SPRT, however, may perform quite unsatisfactory for values between θ_0 and θ_1. This motivated Kiefer and Weiss (1957) to study a different criterion: a sequential test (N^*, D^*) solves the **Kiefer-Weiss problem**, if it minimizes the maximum expected sample size among all (sequential) tests $(N, D) \in \mathcal{K}_{\alpha_0,\alpha_1}$, i.e.

$$\sup_{\theta\in\Theta} I\!E_\theta N \to \min_{(N,D)} .$$

(In the modified Kiefer-Weiss problem $I\!E_\theta N$ is minimized only for a fixed θ.)
Constructing optimal solutions, with respect to these criteria, often has been impossible; therefore usually an asymptotic version with diminishing error probabilities has been considered (e.g. Eisenberg (1982), Huffman (1983), Pavlov (1991)), which will also motivate our generalization defined below.
In several papers, the criteria have been extended to the case of composite hypotheses described by a single one-dimensional parameter of the form "$\theta \leq \theta_0$ and $\theta \geq \theta_1$" (Ghosh (1970), Chapter 3.2).
Restricting considerations on invariant problems, Lai (see Lai (1981) and the references therein) extends the Wald-Wolfowitz situation as well as the modified Kiefer-Weiss problem to composite hypotheses. Pavlov (1991) presents an asymptotic solution to the Kiefer-Weiss problem for very general hypotheses.
Sequential tests are applied in several areas, especially when sampling costs or a small number of specimens to be investigated are of great importance. This can be not only in quality control, but also in such fields like epidemiology or biometrics (e.g. van der Tweel, Kaaks and van Noord (1996), Jennison and Turnbull (1999), or Pöhlmann and Augustin (2001)).

3.2 Sequential testing under interval probability

A natural way to test between two F-probabilities $P_0(\cdot) = [L^{(0)}(\cdot), U^{(0)}(\cdot)]$ and $P_1(\cdot) = [L^{(1)}(\cdot), U^{(1)}]$ considers the decision

$$H_0 : P(\cdot) = P_0(\cdot) \quad \text{versus} \quad H_1 : P(\cdot) = P_1(\cdot) \tag{2}$$

as a testing problem between the corresponding structures $\mathcal{M}_0$ and $\mathcal{M}_1$:

$$H_0 : p(\cdot) \in \mathcal{M}_0 \quad \text{versus} \quad H_1 : p(\cdot) \in \mathcal{M}_1 . \tag{3}$$

So, the task to test between two single, interval-valued hypotheses has been transformed into a classical composite testing problem, and Definition 1 can also be applied in this context.

Note, however, that the hypotheses formulated in (3) are of a very complex form; only in degenerated special cases they can be described by a one-dimensional parameter. As a consequence, the standard methods leading to the construction of optimal sequential tests are no longer directly applicable. Huber (1981, Chapter 10) generalized the Wald-Wolfowitz criterion to interval probability. He succeeded in extending the core of his famous result on the construction of minimax tests (Huber and Strassen (1973)) to the sequential situation, provided that the error probabilities are forced to converge to zero: under certain additional assumptions on the F-probabilities $P_0(\cdot)$ and $P_1(\cdot)$, the optimal procedure for the composite problem (3) can be obtained by considering the optimal procedure for the reduced problem $\bar{H}_0 : p(\cdot) = q_0(\cdot)$ versus $\bar{H}_1 : p(\cdot) = q_1(\cdot)$ where the classical probabilities $q_0(\cdot)$ and $q_1(\cdot)$ are so called least favorable elements of the structures. Quang (1985) has achieved an analogous result for contamination neighborhoods which are 'shrinking' with increasing sample size.

What was already briefly mentioned in Section 3.1 also applies here: the optimal procedure in the sense of the Wald-Wolfowitz criterion may perform quite unsatisfactory "between" the hypotheses. Therefore, in this paper we will consider an extension of the Kiefer-Weiss criterion to interval probability. (The modified Kiefer-Weiss problem can be generalized in an analogous way.) In the spirit of Kiefer and Weiss we have to minimize

$$\sup_{p(\cdot)\in\mathcal{C}(\mathcal{Y})} I\!E_p N \tag{4}$$

with $\mathcal{C}(\mathcal{Y}) = \mathcal{M}_0 \cup \mathcal{M}_1 \cup \mathcal{I}$ as the space of all classical probabilities $p(\cdot)$ lying in $\mathcal{M}_0$ or in $\mathcal{M}_1$ or in an indifference zone $\mathcal{I}$, i.e. a set "between" $\mathcal{M}_0$ and $\mathcal{M}_1$. The set $\mathcal{I}$ has to be specified appropriately: we take $\mathcal{I}$ such that $\mathcal{C}(\mathcal{Y})$ is the envelope of $\mathcal{M}_0 \cup \mathcal{M}_1$, i.e.

$$\mathcal{C}(\mathcal{Y}) := \{p(\cdot) |\ \min_{i=0,1} L_i(A) \le p(A) \le \max_{i=0,1} U_i(A),\ \forall A \in \mathcal{A}\}. \tag{5}$$

Since even in the classical, single parameter situation, the Kiefer-Weiss criterion in its pure form showed to be not tractable, (4) is certainly too complex to allow for a general solution. Therefore, we base our generalization of the Kiefer-Weiss criterion to interval probability on the asymptotic version, which has usually been considered in literature (cf. the references in Section 3.1). Hence we obtain:

Definition 2*:*
A test $(N^*, D^*) \in \mathcal{K}_{\alpha_0,\alpha_1}$ *is called* **asymptotically optimal** *among all tests in* $\mathcal{K}_{\alpha_0,\alpha_1}$ *if, for* $\alpha_0 \to 0$ *and* $\alpha_1 \to 0$*,*

$$\frac{\sup_{p(\cdot)\in\mathcal{C}(\mathcal{Y})} I\!E_p N^*}{\inf_{(N,D)\in\mathcal{K}_{\alpha_0,\alpha_1}} \sup_{p(\cdot)\in\mathcal{C}(\mathcal{Y})} I\!E_p N} = 1 + o(1). \tag{6}$$

□

4 Construction of an asymptotically optimal testing procedure

To construct optimal procedures it may look promising to aim at adopting Huber's result on Wald-Wolfowitz optimal sequential tests under interval probability to the criterion formulated in Definition 2: one could try to reduce the structures to the Huber-Strassen least favorable distributions $q_0(\cdot) \in \mathcal{M}_0$ and $q_1(\cdot) \in \mathcal{M}_1$; and then one would construct the Kiefer-Weiss optimal test between $q_0(\cdot)$ and $q_1(\cdot)$ in the hope that it is also optimal for the testing problem in Equation (3). Unfortunately, as is also demonstrated with Example 1, this conjecture does not work. Apparently, the problem of finding (asymptotically) Kiefer-Weiss optimal procedures has to be based on completely different methods, which will be presented in Theorem 1.

4.1 Main theorem

Before stating the theorem let us shortly describe the basic ideas underlying the procedure.
Sequentially, at each step $\ell \in I\!N$, a new (independent) observation $\{X_\ell = x_\ell\}$ with $x_\ell \in \{y_1, \ldots, y_n\}$ is drawn, and the adapted relative frequency $h^{(\ell-1)}(x_\ell)$ is calculated, based on the first $(\ell - 1)$ observations. This is done in the following way:

$$\begin{aligned} &\text{for } \ell \geq 2: \quad h^{(\ell-1)}(x_\ell) := \tfrac{1}{\ell-1} \textstyle\sum_{j=1}^{\ell-1} 1_{\{X_j = x_\ell\}}, \\ &\qquad\qquad \text{if it leads to a value in } \mathcal{C}(\mathcal{Y}) \text{ (otherwise see below);} \\ &\text{and} \qquad\quad h^{(0)}(x_1) := 1. \end{aligned}$$

It has to be noted that the construction is based on asymptotic considerations. If only few observations have been drawn, it can not be excluded that

$h^{(\ell-1)}(x_\ell)$ would take values not in accordance with $\mathcal{C}(\mathcal{Y})$. They even may be zero, spoiling the whole product in the numerator of $Q_\ell^{(i)}$ (see below) forever. In these cases, $h^{(\ell-1)}(x_\ell)$ has to be restricted to the smallest value being compatible with $\mathcal{C}(\mathcal{Y})$ (cf. Equation (5)).
With these adapted relative frequencies the ratio $Q_\ell^{(i)}$ has to be evaluated at each step ℓ for $i \in \{0,1\}$:

$$Q_\ell^{(i)} = \frac{\prod_{r=1}^{\ell} h^{(r-1)}(x_r)}{\sup_{p(\cdot)\in\mathcal{M}_i} \prod_{r=1}^{\ell} p(x_r)} \tag{7}$$

where $p(x_r) := p(\{X_r = x_r\})$ as the probability of x_r (given H_i resp. $\mathcal{M}_i$).
In $Q_\ell^{(i)}$ we compare, based on the available information up to that time, an estimated probability with the highest probability being in accordance with the hypothesis H_i. If this ratio is for the first time (with respect to ℓ) greater or equal to α_i^{-1}, for one index i (i' say), we call this time point $T^{(i')}$ and the process stops with $N = T^{(i')}$. The decision is to reject the corresponding hypothesis $H_{i'}$, respectively to accept the hypothesis $H_{i''}$ ($i' \neq i''; i', i'' \in \{0,1\}$), that is $D = H_{i''}$.
So we can summarize the procedure in the following theorem:

Theorem 1*:* Let $T^{(i)} := \min\{\ell : Q_\ell^{(i)} \geq \alpha_i^{-1}\}, i \in \{0,1\}$, with $Q_\ell^{(i)}$ as in Equation (7).

The asymptotically optimal testing procedure (N^*, D^*) in the sense of Definition 2 is:

If $T^{(0)} < T^{(1)}$ then $N^* = T^{(0)}$ and $D^* = H_1$ (the decision is for H_1);

if $T^{(0)} > T^{(1)}$ then $N^* = T^{(1)}$ and $D^* = H_0$ (the decision is for H_0).

□

Proof of Theorem 1 :
The proof of this theorem is based on the idea that it is possible to embed the situation under consideration into the general framework of Pavlov (1991). [2]

[2]Pavlov (1991) originally investigates sequential procedures for m hypotheses. With respect to our intended application of his results we confine ourselves to the case $m = 2$.

Given a finite dimensional parameter space Θ, two hypotheses $H_i : \theta \in \Theta_i$, with $\Theta_i \subsetneqq \Theta$, $i \in \{0,1\}$, $\Theta_0 \cap \Theta_1 = \emptyset$, an indifference region $\mathcal{I} = \Theta \backslash (\Theta_0 \cup \Theta_1)$ and error bounds α_0 and α_1, he constructs a test $(N^*, D^*) \in \mathcal{K}_{\alpha_0,\alpha_1}$ with:

$$\frac{\sup\limits_{\theta\in\Theta} I\!E_\theta N^*}{\inf\limits_{(N,D)\in\mathcal{K}_{\alpha_0,\alpha_1}} \sup\limits_{\theta\in\Theta} I\!E_\theta N} = 1 + o(1) \quad . \tag{8}$$

We will show that Pavlov's results also provide a solution to our problem of optimal sequential procedures between interval probabilities. Note that $X_1, X_2, \ldots, X_r, \ldots$ are i.i.d., so that we can write $p(\{X = y_j\})$ instead of $p(\{X_r = y_j\})$ for arbitrary r. To embed our problem into Pavlov's parametric framework we take

$$\theta = (\theta_1, \ldots, \theta_n)$$

with

$$\theta_j := p(\{X = y_j\}) = p(y_j), \ j = 1, \ldots, n,$$

and the constraint $\sum_{j=1}^n \theta_j = 1$. Therefore, every classical probability $p(\cdot) \in \mathcal{C}(\mathcal{Y})$ uniquely corresponds to a certain value $\theta \in \Theta$. In particular, we have $\Theta = \mathcal{C}(\mathcal{Y})$.

With this parametrization Pavlov's optimality criterion coincides with our asymptotic optimality criterion in Equation (6). Therefore, if we transfer Pavlov's main results (Pavlov (1991, Theorem 4.1 and Lemma 4.1)) to our situation, we can conclude that the testing procedure in Theorem 1 is asymptotically optimal in the sense of Equation (6).
To guarantee this transferance of his results, we need to prove further:

a) the identity of the test statistic $T^{(i)}$, given here, with Pavlov's test statistic

b) that Pavlov's conditions (1') to (4') (see Pavlov (1991, p. 283)) are satisfied in the situation given here.

Ad a) Pavlov uses a test statistic based on the ratio

$$\frac{\prod\limits_{r=1}^{\ell} p(x_r||\hat{\theta}_{r-1})}{\sup\limits_{\theta\in H_i} p(x_r||\theta)} \tag{9}$$

where $\hat{\theta}_{r-1}$ is an appropriate estimate for θ, resulting from the first $(r-1)$ observations $x_1, \ldots, x_{r-1}$. If we additionally take into account Pavlov's condition (4'), we know that $\hat{\theta}_{r-1}$ has to be the maximum likelihood estimate for θ.
Under our embedding the denominators in both statistics are equal. Since the maximum likelihood estimates of $(\theta_1, \ldots, \theta_n)$ are the corresponding (adapted) relative frequencies, also the numerators coincide.
Ad b) As mentioned above, Pavlov's condition (4') is now automatically satisfied. The conditions

(1') : Θ is compact and

(2') : the sample space $\mathcal{Y}$ of each draw is compact

are satisfied because in our situation

$$\Theta = \left\{ (\theta_1, \ldots, \theta_n) \,:\, \theta_j = p(y_j) \in [0,1] \text{ and } \sum_{j=1}^{n} \theta_j = 1 \right\}.$$

Therefore, Θ is a closed polyhedron and hence compact. Notice further that $\mathcal{Y} = \{y_1, \ldots, y_n\}$ is finite, and therefore trivially compact.
The function $p(x||\theta)$ is continuous for all (x, θ), as required in Condition (3'). Furthermore, the second demand in Condition (3') is also satiesfied: indeed the Kullback-Leibler information, $\rho(\theta, \varphi) = I\!E_\theta \left(\log \frac{p(x_r||\theta)}{p(x_r||\varphi)} \right)$, is strictly positive for $\theta \neq \varphi$ (see Kullback (1968, p. 14)). □

4.2 Practical aspects and implementation

With the adapted version of the relative frequencies $h^{(r-1)}(x_r)$ for the numerator in $Q_\ell^{(i)}$ no further problems arise.
The denominator of $Q_\ell^{(i)}$ generally can be calculated by a non-linear optimization problem:

$$p(x_1) \cdot \ldots \cdot p(x_\ell) \to \max_{p(\cdot) \in \mathcal{M}_i} \tag{10}$$

subject to the (trivial) linear constraints:

$$p(y_j) \in [0,1], \quad j = 1, \ldots, n, \quad \sum_{j=1}^{n} p(y_j) = 1. \tag{11}$$

For given F-probabilities $P_i(\cdot) = [L^{(i)}(\cdot), U^{(i)}(\cdot)]$, with structures $\mathcal{M}_i$, the conditions $p(\cdot) \in \mathcal{M}_i$ can be transformed with the help of Equation (1) into a system of linear constraints, only using the lower interval limits:

$$p(\bigcup_{j\in J}\{X = y_j\} \geq L^{(i)}(\bigcup_{j\in J}\{X = y_j\}), \qquad \forall J \subseteq \{1, ..., n\}. \tag{12}$$

In the case of an F-PRI, with interval limits $L_j^{(i)}$ and $U_j^{(i)}$ one obtains:

$$p(y_j) = p(\{X = y_j\}) \geq L^{(i)}(\{X = y_j\}) =: L_j^{(i)}, \qquad j = 1, \ldots, n,$$

and

$$p(y_j) = p(\{X = y_j\}) \leq U^{(i)}(\{X = y_j\}) =: U_j^{(i)}, \qquad j = 1, \ldots, n\,.$$

Because of $x_r \in \{y_1, ..., y_n\}$ the objective function (10) can be formulated as follows:

$$p(y_1)^{\ell_1} \cdot \ldots \cdot p(y_n)^{\ell_n} \to \max_{p(\cdot)\in\mathcal{M}_i} \tag{13}$$

with $\ell_j := \sum_{r=1}^{\ell} 1_{\{X_r = y_j\}}$, as the absolute frequency of y_j and $\sum_{j=1}^{n} \ell_j = \ell$.

In general, the maximum can not be given analytically, but this optimization problem can easily be solved by numerical standard procedures.
Taking into account that in Theorem 1 the time points $T^{(i)}, i \in \{0, 1\}$, have to be calculated, it is evident that, as long as $Q_\ell^{(i)} < \alpha_i^{-1}$, both for $i = 0$ and $i = 1$, a further observation has to be drawn. Furthermore, for most of the sequential steps, the following easy-to-handle approximation will suffice: if the objective function in Equation (10) is only roughly estimated at $p_j = L_j^{(i)} (\neq 0)$, the lower limits of the corresponding component of the F-PRI (for $i \in \{0, 1\}$), we obtain:

$$Q_\ell^{(i)} \leq \frac{\prod_{r=1}^{\ell} h^{(r-1)}(x_r)}{{L_1^{(i)}}^{\ell_1} \cdot \ldots \cdot {L_n^{(i)}}^{\ell_n}} =: \overline{Q_\ell}^{(i)}.$$

As long as the upper bound $\overline{Q_\ell}^{(i)}$ is less than α_i^{-1}, for $i = 0$ as well as for $i = 1$, this is also true for $Q_\ell^{(i)}$, and a further observation has to be drawn.

This means, instead of calculating at each step ℓ a non-linear optimization problem, it is sufficient firstly to evaluate $\overline{Q_\ell}^{(i)}$ and to compare it with α_i^{-1}. If both terms are less than α_i^{-1}, a further observation has to be drawn. Only if for at least one i it is greater or equal to α_i^{-1}, we need to start a non-linear optimization routine to check whether the process stops.

4.3 A didactic example

Now let us illustrate the essentials of the procedure in Theorem 1 with an example, which is kept so simple that all calculations can be done by hand.

Example 1*:*

Consider a sample space of three elements:

$$\mathcal{Y} = \{y_1, y_2, y_3\} =: \{1, 2, 3\},$$

and the testing problem

$$H_0 : P(\cdot) = P_0(\cdot) \quad \text{versus} \quad H_1 : P(\cdot) = P_1(\cdot) \tag{14}$$

where $P_0(\cdot)$ and $P_1(\cdot)$ are F-probabilities, with corresponding structures $\mathcal{M}_0$ and $\mathcal{M}_1$, described by the following F-PRIs[3]:

$$\begin{bmatrix} 0.2 & 0.4 \\ 0.2 & 0.4 \\ 0.3 & 0.5 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.4 & 0.6 \\ 0.25 & 0.45 \\ 0.05 & 0.25 \end{bmatrix}. \tag{15}$$

Choosing $\alpha_0 = \alpha_1 = 0.1$, the test statistic $Q_\ell^{(i)}$ leads to a decision if it is greater or equal to 10.

$\ell = 1$:

Now let us assume that the first observation $x_1 = 1$. Then we obtain

$$Q_1^{(i)} = \frac{1}{\sup\limits_{p(\cdot)\in\mathcal{M}_i} p(1)} = \begin{cases} \frac{1}{0.4} = 2.5 & \text{for} \quad i = 0 \\ \frac{1}{0.6} = 1.\overline{6} & \text{for} \quad i = 1\,. \end{cases}$$

[3]Actually, in the case of a sample space with three or less elements, the structure of every F-probability is uniquely determined by the assignments on the singletons; the notions of F-probability and the F-PRI materially coincide.

Because, under H_0 as well as under H_1, the ratio $Q_1^{(i)}$ is less than 10, we have to draw a further observation.

$\ell = 2$:
Let $x_2 = 2$. Now the relative frequency of this observation, resulting from the first observation, would be zero. This is not compatible with $\mathcal{C}(\mathcal{Y})$, because

$$\mathcal{C}(\mathcal{Y}) = \{p(\cdot)\,|\,p(E_j) \in [\underline{C}(E_j), \overline{C}(E_j)] := [\min_{i=0,1} L_j^{(i)}, \min_{i=0,1} U_j^{(i)}]\,,\ j = 1,2,3\}.$$

For our example we obtain

$$\begin{aligned} [\underline{C}(1), \overline{C}(1)] &= [0.20, 0.60] \\ [\underline{C}(2), \overline{C}(2)] &= [0.20, 0.45] \\ [\underline{C}(3), \overline{C}(3)] &= [0.05, 0.50]\,. \end{aligned}$$

So we have to take: $h^{(1)}(x_2) = \underline{C}(2) = 0.2$.
Now $Q_\ell^{(i)}$ results in:

$$Q_2^{(i)} = \frac{1 \cdot 0.2}{\sup\limits_{p(\cdot)\in\mathcal{M}_i} p(1)\cdot p(2)} = \frac{0.2}{\sup\limits_{p(\cdot)\in\mathcal{M}_i} p(1)^1 \cdot p(2)^1 \cdot p(3)^0} \quad .$$

If we only use the upper bound $\overline{Q_2}^{(i)}$ we obtain:

$$\overline{Q_2}^{(i)} = \begin{cases} \frac{0.2}{0.2\cdot 0.2} = 5 < 10 & \text{for} \quad i = 0 \\ \frac{0.2}{0.4\cdot 0.25} = 2 < 10 & \text{for} \quad i = 1, \end{cases}$$

and a further observation has to be drawn.

$\ell = 3$:
Let $x_3 = 1$. Here we obtain a relative frequency of $\frac{1}{2}$, which is compatible with $\mathcal{C}(\mathcal{Y})$ (especially with $[\underline{C}(1), \overline{C}(1)]$), and therefore $h^{(2)}(x_3) = \frac{1}{2}$. By calculating the approximation

$$\overline{Q_3}^{(i)} = \begin{cases} \frac{0.2\cdot 0.5}{0.2^2\cdot 0.2} = 12.5 & \text{for} \quad i = 0 \\ \frac{0.2\cdot 0.5}{0.4^2\cdot 0.25} = 2.5 & \text{for} \quad i = 1, \end{cases}$$

we see that we have to determine the exact value for $Q_3^{(0)}$:

$$Q_3^{(0)} = \frac{0.2 \cdot 0.5}{\sup\limits_{p(\cdot)\in\mathcal{M}_0} p(1)^2 \cdot p(2)^1} = \frac{0.2 \cdot 0.5}{0.4^2 \cdot 0.3} = 2.08\overline{3} < 10.$$

Again we have to continue drawing.

$\ell = 4$:

Let $x_4 = 1$. Here we would obtain a relative frequency of $\frac{2}{3}$ which again is not compatible with $\mathcal{C}(\mathcal{Y})$. We have to restrict $h^{(3)}(x_4)$ to $\overline{C}(1) = 0.6$.

The approximative formula leads to $\overline{Q_4}^{(0)} = 37.5$ and $\overline{Q_4}^{(1)} = 3.75$; therefore we have to determine the exact value of $Q_4^{(0)}$:

$$Q_4^{(0)} = 3.125.$$

$\ell = 7$:

Already with the further observations: $x_5 = 1, x_6 = 1, x_7 = 1$ the procedure stops at:

$$Q_7^{(i)} = \begin{cases} 10.547 > 10 & \text{for} \quad i = 0 \\ \\ 0.794 < 10 & \text{for} \quad i = 1. \end{cases}$$

Now we have: $T^{(0)} = 7$ and $T^{(1)} > 7$ and therefore $N^* = T^{(0)}$. So the decision is for $H_1 : D^* = H_1$.

Let us stay with the example a bit longer and briefly discuss some principle aspects. The structures $\mathcal{M}_0$ and $\mathcal{M}_1$ in the example above can also be connected to a model often used in robust statistics: they can be interpreted as *total variation neighbourhoods* around the centers $p_0(\cdot)$ and $p_1(\cdot)$ with $p_0(y_1) = 0.3$, $p_0(y_2) = 0.3$, $p_0(y_3) = 0.4$ and $p_1(\cdot)$ with $p_1(y_1) = 0.50$, $p_1(y_2) = 0.35$, $p_1(y_3) = 0.15$. From this point of view, $\mathcal{M}_0$ and $\mathcal{M}_1$ are consisting of all classical probabilities which are close to $p_0(\cdot)$ or $p_1(\cdot)$ in the sense that their distance in the total variation norm is less than or equal to 0.1. Then, (14) can be understood as a robust test of the hypotheses $H_0 : p(\cdot) = p_0(\cdot)$ versus $H_1 : p(\cdot) = p_1(\cdot)$, where we de facto test the hypotheses H_i : "$p(\cdot)$ is approximately $p_i(\cdot)$".

An additional fact is worth mentioning: this example also provides a simple counterexample demonstrating that least favourable pairs can not be directly used to construct the optimal test statistic. Huber's (1981) result on the Wald-Wolfowitz optimal testing between interval probabilities can not be transferred to the Kiefer-Weiss criterion considered here.

It can be shown that in this situation $(q_0(\cdot), q_1(\cdot))$ with

$$q_0(y_1) = 0.4, \quad q_0(y_2) = 0.3, \quad q_0(y_3) = 0.3 \quad \text{and}$$

$$q_1(y_1) = 0.4 + \frac{2}{70}, \quad q_1(y_2) = 0.35 - \frac{2}{70}, \quad q_1(y_3) = 0.25$$

is a least favourable pair in the sense of Huber and Strassen (1973). Applying (9) to the hypothesis

$$H_0 : p(\cdot) = q_0(\cdot) \quad \text{and} \quad H_1 : p(\cdot) = q_1(\cdot)$$

derived from the least favourable pair does not lead to the test statistic $Q_\ell^{(i)}$ in Equation (7). The Kiefer-Weiss optimal procedure based on the least favorable pair differs from the optimal test for the interval-valued hypotheses.

5 Concluding remarks

This paper developed a general framework for robust sequential testing of two hypotheses. Using the concept of interval probability we incorporated robustness directly into the formulation of the hypotheses. For these 'cautious hypotheses' we then have, with the Kiefer-Weiss criterion, an unambiguous optimality criterion. This 'ex ante robustification' is very much in the spirit of Huber (1981) who, however, considered a different optimality criterion, namely an extension of the Wald-Wolfowitz criterion to interval probability. For arbitrary interval probabilities on a finite sample space we gave the general form of an Kiefer-Weiss optimal testing procedure and showed how it can be derived in an operational way. Far beyond the robustness considerations originally motivating our research, the generality of our results promises a huge range of potential application. In particular we think of artificial intelligence, where interval probability has shown to be a powerful means to model uncertain expert knowledge.

Several topics of further research suggest themselves. First of all, the procedure proposed evidently needs more detailed investigations from the numerical point of view. Secondly, with respect to application for instance in biometrics, an extension to group sequential tests would be highly desirable. The situation of a fixed sample size at every step is formally contained by appropriately enlarging the underlying sample space $\mathcal{Y}$. Adaptive choice of the sample size at every step is much more difficult; it may even need a complete reconsideration of the issue from the very beginning.

The example above showed that Huber's (1981, p. 273) result can not be directly extended to the Kiefer-Weiss situation: the optimal procedure does not coincide with the optimal test between least favorable elements of the two structures. Therefore, it is still an open question whether the optimal procedure can also be obtained by considering an equivalent testing problem which is easier to be solved.

References

[1] Augustin, T. (2001), Neyman-Pearson testing under interval probability by globally least favorable pairs - Reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, to appear.

[2] Bernard, J. M. (ed.) (2001), Special issue on imprecise probabilities and their applications. *Journal of Statistical Planning and Inference*, to appear.

[3] Christmann, A. (1999), On group sequential tests based on robust location and scale estimators in the two-sample problem. *Computational Statistics 14*, 339–353.

[4] de Cooman, G. and Walley, P. (eds.) (2001), *The Imprecise Probabilities Project.* `http://ippserv.rug.ac.be/`.

[5] Eisenberg, B. (1982), The asymptotic solution of the Kiefer-Weiss problem. *Communications in Statistics - Sequential Ananlysis 1*, 81–88.

[6] Ghosh, B. K. (1970), *Sequential Tests of Statistical Hypotheses.* Addison-Wesley, Reading, Mass.

[7] Huber, P. J. and Strassen, V. (1973), Minimax tests and the Neyman-Pearson lemma for capacities. *Annals of Statistics 1*, 251–263; Correction: 2, 223–224.

[8] Huber, P. J. (1981), *Robust Statistics.* Wiley, New York.

[9] Huffman, M. (1983), An efficient approximate solution to the Kiefer-Weiss problem. *The Annals of Statistics 11*, 306–316.

[10] Irle, A. (1990), *Sequentialanalyse: optimale sequentielle Tests.* Teubner, Stuttgart.

[11] Jennison, C. and Turnbull, B. W. (1999), *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall, Boca Raton.

[12] Kiefer, J. C. and Weiss, L. (1957), Some properties of generalized sequential probability ratio tests. *The Annals of Mathematical Statistics 28*, 57–74.

[13] Kullback, S. (1968), *Information Theory and Statistics.* Dover, New York.

[14] Lai, T. L. (1981), Asymptotic optimality of invariant sequential probability tests. *The Annals of Statistics 9*, 318–333.

[15] Papamarcou, A. and Fine, T. L. (1986), A note on undominated lower probabilities. *The Annals of Probability 14*, 710–723.

[16] Pavlov, I. V. (1991), Sequential procedure of testing composite hypotheses with application to the Kiefer-Weiss problem. *Theory of Probability and its Applications 35*, 280–292.

[17] Pöhlmann, S. and Augustin, T. (2001), A Kiefer-Weiss optimal sign test – Some considerations on a bioequivalence problem from the viewpoint of quality management. In: J. Kunert and G. Trenkler, Eds., *Mathematical Statistics with Applications in Biometry, Festschrift in Honour of Prof. Dr. Siegfried Schach*, Josef Eul, Köln, 179–188.

[18] Quang, P. X. (1985), Robust sequential testing. *The Annals of Statistics 13*, 638–649.

[19] Shafer, G. (1976), *A Mathematical Theory of Evidence.* Princeton University Press, Princeton.

[20] van der Tweel, I., Kaaks, R. and van Noord, P. (1996), Comparison of the single two sided sequential t-test for application in epidemiological studies. *Statistics in Medicine 15*, 2781–2795.

[21] Wald, A. (1947), *Sequential Analysis.* New York, Wiley.

[22] Wald, A. and Wolfowitz, J. (1948), Optimal character of the sequential probability ratio test. *The Annals of Mathematical Statistics 19*, 326–339.

[23] Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall, London.

[24] Weichselberger, K. (2001), *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept.* Physika, Heidelberg.

[25] Weichselberger, K. and Pöhlmann, S. (1990), *A Methodology for Uncertainty in Knowledge Based Systems.* Lecture Notes in Artificial Intelligence 419, Springer, Berlin.

[26] Yager, R. R., Kacprzyk, J. and Fedrizzi, M. (1994), *Advances in the Dempster-Shafer Theory of Evidence.* Wiley, New York.