Combining p-values via averaging

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #21

First posted 20 December 2012 (on arXiv) and 21 November 2017
(as Working Paper). Last revised January 2, 2018.

Project web site:
http://alrw.net

Abstract

This note discusses the problem of multiple testing of a single hypothesis, with a standard goal of combining a number of p-values without making any assumptions about their dependence structure. An old result by Rüschendorf shows that the p-values can be combined by scaling up their arithmetic mean by a factor of 2 (but no smaller factor is sufficient in general). More recent results in mathematical finance show, in addition, that K p-values can be combined by scaling up their geometric mean by a factor of e (for all K) and by scaling up their harmonic mean by a factor of $\ln K$ (for large K). These and other results lead to a generalized version of the Bonferroni–Holm method.

Contents

1 Introduction

Suppose we are testing the same hypothesis using $K \geq 2$ different statistical tests and obtaining p-values $p_1, \ldots, p_K$. How can we combine them into a single p-value?

One of the earliest papers answering this question was Fisher's [5]. However, Fisher's paper assumes that the p-values are independent, whereas we would like to avoid any assumptions besides all p_k, $k = 1, \ldots, K$, being bona fide p-values. Fisher's method has been extended to dependent p-values in, e.g., [3, 13], but the combined p-values obtained in those papers are approximate; in this note we are interested only in precise or conservative p-values.

The simplest way of combining p-values is the Bonferroni method:

$$F(p_1, \ldots, p_K) := K \min(p_1, \ldots, p_K) \tag{1}$$

(when $F(p_1, \ldots, p_K)$ exceeds 1 it can be replaced by 1, but we usually ignore this trivial step). Albeit $F(p_1, \ldots, p_K)$ is a p-value, it has been argued that in many cases it is overly conservative. Rüger [23] extends the Bonferroni method by showing that, for any fixed $k \in \{1, \ldots, K\}$,

$$F(p_1, \ldots, p_K) := \frac{K}{k} p_{(k)} \tag{2}$$

is a p-value, where $p_{(k)}$ is the kth smallest p-value among $p_1, \ldots, p_K$; see [20] for a simpler exposition. Hommel [9] develops this by showing that

$$F(p_1, \ldots, p_K) := \left(1 + \frac{1}{2} + \cdots + \frac{1}{K}\right) \min_{k=1,\ldots,K} \frac{K}{k} p_{(k)} \tag{3}$$

is also a p-value. (Simes [26] improves (3) by removing the first factor on the right-hand side of (3), but he assumes the independence of $p_1, \ldots, p_K$.)

Intuitively, the most natural way to combine K numbers is to average them; my interest in this topic started from this way of combining p-values being used in the method of cross-conformal prediction ([27], (11)), which uses arithmetic mean. None of the functions F in (1), (2), and (3) involves the arithmetic mean $\bar{p} := (p_1 + \cdots + p_K)/K$. Section 2 of this paper draws the reader's attention to a result by Rüschendorf ([24], Theorem 1) showing that $\bar{p}$ is not always a p-value but $2\bar{p}$ is; moreover, the factor of 2 cannot be improved in general. That section also gives a simple proof of the part of Rüschendorf's result stating that $2\bar{p}$ is a bona fide p-value (perhaps conservative). Section 3 considers the case $K = 2$, in which it is very easy to see that the factor of 2 is optimal. Section 4 discusses implications for cross-conformal prediction.

In Section 5 we move on to a generalized notion of mean as axiomatized by Andrei Kolmogorov [12] and adapt various results of robust mathematical finance to combining p-values by averaging in Kolmogorov's wider sense. In particular, to obtain a p-value from given p-values $p_1, \ldots, p_K$, it is sufficient to multiply their geometric mean by e (as noticed by Mattner [18]) and to multiply their harmonic mean by $e \ln K$. More generally, we consider the mean

$M_r(p_1, \dots, p_K)$ defined by $((p_1^r + \dots + p_K^r)/K)^{1/r}$ for $r \in [-\infty, \infty]$; in particular, the results reported in that section cover the Bonferroni method (1), which corresponds to $M_{-\infty}(p_1, \dots, p_K) = \min(p_1, \dots, p_K)$ (see, e.g., [7], (2.3.1)). Section 6 is devoted to the special case $r = -1$ (harmonic mean).

Median is also sometimes regarded as a kind of average. Rüger's (2), applied to $k := \lceil K/2 \rceil$, says that p-values can be combined by scaling up their median by a factor of 2. Therefore, we have the same factor of 2 as in Rüschendorf's [24] result. (Taking $k = \lfloor (K+1)/2 \rfloor = \lceil K/2 \rceil$ is suggested in Section 1.1 of [17].) More generally, the α quantile $p_{(\lceil \alpha K \rceil)}$ becomes a p-value if multiplied by $1/\alpha$.

It is often possible to automatically transform results about multiple testing of a single hypothesis into results about testing multiple hypotheses; the standard procedures are Marcus et al.'s [16] closed testing procedure and its modification by Hommel [10]. In particular, when applied to the Bonferroni method the closed testing procedure gives the well-known method due to Holm [8], which we will refer to as the Bonferroni–Holm method; see, e.g., [10, 11] for its further applications. In Section 7 we briefly discuss a similar application to one of the procedures of Section 5.

Some notation and terminology

If E is a property of elements of a set X, $\mathbf{1}_E : X \to [0, \infty)$ is the indicator function of E: $\mathbf{1}_E(x) = 1$ if x satisfies E and $\mathbf{1}_E(x) = 0$ if not. A function $F : [0,1] \to [0, \infty)$ is *increasing* (resp. *decreasing*) if $F(x_1) \le F(x_2)$ (resp. $F(x_1) \ge F(x_2)$) whenever $x_1 \le x_2$. A function $F : [0,1]^K \to [0, \infty)$ is *increasing* (resp. *decreasing*) if it is increasing (resp. decreasing) in each of its arguments. A set in $[0,1]^K$ is *increasing* (resp. *decreasing*) if its indicator function is increasing (resp. decreasing).

2 Combining p-values by scaled arithmetic mean

A *p-value function* is a random variable P that satisfies

$$\mathbb{P}(P \le \epsilon) \le \epsilon, \quad \forall \epsilon \in [0,1]. \tag{4}$$

The values taken by a p-value function are *p-values* (allowed to be conservative). (In Section 1 the expression "p-value" was loosely used to refer to p-value functions as well.) A *merging function* is an increasing Borel function $F : [0,1]^K \to [0, \infty)$ such that $F(U_1, \dots, U_K)$ is a p-value function, where $U_1, \dots, U_K$ are random variables distributed uniformly on $[0,1]$.

Remark. The requirement that a merging function be Borel does not follow automatically from the requirement that it be increasing: see the remark after Theorem 4.4 in [6] (Theorem 4.4 itself says that every increasing function on $[0,1]^K$ is Lebesgue measurable).

Notice that, for any merging function F, $F(P_1, \dots, P_K)$ is a p-value function whenever $P_1, \dots, P_K$ are p-value functions. Indeed, for each $k \in \{1, \dots, K\}$ we can define a uniformly distributed random variable $U_k \le P_k$ by

$$U_k(\omega) := \mathbb{P}(P_k < P_k(\omega)) + \theta\, \mathbb{P}(P_k = P_k(\omega)), \ \ \omega \in \Omega,$$

where θ is a random variable distributed uniformly on $[0,1]$ and independent of $P_1, \dots, P_k$, and Ω is the underlying probability space extended (if required) to carry such a θ; we then have

$$\mathbb{P}(F(P_1, \dots, P_K) \le \epsilon) \le \mathbb{P}(F(U_1, \dots, U_K) \le \epsilon) \le \epsilon, \quad \forall \epsilon \in [0,1].$$

The following proposition states Rüschendorf's [24] result in terms of merging functions.

Proposition 1. *The function $M : [0,1]^K \to [0,\infty)$ defined by*

$$M(p_1, \dots, p_K) := 2\frac{p_1 + \dots + p_K}{K} \tag{5}$$

is a merging function.

The rest of this section is devoted to a self-contained proof of Proposition 1. A *copular probability measure* is a probability measure on $[0,1]^K$ all of whose marginals are uniform probability measures on $[0,1]$. The *upper copular probability* $\mathbb{C}(E)$ of a Borel set $E \subseteq [0,1]^K$ is defined to be the supremum of $x(E)$, x ranging over the copular probability measures. In terms of $\mathbb{C}$, an increasing Borel function $F : [0,1]^K \to [0,\infty)$ is a merging function if and only if $\mathbb{C}(F \le \epsilon) \le \epsilon$ for all $\epsilon \in [0,1]$. We say that a merging function F is *precise* if $\mathbb{C}(F \le \epsilon) = \epsilon$ for all $\epsilon \in [0,1]$.

For $m \in [0,1]$, define

$$E_m := \left\{ (u_1, \dots, u_K) \in [0,1]^K \;\middle|\; \frac{u_1 + \dots + u_K}{K} \le m \right\} \subseteq [0,1]^K.$$

Proposition 1 can be strengthened: in fact, M is a precise merging function. The original statement of this result, in our notation, is as follows.

Lemma 1 ([24], Theorem 1)**.** *For any $m \in [0,1]$,*

$$\mathbb{C}(E_m) = \min(2m, 1).$$

Remark. In Section 1 we already alluded to an example of a set with a known upper copular probability: the set

$$\left\{ (u_1, \dots, u_K) \in [0,1]^K \;\middle|\; \mathbf{1}_{u_1 \le \alpha} + \dots + \mathbf{1}_{u_K \le \alpha} \ge k \right\},$$

where $\alpha \in [0,1]$ and $k \in \{1, \dots, K\}$, has upper copular probability of $(K/k)\alpha$; this is equivalent to (2) being a precise merging function. Another well-known

example is $H := [0, u_1] \times \cdots \times [0, u_K]$, where $u_1, \ldots, u_K \in [0, 1]$. The upper copular probability of H is $\min(u_1, \ldots, u_K)$. This is known as one of the Fréchet–Hoeffding bounds in the theory of copulas, and is known to be a copula (see, e.g., [21], Exercise 2.35). Lemma 1 is one more example of this kind. Lemma 2 below will give a simple characterization of upper copular probability in the easy case $K = 2$.

Given Lemma 1, the proof of Proposition 1 is trivial: for any $\epsilon \in [0, 1]$,

$$\mathbb{P}(M(U_1, \ldots, U_K) \le \epsilon) = \mathbb{P}\left(\frac{U_1 + \ldots + U_K}{K} \le \frac{\epsilon}{2}\right) = \min(\epsilon, 1) \le \epsilon.$$

Notice that for the proof of Proposition 1 we only need the inequality $\le$ in the lemma. The rest of this section is devoted to the proof of this inequality.

Let $K[0, 1]$ be the sum of K disjoint copies of the interval $[0, 1]$. A (somewhat arbitrary) concrete representation of $K[0, 1]$ is the set $\cup_{k=1}^{K} [2(k-1), 2k-1]$. We will sometimes use the notation $K[0, 1]_k$ for the kth copy of $[0, 1]$ in $K[0, 1]$; so that $K[0, 1]_k = [2(k-1), 2k-1]$ in the concrete representation (but $K[0, 1]_k$ is always identified with $[0, 1]$, via the bijection $u \mapsto u - 2(k-1)$ in the concrete representation). If x is a measure on $[0, 1]^K$, we define x_k to be the projection of x onto the kth coordinate of $[0, 1]^K$,

$$x_k(E) := x([0, 1]^{k-1} \times E \times [0, 1]^{K-k}), \quad E \subseteq [0, 1] \text{ is Borel},$$

and we define Ax to be the measure on $K[0, 1]$ that coincides with x_k on $K[0, 1]_k$ (so that Ax's total mass is K when x is a probability measure). The *uniform measure* on $K[0, 1]$ is the measure on the Borel σ-algebra on $K[0, 1]$ that coincides with the uniform probability measure on each of its components $K[0, 1]_k$ (so that Ax is the uniform measure on $K[0, 1]$ if and only if x is a copular probability measure).

Lemma 1 can be interpreted as a statement about the following infinite-dimensional problem of linear programming (standard in optimal transport [22, Theorem 1.1.1]):

$$cx \to \sup \quad \text{subject to} \quad Ax = b, \ \ x \ge 0, \tag{6}$$

where c is the indicator function of the set E_m, the variable x ranges over all measures on $[0, 1]^K$, cx is understood to be $\int c \mathrm{d}x$, Ax is as defined above, and b is the uniform measure on $K[0, 1]$. The condition $x \ge 0$ is an embellishment without a formal meaning (and emphasizes the fact that measures take only nonnegative values). Lemma 1 says that the value of (6) is $2m$ when $m \le 1/2$.

The formal dual problem to (6) is

$$\lambda b \to \inf \quad \text{subject to} \quad \lambda A \ge c, \tag{7}$$

which we will interpret as follows: the dual variable λ ranges over all Borel functions on $K[0, 1]$, λb is understood to be $\int \lambda \mathrm{d}b$, λA is the function on $[0, 1]^K$ defined by

$$(\lambda A)(u_1, \ldots, u_K) := \lambda_1(u_1) + \cdots + \lambda_K(u_K),$$

where λ_k is the restriction of λ to $K[0,1]_k$, and $\geq$ stands, as usual, for the pointwise inequality.

It is easy to see that the operators $x \mapsto Ax$ and $\lambda \mapsto \lambda A$ are dual, in the sense that $(\lambda A)x = \lambda(Ax)$:

$$(\lambda A)x = \int \lambda_1 \mathrm{d}x_1 + \cdots + \int \lambda_K \mathrm{d}x_K = \int \lambda \mathrm{d}Ax = \lambda(Ax).$$

(This justifies using the same letter for both operators.) As usual, the value of the original problem (6) does not exceed the value of the dual problem (7): indeed, if x satisfies the constraints in (6) and λ satisfies the constraint in (7),

$$cx \leq (\lambda A)x = \lambda(Ax) = \lambda b.$$

Now we have all components for the proof of the inequality $\leq$ in Lemma 1.

Proof of the inequality $\leq$ in Lemma 1. It suffices to prove that the value of the dual problem (7) does not exceed $2m$. Define $\lambda : K[0,1] \to [0,\infty)$ by $\lambda_k(u) := (2/K - u/Km)^+$ for all $k \in \{1,\dots,K\}$, where t^+ is t if $t \geq 0$ and 0 otherwise. (In other words, assuming $m \leq 1/2$, $\lambda_k : [0,1] \to [0,\infty)$ is the function with the subgraph of the smallest area among all functions that are linear when positive and whose graph passes through $(m, 1/K)$.) Since

$$\lambda b = \int \lambda_1 + \cdots + \int \lambda_K \leq 2m$$

(with $=$ in place of the last $\leq$ when $m \leq 1/2$), it remains to prove that the constraint in (7) is satisfied. This is accomplished by the following chain of inequalities:

$$\begin{aligned}
\lambda A(u_1,\dots,u_K) &= \sum_{k=1}^{K} \left(\frac{2}{K} - \frac{u_k}{Km}\right)^+ \geq \left(\sum_{k=1}^{K} \left(\frac{2}{K} - \frac{u_k}{Km}\right)\right)^+ \\
&= (2 - (u_1 + \cdots + u_K)/Km)^+ \geq \mathbf{1}_{2-(u_1+\cdots+u_K)/Km \geq 1} \\
&= \mathbf{1}_{(u_1+\cdots+u_K)/K \leq m} = c(u_1,\dots,u_K). \qquad \square
\end{aligned}$$

3 Case $K = 2$

In the case $K = 2$ upper copular probability admits a simple characterization.

Lemma 2. *If a nonempty Borel set $E \subseteq [0,1]^2$ is decreasing, its upper copular probability is*

$$\mathbb{C}(E) = \min\left(\inf\left\{u_1 + u_2 \,\middle|\, (u_1,u_2) \in [0,1]^2 \setminus E\right\}, 1\right). \tag{8}$$

Lemma 2 implies that the factor 2 in (5) is optimal for $K = 2$: indeed, it shows that the function $M_\alpha(p_1,p_2) := \alpha(p_1 + p_2)$, where $\alpha > 0$, satisfies $\mathbb{C}(M_\alpha \leq \epsilon) = \min(\epsilon/\alpha, 1)$ for all $\epsilon \in [0,1]$; therefore, M_α is a merging function if and only if $\alpha \geq 1$. It is clear that $M_1 = M$ is the only precise merging function among M_α.

Proof of Lemma 2. Let E be a nonempty decreasing Borel set in $[0,1]^2$; suppose $\mathbb{C}(E)$ is strictly less than the right-hand side of (8). Let t be any number strictly between $\mathbb{C}(E)$ and the right-hand side of (8). The copular probability measure concentrated on

$$[(t,0),(0,t)] \cup [(t,t),(1,1)] \tag{9}$$

has a value of at least t on E since E contains $[(t,0),(0,t)]$. Therefore, $\mathbb{C}(E) \geq t$. This contradiction proves the inequality $\geq$ in (8).

As for the opposite inequality, we will check

$$\mathbb{C}(E) \leq \inf\left\{u_1 + \cdots + u_K \,\middle|\, (u_1,\dots,u_K) \in [0,1]^K \setminus E\right\}$$

for an arbitrary $K \geq 2$. Let us assume that E does not contain the set of all $(u_1,\dots,u_K)$ with $u_1 + \cdots + u_K = 1$ (the case when it does is trivial). Choose $\epsilon > 0$ and $(p_1,\dots,p_K) \in [0,1]^K \setminus E$ such that $t := p_1 + \cdots + p_K \in [\epsilon, 1]$ and E contains all $(u_1,\dots,u_K) \in [0,1]^K$ satisfying $u_1 + \ldots + u_K = t - \epsilon$. Since E is decreasing, we have

$$E \subseteq \bigcup_{k=1}^{K} \left\{(u_1,\dots,u_K) \in [0,1]^K \,\middle|\, u_k \leq p_k\right\},$$

and the subadditivity of $\mathbb{C}$ further implies

$$\begin{aligned}\mathbb{C}(E) &\leq \sum_{k=1}^{K} \mathbb{C}\left(\left\{(u_1,\dots,u_K) \in [0,1]^K \,\middle|\, u_k \leq p_k\right\}\right) \\ &= \sum_{k=1}^{K} p_k = t \leq \inf\left\{u_1 + \cdots + u_K \,\middle|\, (u_1,\dots,u_K) \in [0,1]^K \setminus E\right\} + \epsilon.\end{aligned}$$

It remains to notice that ϵ can be chosen arbitrarily small. □

A merging function F_1 *dominates* a merging function F_2 if $F_1 \leq F_2$. The following corollary of Lemma 2 says that, in the case $K = 2$, the merging function (5) is dominated by all precise merging functions. This is not true when $K > 2$: for example, for the Bonferroni function (1) we have $M(p,\dots,p) = 2p < Kp = F(p,\dots,p)$. (The Bonferroni merging function being precise, $\mathbb{C}(F \leq \epsilon) = \epsilon$, is witnessed by the probability distribution on $[0,1]^K$ that is uniform inside the K-dimensional rectangles $[0,\epsilon/K] \times [\epsilon/K,1] \times \cdots \times [\epsilon/K,1]$,..., $[\epsilon/K,1] \times \cdots \times [\epsilon/K,1] \times [0,\epsilon/K]$, $[\epsilon/K,1] \times \cdots \times [\epsilon/K,1]$ and assigns to them probabilities ϵ/K,..., ϵ/K, $1-\epsilon$, respectively.)

Corollary 1. *When $K = 2$, any precise merging function dominates M.*

Proof. Let $F : [0,1]^2 \to [0,\infty)$ be a merging function that does not dominate M. Choose $(u_1,u_2) \in [0,1]^2$ such that $F(u_1,u_2) > u_1 + u_2$ and choose $\epsilon \in (u_1 + u_2, F(u_1,u_2))$. Since $\{F \leq \epsilon\}$ does not contain (u_1,u_2), we have $\mathbb{C}(F \leq \epsilon) \leq u_1 + u_2 < \epsilon$, and so F is not precise. □

We can easily deduce from Lemma 2 that the merging function (1) is precise for an even K.

Corollary 2. *If K is even, the merging function* (1) *is precise.*

Proof. To check that $\mathbb{C}(M \le \epsilon) = \epsilon$ for $\epsilon \in [0,1]$, consider the product of $K/2$ copular probability measure on $[0,1]^2$ concentrated on (9) for $t := \epsilon$. □

To drop the assumption that K is even in Corollary 2, we can follow [22, Example 3.6.4] or [25, Fig. 1]: first consider the case $K = 3$, where we can define

$$\begin{aligned} p_1 &:= U \\ p_2 &:= U + 0.5\epsilon\,\mathbf{1}_{[0,\epsilon/2]}(U) - 0.5\epsilon\,\mathbf{1}_{(\epsilon/2,\epsilon]}(U) \\ p_3 &:= U - 3\,\mathbf{1}_{[0,\epsilon]}\,U + \epsilon\,\mathbf{1}_{[0,\epsilon/2)}(U) + 2\epsilon\,\mathbf{1}_{(\epsilon/2,\epsilon]}(U) \end{aligned} \tag{10}$$

with U is distributed uniformly on $[0,1]$, so that $\mathbb{P}(M(p_1,p_2,p_3) \le \epsilon) = \epsilon$; and then represent an odd $K \ge 2$ as sum of 2s and 3.

4 Connections with conformal prediction

This section assumes the knowledge of basic definitions of conformal prediction (see, e.g., [28]). The method of cross-conformal prediction, already mentioned in Section 1, defines putative p-values as, essentially, the arithmetic means of the p-values computed from K folds (cf. [27], (11)). The experiments reported in [27] confirm the empirical validity of the arithmetic means in the sense of approximately satisfying (4), although Appendix A of [27] gives examples showing that in general (4) may be violated. In their recent paper [14] Linusson et al. give examples of particularly unstable randomized underlying algorithms which make the arithmetic means used in the method of conformal prediction violate (4) in their experimental studies. This makes the phenomenon discussed theoretically in Appendix A of [27] a practical issue. Proposition 1 shows that there is a limit to the degree to which the validity of cross-conformal predictors can be violated, even for the most unstable underlying algorithms: namely, we always have

$$\mathbb{P}(P \le \epsilon) \le 2\epsilon, \quad \forall \epsilon \in [0,1], \tag{11}$$

where P is the arithmetic mean of p-values.

In fact, cross-conformal predictors output only approximate arithmetic means of p-values; the precise expression ([27], (11)) for their putative p-values p is

$$p = \bar{p} + \frac{K-1}{l+1}(\bar{p} - 1), \tag{12}$$

where l is the size of the training set, K is the number of equal-sized folds, and $\bar{p}$ is the arithmetic mean of the p-values computed from the different folds. Expressing $\bar{p}$ via p,

$$\bar{p} = \frac{l+1}{l+K}\left(p + \frac{K-1}{l+1}\right)$$

and using Proposition 1, we obtain

$$\mathbb{P}\left(\frac{l+1}{l+K}\left(p+\frac{K-1}{l+1}\right)\le\delta\right)\le 2\delta$$

for any $\delta>0$. Rewriting the last inequality as

$$\mathbb{P}\left(p\le\frac{l+K}{l+1}\delta-\frac{K-1}{l+1}\right)\le 2\delta$$

and setting

$$\epsilon:=\frac{l+K}{l+1}\delta-\frac{K-1}{l+1},$$

we finally obtain

$$\mathbb{P}\left(p\le\epsilon\right)\le 2\epsilon+2\frac{K-1}{l+K}(1-\epsilon) \tag{13}$$

for any $\epsilon>0$. The last equation, (13), is the precise version of (11) in the case where p is the putative p-value output by a cross-conformal predictor; we can see that (13) is marginally weaker.

In the rest of this section we will ignore the difference between the expression (12) and the arithmetic mean $\bar{p}$, and it will be convenient to generalize the definition of $\bar{p}$ as $\bar{p}:=\int p_\theta P(d\theta)$, where p_θ is now a family of uniformly distributed random variables indexed by, and measurable in, $\theta\in\Theta$, where Θ is a measurable space and P is a probability measure on Θ. We will refer to the uniform distribution on $[0,1]$ as U. Let us say that a random variable $\xi\in[0,1]$ is *stochastically less variable than* U if $\mathbb{E}(h(\xi))\le\int_0^1 h(u)du$ for any convex function h on $[0,1]$. The following result is a restatement of Theorem 1 in [19].

Proposition 2. *It is always true that $\bar{p}$ is stochastically less variable than U.*

Proof. For any convex h,

$$\begin{aligned}
\mathbb{E}(h(\bar{p})) &= \mathbb{E}\left(h\left(\int p_\theta P(d\theta)\right)\right) && \text{by definition}\\
&\le \mathbb{E}\left(\int h(p_\theta)P(d\theta)\right) && \text{by Jensen's inequality}\\
&= \int \mathbb{E}(h(p_\theta))P(d\theta) && \text{by Fubini's theorem}\\
&= \int\left(\int_0^1 h(u)du\right)P(d\theta) && \text{as } p_\theta\sim U\\
&= \int_0^1 h(u)du. && \square
\end{aligned}$$

Meng ([19], Corollary 1) proves the following stronger version of Rüschendorf's result, applicable to any random variable that is stochastically less variable than U, not just $\bar{p}$.

Proposition 3. *If G is the distribution function of a random variable that is stochastically less variable than U, then, for all $\epsilon \in [0,1]$,*

$$\epsilon - \sqrt{\epsilon^2 - 2\int_0^\epsilon G(u)du} \le G(\epsilon) \le \epsilon + \sqrt{\epsilon^2 - 2\int_0^\epsilon G(u)du} \le 2\epsilon \tag{14}$$

(in particular, the expression under the square root signs is always nonnegative).

For example, (14) says that if at some ϵ the property of validity of $\bar{p}$ as a p-value is violated to the greatest degree, $\mathbb{P}(\bar{p} \le \epsilon) = 2\epsilon$, then, at all smaller significance levels $\epsilon' < \epsilon$, $\bar{p}$ must be extremely conservative, $\mathbb{P}(\bar{p} \le \epsilon') = 0$.

Applying Proposition 2 to $h(u) := u$ and $h(u) := -u$, we obtain $\mathbb{E}\,\bar{p} = 1/2$. Therefore, $\bar{p}$ has the right expectation but is more concentrated around it than U is. There are two particularly interesting special cases that have been discussed in literature. One is where $\bar{p}$ is a constant (necessarily $1/2$): see, e.g., [22, Example 3.6.4] (we already alluded to this result at the end of Section 3; now the ϵ in (10) should be set to 1). The other is where $\bar{p}$ is distributed as the sum K independent random variables each distributed uniformly on $[0,1]$; the resulting distribution of $\bar{p}$, which we now again assume to be $(p_1 + \cdots + p_K)/K$, is then known as the Bates distribution [1] and is concentrated around $1/2$ for large K. In the context of cross-conformal prediction, this corresponds to the extreme case where the p-value for each fold is computed using a randomized algorithm that does not pay any attention to the data. In general, we need to be careful when using randomized algorithms as underlying algorithms in cross-conformal prediction [14].

5 More general averages

So far we have only considered averaging in the sense of arithmetic mean. A much more general notion of averaging, axiomatized by Kolmogorov [12], is

$$M_\phi(p_1, \dots, p_K) := \psi\left(\frac{\phi(p_1) + \cdots \phi(p_K)}{K}\right), \tag{15}$$

where $\phi : [0,1] \to [-\infty, \infty]$ is a continuous strictly monotonic function and ψ is its inverse (with the domain $\phi([0,1])$). For example, arithmetic mean corresponds to the identity function $\phi(p) = p$, geometric mean corresponds to $\phi(p) = \ln p$, and harmonic mean corresponds $\phi(p) = 1/p$. In this section we will extend Proposition 1 to the general class of means (15).

The main results of this section, Theorems 1 and 2, will be very simple corollaries of known results in the field of mathematical finance dealing with robust versions of the Value at Risk (VaR); the origin of this field lies in a problem posed by Kolmogorov (see, e.g., [15]).

Theorem 1. *Suppose a continuous strictly monotonic $\phi : [0,1] \to [-\infty, \infty]$ is integrable, i.e., $\int_0^1 |\phi(u)|\,du < \infty$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(M_\phi(p_1, \dots, p_K) \le \psi\left(\frac{1}{\epsilon}\int_0^\epsilon \phi(u)du\right)\right) \le \epsilon. \tag{16}$$

Proof. Without loss of generality we can, and will, assume that ϕ is strictly increasing. Indeed, if ϕ is strictly decreasing, we can redefine $\phi := -\phi$ and $\psi(u) := \psi(-u)$ and notice that the statement of Proposition 1 for new ϕ and ψ will imply the analogous statement for the original ϕ and ψ. In the rest of this proof, ϕ and, therefore, ψ are assumed monotonically increasing.

Theorem 1 of [2] implies

$$\sum_{k=1}^{K} \mathrm{LTVaR}_\epsilon(X_k) \le \mathrm{VaR}_\epsilon(S) \le \sum_{k=1}^{K} \mathrm{TVaR}_\epsilon(X_k), \tag{17}$$

where $S := X_1 + \cdots + X_K$, all X_k and S are assumed to have continuous and strictly decreasing distribution functions and to possess the first moment,

$$\begin{aligned} \mathrm{TVaR}_\epsilon(X) &:= \frac{1}{1-\epsilon} \int_\epsilon^1 \mathrm{VaR}_u(X) du, \\ \mathrm{LTVaR}_\epsilon(X) &:= \frac{1}{\epsilon} \int_0^\epsilon \mathrm{VaR}_u(X) du, \\ \mathrm{VaR}_\epsilon(X) &\in F_X^{-1}(\epsilon) \end{aligned}$$

(the set $F_X^{-1}(\epsilon)$ is a singleton and, therefore, $\mathrm{VaR}_\epsilon(X)$ is well-defined), and F_X is the distribution function of X. We will only use the lower bound in (17), applying it to the random variables $X_k := \phi(p_k)$. They possess the first moment since we assumed $\int_0^1 |\phi(u)|\, du < \infty$.

We have

$$F_{X_k}(x) = \mathbb{P}(\phi(p_k) \le x) = \mathbb{P}(p_k \le \psi(x)) = \psi(x)$$

and, therefore, $\mathrm{VaR}_u(X_k) = \psi^{-1}(u) = \phi(u)$. This gives

$$\mathrm{LTVaR}_\epsilon(X_k) = \frac{1}{\epsilon} \int_0^\epsilon \mathrm{VaR}_u(X_k) du = \frac{1}{\epsilon} \int_0^\epsilon \phi(u) du.$$

Plugging this into the lower bound in (17) gives

$$\frac{K}{\epsilon} \int_0^\epsilon \phi(u) du \le F_S^{-1}(\epsilon),$$

where $S = \phi(p_1) + \cdots + \phi(p_K)$ and the singleton $F_S^{-1}(\epsilon)$ is identified with its only element. The last inequality can be rewritten as

$$\mathbb{P}\left(S \le \frac{K}{\epsilon} \int_0^\epsilon \phi(u) du \right) \le \epsilon,$$

which is equivalent to

$$\mathbb{P}\left(\psi(S/K) \le \psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) du \right) \right) \le \epsilon,$$

which in turn is equivalent to (16). □

A special case of (15) is

$$M_r(p_1, \ldots, p_K) := \left(\frac{p_1^r + \cdots + p_K^r}{K} \right)^{1/r},$$

where $r \in \mathbb{R} \setminus \{0\}$ and the following standard conventions are used: $0^c := \infty$ for $c < 0$, $0^c := 0$ for $c > 0$, $\infty + c := \infty$ for $c \in \mathbb{R} \cup \{\infty\}$, and $\infty^c := 0$ for $c < 0$. The case $r = 0$ (considered in [18]) is treated separately (as the limit as $r \to 0$):

$$M_0(p_1, \ldots, p_K) := \exp\left(\frac{\ln p_1 + \cdots \ln p_K}{K} \right),$$

where, as usual, $\ln 0 := -\infty$, $-\infty + c := -\infty$ for $c \in \mathbb{R} \cup \{-\infty\}$, and $\exp(-\infty) := 0$. It is also natural to set

$$M_\infty(p_1, \ldots, p_K) := \max(p_1, \ldots, p_K),$$
$$M_{-\infty}(p_1, \ldots, p_K) := \min(p_1, \ldots, p_K).$$

The most important special cases of M_r are perhaps those corresponding to $r = -\infty$ (minimum), $r = -1$ (harmonic mean), $r = 0$ (geometric mean), $r = 1$ (arithmetic mean), and $r = \infty$ (maximum); the cases $r \in \{-1, 0, 1\}$ are known as Platonic means.

Corollary 3. *Let $r \in (-1, \infty]$. Then the function*

$$M(p_1, \ldots, p_K) := (r+1)^{1/r} M_r(p_1, \ldots, p_K)$$

is a merging function.

The expression $(r+1)^{1/r}$ is understood to be $e = \lim_{r \to 0}(r+1)^{1/r}$ when $r = 0$ and $1 = \lim_{r \to \infty}(r+1)^{1/r}$ when $r = \infty$.

Proof. Evaluating the term

$$\psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) du \right) \tag{18}$$

in (16), we obtain:

- when $r = 0$,

$$\psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) du \right) = \exp\left(\frac{1}{\epsilon} \int_0^\epsilon \ln u du \right) = \exp\left(\frac{1}{\epsilon} \left[u \ln u - u \right]_0^\epsilon \right) = \epsilon / e;$$

- when $r \neq 0$,

$$\psi\left(\frac{1}{\epsilon} \int_0^\epsilon \phi(u) du \right) = \left(\frac{1}{\epsilon} \int_0^\epsilon u^r du \right)^{1/r} = \left(\frac{1}{\epsilon} \left[\frac{u^{r+1}}{r+1} \right]_0^\epsilon \right)^{1/r} = (r+1)^{-1/r} \epsilon.$$

□

The condition $r > -1$ in Corollary 3 ensures that the term (18) is finite, and also that the condition $\int_0^1 |\phi(u)|\, du < \infty$ in Theorem 1 is satisfied. For arithmetic mean it gives the same result as before: namely, Proposition 1 is a special case of Corollary 3 corresponding to $r = 1$. However, the condition rules out harmonic mean (for which $r = -1$) and the minimum ($r = -\infty$). The next simple corollary of another known result will cover those cases as well.

Theorem 2. *Suppose $\phi : [0,1] \to [-\infty, \infty]$ is a strictly decreasing continuous function satisfying $\phi(0) = \infty$. Then, for any $\epsilon \in [0,1]$ such that $\phi(\epsilon) \geq 0$,*

$$\mathbb{P}\left(M_\phi(p_1, \ldots, p_K) \leq \epsilon\right) \leq \inf_{t \in (0, \phi(\epsilon)]} \frac{\int_{\phi(\epsilon)-t}^{\phi(\epsilon)+(K-1)t} \psi(u) du}{t}. \tag{19}$$

Proof. We will apply Theorem 4.2 of [4] in our situation where the function ϕ (and, therefore, ψ as well) in (15) is decreasing. Letting $X_k := \phi(p_k)$ and using the notation m_+ (used in Theorem 4.2 of [4]), we have, by the definition of m_+,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^K X_k < s\right) &\geq m_+(s), \\
\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^K \phi(p_k) < s/K\right) &\geq m_+(s), \\
\mathbb{P}\left(M_\phi(p_1, \ldots, p_K) > \psi(s/K)\right) &\geq m_+(s), \\
\mathbb{P}\left(M_\phi(p_1, \ldots, p_K) \leq \psi(s/K)\right) &\leq 1 - m_+(s).
\end{aligned} \tag{20}$$

The lower bound on $m_+(s)$ given in Theorem 4.2 of [4] involves $1 - F(x)$, where F is the common distribution function of X_k, and in our current context we have:

$$1 - F(x) = \mathbb{P}(X_k > x) = \mathbb{P}(\phi(p_k) > x) = \mathbb{P}(p_k < \psi(x)) = \psi(x). \tag{21}$$

The last inequality and chain of equalities in combination with Theorem 4.2 of [4] give

$$\mathbb{P}\left(M_\phi(p_1, \ldots, p_K) \leq \psi(s/K)\right) \leq K \inf_{r \in [0, s/K)} \frac{\int_r^{s-(K-1)r} \psi(x) dx}{s - Kr}.$$

Setting $\epsilon := \psi(s/K) \in [\psi(\infty), \psi(0)]$ (so that it is essential that $\psi(\infty) = 0$), we obtain, using $s = K\phi(\epsilon)$,

$$\mathbb{P}\left(M_\phi(p_1, \ldots, p_K) \leq \epsilon\right) \leq K \inf_{r \in [0, \phi(\epsilon))} \frac{\int_r^{K\phi(\epsilon)-(K-1)r} \psi(x) dx}{(\phi(\epsilon) - r)K}.$$

Setting $t := \phi(\epsilon) - r$ and renaming x to u, this can be rewritten as (19). □

As $t \to 0$, the upper bound in (19) is not informative since, for $t \approx 0$,

$$\frac{\int_{\phi(\epsilon)-t}^{\phi(\epsilon)+(K-1)t} \psi(u)du}{t} \approx \frac{Kt\psi(\phi(\epsilon))}{t} = K\epsilon,$$

which is dominated by the Bonferroni bound. On the other hand, the upper bound is informative when $t = \phi(\epsilon)$ provided the integral is convergent. For example, we have the following corollary.

Corollary 4. *For $r < -1$, we have, for any $\epsilon \in [0,1]$,*

$$\mathbb{P}\left(M_r(p_1,\dots,p_K) \le \epsilon\right) \le \frac{r}{r+1} K^{1+1/r}\epsilon. \tag{22}$$

Proof. By Theorem 2 applied to $\phi(u) := u^r$, $r < -1$, we have:

$$\begin{aligned}\mathbb{P}\left(M_\phi(p_1,\dots,p_K) \le \epsilon\right) &\le \frac{\int_0^{K\phi(\epsilon)} \psi(u)du}{\phi(\epsilon)} = \frac{\int_0^{K\epsilon^r} u^{1/r}du}{\epsilon^r} \\ &= \epsilon^{-r}\left[\frac{u^{1/r+1}}{1/r+1}\right]_0^{K\epsilon^r} = \frac{r}{r+1}K^{1+1/r}\epsilon. \qquad \square\end{aligned}$$

Corollary 4 includes the Bonferroni bound (1): for $r := -\infty$, the right-hand side of (22) is $K\epsilon$. On the other hand, it does not cover the case $r = -1$ directly, although it easily implies it.

Corollary 5. *Suppose $K > 2$. For any $\epsilon \in [0,1]$,*

$$\mathbb{P}\left(M_{-1}(p_1,\dots,p_K) \le \epsilon\right) \le (e\ln K)\epsilon. \tag{23}$$

Proof. Let us find the smallest value of the coefficient $\frac{r}{r+1}K^{1+1/r}$ in front of ϵ in (22). Setting the derivative in r of the logarithm of this coefficient to 0, we obtain a linear equation whose solution is

$$r = \frac{\ln K}{1-\ln K}. \tag{24}$$

Plugging this into the coefficient gives

$$\frac{\frac{\ln K}{1-\ln K}}{\frac{\ln K}{1-\ln K}+1} K^{1+\frac{1-\ln K}{\ln K}} = \frac{\ln K}{\ln K + 1 - \ln K} K^{\frac{1}{\ln K}} = e\ln K.$$

It remains to notice that r defined by (24) satisfies $r < -1$ and apply the inequality $M_r \le M_{-1}$ ([7], Theorem 16). $\square$

We stated Corollary 5 in a weakened form because of the importance of harmonic mean as one of the three Platonic means, but the proof shows that in fact M_{-1} in (23) can be replaced by $M_r \le M_{-1}$ for

$$r := \frac{\ln K}{1-\ln K} < -1$$

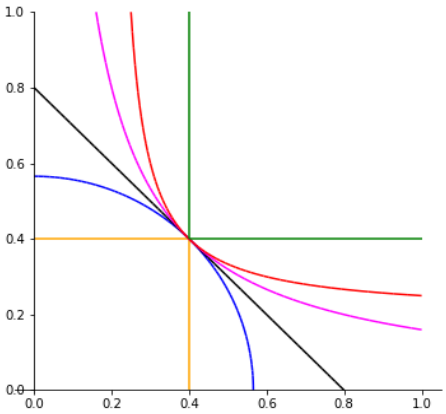

Figure 1: The parts of the spheres $M_r = \epsilon$ for $K = 2$, $\epsilon = 0.4$, and $r \in \{\infty, 2, 1, 0, -1, -\infty\}$ inside the square $[0,1]^2$. The innermost (orange) sphere is $M_\infty = \epsilon$, the next one is $M_2 = \epsilon$ (blue, Euclidean sphere, corresponding to quadratic mean), then $M_1 = \epsilon$ (black and straight, corresponding to arithmetic mean), then $M_0 = \epsilon$ (magenta, corresponding to geometric mean), then $M_{-1} = \epsilon$ (red, corresponding to harmonic mean), and finally $M_{-\infty} = \epsilon$ (outermost and green, corresponding to minimum).

(cf. (24); the inequality assumes $K > 2$).

The discussion in Section 3 shows that the bounds in Corollaries 3, 4, and 5 are far from optimal in the case $K = 2$. Those corollaries can be restated as saying that the function

$$M(p_1, \dots, p_K) := \begin{cases} (r+1)^{1/r} M_r(p_1, \dots, p_K) & \text{if } r > -1 \\ (e \ln K) M_r(p_1, \dots, p_K) & \text{if } r = -1 \\ \frac{r}{r+1} K^{1+1/r} M_r(p_1, \dots, p_K) & \text{if } r < -1 \end{cases}$$

(the first case becoming $eM_0(p_1, \dots, p_K)$ when $r = 0$) is a merging function. Applying Lemma 2 to the balls $E := \{(p_1, p_2) \mid M_r(p_1, p_2) \le \epsilon\}$ (whose boundaries are shown in Figure 1), we can see that in fact the smaller function

$$M(p_1, p_2) := \begin{cases} 2^{1/r} M_r(p_1, p_2) & \text{if } r \ge 1 \\ 2 M_r(p_1, p_2) & \text{if } r \le 1 \end{cases} \tag{25}$$

is a merging function when $K = 2$.

6 Harmonic mean

The tight version of the bound (22) can be derived from (and is essentially given by) Theorem 3.4 in [29], but in this version of the paper we provide details only for the case of harmonic mean, $r = -1$; this will allow us to deduce an asymptotically tight version of Corollary 5: see Corollary 6 below.

Proposition 4. *For any $\epsilon \in [0,1]$,*

$$\mathbb{P}\left(M_{-1}(p_1,\dots,p_K) \le \alpha_K \epsilon\right) \le \epsilon, \tag{26}$$

where α_K is the unique solution of the equation

$$\ln\left(\frac{2(K-1)}{1+\sqrt{1-4\alpha(K-1)/K}} - (K-1)\right) = \frac{K-2+K\sqrt{1-4\alpha(K-1)/K}}{2(K-1)\alpha} \tag{27}$$

in $\alpha \in (0, \frac{K}{4(K-1)}]$. For each $K > 1$, the constant α_K in (26), as defined by (27), is optimal.

Notice that for $K := 2$ Proposition 4 gives us the value $\alpha_K = 0.5$ that agrees with (25) for $r := -1$.

Proof. First let us check that equation (27) is well-defined and has a unique solution. Since $\alpha \in (0, \frac{K}{4(K-1)}]$, the square roots in (27) are well-defined, and it is clear that the expression under the ln sign is positive. It is obvious that the left-hand side of (27) is strictly increasing in α, and differentiating the right-hand side of (27) shows that it is decreasing in α (the derivative is the sum of two nonpositive terms). For $\alpha = 0$ the left hand side is less than the right-hand side ($-\infty < \infty$), and for $\alpha = \frac{K}{4(K-1)}$ the left hand side is equal to or greater than the right-hand side ($\ln(K-1) \ge 2 - \frac{4}{K}$ is true for all integer $K \ge 2$). Therefore, α_K is well-defined.

We proceed as at the beginning of the proof of Theorem 2 setting $\phi(u) := 1/u$, $u \in [0,1]$, and $\psi(u) := 1/u$, $u \in [1,\infty]$. Both ϕ and ψ are decreasing. Setting $X_k := \phi(p_k)$ we obtain random variables whose distribution function is $F = 1 - \psi$ (cf. (21)) or, in more detail,

$$F(x) = \begin{cases} 1 - 1/x & \text{if } x \ge 1 \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

The density $1/x^2$ is decreasing over the support $[1,\infty)$ of X_k. Introducing a new variable $\epsilon := \psi(s/K) = K/s$ in (20), as before, we obtain

$$\mathbb{P}\left(M_\phi(p_1,\dots,p_K) \le \epsilon\right) \le 1 - m_+(K/\epsilon). \tag{29}$$

By Theorem 3.4 in [29], $m_+(K/\epsilon) = \Phi^{-1}(K/\epsilon)$, where $\Phi(a) = H_a(c(a))$ (provided $c(a) > 0$, which will be the case),

$$H_a(x) := (K-1)F^{-1}(a + (K-1)x) + F^{-1}(1-x), \quad a \in [0,1],$$

and

$$c(a) := \inf\left\{ c \in \left(0, \frac{1-a}{K}\right] \,\middle|\, \int_c^{(1-a)/K} H_a(t)dt \geq \left(\frac{1-a}{K} - c\right) H_a(c) \right\}. \quad (30)$$

By (28), we have $F^{-1}(t) = 1/(1-t)$ for $t \in [0,1]$, and so

$$H_a(x) = \frac{K-1}{1-a-(K-1)x} + \frac{1}{x};$$

this is well-defined for $x \in [0, (1-a)/K]$ since the denominator is

$$1 - a - (K-1)x \geq 1 - a - (K-1)\frac{1-a}{K} = \frac{1-a}{K} \geq 0.$$

We can see that H_a is a convex function on its domain $x \in [0,(1-a)/K]$, and its value is ∞ at $x = 0$. Geometrically, the inequality "$\geq$" in (30) means that the average of H_a over $[c,(1-a)/K]$ is at least $H_a(c)$, which implies that we indeed have $c(a) > 0$ (as claimed earlier).

To summarize the previous paragraph, we can rewrite (29) as

$$\mathbb{P}\left(M_\phi(p_1,\ldots,p_K) \leq \epsilon\right) \leq 1 - a,$$

where $a = \Phi^{-1}(K/\epsilon)$ is the solution to the equation $\Phi(a) = K/\epsilon$, i.e., $H_a(c(a)) = K/\epsilon$ (we will see that there is a unique solution). Because of the continuity of H_a, these values a and $c = c(a)$ will satisfy the simultaneous equations

$$\begin{cases} H_a(c) = K/\epsilon \\ \int_c^{(1-a)/K} H_a(t)dt = \left(\frac{1-a}{K} - c\right) K/\epsilon, \end{cases}$$

i.e.,

$$\begin{cases} \frac{K-1}{1-a-(K-1)c} + \frac{1}{c} = \frac{K}{\epsilon} \\ \int_c^{(1-a)/K} \left(\frac{K-1}{1-a-(K-1)t} + \frac{1}{t}\right) dt = \left(\frac{1-a}{K} - c\right) \frac{K}{\epsilon}. \end{cases}$$

The first equation is quadratic in c, simplifying to

$$K(K-1)c^2 - K(1-a)c + (1-a)\epsilon = 0$$

and giving us two solutions:

$$c = \frac{K(1-a) \pm \sqrt{K^2(1-a)^2 - 4K(K-1)(1-a)\epsilon}}{2K(K-1)}. \quad (31)$$

To find which solution is the right one notice that the minimum of H_a is attained at the point x where $H'_a(x) = 0$, which gives us

$$x = \frac{1-a}{2(K-1)}.$$

Since our solution must satisfy $c > x$ for that x, we can see that the $\pm$ in (31) is in fact $+$.

The integral in the second equation is

$$[-\ln(1-a-(K-1)t) + \ln t]_c^{(1-a)/K} = \ln\left(\frac{1-a}{c} - (K-1)\right),$$

and so that equation becomes

$$\ln\left(\frac{1-a}{c} - (K-1)\right) = \left(\frac{1-a}{K} - c\right)\frac{K}{\epsilon}.$$

Setting $R := 1-a$ and plugging (31) (with $+$ in place of $\pm$) into the last equation, we obtain the equation

$$\ln\left(\frac{2K(K-1)}{K+\sqrt{K^2-4K(K-1)\epsilon/R}} - (K-1)\right) = \left(\frac{R}{K} - \frac{KR+\sqrt{K^2R^2-4K(K-1)R\epsilon}}{2K(K-1)}\right)\frac{K}{\epsilon}$$

for R. In terms of $\alpha := \epsilon/R$ we get, after some simplification,

$$\ln\left(\frac{2(K-1)}{1+\sqrt{1-4\alpha(K-1)/K}} - (K-1)\right) = \frac{K-2+K\sqrt{1-4\alpha(K-1)/K}}{2(K-1)\alpha}.$$

This coincides with (27). □

Asymptotically, Proposition 4 is stronger than Corollary 5:

Lemma 3. *As $K \to \infty$, $\alpha_K \sim 1/\ln K$.*

Proof. For an arbitrarily small $\epsilon > 0$, our goal is to prove that $\alpha_K \in [(1-\epsilon)/\ln K, (1+\epsilon)/\ln K]$ from some K on. In other words, that the left-hand side of (27) is less than or equal to the right-hand side for $\alpha = (1-\epsilon)/\ln K$, and that their order is reversed for $\alpha = (1+\epsilon)/\ln K$, from some K on.

For $\alpha = (1-\epsilon)/\ln K$, the right-hand side of (27) is asymptotically equivalent to

$$\frac{2K-2}{2(K-1)\alpha} = \frac{1}{\alpha} = \frac{\ln K}{1-\epsilon}.$$

The relation

$$\frac{2}{1+\sqrt{1-x}} = 1 + \frac{x}{4} + O(x^2) \tag{32}$$

(as $x \to 0$) implies that the left-hand side is asymptotically equivalent to

$$\ln\left((K-1)\left(1+\alpha\frac{K-1}{K}\right)-(K-1)\right)=\ln\left((K-1)\alpha\frac{K-1}{K}\right)\sim\ln K. \tag{33}$$

This proves the first statement.

For $\alpha=(1+\epsilon)/\ln K$, the right-hand side of (27) is asymptotically equivalent to

$$\frac{2K-2}{2(K-1)\alpha}=\frac{1}{\alpha}=\frac{\ln K}{1+\epsilon},$$

and we still have (33). This proves the second statement. □

The following analogue of Corollary 5 is tighter but only works for large K.

Corollary 6. *Suppose $K>345$. For any $\epsilon\in[0,1]$,*

$$\mathbb{P}\left(M_{-1}(p_1,\dots,p_K)\le\epsilon\right)\le(\ln K)\epsilon.$$

By Lemma 3, it is impossible to replace $\ln K$ by $c\ln K$ for any $c<1$ and for any lower bound on K.

Proof. Let us check that $\alpha_K\ge 1/\ln K$ for $K\ge 6000$ (and after that we can check numerically that $\alpha_K\ge 1/\ln K$ also holds for all $K>345$). It suffices to show that the left-hand side of (27) is less than or equal to the right-hand side for $\alpha=1/\ln K$; this assumes that $\alpha=1/\ln K$ is in the range $(0,\frac{K}{4(K-1)}]$ of α, which is the case for $K\ge 60$.

We will use the modification

$$\frac{2}{1+\sqrt{1-x}}\le 1+\frac{x}{3} \tag{34}$$

of (32), which is valid for $x\in[0,0.46]$, and the inequality

$$\sqrt{1-x}\ge 1-\frac{x}{2}-\frac{x^2}{5}, \tag{35}$$

which is valid for $x\in[0,0.66]$. Applying the inequalities (34) and (35) to (27) (the application being valid for $K\ge 6000$ and $K\ge 440$, respectively), we reduce our task to proving

$$\ln\left((K-1)\left(1+\frac{4}{3}\alpha(K-1)/K\right)-(K-1)\right)\le\frac{K-2+K\left(1-2\alpha(K-1)/K-\frac{16}{5}\left(\alpha\frac{K-1}{K}\right)^2\right)}{2(K-1)\alpha}.$$

The last inequality can be rewritten (for $\alpha=1/\ln K$) as

$$\ln\frac{4}{3}+2\ln(K-1)-\ln\ln K\le 2\ln K-1-\frac{8}{5}\frac{K-1}{K\ln K}$$

and holds for $K\ge 150$. □

Algorithm 1 Generalized Bonferroni–Holm procedure

Require: A significance level $\epsilon > 0$ and parameter $r < -1$ (or, w.l.o.g., (36)).
Require: A sequence of p-values $p_1, \dots, p_K$ ordered as $p_{k_1} \le \dots \le p_{k_K}$.
  for $k = 1, \dots, K$ **do**
    `reject` := `true`
    $I := \{k\}$
    for $i = K, \dots, 1, 0$ **do**
      if $\frac{r}{r+1} |I|^{1+1/r} M_r(p_I) > \epsilon$ **then**
        `reject` := `false`
      end if
      $I := I \cup \{k_i\}$
    end for
    if `reject` = `true` **then**
      reject H_k
    end if
  end for

7 Application to testing multiple hypotheses

In this section we apply the results of the previous sections, concerning multiple testing of a single hypothesis, to testing multiple hypotheses. Namely, we will arrive at a generalization of the Bonferroni–Holm method [8]. Fix a parameter

$$r \le \frac{\ln K}{1 - \ln K} \tag{36}$$

(cf. (24)); the Bonferroni–Holm case is $r := -\infty$.

Suppose p_k is a p-value for testing a composite null hypothesis H_k (meaning that, for any $\epsilon > 0$, $\mathbb{P}(p_k \le \epsilon) \le \epsilon$ under H_k). For $I \subseteq \{1, \dots, K\}$, let H_I be the hypothesis

$$H_I := (\cap_{k \in I} H_k) \cap \left(\cap_{k \in \{1,\dots,K\} \setminus I} H_k^c\right),$$

where H_k^c is the complement of H_k.

Fix a significance level ϵ. Let us reject H_I when

$$\frac{r}{r+1} |I|^{1+1/r} M_r(p_I) \le \epsilon,$$

where p_I is the vector of p_k for $k \in I$; by Corollary 4, the probability of error will be at most ϵ. If we now reject H_k when all H_I with $I \supseteq \{k\}$ are rejected, the family-wise error rate (FWER) will be at most ϵ. This gives the procedure described as Algorithm 1, in which $(k_1, \dots, k_K)$ stands for a fixed permutation of $\{1, \dots, K\}$ such that $p_{k_1} \le \dots \le p_{k_K}$.

An alternative representation of the generalized Bonferroni–Holm procedure given as Algorithm 1 is in terms of adjusting the p-values $p_1, \dots, p_K$ to new p-values $p_1^*, \dots, p_K^*$ that are valid in the sense of the FWER: we are guaranteed

Algorithm 2 Generalized Bonferroni–Holm procedure for adjusting p-values

Require: A parameter $r < -1$ (or, w.l.o.g., (36)).
Require: A sequence of p-values $p_1, \ldots, p_K$ ordered as $p_{k_1} \leq \cdots \leq p_{k_K}$.
  for $k = 1, \ldots, K$ **do**
    $p_k^* := 0$
    $I := \{k\}$
    for $i = K, \ldots, 1, 0$ **do**
      if $\frac{r}{r+1} |I|^{1+1/r} M_r(p_I) > p_k^*$ **then**
        $p_k^* := \frac{r}{r+1} |I|^{1+1/r} M_r(p_I)$
      end if
      $I := I \cup \{k_i\}$
    end for
  end for

to have $\mathbb{P}(\min_{k \in I} p_k^* \leq \epsilon) \leq \epsilon$ for all $\epsilon \in [0, 1]$, where I is the set of the indices k of the true hypotheses H_k. The adjusted p-values can be defined as

$$p_k^* := \max_{k \in I \subseteq \{1, \ldots, K\}} \frac{r}{r+1} |I|^{1+1/r} M_r(p_I)$$

and computed using Algorithm 2.

Acknowledgments

I am grateful to Dave Cohen, Alessio Sancetta, Wouter Koolen, and Lutz Mattner for their advice. This work was partially supported by the Cyprus Research Promotion Foundation.

References

[1] Grace E. Bates. Joint distributions of time intervals for the occurrence of successive accidents in a generalized Polya scheme. *Annals of Mathematical Statistics*, 26:705–720, 1955.

[2] Carole Bernard, Ludger Rüschendorf, and Steven Vanduffel. Value-at-Risk bounds with variance constraints. *Journal of Risk and Insurance*, 84:923–959, 2017.

[3] Morton B. Brown. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 31:987–992, 1975.

[4] Paul Embrechts and Giovanni Puccetti. Bounds for functions of dependent risks. *Finance and Stochastics*, 10:341–352, 2006.

[5] Ronald A. Fisher. Combining independent tests of significance. *American Statistician*, 2:30, 1948. This is the answer to Question 14 in Frederick Mosteller's "Questions and Answers" column.

[6] Benjamin T. Graham and Geoffrey R. Grimmett. Influence and sharp-threshold theorems for monotonic measures. *Annals of Probability*, 34:1726–1745, 2006.

[7] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, England, second edition, 1952.

[8] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[9] Gerhard Hommel. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal*, 25:423–430, 1983.

[10] Gerhard Hommel. Multiple test procedures for arbitrary dependence structures. *Metrika*, 33:321–336, 1986.

[11] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386, 1988.

[12] Andrei N. Kolmogorov. Sur la notion de la moyenne. *Atti della Reale Accademia Nazionale dei Lincei. Classe di scienze fisiche, matematiche, e naturali. Rendiconti Serie VI*, 12(9):388–391, 1930.

[13] James T. Kost and Michael P. McDermott. Combining dependent p-values. *Statistics and Probability Letters*, 60:183–190, 2002.

[14] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, and Tuve Löfström. On the calibration of aggregated conformal predictors. *Proceedings of Machine Learning Research*, 60:154–173, 2017.

[15] G. D. Makarov. Estimates for the distribution function of the sum of two random variables with given marginal distributions. *Theory of Probability and its Applications*, 26:803–806, 1981.

[16] Ruth Marcus, Eric Peritz, and K. Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660, 1976.

[17] Lutz Mattner. Combining individually valid and conditionally i.i.d. P-variables. Technical Report `arXiv:1008.5143 [stat.ME]`, `arXiv.org` e-Print archive, August 2011.

[18] Lutz Mattner. Combining individually valid and arbitrarily dependent P-variables. In *Abstract-book of the Tenth German Probability and Statistics Days*, page 104, Mainz, Germany, March 2012. Institut für Mathematik, Johannes Gutenberg-Universität Mainz.

[19] Xiao-Li Meng. Posterior predictive p-values. *Annals of Statistics*, 22:1142–1160, 1993.

[20] Dietrich Morgenstern. Berechnung des maximalen Signifikanzniveaus des Testes "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 27:285–286, 1980.

[21] Roger B. Nelsen. *An Introduction to Copulas*. Springer, New York, second edition, 2006.

[22] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems*. Springer, New York, 1998. Volume I: Theory; Volume II: Applications.

[23] Bernhard Rüger. Das maximale Signifikanzniveau des Tests "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 25:171–178, 1978.

[24] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14:623–632, 1982.

[25] Ludger Rüschendorf and Ludger Uckelmann. Variance minimization and random variables with constant sum. In Carles M. Cuadras, Josep Fortiana, and José A. Rodríguez-Lallena, editors, *Distributions with Given Marginals and Statistical Modelling*, pages 211–222. Kluwer, Dordrecht, Netherlands, 2002.

[26] R. John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.

[27] Vladimir Vovk. Cross-conformal predictors, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 6, August 2012.

[28] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[29] Ruodu Wang, Liang Peng, and Jingping Yang. Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities. *Finance and Stochastics*, 17:395–417, 2013.