

# Lecture 2: Merging evidence and FWER multiple testing

Eugenio Clerico

November 2025

**Merging functions.** In many statistical applications one does not work with a single p-value, but with several of them (e.g., coming from multiple tests, different data splits, or different experiments). In such situations, one often wants to combine the available p-values into a single summary p-value. This sort of procedure is called *merging*. We will focus here on the case where no assumptions are made about the dependence structure between the p-values to be merged.

Fix an integer  $K \geq 1$ . A *p-merging function* is a measurable, non-decreasing  $F : [0, \infty]^K \rightarrow [0, \infty]$ , such that for every data space  $\mathcal{X}$ , any hypothesis  $\mathcal{H}$  on  $\mathcal{X}$ , any  $K$ -tuple  $\mathbf{p} = (p_1, \dots, p_K)$  of p-variables for  $\mathcal{H}$ , the mapping  $x \mapsto F(p_1(x), \dots, p_K(x))$  is a p-variable for  $\mathcal{H}$ . Perhaps, the simplest example is the coordinate-wise maximum, namely

$$F(\mathbf{u}) = \max\{u_1, \dots, u_K\}.$$

It is indeed easily checked that if  $\mathbf{p}$  is a tuple of p-variables, then for any  $\alpha \geq 0$  and any  $Q$  in the null we have

$$Q(F(\mathbf{p}) \leq \alpha) \leq \min_{1 \leq k \leq K} Q(p_k \leq \alpha) \leq \alpha.$$

Another classical example is the Bonferroni merging function,

$$F(\mathbf{u}) = K \min\{u_1, \dots, u_K\}. \tag{1}$$

Its validity follows immediately from the union bound:

$$Q(F(\mathbf{p}) \leq \alpha) \leq \sum_{k=1}^K Q(p_k \leq \alpha/K) \leq \alpha.$$

A completely analogous notion exists for e-values. We call a measurable, non-decreasing function  $F : [0, \infty]^K \rightarrow [0, \infty]$  an *e-merging function* if for any space  $\mathcal{X}$ , any  $\mathcal{H}$  hypothesis on  $\mathcal{X}$ , and every  $K$ -tuple  $\mathbf{E} = (E_1, \dots, E_K)$  of e-variables for  $\mathcal{H}$ , the mapping  $x \mapsto F(E_1(x), \dots, E_K(x))$  is also an e-variable for  $\mathcal{H}$ . Because e-variables are characterised by the simple linear constraint  $\mathbb{E}_Q[E] \leq 1$  under every  $Q$  in the null, it turns out that the e-merging functions admit an explicit characterisation. More precisely, a measurable, non-negative, and non-decreasing function  $F$  is an e-merging function if, and only if, there exists a  $K$ -tuple  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$  of non-negative coefficients whose sum is at most 1, such that  $F$  is everywhere dominated by the affine functional

$$F_{\boldsymbol{\lambda}} : \mathbf{u} \mapsto 1 + \sum_{k=1}^K \lambda_k (u_k - 1). \tag{2}$$

A short proof of this characterisation is provided at the end of these notes.<sup>1</sup>

As a final remark, we note that e-merging functions naturally induce p-merging functions via the use of calibrators. Indeed, if  $F$  is an e-merging function, given any  $K$  calibrators  $f_1, \dots, f_K$  we have that

$$\tilde{F} : \mathbf{u} \mapsto \frac{1}{F(f_1(u_1), \dots, f_K(u_K))}$$

defines a p-merging function, mapping a  $K$ -tuple of p-variables into a post-hoc p-variable.

---

<sup>1</sup>The original proof is much longer and leverages optimal transport arguments. It can be found in Ruodu Wang's paper *The only admissible way of merging arbitrary e-values*.

**Multiple testing.** Multiple testing refers to the situation in which the same data set is used to test a family of hypotheses simultaneously.

Let  $\mathcal{X}$  be the data space, and let  $\{\mathcal{H}_i\}_{1 \leq i \leq N}$  be a finite family of hypotheses on  $\mathcal{X}$ . Suppose that the data are generated according to some distribution  $Q$ . We denote by

$$I_Q = \{i : Q \in \mathcal{H}_i\}$$

the set of indices corresponding to the hypotheses that are actually *true* under  $Q$ . A multiple testing procedure analyses the observed data and returns a set  $R \subseteq \mathcal{I}$  of indices, corresponding to the hypotheses to be *rejected*. Ideally, one wishes to avoid rejecting any true hypothesis. Thus, given a target level  $\alpha \in (0, 1)$ , if the data are generated from  $Q$  we aim for a test that rejects some index  $i \in I_Q$  with probability at most  $\alpha$ :

$$Q(R \cap I_Q \neq \emptyset) \leq \alpha. \quad (3)$$

Since it is not known in practice what is the distribution  $Q$  generating the data, a natural requirement in order to ensure (3) is

$$\sup_{i \in \mathcal{I}} \sup_{Q \in \mathcal{H}_i} Q(R \cap I_Q \neq \emptyset) \leq \alpha. \quad (4)$$

A testing procedure satisfying (4) is said to control the *family-wise error rate* (FWER) at level  $\alpha$ .

**Closed testing.** A general and powerful method to achieve FWER control (indeed the only admissible one, in the sense that no other procedure can uniformly outperform it) is the *closed testing* principle.

For every non-empty subset  $I \subseteq \mathcal{I}$ , fix a level- $\alpha$  test  $\phi_I$  for the hypothesis

$$\mathcal{H}_I = \bigcap_{i \in I} \mathcal{H}_i.$$

More explicitly,  $\phi_I$  is a binary random variable whose outcomes can be 0 or 1. The test rejects  $\mathcal{H}_I$  only when  $\phi_I(x) = 1$ , where  $x$  is the data set. The significance level  $\alpha$  ensures that

$$\sup_{Q \in \mathcal{H}_I} Q(\phi_I = 1) \leq \alpha.$$

The closed testing procedure rejects a single hypothesis  $\mathcal{H}_i$  if, and only if,  $\mathcal{H}_I$  is rejected for every  $I \ni i$ . In terms of the tests  $\phi_I$ , this is equivalent to say that the resulting test is induced by the binary variable

$$\phi_i^* = \min_{I \ni i} \phi_I.$$

Each local test  $\phi_i^*$  has significance level  $\alpha$ , and the whole procedure controls the FWER at level  $\alpha$ .

To verify the FWER control guarantee, fix any  $Q$  with  $I_Q \neq \emptyset$ . Then,  $R \cap I_Q = \{i \in I_Q : \phi_i^* = 1\}$  is empty only when  $\max_{i \in I_Q} \phi_i^* = 0$ . It follows that  $Q(R \cap I_Q \neq \emptyset) = Q(\max_{i \in I_Q} \phi_i^* = 1)$ . For every  $i \in I_Q$ , we have that  $\phi_i^* \leq \phi_{I_Q}$ . In particular,  $\max_{i \in I_Q} \phi_i^* \leq \phi_{I_Q}$ . From this we see that

$$Q(R \cap I_Q \neq \emptyset) = Q\left(\max_{i \in I_Q} \phi_i^* = 1\right) \leq Q(\phi_{I_Q} = 1) \leq \alpha,$$

since  $\phi_{I_Q}$  define a valid test at level  $\alpha$  for  $\mathcal{H}_{I_Q}$ , and  $Q \in \mathcal{H}_{I_Q}$  by definition.

**Using p-values and e-values.** Closed testing can be implemented both with p-values and with e-values. Suppose first that for every non-empty  $I \subseteq \mathcal{I}$  we are given a p-variable  $p_I$  for the intersection hypothesis  $\mathcal{H}_I = \bigcap_{i \in I} \mathcal{H}_i$ . In the closed procedure, we reject  $\mathcal{H}_I$  whenever  $p_I(X) \leq \alpha$ , and subsequently reject  $\mathcal{H}_i$  only if  $\mathcal{H}_I$  is rejected for all  $I \ni i$ . Equivalently, the induced local p-value is

$$p_i^*(X) = \max_{I \ni i} p_I(X),$$

and we reject  $\mathcal{H}_i$  if, and only if,  $p_i^*(X) \leq \alpha$ .

To show that this satisfies FWER control we proceed as earlier. Fix any distribution  $Q$  with  $I_Q \neq \emptyset$ . Then  $R \cap I_Q \neq \emptyset$  implies that  $\min_{i \in I_Q} p_i^* \leq \alpha$ . Since  $\min_{i \in I_Q} p_i^* \geq p_{I_Q}$ ,  $Q(R \cap I_Q \neq \emptyset) \leq Q(p_{I_Q} \leq \alpha) \leq \alpha$ , as  $Q \in \mathcal{H}_{I_Q}$ .

A completely analogous construction works with e-values. Suppose we are given e-variables  $E_I$  for the intersection hypotheses. In the closed procedure we reject  $\mathcal{H}_I$  whenever  $E_I(x) \geq 1/\alpha$ . The induced local e-value is

$$E_i^*(X) = \min_{I \ni i} E_I(X),$$

and we reject  $\mathcal{H}_i$  whenever  $E_i^* \geq 1/\alpha$ . Repeating essentially the same argument as above shows that the procedure controls the FWER at level  $\alpha$ .

Note that in both the p-value and e-value formulations, the only structural requirement is that  $p_I$  (respectively,  $E_I$ ) is a valid p-variable (e-variable) for each intersection hypothesis. Producing such valid objects is precisely where *merging functions* enter the picture: they provide general, assumption-free ways to construct  $p_I$  or  $E_I$  for arbitrary  $I$  starting from the individual  $p_i$  or  $E_i$ . We now illustrate this with the classical Holm method.

**Holm closed testing.** A direct application of closed testing requires specifying a valid p-variable  $p_I$  for every non-empty  $I \subseteq \mathcal{I}$ . Since there are  $2^N - 1$  such subsets, this may appear computationally infeasible. The Holm method is a technique to perform closed testing using only  $N$  p-values, those relative to the hypotheses  $\mathcal{H}_i$  to be tested.

So, let us assume that we are given a p-variable  $p_i$  for each  $\mathcal{H}_i$ . For any non-empty  $I \subset \mathcal{I}$ , we leverage the Bonferroni merging function (1) and define

$$p_I = |I| \min_{i \in I} p_i.$$

Each  $p_i$  is a p-variable for  $\mathcal{H}_i$ , and so in particular for  $\mathcal{H}_I \subseteq \mathcal{H}_i$ . Thanks to the fact that we are using a p-merging function, we have that  $p_I$  is also a p-variable for  $\mathcal{H}_I$ . We can therefore apply the closed testing principle, and define

$$p_i^* = \max_{I \ni i} p_I = \max_{I \ni i} (|I| \min_{j \in I} p_j).$$

We reject  $\mathcal{H}_i$  if, and only if,  $p_i^*(X) \leq \alpha$ , where  $X$  is the observed dataset.

At first sight this expression still seems to require evaluating  $p_I(X)$  over exponentially many subsets  $I$ . However, the Holm procedure has a special structure that makes the computation extremely efficient.

To simplify the analysis, without loss of generality we assume that the p-values are already ordered so that

$$p_1(X) \geq p_2(X) \geq \dots \geq p_N(X).$$

If this is not the case, one can simply re-label the hypotheses. Now, let us start to check if we can reject  $\mathcal{H}_N$ . Picking  $I = \mathcal{I}$ , we see that a necessary condition for  $\mathcal{H}_N$  to be rejected is that

$$p_{\mathcal{I}}(X) = \min_{i \in \mathcal{I}} p_i(X) = p_N(X) \leq \alpha/N.$$

Actually, we can also see that if  $p_N(X) > \alpha/N$ , then no hypothesis can be rejected, because being unable to reject  $\mathcal{H}_{\mathcal{I}}$  means that no rejection is possible. On the other hand, if  $p_N(X) \leq \alpha/N$ , then for any  $I \ni N$ , we have that

$$p_I(X) = \min_{i \in I} p_i(X) = p_N(X) \leq \alpha/N \leq \alpha/|I|,$$

and so  $\mathcal{H}_N$  can be rejected.

If  $\mathcal{H}_N$  has been rejected, then we can check if  $\mathcal{H}_{N-1}$  can be rejected. Since we have rejected  $\mathcal{H}_N$ , we already know that, for any  $I \supseteq \{N-1, N\}$ ,

$$p_I(X) = p_N(X) \leq \alpha/N \leq \alpha/|I|.$$

So, only the sets  $I \subseteq \{1, \dots, N-1\}$  containing  $i$  need to be checked. Proceeding exactly as before (replacing  $N$  with  $N-1$ ), we see that  $\mathcal{H}_{N-1}$  is rejected if, and only if

$$p_{N-1}(X) \leq \alpha/(N-1).$$

If  $\mathcal{H}_{N-1}$  is rejected, one can go further, and check if  $\mathcal{H}_{N-2}$  can be rejected, but if  $\mathcal{H}_{N-1}$  cannot be rejected one has to stop and no further hypothesis can be rejected.

So, one can proceed in this sequential way, first checking if we can reject  $\mathcal{H}_N$ , then  $\mathcal{H}_{N-1}$ , and so forth, stopping the procedure whenever a hypothesis is not rejected. This requires at most  $N$  steps (if everything is rejected).

As a summary, an equivalent formulation of the Holm procedure with p-values is the following. First we order the p-values and re-label the hypotheses, so that  $p_N(X) \leq p_{N-1}(X) \dots \leq p_1(X)$ . We let  $i^*$  be the largest index such that for all  $i \geq i^*$  we have that

$$ip_i(X) \leq \alpha,$$

(or  $i^* = N + 1$  if the condition is not met by any  $i$ ). We then reject the indices in

$$R = \{i \in \mathcal{I} : i \geq i^*\}.$$

As a remark, note that the threshold  $\alpha/N$  can look extremely conservative when  $N$  is large, and it must be smaller than the smallest p-value for the procedure to reject anything. This is true, but it is not a flaw: it is simply how FWER works. The goal is to ensure that there is *not even a single* false positive. To get some intuition, imagine a situation in which all hypotheses are true and the p-values are independent and uniformly distributed. When  $N$  grows, the random fluctuations make it increasingly likely that some p-value will be extremely small. With many tests, we are almost guaranteed to see an apparently “significant” value just by chance. Therefore, the threshold must shrink accordingly to prevent these random fluctuations from causing any false discovery. For situations where such strong guarantees are not required, one typically considers less conservative forms of control than FWER, which lie outside the scope of these notes.

**Holm with e-values.** Similarly to the case of p-values, we can also perform closed testing with e-values avoiding exponential computational complexity. Suppose that we have  $N$  e-variables,  $E_1, \dots, E_N$ , one for each hypothesis  $\mathcal{H}_i$ . As discussed earlier, the best way to merge them is via an affine mapping. Here we will consider the arithmetic mean, which is in the form (2) for  $\lambda = \mathbf{1}/K$ . So, for any  $I \subseteq \mathcal{I}$ , define the local e-variable

$$E_I = \frac{1}{|I|} \sum_{i \in I} E_i.$$

Following the closed testing principle, we define for each  $i$

$$E_i^* = \min_{I \ni i} E_I = \min_{I \ni i} \frac{1}{|I|} \sum_{j \in I} E_j,$$

and reject  $\mathcal{H}_i$  if and only if  $E_i^*(X) \geq 1/\alpha$ .

To see how this can lead to an efficient algorithm, we remark that for each  $i$  the condition  $E_i^*(X) \geq 1/\alpha$  holds if, and only if,

$$E_i(X) \geq \frac{1}{\alpha} + \max_{I \ni i} \sum_{j \in I \setminus \{i\}} \left( E_j(X) - \frac{1}{\alpha} \right).$$

The maximum is achieved by the set  $I_\alpha \cup \{i\}$ , with

$$I_\alpha = \{j \in \mathcal{I} : E_j(X) < 1/\alpha\}.$$

Note that if  $i \in I_\alpha$  then  $E_i(X) < 1/\alpha$  and so  $\mathcal{H}_i$  will be rejected. In particular, the rejection set can be written as

$$R = \left\{ i \notin I_\alpha : E_i(X) \geq \frac{1}{\alpha} + \sum_{j \in I_\alpha} \left( \frac{1}{\alpha} - E_j(X) \right) \right\}.$$

This whole procedure can be summarised even more compactly. Define

$$\delta_\alpha = \sum_{i \in \mathcal{I}} \max\{0, \frac{1}{\alpha} - E_i(X)\}. \quad (5)$$

Then, the rejection set is simply

$$R = \{i \in \mathcal{I} : E_i(X) \geq \frac{1}{\alpha} + \delta_\alpha\}. \quad (6)$$

**Comments.** In general, neither the p-value nor the e-value version of Holm's method uniformly dominates the other. The classical p-value formulation can be more powerful when the given p-values are sharp, as the e-values are often more conservative. On the other hand, aggregating evidence through merging tends to favour the e-value approach, which can tilt the balance in favour of e-values in certain settings.

When the p-variables under consideration are post-hoc p-variables (that is, inverses of e-variables) the e-value version of Holm's method always improves upon its p-value counterpart. Let us see this more explicitly and suppose that  $p_1, \dots, p_N$  are post-hoc p-variables. The p-value method can now be viewed as an e-value closed testing procedure based on the e-variables  $E_i = 1/p_i$  merged through the function  $\mathbf{u} \mapsto \frac{1}{K} \max\{1, \dots, u_K\}$ , which is dominated by the arithmetic mean. Consequently, its induced local e-values  $\tilde{E}_i^*$  satisfy  $\tilde{E}_i^* \leq E_i^*$  (with  $E_i^*$  the local e-values of the Holm procedure), with strict inequality unless all  $E_i$  coincide.

The two merging functions used highlight an important conceptual difference between the two approaches. The p-value based procedure effectively look at every hypothesis *on its own* and does not combine evidence ( $p_I$  only depends on the smallest  $p_i$ , with  $i \in I$ , and all the evidence brought by the other p-values is lost). In contrast, the e-value formulation *aggregates* evidence: every  $E_i$  (with  $i \in I$ ) contributes to the merged quantity  $E_I$ , and so the effective threshold  $1/\alpha + \delta_\alpha$  used in (6) takes into account the collective behaviour of all e-values. In extreme cases, this leads to completely different outcomes: if every  $E_i(X) = 1/\alpha$  for all indices  $i$ , then the e-value approach rejects all the hypotheses, while the p-value Holm procedure does not reject any of them, unless  $N = 1$  (each p-value equals  $\alpha$ , while the threshold is  $\alpha/N$ ). This example illustrates how the e-value approach focuses on joint evidence: observing many e-values simultaneously above  $1/\alpha$  is extremely unlikely under the null  $\mathcal{H}_I$ , and thus constitutes strong *collective* evidence. In contrast, the p-value based method triggers a rejection as soon as a *single* p-value falls below the threshold, which forces the threshold itself to be very small in order to protect against random fluctuations across many tests.

As a final remark, the e-value Holm procedure illustrates that small e-values (even below 1) still play an important role when evidence is aggregated. Indeed, while such values would not trigger a rejection in an experiment testing a single hypothesis, they directly influence the threshold correction  $\delta_\alpha$ , defined in (5), and hence affect the overall decision, in multiple testing. Thus, with e-values even small contribution matter once several source of evidence are combined.

**Proof of the e-merging characterisation.** Let  $F : [0, \infty]^K \rightarrow [0, \infty]$  be an e-merging function. We need to prove that it is dominated by an affine function in the form (2). We will use the following lemma.

**Lemma 1.** *Fix any finite family of vectors  $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)} \in [0, \infty]^K$ , and non-negative weights  $q_1, \dots, q_n$ , such that  $\sum_{j=1}^n q_j = 1$ . Let  $\mathbf{1}$  denote the vector  $(1, \dots, 1) \in \mathbb{R}^K$ . Then, we have that*

$$\sum_{j=1}^n q_j \mathbf{u}^{(j)} = \mathbf{1} \quad \implies \quad \sum_{j=1}^n q_j F(\mathbf{u}^{(j)}) \leq 1.$$

*Proof.* Let  $\mathcal{X} = \{x_1, \dots, x_n\}$  be a finite set, and consider the hypothesis  $\mathcal{H} = \{Q\}$ , with  $Q = \sum_{j=1}^n q_j \delta_{x_j}$ . Define  $K$  random variables  $E_1, \dots, E_K$  on  $\mathcal{X}$  via  $E_k(x_j) = u_k^{(j)}$ . Then,

$$\mathbb{E}_Q[E_k] = \sum_{j=1}^n q_j E_k(x_j) = \sum_{j=1}^n q_j u_k^{(j)} = 1,$$

and so each  $E_k$  is an e-variable for  $\mathcal{H}$ . We have

$$\sum_{j=1}^n q_j F(\mathbf{u}^{(j)}) = \sum_{j=1}^n q_j F(E_1(x_j), \dots, E_K(x_j)) = \mathbb{E}_Q[F(E_1, \dots, E_K)] \leq 1,$$

since  $F$  being an e-merging function implies that  $F(E_1, \dots, E_K)$  is an e-variable for  $\mathcal{H}$ .  $\square$

Now, let  $\hat{F}$  denote the concave envelope of  $F$ , namely the smallest concave function that dominates  $F$  everywhere. A standard characterisation gives

$$\hat{F}(\mathbf{u}) = \sup_{n \geq 1} \sup_{q \in \Delta_n} \sup \left\{ \sum_{j=1}^n q_j F(\mathbf{u}^{(j)}) : \mathbf{u}^{(j)} \in [0, \infty]^K \forall j, \sum_{j=1}^n q_j \mathbf{u}^{(j)} = \mathbf{u} \right\}, \quad (7)$$

where  $\Delta_n = \{q \in [0, 1]^n : \sum_{j=1}^n q_j = 1\}$ .

Now, the key observation is that thanks to Lemma 1 and the characterisation (7), we have that

$$\hat{F}(\mathbf{1}) \leq 1.$$

Then, applying the supporting hyperplane theorem to the hypograph of  $\hat{F}$  (which is a convex set) we deduce that  $\hat{F}$  is dominated by an affine function  $G$  such that  $G(\mathbf{1}) = 1$ , namely a  $G$  in the form

$$G(\mathbf{u}) = 1 + \sum_{k=1}^K \lambda_k (u_k - 1),$$

with  $\boldsymbol{\lambda} \in [-\infty, \infty]^K$ . Note that  $F$  (and so  $\hat{F}$ ) is non-negative on the whole domain  $[0, \infty]^K$ , and so  $G$  as well must be non-negative. Imposing that both  $G(\mathbf{0})$  and  $G(\boldsymbol{\infty})$  are non-negative implies that the components of  $\boldsymbol{\lambda}$  are non-negative and have sum at most 1, as required.<sup>2</sup>

**Bibliography.** The discussion on merging e-values mostly follows the beginning of Chapter 8 in Aditya Ramdas and Ruodu Wang's book *Hypothesis Testing with E-values*, although the proof of the e-merging function characterisation is novel. Chapter 12 discusses in greater details p-merging, and its connections with e-merging. The closure principle and the derivation of the Holm procedure for p-values can be found on most textbooks about multiple testing. The Holm procedure with e-values is presented in the first part of *Family-wise error rate control with e-values* by Will Hartog and Lihua Lei.

---

<sup>2</sup>As usual, we used the bold font for vectors in  $[0, \infty]^K$ , so  $\mathbf{0} = (0, \dots, 0)$  and  $\boldsymbol{\infty} = (\infty, \dots, \infty)$ .