

Beyond Neyman-Pearson: e-values enable hypothesis testing with a data-driven alpha

Peter Grünwald¹²

¹Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

²Leiden University, Leiden, The Netherlands

April 4, 2024

Abstract

A standard practice in statistical hypothesis testing is to mention the p-value alongside the accept/reject decision. We show the advantages of mentioning an e-value instead. With p-values, it is not clear how to use an extreme observation (e.g. $P \ll \alpha$) for getting better frequentist decisions. With e-values it is straightforward, since they provide Type-I risk control in a generalized Neyman-Pearson setting with the decision task (a general loss function) determined post-hoc, after observation of the data — thereby providing a handle on ‘roving α ’s’. When Type-II risks are taken into consideration, the only admissible decision rules in the post-hoc setting turn out to be e-value-based. Similarly, if the loss incurred when specifying a faulty confidence interval is not fixed in advance, standard confidence intervals and distributions may fail whereas e-confidence sets and e-posteriors still provide valid risk guarantees. Sufficiently powerful e-values have by now been developed for a range of classical testing problems. We discuss the main challenges for wider development and deployment.

We perform a null hypothesis test with significance level α and we observe a p-value $P \ll \alpha$. Why aren’t we allowed to say “we have rejected the null at level P ”? While a continuous source of bewilderment to the applied scientist, professional statisticians understand the reason: to get a Type-I error probability guarantee of α — a cornerstone of the Neyman-Pearson (NP) theory of testing — we must set α in advance. But this immediately raises another question: why should the p-value be mentioned at all in scientific papers, next to the reject/accept decision for the pre-specified α [4, 20]? The prevailing attitude is to accept this standard practice, on the grounds that it “provides more information” — as explicitly stated by, for example, Lehmann [25], one of NP theory’s main contributors. But this is problematic: there is nothing in NP theory to tell us what the decision-theoretic consequences of ‘ $P \ll \alpha$ ’ could be, whereas at the same time, the fundamental motivation behind NP theory *is* decision-theoretic: according to [32], “[all of] *mathematical statistics deals with problems relating to performance characteristics of rules of inductive behavior* [i.e. decision rules] *based on random experiments*”. There is no simple way though to translate observation of a P with $P \ll \alpha$ into better decisions: as is well-known and reviewed below ((4) and (28)), intuitive and common decision-theoretic interpretations of $P \ll \alpha$ are usually just wrong. We are therefore faced with a standard practice in NP testing that, according to (strict, behaviorist) NP theory, is not part of mathematical statistics!

E as an alternative for P In our main result, Theorem 1, we show that this issue can be resolved *by mentioning e-values rather than p-values* next to the accept/reject decision. E-values [13, 52, 42, 56, 37] are a recently popularized alternative for p-values that are related to, but far more general than, likelihood ratios. Importantly, as reviewed in Example 2 below, for any NP test with the accept/reject-decision based on a p-value, the exact same test can be implemented by basing the decision on an e-value. Thus there is no a priori reason why one should accompany the decision of a NP test with a p-value rather than an e-value. But, in contrast to the p-value, the e-value has a clear decision-theoretic justification that remains valid if decision tasks are formulated *post-hoc*, i.e. after seeing, and in light of, the data. Concretely, after the result of a study has been published, and when new circumstances prevail, one conceivably might contemplate different actions, with different associated losses, than originally planned. For example, a study about vaccine efficacy (VE) in a pandemic may have been set up as a test between null hypothesis $VE \leq 30\%$ and alternative $VE \geq 50\%$ [45]. The original plan was to vaccinate all people above 60 years of age if the null is rejected. But suppose the null actually gets rejected with a very small p-value $\ll \alpha$, and at the same time the virus’ reproduction rate may be much higher than anticipated. Based on *both* the observed data (summarized by P) *and* the changed circumstances, one might now contemplate a new action, vaccinate everyone over 40, with higher losses if the alternative is false and higher pay-offs if it is true. E-values can be used unproblematically for such a post-hoc formulated decision task; p-values cannot. A second example is simply the fact that scientific results are *published* and remain on record so as to be useful for future deployment: a company contemplating to produce medication X may find a publication about the efficacy of X that is, say, 15 years old. Back then, in two independent studies the null (no efficacy) was rejected at the given $\alpha = 0.05$, but producing X would have been prohibitively expensive so this finding was not acted upon. But recently the company managed a technological breakthrough making production of X much cheaper. Had α been smaller than 0.01, they would now decide to take X into production. But now suppose that in both original studies, $P < 0.01$ yet $\alpha = 0.05$. The upshot of Example 1, Proposition 1 and Theorem 1 below is that, if one had observed $S^{-1} < 0.01$ for an e-variable S then acting anyway, despite the changed circumstances is *Type-I risk safe*, in the precise sense of (1) below; but doing this based on $P < 0.01$ is unsafe in the sense that no clear risk (performance) bounds can be given when engaging in such behavior.

From Testing to Estimation with Confidence: the e-posterior The medication X example is too simplistic: it only deals with rejecting the null of ‘no efficacy’. In reality one wants to take effect sizes into account as well when making decisions. Section 3 shows that e-value methods extend to that setting as well. Upon observing data from a statistical model with parameter of interest θ , the question now becomes how to properly interpret the statement “ $\theta \in CS_\alpha$ ”, where CS_α is a $(1 - \alpha)$ -confidence set, usually an interval. The correct, basic interpretation only says that, when repeatedly performing studies, the true parameter will lie in CS_α in a fraction of about $1 - \alpha$ studies. But practitioners want more, and indeed, CS’s are often given an evidential interpretation — one outputs not one but a system of confidence intervals, one for each of a series of coefficients such as 80%, 90%, 95%, 99%, or even a full *confidence distribution* [6, 41] and this, it is said “*summarizes what the data tell us about θ , given the model*” [8, page 227] or “the information about the parameter” [24]. As our second contribution, we show there is benefit in replacing standard CS’s by e-CS’s, and

confidence distributions by *e*-posteriors [15]: again, these stand on firmer decision-theoretic ground.

BIND Assumption underlying p-values and standard CSs While so far we highlighted the problems with post-hoc determined loss functions, below we show that decisions based on P’s (Section 1) and CS’s (Section 3) may already become unsafe, in the Type-I risk sense of (1), as soon as the decision task involves a ‘Type-I’ loss function that can take on more than two values, even if this loss function *is* determined in advance. Essentially, we can only be sure that decisions based on P’s and CS’s are reliable if both (1) the loss function is binary-valued (B) and (2), it is determined in advance, or at least independently (IND) of the observed data. Thus, they really operate under a BIND (binary + independence) assumption. E-values and -posteriors lead to decisions that retain Type-I risk safety if BIND is violated.

Technical Contribution and Contents To obtain frequentist guarantees without BIND we first need to reformulate NP testing in terms of losses and risks rather than errors and error probabilities, an idea going back to Wald’s seminal 1939 paper introducing statistical decision theory [54]. But while Wald lets go off the Type-I/II error distinction as soon as he allows for more than two actions, we stick with Type-I and Type-II risks (replacing Type-I and Type-II error probabilities, respectively) and show that the e-value is then the natural statistic to base decisions upon, and remains so if the decision task is determined post-hoc. Thus, our *GNP* (*Generalized Neyman-Pearson*) Theory follows a path opened up by Wald but apparently not pursued further thereafter. In Section 1 we informally present this reformulation, show how P-based procedures get in trouble if BIND is violated, introduce e-values and explain how, when combined with a *maximally compatible decision rule*, they guarantee Type-I risk safety even without BIND. Section 2 then formalizes the reasoning and presents our main result, Theorem 1. Among all Type-I risk safe decision rules, we aim only for those that have *admissible* Type-II risk behavior; we call a rule admissible if there exists no other decision rule that is never worse and sometimes strictly better. Theorem 1, which has the flavour of a *complete class theorem* [3, 9] shows that, under mild regularity conditions, the set of admissible decision rules are precisely those that are based on some e-variable S via a maximally compatible decision rule. Section 3 extends our findings to confidence intervals and distributions (CD’s). CD’s can be replaced by e-posteriors, a novel notion treated in much more detail in my recent paper [15], which may be viewed as a companion to this one, more oriented towards a Bayesian-inclined readership.

An Important Caveat Systematic development of e-values has only started very recently (in 2019). While a lot of progress has been made, and by now useful (\approx powerful) e-values are available for a number of practically important parametric and nonparametric testing and estimation problems, there is still an enormously wide range of problems for which p-values — systematically developed since the 1930s — exist yet e-values have not yet been developed. We briefly review initial success stories and current challenges in Section 4, informing the final Section 5 which indicates the way forward and re-interprets our findings as establishing a *quasi-conditional* paradigm. All longer mathematical derivations and proofs are delegated to the Supporting Information Appendix (SI).

1 Generalized Neyman-Pearson Theory

1.1 Losses instead of Errors

In the basic NP setting, we observe data Y taking values in some set \mathcal{Y} , with both the null hypothesis $\mathcal{H}(\underline{0})$ and the alternative $\mathcal{H}(\underline{1})$ being represented as collections of distributions for Y . NP [34] tell us to fix some α and then adopt the decision rule that, among all decision rules with Type-I error bounded by α , minimizes the Type-II error. Following Wald [54], we re-interpret this procedure in terms of a nonnegative loss function $L(\cdot, \cdot)$, with $L(\kappa, a)$ denoting the loss made by action a if κ is the true state of nature. We have $\kappa \in \{\underline{0}, \underline{1}\}$ and $\mathcal{A} = \{0, 1\}$, $\kappa = \underline{0}$ representing that the null is correct, $\kappa = \underline{1}$ that the alternative is correct, $a = 0$ standing for ‘accept’ and $a = 1$ for ‘reject’ the null. We invariably assume $L(\underline{0}, 1) > L(\underline{0}, 0) \geq 0, L(\underline{1}, 0) > L(\underline{1}, 1) \geq 0$. ‘Of course’ (as Wald writes) we may want to set $L(\underline{0}, 0) = L(\underline{1}, 1) = 0$ and we will do this for now, but it is not required for the subsequent developments. In this formulation, the usual α -Type-I error guarantee is replaced by an ℓ -Type-I risk guarantee. Formally, we fix an ℓ in advance of observing the data and we say that decision rule δ (i.e. a test), defined as a function from \mathcal{Y} to \mathcal{A} , is *Type-I risk safe* if

$$\sup_{P_0 \in \mathcal{H}(\underline{0})} \mathbf{E}_{Y \sim P_0}[L(\underline{0}, \delta(Y))] \leq \ell, \quad (1)$$

where, for $j = 0, 1$, $P_j \in \mathcal{H}(j)$, $\mathbf{E}_{Y \sim P_j}[L(\underline{0}, \delta(Y))]$ is called the *risk of P_j* , i.e. the expected loss under P_j . Following NP again, with again ‘error probability’ replaced by ‘risk’, we now postulate that among all Type-I risk safe decision rules δ , we ideally want to pick one that has small *worst-case Type-II risk*

$$\sup_{P_1 \in \mathcal{H}(\underline{1})} \mathbf{E}_{Y \sim P_1}[L(\underline{1}, \delta(Y))]. \quad (2)$$

(1) expresses that, whatever we decide, we want to make sure that our risk (expected loss) under the null is no larger than ℓ . In a standard level- α test, one rejects the null if $\mathbf{P}(y)$, the p-value corresponding to data y , satisfies $\mathbf{P}(y) \leq \alpha$. A corresponding decision rule in terms of loss functions is to reject the null whenever the observed $\mathbf{P}(y)$ satisfies

$$\mathbf{P}(y) \cdot L(\underline{0}, 1) \leq \ell. \quad (3)$$

We get exactly the same behavior as for the standard level α -test if we set $L(\underline{0}, 1) = \ell/\alpha$. For example, for $\alpha = 0.05$ we can set $\ell = 1$ and then $L(\underline{0}, 1) := 20$; then, just like in NP testing (3) tells us to pick a δ° which rejects the null if $\mathbf{P} \leq 0.05$. If \mathbf{P} is defined so that δ° is UMP (uniformly most powerful), then combined with any loss function $L(\underline{1}, 0) > 0$, δ° will also minimize the worst-case Type-II risk (2) among all δ that satisfy Type-I error probability $\leq \alpha$: up until now we have merely reformulated standard NP theory.

Actions of Varying Intensity But now suppose we have *more than two* actions available. For example, consider four alternative actions: accept the null (retain the status quo), take mild action (e.g. vaccinate all people over 60), take more drastic action (vaccinate everyone over 40) and extreme action (vaccinate the whole population). We consider this question, too, in terms of Type-I and Type-II risk and confidence — thereby taking a different direction than standard decision theory. For example, our action space could now be $\mathcal{A}_b = \{0, 1, 2, 3\}$ with loss function $L_b(\underline{0}, 0) = 0, L_b(\underline{0}, 1) = 20\ell, L_b(\underline{0}, 2) = 100\ell, L_b(\underline{0}, 3) = 500\ell$ and $L_b(\underline{1}, 3) <$

$L_b(\underline{1}, 2) < L_b(\underline{1}, 1) < L_b(\underline{1}, 0) = \ell$. More generally, as long as Type-I loss is increasing in a and Type-II loss is decreasing, such an extension of the NP setting makes intuitive sense.

In terms of p-values, the straightforward extension of (3) to this multi-action case would be to play action a where a is the largest value such that

$$P(y) \cdot L_b(\underline{Q}, a) \leq \ell. \quad (4)$$

But, assuming our p-value is strict so that it has a uniform distribution under the null, this gives a Type-I risk of

$$\mathbf{E}_{Y \sim P_0}[L_b(\underline{Q}, \delta(Y))] = \left(\frac{1}{20} - \frac{1}{100}\right) \cdot 20\ell + \left(\frac{1}{100} - \frac{1}{500}\right) \cdot 100\ell + \frac{1}{500} \cdot 500\ell = 2.6\ell, \quad (5)$$

violating the guarantee we aimed to impose and showing that a naive p-value based procedure does not work. The problem gets exacerbated if we allow for more than four actions: in the SI we show that the expected loss of the naive procedure (4) may go to ∞ as we add additional actions with $L_b(\underline{Q}, a)$ increasing and $L_b(\underline{1}, a)$ decreasing in a . There we also show that an obvious ‘fix’, namely modifying (4) to make sure that for each action a , $L_b(\underline{Q}, a)$ gets multiplied by exactly the probability that action a is taken, does not solve this issue.

Post-Hoc Loss Functions Allowing more than two actions is really just a warm-up to a further extension which arguably better models what often happens in, for example, medical practice: the post-hoc determination or modification of a decision task, after seeing the data and dependent on the data, such as in the vaccine efficacy example in the introduction. That is, there is really an underlying class (whose definition may be unknowable) of loss functions $L_b(\cdot, \cdot)$ with associated action spaces \mathcal{A}_b , and the decision-maker (DM) is posed a particular decision task $L_b(\cdot, \cdot)$ where b , indexing the loss actually used, is really the outcome of a random variable $B = b$, whose distribution may depend on the data in all kinds of ways. The actual $B = b$ that is presented is thus random and only fixed *after* the study result has become available; i.e. ‘post-hoc’. Crucially, the process determining the actual value of B is typically murky; nobody knows exactly what loss function would have been considered in what alternative circumstances; DM only knows the loss function finally arrived at.

Again, with p-values, we might be tempted to pick the largest action a such that (4) holds, where now b is really the (observed, known) outcome of random variable B whose definition is itself unknown. Now, even if for each b , L_b allows for only two actions, so that the problem superficially resembles the standard NP setting, using (4) can have disastrous consequences in the post-hoc setting, as the following example shows.

Example 1 Suppose there are three loss functions L_b , for $b \in \mathcal{B} = \{1, 2, 3\}$, with corresponding actions $\mathcal{A}_b = \{0, b\}$. We set $L_1(\underline{Q}, 1) = 20\ell$, $L_2(\underline{Q}, 2) = 100\ell$, $L_3(\underline{Q}, 3) = 500\ell$, $L_b(\underline{Q}, 0) = 0$, $L_b(\underline{1}, 0) := \ell$ for all $b \in \mathcal{B}$, and $L_b(\underline{1}, b)$ strictly decreasing in b . This is like the previous example, but rather than always being able to choose one among four actions, the very set of choices that is presented to DM via setting $B = b$ might depend on the data Y or on external situations. One cannot rule out that this is done in an unfavourable manner — if the data suggest strong evidence then the policy developers (e.g. a pandemic outbreak management team) might only suggest actions with drastic consequences. Suppose, for example, that if $P > 0.02$, the DMs are presented loss L_1 ; if $0.001 < P \leq 0.02$ they are presented loss

L_2 ; and if $P \leq 0.001$ they are presented loss L_3 . Using (4), we then get (assuming again uniform P) a Type-I risk of

$$\mathbf{E}_{Y \sim P_0}[L(\underline{0}, \delta(y))] = (0.05 - 0.02) \cdot 20\ell + (0.02 - 0.001) \cdot 100\ell + 0.001 \cdot 500\ell = 3 \cdot \ell.$$

As in (5) the resulting decision rule (4) is not Type-I risk safe, and again, the Type-I risk can even go to infinity with the number of potential actions.

1.2 E-Values to the Rescue

Reporting evidence as e-values (as defined by (6)) rather than p-values solves both the multiple action and post-hoc-loss issue identified above. An *e-value* is the value of a special type of statistic called an *e-variable*. An e-variable is any *nonnegative* random variable $S = S(Y)$ that can be written as a function of the observed Y and that satisfies the inequality:

$$\text{for all } P \in \mathcal{H}(\underline{0}): \mathbf{E}_P[S] \leq 1. \quad (6)$$

The e-variable's simplest application is in defining tests: the S -based hypothesis test at level α is defined to reject the null iff $S \geq 1/\alpha$. Since for any e-variable S , all $P \in \mathcal{H}(\underline{0})$, by Markov's inequality, $P(S \geq 1/\alpha) \leq \alpha$, with such a test we get a Type-I error guarantee of α , with the advantage that (as shown by [13, 52]) unlike with p-values, the Type-I error guarantee remains valid under *optional continuation*, i.e. deciding based on a study result whether new studies should be undertaken and if so, multiplying the corresponding e-values. The term 'e-variable' was coined in 2019 [13, 52] but their history is older, as described by [13, 37].

We may now simply pick any e-variable S we like and replace decision rule (4) by the following *maximally compatible* alternative rule: upon observing data $Y = y$ and loss function indexed by $B = b$ with accompanying maximum imposed risk bound ℓ , select the *largest* a for which

$$S^{-1}(y) \cdot L_b(\underline{0}, a) \leq \ell, \text{ i.e. } L_b(\underline{0}, a) \leq S(y) \cdot \ell, \quad (7)$$

where we adopt the (in our setting harmless) convention that, for $u = 0$ and $v \geq 0$, $u^{-1}v := 0$ if $v = 0$ and $u^{-1}v = \infty$ if $v > 0$ (in Section 5 we discuss where ℓ comes from). Theorem 1 below gives conditions under which (7) has a unique solution. For the original NP setting of two actions, (7) is simply the p-value based rule (4) with P replaced by $1/S$, illustrating that *large* e-values correspond to evidence against the null. But in contrast to the p-value based rule, with the e-based rule, no matter what e-variable S we take (as long as it is itself chosen before data are observed), no matter how many actions \mathcal{A} contains, no matter the process determining the loss B , we have the Type-I risk guarantee (1) (Theorem 1 below): replacing P by $1/S$ resolves the BIND problem. Of course, this raises the question whether p-values cannot be used safely for Type-I risk after all, in a manner different from (4). The only such method we know of is to first convert a p-value into an e-value and then use (7) after all. As discussed in the SI, the e-values resulting from such a conversion are usually suboptimal, so we prefer to design and use e-values directly.

Example 2 [The NP and LR E-Variables] As with p-values, many different e-variables can be defined for the same $\mathcal{H}(\underline{0})$. As discussed by [42], an extreme choice is to start with a fixed level α and p-value P and to set $S^{\text{NP}(\alpha)} := (1/\alpha)$ if $P \leq \alpha$ and $S^{\text{NP}(\alpha)} = 0$ otherwise. Clearly

$\mathbf{E}_{Y \sim P_0}[S^{\text{NP}(\alpha)}] \leq \alpha(1/\alpha) = 1$ so $S^{\text{NP}(\alpha)}$ is an e-variable. In the case of a classical, 2-action NP problem as defined underneath (3), the test (7) based on e-variable $S = S^{\text{NP}(\alpha)}$ will lead to $a = 1$ (reject the null) exactly iff the classical NP test based on P does. This shows that any P -based NP test can also be arrived at using (7) with a special e-value: nothing is lost by replacing p-values with e-values. Still, in case there are more than 2 actions and/or post-hoc decisions, while preserving the ℓ -Type-I risk guarantee, decisions based on $S^{\text{NP}(\alpha)}$ may not be very wise in the Type-II risk sense. For example, with the loss function used in (5) and $\alpha = 0.05$, we get that even for very small underlying P (i.e. extreme data), we will still choose action 1 whereas it seems more reasonable to select more extreme actions, minimizing Type-II loss, as the evidence against the null gets stronger. In case $\mathcal{H}(\underline{0}) = \{P_0\}$ and $\mathcal{H}(\underline{1}) = \{P_1\}$ are simple, this can be achieved by taking S to be a *likelihood ratio*: assuming P_j has density p_j ,

$$S^{\text{LR}} := \frac{p_1(Y)}{p_0(Y)} \quad (8)$$

which is immediately seen to be an e-variable:

$$\mathbf{E}_{P_0}[S^{\text{LR}}] = \int p_0(y) \frac{p_1(y)}{p_0(y)} d\mu = \int p_1(y) d\mu = 1, \quad (9)$$

i.e. to satisfy (6). We can compare $S^{\text{NP}(\alpha)}$ and S^{LR} if P underlying $S^{\text{NP}(\alpha)}$ is itself a monotonic function of the likelihood ratio S^{LR} , as it will be for the standard optimal power NP test. In the decision task above (4), when used in (7), $S^{\text{NP}(\alpha)}$ can, for each α , select at most 2 actions whereas S^{LR} leads to selection of action 0, 1, 2 or 3 depending on the amount of evidence, at the price of imposing a larger threshold before any particular action is selected compared to the S^α that is optimal for that action (e.g. $S^{0.05}$ is optimal for action 1 in this sense). We will see more sophisticated e-variables in Example 4 and refer to numerous further examples of useful e-variables in Section 4.

2 Mathematical Formalization and Results

2.1 Type-I Risk Safety and Compatibility

Let $\mathcal{H}(\underline{0})$, the null hypothesis, be a set of probability distributions for random Y taking values in a *outcome space* \mathcal{Y} .

Definition 1 A GNP (Generalized Neyman-Pearson) testing problem *relative to* $\mathcal{H}(\underline{0})$ is a tuple $(\mathcal{B}, \{(\mathcal{A}_b, L_b(\underline{0}, \cdot) : \mathcal{A}_b \rightarrow \mathbb{R}_0^+) : b \in \mathcal{B}\})$ where for all $b \in \mathcal{B}$, we call $L_b(\underline{0}, \cdot)$ the Type-I loss indexed by b with action space \mathcal{A}_b .

In Section 3 we extend the definition to uncertainty quantification beyond testing. Relative to any given GNP testing problem, we further define a *decision rule* to be any collection of functions $\{\delta_b : b \in \mathcal{B}\}$, where $\delta_b(y)$ denotes the $a \in \mathcal{A}_b$ picked when loss function indexed by $B = b$ (i.e. L_b) is presented and $Y = y$ is observed. Let δ be any decision rule and let $S = S(Y)$ be any e-variable. We call δ *compatible with* S if

$$L_b(\underline{0}, \delta_b(y)) \leq S(y) \text{ for all } y \in \mathcal{Y}, b \in \mathcal{B}. \quad (10)$$

We now prepare the definition of Type-I risk safety for GNP decision problems. First, we note that in general, the threshold ℓ a DM would like to impose on the risk via (7) when confronted

with loss function L_b may be an arbitrary positive real. However, using this maximal rule (7), for every observed $Y = y$ and $B = b$, the exact same decision will be taken if we normalize all losses, using L'_b with $L'_b(\underline{0}, a) = L_b(\underline{0}, a)/\ell$ instead of L_b and $\ell' = 1$ instead of ℓ . Hence, without loss of generality, from now on we simplify the treatment by taking $\ell = 1$ (in the SI we discuss in more detail why this is not harmful). With this in mind, consider a concrete setting in which the actual loss function L_B with index B presented to DM is determined in a data-dependent manner (perhaps by some policy makers, perhaps completely implicitly). Since we do not know the definition of B , i.e. how the choice is made, we want to ensure that the analogue of Eq. 1 holds, in the worst case, over all possible choices. Thus, as a first attempt, we may extend Eq. 1, by defining δ to be Type-I risk safe if

$$\sup_{P_0 \in \mathcal{H}(\underline{0})} \mathbf{E}_{P_0} \left[\sup_{b \in \mathcal{B}} L_b(\underline{0}, \delta_b(Y)) \right] \leq 1. \quad (11)$$

As discussed in the SI, the expectation might be undefined for pathological choices of the set \mathcal{B} and the functions $\{L_b : b \in \mathcal{B}\}$ — after all, we have not restricted the choice of \mathcal{B} and L_b at all. We can simply avoid this issue by slightly modifying the definition: we define δ to be *Type-I risk safe* if there exists a function $U : \mathcal{Y} \rightarrow \mathbb{R}_0^+$ such that for all $P_0 \in \mathcal{H}(\underline{0})$, $\mathbf{E}_{P_0}[U(Y)]$ is well-defined, and for all $y \in \mathcal{Y}$,

$$\sup_{b \in \mathcal{B}} L_b(\underline{0}, \delta_b(y)) \leq U(y) ; \quad \sup_{P_0 \in \mathcal{H}(\underline{0})} \mathbf{E}_{P_0} [U(Y)] \leq 1. \quad (12)$$

E-Variable Compatibility \Leftrightarrow **Type-I Risk Safety** In NP Theory, Type-I error guarantees come first — we look for an optimal decision rule among all rules that have the desired Type-I error guarantee. Analogously, here we first restrict our search for ‘good’ decision rules to those that are Type-I risk safe for the given decision problem. How to find these? Realizing that the second equation in (12) expresses that U is an e-variable, and the first equation says that δ is compatible with this e-variable, we see that the Type-I risk safe decision rules are exactly those that are compatible with an e-variable, thereby explaining the importance of e-variables to generalized NP testing. Formally, we have just proved the following trivial consequence of our definitions:

Proposition 1 *Fix an arbitrary GNP testing problem. For every δ defined relative to this problem:*

1. *For every e-variable S for $\mathcal{H}(\underline{0})$: if δ is compatible with S , then δ is Type-I risk safe.*
2. *Suppose that δ is Type-I risk safe. Let $S = U$ be as in (12) (in standard cases we can simply take $S(y) = \sup_{b \in \mathcal{B}} L_b(\underline{0}, \delta_b(y))$). Then S is an e-variable for $\mathcal{H}(\underline{0})$, and δ is compatible with S .*

2.2 Admissibility

We now turn to Type-II losses. The reader may have wondered why the specification of Type-II loss functions $L_b(\underline{1}, a) : \mathcal{A}_b \rightarrow \mathbb{R}$ as in (2) was not made part of Definition 1. The following crucial observation implies that this is superfluous, thereby greatly satisfying the treatment: suppose there were two actions $a, a' \in \mathcal{A}_b$ such that $L_b(\underline{0}, a') > L_b(\underline{0}, a)$ and $L_b(\underline{1}, a') > L_b(\underline{1}, a)$. Then any rational DM would always prefer a over a' , and hence never

want to play a' . We can thus take a' out of the set \mathcal{A}_b without affecting the set of decisions that a DM might ever want to consider. Assuming that \mathcal{A}_b has been pre-processed like this, we automatically obtain that the larger the Type-I loss of an action, the smaller the Type-II loss, allowing us to refrain from specifying $L_b(\underline{1}, \cdot)$: we may thus call a decision rule δ° *Type-II strictly better* than δ if for all $b \in \mathcal{B}$, all $P \in \mathcal{H}(\underline{0})$, we have

$$P(L_b(\underline{0}, \delta_b^\circ(Y)) < L_b(\underline{0}, \delta_b(Y))) = 0 \quad (13)$$

whereas there exist $b \in \mathcal{B}, P \in \mathcal{H}(\underline{0})$ such that

$$P(L_b(\underline{0}, \delta_b^\circ(Y)) > L_b(\underline{0}, \delta_b(Y))) > 0. \quad (14)$$

If \mathcal{Y} is uncountable, L_b, δ° and δ could again be picked in highly pathological ways, such that the probabilities above are undefined. This is fully resolved by the generalization of (13) and (14) given in the SI.

Clearly, if both δ° and δ are Type-I risk safe and δ° is Type-II strictly better than δ , we would always prefer playing δ° over δ . We may say that δ is *inadmissible*. Formally, for any decision rule δ we say that it is *admissible* if it is Type-I risk safe and no other Type-I risk safe decision rule is Type-II strictly better.

Main Result This admissibility notion is reminiscent of standard admissibility notions in classical statistical decision theory, and the theorem below is in the spirit of a *complete class theorem* [3, 9] expressing that in searching for reasonable (i.e., admissible) decision rules in GNP problems we may restrict ourselves to those based on e-variables via *maximally compatible decision rules*. Formally, we call a decision rule δ *maximally compatible* with e-variable S relative to a given GNP testing problem, if it is compatible with S and there exists no decision rule δ° such that δ° is also compatible with S yet δ° is Type-II strictly better than δ . We will relate this to the earlier informal definition of maximum compatibility ((7)) further below.

To state the theorem, we need one more concept: we call a GNP testing problem *rich* relative to e-variable $S = S(Y)$ if for every s in the co-domain of S , there exist $b \in \mathcal{B}$ and $a \in \mathcal{A}_b$ such that $L_b(\underline{0}, a) = s$. An example of a simple GNP testing problem that is rich relative to any e-variable at all is obtained whenever $\mathcal{B} = \{\text{SQ}\} \cup \mathcal{B}'$, for arbitrary \mathcal{B}' , where $\mathcal{A}_{\text{SQ}} = \mathbb{R}_0^+$ and $L_{\text{SQ}}(\underline{0}, a) = a^2$ (the squared error loss — richness follows since it can take on any value in \mathbb{R}_0^+). An example of a GNP testing problem that is rich relative to e-variable $S^{\text{NP}(\alpha)}$ of Example 2 is given by $\mathcal{B} = \{\text{NP}\} \cup \mathcal{B}'$, for arbitrary \mathcal{B}' , where $\mathcal{A}_{\text{NP}} = \{0, 1\}$ and $L_{\text{NP}}(\underline{0}, 0) = 0, L_{\text{NP}}(\underline{0}, 1) = 1/\alpha$ (if $\mathcal{B}' = \emptyset$, this is the classical NP setting of Section 1 again): choose $B = \text{NP}, a = 0$ if $S^{\text{NP}(\alpha)} = 0$, and choose $B = \text{NP}, a = 1$ if $S^{\text{NP}(\alpha)} = 1/\alpha$.

Theorem 1 *Consider a GNP testing problem. Then:*

1. *If δ is an admissible decision rule, then there exists an e-variable S such that δ is a maximally compatible decision rule for S .*
2. *As a partial converse, suppose that δ is a maximally compatible decision rule for some e-variable S . If (a) all $P \in \mathcal{H}(\underline{0})$ are mutually absolutely continuous (see below) and (b) S is sharp relative to the given testing problem, i.e. $\mathbf{E}_{P_0}[S] = 1$ for some $P_0 \in \mathcal{H}(\underline{0})$, and (c) the GNP testing problem is rich relative to S , then δ is admissible.*

Part 1 shows that we can restrict our search for admissible decision rules to the ones that are maximally compatible for some e-variable S . Part 2 is in essence a converse, showing that, under some regularity conditions, maximally compatible decision rules must be admissible. All three conditions required are weak: (a) Two distributions P, P' are mutually absolutely continuous if ‘they agree on what is practically impossible’, i.e. for each event \mathcal{E} , we have $P(\mathcal{E}) = 0$ iff $P'(\mathcal{E}) = 0$. Most standard parametric families are absolutely continuous or can be made such by excluding the boundary of the parameter space. (b) Sharpness of S expresses that S cannot be uniformly improved — a mild requirement satisfied by all e-variables considered in this paper (and also in most other papers on e-variables [37, 13]). (c) Richness relative to the S considered holds in all examples encountered in this paper (see Example 3 below for further illustration). More importantly perhaps, for any sharp e-variable S which we might want to base our decisions on, we can trivially *enlarge* any given GNP testing problem by adding one particular loss function so that the extended GNP decision problem will automatically be rich relative to S , and Part 2 of Theorem 1 can then be applied. In the SI we explain how this enlargement works and why it is a reasonable operation.

The theorem thus expressing that maximally compatible δ tend to coincide with admissible δ , we would still like to be assured that such maximally compatible δ exist in wide generality. To briefly illustrate that this is the case, at the same time connecting the formal notion to the earlier informal definition based on (7), let us consider what we will call *simple* GNP testing problems. All GNP testing problems encountered in this and the previous section are simple. They are defined as those GNP testing problems for which, (i) for all $b \in \mathcal{B}$, \mathcal{A}_b is a finite union of closed intervals in $\mathbb{R}_0^+ \cup \{\infty\}$ (in particular this includes the case that \mathcal{A}_b is finite); (ii) the Type-I loss $L_b(\underline{0}, a)$ is monotonically and, on each interval, continuously increasing in a , and (iii) all $P \in \mathcal{P}$ are mutually absolutely continuous. The following is easily checked: for arbitrary e-variable S , such simple GNP decision problems must have a maximally compatible δ^* relative to S that generalizes (7), with our simplification $\ell = 1$: δ^* is the rule which selects, when presented $Y = y, B = b$,

$$\delta_b^*(y) := \text{largest } a \in \mathcal{A}_b \text{ with } L_b(\underline{0}, a) \leq S(y). \quad (15)$$

Moreover, this maximally compatible δ^* is essentially unique, i.e. if δ^*, δ' are both maximally compatible, then for all $P \in \mathcal{H}(\underline{0})$, we have $P(\delta^* \neq \delta') = 0$.

Example 3 Consider a simple vs. simple testing problem with $\mathcal{H}(\underline{0}) = \{P_0\}, \mathcal{H}(\underline{1}) = \{P_1\}$. Let $\mathbb{P}(Y)$ be a strict p-value, i.e. $P_0(\mathbb{P} \leq \alpha) = \alpha$ for $\alpha \in [0, 1]$, that is monotonically and continuously decreasing in the likelihood ratio $S^{\text{LR}}(Y)$; use of such a p-value is standard in NP testing with continuous-valued outcome spaces. Consider the following variation of Example 1: $\mathcal{B} \subset \mathbb{R}_0^+$ with for $b \in \mathcal{B}$, $\mathcal{A}_b = \{0, 1\}$ and $L_b(\underline{0}, 0) = 0, L_b(\underline{0}, 1) = b$. Take arbitrary but fixed $0 < \alpha < 1$. Then the maximally compatible decision rule δ^* as in (15) relative to e-variable $S^{\text{NP}(\alpha)}$ is sharp. When presented with loss function L_b , this δ^* always plays 0 if $b > 1/\alpha$. If $b \leq 1/\alpha$, it plays 1 if $b \leq S^{\text{NP}(\alpha)}$ (i.e. if $S^{\text{NP}(\alpha)} = 1/\alpha$, i.e. if $\mathbb{P} \leq \alpha$) and 0 otherwise (i.e. if $S^{\text{NP}(\alpha)} = 0$, i.e. if $\mathbb{P} > \alpha$). By Part 2 of Theorem 1, this δ^* is admissible if \mathcal{B} contains $b = 1/\alpha$, which ensures richness relative to $S^{\text{NP}(\alpha)}$.

In contrast, consider the δ^* as in (15) based on the likelihood ratio e-variable S^{LR} , which is also sharp. When presented L_b , this decision rule plays 1 if $b \leq S^{\text{LR}}$ and 0 otherwise. If we set $\mathcal{B} = \mathbb{R}_0^+$, we have richness relative to S^{LR} so by Theorem 1, this δ^* is admissible as well. In this case though, admissibility of δ^* may fail if we take \mathcal{B} a strict subset of \mathbb{R}_0^+ .

3 Robust Confidence via the E-Posterior

Now let us consider a statistical model \mathcal{P} partitioned according to a parameter of interest $\theta \in \Theta$, with $\phi : \mathcal{P} \rightarrow \Theta$ indicating the parameter corresponding to each P ; for example, $\theta = \phi(P)$ might be the mean of P , or, if $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric model, ϕ might simply denote the parameterization function, $\phi(P_\theta) = \theta$. Any collection of p-values $\{P_\theta : \theta \in \Theta\}$, with P_θ a p-value for the null $\mathcal{H}(\theta) := \{P \in \mathcal{P} : \phi(P) = \theta\}$ can be used to build a valid $(1 - \alpha)$ confidence set, by setting $\text{CS}_\alpha(Y) = \{\theta : P_\theta(Y) > \alpha\}$ to be the set of θ 's that would not have been rejected at the given level α . For simplicity, we restrict attention to scalar $\Theta \subseteq \mathbb{R}$; then the CS_α will usually be intervals, and indeed this p-value based construction is a standard way to construct such intervals. Analogously [58, 37], any *e-collection*, i.e. a collection of e-variables $\{S_\theta : \theta \in \Theta\}$ such that S_θ is an e-variable for the ‘null’ $\mathcal{H}(\theta)$ (by this we mean that S_θ must satisfy (6), i.e. $\mathbf{E}_P[S_\theta] \leq 1$, for all $P \in \mathcal{H}(\theta)$) can be used to build an equally valid, usually larger, *e-based* $(1 - \alpha)$ -confidence set (again, for scalar θ this usually becomes an interval), one for each α , by setting $\text{CS}_\alpha^*(Y) = \{\theta : S_\theta(Y) < 1/\alpha\}$ as the set of θ 's that would not have been rejected at level α with an e-value based test. Below we first give a simple example. We then, in Section 3.3.1 retrace the steps of Section 1 and Section 2, re-interpreting confidence sets in terms of actions with associated losses and risks. Section 3.3.2 and 3.3.3 show that, once again, if losses are determined post-hoc (BIND is violated), then standard confidence intervals lose their validity whereas e-based confidence intervals remain Type-I risk safe. Relatedly, without BIND, decisions based on *confidence distributions* can be unsafe, but those based on the e-posterior — a means of summarizing e-CS's for all α 's at once — remain Type-I risk safe.

Example 4 Consider the normal location family: data are $Y = X^n$ where, under P_θ , $X^n = (X_1, \dots, X_n)$ are i.i.d. $\sim N(\theta, 1)$. We consider e-based confidence intervals based on various e-collections. [15] gives various suitable collections, but for simplicity we here stick to a single, simple choice, taken from Example 8 of [15], that, like the standard CI, is symmetric around the MLE $\hat{\theta}(X^n) = n^{-1} \sum X_i$. Fix *anticipated* sample size n^* and confidence level $0 < \alpha^* < 1$. For each θ we define $\theta^- < \theta$ and $\theta^+ > \theta$ to satisfy

$$\frac{1}{2}n^*(\theta - \theta^+)^2 = \frac{1}{2}n^*(\theta - \theta^-)^2 = \log \frac{2}{\alpha^*}. \quad (16)$$

Now define e-variables $S_\theta^-(y) = p_{\theta^-}(y)/p_\theta(y)$, $S_\theta^+(y) = p_{\theta^+}(y)/p_\theta(y)$ and $S_\theta(y) = (1/2)(S_\theta^-(y) + S_\theta^+(y))$. These choices can be motivated based on the fact that S_θ^- and S_θ^+ are also uniformly most powerful Bayes factors [23, 15] and hence reasonable e-variables for 1-sided CSS. We continue with S_θ for two-sided CSS. As is proved analogously to (9), S_θ remains an e-variable even if neither the actual sample size n nor the to-be-used significance level $0 < \alpha < 1$ are equal to the hoped-for n^* and α^* ; more on this below. In the SI we show that a sufficient condition for $S_\theta(Y) \geq \alpha^{-1}$, i.e. for $\theta \notin \text{CS}_\alpha^*(Y)$ is that

$$\begin{aligned} |\theta - \hat{\theta}| &\geq \sqrt{\frac{2}{n} \cdot \log \frac{2}{\alpha}} \cdot g(c), \text{ with} \\ c &= \frac{n^*}{n} \cdot \frac{\log(2/\alpha)}{\log(2/\alpha^*)}, \quad g(c) = \frac{1}{2} \left(c^{1/2} + c^{-1/2} \right). \end{aligned} \quad (17)$$

As explained in the SI, for fixed α , (17) is tight for all but the smallest n . Thus, the e-based confidence interval $\text{CS}_\alpha^*(Y)$ has width $|\theta - \hat{\theta}| \asymp 1/\sqrt{n}$, of the same order as the region for the standard Neyman-Pearson test, with a factor $g(c)$ depending on how well aligned n, n^*, α and α^* are: $g(c)$ and hence the width is minimized, if $c = 1$ (and then $g(c) = 1$) which is the case if $n = n^*$ and $\alpha = \alpha^*$. At $\alpha = .05$, we get in this optimal case that $S_\theta(Y) \geq \alpha^{-1}$ if $|\theta - \hat{\theta}| \geq \sqrt{(2 \log 40)/n} \approx 2.72/\sqrt{n}$, making the e-based CI wider than the standard CI by a constant factor of $\approx 2.72/1.96 \approx 1.4$ (see Figure 1).

E-Processes Why do we not simply set n^* in the definition of S_θ actual to the actual n ? The reason is that we allow the actual n to be unknown in advance, and even to be random (i.e. a stopping time with unknown definition): it is easily seen that, if $Y = X^\tau$ with $X^\tau = (X_1, \dots, X_\tau)$ for a stopping time τ (whose definition may be unknown to DM), then $S_\theta(X^\tau)$ is still an e-variable. Formally, $S_\theta(X^1), S_\theta(X^2), \dots$ constitutes an *e-process* in the sense of [38, 37]. Thus, we can use S_θ without knowing the definition of τ , and in particular, τ may be unequal to n^* — such as an extension of e-variables to be used with arbitrary stopping times is often, but not always possible [13]; whenever it is, it provides an additional bonus over use of standard p-variables in testing, which require the stopping time to be set in advance. As stated, assuming that we base the S_θ on the correct n and α , this e-based confidence interval is about 1.4 times as wide as the standard one; the inevitable (yet, I feel, worthwhile!) price to pay for the added flexibility and robustness: in contrast to the standard one, we can use the e-based interval for unknown n (or τ) as well, and we can also use it to get valid confidence intervals for B if BIND is violated, as we proceed to show.

3.1 Reformulating Coverage in terms of Type-I Risk

We now generalize the definition of GNP testing problem so that (besides much else) it also allows for estimation with confidence intervals.

Definition 2 Fix a set of distributions \mathcal{P} for Y , a set Θ and a function $\phi : \mathcal{P} \rightarrow \Theta$ mapping $P \in \mathcal{P}$ to property $\phi(P) \in \Theta$ as above. A GNP (Generalized Neyman-Pearson) decision problem relative to \mathcal{P} , Θ and ϕ is a tuple $(\mathcal{B}, \{(\mathcal{A}_b, L_b : \Theta \times \mathcal{A}_b \rightarrow \mathbb{R}_0^+) : b \in \mathcal{B}\})$.

A GNP decision problem is really a set of GNP testing problems, one for each $\theta \in \Theta$: we recover Definition 1 by taking a singleton $\Theta = \{\underline{0}\}$, $\mathcal{P} = \mathcal{H}(\underline{0})$ and $\phi(P) = \underline{0}$ for all $P \in \mathcal{H}(\underline{0})$. For general $\theta \in \Theta$, the θ -testing problem corresponding to the GNP decision problem is the testing problem $(\mathcal{B}, \{(\mathcal{A}_b, L_b(\theta, \cdot) : \mathcal{A}_b \rightarrow \mathbb{R}_0^+) : b \in \mathcal{B}\})$ with null hypothesis $\mathcal{H}(\theta) = \{P : \phi(P) = \theta\}$ and with $L_b(\theta, \cdot)$ in the role of $L_b(\underline{0}, \cdot)$. All definitions for GNP testing problems are now easily extended to GNP decision problems by requiring them to hold for the corresponding θ -testing problem, for all $\theta \in \Theta$. In particular, we say that decision rule δ is compatible with e-collection $\{S_\theta : \theta \in \Theta\}$ if we have for all $y \in \mathcal{Y}$, $b \in \mathcal{B}$ that

$$\forall \theta \in \Theta : L_b(\theta, \delta_b(y)) \leq S_\theta(y). \quad (18)$$

The definition of Type-I risk safety is extended analogously from (12): δ is Type-I risk safe iff there exists an e-collection $S = \{S_\theta : \theta \in \Theta\}$ such that δ is compatible with S . If the expectation below is well-defined (which it will be in the confidence interval setting), Type-I risk safety is then clearly equivalent to the corresponding generalization of (11):

$$\sup_{\theta \in \Theta} \sup_{P \in \mathcal{H}(\theta)} \mathbf{E}_P \left[\sup_{b \in \mathcal{B}} L_b(\theta, \delta_b(Y)) \right] \leq 1. \quad (19)$$

Admissibility is extended analogously: we call a decision rule δ° *Type-II strictly better* than δ if for all $\theta \in \Theta$, the corresponding θ -testing problem satisfies (13) with $\underline{0}$ replaced by θ , whereas there exist $\theta \in \Theta$, $b \in \mathcal{B}$, $P \in \mathcal{H}(\underline{0})$ such that the corresponding θ -testing problem satisfies (14) with $\underline{0}$ replaced by θ . The definition of admissibility and maximum compatibility are now based on this extended notion of Type-II strictly-betterness and otherwise unchanged; we further extend the notions of sharpness and richness to this generalized setting and provide a generalization of Theorem 1 to full GNP decision problems in the SI Appendix.

Confidence Intervals as Actions We now instantiate the above to estimation of confidence intervals. Given a probability model \mathcal{P} and parameter of interest $\theta \in \Theta \subset \mathbb{R}$ with $\theta = \phi(P)$ as above, take the GNP decision problem with this Θ and ϕ , and with $\mathcal{B} = [1, \infty)$, $\mathcal{A}_b = \{[\theta_L, \theta_R] : \theta_L, \theta_R \in \Theta, \theta_L \leq \theta_R\}$,

$$L_b(\theta, [\theta_L, \theta_R]) = b \cdot \mathbf{1}_{\theta \notin [\theta_L, \theta_R]}. \quad (20)$$

Thus, we incur a Type-I loss, if the sampling distribution θ is not in the interval $[\theta_L, \theta_R]$ we specified, and b determines how bad such a mistake is — this may again be data-dependent: we assume once again that we are presented $B = b$ via a random and potentially unknowable process, and we want to obtain the Type-I risk guarantee (19), which instantiates to

$$\sup_{\theta \in \Theta} \sup_{P \in \mathcal{H}(\theta)} \mathbf{E}_P[\sup_{b \in \mathcal{B}} b \cdot \mathbf{1}_{\theta \notin \delta_b(Y)}] \leq 1, \quad (21)$$

where $\delta_b(Y) = [\theta_L(Y, b), \theta_R(Y, b)]$. Among all decision rules (i.e. confidence intervals) δ satisfying (21), we want to find the narrowest ones. Our definition of Type-II strictly better above ‘automatically’ accounts for this: the extended definition of Type-II betterness implies that $[\theta_L, \theta_R]$ is Type-II strictly better than $[\theta'_L, \theta'_R]$ iff $[\theta_L, \theta_R]$ is a proper subset of $[\theta'_L, \theta'_R]$.

If we may assume that BIND holds we can take the supremum over \mathcal{B} in (21) out of the expectation, i.e. $b\mathbf{E}_P[\mathbf{1}_{\theta \notin \delta_b(Y)}] \leq 1$ must hold for all fixed $\theta, P \in \mathcal{H}(\theta)$ and b . We may then think of b as set in advance: if we set $b = 1/\alpha$ for some $0 < \alpha \leq 1$, then the requirement says that $\delta_b(Y)$ is a standard $(1 - \alpha)$ -confidence interval. Thus, under BIND, standard confidence intervals δ_b coincide with Type-I risk safe confidence intervals as defined above: just as for p-value based tests in Section 1, under BIND the new setting is simply an equivalent reformulation of the existing theory of confidence sets. Yet, again, if BIND is violated, then standard confidence intervals are not Type-I safe any more, whereas e-based confidence intervals still are.

Example 5 [Ex. 4, Continued] Suppose you observe $Y = y$, $B = b$. Let us use the e-confidence intervals as defined relative to a particular anticipated n^* and α^* . Using (17) and substituting $1/b$ for α (so that now $c = (n^*/n) \cdot (\log(2b)/(\log(2/\alpha^*)))$), gives that $\forall \theta \in \Theta : L_b(\theta, \delta_b(y)) \leq S_\theta$ with $\delta_b(y) = [\theta_L, \theta_R]$ (i.e. compatibility ((18)) holds and hence Type-I risk safety (21) holds as well) as soon as $\theta_L \leq \hat{\theta} - A$ and $\theta_R \geq \hat{\theta} + A$ where

$$A = \sqrt{\frac{2}{n}} \cdot \sqrt{\log(2b)} \cdot g(c). \quad (22)$$

We may choose $\delta_B(Y) = [\hat{\theta} - A, \hat{\theta} + A]$ to satisfy this with equality to make the interval as narrow as possible, making our interval admissible. We are then guaranteed Type-I safety, (21), irrespective of the definition of B . In contrast, it is not clear how to construct Type-I

safe CI's for data-dependent B without e-values. We might be tempted to do this based on *confidence distributions* (CD's) [6, 41] that summarize confidence intervals for each α into a posterior-like quantity, or *objective Bayes posteriors* [3], but as we now show, this can have bad results.

3.2 CD's and O'Bayes Posteriors are not valid Post-Hoc

Consider the normal location family again. With the standard (uniform, improper) 'objective Bayes' prior for this family and data $Y = x^n$, the posterior $W^\circ | Y = x^n$ has a normal density $w^\circ(\theta | x^n)$ with mean and median equal to the MLE $\hat{\theta}(x^n)$ and variance $1/n$ [3]. In this case the objective Bayes posterior also coincides with the *fiducial* [17] and the *confidence distribution* (CD) [41] based on x^n , and has an exact coverage property: if we let $[\theta_L, \theta_R]$ represent the standard $(1 - \alpha)$ -Bayesian credible interval based on $w^\circ(\theta | x^n)$, i.e. taken symmetrically around the MLE $\hat{\theta}(x^n)$ then this coincides exactly with the standard $(1 - \alpha)$ -confidence interval, e.g. for $\alpha = 0.05$, we have $\theta_L = \hat{\theta}(x^n) - 1.96/\sqrt{n}$, $\theta_R = \hat{\theta}(x^n) + 1.96/\sqrt{n}$.

The question is now how to base inferences on the objective Bayes or CD's within our current GNP decision problem, i.e. if the goal is to come up with an interval as narrow as possible that contains the true θ , where making a mistake is weighted by some B that is determined post-hoc. Upon observing $Y = x^n$ and $B = b$, and based on the CD $w^\circ(\theta | Y)$, one would presumably pick the smallest interval symmetric around $\hat{\theta}$ for which the Bayes posterior satisfies the required risk bound, i.e. $\delta'_b(Y) = [\hat{\theta} - A, \hat{\theta} + A]$ where A , depending on b , is the smallest number such that

$$\mathbf{E}_{\bar{\theta} \sim W^\circ | Y = x^n} [L_b(\bar{\theta}, \delta'_b(x^n))] \leq 1, \quad (23)$$

$$\text{i.e. } \mathbf{E}_{\bar{\theta} \sim W^\circ | Y = x^n} [b \cdot \mathbf{1}_{\bar{\theta} \notin [\hat{\theta} - A, \hat{\theta} + A]}] \leq 1, \quad (24)$$

with b the observed value taken by B ; and for this smallest A , (24) holds with equality. Since $W^\circ(\bar{\theta} \notin [\hat{\theta} - A, \hat{\theta} + A] | Y = x^n) = 1/b$, this $\delta'_b(Y)$ is equal to the standard $(1 - \alpha)$ -confidence interval for $\alpha = 1/b$. The intuitive appeal for choosing this δ is clear: (24) expresses that as a DM one can expect the loss given the data to be bounded by 1; one simply wants to pick the smallest, most informative interval for which this holds true. Yet the *real* expectation of the loss may very well be different from (24) — assuming that B is a fixed function of Y , it is given by

$$\mathbf{E}_{Y \sim P_{\theta^*}} [B(Y) \cdot \mathbf{1}_{\theta^* \notin \delta'_{B(Y)}(Y)}], \quad (25)$$

with θ^* indexing the true sampling distribution. This quantity may be much larger than 1 if B is dependent on Y . We provide a simple yet extreme example (inspired by the less extreme Example 8 of [14]) with $n = 1$, $Y = X_1$ (equivalently, think of Y as the Z-score corresponding to a larger sample): fix any $\epsilon > 0$. If, whenever $Y \geq \epsilon$, we set $B := 1/(2F_0(-Y + \epsilon^2/Y))$ where F_0 is the CDF of a standard normal, then, as demonstrated in the SI, under $\theta^* = 0$, (25) evaluates to ∞ , irrespective of the definition of $B(Y)$ for $Y < \epsilon$. In particular we may set $B = 1$ for such Y , corresponding to the decision problem being 'called off', because the required bound (24) is then achieved trivially by issuing the empty interval.

Repercussions for Neyman's Inductive Behavior This discrepancy between what one *believes* will happen according to a posterior (risk bounded by 1) and what actually will happen (potentially infinite risk) has repercussions for Neyman's interpretation of statistics

as long-run performance guarantees of inductive behavior. To illustrate, imagine a DM who is confronted with such a decision problem many times (each time j the underlying $\theta_{(j)}$ with $Y_{(j)} \sim P_{\theta_{(j)}}$ and the sample size $n_{(j)}$ and the importance function $B_{(j)}$ may be different). Then, based on (24) she might think to have, by the law of large numbers, the guarantee that, almost surely,

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \sum_{j=1}^m B_{(j)} \cdot \mathbf{1}_{\theta \notin [\theta_L(Y_{(j)}), \theta_R(Y_{(j)})]} \leq 1. \quad (26)$$

Unfortunately however, this statement is likely false if in reality there is a dependence between $B_{(j)}$ and $Y_{(j)}$. In the SI we show that, based on the example above, the average in (26) may in fact a.s. converge to infinity, even though the individual $B_{(j)}$'s look pretty innocuous. A first reaction may be to require the DM to address this problem by modeling the dependency between $B_{(j)}$ and $Y_{(j)}$. But the precise relation may be unknowable, and then it is not clear how to do this. To avoid the issue one may output e-based CIs or, equivalently but perhaps more illuminatingly, CIs based on the *e-posterior* that we now introduce.

3.3 The E-Posterior remains valid Post-Hoc

Let $S = \{S_\theta : \theta \in \Theta\}$ be an e-collection. Just like it is tempting to interpret a ‘system’ of confidence intervals, one for each α , i.e. a CD, as a type of ‘posterior’, one can also view the S_θ -reciprocal $\bar{P}(\theta | Y) := S_\theta^{-1}(Y)$ as a type of ‘posterior representation of uncertainty’ for parameter θ . This idea has been conceived of independently by [57] and [15], who called $\bar{P}(\theta | Y)$ the *e-posterior*. The crucial difference between e-posteriors and CDs is that the former enable valid inferences under specific post-hoc, data-dependent assessments of Type-I risk, whereas standard CD’s can only be validly used as in (23) if BIND holds. We thus recommend e-posteriors, like Cox [6] did for CD’s, as a summary of estimation uncertainty — but a summary that is significantly more robust than that provided by CD’s.

Using the e-posterior we can re-express compatibility, (18), as

$$\sup_{\theta \in \Theta} \bar{P}(\theta | y) \cdot L_b(\theta, \delta_b(y)) \leq \ell, \quad (27)$$

with conventions about $0 \cdot \infty$ as underneath (7) and $\ell = 1$. We already know that δ satisfying (27) are Type-I risk safe irrespective of how B is defined. The rewrite suggests an analogy to the Bayes posterior risk assessment, (23): if we replace objective Bayes/CD-posterior *expectation* by e-posterior *maximum*, we get Type-I risk safety without the BIND assumption.

[15] shows that, for general bounds ℓ and with L_b replaced by general loss functions, without Type-I/II-dichotomies, assessment (27) is meaningful and provides a non-Bayesian alternative for Bayes-posterior expected loss assessment. In that paper, I also list a variety of e-posteriors, including an extension of the one of Example 4 to general exponential families, and point out deeper relations between e-posteriors and Bayesian posteriors. In the present paper, we merely present the e-posterior as a graphical tool which summarizes the e-based confidence intervals as given by (18) and helps to visualize how they relate to standard confidence intervals: see Figure 1.

4 State of the Art

The modern development of e-values and e-processes started only in 2019 when first versions of the four ground-breaking papers [13, 52, 42, 56] appeared on arxiv. Since then, development

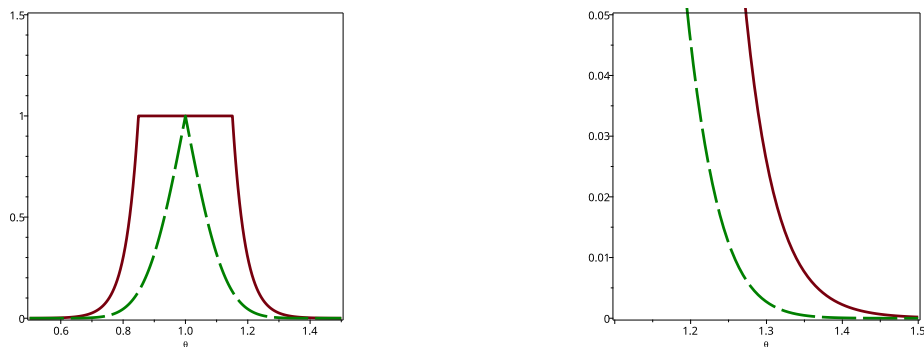


Figure 1: The solid line depicts the e-posterior corresponding to the e-collection of Example 4 capped at 1, i.e. $\min\{1, \bar{P}(\theta | y)\}$, for data $y = x^n$ with $n = 100$ and MLE $\hat{\theta}(y) = 1.00$. The dashed green line depicts twice the tail area of the objective Bayes posterior (CD) $W^\circ | Y = x^n$ of (24), given by $f(\theta) := 2W^\circ(\bar{\theta} \geq \theta | y) = 2 \int_\theta^\infty w^\circ(\bar{\theta} | y) d\theta$. The standard two-sided $(1 - \alpha)$ -confidence interval is given by $[\theta_L, \theta_R]$ where $\theta_L < \hat{\theta} = 1.00$ is the leftmost θ at which the dashed green curve takes value α , and θ_R is the rightmost such θ . The $(1 - \alpha)$ -e-confidence interval based on S_θ as defined above (17), and with boundaries approximately equal to, (17), is given by the $[\theta_L, \theta_R]$ at which the solid line takes value α . The right picture zooms in on the right for $\alpha \leq 0.05$. As expected, the dashed line hits 0.05 for $\theta_R = \hat{\theta}(x^n) + 1.96/\sqrt{n} = 1.196$, the solid (e-based) line at $\theta_R = \hat{\theta}(x^n) + 2.72/\sqrt{n} = 1.272$.

has been remarkably fast, often centering around GRO (*growth-rate optimal*) e-variables and -processes (see [13, 37] who also provide more historical context). Growth rate (also called *e-power* [60]) is a natural analogue of power in the e-process setting, in which sample sizes are not fixed in advance, related to the minimum sample size needed to reject the null and CI width. As a first step, GRO e-methods were successfully developed for basic workhorses of statistics such as the z-test (the one appearing in Example 4 has GRO status), the t-test [13], the test of two proportions [49] and the logrank test [46], which has been successfully deployed in a ‘live’ meta-analysis of ongoing clinical trials [47]. The t-test setting has been extended to general tests and CI’s with group invariances and linear and nonparametric Gaussian process regression [35, 26, 31]. The test of two proportions has been extended to k -sample tests of general exponential families [18], CMH tests [50] and to conditional independence (of X and Y given Z) testing under a *model-X assumption* (combineable with arbitrary models for $Y | X, Z$) [16]. E-variables that are not GRO yet still have good power-like properties have been quite successful in multiple testing applications [55, 22] as well as several other nonparametric problems [36, 57, 19] — the recent overview [37] provides a comprehensive list. In all applications mentioned, the qualitative behavior is similar to that of Example 4, with CI widths of order $O(f(n)/\sqrt{n})$ per parameter of interest, $f(n)$ varying from constant to $O(\sqrt{\log n})$, depending on the specific application.

(Current) Limitations and Challenges These initial successes notwithstanding, the development of e-values is, of course, still in its infancy, competing with almost a century of p-value development. As such, many challenges remain. To appreciate these, we first note that the aforementioned GRO-type approaches can in principle be made competitive, in terms of sample sizes needed to draw a conclusion, with classical ones that rely on BIND — see below; sometimes they even significantly beat such classical methods (e.g. [50, 57]). Also, [13] shows that GRO e-values exist and can be calculated for very general testing problems. Yet in general, this calculation is not efficient. In some cases (such as [35, 50, 16] mentioned above), they admit an analytical and hence efficiently calculable expression, but for others, they do not. These hard cases include regression (i.e. $Y = f_\theta(X, Z) + \text{NOISE}$) that involves a nonlinearity, such as GLMs and Cox proportional hazards, whenever the variable X to be tested (e.g. treatment vs. control) does *not* satisfy the model-X assumption (conditional distribution of X given Z known). While model-X is automatically satisfied in clinical trials, there are of course many important cases in which it is not. *Universal inference* [56] provides an alternative generic e-design method that does lead to efficiently calculable e-values in such cases, but in regression problems it is not competitive in terms of power with classical methods for medium- to high-dimensional models [48] — its strength has rather been to provide e-values for complex $\mathcal{H}(\mathbb{Q})$ that have simply eluded classical testing [10].

Challenges – II With GRO-type methods one can obtain comparable performance in terms of power as compared to classical approaches. In many (not all) settings though, one needs to engage in optional stopping to achieve this. For a broad class of e-values, this is no problem ([13] provides a detailed analysis): all coverage and Type-I risk guarantees are retained under such optional stopping. Still, it points towards a second challenge for e-methods, sociological/psychological rather than statistical: it requires researchers to think differently, and this is, of course, always difficult to accomplish. In this respect, the tech industry is at the forefront: anytime-valid methods based on e-processes have been adopted

by several major tech companies [26].

5 Discussion, Future Work, Conclusion

We provide a few concluding remarks. First we analyze in what sense we solved the ‘roving α ’ issue that motivated this work. Second, we discuss related work. Third, we suggest a ‘road ahead’ for e-methods.

Roving α Revisited: the Quasi-Conditional Paradigm Assume we have a prior on $\mathcal{H}(\underline{0})$ and $\mathcal{H}(\underline{1})$ and priors W_0 and W_1 on the distributions inside these hypotheses. We can then use Bayes’ theorem to calculate the Bayes posterior $P(\mathcal{H}(\underline{0}) \mid Y)$ based on data Y . Suppose we reject the null if $P(\mathcal{H}(\underline{0}) \mid Y) \leq \alpha$. We may then define, for all y for which this holds, i.e. for which we reject the null, the *conditional Type-I error probability* $\hat{\alpha}$ to be simply equal to this posterior probability, $\hat{\alpha} := \hat{\alpha}_{|y} := P(H_0 \mid Y = y)$. This implies that, for any fixed $\alpha_0 \leq \alpha$, for any long sequence of studies, with probability tending to one,

$$\begin{aligned} &\text{“among all studies with } \hat{\alpha} \leq \alpha_0, \text{ we make a} \\ &\text{Type-I error at most a fraction } \alpha_0 \text{ of the time”}. \end{aligned} \tag{28}$$

Such a fully conditional statement, with post-hoc determined $\hat{\alpha}_{|Y}$, is only correct if the priors can be fully trusted, i.e. if one accepts a fully subjective Bayesian stance. It would definitely be incorrect if we set $\hat{\alpha}_{|Y}$ either to a p-value or the reciprocal of an e-value based on Y . Still, as we have seen, if we instead use e-values to perform a data-dependent action, which is allowed to get more extreme (higher Type-I loss) as our evidence against the null increases (higher e-value) according to the maximally compatible rule (which in simple cases is given by (7)), then we *do* get an ‘unconditionally’ valid bound on Type-I risk. Thus, using e-values, setting a roving α to be equal to $\hat{\alpha} := \ell/S(y)$ for the observed y is still incorrect if we interpret it as expressing (28); but it is correct if we interpret it as setting a ‘roving bound’ of $\ell/\hat{\alpha}$ on the Type-I loss $L_B(\underline{0}, a)$ we dare to make: if we make sure to pick a so that $L_B(\underline{0}, a) \leq \ell/\hat{\alpha}$, then we have compatibility and hence Type-I risk safety, (11). Note that B is allowed to be any function of, hence ‘conditional on’ data; but its performance is evaluated ‘unconditionally’, i.e. by means of (11) which is an unconditional expectation. This *quasi-conditional stance*, explained further in [15], provides a middle ground between fully Bayesian and traditional Neyman-Pearson-Wald type methods and analysis.

Where does the Type-I risk bound ℓ come from? Whereas B may arbitrarily depend on data Y , the upper bound ℓ in (1) has to be set independently of Y after all. It may still vary from decision problem to decision problem though (in the SI Appendix we explain what this means in terms of Neyman’s inductive behavior paradigm and we explain that setting $\ell = 1$, as was done for mathematical convenience in Section 2, is unproblematic). In many practical testing problems, we might expect that for all $b \in \mathcal{B}$, \mathcal{A}_b contains a special action 0, which stands for ‘do nothing’ (keep status quo), which would then have the same Type-II loss under all $b \in \mathcal{B}$, i.e. there is an ℓ' such that for all $b \in \mathcal{A}_b$, $L_b(\underline{1}, 0) = \ell'$. We might then simply set $\ell = \ell'$, making sure that we can expect our result (with all costs and benefits incorporated), whatever action we take, to be no worse than “the cost of doing nothing when we really should have done something”.

Related Work: Inferential Models Like we do in Example 4, Martin, Liu and collaborators [28, 27] point out discrepancies between what one would expect to be a valid confidence set according to a fiducial, CD or Bayesian posterior and what are actually valid confidence sets according to the unknown, true distribution. They provide *inferential models (IMs)* as a safer alternative. Unlike e-posteriors, the specific IMs proposed by [28] still work under the BIND assumption and thus will not provide reliable inferences if BIND does not hold. But it may very well be that some other IMs (IMs constitute a family of methods, not a single method) essentially behave like e-posteriors.

The Road Ahead Future work will include a further investigation of the ‘quasi-conditional’ idea launched above, as well as of the precise relation to Martin’s IMs and other related uses of e-variables such as [1] who, like us, employ e-values with a Type-I/II-error distinction with more than 2 actions.

Another unresolved fundamental issue is this: most practitioners still interpret p-values in a Fisherian way, as a notion of *evidence* against $\mathcal{H}(0)$. Although this interpretation has always been controversial, it is to some extent, and with caveats (such as ‘single isolated small p-value does not give substantial evidence’ [29] or ‘only work with special, *evidential* p-values [12]’), adopted by highly accomplished statisticians, including the late Sir David Cox [7, 30]. Even Neyman [33] has written ‘my own preferred substitute for ‘do not reject H ’ is ‘no evidence against H is found’. In light of the present results, one may ask if, perhaps, *e-values are more suitable than p-values* as such a measure. We preliminarily conjecture they are, and motivate this in the SI — although a proper analysis of such a claim warrants a separate paper, which we hope to provide in the future.

Perhaps more important for practice than all of this though, in light of Section 4 above, is the further development of practically useful e-variables for standard settings (such as GLMs) in which they are not yet available, as well as more accompanying software such as [51].

Conclusion: A different kind of Robustness Standard P and CS-based decision rely on BIND, an assumption that will often be false or unverifiable at the time study results are published. In this paper we showed that e-values provide valid error and risk guarantees without making such assumptions, and are therefore *robust* tools for inference. But whereas ‘robustness’ usually refers to robust inference in the presence of outliers, or model structure or noise process misspecification, this is a different, much less studied form of robustness: robustness in terms of the actual decision task that the study results will be used to solve.

Acknowledgements The author would like to thank an anonymous referee and Dr. W. Koolen, who both independently alerted him to the fact that, without essential loss of generality, one may assume Type-II loss to decrease whenever Type-I loss increases.

References

- [1] S. Bates, M. I. Jordan, M. Sklar, and J. Soloff. Principal-agent hypothesis testing. *arXiv:2205.06812*, 2022.
- [2] J. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence (with discussion and rejoinder). *Journal of the American Statistical Association*, 82(397):112–122,135–139, 1987.

- [3] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 1985.
- [4] J.O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–12, 2003.
- [5] Patrick Billingsley. *Probability and Measure*. Wiley, third edition, 1995.
- [6] David R Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.
- [7] David R. Cox. In gentle praise of significance tests, 2018. Talk given at RSS 2018 keynote conference session.
- [8] D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.
- [9] Haosui Duanmu and Daniel M. Roy. On extended admissible procedures and their nonstandard Bayes risk. *Annals of Statistics*, 49(4):2053–2078, 2021.
- [10] Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal inference meets random projections: a scalable test for log-concavity. *arXiv:2111.09254*, 2021.
- [11] A.W.F. Edwards. *Likelihood*. Cambridge University Press, 1984.
- [12] Sander Greenland. Divergence vs. decision p-values: A distinction worth making in theory and keeping in practice. *Scandinavian Journal of Statistics*, 2022.
- [13] P. Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society, Series B*, 2024. to appear, with discussion; also arXiv:1906.07801.
- [14] Peter Grünwald. Safe probability. *Journal of Statistical Planning and Inference*, 2018.
- [15] Peter Grünwald. The E-posterior. *Phil. Trans. Roy. Soc. A*, 2023.
- [16] Peter Grünwald, Alexander Henzi, and Tyron Lardy. Anytime-valid tests of conditional independence under model-X. *Journal of the American Statistical Association*, 2023.
- [17] J. Hannig, H. Iyer, R.C.S. Lai, and T.C.M. Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361, 2016.
- [18] Yunda Hao, Peter Grünwald, Tyron Lardy, Long Long, and Reuben Adams. E-values for k-sample tests with exponential families. *Sankhya A*, 2024.
- [19] Alexander Henzi and Johanna F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 2021.
- [20] R. Hubbard. Alphabet soup: Blurring the distinctions between p 's and α 's in psychological research. *Theory and Psychology*, 14(3):295–327, 2004.
- [21] R. Hubbard and M.J. Bayarri. Confusion over measures of evidence (p 's) versus errors (α 's) in classical statistical testing. *The American Statistician*, 57:171—177, 2003.

- [22] Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. E-values as unnormalized weights in multiple testing. *arXiv:2204.12447*, 2022.
- [23] Valen E Johnson. Uniformly most powerful Bayesian tests. *Annals of Statistics*, 41(4), 2013.
- [24] E.L. Lehmann. *Testing Statistical Hypotheses*. Wiley, first edition, 1959.
- [25] E.L. Lehmann. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249, 1993.
- [26] Michael Lindon, Dae Woong Ham, Martin Tingley, and Iavor Bojinov. Anytime-valid linear models and regression adjusted causal inference in randomized experiments. *arXiv:2210.08589*, 2022.
- [27] Ryan Martin. Inferential models and the decision-theoretic implications of the validity property. *arXiv preprint arXiv:2112.13247*2009, 2021.
- [28] Ryan Martin and Chuanhai Liu. *Inferential models: reasoning with uncertainty*. CRC Press, 2015.
- [29] Deborah G Mayo. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press, 2018.
- [30] Deborah G Mayo and David R Cox. Frequentist statistics as a theory of inductive inference. *Lecture Notes-Monograph Series 2nd Lehmann Symposium*, pages 77–97, 2006.
- [31] Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified Gaussian processes. In *Algorithmic Learning Theory*. PMLR, 2021.
- [32] J. Neyman. *First Course in Probability and Statistics*. 1950.
- [33] J. Neyman. Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in Statistics: Theory and Methods*, 5(8):737–751, 1976.
- [34] J. Neyman and E. S Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A.*, 231:289—337, 1933.
- [35] Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter Grünwald. E-statistics, group invariance and anytime valid testing. *arXiv:2208.07610*, 2022.
- [36] Aleksandr Podkopaev, Patrick Bloebaum, Shiva Kasiviswanathan, and Aaditya Ramdas. Sequential kernelized independence testing. In *International Conference on Machine Learning*, 2023.
- [37] Aaditya Ramdas, Peter Grünwald, Volodya Vovk, and Glenn Shafer. game-theoretic statistics and safe anytime-valid inference. *statistical science*, 2023. To appear.
- [38] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020.
- [39] Richard Royall. *Statistical evidence: a likelihood paradigm*. Chapman and Hall, 1997.

- [40] Alexander Ly Samuel Pawel and Eric-Jan Wagenmakers. Evidential calibration of confidence intervals. *The American Statistician*, 78(1):47–57, 2024.
- [41] T. Schweder and N. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, 2016.
- [42] G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, 2021. With Discussion.
- [43] G. Shafer and V. Vovk. *Game-Theoretic Probability: Theory and Applications to Prediction, Science and Finance*. Wiley, 2019.
- [44] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.
- [45] J. ter Schure and P. Grünwald. ALL-IN meta-analysis: breathing life into living systematic reviews. *F1000Research*, 11(549), 2022.
- [46] J. ter Schure, M.F. Perez-Ortiz, A. Ly, and P. Grünwald. Safe logrank test: Error control under continuous monitoring with unlimited horizon. *arXiv:1906.07801*, 2021.
- [47] Judith Ter Schure, Alexander Ly, Lisa Belin, Christine S Benn, Marc JM Bonten, Jeffrey D Cirillo, Johanna AA Damen, Inês Fronteira, Kelly D Hendriks, Anna Paula Junqueira-Kipnis, André Kipnis, Odile Launay, Jose Euberto Mendez-Reyes, Mihai G Netea, Sebastian Nielsen, Caryn M Upton, Gerben van den Hoogen, Jesper M Weehuizen, Peter D Grünwald, and CH (Henri) van Werkhoven. Bacillus Calmette-Guérin vaccine to reduce COVID-19 infections and hospitalisations in healthcare workers. *medRxiv*, pages 2022–12, 2022.
- [48] Timmy Tse and Anthony C Davison. A note on universal inference. *Stat*, 11(1):e501, 2022.
- [49] Rosanne Turner and Peter Grünwald. Anytime-valid confidence intervals for contingency tables and beyond. *Statistics and Probability Letters*, 2023.
- [50] Rosanne Turner and Peter Grünwald. Safe sequential testing and effect estimation in stratified count data. In *Annual AI and Statistics Conference*, 2023.
- [51] Rosanne Turner, Alexander Ly, Muriel-Felipe Ortiz-Perez, Judith ter Schure, and Peter Grünwald. R-package `safestats`, 2022. CRAN.
- [52] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 2021.
- [53] Eric-Jan Wagenmakers, Quentin F Gronau, Fabian Dablander, and Alexander Etz. The support interval. *Erkenntnis*, pages 1–13, 2020.
- [54] Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10:299–326, 1939.
- [55] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 2022.

- [56] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.
- [57] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 2024. With discussion.
- [58] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. *arXiv:2203.12572*, 2022.
- [59] C. Yanofsky. It’s chancy, 2019. Blog Post February 5th, 2019.
- [60] Zhenyuan Zhang, Aaditya Ramdas, and Ruodu Wang. When do exact and powerful p-values and e-values exist? *arXiv:2305.16539*, 2023.

Supporting Information Appendix

A Supporting Information for Section 1 and Section 2.2.1

Unbounded expected loss based on (4) and an ‘improvement’ of (4) This issue is best illustrated by (but certainly not limited to) a discrete-valued p-value P that can take values $1, 1/2, 1/4, 1/8, \dots, 1/2^k$ for some $k > 0$ and that is piece-wise strict, i.e. it satisfies $P_0(P \leq \alpha) = \alpha$ for $\alpha \in \{1, 1/2, \dots, 1/2^k\}$. Consider a GNP decision task as in Section 2.1 with loss function satisfying $L_b(\underline{0}, a) = 2a$, for $a \in \mathcal{A}_b = \{0, 1\ell, 2\ell, 4\ell, \dots, 2^k\ell\}$. Based on (4), upon observing $P = 2^{-c}$, one would take action 2^c . The resulting expected loss, analogously to (5), is given by $\sum_{c=1}^k 2 \cdot 2^{-c} 2^c = 2k$ which goes to ∞ as we make k larger — showing that the expected loss can be unbounded if we base decisions on (4). Now, instead of using (4) it may seem more reasonable to pick the largest a such that

$$Q(y) \cdot L_b(\underline{0}, a) \leq \ell, \tag{29}$$

where $Q(y) = P(y)/2$: with this modification, for each $a \in \mathcal{A}_b$, we end up multiplying $L_b(\underline{0}, a)$ in (29) with exactly the probability that a will be selected (rather than, as in (4), with a larger probability). For example, $a = 2^c$ will be selected if $Q(y) = 2^{-c}$; this happens iff $P(y) = 2^{-c+1}$, which happens with exactly probability 2^{-c} , so with probability $Q(y)$. Yet still, using (29) leads to unbounded expected loss: in the above sample the expected loss is now k rather than $2k$, still growing linearly in k .

Standard conversions of p-values into e-values are sub-optimal [44, 43, 52] have studied functions that convert arbitrary p-values to e-variables. These *calibrators* are strictly decreasing functions f , such that $S(Y) := f(P(Y))$ is an e-variable whenever P is a p-value. Calibrators invariably have the property that as $P \downarrow 0$, $f(P)$ grows towards ∞ more slowly than $1/P(Y)$ (note that, for any e-variable S , $1/S$ is a (conservative) p-value but the converse does not hold). For example, [43], $f(P) = 1/\sqrt{P} - 1$ is a calibrator. Given that such calibrators exist, one might wonder if in this paper we are really merely advocating to change the *scale* at which evidence against the null is expressed: isn’t it sufficient to take a p-value to express evidence, convert it to an e-value, and then use the maximally compatible decision rule (7)? If so, this would undercut our arguments for using e vs. p. The answer is that, while calibrating p-values and then using (7) does give us Type-I risk safety, it is still not advisable because e-variables arising from calibrated p-values are typically far from optimal. Intuitively, a ‘good’ e-variable, relative to a given alternative, is one that tends to be large (provide much evidence) if the alternative is true. This can be formalized in terms of power or, as stated in the main text, GRO. For example, suppose we aim to test null $\theta \leq 0$ against alternative $\theta \geq \delta$. If we take the 1-sided e-variables S_θ^+ for the normal location family as defined underneath (16) with $\alpha^* = \alpha$ and $n^* = n$ specified correctly, then to get power 0.8 we need a factor of ≈ 1.75 more data points than if we use the standard UMP NP test (this follows from the derivation in [13, Appendix B.6]; as explained there, the factor can be significantly reduced by optional stopping). If, instead, we use the p-value corresponding to this UMP test directly and turn it into an e-value by the above calibrator, we need a factor of ≈ 3.0 more data [13, Section 7]. The reason for this discrepancy is that calibrators work for arbitrary p-values and are thus ‘blind’ to the underlying sampling model (in this case, normal location). In order to get high power it is invariably (much) better to use e-values designed for the underlying model

directly — so, importantly, the distinction between e and p is not just a matter of scale and designing ‘good’ e-values (with good GRO properties) is a nontrivial task — we cannot just take any given p-value with good power-properties and calibrate.

Why normalizing ℓ to 1 in (11) and (12) is not harmful This is further discussed, in a wider context, in the Supporting Information for Section 5 further below.

B Supporting Information for Section 2.2.2 until and including Section 3.3.1

In this section we state and prove Theorem 3, a general result which has Theorem 1 of Section 2 as a special case, extending it to the case of general GNP decision problems (e.g. confidence intervals) as defined in Definition 2, Section 3. However, in the first subsection below, we first state and prove a simpler form of the theorem which works for *simple* GNP decision problems, as defined in Section 2, simplified even further by requiring countable \mathcal{Y} . This allows us to strip away all issues about ‘almost surely, measurability’ etc. and focus on the key idea, which lies in the fundamental Lemma 1. The proof of the general Theorem 3 is based on essentially the same key insight but requires substantial additional notations and quantifications. We prepare these in Section B.B.2 below, where we also explain why the probabilities in (13), (14) (used to define Type-II strictly-better-than and admissibility) as well as the expectation (11) may be undefined in pathological cases, and we extend the definitions of Type-II-strictly better and admissibility to ensure that these are always well-defined. We then state and prove Lemma 2, the general version of Lemma 1 in Section B.B.3, and state and prove the general theorem in Section B.B.4. Before we do all this though, we explain, as promised underneath Theorem 1 in the main text, how we can make any GNP testing problem rich by adding a single additional loss function and why this makes the condition of richness a reasonable one. This is illustrated by Example 6 which indicates why a condition like richness is necessary and thereby gives a (very) high-level intuition to the proof.

Enforcing Richness relative to S : why richness is a weak condition, and why it is needed For any sharp e-variable S which we might want to base our decisions on, we can trivially *enlarge* any given GNP testing problem $(\mathcal{B}', \{(A_b, L_b(\underline{0}, \cdot) : A_b \rightarrow \mathbb{R}_0^+) : b \in \mathcal{B}'\})$ by setting $\mathcal{B} := \mathcal{B}' \cup \{\text{ID}(S)\}$, adding a loss function indexed by $\text{ID}(S)$ with action space $\mathcal{A}_{\text{ID}(S)}$ set equal to the co-domain of S and, for $s \in \mathcal{A}_{\text{ID}(S)}$, we set $L_{\text{ID}(S)}(\underline{0}, s) := s$. Then the extended GNP decision problem will automatically be rich relative S , and Part 2 of Theorem 1 can be applied. In reality, DM usually is not aware of the full details of the problem anyway, being only presented one particular loss function L_b , an element of a set $\{L_b : b \in \mathcal{B}'\}$ that is unknown: DM will only know the definition of the particular function L_b that she is presented with. Thus, assuming the set already contains this additional, special loss $L_{\text{ID}(S)}$ does not really impose any additional condition on the DM and only serves to make the analysis more robust, it thus seems a reasonable assumption. It gives an imagined adversary who chooses $b = B(Y)$ as function of Y more power, and as illustrated by the example below, without something like this added power, the theorem simply cannot hold. As such is analogous to (but not the same as) allowing an adversary to randomize between actions, as required for the minimax theorem in game theory. To take the analogy to the minimax theorem even

further, we note that, just like in that theorem, one side of the proof (Part 1) is in essence trivial whereas the other side (Part 2) requires a sophisticated argument.

Example 6 Consider a GNP testing problem and e-variable S defined as follows:

1. $\mathcal{Y} = \{0, 10, 20\}$; $\mathcal{H}(\underline{0}) = \{P_0\}$ with $P_0(Y = 10) = 1/20$ and $P_0(Y = 20) = 1/40$.
2. $\mathcal{B} = \{b_1\}$, $\mathcal{A}_{b_1} = \{0, 9, 19, 21\}$, and for all $a \in \mathcal{A}_{b_1}$, we set $L_{b_1}(\underline{0}, a) = a$.
3. $S(y) := y$.

We note that S is a sharp e-variable, but the GNP testing problem is *not* rich relative to S . And indeed, the conclusion of Part 2 of Theorem 1 is violated here:

$$\delta_{b_1}(0) = 0; \delta_{b_1}(10) = 9; \delta_{b_1}(20) = 19$$

is seen to be a decision rule that is maximally compatible with S , but it is not admissible: the decision rule

$$\delta'_{b_1}(0) = 0; \delta'_{b_1}(10) = 9; \delta'_{b_1}(20) = 21$$

is Type-II strictly better than δ yet still Type-I risk safe, since $\mathbf{E}_{P_0}[L_{b_1}(\underline{0}, \delta'_{b_1}(Y))] = 9/20 + 21/40 = 39/40 < 1$. So we have a decision rule that is maximally compatible relative to a sharp e-variable yet not admissible — this shows some additional condition such as richness is necessary. To get an initial idea why ‘richness’ does the trick, let’s enlarge \mathcal{B} as above to make the resulting GNP testing problem rich relative to S . That is, we add loss function indexed by $b_2 := \text{ID}(S)$, so that $\mathcal{A}_{b_2} = \{0, 10, 20\}$, and for all $a \in \mathcal{A}_{b_2}$, we have $L_{b_2}(\underline{0}, a) = a$. Decision rule δ above was maximally compatible in the original problem, and the only way to extend it to the enlarged problem while keeping it maximally compatible is to set $\delta_{b_2}(y) = y$ for all $y \in \mathcal{Y}$. But now, in this enlarged problem, δ has become admissible! Rather than proving this in full generality we will just show that the decision rule δ' above that witnessed δ ’s inadmissibility in the original problem will not witness it any more in the enlarged problem. To see why, note that to witness inadmissibility of δ , we must have that δ' is Type-II strictly better than δ and at the same time Type-I risk safe. The only way to extend δ' to the enlarged problem while keeping it strictly better than δ is to set it such that $\delta'_{b_2}(y) \geq \delta_{b_2}(y)$ for all $y \in \mathcal{Y}$. But then it is not Type-I risk safe any more, so this extended δ' does not show δ to be inadmissible! To see why the extended δ' is not Type-I risk safe any more, note that, by adding loss L_{b_2} , we gave the imagined adversary more power: upon observing $y = 10$, the adversary can now choose $B = b_2$, and upon observing $y = 20$, she can choose $B = b_1$. Then $\mathbf{E}_{P_0}[L_B(\underline{0}, \delta'_B(Y))] \geq (1/20) \cdot 10 + (1/40) \cdot 21 = 21/40 > 1$, so δ' is not Type-I risk safe. Lemma 1 and its generalization Lemma 2 further below formalize this idea and are the key to proving Part 2 of Theorem 1 in the main text and its generalization Theorem 3 below.

B.1 Theorem 1 for countable \mathcal{Y} and $\mathcal{H}(\underline{0})$ with full support

Throughout this subsection we assume that we deal with a GNP testing problem that is simple in the sense of Section 2, with countable \mathcal{Y} and $\mathcal{H}(\underline{0})$ with full support. ‘Full support’ simply means that for all $y \in \mathcal{Y}$, all $P \in \mathcal{H}(\underline{0})$, we have $P(Y = y) > 0$. Because the testing problem is simple, we can define maximum compatibility in terms of (15).

So, fix any simple GNP testing problem of this type. For any given random variables $U = u(Y), V = v(Y)$, we write $U \leq V$ as an abbreviation of: for all $y \in \mathcal{Y}$, $u(y) \leq v(y)$; similarly for $U < V$ and $U = V$.

Key Concept and Lemma: Equalizing Maximal Compatibility For any function $B : \mathcal{Y} \rightarrow \mathcal{B}$, we say that decision rule δ° is *equalizing-maximally compatible* relative to e-variable S when restricted to B if

$$\text{for all } y \in \mathcal{Y} : L_{B(y)}(\underline{0}, \delta_{B(y)}^\circ(y)) = S(y), \text{ i.e. } L_B(\underline{0}, \delta_B^\circ) = S. \quad (30)$$

The following lemma is the key insight needed to prove Theorem 2 below and hence the special case of Theorem 1 in the main text with a simple GNP testing problem and countable \mathcal{Y} . It will later be generalized to Lemma 2 which plays the same key role for the general result Theorem 3.

Lemma 1 *Fix any simple GNP testing problem as above. Suppose that S is a sharp e-variable and δ° is a Type-I risk safe decision rule such that there exists a function $B : \mathcal{Y} \rightarrow \mathcal{B}$ so that δ° is equalizing-maximally compatible relative to S when restricted to B , as above. Then δ° is fully compatible with S , i.e. for all $b \in \mathcal{B}$, we have: $L_b(\underline{0}, \delta_b^\circ) \leq S$.*

To understand the lemma, suppose we are given some e-variable S and some Type-I risk safe δ . Then δ need not be compatible with S (it must be compatible with *some* S' , but not necessarily *this* S). The lemma says that if δ is in some specific sense ‘partially’ compatible with S though, namely for a specific B , and S is sharp, then it must be fully compatible with S after all. Now, in the case where the GNP testing problem has been made rich relative to S by adding the special loss function $\text{ID}(S)$ above, we would typically apply this lemma with $B = \text{ID}(S)$, i.e. B is a constant, independent of Y ; but the lemma works even if B may vary with Y . The surprising thing here is that compatibility relative to B (which may even be the constant $\text{ID}(S)$) has repercussions for the behavior of $\delta_{b'}$ for *all* $b' \in \mathcal{B}$.

The lemma immediately leads to the following corollary:

Corollary 1 *Fix any simple GNP testing problem as above. If a decision rule δ^* is maximally compatible relative to a sharp e-variable S (i.e. (15) holds) and equalizing-maximally compatible relative to the same S when restricted to some function $B : \mathcal{Y} \rightarrow \mathcal{B}$ then any δ° that is equalizing-maximally compatible relative to S when restricted to B and Type-I risk safe must also be fully compatible with S and hence, since δ^* is maximally compatible, satisfy $\delta_b^\circ \leq \delta_b^*$ for all $b \in \mathcal{B}$.*

Proof: [of Lemma 2] By Proposition 1, there must be some e-variable S' such that δ° is compatible with S' , i.e.

$$\text{for all } b \in \mathcal{B} : L_b(\underline{0}, \delta_b^\circ) \leq S'. \quad (31)$$

By equalizing-maximal compatibility relative to S when restricted to B , transitivity, and weakening (31), we must therefore also have

$$S = L_B(\underline{0}, \delta_B^\circ) \leq S',$$

so that $S \leq S'$. Suppose by means of contradiction that, even stronger, there is $y \in \mathcal{Y}$ such that $S(y) < S'(y)$. We know that for some $P_0 \in \mathcal{H}(\underline{0})$, $\mathbf{E}_{P_0}[S] = 1$. But then $\mathbf{E}_{P_0}[S'] > 1$, so S' is not an e-variable and we have arrived at a contradiction. Since we already established $S \leq S'$ it follows that $S = S'$. But then using the inequality in (31) shows that for all $b \in \mathcal{B}$, we have $L_b(\underline{0}, \delta_b^\circ) \leq S$ and the lemma is proved. \square

Armed with this result, we can now state and prove a restricted version of Theorem 1.

Theorem 2 Consider any simple GNP testing problem as above, with countable \mathcal{Y} and all $P \in \mathcal{H}(\underline{0})$ having full support on \mathcal{Y} .

1. Suppose that decision rule δ is admissible. Then there exists an e-variable S such that δ is maximally compatible with S .
2. Suppose that S is a sharp e-variable S and δ^* is maximally compatible relative to S (such a δ^* exists because we assume the GNP testing is simple); and assume further that the GNP testing problem is rich relative to S . Then δ^* is admissible.

Proof: [of Theorem 2]

Part 1. Suppose that δ is admissible. Then δ is by definition Type-I risk safe. By Proposition 1 there must be an e-variable S such that δ is compatible with S . Since δ is admissible, every strictly better δ' is not Type-I risk safe, hence cannot be compatible with any e-variable; in particular, δ' is not compatible with S . Hence there exists no δ' that is strictly better than δ and compatible with S . It follows that δ is maximally compatible with S .

Part 2. Let δ^* be maximally compatible and let δ° be another Type-I risk safe decision rule. We will show that δ° cannot be Type-II strictly better than δ^* ; this implies the result.

By our S -relative richness assumption and the construction of δ^* , we know that there exists a function $B : \mathcal{Y} \rightarrow \mathcal{B}$ with:

$$L_B(\underline{0}, \delta_B^*) = S. \quad (32)$$

We may assume that δ° satisfies $\delta_B^* \leq \delta_B^\circ$, otherwise we already know that δ° is not Type-II strictly-better. Now suppose by means of contradiction that for some $y \in \mathcal{Y}$, we have $\delta_{B(y)}^*(y) < \delta_{B(y)}^\circ(y)$. We then have for the $P_0 \in \mathcal{H}(\underline{0})$ with $\mathbf{E}_{P_0}[S] = 1$ (which exists by sharpness) that

$$1 = \mathbf{E}_{P_0}[S] = \mathbf{E}_{P_0}[L_B(\underline{0}, \delta_B^*)] < \mathbf{E}_{P_0}[L_B(\underline{0}, \delta_B^\circ)],$$

contradicting our assumption that δ° is Type-I risk safe. We may thus assume $\delta_B^\circ = \delta_B^*$.

The corollary of Lemma 1 above now implies that $\delta_b^\circ \leq \delta_b^*$ for all $b \in \mathcal{B}$ (i.e. not just for B), hence δ° is not Type-II strictly better than δ^* ; the theorem is proved. \square

B.2 Preparing the General Proof of Theorem 1 and 3

Almost sure inequality Fix any GNP decision problem with parameter set Θ , as in the general Definition 2. In particular in some applications we may have $\Theta = \{\underline{0}\}$, then we really deal with a GNP testing problem and the notation \leq_θ that we will now define can in such cases be replaced by $\leq_{\underline{0}}$.

For all $\theta \in \Theta$, for functions $U, V : \mathcal{Y} \rightarrow \mathbb{R}_0^+$ we define

$$U(Y) \leq_\theta V(Y) \quad (33)$$

to mean that for all $P \in \mathcal{H}(\theta)$, for all $\epsilon > 0$ and every measurable set $\mathcal{E} \subset \mathcal{Y}$ such that for all $y \in \mathcal{E}$, $U(y) > V(y) + \epsilon$, we have $P(\mathcal{E}) = 0$. Similarly,

$$U(Y) <_\theta V(Y) \quad (34)$$

is defined to mean that $U(Y) \leq_\theta V(Y)$ and there exist $P \in \mathcal{H}(\theta)$, $\epsilon > 0$ and measurable set $\mathcal{E} \subset \mathcal{Y}$ such that for all $y \in \mathcal{E}$, $U(Y) \leq V(Y) - \epsilon$, and we have $P(\mathcal{E}) > 0$. Note that statements (33) and (34) are well-defined even if U or V are not measurable so that $U(Y)$ or $V(Y)$ are not random variables. Nevertheless, we shall abuse notation by abbreviating $U(Y)$ to U and $V(Y)$, just like we do for random variables. Note that, if the events inside the probabilities below are measurable after all, then we have

$$U \leq_\theta V \Leftrightarrow \forall P \in \mathcal{H}(\theta) : P(U \leq V) = 1. \quad (35)$$

We also write $U >_\theta V$ if it is not the case that $U \leq_\theta V$; we write $U \geq_\theta V$ if it is not the case that $U <_\theta V$; and $U =_\theta V$ if $U \leq_\theta V$ and $U \geq_\theta V$; if V and U are measurable then clearly the corresponding analogues to (35) hold as well, e.g.

$$U <_\theta V \Leftrightarrow \forall P \in \mathcal{H}(\theta) : P(U \leq V) = 1 \text{ and } \exists P \in \mathcal{H}(\theta) : P(U < V) > 0. \quad (36)$$

It is easily checked that $=_\theta$ establishes an equivalence relation on functions of Y , and relative to this relation, \leq_θ is a partial order and $<_\theta$ is the corresponding strict order, i.e $U <_\theta V$ iff $U \leq_\theta V$ and not $U =_\theta V$. We shall freely use standard properties of this partial order (such as transitivity) below.

Moreover, we introduce the additional notation, for each function $B : \mathcal{Y} \rightarrow \mathcal{B}$, where, in line with the above, we abbreviate $B(Y)$ to B and $\delta_{B(Y)}(Y)$ to δ_B :

$$\begin{aligned} \delta_B^\circ \leq_L \delta_B &\Leftrightarrow \forall \theta \in \Theta : L_B(\theta, \delta_B^\circ) \leq_\theta L_B(\theta, \delta_B), \\ \delta^\circ \leq_L \delta &\Leftrightarrow \forall \theta \in \Theta, b \in \mathcal{B} : L_b(\theta, \delta_b^\circ) \leq_\theta L_b(\theta, \delta_b), \end{aligned}$$

as such avoiding the cumbersome expression on the right whenever we can. Analogously,

$$\begin{aligned} \delta_B^\circ <_L \delta_B &\Leftrightarrow \delta_B^\circ \leq_L \delta_B \text{ and } \exists \theta \in \Theta : L_B(\theta, \delta_B^\circ) <_\theta L_B(\theta, \delta_B), \\ \delta^\circ <_L \delta &\Leftrightarrow \delta^\circ \leq_L \delta \text{ and } \exists \theta \in \Theta, b \in \mathcal{B} : L_b(\theta, \delta_b^\circ) <_\theta L_b(\theta, \delta_b), \end{aligned}$$

and correspondingly with \geq_L and $>_L$. Finally, $\delta_B^\circ =_L \delta_B$ is defined to be equivalent to ' $\delta_B^\circ \geq_L \delta_B$ and not $\delta_B^\circ >_L \delta_B$ '; similarly for ' $\delta^\circ =_L \delta$ '.

Generalized Admissibility, Maximal Compatibility We can now generalize the definitions of *Type-II strictly-better-than* and *admissibility* for general GNP decision problems in the main text: formally, we say decision rule δ is *Type-II strictly better* than δ° , simply if we have

$$\delta^\circ <_L \delta.$$

As before, a decision rule δ° is *admissible* if it is Type-I risk safe (according to the definition underneath (18) in the main text), and there is no other Type-I risk safe decision rule that is Type-II strictly better than δ .

We also extend the definition of maximally compatible decision rule in the same way: formally, a *maximally compatible decision rule* relative to a given GNP decision problem and e-variable S is any compatible decision rule δ° which further satisfies that there is no other decision rule δ that is also compatible with S and that is Type-II strictly better than δ° , with the extended definition of Type-II strictly better given above.

We see that in the case of a GNP testing problem, whenever the events $L_b(\underline{0}, \delta_b^\circ) > L_b(\underline{0}, \delta_b)$ are measurable for all $b \in \mathcal{B}$ (in particular whenever \mathcal{Y} is countable), the probabilities

in (13) and (14) are well-defined and then the definition of strictly-better-than coincides with the one given in the main text. But it continues to be valid in case we pick pathological \mathcal{B} , $\{L_b : b \in \mathcal{B}\}$ for which the events above are nonmeasurable — which cannot be ruled out since we made no restrictions on the functions L_b , δ° and δ . Similarly, by replacing the definition (11) in the main text by (12) we also make sure that Type-I risk safety is well-defined irrespective of whether $\sup_{b \in \mathcal{B}} L_b(\underline{0}, \delta_b(Y))$ is measurable or not (we could also have avoided such measurability issues using inner- and outer-measure [5, Section 1.3] but this does not simplify the treatment so we decided against it).

In the same way, for a general GNP decision problem, the definition of strictly-better-than given here generalizes the one given in the main text below (19) to the case where the events involved may be nonmeasurable; as a consequence, the definitions of admissibility for GNP testing and decision problems, and the definition of maximum compatibility relative to S for GNP testing problems given in the main text, are all generalized by the definitions given here based on the generalized notion of strictly-better-than, and are valid irrespective of the measurability of the functions and events involved.

Crucially for the proof of Lemma 2 below, the ordering relation $\leq_{\underline{0}}$ is strong enough to imply inequality in expectation:

Proposition 2 *Consider a GNP testing problem with null hypothesis $\mathcal{H}(\underline{0})$ such that all $P \in \mathcal{H}(\underline{0})$ are absolutely mutually continuous. Let $S = S(Y)$ and $S' = S'(Y)$ be nonnegative random variables such that for all $P \in \mathcal{H}(\underline{0})$, $\mathbf{E}_P[S]$ is finite. Suppose $S \leq_{\underline{0}} S'$. Then (a) for all $P \in \mathcal{H}(\underline{0})$ we have $\mathbf{E}_P[S] \leq \mathbf{E}_P[S']$. Further, suppose $S <_{\underline{0}} S'$. Then (b) for every $P \in \mathcal{H}(\underline{0})$ we have $\mathbf{E}_P[S] < \mathbf{E}_P[S']$.*

Proof: (a) According to definition (33), for all $P \in \mathcal{H}(\underline{0})$, $\epsilon > 0$ and every measurable set $\mathcal{E} \subset \mathcal{Y}$ with for all $y \in \mathcal{E}$, $S(y) > S'(y) + \epsilon$, we have

$$\mathbf{E}_P[S] = \mathbf{E}_P[\mathbf{1}_{Y \in \mathcal{E}} \cdot S + \mathbf{1}_{Y \notin \mathcal{E}} \cdot S] = \mathbf{E}_P[\mathbf{1}_{Y \notin \mathcal{E}} \cdot S] \leq \mathbf{E}_P[\mathbf{1}_{Y \notin \mathcal{E}} \cdot (S' + \epsilon)] = \mathbf{E}_P[S' + \epsilon]$$

and the result follows.

(b) Fix $P \in \mathcal{H}(\underline{0})$. According to definition (33), for each $\epsilon > 0$ and measurable set \mathcal{E} with for all $y \in \mathcal{E}$, $S(y) > S'(y) + \epsilon$, we have $P(\bar{\mathcal{E}}) = 1$ (with $\bar{\cdot}$ denoting complement), whereas there exist $\delta > 0$ and $Q \in \mathcal{H}(\underline{0})$ and measurable \mathcal{F} such that $S(y) < S'(y) - \delta$ on \mathcal{F} and $Q(\mathcal{F}) > 0$. By mutual absolute continuity, we have $P(\mathcal{F}) > 0$ as well, and therefore:

$$\begin{aligned} \mathbf{E}_P[S] &= \mathbf{E}_P[\mathbf{1}_{Y \in \bar{\mathcal{E}}} \cdot S] = \mathbf{E}_P[\mathbf{1}_{Y \in \bar{\mathcal{E}} \cap \mathcal{F}} \cdot S + \mathbf{1}_{Y \in \bar{\mathcal{E}} \cap \bar{\mathcal{F}}} \cdot S] \\ &\leq \mathbf{E}_P[\mathbf{1}_{Y \in \bar{\mathcal{E}} \cap \mathcal{F}} \cdot (S' - \delta) + \mathbf{1}_{Y \in \bar{\mathcal{E}} \cap \bar{\mathcal{F}}} \cdot (S' + \epsilon)] \leq \mathbf{E}_P[S'] - P(\mathcal{F})\delta + \epsilon. \end{aligned}$$

Since this holds for fixed $\delta > 0$ and for every $\epsilon > 0$, the result follows. \square

B.3 General Form of Equalizing Maximal Compatibility Lemma

Consider any GNP testing problem. Fix any function $B : \mathcal{Y} \rightarrow \mathcal{B}$. Generalizing (30), we say that decision rule δ° is *a.s. equalizing-maximally compatible* relative to e-variable S when restricted to B if

$$L_B(\underline{0}, \delta_B^\circ) =_{\underline{0}} S, \tag{37}$$

where here and below, ‘a.s.’ stands for ‘almost surely’. The following lemma, generalizing Lemma 1, is the key insight needed to prove Theorem 3 below and hence its simplification Theorem 1 in the main text.

Lemma 2 Fix any GNP testing problem. Suppose that S is a sharp e-variable and δ° is a Type-I risk safe decision rule such that there exists a function $B : \mathcal{Y} \rightarrow \mathcal{B}$ so that δ° is a.s. equalizing-maximally compatible relative to S when restricted to B , as in (37). Then δ° is a.s. fully compatible with S , i.e. for all $b \in \mathcal{B}$, $L_b(\underline{0}, \delta_b^\circ) \leq_0 S$.

Just like in the simplified case with countable \mathcal{Y} , this immediately leads to a relevant corollary:

Corollary 2 Fix any GNP testing problem. If a decision rule δ^* is maximally compatible relative to a sharp e-variable S and a.s. equalizing-maximally compatible when restricted to some function $B : \mathcal{Y} \rightarrow \mathcal{B}$ then any δ° that is a.s. equalizing-maximally compatible relative to S when restricted to B and Type-I risk safe must also be a.s. fully compatible with S , i.e. for all $b \in \mathcal{B}$, $L_b(\underline{0}, \delta_b^\circ) \leq_0 L_b(\underline{0}, \delta_b^*)$.

Proof: [of Lemma 2] By Proposition 1, there must be some e-variable S' such that δ° is compatible with S' , i.e.

$$\text{for all } b \in \mathcal{B}, y \in \mathcal{Y}: L_b(\underline{0}, \delta_b^\circ(y)) \leq S'(y). \quad (38)$$

By a.s. equalizing-maximal compatibility relative to S when restricted to B , transitivity, and weakening (38), we must therefore also have

$$S =_0 L_B(\underline{0}, \delta_B^\circ) \leq_0 S', \quad (39)$$

so that $S \leq_0 S'$. Suppose by means of contradiction that, even stronger, $S <_0 S'$. We know that for some $P_0 \in \mathcal{H}(\underline{0})$, $\mathbf{E}_{P_0}[S] = 1$. But then Proposition 2 gives that $\mathbf{E}_{P_0}[S'] > 1$, so S' is not an e-variable and we have arrived at a contradiction. Since we already established $S \leq_0 S'$ it follows that $S =_0 S'$. But then using the inequality in (38) gives for all $b \in \mathcal{B}$, $L_b(\underline{0}, \delta_b^\circ) \leq_0 S$, and the lemma is proved. \square

B.4 Extension of Theorem 1 to general GNP decision problems

First, we extend the definitions of richness and sharpness from GNP testing to decision problems in the obvious manner, by inserting ‘for all’ quantifiers: we say that a GNP decision problem is *rich* relative to e-collection $\{S_\theta : \theta \in \Theta\}$ if for all $\theta \in \Theta$, the corresponding θ -testing problem (as defined in the main text underneath Definition 2) is rich relative to S_θ . Relative to a given GNP decision problem, we say that e-collection $\{S_\theta : \theta \in \Theta\}$ is sharp if for all $\theta \in \Theta$, S_θ is sharp relative to the corresponding θ -testing problem.

Theorem 3 Consider any GNP decision problem.

1. Suppose that decision rule δ is admissible. Then there exists an e-collection $S = \{S_\theta : \theta \in \Theta\}$ such that δ is maximally compatible with S .
2. Suppose that δ^* is a maximally compatible decision rule relative to some e-collection $S = \{S_\theta : \theta \in \Theta\}$. If (a) all $P \in \mathcal{P}$ are mutually absolutely continuous and (b) S is sharp relative to the given GNP decision problem, and (c) the GNP decision problem is rich relative to S , then δ^* is admissible.

Proof: [of Theorem 3]

Part 1. Suppose that δ is admissible. Then δ is by definition Type-I risk safe. But then by the definition in the main text underneath (18) there must be an e-collection $S = \{S_\theta : \theta \in \Theta\}$ such that δ is compatible with S . Since δ is admissible, for every δ' with $\delta' >_L \delta$ (i.e. δ' is strictly better than δ) we have that δ' is not Type-I risk safe. But then δ' cannot be compatible with any e-collection, in particular it cannot be compatible with S . Hence, there exist no δ' that is strictly better than δ and also compatible with S ; hence δ is maximally compatible.

Part 2.

Let δ^* be maximally compatible relative to S and let δ° be another Type-I risk safe decision rule. We will show that δ° cannot be Type-II strictly better than δ^* ; this implies the result.

By our relative richness assumption and the construction of δ^* , we know that for all $\theta \in \Theta$, there exists a function $B : \mathcal{Y} \rightarrow \mathcal{B}$ with:

$$\text{for all } y \in \mathcal{Y}: L_{B(y)}(\theta, \delta_{B(y)}^*(y)) = S_\theta(y). \quad (40)$$

We may assume that δ° satisfies $\delta_B^\circ \geq_L \delta_B^*$, otherwise we already know that it's not Type-II strictly-better. So, in particular, $L_B(\theta, \delta_B^\circ) \geq_\theta L_B(\theta, \delta_B^*)$. Now suppose by means of contradiction that $L_B(\theta, \delta_B^\circ) >_\theta L_B(\theta, \delta_B^*)$ for some $\theta \in \Theta$. Since δ° is Type-I risk safe, it must by definition also be compatible with an e-collection $S' = \{S'_\theta : \theta \in \Theta\}$ so then we also have $S'_\theta >_\theta L_B(\theta, \delta_B^*)$. By Proposition 2, using the assumption of mutual absolute continuity, we then have for the $P \in \mathcal{H}(\theta)$ with $\mathbf{E}_P[S_\theta] = 1$ (which must exist by sharpness) that

$$1 = \mathbf{E}_P[S_\theta] = \mathbf{E}_P[L_B(\theta, \delta_B^*)] < \mathbf{E}_P[S'_\theta]$$

so S'_θ is not an e-variable and hence S' is not an e-collection, contradicting our assumptions (we note that all quantities inside the equation must be measurable, because S_θ and S'_θ are both e-variables, and hence measurable by definition). We may thus assume $L(\theta, \delta_B^\circ) =_\theta L(\theta, \delta_B^*)$.

The corollary of Lemma 2 above, applied with the corresponding θ -GNP testing problem, now implies that for all $b \in \mathcal{B}$ (hence not just for B !) we have $L_b(\theta, \delta_b^\circ) \leq_\theta L_b(\theta, \delta_b^*)$. Since we can make this argument for all $\theta \in \Theta$, it follows that for all $b \in \mathcal{B}$, $\delta_b^\circ \leq_L \delta_b^*$. Therefore δ° is not Type-II strictly better than δ^* ; the theorem is proved. \square

C Supporting Information for Section 3.3.2

Proof for Claim underneath (25) Fix some $\epsilon > 0$. For simplicity we fix $\theta^* = 0$ and $n = 1$ (so that $Y = X_1 = \hat{\theta}(X_1)$); extension of the following argument to general sampling distributions θ^* and $n > 1$ is straightforward (for $\theta^* \neq 0$, use $Y' = Y - \theta^*$; for $n > 1$, simply adjust the variance).

We will construct $B(y)$ such that if $y = \epsilon$, the CI corresponding to $B(y)$ will be a single point at ϵ ; if y gets larger, the CI widens but no matter how large y , it will never cover the true $\theta^* = 0$. To this end, fix any strictly positive, strictly decreasing function g_0 with $g_0(\epsilon) = \epsilon$. We will take $B(y)$ such that $\delta_{B(y)}(Y)$ has as its left-end $\theta_L = g_0(y) = y - h(y)$ where $h(y) := y - g_0(y)$. The CI being by definition symmetric around y , we must then have $\theta_R = y + h(y) = 2y - g_0(y)$. Since for any $0 \leq \alpha \leq 1$, the $(1 - \alpha)$ -CI coincides with

the $(1 - \alpha)$ credible interval taken symmetrically around the MLE $\hat{\theta} = y$, both its left- and right-tail must have posterior weight $\alpha/2$. Since the posterior has a Gaussian density with mean y and variance 1, we must thus have $\alpha/2 = \int_{-\infty}^{g_0(y)} f_y(u) du$, where we by denote f_μ the density of a normal with variance 1 and mean μ , so $\alpha = 2F_y(g_0(y)) = 2F_0(-y + g_0(y))$, where F_μ is the CDF of a normal with mean μ and variance 1. It follows that $B(y)$ must be equal to $1/\alpha = 1/(2F_0(-y + g_0(y)))$

If data are actually sampled from $\theta^* = 0$, then the expected loss we *actually* make can be calculated in steps as follows:

$$\begin{aligned} \mathbf{E}_{Y \sim P_{\theta^*}}[L_{B(Y)}(\theta^*, \delta_{B(Y)}(Y))] &= \mathbf{E}_{Y \sim P_0}[B(Y) \cdot \mathbf{1}_{0 \notin \delta_{(B)}(Y)}] \geq \mathbf{E}_{Y \sim P_0}[B(Y) \cdot \mathbf{1}_{Y \geq \epsilon}] \\ &= \int_{\epsilon}^{\infty} f_0(y) \cdot \frac{1}{2 \cdot F_0(g_0(y) - y)} dy \\ &\geq \frac{1}{2} \cdot \int_{\epsilon}^{\infty} \exp\left(-\frac{y^2}{2}\right) \cdot \exp\left(\frac{(y - g_0(y))^2}{2}\right) (y - g_0(y)) dy \\ &\geq \sqrt{\frac{\pi}{2}} \cdot \int_{\epsilon}^{\infty} \exp(-yg_0(y)) \cdot (y - g_0(y)) dy, \end{aligned}$$

where we used the standard result that, with P_0 denoting a standard normal distribution, $P_0(Y \geq c) \leq \exp(-c^2/2)/(c \cdot \sqrt{2\pi})$. Clearly the integral diverges for many choices of g_0 satisfying our requirements; for example, we can take $g_0(y) = \epsilon^2/y$ (which works for all $\epsilon > 0$) or (if we want to make the probability of large B smaller) we can set $g_0(y) = \epsilon \cdot (\log(y + \exp(\epsilon)) - \epsilon)/y$ if ϵ is set to 2; then $\exp(-yg_0(y)) = (y + \exp(2) - 2)^{-2}$. In the table in the main text we took the former choice to see how a typical sample of the B 's, and corresponding α 's and CI's might look like.

Proof that the average in (26) may converge to infinity with probability 1 Assume a sequence of independent studies $Y_{(1)}, Y_{(2)}, \dots$, all of which are of the form $Y_{(j)} = (X_{(j),1})$ and thus have sample size 1 (we can treat larger sample sizes, and simple sizes varying from study to study, by thinking of the $Y_{(j)}$ as z -scores summarizing studies of varying sample size). Instantiate $\epsilon = 0.01$ and set $B_{(j)} := 1/(2F_0(-Y_{(j)} + \epsilon^2/Y_{(j)}))$ as above if $Y_{(j)} \geq \epsilon$ and $B_{(j)} = 1$ otherwise. Suppose that the $Y_{(j)}$ are all independently sampled from the same $\theta^* = 0$. Here is a sample (generated i.i.d. by R) of 20 corresponding $B_{(j)}$ (recall that for each j , the corresponding produced interval $\delta_{B_{(j)}}(Y_{(j)})$ is equal to the standard $(1 - \alpha_{(j)})$ -confidence interval, with $\alpha_{(j)} = 1/B_{(j)}$):

$$\begin{aligned} &1.15, 1, 3.44, 1.09, 1.91, 4.17, 10.40, 1.11, 1, 1, 1 \\ &1.47, 1.31, 1, 1, 1, 2.28, 1.76, 1, 1, 1 \end{aligned} \tag{41}$$

While the sequence looks rather innocuous, using (25), with A in $\hat{\theta} \pm A$ chosen by (24), we see the limit in (26) will go a.s. to ∞ rather than to 1. The example was deliberately designed to give an extreme discrepancy — in more realistic examples, the difference will presumably not be infinite but without knowing the dependency between Y and B there is no way to assess it.

Proof for Example 4

We first treat e-variables corresponding to one-sided tests, for which we can give exact results. To this end, let $\theta^- < \theta < \theta^+$ be

$$\frac{1}{2}n^*(\theta - \theta^+)^2 = \frac{1}{2}n^*(\theta - \theta^-)^2 = \log \frac{1}{\alpha^*}. \quad (42)$$

(note that $\log(2/\alpha^*)$ in (42) in the main text has been replaced by $\log(1/\alpha^*)$ here).

The *uniformly most powerful Bayes factor* [23] for a 1-sided test at sample size n^* and level α^* of $\mathcal{H}(0) = \{P_\theta\}$ vs. $\mathcal{H}(1) = \{P_{\theta'} : \theta' > \theta\}$ (or, $\mathcal{H}(1) = \{P_{\theta'} : \theta' < \theta\}$ respectively) is given by $S_\theta^+ := \frac{p_{\theta^+}(y)}{p_\theta(y)}$ (respectively, $S_\theta^- := \frac{p_{\theta^-}(y)}{p_\theta(y)}$). Straightforward rewriting now gives:

$$S_\theta^+(y) = \frac{p_{\theta^+}(y)}{p_\theta(y)} = \frac{e^{-n \log \frac{p_{\hat{\theta}^+}(y)}{p_{\theta^+}(y)}}}{e^{-n \log \frac{p_{\hat{\theta}^-}(y)}{p_\theta(y)}}} = \frac{e^{-(n/2)(\hat{\theta} - \theta - U)^2}}{e^{-(n/2)(\hat{\theta} - \theta)^2}} = \frac{e^{n \cdot (\hat{\theta} - \theta)U}}{e^{nU^2/2}} = e^{-nU^2/2 + n(\hat{\theta} - \theta)U} \quad (43)$$

where

$$U = \sqrt{\frac{2(\log(1/\alpha^*))}{n^*}} = \sqrt{\frac{2 \cdot c \cdot \log(1/\alpha)}{n}} \text{ with } c = \frac{n^*}{n} \cdot \frac{\log(1/\alpha)}{\log(1/\alpha^*)}.$$

We see that S_θ^+ is strictly increasing in $\hat{\theta}$, so it is $\geq 1/\alpha$ iff $\hat{\theta} \geq \theta_R$, where θ_R is the solution to

$$e^{-nU^2/2 + n(\theta_R - \theta)U} = \frac{1}{\alpha}.$$

Straightforward calculation shows that this is the case iff $\theta_R - \theta$ is equal to

$$\sqrt{\frac{2}{n} \cdot \log(1/\alpha)} \cdot g(c) \text{ with } g(c) = \frac{c+1}{2\sqrt{c}} = \frac{1}{2} (c^{1/2} + c^{-1/2}). \quad (44)$$

An analogous calculation gives that S_θ^- is decreasing in $\hat{\theta}$ and $\geq 1/\alpha$ iff $\hat{\theta} \leq \theta_L$, with $\theta - \theta_L$ equal to (44).

For the two-sided e-variable $S_\theta = (1/2)S_\theta^+ + (1/2)S_\theta^-$, a sufficient condition for $S_\theta \geq 1/\alpha$ is then that

$$S_\theta^+ \geq \frac{2}{\alpha} \text{ or } S_\theta^- \geq \frac{2}{\alpha}.$$

Therefore, if we apply the above with $\alpha^{*'}$ set to $\alpha^*/2$ and α' set to $\alpha/2$, we get that $\theta^+, \theta^-, S_\theta^+, S_\theta^-, S_\theta$ and c are now defined as in the main text, and a sufficient condition for $S_\theta \geq 1/\alpha$ is that

$$|\hat{\theta} - \theta| \geq \sqrt{\frac{2}{n} \cdot \log \frac{2}{\alpha}} \cdot g(c), \quad (45)$$

which was our claim in the main text. Since for fixed α , for all but the smallest n , whenever $S_\theta^+ \geq \frac{2}{\alpha}$, we have that S_θ^- must be very close to 0, since it decreases exponentially in n (and the same with S_θ^+ and S_θ^- interchanged), we find that (45) is quite tight in practice.

D Supporting Information for Section 5

On the Type-I Risk Upper Bound ℓ Here we discuss why normalizing ℓ to 1 in (11) and (12) is not harmful, and what we mean when we say (as we did in the discussion Section 5) that ‘ ℓ can be chosen differently from problem to problem, but it needs to be chosen independently of the data observed in that problem’.

Suppose you are a statistician, performing hypotheses tests within a variety of domains. Let $Y_{(1)}, Y_{(2)}, \dots$ be the sequence of samples, taking values in potentially different sets $\mathcal{Y}_{(1)}, \mathcal{Y}_{(2)}, \dots$, and associated with different null hypotheses $\mathcal{H}_{(0)}(\underline{0}), \mathcal{H}_{(1)}(\underline{0}), \dots$ and associated GNP testing problems, that you are confronted with in your professional career. We will assume the $Y_{(j)}$ are all independent. Each time j that you perform a hypothesis test, policy makers provide you with an upper bound $\ell_{(j)}$ (e.g. set equal to $L(\underline{1}, 0)$, the cost of maintaining the status quo and doing nothing, see Section 5), that may depend on previous outcomes but, given $Y_{(1)}, \dots, Y_{(j-1)}$, must be independent of $Y_{(j)}$. You also are given the loss function $L_{B_{(j)}}$ with associated action space $\mathcal{A}_{B_{(j)}}$. You change this into loss function $L'_{B_{(j)}} := L_{B_{(j)}}/\ell_{(j)}$, and you advise an action by applying a maximally compatible decision rule $\delta_{(j)}$ relative to $L'_{B_{(j)}}$. By multiplying both sides in the definition of Type-I risk safety ((11) or (12)) with $\ell_{(j)}$ again, we see that Theorem 1 implies that at each time j , your Type-I risk is bounded by $\ell_{(j)}$, i.e. $\sup_{P \in \mathcal{H}_{(j)}(\underline{0})} \mathbf{E}_P[L_{B_{(j)}}(\underline{0}, \delta_{B_{(j)}}(Y_{(j)}))] \leq \ell_{(j)}$.

Now, suppose that an outside evaluating agency is interested in your performance whenever the imposed bound is close to some specific ℓ^* . Thus, after you have engaged in m hypothesis testing problems, they look at the subset $\mathcal{I}_{m,\delta} := \{j \in [m] : |\ell_{(j)} - \ell^*| \leq \delta\}$ for some small $\delta > 0$. Now, let us assume that the process determining the $\ell_{(j)}$ s is such that $\lim_{m \rightarrow \infty} |\mathcal{I}_{m,\delta}| = \infty$, almost surely, i.e. a risk bound close to ℓ^* will eventually be chosen infinitely often. Then, by the strong law of large numbers, we also have that

$$\limsup_{m \rightarrow \infty} \frac{1}{|\mathcal{I}_{m,\delta}|} \sum_{j \in \mathcal{I}_{m,\delta}} L_{B_{(j)}}(\underline{0}, \delta_{B_{(j)}}(Y_{(j)})) \leq \ell^* + \delta.$$

Thus, your e-value based statistical hypothesis tests have a Neymanian inductive behavior interpretation: as long as the bounds $\ell_{(j)}$ themselves do not depend on data $Y_{(j)}$, then in the long run, among all tests in which the imposed bound was within δ of ℓ^* , you will achieve average loss that is also within δ of ℓ^* . In particular the normalization to $\ell_{(j)} = 1$ in the definitions in Section 2 does not affect this guarantee.

D.1 Evidential Interpretation of E-Values

Most practitioners still interpret p-values in a Fisherian way, as a notion of evidence against the null. Although this interpretation has always been highly controversial, it is to some extent, and with caveats (such as ‘single isolated small p-value does not give substantial evidence’ [29] or ‘only work with special, *evidential* p-values [12]’), adopted by highly accomplished statisticians, including the late Sir David Cox [7, 30]. Even Neyman [33] has written ‘my own preferred substitute for ‘do not reject H ’ is ‘no evidence against H is found’. In light of the results of this paper, one may ask if, perhaps, *e-values are more suitable than p-values* as such a measure. Although a proper analysis of such a claim warrants (at the very least) a separate paper, we briefly make the case here. At first sight this question may seem orthogonal to the Neymanian ‘inductive behavior’ stance adopted in this paper— as has often

been noted [2, 21, 4, 20], Fisher’s and Neyman’s interpretations of testing seem incompatible. Nevertheless (echoing a point made by error statisticians [29] and likelihoodists [39] alike), for any notion of ‘evidence the data provide about a hypothesis $\mathcal{H}(\underline{0})$ ’ to be meaningful at all, there have to be circumstances, perhaps idealized, in which additional knowledge κ is available, and together with κ , the evidence can be operationalized into reliable decisions (for, if there were no such circumstances, obtaining ‘high evidence’ for or against a claim could never have any empirical meaning whatsoever...). For the likelihoodists’s notion of evidence [39, 11], i.e. a likelihood ratio between simple $\mathcal{H}(\underline{0})$ and $\mathcal{H}(\underline{1})$, this additional knowledge κ would be a trustworthy prior probability on $\{\mathcal{H}(\underline{0}), \mathcal{H}(\underline{1})\}$ — once this is supplied, a DM can use Bayes’ theorem to come up with a posterior which can then lead to optimal decisions against arbitrary loss functions. For the notion of evidence against $\mathcal{H}(\underline{0})$ as a p-value, this κ would comprise a guarantee that a specific, a priori fixed and known sampling plan, would have been followed (otherwise the p-value would be undefined), and an a priori specified α , and knowledge that the decision would be of the simple form ‘accept’/‘reject’. This κ , however, is additional knowledge of a *very* specific kind (essentially, what we called the BIND assumption). In other situations, it is not clear at all how to operationalize evidence-by-p-value into decisions. Now, if we accept e-values as evidence against the null, the set of circumstances under which we can operationalize the evidence is much wider, as shown in this paper. Having thus direct empirical content in a wider variety of situations, e would seem preferable over P (a).

Note that I am not saying that evidence should *invariably* be a ‘stepping-stone’ towards a decision¹; evidence seems a more general notion than that. I am only saying that if there are broad sets of circumstances in which it *is* a stepping stone, this may be a good rather than a bad thing.

Add to this: (b) if $\mathcal{H}(\underline{0})$ and $\mathcal{H}(\underline{1})$ are simple, the e-value *coincides* with the likelihood ratio, i.e. the main competing notion of evidence; (c) if $\mathcal{H}(\underline{0})$ is simple yet $\mathcal{H}(\underline{1})$ is not, a special type of e-value coincides with a recently proposed Bayesian notion of evidence (the *support interval* [40, 53]); (d) unlike Bayesian methods, e-values can be constructed even if no clear alternative can be formulated and if the setting is highly nonparametric; and (e) in contrast to p-values, e-values remain meaningful if some details of the sampling plan are unknown or unknowable and if information from several interdependent studies is combined [13, 37]. In fact, this may be the most important observation: if a scientific study is performed, and, *because* the scientific study seemed promising, a second study was performed, then we would lie to report the evidence against the null provided by both studies taken together. Yet, while for e-values this is no problem (we can multiply the e-values of the individual studies), it is next to impossible to calculate a valid p-value for the two studies taken together — this is the main point of [13]. The fact that they cannot be calculated in such a standard scenario would seem to make them unsuitable as a notion of evidence. If we tae (a)—(d) together though, the case for e-values as evidence seems strong.

A similar comment pertains to Mayo’s *error statistics* philosophy with its concept of *severe testing* [30, 29, 59]: currently, Mayo’s notion of severity is, at least in simple cases, indirectly based on p-values [29, page 144]. In light of the above, it might be preferable to use e-values instead.

¹Thanks to a referee for prompting this important clarification.