Synthese (2024) 203:12
https://doi.org/10.1007/s11229-023-04434-6

ORIGINAL RESEARCH

Entropic taming of the Look Elsewhere Effect

Miklós Rédei[1] · **Márton Gömöri**[2,3]

Received: 28 July 2022 / Accepted: 16 November 2023 / Published online: 21 December 2023

Abstract
To mitigate the Look Elsewhere Effect in multiple hypothesis testing using p-values, the paper suggests an "entropic correction" of the significance level at which the null hypothesis is rejected. The proposed correction uses the entropic uncertainty associated with the probability measure that expresses the prior-to-test probabilities expressing how likely the confirming evidence may occur at values of the parameter. When the prior-to-test probability is uniform (embodying maximal uncertainty) the entropic correction coincides with the Bonferroni correction. When the prior-to-test probability embodies maximal certainty (is concentrated on a single value of the parameter at which the evidence is obtained), the entropic correction overrides the Look Elsewhere Effect completely by not requiring any correction of significance. The intermediate situation is illustrated by a simple hypothetical example. Interpreting the prior-to-test probability subjectively allows a Bayesian spirit enter the frequentist multiple hypothesis testing in a disciplined manner. If the prior-to-test probability is determined objectively, the entropic correction makes possible to take into account in a technically explicit way the background theoretical knowledge relevant for the test.

Keywords Hypothesis testing · Look Elsewhere Effect · Entropy

✉ Miklós Rédei
m.redei@lse.ac.uk

Márton Gömöri
gomorim@gmail.com

[1] Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK

[2] Department of Logic, Institute of Philosophy, Loránd Eötvös University, Múzeum krt. 4/i, Budapest 1088, Hungary

[3] Institute of Philosophy, Research Center for the Humanities, Tóth Kálmán u. 4, Budapest 1097, Hungary

1 The main suggestion of the paper

The Look Elsewhere Effect is an important phenomenon in multiple (parameter dependent) statistical hypothesis testing based on p-values. Given that this form of hypothesis testing is widespread in empirical sciences, the problem of how to mitigate it is a standard topic in statistics, where it is treated sometimes under the heading "Family-wise Error Rate" (see e.g. Bayer & Seljak 2020; Benjamini & Hochberg, 1995; Foster et al., 2006; Hochberg, 1988; Lehmann & Romano, 2005a, 2005b; List et al., 2019).

The problem is especially acute in experimental high-energy particle physics (Dawid, 2015, 2017; Gross & Vitells, 2010; Lyons, 2008, 2013). The Look Elsewhere Effect is one of the reasons why in high-energy particle physics the significance level of rejecting a hypothesis is required to be very high in comparison with typical significance levels adopted in other sciences. One can argue however—as Dawid does (Dawid, 2015, 2017), based on the analysis of the experimental confirmation of the Higgs particle—that the usual reasoning referring to the Look Elsewhere Effect to motivate the demand of high significance in particle physics can be too much decoupled from the theoretical background of the experiment. Dawid then suggests both a simple (Dawid, 2015) and a more elaborate (Dawid, 2017) Bayesian way of taking into account the physicists' prior-to-test theoretical belief in the truth of the theory under test to mitigate the Look Elsewhere Effect.

Motivated by Dawid's analysis, in this paper we suggest another Bayesian way of mitigating the Look Elsewhere Effect. Our main idea is to adjust the significance level by taking into account a possible prior-to-test theoretical knowledge that indicates the probability with which the (dis)confirming phenomenon might be found at different parameter values in the test. We demand the adjustment to satisfy three requirements:

(i) The correction should be sensitive to the uncertainty embodied by the probability measure ρ that represents the scientists' prior theoretical knowledge (belief) about where (at which parameter value t) the confirming evidence might be found in a test: the higher the uncertainty of ρ, the higher significance is required. The uncertainty embodied by the probability measure ρ should be justifiable on independent grounds as a quantity that expresses the uncertainty of ρ.

(ii) The correction should also be sensitive to the prior probability $\rho(t)$ of the observed event x_t to be found in the test at a specific parameter value t: the higher $\rho(t)$ the smaller significance is required.

(iii) The adjustment should yield the "Bonferroni correction" (Bayer & Seljak, 2020; Foster et al., 2006; Lehmann & Romano, 2005a) in the case when ρ is the uniform probability measure on the parameter space (which we assume to be finite, having $N > 1$ elements). Moreover, the uncertainty and the correction should be such that the correction depends on ρ in a continuous manner; in particular, as ρ approaches (pointwise) the uniform probability, the correction should approach the Bonferroni correction.

Taking the entropy of ρ as a measure of uncertainty of ρ, we define a quantitative entropic correction of significance that depends on the entropy of ρ, on the value $\rho(t)$ and on the number N of parameters, and which satisfies the conditions (i)–(iii). Thus

the proposed entropic correction mitigates the Look Elsewhere Effect in a way that reflects specific structural features of the prior knowledge represented by ρ, yielding the Bonferroni correction as a special case when the entropic uncertainty is maximal: when ρ is the uniform probability on the parameters. The correction also is intuitively correct at the other end of the uncertainty scale: when ρ is totally concentrated at a particular value of the parameter and hence represents maximal certainty (its entropy is zero). In this case the suggested entropic correction is zero: no need to adjust the p-value as a result of the Look Elsewhere Effect. In this extreme situation the Bayesian certainty overrides the Look Elsewhere Effect completely. In the intermediate cases the entropic correction depends very sensitively on the global features of ρ and the value $\rho(t)$. We illustrate the intermediate case on a simple hypothetical example.

To make the paper reasonably self-contained, Sect. 2 recalls the main idea of statistical theory testing, including the notion of p-values. Section 3 describes the Look Elsewhere Effect and the standard way of correction of significance intended to mitigate the effect. Section 4 introduces the entropic correction of significance, and Sect. 6 contains some concluding comments.

2 Statistical hypothesis testing and *p*-values

In science one frequently tests a new theory T' using statistical methods. The typical situation is that in these tests T' is evaluated not in isolation but in comparison with the old theory T (called "background theory" or "null hypothesis"), and the comparison is in terms of so called "p-values". This testing procedure is a standard topic in statistics: it is described e.g. in Chap. 26 in Freedman et al. (2007); the chapter "Hypothesis Tests and Corroboration" in Hartmann and Sprenger (2019), the Sect. 18.4 of Sprenger (2016), and Romeijn (2017) provide a compact summary with a philosophical flavor; and the review in Sect. 6 of Lyons (2013) is from the perspective of particle physics. To simplify the discussion, we call this way of comparative testing of T' *Contrastive Statistical Confirmation of* T' *using* p*-values*, "T-CS-confirmation of T'", for short. The confirmation procedure is summarized below, following the description in Sects. 6.2 and 6.3 in Lyons (2013), but using a different and more explicit notation.

A probability measure space $(X, \mathcal{S}, p)$ together with a random variable $f: X \to \mathbb{R}$ is called a *statistical test theory* of theory T with respect to f, and f is called the *statistical data function* (also called "test statistics"), if, assuming T to be true, one infers from T that the probability measure p on $\mathcal{S}$ specifies what one expects to find about the values of f in empirically executable trials with results in X: Assuming T is true, one expects to find $f(x)$ to lie more frequently in sets d of real numbers for which $p(f^{-1}(d))$ is a "large" probability (f^{-1} is the inverse image function of f), and one regards T being in tension with the observation if the observed values of f fall in a d such that $p(f^{-1}(d))$ is "small". What is a "small" and what is a "large" probability depends on the particular testing situation, and in each testing situation the definition of "small" and "large" gets specified by fixing what are called the "significance level" of the test: If $p(f^{-1}(d))$ is smaller than a conventionally chosen significance level, then it is "small", otherwise it is "large". The question of how one can decide whether $(X, \mathcal{S}, p)$ with a particular test statistics f is a good statistical test theory for T does

not have a general answer. It is the task of the specific sciences in which T is a theory to come up with a $(X, \mathcal{S}, p)$ and a suitable f that can serve as a good statistical test theory for T. The implementation of this idea of comparing T and T' in terms of statistical test theories starts with the definition of the p-value:

Definition 1 $(p\text{-}value)$ Let $(X, \mathcal{S}, p, f)$ be a statistical test theory for T. For an $x \in X$ consider the event E_x^+ that the value of f is equal or larger than $f(x)$:

$$E_x^+ \doteq \{y \in X : f(y) \geq f(x)\} \tag{1}$$

Then the function $p_f^+ : X \to [0, 1]$ defined by

$$p_f^+(x) \doteq p(E_x^+) = p(\{y \in X : f(y) \geq f(x)\}) \tag{2}$$

is called the *positive side p_f-value function*. The event E_x^- defined by

$$E_x^- \doteq \{y \in X : f(y) \leq f(x)\} \tag{3}$$

leads to the analogously defined function $p_f^- : X \to [0, 1]$ defined by

$$p_f^-(x) \doteq p(E_x^-) = p(\{y \in X : f(y) \leq f(x)\}) \tag{4}$$

and

$$p_f^{+,-}(x) \doteq 2 \cdot \min\{p_f^-(x), p_f^+(x)\} \tag{5}$$

is the *two-sided p_f-function*.

Different theories and test situations use different statistical data functions. In a particular application the statistical data function is chosen in such a way that the p-value can be interpreted as the measure of *incompatibility* of theory T with the empirical test result: The smaller the p-value $p_f^+(x)$, the *less* compatible theory T is with the empirical test result x; the larger the p-value $p_f^+(x)$, the *more* compatible theory T is with the empirical test result x. This interpretation of $p_f^+(x)$ depends very sensitively on f, and, depending on the specific situation, one might have to use $p_f^-(x)$ to interpret it the same way: the smaller $p_f^-(x)$, the less compatible T is with the empirical test result x. Sometimes a two-sided p_f-function has to be used as p-value in order for a similar interpretation of the p-value to be justified.

The typical procedure of testing T' in contrast to T using p-values has the assumptions and structure described in the following definition of T-CS confirmation of T':

Definition 2 $(T\text{-}CS\ confirmation\ of\ T')$

1. One specifies

   (i) A statistical test theory $(X, \mathcal{S}, p, f)$ for T.

(ii) A statistical test theory $(X, \mathcal{S}, p', f)$ for T'.

2. One assumes that the following hold:

(2.i) The positive side p_f-value function p_f^+ has the interpretation of indicating the (in)compatibility of the evidence x with theory T.

(2.ii) The negative side p_f-value function p'^-_f has the interpretation of indicating the (in)compatibility of the evidence x with theory T'.

(2.iii) Given a real number r, the probability that T' predicts for the event that the value of the test function is smaller than r is smaller than the probability predicted by theory T for this event:

$$p'(f^{-1}((-\infty, r])) < p(f^{-1}((-\infty, r])) \qquad \text{for all } r \in \mathbb{R} \tag{6}$$

3. Under these assumptions we say that T' *is* T*-CS-confirmed by evidence* $x \in X$ if:

(a) For some "small" real number $\alpha > 0$ we have

$$p_f^+(x) \leq \alpha \tag{7}$$

(b) For some "large" real number $\beta > 0$ we have

$$p'^-_f(x) \geq \beta \tag{8}$$

In view of assumption (2.iii) the conditions (7)–(8) do indeed positively distinguish T' from T because

- 3. (a) says: It is very unlikely to obtain the value $f(x)$ or more extreme (larger) value of the test if T is true; so we exclude T on this ground.
- 3. (b) says: It is *not* very unlikely to obtain the value $f(x)$ or more extreme (smaller) value of the test if T' is true, so we do *not* exclude T' on this ground.

The numbers α, β are called *significance* numbers, their choice is a matter of convention. The smaller α and the larger β, the more strongly the evidence x T-CS-confirms T': A small α (high significance) makes it more difficult to exclude the null hypothesis; a large β [called "the exclusion level" for the new theory T' (Lyons, 2013)] makes it easier to exclude the new theory T'. In many applications it is customary to take $\alpha = 0.02$. In experimental particle physics typical values of α are taken to be the numbers which correspond in case of a probability measure given by a Gaussian density to the probability that the value of the identity function as random variable is outside of the symmetric interval of length $k\sigma$ around its expectation value, where σ is the variance of the Gaussian density. The value $k = 5$ results in $\alpha = 3 \times 10^{-7}$. This particular "$5\sigma$" value for α is regarded a very demanding p-value, and is the convention accepted in particle physics. One of the reasons for choosing α very small is the Look Elsewhere Effect to be discussed below. The exclusion level β is typically taken to be much larger than α Lyons (2013); hence it is easier for the new theory T' to be incompatible with evidence than it is for T. This is reasonable because T is typically well-corroborated, hence discarding it needs strong evidence (high degree

of incompatibility with evidence); whereas "... incorrect exclusion of New Physics is not regarded as so dramatic" Lyons (2013).

The relation of p-values $p^{+}_{f}(x)$ and $p'^{-}_{f}(x)$ to α and β can be different from the one expressed by Eqs. (7)–(8). How to interpret the situation when the p-value of the null hypothesis T is not smaller than α, is a somewhat controversial matter (see Hartmann & Sprenger, 2019, pp. 229–230 and the references there). The Table 2 in Lyons (2013, p. 21) summarizes all the logically possible cases. In our notation the possibilities listed in Table 2 in Lyons (2013, p. 21) are reproduced below.

		Decision	If T true	If T' true
$p^{+}_{f}(x) \leq \alpha$	$p'^{-}_{f}(x) > \beta$	T' is T-CS confirmed	Type I error	Good choice
$p^{+}_{f}(x) > \alpha$	$p'^{-}_{f}(x) \leq \beta$	T' is T-CS DISconfirmed	Correct choice	Type II error
$p^{+}_{f}(x) > \alpha$	$p'^{-}_{f}(x) > \beta$	No decision	Weak test	Weak test
$p^{+}_{f}(x) \leq \alpha$	$p'^{-}_{f}(x) \leq \beta$	Ambiguous	Ambiguous	Ambiguous

2.1 Illustrative example: coin flipping

To illustrate the concept of T-CS-confirmation of T' we describe a simple instance of the paradigm example (Maxwell, 1994) of coin flipping. Let T' be the theory that a coin is *not* fair but biased towards its Head in 4:1 proportion, and let T be the theory that the coin is fair. The task is to T-CS-confirm T'. Let the probability space component of the statistical test theory for T be $(X_M, \mathcal{S}_M, p_M)$, where the set X_M is the set of all possible outcomes of flipping the coin M times and p_M is the uniform probability measure:

$$X_M \doteq \{(x_1, \ldots x_M) : x_i = H, T; i = 1, \ldots M\} \tag{9}$$

$$p_M(\{(x_1, \ldots x_M)\}) \doteq \frac{1}{2^M} \tag{10}$$

Let the statistical data function f be the number of heads thrown in M flips:

$$f((x_1, \ldots x_M)) \doteq \#\{x_i : i = 1, \ldots M; x_i = H\} \tag{11}$$

Here #{} denotes the number of elements in the set { }.

Let the probability space component of the statistical test theory for T' be $(X_M, \mathcal{S}_M, p'_M)$, where p'_M is the probability measure describing M throws with a coin that is biased towards Head in 4:1 proportion:

$$p'_M(\{(x_1, \ldots x_M)\}) \doteq \left(\frac{4}{5}\right)^k \left(\frac{1}{5}\right)^{M-k} \tag{12}$$

where k is the number of H's among $(x_1, \ldots x_M)$.

Suppose the outcome of 10 throws is

$$x = (H, H, T, H, H, H, H, H, H, H) \tag{13}$$

This outcome contains 9 Heads. Elementary calculation shows

$$p_{10f}^{+}(x) \approx 0.01074 \tag{14}$$

$$p'_{10f}^{-}(x) \approx 0.89263 \tag{15}$$

Thus choosing $\alpha = 0.03$ and $\beta = 0.7$ one can conclude that the evidence of throwing 9 Heads in 10 throws T-CS-confirms the hypothesis T' that the coin is biased in favor of Heads in proportion to 4:1—in comparison with the null hypothesis T that the coin is fair.

3 Parameter-dependent contrastive statistical confirmation and the Look Elsewhere Effect

In some situations in which T' is to be T-CS confirmed, theory T' has a free parameter t with values in a range R, and the theory's claim to be tested has the form "*For a specific value of parameter t, property* Φ *holds*". A T-CS-type confirmation of T' in this case consists in establishing that T' is T-CS-confirmed at some value t_0 of the parameter. To this end, one makes all the concepts featuring in T-CS testing parameter-dependent, and one stipulates certain connections between the parametrized notions in order to compare the theories T and T'.

To be specific, one starts with inferring from T a probability space $(X_t, \mathcal{S}_t, p_t)$ for each t in R with statistical data function f_t, and proceeds with inferring from T' = "*For a specific value of parameter t, property* Φ *holds*" a statistical test theory $(X_t, \mathcal{S}_t, p'_t)$ (with statistical data function f_t) for each parameter value t. To simplify the situation we assume that (i) the random events at every value of the parameter are of the "same sort"; (iia) the background theory T predicts the same probability for the same type of event at different parameter values and (iib) that the same holds for T'; (iii) that the comparison is in terms of the "same sort" of test function at every parameter value. The technical formulations of these assumptions are summarized in (I)–(III) below. (For the notions of isomorphism of measurable spaces and of probability measure spaces featuring in (I) and (II) see Bogachev (2007, p. 275), for a compact summary of these notions see Rédei and Gyenis (2021).)

Assumptions

For all parameter values t we have:

(I) The measurable spaces $(X_t, \mathcal{S}_t)$ are isomorphic to $(X, \mathcal{S})$ via the isomorphisms $h_t : X_t \to X$.

(IIa) The $(X_t, \mathcal{S}_t, p_t)$ is measure theoretically (strictly) isomorphic to $(X, \mathcal{S}, p)$ via the isomorphism $h_t : X_t \to X$.

(IIb) The $(X_t, \mathcal{S}_t, p'_t)$ is measure theoretically (strictly) isomorphic to $(X, \mathcal{S}, p')$ via the isomorphism $h_t : X_t \to X$.

(III) The statistical data functions f_t do not depend on t: There is a statistical data function $f: X \to \mathbb{R}$ such that $f_t = h_t \circ f$; i.e. $f_t(x_t) = f(h_t(x_t))$ for all $x_t \in X_t$.

One then finds empirical evidence $x_t \in X_t$ for each parameter value t in the range R. One can define t-dependent p-values $p^+_{f_t}(x_t)$ and $p'^-_{f_t}(x_t)$ in complete analogy with Eqs. (2)–(4). The numbers $p^+_{f_t}(x_t)$ and $p'^-_{f_t}(x_t)$ are called "local p-values". In view of the assumptions (I)–(III) they can be written as

$$p^+_{f_t}(x_t) = p^+_f(h_t(x_t)) \qquad \text{and} \qquad p'^-_{f_t}(x_t) = p'^-_f(h_t(x_t)) \tag{16}$$

Assume that one has obtained as empirical evidence a set $\mathcal{E} = \{x_t \in X_t : t \in R\}$ of elements, and one finds an $x_{t_0} \in \mathcal{E}$ for which the analogues of (7)–(8) hold for the local p-values:

$$p^+_f(h_{t_0}(x_{t_0})) \leq \alpha \tag{17}$$

$$p'^-_f(h'_{t_0}(x_{t_0})) \geq \beta \tag{18}$$

and no other element x_t in the evidence set $\mathcal{E}$ satisfies (17)–(18). Can one now regard T' = "*For a specific value of parameter* t, *property* Φ *holds*" as T-CS-confirmed by evidence x_{t_0}, declaring t_0 as the parameter value at which Φ holds? The problem with doing so without further consideration is that, if the parameter space R is large and we collect evidence at every value of the parameter, then it becomes more and more likely that we find evidence x_t at *some* parameter value t that has the same or smaller (local) p-value as x_{t_0} has—even if T is true and T' is not. Thus the danger of making a Type I error increases because the local p-value $p^+_f(h_{t_0}(x_{t_0}))$ at the particular parameter value t_0 is an overestimation of the degree of incompatibility of T with evidence x_{t_0}. This "dilution of significance" in case of a parameter-dependent testing is called the "Look Elsewhere Effect" (Bayer & Seljak, 2020; Dawid, 2015; Gross & Vitells, 2010; Lyons, 2008, 2013).

The customary way to mitigate the "dilution of significance" of local p-values is to introduce a "global" p-value $p_G(x_{t_0})$ and, for rejecting the background theory in view of evidence x_{t_0}, demand that this global p-value be below a chosen threshold significance α. This demand then requires the local p-value to be below a significance δ that is much smaller than α. By definition, the global p-value $p_G(x_{t_0})$ of x_{t_0} is the probability that in testing at each parameter value we find at least one event x_t at some parameter value t such that the local p-value $p^+_f(h_t(x_t))$ of x_t is less than or equal to the local p-value $p^+_f(h_{t_o}(x_{t_0}))$ of x_{t_0}. The ratio

$$\frac{p_G(x_{t_0})}{p^+_f(h_{t_o}(x_{t_0}))} \tag{19}$$

is called the *trial factor* (Bayer & Seljak, 2020; Gross & Vitells, 2010).

Calculating the trial factor is typically difficult in applications. This is because the global p-value of an observed event x_{t_0} is *not* determined by the local probability

spaces $(X_t, \mathcal{S}_t, p_t)$, $(X_t, \mathcal{S}_t, p'_t)$ alone, not even when one makes the assumptions (I)–(III). To obtain the global p-value one has to aggregate the local probability spaces $(X_t, \mathcal{S}_t, p_t)$, $(X_t, \mathcal{S}_t, p'_t)$ into a single probability space; in particular one has to assume some additional hypothesis about what probabilistic dependencies exist among probabilities of events at different values of the parameter. It is not always obvious what the dependencies are—this depends on the specific features of the whole testing situation. If R is a finite set containing N number of parameter values (which we have assumed), and one takes as the aggregated probability space the product of the local probability spaces, with the product probability measure as the aggregated probability measure, which expresses that the events at different parameter values are independent, then one can calculate $p_G(x_{t_0})$ and obtain (see the Appendix for details):

$$p_G(x_{t_0}) = 1 - [1 - p_f^+(h_{t_0}(x_{t_0}))]^N \tag{20}$$

For large enough N, the global p-value in (20) can be large (close to 1) even if $p_f^+(h_{t_0}(x_{t_0}))$ is small. So if there are a large number of parameter values at which evidence is collected, it is very likely that one finds evidence at some parameter value with significance equal or higher than the significance displayed by the local p-value of x_{t_0}, even if T is true. For instance, taking $p_f^+(h_{t_0}(x_{t_0})) = 3 \times 10^{-7}$ (the 5σ value), and 100 parameter values at which evidence is collected, we have

$$1 - [1 - p_f^+(h_{t_0}(x_{t_0}))]^{100} \approx 3 \times 10^{-5} \tag{21}$$

So the local p-value gets diluted by two orders of magnitude, resulting in the trial factor to be about 100. Thus, in the case of dilution (20), in order for a given local significance δ *not* to dilute above the global significance α, the local significance δ must satisfy

$$\delta \leq 1 - [1 - \alpha]^{\frac{1}{N}} \tag{22}$$

Since in general the calculation of the global p-value is difficult and an analytic expression corresponding to (22) is not available, procedures have been developed with the aim of controlling the Look Elsewhere Effect in a simple manner [Chap. 9 in Lehmann and Romano (2005a) gives a review of these procedures, in particular the procedure suggested by Benjamini and Hochberg (1995); see also Efron (2008)]. The simplest, one-step procedure is the *Bonferroni correction* (Bayer & Seljak, 2020; Foster et al., 2006; Lehmann & Romano, 2005a). In this correction one sets the local significance level δ as

$$\delta \doteq \frac{\alpha}{N} \tag{23}$$

For the T-CS-confirmation of "Φ *holds at value* t_0" by the observation x_{t_0}, with significance α, one then requires that the local p-value $p_f^+(h_{t_0}(x_{t_0}))$ satisfies

$$p_f^+(h_{t_0}(x_{t_0})) \leq \delta = \frac{\alpha}{N} \tag{24}$$

If N is large, then for the *global* significance to be less than α, the *local* p-value $p_f^+(h_{t_0}(x_{t_0}))$ must be very small. This is one of the reasons why *local* p-values are required in particle physics to be less than the very demanding 3×10^{-7} (the 5σ value).

The Bonferroni correction is very demanding: one can show Lehmann and Romano (2005a, p. 350) that it is the strongest correction in the sense that any dilution of δ is bounded by α if $\delta \leq \frac{\alpha}{N}$. Thus, controlling the dilution of significance α by the Bonferroni correction in order to control Type I error one may over-compensate, making Type II error more likely. This can happen especially for large N. (In Sect. 4.1 we will see an example showing that the Bonferroni correction can make it *impossible* to discard the null theory T in a given test situation.) Motivated by this, other, multiple-step correction procedures have been designed (see Chap. 9 in Lehmann & Romano, 2005a). A common presupposition shared by the one-step Bonferroni and the other, multi-step corrections is that we have no prior-to-test information whatsoever on the value of the parameter at which the evidence confirming the new theory might happen. But this presupposition is not always justified. There could be situations in which one does have expectations about how likely it is that the confirming evidence occurs at different parameter values. In such cases the adjustment of the local p-value should take into account this information. Neither the Bonferroni, nor the multi-step corrections have the tools to incorporate such information. In the next section we make a suggestion for a modification of the Bonferroni correction that is sensitive for such extra information.

4 Entropic taming of the Look Elsewhere Effect

Instead of requiring the local p-value $p_f^+(h_{t_0}(x_{t_0}))$ to satisfy the Bonferroni corrected significance condition (24) it is reasonable to require a more fine-tuned condition

$$p_f^+(h_{t_0}(x_{t_0})) \leq \delta = \frac{\alpha}{C} \tag{25}$$

where C is a number that contains information on how likely it is that the claim of T' is true at parameter value t_0. If one has a theoretical probability measure on R that indicates, before any actual test, how likely it is that the claim of the theory is true at specific parameter values, then C should reflect two features of this probability measure: (i) how much uncertainty the probability measure itself represents; (ii) what the probability of the particular parameter value t_0 is at which evidence x_{t_0} obtains: the larger the uncertainty, the larger C should be; the larger the probability of the particular parameter value t_0 at which evidence x_{t_0} obtains, the smaller C should be. We also demand that (iii) the correction (25) yields the Bonferroni correction (23) when the uncertainty represented by ρ is maximal: for a ρ that is uniform on the parameters, $C = N$ should hold.

A possible specification of this idea is the following. Let ρ be the probability on R; $\rho(t)$ is to be interpreted as the prior-to-test probability that the true parameter value is t. Taking the usual entropy $H(\rho) = -\sum_{t \in R} \rho(t) \log(\rho(t))$ as the uncertainty of the probability measure ρ, and assuming that $\rho(t_0) \neq 0$, the constant C_ρ defined by

$$C_\rho \doteq 1 + \frac{N-1}{N \log(N)} \cdot \frac{H(\rho)}{\rho(t_0)} \tag{26}$$

is a quantity that satisfies the requirements (i)–(iii): For fixed N, C_ρ is greater the larger the uncertainty expressed by the entropy of ρ, and C_ρ is smaller the more probable it is before the test that the confirming test result is at parameter value t_0. If ρ is the uniform distribution: $\rho(t) = 1/N$, $(t = 1, \ldots N)$, then $H(\rho) = \log(N)$, and $C_\rho = N$, thus yielding the Bonferroni correction. On the other extreme, if ρ is peaked at t_0, i.e. $\rho(t_0) = 1$, then $H(\rho)$ is zero, hence $C_\rho = 1$, i.e. no correction of the global significance α is required. In this case the Look Elsewhere Effect is neutralized completely by the prior-to-test certainty that the parameter at which the confirming evidence happens is t_0. In intermediate cases the entropic correction factor C_ρ given by (26) can take any value between N and 1 because the entropy function is continuous in the probability profile $(\rho(t_1), \rho(t_2), \ldots \rho(t_i) \ldots \rho(t_N))$: if $\rho(t_i) \to 1$, (and therefore $\rho(t_j) \to 0$ $(j \neq i)$), then $H(\rho) \to 0$. So, for ρ peaked enough around t_0, the entropy $H(\rho)$ will be small enough to entail that the entropic correction results in a less demanding significance threshold than the one coming from the Bonferroni correction.

If $\rho(t_0) = 0$, i.e. if the prior-to-test probability of the true parameter value being t_0 is zero, then one should distinguish two sub-cases:

(i) If the entropy $H(\rho)$ is non-zero, then one can choose a "cut-off" $\epsilon > 0$ to be put in place of $\rho(t_0)$ in the formula (26) that specifies C_ρ, yielding the required significance C_ϵ in this situation.

(ii) The entropy $H(\rho)$ also is zero. This is an extreme situation because now ρ is totally concentrated at a parameter t' that *differs* from t_0 where the evidence x_{t_0} is found. So one has maximal prior-to-test probabilistic certainty that the confirming evidence obtains at parameter t'; yet in the test the confirming evidence is found at parameter $t_0 \neq t'$. In this extreme, unexpected situation it is reasonable to demand a maximal (i.e. Bonferroni) correction of the significance.

Thus, instead of requiring (24), the entropic correction requires the local p-value to satisfy

$$p_f^+(h_{t_0}(x_{t_0})) \leq \frac{\alpha}{1 + \frac{N-1}{N \log(N)} \cdot \frac{H(\rho)}{\rho(t_0)}} \qquad \text{if } \rho(t_0) \neq 0 \tag{27}$$

$$p_f^+(h_{t_0}(x_{t_0})) \leq \frac{\alpha}{1 + \frac{N-1}{N \log(N)} \cdot \frac{H(\rho)}{\epsilon}} \qquad \text{if } \rho(t_0) = 0, \quad H(\rho) \neq 0 \tag{28}$$

$$p_f^+(h_{t_0}(x_{t_0})) \leq \frac{\alpha}{N} \qquad \text{if } \rho(t_0) = 0, \quad H(\rho) = 0 \tag{29}$$

We call C_ρ given by (26) the *entropic correction factor* and α/C_ρ the *entropic correction of significance*. C_ϵ denotes the entropic correction factor defined by the cutoff (i.e. C_ϵ is the denominator in the right hand side of (28)).

Similar considerations apply to the exclusion level condition expressed by Eq. (8): The significance β also gets diluted by the Look Elsewhere Effect. That is to say, as the number of parameters increases it becomes more and more likely that one finds

an x_{t_0} at some parameter value t_0 such that the local p-value $p'^{-}_{f}(h_{t_0}(x_{t_0}))$ is larger than β even if T' is false. So, the local p-value $p'^{-}_{f}(h_{t_0}(x_{t_0}))$ is an overestimation of a possible compatibility of T' with the evidence x_{t_0}. Thus one needs a correction of β; but, in contrast to α, the correction should *increase* β. To correct the significance β using the entropic correction factor C_ρ one can demand

$$p'^{-}_{f}(h_{t_0}(x_{t_0})) \geq 1 - \frac{1-\beta}{C_\rho} = 1 - \frac{1-\beta}{1 + \frac{N-1}{N \log(N)} \cdot \frac{H(\rho)}{\rho(t_0)}} \quad \text{if } \rho(t_0) \neq 0,\ H(\rho) \neq 0 \tag{30}$$

Just like in the case of the entropic compensation of α, this correction possesses the intuitively desirable properties: given N, the larger the entropy of ρ, the larger the correction of β (i.e. the larger the right hand side of (30)); and the larger the prior-to-test probability $\rho(t_0)$ that the confirming evidence is at parameter t_0, the smaller the correction of β (i.e. the smaller the right hand side of (30)). If ρ is the uniform probability hence the entropy is maximal and is equal to $\log(N)$, the entropic correction of β reaches its maximum $1 - \frac{1-\beta}{N}$. This corresponds to the Bonferroni correction of α. If the entropy of ρ is zero and thus ρ is concentrated on t_0, entailing $C_\rho = 1$, then the entropic compensation given by (30) leaves the significance β unchanged—the Bayesian certainty overrides the Look Elsewhere Effect. In intermediate cases the entropic correction of β provided by (30) forces the local p-value $p'^{-}_{f}(h_{t_0}(x_{t_0}))$ to be larger than β by an amount that is sensitive to the global properties of ρ and the prior-to-test probability $\rho(t_0)$. This will be illustrated in Sect. 4.1.

If $\rho(t_0) = 0$, then, according to the two sub-cases $H(\rho) \neq 0$ and $H(\rho) = 0$ distinguished above in connection with the correction of α, the entropic corrections of β corresponding to (28) and (29) are:

$$p'^{-}_{f}(h_{t_0}(x_{t_0})) \geq 1 - \frac{1-\beta}{C_\epsilon} \quad \text{if } \rho(t_0) = 0, \quad H(\rho) \neq 0 \tag{31}$$

$$p'^{-}_{f}(h_{t_0}(x_{t_0})) \geq 1 - \frac{1-\beta}{N} \quad \text{if } \rho(t_0) = 0, \quad H(\rho) = 0 \tag{32}$$

To summarize: If one has a prior-to-test probability measure ρ indicating the probability $\rho(t)$ of finding a confirming evidence at parameter value t, then evidence x_{t_0} T-CS-confirms T' at parameter value t_0 with significance α and exclusion level β if conditions (27)–(29) and (30)–(32) hold, and this T-CS confirmation of T' has taken into account the Look Elsewhere Effect tamed entropically using the information contained in ρ.

4.1 Illustrative example: coin-flipping at different temperatures

To illustrate the entropic taming consider the parametrized version of the coin-flipping example in Sect. 2: Assume that now the theory T' is that a coin is fair but at a particular temperature in the range of 0–100 °C becomes biased towards Head in 4:1 proportion. Theory T is that the coin is fair at any temperature. The task is to T-CS-confirm T'.

Assume that the temperature range is divided into 100 bins d_i of equal length (1 degree) and ten flips are made at a temperature $t_i \in d_i$ $(i = 1, \ldots, 100)$ in each

bin. The probability spaces $(X_{t_i}, \mathcal{S}_{t_i}, p_{t_i})$ $(i = 1, \ldots, 100)$ in this case are taken to be isomorphic via h_{t_i} with the probability space describing ten flips with uniform probability (the space (9)–(10) with $M = 10$). The probability spaces $(X_{t_i}, \mathcal{S}_{t_i}, p'_{t_i})$ $(i = 1, \ldots, 100)$ are taken to be isomorphic with the probability space describing 10 flips with the 4:1 bias towards Head, with p' given by (12). Suppose that the evidence is a set of flips with the coin ten times at each temperature t_i, and that this set contains the outcome described by Eq. (13), obtained at temperature 50°:

$$x_{50^\circ} = (H, H, T, H, H, H, H, H, H, H)_{50^\circ} \tag{33}$$

The local p-value of x_{50° was calculated in Sect. 3 and is given by Eq. (14):

$$p_{50^\circ}{}^{+}_{f_{50^\circ}}(x_{50^\circ}) = p_{10}{}^{+}_{f}(h_{50^\circ}(x_{50^\circ})) \tag{34}$$

$$= p_{10}{}^{+}_{f}((H, H, T, H, H, H, H, H, H, H)) \tag{35}$$

$$= 0.01074 \tag{36}$$

This local p-value was low enough to reject the null-theory of fairness of the coin at significance level 0.03 in the case of the parameter-free test. In the current, parametric case, the Bonferroni correction would not allow this conclusion because the Bonferroni correction would require the local p-value to be less than $0.03/100 = 0.0003$. In fact, since the smallest non-zero probability of an event now is 2^{-10}, and $2^{-10} > 0.0003$, T' could not be T-CS-confirmed *at all* by throwing 10 times with significance required by the Bonferroni correction because it is impossible to reject the null hypothesis with the required significance.

But suppose we have the following prior-to-flipping probability ρ about the temperature where the bias might occur:

$$\rho(t_i) = 0.05/98 \qquad t_i \neq 50^\circ, 25^\circ \tag{37}$$

$$\rho(50^\circ) = 0.9 \tag{38}$$

$$\rho(25^\circ) = 0.05 \tag{39}$$

The entropic corrections of the significance and exclusion levels can be easily calculated:

(i) The entropic correction of the 0.03 significance for x_{50° for such a ρ is 0.02611. Since the local p-value 0.01074 of the evidence x_{50° is below 0.02611, this evidence allows the rejection of T.

(ii) Calculating the entropic correction of the exclusion level $\beta = 0.7$ one obtains 0.73890 (rounded up to 5 decimals). Since the local p-value $p'_{10}{}^{-}_{f}(x) = 0.89263$ in Eq. (15) is above this entropically corrected exclusion level, T' is not rejected at parameter value 50°.

In view of (i)–(ii) above, theory T' is T-CS-confirmed at 50° by x_{50°. This confirmation is at significance 0.02611 and exclusion level 0.73890, where these levels take into account the Look Elsewhere Effect, tamed in an entropic manner.

The entropic correction of significance for an event consisting of nine Heads and one Tail, like x_{50° but occurring at 25° degrees, can also be calculated: it is 0.00814. This value is below of the local p-value 0.01074 of the evidence x_{50°. Hence, if the evidence of throwing 9 Heads in 10 throws were found at 25°, one would *not* be allowed to reject the hypothesis T that the coin is fair at this temperature and thus one would not be able to conclude that T' is T-CS-confirmed at 25°. This negative conclusion would be the result of the Look Elsewhere Effect—even after having tamed the Look Elsewhere Effect in the entropic manner.

5 Assessing the performance of the entropic correction

If the coin in the example of the previous section does indeed change its non-biased character at $t = 50^\circ$, then the ρ assigning high (0.9) probability to $t = 50^\circ$, is a "good" prior because it leads to the correct decision. And the example also shows that this good decision is *not* possible on the basis of the Bonferroni correction (with the chosen significance levels)—so the entropic correction performs better than the Bonferroni correction in this case. If the prior assigning 0.9 probability to $t = 50^\circ$ is a very *bad* reflection/anticipation of the parameter value at which the new theory is in fact true (because the coin changes its non-biased character *not* at $t = 50^\circ$, yet the prior assigns high probability to this), then rejection/acceptance based on the entropic correction (on the basis of the evidence of throwing 9 Heads) is an error. In this latter case the entropic correction did not perform well. This observation leads to the general question of how the taming of the Look Elsewhere Effect using the entropic correction compares with using the Bonferroni correction in general in the sense that using the entropic correction is more likely to result in rejecting hypotheses that are more likely false and is more likely to result in accepting hypotheses that are more likely true.[1] To formulate this question more precisely, one needs a measure of good performance of a correction of significance in a parametric T-CS confirmation. Below we define explicitly a measure that seems suitable to reflect the degree of good performance—without claiming that it is the only one that does so.

Suppose ρ is understood as a probability measure that reflects some objective chances (e.g. frequencies) with which T' is true at the parameter values. Suppose one tests T' at each parameter via a T-CS confirmation, using both the entropic and the Bonferroni corrections of the significance levels α and β. Let $P_{Ent}^{(\alpha,\beta)}(t)$ be the probability with which T' is confirmed when T-CS-tested at parameter t using the entropic corrections of α and β, and $P_{Bonf}^{(\alpha,\beta)}(t)$ be the probability with which T' is confirmed when T-CS-tested at parameter t using the Bonferroni corrections of α and β. At t the entropic correction performs better than the Bonferroni correction if $P_{Ent}^{(\alpha,\beta)}(t)$ is closer to $\rho(t)$ than $P_{Bonf}^{(\alpha,\beta)}(t)$ is, i.e. if

$$|P_{Ent}^{(\alpha,\beta)}(t) - \rho(t)| \leq |P_{Bonf}^{(\alpha,\beta)}(t) - \rho(t)| \tag{40}$$

[1] We thank an anonymous referee who suggested to investigate this issue in more detail.

The ρ-expected values of the left and right-hand sides of the inequality Eq. (40) are reasonable indicators of how well the entropic and Bonferroni corrections perform at this ρ at the significance levels α and β. So we define performance indicators $I^{Ent}_{\rho;(\alpha,\beta)}$ and $I^{Bonf}_{\rho;(\alpha,\beta)}$ by

$$I^{Ent}_{\rho;(\alpha,\beta)} \doteq \sum_t \rho(t)|P^{(\alpha,\beta)}_{Ent}(t) - \rho(t)| \tag{41}$$

$$I^{Bonf}_{\rho;(\alpha,\beta)} \doteq \sum_t \rho(t)|P^{(\alpha,\beta)}_{Bonf}(t) - \rho(t)| \tag{42}$$

And we say:

Definition 3

- The entropic correction performs better than the Bonferroni correction at ρ at the significance levels (α, β) if

$$I^{Ent}_{\rho;(\alpha,\beta)} \leq I^{Bonf}_{\rho;(\alpha,\beta)} \tag{43}$$

- The entropic correction performs better than the Bonferroni correction *overall* if for all ρ and for all significance levels α and β the inequality (43) holds.

Problem 1 *Is the overall performance of the entropic correction better than the overall performance of the Bonferroni correction?*

Problem 1 seems to be a difficult one; at any rate we are unable to solve it. What we can do is to offer numerical evidence in the particular case of the ρ specified in our example in Sect. 4.1 (see Eqs. (37)–(39)) that, for this ρ, and for the significance levels $\alpha = 0.3$ and $\beta = 0.7$, the entropic correction performs better than the Bonferroni correction. In this particular case the performance indicators $I^{Ent}_{\rho;(0.3,0.7)}$ and $I^{Bonf}_{\rho;(0.3,0.7)}$ can be calculated explicitly. We calculated numerically these numbers for five different testing situations: when we use as evidence outcomes of 10, 15, 20, 25 and 30 flips. Some details of the calculations are in Sect. 2 of the Appendix. The result of the calculations show (see Table 2 in Sect. 2 of the Appendix) that the entropic correction performs better in the case of this particular ρ and at these significance levels in each of these five testing scenarios:

Number of flips	$I^{Ent}_{\rho;(0.3,0.7)}$	$I^{Bonf}_{\rho;(0.3,0.7)}$
10	0.506838084	0.81252551
15	0.489350944	0.783931589
20	0.301453498	0.80315755
25	0.312234809	0.790270514
30	0.465918806	0.803975713

Comments on the numerical calculations:

(i) While in our example the performance of the entropic correction is better than the performance of the Bonferroni correction in the sense that in each one of the testing scenarios $I^{Ent}_{\rho;(0.3,0.7)} < I^{Bonf}_{\rho;(0.3,0.7)}$ holds, the numerical calculations also show that it is *not* true that in each one of the testing scenarios the performance of the entropic correction is better *at every* single parameter: it is *not* true that $|P^{(0.3,0.7)}_{Ent}(t) - \rho(t)| \leq |P^{(0.3,0.7)}_{Bonf}(t) - \rho(t)|$ holds for *all* t in each one of the testing scenarios. The reason for this is that when $\rho(t)$ is very small, the entropic correction can be stronger than the Bonferroni correction, and this can make it too hard to T-CS confirm the hypothesis at that parameter using the entropic correction. In our example this happens in the testing scenario in which the test is a sequence of tosses of length 30 and when it comes to the confirmation of the coin transforming into a biased one at a temperature different from 50° and 25° with a probability of $0.05/98 = 0.510204 \times 10^{-3}$. In this case we have: $P^{(0.3,0.7)}_{Ent}(t) = 0.632533 \times 10^{-6}$ and $P^{(0.3,0.7)}_{Bonf}(t) = 0.539747 \times 10^{-5}$. But since the transformation of the coin into biased one at temperatures different from 50° and 25° happen with a very low probability, the average performance of the entropic correction is less affected by the low probability of confirming the hypothesis at such a parameter.

(ii) The observation in (i) shows that in our example the entropic correction performs better than the Bonferroni correction at parameters t for which $\rho(t)$ is higher. This indicates a general feature of the entropic correction: if ρ is a probability measure that reflects some objective chances (e.g. frequencies) with which T' is true at the parameter values, then the T-CS confirmation using the entropic correction is more likely to confirm the hypothesis as true at parameters at which the hypothesis are objectively more likely true than is the T-CS confirmation using the Bonferroni correction.

One also can raise the question of how the entropic correction performs in comparison with other corrections available in the literature that are different from the Bonferroni correction. We do not have any result in this direction.

6 Concluding comments

The entropic taming of the Look Elsewhere Effect allows one to introduce in a disciplined and technically explicit manner a Bayesian component in the otherwise frequentist statistical theory testing. This is in the spirit of Dawid: "... keeping in mind a Bayesian framework can help making the right choices in framing data analysis" Dawid (2017). But the kind of Bayesianism the entropic correction represents differs significantly from the sort of Bayesianism present in Dawid (2015, 2017). Dawid's analysis in both Dawid (2015, 2017) involves the subjective prior probabilities about the truth of the null hypothesis and of the theory to be tested. Such subjective priors are completely absent in the entropic correction of the significance levels: Degrees of beliefs enter the entropic correction of significance only through the prior-to-test probability ρ representing expectations about where the confirming evidence might occur; no prior degree of belief in the theory to be tested or in the null hypothesis is

assumed. Also, Dawid's analysis in Dawid (2017) assumes explicitly that "... the probability of the existence of a Higgs particle is uniformly distributed over the allowed mass region..." Dawid (2017). Using our terminology, this amounts to assuming that ρ is the uniform probability—and the main point in the entropic correction is that ρ need not be uniform.

That the entropic correction of the Look Elsewhere Effect does not involve the prior belief in the theory under test has the advantage that it is applicable even when the prior belief in the new theory T' is not very high: Even if one does not have a high degree of belief in the truth of T', one might have a high probability that *if* T' is true, the conforming evidence obtains with higher probability at some specific parameter values. This allows the taming of the Look Elsewhere Effect in the entropic way in situations when the overall confidence in the truth of T' is low or moderate.

"Uncertainty" is not a uniquely determined notion and there are measures different from entropy that express how uncertain or "peaked" a probability measure is. So there seems to exist in principle an avenue for different types of Bayesian taming of the Look Elsewhere Effect, where "type" refers to the type of uncertainty one uses to measure the uncertainty of a Bayesian ρ. So no claim is made here about the uniqueness of the suggested entropic correction as the only one that satisfies the desiderata formulated in Sect. 1. But we do not have an example of a measure of uncertainty of a probability measure and corrections of significance that differ from the entropic one and which satisfy the desiderata in section. So we leave the question about the existence of such a correction as an open issue. One also might consider further strengthening the demands (i)–(iii) on a correction formulated in Sect. 1. We also leave it for future analysis in which direction such strengthening are feasible or desirable.

Another possible topic for future investigation is the situation when the parameter space is not finite; in particular when it is uncountably infinite. This situation presents several technical and conceptual difficulties. In such a situation the Bonferroni correction is not applicable and it is unclear to us how one could tame the Look Elsewhere Effect along the lines suggested in this paper.

The entropic taming of the Look Elsewhere Effect itself is neutral with respect to the source of the probability measure ρ: One can interpret ρ entirely subjectively, as expressing a rather personal conviction (degree of belief) of the experimenter (e.g. physicist) carrying out the test. But the probability measure ρ can also be the prediction of (or informed by) a well-corroborated framework theory that provides a probabilistic prediction regarding the chances of finding specific systems with certain parameter values. In this latter case the entropic taming of the Look Elsewhere Effect is in the spirit of the advice of the American Statistical Association (Wasserstein & Lazar, 2016), which emphasizes the importance of taking into account the broader scientific context in which statistical analysis using p-values takes place.

Acknowledgements Research supported by the Hungarian National Research, Development and Innovation Office, contract number: K-134275. M. Rédei thanks the Hamburg Institute for Advanced Study https://hias-hamburg.de/, where the final revision of the paper was completed.

Funding This work was supported by the Hungarian National Research, Development and Innovation Office ("Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal"), Grant number: K-134275. Both authors are members of the team that is the recipient of this grant.

Declarations

Conflict of interest Both authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Appendix

Calculation of the global *p*-value when the probabilities of events at different parameter values are independent

As an explicit illustration of the Look Elsewhere Effect and the global p-value, we recall here the well-known calculation of the global p-value of x_{t_0} in the case when the probabilities of events at different parameter values are independent. The calculations (especially in lines (46)–(49)) below use the assumptions (I)–(III).

Let $(\times_{i=1}^{N} X_{t_i}, \times_{i=1}^{N} \mathcal{S}_{t_i})$ be the measurable space that describes the possible outcomes of tests carried out at parameter values in R. Then

$$E_{t_0} = \{y_{t_o} \in X_{t_0} : f_{t_0}(y_{t_0}) \geq f_{t_0}(x_{t_0})\} \tag{44}$$

is the event in $\mathcal{S}_{t_0}$ that the result of the empirical test x_{t_0} has value $f_{t_0}(x_{t_0})$ or a larger value. The complement $E_{t_0}^{\perp}$ of E_{t_0} in X_{t_0} is the event in $\mathcal{S}_t$ that E_{t_0} does not happen at parameter value t_0. The event $({h_t}^{-1} \circ h_{t_0})(E_{t_0}^{\perp})$ is the event that the event E_{t_0} does not happen at parameter value t, and the event that E_{t_0} does not happen at *any* parameter value is represented by $\overline{E_0}$, where

$$\overline{E_0} \doteq \times_{i=1}^{N} \left[({h_{t_i}}^{-1} \circ h_{t_0})(E_{t_0}^{\perp}) \right] \in \times_{i=1}^{N} \mathcal{S}_{t_i} \tag{45}$$

So if one takes the product probability measure $\times_{i=1}^{N} p_{t_i}$ on $\times_{i=1}^{N} \mathcal{S}_{t_i}$, then the probability that E_{t_0} does not happen at *any* value of the parameter is

$$\begin{aligned} \times_{i=1}^{N}\, p_{t_i}(\overline{E_0}) &= \Pi_i\, p_{t_i}(({h_{t_i}}^{-1} \circ h_{t_0})(E_{t_0}^{\perp})) && (46)\\ &= \Pi_i\, p((h_{t_0})(E_{t_0}^{\perp})) = \Pi_i\, p(X \setminus h_{t_0}(E_{t_0})) && (47)\\ &= \Pi_i [1 - p(h_{t_0}(E_{t_0}))] = [1 - p(h_{t_0}(E_{t_0}))]^N && (48)\\ &= [1 - p_f^{+}(x_{t_0})]^N && (49) \end{aligned}$$

Table 1 Testing biasedness of coin at temperature 50 degrees centigrade, using entropic and Bonferroni corrections and flipping coin 15 times

# of Heads	$p^+(n)$	$p^-(n)$	E-conf.?	B-conf.?
0	1	0.32768×10^{-10}	0	0
1	0.999542236	0.199885×10^{-8}	0	0
2	0.999511719	0.570491×10^{-7}	0	0
3	0.996307373	0.101125×10^{-5}	0	0
4	0.982421875	0.124617×10^{-4}	0	0
5	0.940765381	0.113226×10^{-3}	0	0
6	0.849121094	0.784985×10^{-3}	0	0
7	0.696380615	0.423975×10^{-2}	0	0
8	0.5	$0.18058807 \times 10^{-1}$	0	0
9	0.303619385	0.6105143×10^{-1}	0	0
10	0.150878906	0.164233724	0	0
11	$0.59234619 \times 10^{-1}$	0.351837895	0	0
12	$0.17578125 \times 10^{-1}$	0.601976791	0	0
13	0.3692627×10^{-2}	0.832874233	1	0
14	0.488281×10^{-3}	0.964815628	1	0
15	0.305176×10^{-4}	1	1	1

So the probability that E_{t_0} does happen at *some* value of the parameter is

$$1 - \times_{i=1}^{N} p_{t_i}(\overline{E_0}) = 1 - [1 - p_f^+(x_{t_0})]^N \tag{50}$$

Numerical calculations

We show here how to calculate the performance indicators $I^{Ent}_{\rho;(\alpha,\beta)}$ and $I^{Bonf}_{\rho;(\alpha,\beta)}$ defined in Sect. 5, in the case of ρ specified by Eqs. (37)–(39) in our example in Sect. 4.1, for the significance levels $\alpha = 0.3$ and $\beta = 0.7$. We show the steps in the calculations in the testing scenario in which the testing is in terms of $M = 15$ flips of the coin.

Table 1 shows the numerical results in case of testing biasedness of the coin at 50° degrees. The entropic corrections of the significance levels $\alpha = 0.3$ and $\beta = 0.7$ at this temperature are: 0.026110449 and 0.738895514. Table 1 has the following data:

1. The first column in Table 1 shows the possible numbers of Heads in 15 flips.
2. The second column shows the *positive* side p-values $p^+(n)$ of the events of obtaining n Heads in 15 flips with the *unbiased* coin. The formula used to calculate this number is:

$$p^+(n) = 1 - \sum_{i=0}^{n-1} B(i, 15, 0.5) \tag{51}$$

Here, and in what follows, $B(j, k, r)$ is the Binomial distribution with parameter $r \in [0, 1]$:

$$B(j, k, r) = \binom{k}{j} r^j (1 - r)^{k-j} \tag{52}$$

3. The third column shows the *negative* side p-values $p^-(n)$ of the events of obtaining n Heads in 15 flips with the *biased* coin. The formula used to calculate this number is:

$$p^-(n) = \sum_{i=0}^{n} B(i, 15, 0.8) \tag{53}$$

4. The fourth column *E-conf.?* contains
   - 1 in row with n Heads if n Heads in a sequence of 15 flips T-CS confirms T' (= the coin turns biased at temperature 50° degrees) after the significance has been corrected in an *entropic* manner;
   - 0 otherwise.
5. The fifth column *B-conf.?* contains
   - 1 in row with n Heads if n Heads in a sequence of 15 flips T-CS confirms T' (= the coin turns biased at temperature 50° degrees) after the significance has been *Bonferroni* corrected;
   - 0 otherwise.

Let g_E be the function that assigns to the number n of Heads in the table the corresponding value 1 or 0 in column 4 (indicating whether T' is T-CS confirmed by n Heads in 15 flips, after the entropic correction of significance); and, let g_B the similarly defined function for the confirmation using the Bonferroni correction.

Let p_{50} be the probability measure that gives the probability $p_{50}(n)$ that n Heads occur in 15 flips at temperature 50°. Since the probability that the coin turns into the biased one at 50° degrees is 0.9, we have:

$$p_{50}(n) = 0.9 \cdot B(n, 15, 0.8) + 0.1 \cdot B(n, 15, 0.5) \tag{54}$$

Then the expectation value of g_E with respect to p_{50} yields the probability $P_{Ent}^{(0.3,0.7)}(50°)$ with which T' is confirmed when T-CS-tested at parameter 50° using the entropic correction:

$$\begin{aligned} P_{Ent}^{(0.3,0.7)}(50°) &= \sum_{i=0}^{15} p_{50}(i) g_E(i) && (55)\\ &= p_{50}(13) g_E(13) + p_{50}(14) g_E(14) + p_{50}(15) g_E(15) && (56)\\ &= 0.358590151 && (57) \end{aligned}$$

Table 2 Numerical values of performance indicators of entropic and Bonferroni corrections in case of flipping coin 10, 15, 20, 25 and 30 times. Smaller value indicates better perfromance

Number of flips	$I^{Ent}_{\rho;(0.3,0.7)}$	$I^{Bonf}_{\rho;(0.3,0.7)}$
10	0.506838084	0.81252551
15	0.489350944	0.783931589
20	0.301453498	0.80315755
25	0.312234809	0.790270514
30	0.465918806	0.803975713

Exactly the same type of calculation can be used to obtain the probability $P^{(0.3,0.7)}_{Ent}(25°)$ with which T' is confirmed when T-CS-tested at parameter 25°, and the probability $P^{(0.3,0.7)}_{Ent}(t')$ with which T' is confirmed when T-CS-tested at parameters $t' \neq 50°, 25°$, using the entropic correction. The results of the calculation are:

$$P^{(0.3,0.7)}_{Ent}(25°) = 0.8820156 \times 10^{-2} \tag{58}$$

$$P^{(0.3,0.7)}_{Ent}(t') = 0.484532 \times 10^{-4} \tag{59}$$

The entropic performance indicator $I^{Ent}_{\rho;(0.3,0.7)}$ can now be calculated:

$$I^{Ent}_{\rho;(0.3,0.7)} \doteq \sum_t \rho(t)|P^{(0.3,0.7)}_{Ent}(t) - \rho(t)| \tag{60}$$

$$= 0.9\left|P^{(0.3,0.7)}_{Ent}(50°) - 0.9\right| + 0.05\left|P^{(0.3,0.7)}_{Ent}(25°) - 0.05\right| \tag{61}$$

$$+98 \cdot 0.05/98\left|P^{(0.3,0.7)}_{Ent}(t') - 0.05/98\right| \tag{62}$$

$$= 0.489350944 \tag{63}$$

An entirely analogous calculation yields the performance indicator of the Bonferroni correction:

$$I^{Bonf}_{\rho;(0.3,0.7)} = 0.783931589 \tag{64}$$

The calculation of the performance indicators in case of 10, 20, 25 and 30 flips proceeds exactly along the above lines. The results are provided in Table 2.

References

Bayer, A. E., & Seljak, U. (2020). The look-elsewhere effect from a unified Bayesian and frequentist perspective. *Journal of Cosmology and Astroparticle Physics, 2020*(10), 1–22.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289–300.

Bogachev, V. I. (2007). *Measure theory* (Vol. II). Springer.

Dawid, R. (2015). Higgs discovery and the Look Elsewhere Effect. *Philosophy of Science, 82*, 76–86.

Dawid, R. (2017). Bayesian perspectives on the discovery of the Higgs particle. *Synthese, 194*, 377–394.

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics, 2*, 197–223.

Foster, J. J., Barkus, E., & Yavorsky, C. (2006). *Understanding and using advanced statistics*. Sage.
Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed). W.W. Norton & Company.
Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *European Physical Journal C, 70*, 525–530.
Hartmann, S., & Sprenger, J. (2019). *Bayesian philosophy of science*. Oxford University Press.
Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800–802.
Lehmann, E. L., & Romano, J. P. (2005a). Generalizations of the familywise error rate. *The Annals of Statistics, 33*, 1138–1154.
Lehmann, E. L., & Romano, J. P. (2005b). *Testing statistical hypotheses*. Springer texts in statistics (3rd ed.). Springer.
List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics, 22*, 773–793.
Lyons, L. (2008). Open statistical issues in particle physics. *The Annals of Applied Statistics, 2*, 887–915.
Lyons, L. (2013). Bayes and frequentism: A particle physicist's perspective. *Contemporary Physics, 54*, 1–16.
Maxwell, N. P. (1994). A coin-flipping exercise to introduce the *p*-value. *Journal of Statistics Education, 2*, 1–4.
Rédei, M., & Gyenis, Z. (2021). The maxim of probabilism, with special regard to Reichenbach. *Synthese, 199*, 8857–8874.
Romeijn, J.-W. (2017). Philosophy of statistics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 edition). Metaphysics Research Lab, Stanford University.
Sprenger, J. (2016). Bayesianism vs frequentism in statistical inference. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (Chap. 18, pp. 382–405). Oxford University Press.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*, 129–133.