DOI: 10.1017/psa.2024.13
This is a manuscript accepted for publication in *Philosophy of Science*.
This version may be subject to change during the production process.

Bamboozled by Bonferroni

Conor Mayo-Wilson
University of Washington, Department of Philosophy, Seattle, USA

Abstract

When many statistical hypotheses are evaluated simultaneously, statisticians often recommend adjusting (or "correcting") standard hypothesis tests. In this paper, I (1) distinguish two senses of adjustment, (2) investigate the prudential and epistemic goals that adjustment might achieve, and (3) identify conditions under which a researcher should *not* adjust for multiplicity in the two senses I identify. I tentatively conclude that the goals of scientists and the public may be misaligned with the decision criteria used to evaluate multiple testing regimes.

Imagine a pharmaceutical company spends years developing a new cancer treatment. Because of the expense of drug development, the company collects extensive data during human trials. In particular, researchers collect data about hundreds of health outcomes other than cancer. When the data is analyzed, researchers find that treatment is associated with a reduction in breast cancer. Here's an instance of a more general question.

Question: Should the pharmaceutical researchers alter their methods for analyzing the cancer data because the treatment for efficacy was assessed in *many* other ways?

According to many statisticians and scientists, "yes." Let *multiplicity* refer to the act of evaluating many statistical hypotheses simultaneously. When multiplicity occurs, many statisticians and scientists recommend "correcting"[1] p-values so as to reduce the number of false positive results.[2] Although Bayesian statisticians reject the use of p-values, many likewise argue that one's statistical methods should be adjusted for multiplicity.[3] This

[1] Henceforth, I say "adjust" rather than "correct" so as to avoid suggesting that adjustment is good or obligatory.

[2] See [Lehmann and Romano, 2008, Chapter 9]. The most common classical techniques are Bonferroni's method and [Benjamini and Hochberg, 1995]'s procedure.

[3] See [Berry and Hochberg, 1999] and [Scott and Berger, 2006] for discussions of Bayesian approaches to multiplicity.

raises a very general question.

Central Question: Under what conditions, if any, should statistical methods be adjusted for multiplicity? In what way should they be adjusted? And why?

The central question is important because, as our computational power grows, so does our ability to evaluate thousands of policy-relevant statistical hypotheses in a matter of minutes.

Although statisticians have investigated the reliability of many adjustment procedures, few have clarified the central question. What exactly is adjustment? Can "adjustment" be defined without reference to particular statistical methods? If "adjusting" means "changing reported p-values", then devout Bayesian statisticians never adjust for multiplicity, as they avoid calculating p-values![4] So is there a sense of "adjustment" that renders classical and Bayesian approaches comparable?

The normative dimensions of the central question have also yet to be clarified. In what sense "should" one adjust for multiplicity? Is adjustment *rationally* required to achieve certain goals? If so, which goals? Is adjustment *epistemically* required to respect one's evidence? Is it *scientifically* required by norms of scientific inquiry? Is it *ethically* obligatory? If adjustment is not obligatory, is it permissible or good in any sense?[5]

Finally, answers to those normative questions depend upon *who* or *what* is adjusting. Researchers can adjust reported p-values. But so can journal editors. Grant-giving agencies – like the National Institutes of Health (NIH) – can also adjust for multiplicity in various ways. Which, if any, of these decision-making bodies should adjust?

The main contribution of this paper is to (1) distinguish two senses of adjustment, (2) investigate the prudential and epistemic goals that adjustment might achieve, and (3) formulate more precise versions of the central question. I also prove a new theorem characterizing when adjustment is *impermissible*. I tentatively conclude that there is a mismatch between the goals of scientists (both individually and collectively) and the guarantees of existing adjustment procedures. This paper, thus, is a call for further research: we must either prove existing adjustment methods achieve goals

[4]See [Rubin, 2021] for an attempt to answer the central question when "adjustment" is interpreted narrowly about significance levels.

[5]Philosophers have focused on evidential questions. See [Kotzen, 2013] and [Mayo, 2018]. In contrast, most statisticians employ a quasi-decision-theoretic framework, which seems most suited for questions of rationality.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

of actual scientific interest, or develop alternative procedures.

1 Basic Model

To distinguish types of adjustment, I introduce a model. Suppose N hypotheses are under investigation. Assume that any subset of the N hypotheses might be true. Let $\Theta = \{0,1\}^N$ be the set of all binary strings/vectors of length N. A vector $\theta \in \Theta$, therefore, specifies which of the N hypotheses are true and which are false. Let $H_k = \{\theta \in \Theta : \theta_k = 0\}$ be the set of vectors that say the k^{th} hypothesis is true.

Suppose that, for each hypothesis H_k, there is some experiment X_k that *could* be conducted (or observation that *could* be made); researchers believe X_k could be informative about whether H_k holds. Formally, X_k is a random variable, and for each $\theta \in \Theta$, let $\mathbb{P}_\theta(X_1, \ldots, X_N)$ denote the probability measure that specifies the chances of various experimental outcomes.

For simplicity, assume that, for all $\theta \in \Theta$, the N experiments are *mutually independent* with respect to $\mathbb{P}_\theta$. In symbols, let $\vec{X} = \langle X_{i_1}, X_{i_2}, \ldots, X_{i_k} \rangle$ be a random vector, representing some subset of the N experiments. Then for all sequences $\vec{x} = (x_{i_1}, \ldots x_{i_k})$ representing the outcome of those $k \leq N$ experiments:

$$\mathbb{P}_\theta(\vec{X} = \vec{x}) = \prod_{j \leq k} \mathbb{P}_\theta(X_{i_j} = x_{i_j}). \tag{1}$$

Further, suppose that the truth or falsity of the H_k entirely determines the probabilities of the possible outcomes of the k^{th} experiment, i.e., for all $k \leq N$ and all $r \in \{0,1\}$, there is a probability distribution $\mathbb{P}_{k,r}$ such that $\mathbb{P}_\theta(X_k = x_k) = \mathbb{P}_{k,\theta_k}(X_k = x_k)$. Together with the assumption of mutual independence, this entails that:

$$\mathbb{P}_\theta(\vec{X} = \vec{x}) = \prod_{j \leq k} \mathbb{P}_{i_j,\theta_{i_j}}(X_{i_j} = x_{i_j}) \text{ for all } \theta \in \Theta. \tag{2}$$

To assess whether a decision-maker should adjust for multiplicity, compare two types of situations. In the first, the decision-maker learns the outcome of a proper subset of the N tests.

For simplicity, suppose that the researcher learns only the value of X_1. In the second, she learns the values of *all* N variables. Say that the decision-maker should adjust for multiplicity if her (1) beliefs or (2) decisions about H_1 should differ in those two situations. Let's clarify those two senses of adjustment.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

2 Belief

For the Bayesian, beliefs are modeled by posterior probabilities, and so a Bayesian adjusts for multiplicity if there is a value x_1 of X_1 such that

$$P(H_1|X_1 = x_1) \neq P(H_k|X_1 = x_1, \ldots X_N = x_N) \tag{3}$$

for all values $x_2, \ldots x_N$ of $X_2, \ldots X_N$ for which $P(X_1 = x_1, \ldots X_N = x_N) > 0$. One could distinguish a weaker sense of adjustment, whereby Equation 3 holds for *some* values of $X_2, \ldots X_N$. For critics of Bayesianism, one can replace the probability functions in Equation 3 with another object representing belief.[6]

Should one ever adjust for multiplicity, in the strong sense just identified? Yes. Consider a Bayesian researcher who regards the hypotheses as dependent, so that learning about one hypothesis provides evidence about another. For example, suppose our hypothetical pharmaceutical researchers consider two hypotheses: (1) The treatment is not effective in 33-year-old women and (2) The treatment is not effective in 34-year-old women. A researcher might reasonably believe that the first hypothesis is true if and only if the second is. If so, acquiring data about 33-year-old women would provide evidence about the efficacy of the treatment for 34-year-old women. Here's a toy model to illustrate such adjustment.

Example 1: Suppose each X_k is a binary random variable that represents a test to retain or reject H_k. Assume there are $\alpha, \beta \in (0, 1)$ such that for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(X_k = 1) = \begin{cases} \alpha & \text{if } \theta_k = 0 \\ 1 - \beta & \text{if } \theta_k = 1 \end{cases}$$

That is, each test X_k has a Type I error of α and Type II error of β.

To model a researcher who believes the hypotheses to be dependent, suppose that the researcher assigns positive probability to precisely two vectors in Θ, namely, $\mathbf{0} = \langle 0, \ldots, 0 \rangle$ – which says each H_k is true – and $\mathbf{1} = \langle 1, \ldots, 1 \rangle$ – which says each H_k is false. If $\pi = P(\mathbf{0}) = 1 - P(\mathbf{1})$ represents the researcher's prior degree of belief that all hypotheses are true, then her posterior probability in H_1 if she learns only that the first test is negative equals the following:

$$P(H_1|X_1 = 0) = \frac{\pi \cdot (1 - \alpha)}{\pi \cdot (1 - \alpha) + (1 - \pi) \cdot \beta}. \tag{4}$$

[6]For instance, one might represent belief using orderings [Mayo-Wilson and Saraf, 2022], ranking functions [Spohn, 2012], Dempster-Shafer functions [Dempster, 1968], etc.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

In contrast, if she learns *two* tests are negative, her posterior is:

$$P(H_1|X_1 = 0, X_2 = 0) = \frac{\pi \cdot (1-\alpha)^2}{\pi \cdot (1-\alpha)^2 + (1-\pi) \cdot \beta^2} \tag{5}$$

Finally, if she learns the second test is positive, her posterior will be:

$$P(H_1|X_1 = 0, X_2 = 1) = \frac{\pi \cdot (1-\alpha) \cdot \alpha}{\pi \cdot (1-\alpha) \cdot \alpha + (1-\pi) \cdot \beta \cdot (1-\beta)} \tag{6}$$

If $0 < \pi < 1$, then Equation 4 equals both Equation 5 and Equation 6 if and only if $\alpha = (1-\beta)$. If $\alpha \neq (1-\beta)$, therefore, the Bayesian researcher adjusts for multiplicity in the strong sense defined in Equation 3.[7]

□

Example 1 illustrates the commonsense idea that when one believes two hypotheses stand or fall together, evidence for/against one hypothesis is evidence for/against the other. Thus, a Bayesian researcher will adjust for multiplicity. Similarly, if the researcher believes evidence for one hypothesis is evidence *against* another, she will adjust for multiplicity, as can be shown by analogous calculations.

In short, if a researcher believes several hypotheses are dependent, she will typically adjust her beliefs for multiplicity. Conversely, if the researcher regards the hypotheses as mutually *independent* then she will *not* adjust for multiplicity; in that case, it is easy to check that $P(H_1|X_1) = P(H_1|X_1, \ldots X_N)$ (again, assuming Equation 1 holds).[8]

On one hand, these results about the relationship between adjustment and dependence in the toy Bayesian model above are not surprising. They illustrate the intuition that a researcher who wants to know what to believe about the effects of cigar smoking (i) will typically adjust her belief if she acquires data about the effects of cigarette smoking but (ii) will *not* adjust her beliefs if she acquires data about implicit bias.

On the other hand, the results begin to answer the central question. In particular, they answer the objection that there is no principled way to determine when to adjust [Perneger, 1998]. This objection is typically leveled

[7] Technically, our definition of adjustment compares the case in which the researcher learns X_1 to the case in which she learns the value of *all* N variables. The above equations show the researcher adjust when there are precisely $N = 2$ hypotheses under investigation, but similar calculations show adjustment is necessary when $N > 2$ as well.

[8] See [Berry and Hochberg, 1999, p. 218] for the calculation and discussion of the conditions under mutual independence of the hypotheses is reasonable. The hypotheses are mutually independent with respect to P if $P(\bigcap_{i \in I} \theta_i = r_i) = \Pi_{i \in I} P(\theta_i = r_i)$ for all $I \subseteq \{1, \ldots, N\}$ and all binary vectors $(r_i)_{i \in I}$.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

against classical methods – like Bonferroni's or Benjamini-Hochberg's – that recommend adjusting significance thresholds downward as the number of hypotheses increases. Yet the objection applies equally to a simple objective Bayesian method that I discuss below; the method adjusts for multiplicity by uniformly decreasing the prior probabilities assigned to hypotheses as the number of hypotheses grow.

According to critics, the justification of such methods implies that one should adjust/"correct" for any chosen set of hypotheses. But that's absurd because one would be required to adjust for *every statistical hypothesis that has ever been formulated*. This motivates thinking the answer to the central question is "One should never adjust for multiplicity, and intuitions to the contrary are misleading."

The toy results above show how simple Bayesian thinking can partially answer the objection. Prior evidence or background theory may tell us that certain hypotheses are dependent, and in such cases, belief adjustment will almost certainly be necessary. Further research should investigate whether the most common classical adjustment methods (see subsection 3.2) can ever be interpreted as reflecting belief adjustment.

One might object that the above definition of adjusting "belief" is too simple to model some common statistical practices. The problem is that the *same* probability measure P appears on both sides of Equation 3. So the definition is inapplicable for assessing whether "objective" Bayesian methods require adjustment.

Recall, objective Bayesians maintain that the prior probability that one assigns to hypothesis H may vary with the hypothesis space in which H is embedded. For example, consider an attempt to identify which genes are associated with which heritable diseases. For each gene and disease under investigation, researchers may investigate a hypothesis $H_{g,d}$ of the form "Gene g is associated with the disease d." In an objective Bayesian analysis, each hypothesis $H_{g,d}$ will typically receive lower prior probability if there are $20,000$ genes under investigation than it would receive if there were $10,000$ genes under consideration.

I will not compare the merits of objective versus subjective Bayesian analysis.[9] But simple objective Bayesian adjustment methods deserve further scrutiny. Imagine our hypothetical pharmaceutical researcher wonders about the effect of El Niño on the stock market. The mere contemplation of a new hypothesis should not automatically cause the researcher to become less confident in the efficacy of new cancer treatment.

[9] See [Goldstein, 2006] and [Berger, 2006] for opposing views.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Yet considering additional – logically independent — hypotheses can affect an objective Bayesian's prior probabilities if those probabilities are chosen in a mechanical fashion as a function of the number of hypotheses.

Objective Bayesians might respond that a prior distribution need not represent anyone's *beliefs*.[10] Rather, a prior should be treated as part of a *decision rule*. I agree, and I consider decision-making below. For now, note that it is similarly implausible that a pharmaceutical researcher should adjust her decisions about the efficacy of the cancer treatment after contemplating of El Niño. Saying the researcher's prior need not represent her beliefs does not explain why adjustment is not necessary.

3 Decision

Scientists are rarely satisfied with an answer to the question, "What should I believe?" They also want to know, "What should I do?" For instance, an experimentalist might want to know which experiment she should conduct next.

Imagine that, for each hypothesis H_k, there is some set of acts A_k that the researcher might take. For instance, a researcher might announce that the hypothesis H_k has been rejected or that it's been retained. She might collect more data about H_k or cease an experiment. And so on.

I call elements of A_k *component acts*, and I define a *strategy* to be a set S of component acts such that, for all k, either $S \cap A_k$ is a singleton or empty. That is, at most one act can be taken with respect to a hypothesis. A *decision rule* d maps subsets of (values of) the observable variables $X_1, \ldots X_N$ to strategies. I require that $d(X_{k_1} = x_{k_1}, \ldots, X_{k_m} = x_{k_m})$ contains precisely one element from each of the sets A_{k_m}. That requirement says that a decision rule specifies actions only with respect hypotheses for which the researcher has collected data, and that if researcher observes X_k, then she must take some action in A_k.

I say that a decision rule d adjusts for multiplicity if there is some x_1 such that

$$d(x_1) \notin d(x_1, \ldots x_N) \tag{7}$$

for all values $x_2, \ldots x_N$ of $X_2, \ldots X_N$.

Do any plausible decision rules require adjusting? Again, yes. For a Bayesian, reporting one's posterior probabilities is a decision. So belief ad-

[10] See [Gelman and Shalizi, 2013] for alternative interpretations of prior probabilities used in Bayesian analyses.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

justment is a special case of decision-adjustment. A better question is, "can there be decision-adjustment without belief adjustment, and what goals, if any, does decision-adjustment achieve?"

Before discussing the standard approach for evaluating testing procedures (in terms of the family-wise error rate or false discovery rate), I begin with the most naïve, decision-theoretic approach for answering these questions. The naïve approach is worth sketching because (1) it is, I think, the correct approach when it can be employed,[11] and (2) it helps one identify the oddness of the goals that are presumed in standard discussions of adjustment.

3.1 A Naïve Approach

Suppose a researcher assigns a utility $u(S, \theta)$ to each strategy S and vector $\theta \in \Theta$ specifying which of the N hypotheses are true. If we fix a vector $\theta \in \Theta$, then the researcher's expected utility (with respect to $\mathbb{P}_\theta$) can be defined straightforwardly, whether she decides to observe one variable or all N variables:[12]

$$\mathbb{E}^1_\theta[d] = \sum_{x_1 \in \mathcal{X}_1} \mathbb{P}_\theta(X_1 = x_1) \cdot u(d(x_1), \theta)$$
$$\mathbb{E}^N_\theta[d] = \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}_\theta(\vec{X} = \vec{x}) \cdot u(d(\vec{x}), \theta)$$

Here, $\mathcal{X}_1$ is the range of X_1 and $\mathcal{X}$ is the range of the random vector $\vec{X} = (X_1, \ldots, X_N)$. One can now apply standard decision-theoretic terms to identify different senses in which a decision rule is good or bad.

For instance, a researcher might desire a *maximin* decision rule, i.e., a rule d such that $\min_{\theta \in \theta} \mathbb{E}^j_\theta[d] \geq \min_{\theta \in \theta} \mathbb{E}^j_\theta[e]$ for all decision rules e, where $j = 1$ or $j = N$. Alternatively, she might be a Bayesian, i.e., she might always select a (subjective) expected utility maximizing strategy with respect to her posterior. Recall, the subjective expected utility of a strategy S with respect to a measure P is given by:

$$\mathbb{E}_P[S] := \sum_{\theta \in \Theta} P(\theta) \cdot u(S, \theta) \tag{8}$$

[11] See [Muller et al., 2006] for a defense of this decision-theoretic approach.

[12] For simplicity, I assume all of the sets in this paper are finite, including Θ, the ranges of the random variables $X_1, \ldots X_n$, and the range of all decision rules. Under appropriate measure-theoretic assumptions, the sums in the paper can be replaced with integrals if one is interested in extending these ideas to continuous spaces.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Thus, there is a Bayesian who will adjust for multiplicity if there is a probability measure P, utility function u, and experimental outcomes $\vec{x} = (x_1, \ldots, x_N) \in \mathcal{X}$ such that three conditions hold:

1. $P(\vec{X} = \vec{x}) > 0$,
2. a_1 maximizes $\mathbb{E}_{P(\cdot|X_1=x_1)}[a]$ over all $a \in A_1$, and
3. $a_1 \notin S$ for some S that maximizes $\mathbb{E}_{P(\cdot|\vec{X}=\vec{x})}[T]$, where T ranges over strategies containing a component act in every A_k.

We can now make the central question more precise in a second way. For which utility functions do standard non-probabilistic decision rules like maximin adjust for multiplicity in the sense of Equation 7? Similarly, for which priors and utility functions does an expected utility maximizer adjust for multiplicity?

For simplicity, assume that a decision-maker's utilities are *separable* across component acts in the following sense.[13] Assume that, for each hypothesis H_k, there is a "component" utility function $u_k : A_k \times \{0, 1\} \to \mathbb{R}$ that specifies the utilities $u(a, 0)$ and $u(a, 1)$ of taking action $a \in A_k$ when H_k is true and false respectively. Further, suppose that the utility of a strategy $u(S, \theta)$ in state θ is the sum of the utilities of component acts, i.e.,

$$u(S, \theta) = \sum_{k \leq N} \sum_{a \in S \cap A_k} u_k(a, \theta_k) \tag{9}$$

Utilities are separable when (a) the decision-maker can take component acts in parallel and (b) payoffs for taking different component acts do not interact. Such assumptions are most plausible when two conditions are met. First, acts are cheap or the decision-maker has plentiful resources (and so pursuing multiple projects in parallel is not prohibitively costly). Second, the hypotheses concern unrelated phenomena (so that the important theoretical consequences of a conjunction of hypotheses is the union of the theoretical consequences of the conjuncts). If the decision-maker is a grant-making institution like the NSF or NIH, then utilities associated with projects in different scientific fields are plausibly separable. The size of the institution makes funding projects in parallel possible, and it is rare to find results in two disparate scientific fields that, when taken together, yield important insights that neither result yields by itself.

The next theorem suggests that, when utilities are separable, adjustment is never obligatory, and it is sometimes impermissible.[14]

[13] See [Cohen and Sackrowitz, 2005] for a similar assumption.

[14] See online supplemental materials for a proof.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Theorem 1. *Suppose utilities are separable in the sense of Equation 9. Then there are maximin rules that do not adjust for multiplicity. If in addition the hypotheses of* Θ *as mutually independent with respect to* P*, then one can maximize (subjective) expected utility with respect to* P *without adjusting. It follows that if the maximin rule is unique, then no decision rule that adjusts is maximin. Similar remarks apply to expected utility maximization.*

One might object that individual scientists will rarely have separable utilities for the reasons identified above. Component acts are often costly: pursuing one project typically comes at the expense of pursuing another. And even if the component acts are cheap (e.g., making an announcement), it is rare that scientists investigate hypotheses that are so unrelated that, if the conjunction were true, no further important insights would follow. Scientists are highly specialized, and thus, they typically study hypotheses that are related.

However, I have not identified necessary conditions for separability; utility functions might be (approximately) separable for other reasons. More importantly, Theorem 1 yields sufficient conditions for non-adjustment, not necessary ones. So a suspicion that Theorem 1 is rarely applicable does not justify decision adjustment for individual researchers. The theorem shifts the burden to providing a positive argument for adjustment.

The reader might speculate that, given the extensive research on multiplicity, statisticians have (i) identified utility functions that plausibly represent the interests of scientists and (ii) shown that common adjustment procedures are uniquely maximin, or expected utility maximizing with respect to those utility functions. Unfortunately, that's not the case. Some classical procedures for multiple testing are, in fact, *inadmissible* (i.e., weakly dominated) for plausible utility/loss functions.[15] Thus, the criteria used to justify standard classical testing procedures is more complex than it might initially seem; I turn to those criteria now.

3.2 Family-Wise Error Rates and False Discovery Rates

Classical approaches to multiple testing typically aim to control either the *family-wise error rate* (FWER) – which is the probability that a series of tests yields at least one false positive – or the *false discovery rate* (FDR) – which is the expected *proportion* of rejected null hypotheses that are true.

Statisticians routinely say that the FWER is rarely of interest. I agree. The FWER is almost always maximized when all null hypotheses are false.

[15] Again, see [Cohen and Sackrowitz, 2005].

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

But in many applications, researchers know that at least one null hypothesis is false. Consider again genome-wide association studies that investigate the associations between thousands of genes and multiple heritable diseases. If at least one disease is known to be heritable and genes are the mechanism for inheritance, then there must be at least one gene that is associated with at least one disease!

Thus, some researchers now insist that multiple testing regimes should control the FDR. If the FDR is identical to one's loss function, are existing regimes maximin? Do they ever minimize subjective expected loss? The answer to both questions is clearly "no." One minimizes the FDR (or FWER) by retaining all null hypotheses. Thus, as is standard in classical hypothesis testing, existing multiple-testing procedures typically (i) fix a threshold for FDR and (ii) attempt to maximize power (i.e., the probability of a false negative) subject to the constraint that the FDR is below the threshold. Assuming utility is identified with (some kind of) power, statisticians have identified testing regimes that are maximin among the set of procedures that maintain FDR and/or FWER below a threshold.[16]

I will not rehearse standard objections to maximin reasoning,[17] nor to the bizarre two-step procedure in which one first culls testing procedures using FDR and then applies maximin. Instead, I emphasize that the decision criteria just described (1) treat all *null* hypotheses equally, (2) treat null hypotheses *differently* from alternatives, and (3) ignore effect sizes. However, there are virtually no circumstances in which such equal treatment and dismissal of effect size reflects either scientific or public interest.

Consider a recent influential genome-wide study in which researchers tested roughly 14,000 genes for associations with seven common diseases, which included bipolar disorder and Crohn's disease [Consortium, 2007]. Although the authors of the study reported adjusted p-values, they also laudably applied many statistical techniques, incorporated background genetic knowledge, and avoided making policy recommendations based solely on adjusted p-values. Why did they not simply apply a testing procedure with good power subject to control of FDR?

All seven diseases they considered are serious, but the incidence varies widely, as does the cost and efficacy of available treatments. From a public health perspective, therefore, it would be inappropriate to treat every hy-

[16] Just as there are multiple notions of "Type I error" when many hypotheses are tested (e.g., FWER or FDR), so there are multiple notions of "power" that might be invoked, such as the probability of *at least one* false negative, the "average" power, and more. For a discussion and proof of optimality of certain classical procedures, see [Rosset et al., 2022].

[17] See [Savage, 1954, Chapters 9-10], for example.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

pothesis of the form "Gene g is associated with disease d" equally and to ignore the strength of such associations.

One might object that the severity of the diseases does not affect the *evidence* for the various hypotheses. Does adjustment somehow reflect one's evidence?

Answering that question is beyond the scope of this paper; I lack the space to explore the relationship among evidence, belief, and decision.[18] But I am skeptical of both (a) the importance of the question and (b) an answer that involves classical procedures that control FDR or FWER.

Concerning (a), philosophers and scientists alike should be wary of directives to ignore the suffering caused by diseases and instead coldly evaluate only the evidence for empirical hypotheses. I admit that a subjective expected utility analysis of genome wide studies seems daunting. I have no idea how to define a prior over a roughly 100,000 (i.e., approximately $7 \cdot 14,000$) dimensional parameter space that incorporates expert knowledge. Nor do I have any idea how to define a utility function that balances considerations of the severity and incidence of different diseases. But I stress that mechanical use of multiple testing procedures amounts to a refusal to engage with questions of ethical importance, not an answer.

Concerning (b), like many classical procedures, decision criteria that first cull tests by FWER or FDR treat null hypotheses differently from the alternatives. But if evidential strength is divorced from pragmatic and ethical considerations, it is hard to see how the asymmetric treatment of null and alternative hypotheses could reflect anything evidential: what could distinguish a hypothesis H from its negation $\neg H$, evidentially speaking?

4 Conclusions

The goals of scientists and of the public may be misaligned with the decision criteria used to evaluate multiple testing regimes. Thus, I urge two broad projects for future research.

First, in scientific contexts in which large numbers of statistical hypotheses can be tested, scientists and philosophers must study the interests of the affected parties. The differential funding provided for medical research – in comparison to academic philosophy, for instance – is typically justified by its social importance. Scientists should make good on that promise to advance collective interests.[19]

[18] [Royall, 1997] clearly distinguishes questions about belief, decision, and evidence.

[19] See also [Longino, 1990], [Kitcher, 2003], and [Douglas, 2009].

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Second, statisticians must prove existing testing procedures advance the interests of affected parties, or they must develop alternative procedures altogether. Otherwise, we all stand to be bamboozled by Bonferroni.

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. doi: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. Publisher: Wiley Online Library.

James Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006. doi: https://doi.org/10.1214/06-BA115.

Donald A. Berry and Yosef Hochberg. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1-2):215–227, 1999. doi: https://doi.org/10.1016/S0378-3758(99)00044-0. Publisher: Elsevier.

Arthur Cohen and Harold B. Sackrowitz. Characterization of Bayes procedures for multiple endpoint problems and inadmissibility of the step-up procedure. *The Annals of Statistics*, 33(1):145–158, 2005. doi: https://doi.org/10.1214/009053604000000986.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007. doi: https://doi.org/10.1038/nature05911.

Arthur P. Dempster. Upper and lower probabilities generated by a random closed interval. *The Annals of Mathematical Statistics*, 39(3):957–966, 1968.

Heather Douglas. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, Pittsburgh (Pa.), first edition edition, July 2009.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013. doi: https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Michael Goldstein. Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, 1(3):403–420, 2006. doi: https://doi.org/10.1214/06-BA116.

Philip Kitcher. *Science, Truth, and Democracy*. Oxford University Press, Oxford, September 2003.

Matthew Kotzen. Multiple Studies and Evidential Defeat. *Nous*, 47(1): 154–180, 2013. doi: https://doi.org/10.1111/j.1468-0068.2010.00824.x.

Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, third edition, 2008.

Helen E. Longino. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press, 1990.

Deborah G. Mayo. *Statistical inference as severe testing*. Cambridge: Cambridge University Press, 2018.

Conor Mayo-Wilson and Aditya Saraf. Robust Bayesianism and Likelihoodism. https://doi.org/10.48550/arXiv.2009.03879, 2022. URL https://arxiv.org/abs/2009.03879.

Peter Muller, Giovanni Parmigiani, and Kenneth Rice. FDR and Bayesian multiple comparisons rules. Alicante, Spain, June 2006. URL https://biostats.bepress.com/jhubiostat/paper115/.

Thomas V. Perneger. What's wrong with Bonferroni adjustments. *BMJ: British Medical Journal*, 316(7139):1236–1238, 1998. doi: https://doi.org/10.1136/bmj.316.7139.1236.

Saharon Rosset, Ruth Heller, Amichai Painsky, and Ehud Aharoni. Optimal and maximin procedures for multiple testing problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1105–1128, 2022. doi: https://doi.org/10.1111/rssb.12507. Publisher: Oxford University Press.

Richard Royall. *Statistical evidence: a likelihood paradigm*, volume 71 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 1 edition, 1997.

Mark Rubin. When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, 199(3-4):10969–11000, 2021. doi: https://doi.org/10.1007/s11229-021-03276-4. Publisher: Springer.

Leonard J. Savage. *The foundation of statistics*. Dover publications, 1954.

James G. Scott and James O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of statistical planning and inference*, 136(7): 2144–2162, 2006. doi: https://doi.org/10.1016/j.jspi.2005.08.031. Publisher: Elsevier.

Wolfgang Spohn. *The laws of belief: Ranking theory and its philosophical applications*. Oxford University Press, 2012.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Online Appendix to Bamboozled by Bonferroni

Conor Mayo-Wilson

In this document, we prove Theorem 1 in the body of the paper. To ensure this online appendix is self-contained, some definitions appearing in the body of the paper are copied below.

1 Basic Model

Suppose N hypotheses are under investigation, and let $\Theta = \{0,1\}^N$ be the set of all binary strings of length N. A vector $\theta \in \Theta$ specifies which of the N hypotheses are true. For each $k \leq N$, let $H_k = \{\theta \in \Theta : \theta_k = 0\}$ be the set of vectors that say the k^{th} hypothesis is true. For each $k \leq N$, let X_k be a random variable representing an experiment. For each $\theta \in \Theta$, let $\mathbb{P}_\theta(X_1, \ldots, X_N)$ denote the probability measure that specifies the chances of various experimental outcomes.

We assume that, for all $\theta \in \Theta$, the N experiments are *mutually independent* with respect to $\mathbb{P}_\theta$. In symbols, let $\vec{X} = \langle X_{i_1}, X_{i_2}, \ldots, X_{i_k} \rangle$ be a random vector, representing some subset of the N experiments. Then:

$$\mathbb{P}_\theta(\vec{X} = \vec{x}) = \prod_{j \leq k} \mathbb{P}_\theta(X_{i_j} = x_{i_j}) \tag{1}$$

for all $\vec{x} = (x_{i_1}, \ldots x_{i_k})$. Further, suppose that the truth or falsity of the H_k determines the probabilities of the possible outcomes of the k^{th} experiment, i.e., for all $k \leq N$ and all $r \in \{0,1\}$, there is a probability distribution $\mathbb{P}_{k,r}$ such that $\mathbb{P}_\theta(X_k = x_k) = \mathbb{P}_{k,\theta_k}(X_k = x_k)$. Together with the assumption of mutual independence, this entails that:

$$\mathbb{P}_\theta(\vec{X} = \vec{x}) = \prod_{j \leq k} \mathbb{P}_{i_j,\theta_{i_j}}(X_{i_j} = x_{i_j}) \text{ for all } \theta \in \Theta. \tag{2}$$

1.1 Decision Adjustment

For each $k \leq n$, let A_k denote a set of *component acts*, and define a *strategy* to be a set S of component acts such that, for all k, either $S \cap A_k$ is a

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

singleton or empty. That is, at most one act can be taken with respect to a hypothesis H_k. A *decision rule* d maps subsets of (values of) the observable variables $X_1, \ldots X_N$ to strategies. I require that $d(X_{k_1} = x_{k_1}, \ldots, X_{k_m} = x_{k_m})$ contains precisely one element from each of the sets A_{k_m}.

A decision rule d adjusts for multiplicity if there is some x_1 such that

$$d(x_1) \notin d(x_1, \ldots x_N) \tag{3}$$

for all values $x_2, \ldots x_N$ of $X_2, \ldots X_N$.

1.2 Maximin and Bayes Rules

Suppose a researcher assigns a utility $u(S, \theta)$ to each strategy S and vector $\theta \in \Theta$ specifying which of the N hypotheses are true. If we fix a vector $\theta \in \Theta$, then the researcher's expected utility (with respect to $\mathbb{P}_\theta$) can be defined straightforwardly, whether she decides to observe one variable or all N variables:[1]

$$\mathbb{E}^1_\theta[d] = \sum_{x_1 \in \mathcal{X}_1} \mathbb{P}_\theta(X_1 = x_1) \cdot u(d(x_1), \theta)$$
$$\mathbb{E}^N_\theta[d] = \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}_\theta(\vec{X} = \vec{x}) \cdot u(d(\vec{x}), \theta)$$

Here, $\mathcal{X}_1$ is the range of X_1 and $\mathcal{X}$ is the range of the random vector $\vec{X} = (X_1, \ldots, X_N)$.

A decision rule d is called *maximin* if $\min_{\theta \in \theta} \mathbb{E}^j_\theta[d] \geq \min_{\theta \in \theta} \mathbb{E}^j_\theta[e]$ for all decision rules e, where $j = 1$ or $j = N$.

Recall, the subjective expected utility of a strategy S with respect to a measure P is given by:

$$\mathbb{E}_P[S] := \sum_{\theta \in \Theta} P(\theta) \cdot u(S, \theta) \tag{4}$$

Thus, there is a Bayesian who will adjust for multiplicity if there is a probability measure P, utility function u, and experimental outcomes $\vec{x} = (x_1, \ldots, x_N) \in \mathcal{X}$ such that three conditions hold:

1. $P(\vec{X} = \vec{x}) > 0$,

[1] For simplicity, I assume all of the sets in this paper are finite, including Θ, the ranges of the random variables $X_1, \ldots X_n$, and the range of all decision rules. Under appropriate measure-theoretic assumptions, the sums in the paper can be replaced with integrals if one is interested in extending these ideas to continuous spaces.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

2. a_1 maximizes $\mathbb{E}_{P(\cdot|X_1=x_1)}[a]$ over all $a \in A_1$, and

3. $a_1 \notin S$ for some S that maximizes $\mathbb{E}_{P(\cdot|\vec{X}=\vec{x})}[T]$, where T ranges over strategies containing a component act in every A_k.

For simplicity, assume that a decision-maker's utilities are *separable* across component acts in the following sense. Assume that, for each hypothesis H_k, there is a "component" utility function $u_k : A_k \times \{0,1\} \to \mathbb{R}$ that specifies the utilities $u(a,0)$ and $u(a,1)$ of taking action $a \in A_k$ when H_k is true and false respectively. Further, suppose that the utility of a strategy $u(S,\theta)$ in state θ is the sum of the utilities of component acts, i.e.,

$$u(S,\theta) = \sum_{k \leq N} \sum_{a \in S \cap A_k} u_k(a, \theta_k) \tag{5}$$

2 Theorem and Proof

Theorem 1. *Suppose utilities are separable in the sense of Equation 5. Then there are maximin rules that do not adjust for multiplicity. If in addition the hypotheses of Θ as mutually independent with respect to P, then one can maximize (subjective) expected utility with respect to P without adjusting. It follows that if the maximin rule is unique, then no decision rule that adjusts is maximin. Similar remarks apply to expected utility maximization.*

Before proving the theorem, we introduce some notation. Given any decision rule d and $k \leq N$, we define a function $d_k : \mathcal{X} \to A_k$ by $d_k(\vec{y}) := A_k \cap d(\vec{y})$. In other words, d_k picks out the k^{th} component act from each strategy recommended by d.

$$\begin{aligned}
\mathbb{E}^N_\theta[d] &= \sum_{\vec{y} \in \mathcal{X}} \mathbb{P}_\theta(\vec{y}) \cdot u(d(\vec{y}), \theta) \\
&= (\mathbb{P}_\theta(\vec{x}) \cdot u(d(\vec{x}), \theta)) + \left(\sum_{\vec{y} \neq \vec{x}} \mathbb{P}_\theta(\vec{y}) \cdot u(d(\vec{y}), \theta) \right) \\
&= \left(\sum_{1 \leq k \leq N} \mathbb{P}_\theta(\vec{x}) \cdot u_k(d_k(\vec{x}), \theta) \right) + \left(\sum_{\vec{y} \neq \vec{x}} \mathbb{P}_\theta(\vec{y}) \cdot u(d(\vec{y}), \theta) \right) \\
&\quad \text{by separability} \\
&= (\mathbb{P}_\theta(\vec{x}) \cdot u_1(d_1(\vec{x}), \theta)) + \left(\sum_{1 < k \leq N} \mathbb{P}_\theta(\vec{x}) \cdot u_k(d_k(\vec{x}), \theta) \right) + \left(\sum_{\vec{y} \neq \vec{x}} \mathbb{P}_\theta(\vec{y}) \cdot u(d(\vec{y}), \theta) \right)
\end{aligned}$$

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Call the first, second, and third summands in the previous equation $T_1(\theta, \vec{x}, d)$, $T_2(\theta, \vec{x}, d)$, and $T_3(\theta, \vec{x}, d)$ respectively.

Proof of Theorem 1: The outline of the proof is identical for both maximin and subjective expected utility (SEU) maximization. We first pick any decision rule d that is maximin (or maximizes SEU). Such a rule exists because we have assumed all the relevant sets to be finite. If d does *not* adjust for multiplicity, we're done. Otherwise, there is some vector $\vec{x} = (x_1, \ldots x_N)$ such that $d(x_1) \notin d(\vec{x})$. Define a new decision rule – call it e – such that e is alike d in all respects except the following. Let $a_1 \in A_1$ be such that $d(x_1) = \{a_1\}$, and let b_1 be the unique element of $A_1 \cap d(\vec{x})$. Define $e(\vec{x}) = (d(\vec{x}) \setminus b_1) \cup \{a_1\}$. And as we said, define $e(\vec{y}) = d(\vec{y})$ for all $\vec{y} \neq \vec{x}$ (regardless of length). We claim that e is also maximin (or maximizes SEU). By repeating this process some finite number of times, we'll obtain a decision rule that is maximin (or maximizes SEU) and that does not adjust for multiplicity.

First, we consider the case in which d is maximin. Because d itself is maximin, to show that e is maximin, it suffices to show that:

$$\min_{\theta \in \Theta} \mathbb{E}^1_\theta[e] \geq \min_{\theta \in \Theta} \mathbb{E}^1_\theta[d] \text{ and} \tag{6}$$

$$\min_{\theta \in \Theta} \mathbb{E}^N_\theta[e] \geq \min_{\theta \in \Theta} \mathbb{E}^N_\theta[d] \tag{7}$$

The first equation follows immediately from the definition of e since $e(x) = d(x)$ for all $x \in \mathcal{X}_1$, i.e., the values of e and d do not differ on vectors of length 1. So we need to show only that $\min_{\theta \in \Theta} \mathbb{E}^N_\theta[e] \geq \min_{\theta \in \Theta} \mathbb{E}^N_\theta[d]$.

Using the decomposition described above, we first show that $T_2(\theta, \vec{x}, d) = T_2(\theta, \vec{x}, e)$ and that $T_3(\theta, \vec{x}, d) = T_3(\theta, \vec{x}, e)$ for all θ and $\vec{x}$.

To show $T_2(\theta, \vec{x}, d) = T_2(\theta, \vec{x}, e)$ for all θ, let θ be arbitrary. Notice first that, by the definition of e, we know that $d_k(\vec{y}) = e_k(\vec{y})$ for all $k > 1$ and for all $\vec{y}$ (including $\vec{x}$). It follows that for all θ and all $\vec{y}$:

$$\sum_{1 < k \leq N} \mathbb{P}_\theta(\vec{y}) \cdot u_k(d_k(\vec{y}), \theta) = \sum_{1 < k \leq N} \mathbb{P}_\theta(\vec{y}) \cdot u_k(e_k(\vec{y}), \theta) \tag{8}$$

which is exactly what $T_2(\theta, \vec{x}, d) = T_2(\theta, \vec{x}, e)$ asserts.

To show $T_3(\theta, \vec{x}, d) = T_3(\theta, \vec{x}, e)$, again note that by definition of e, we know that $d_1(\vec{y}) = e_1(\vec{y})$ for all $\vec{y} \neq \vec{x}$. It follows that:

$$\mathbb{P}_\theta(\vec{y}) \cdot u(d_1(\vec{y}), \theta) = \mathbb{P}_\theta(\vec{y}) \cdot u(e_1(\vec{y}), \theta) \text{ for all } \theta \text{ and all } \vec{y} \neq \vec{x}. \tag{9}$$

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Equation 9 and Equation 8 together entail

$$\sum_{1\leq k\leq n} \mathbb{P}_\theta(\vec{y}) \cdot u_k(d_k(\vec{y}), \theta) = \sum_{1\leq k\leq n} \mathbb{P}_\theta(\vec{y}) \cdot u_k(e_k(\vec{y}), \theta) \text{ for all } \theta \text{ and } \vec{y} \neq \vec{x} \tag{10}$$

Because u is separable, Equation 10 implies that for all $\vec{y} \neq \vec{x}$

$$\mathbb{P}_\theta(\vec{y}) \cdot u(d(\vec{y}), \theta) = \mathbb{P}_\theta(\vec{y}) \cdot u(e(\vec{y}), \theta) \text{ for all } \theta \text{ and } \vec{y} \neq \vec{x} \tag{11}$$

And that immediately entails:

$$\sum_{\vec{y}\neq\vec{x}} \mathbb{P}_\theta(\vec{y}) \cdot u(d(\vec{y}), \theta) = \sum_{\vec{y}\neq\vec{x}} \mathbb{P}_\theta(\vec{y}) \cdot u(e(\vec{y}), \theta) \text{ for all } \theta \text{ and } \vec{y} \neq \vec{x} \tag{12}$$

Notice the previous equation asserts $T_3(\theta, \vec{x}, d) = T_3(\theta, \vec{x}, e)$, as desired.

So to show that e is maximin, it therefore suffices to show that $\min_{\theta\in\Theta} T_1(\theta, \vec{x}, e) \geq \min_{\theta\in\Theta} T_1(\theta, \vec{x}, d)$, where recall:

$$T_1(\theta, \vec{x}, e) = \mathbb{P}_\theta(\vec{x}) \cdot u_1(e_1(\vec{x}), \theta)$$

and similarly for $T_1(\theta, \vec{x}, d)$.

For the sake of contradiction, suppose that

$$\min_{\theta\in\Theta} \mathbb{P}_\theta(\vec{x}) \cdot u_1(e_1(\vec{x}), \theta) < \min_{\theta\in\Theta} \mathbb{P}_\theta(\vec{x}) \cdot u_1(d_1(\vec{x}), \theta) \tag{13}$$

Because the likelihood function factors (by Equation 1), it follows that

$$\min_{\theta\in\Theta} \left(\mathbb{P}_\theta(x_1) \cdot \prod_{k\geq 2} \mathbb{P}_\theta(x_k) \right) \cdot u_1(e_1(\vec{x}), \theta) < \min_{\theta\in\Theta} \left(\mathbb{P}_\theta(x_1) \cdot \prod_{k\geq 2} \mathbb{P}_\theta(x_k) \right) \cdot u_1(d_1(\vec{x}), \theta)$$

That inequality cannot be strict unless $\prod_{k\geq 2} \mathbb{P}_\theta(x_k) > 0$ for at least one θ.
It follows that:

$$\min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u_1(e_1(\vec{x}), \theta) < \min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u_1(d_1(\vec{x}), \theta)$$

Recall, $d_1(\vec{x}) = \{b_1\}$, and so the last equation becomes:

$$\min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u_1(e_1(\vec{x}), \theta) < \min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u_1(b_1, \theta)$$

By separability, the previous equation entails:

$$\min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u(e(x_1), \theta) < \min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u(\{b_1\}, \theta)$$

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

And since $e(x_1) = d(x_1)$, we obtain that:

$$\min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u(d(x_1), \theta) < \min_{\theta\in\Theta} \mathbb{P}_\theta(x_1) \cdot u(\{b_1\}, \theta)$$

Now if we add $\sum_{y\in\mathcal{X}_1\setminus\{x_1\}} \mathbb{P}_\theta(y) \cdot u(d(y), \theta)$ under the minimum on both sides of the equation, we get:

$$\min_{\theta\in\Theta} \left(\sum_{y\in\mathcal{X}_1\setminus\{x_1\}} \mathbb{P}_\theta(y) \cdot u(d(y), \theta) \right) + \mathbb{P}_\theta(x_1) \cdot u(d(x_1), \theta) <$$
$$\min_{\theta\in\Theta} \left(\sum_{y\in\mathcal{X}_1\setminus\{x_1\}} \mathbb{P}_\theta(y) \cdot u(d(y), \theta) \right) + \mathbb{P}_\theta(x_1) \cdot u(\{b_1\}, \theta)$$

The left-hand side of that inequality is $\min_{\theta\in\Theta} \mathbb{E}^1_\theta[d]$. And if we let f be the decision rule that is exactly alike d except $f(x_1) = \{b_1\}$, then the right-hand side is $\min_{\theta\in\Theta} \mathbb{E}^1_\theta[f]$. So we've shown:

$$\min_{\theta\in\Theta} \mathbb{E}^1_\theta[d] < \min_{\theta\ \in\Theta} \mathbb{E}^1_\theta[f]$$

which contradicts the assumption that d is maximin. That finishes the proof of the claim about maximin.

Next we prove the claim about expected utility maximization. Suppose that (I) d adjusts for multiplicity maximizes SEU with respect to the probability measure P and (II) that the hypotheses (i.e., members of Θ) are mutually independent with respect to P. To say that d maximizes SEU with respect to P means that

1. $\mathbb{E}_{P(\cdot|X_1=y)}[d(y)] \geq \mathbb{E}_{P(\cdot|X_1=y)}[a_1]$ for all $a_1 \in A_1$ and all $y \in \mathcal{X}_1$, and
2. $\mathbb{E}_{P(\cdot|\vec{X}=\vec{y})}[d(\vec{y})] \geq \mathbb{E}_{P(\cdot|\vec{X}=\vec{y})}[S]$ for all for all strategies $S \subset \bigcup_{k\leq N} A_k$ and all $\vec{y} \in \mathcal{X}$.

As above, let $\vec{x}$ be the vector witnessing the fact that d adjusts for multiplicity, and define a decision rule e as in the first half of the proof.

Because $e(y) = d(y)$ for all $y \in \mathcal{X}_1$, it follows immediately that $e(y)$ maximizes SEU with respect to $P(\cdot|X_1 = y)$ for all $y \in \mathcal{X}_1$ (because $d(y)$ is a maximizer!).

So it remains to be shown that $e(\vec{y})$ maximizes SEU with respect to $P(\cdot|\vec{X} = \vec{y})$ for all $\vec{y} \in \mathcal{X}$. Because $e(\vec{y}) = d(\vec{y})$ for all $\vec{y} \neq \vec{x}$ and because d is an SEU maximizer, it suffices to show that

$$\mathbb{E}_{P(\cdot|\vec{X}]=\vec{x})}[e(\vec{x})] \geq \mathbb{E}_{P(\cdot|\vec{X}=\vec{x})}[d(\vec{x})]$$

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

To show that, notice we can decompose $\mathbb{E}_{P(\cdot|\vec{X}=\vec{x})}[e(\vec{x})]$ as follows:

$$\begin{aligned}\mathbb{E}_{P(\cdot|\vec{X}=\vec{x})}[e(\vec{x})] &= \sum_{\theta\in\Theta} P(\theta|\vec{X}=\vec{x})\cdot u(e(\vec{x}),\theta)\\ &= \sum_{\theta\in\Theta}\sum_{k\leq N} P(\theta|\vec{X}=\vec{x})\cdot u_k(e_k(\vec{x}),\theta_k) \quad \text{by separability}\\ &= \sum_{\theta\in\Theta} P(\theta|\vec{X}=\vec{x})\cdot u_1(e_1(\vec{x}),\theta_1) + \sum_{\theta\in\Theta}\sum_{1<k\leq N} P(\theta|\vec{X}=\vec{x})\cdot u_k(e_k(\vec{x}),\theta_k)\end{aligned}$$

Now notice that because $e_k(\vec{y}) = d_k(\vec{y})$ for all $k > 1$, the second summand above – that is, the double sum – is equal to the same term in which d_k is substituted for e_k. So it suffices to show that

$$\sum_{\theta\in\Theta} P(\theta|\vec{X}=\vec{x})\cdot u_1(e_1(\vec{x}),\theta_1) \geq \sum_{\theta\in\Theta} P(\theta|\vec{X}=\vec{x})\cdot u_1(d_1(\vec{x}),\theta_1) \tag{14}$$

By Bayes rule and our assumptions about mutual independence of the hypotheses (and of the random variables), we have that for all θ:

$$\begin{aligned}P(\theta|\vec{X}=\vec{x}) &= \frac{\mathbb{P}_\theta(\vec{X}=\vec{x})\cdot P(\theta)}{P(\vec{X}=\vec{x})}\\ &= \frac{\prod_{k\leq N}\mathbb{P}_{\theta_k}(X_k=x_k)\cdot P(\theta_k)}{P(\vec{X}=\vec{x})}\\ &= \frac{\prod_{k\leq N} P(X_k=x_k|\theta_k)\cdot P(\theta_k)}{P(\vec{X}=\vec{x})}\\ &= \frac{\prod_{k\leq N} P(\theta_k|X_k=x_k)\cdot P(X_k=x_k)}{P(\vec{X}=\vec{x})}\\ &= \frac{\prod_{k\leq N} P(\theta_k|X_k=x_k)\cdot\prod_{k\leq N} P(X_k=x_k)}{P(\vec{X}=\vec{x})}\\ &= \frac{\prod_{k\leq N} P(X_k=x_k)}{P(\vec{X}=\vec{x})}\cdot\prod_{k\leq N} P(\theta_k|X_k=x_k)\end{aligned}$$

It follows that Equation 14 holds if and only if:

$$\sum_{\theta\in\Theta}\prod_{k\leq N} P(\theta_k|X_k=x_k)\cdot u_1(e_1(\vec{x}),\theta_1) \geq \sum_{\theta\in\Theta}\prod_{k\leq N} P(\theta_k|X_k=x_k)\cdot u_1(d_1(\vec{x}),\theta_1) \tag{15}$$

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

Recall that $e_1(\vec{x}) = d(x_1)$ by construction, and so the last inequality holds if and only if

$$\sum_{\theta \in \Theta} \prod_{k \leq N} P(\theta_k | X_k = x_k) \cdot u(d(x_1), \theta_1) \geq \sum_{\theta \in \Theta} \prod_{k \leq N} P(\theta_k | X_k = x_k) \cdot u_1(b_1, \theta_1) \tag{16}$$

Now rewrite the term on the left-hand-side of Equation 16. To do so, perform the outside sum in two steps, by first summing over values of θ_1 and then by summing over the values of $\theta_2, \ldots, \theta_N$. In other words, observe that we can rewrite the left-hand-side of the equation as follows:

$$\begin{aligned}
&\sum_{\theta \in \Theta} \prod_{k \leq N} P(\theta_k | X_k = x_k) \cdot u_1(d(x_1), \theta_1) \\
&= \sum_{\theta_1} \sum_{\theta_2, \ldots \theta_N} \prod_{k \leq N} P(\theta_k | X_k = x_k) \cdot u_1(d(x_1), \theta_1) \\
&= \sum_{\theta_1} \sum_{\theta_2, \ldots \theta_N} (P(\theta_1 | X_1 = x_1) \cdot u_1(d(x_1), \theta_1)) \cdot \left(\prod_{1 < k \leq N} P(\theta_k | X_k = x_k) \right) \\
&= \sum_{\theta_2, \ldots \theta_N} \sum_{\theta_1} (P(\theta_1 | X_1 = x_1) \cdot u_1(d(x_1), \theta_1)) \cdot \left(\prod_{1 < k \leq N} P(\theta_k | X_k = x_k) \right) \\
&\text{by reordering the sums} \\
&= \sum_{\theta_2, \ldots \theta_N} \left(\prod_{1 < k \leq N} P(\theta_k | X_k = x_k) \cdot \left(\sum_{\theta_1} P(\theta_1 | X_1 = x_1) \cdot u_1(d(x_1), \theta_1) \right) \right) \\
&= \sum_{\theta_2, \ldots \theta_N} \left(\prod_{1 < k \leq N} P(\theta_k | X_k = x_k) \cdot \left(\sum_{\upsilon \in \Theta : \upsilon_1 = \theta_1} P(\upsilon | X_1 = x_1) \cdot u(d(x_1), \upsilon) \right) \right) \\
&\text{as } u(d(x_1), \theta_1)) = u_1(d(x_1), \upsilon) \text{ if } \upsilon_1 = \theta_1 \text{ by separability} \\
&= \sum_{\theta_2, \ldots \theta_N} \prod_{1 < k \leq N} P(\theta_k | X_k = x_k) \cdot \mathbb{E}_{P(\cdot | X_1 = x_1)}[d(x_1)] \\
&= \mathbb{E}_{P(\cdot | X_1 = x_1)}[d(x_1)] \cdot \sum_{\theta_2, \ldots \theta_N} \prod_{1 < k \leq N} P(\theta_k | X_k = x_k)
\end{aligned}$$

□

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press