The New England Journal of Statistics in Data Science Volume 1, 126–137 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS4

Four Types of Frequentism and Their Interplay with Bayesianism

James BERGER

1. INTRODUCTION

The majority of statisticians and data scientists declare themselves to be frequentists, but they often mean very different things by this declaration. Indeed, while I.J. Good identified 46,656 potential types of Bayesians [13], there may be even more potential types of frequentists. This paper restricts attention to the four most common types of frequentism, discussed in Section 2.

The paper has several goals:

- Highlight and compare the major different types of frequentistism.
- Relate the different types of frequentistism with Bayesianism; some types are more compatible with Bayesianism than others. The focus of these discussions is in determining which types of frequentism are most useful to Bayesians. These discussions are given at the end of each subsection.
- Evaluate a number of common statistical scenarios from the different frequentist perspectives. This results in some (perhaps) surprising findings in common situations such as multiple hypothesis testing and sequential endpoint testing (Section 3.2).
- Evaluate certain Bayesian procedures, such as the use of odds in testing (Section 3.4), from these frequentist perspectives.

The focus here is not on studying general approaches to statistical analysis; we consider specific examples of statistical analysis to illustrate the issues, but do not focus on general theories. For instance, there has recently been great interest (generated, in part, by the Bayesian, Fiducial & Frequentist (BFF) series of meetings) in developing Confidence Distribution analysis (cf. [26]) and Generalized Fiducial analysis (cf. [14]), but application of these methods to specific contexts could result in different types of frequentism being utilized.

Caveat 1. Most of the concepts in the paper have been extensively discussed over hundreds of year. We do not attempt to trace this history; instead we have only the pedagogical goal of trying to clarify the concepts that have emerged. The clarification is most easily done with simple examples; indeed, all examples in the paper only consider one-dimensional parameters.

Caveat 2. The word frequentist is traditionally viewed as referring to some type of long-run average (long-run frequency), and we restrict consideration in this paper to only that notion. Many people today also use the word frequentist to refer to what are essentially Fisherian concepts [12] that do not necessarily involve a long-run average. A recent example is [18, 19] whose interesting statistical philosophy is based on a mix of Fisherian and long-run average concepts.

2. FOUR TYPES OF FREQUENTISM

The four types of frequentism that we address are each defined and illustrated (through numerous examples) in a subsection herein. Each subsection includes a discussion of the relationship of the corresponding principle with Bayesianism.

2.1 Type I. Empirical Frequentism

Empirical frequentist principle. *In repeated practical use of a statistical procedure, the long-run average actual accuracy achieved should not be less than (and ideally should equal) the long-run average reported accuracy, in the sense that the difference of the two should go to zero.*

We do not attempt a formal mathematical statement of this principle, because many variants are possible. Instead we illustrate this (and later principles) through a variety of examples.

Assertion (to be justified as we proceed)**.** While other frequentist notions have value, this is the gold standard for frequentist evaluation. (An improvement is conditional frequentism – see Section 2.4 – but this is so much more complex that we mainly focus on satisfying the empirical frequentist principle in this paper.) Indeed, Neyman repeatedly pointed out – see, e.g., [20] – that the motivation for the frequentist principle is in repeated use of a procedure on differing real problems and not use on imaginary repetitions of one problem, as is often taught in textbooks.

2.1.1 Confidence Intervals

Consider a sequence of real problems $E_1, E_2, \ldots$, where E_i is an experiment yielding data x_i that arises probabilistically from a distribution having unknown parameter θ_i, both of which can vary from experiment to experiment; we *are not* (here) considering the usual frequentist notion of

studying repetitions of a fixed experiment with a given distribution and a fixed unknown θ.

The scenario considered, in this section, is that of producing confidence intervals for the θ_i, so the result of each analysis is a confidence interval $C_i(x_i)$, with stated confidence (of containing θ_i) equal to $1-\alpha_i(x_i)$. We are not defining 'confidence' here – it could be either frequentist or Bayesian, for instance – and we allow the stated confidence to depend on the data.

Suppose one eventually learns if $C_i(x_i)$ contains θ_i or not (say we are targeting stock prices in the future, and eventually learn them). The empirical frequentist principle could be formulated, in this context, as saying that

$$\lim_{N\to\infty}\Big[\frac{1}{N}\sum_{i=1}^{N}(1-\alpha_i(x_i))- \frac{\#\text{ times } C_i \text{ contained } \theta_i}{N}\Big]=0\,. \tag{2.1}$$

Thus the difference between the average reported confidence and the average attained coverage of the intervals should go to zero. It is often viewed as being acceptable to be conservative, which would happen, for instance, if the bracketed term in (2.1) were less than zero for sufficiently large N.

What is random in (2.1) will depend on context; typically $(x_1, x_2, \ldots, x_N)$ will be random, with a joint distribution specified by the distributions in $\{E_1, \ldots, E_N\}$, given $(\theta_1, \ldots, \theta_N)$. But sometimes the θ_i will also be random. (And sometimes neither the x_i nor θ_i are random; in finite population settings, for instance, both are considered fixed and the randomness comes from the random mechanism by which subjects are selected to be in the sample.) Such considerations arise when trying to prove that (2.1) holds, but the condition itself does not depend on any notion of randomness.

The textbook notion of confidence is, however, one possible report and can satisfy the empirical frequentist principle. Indeed, suppose that, in E_i,

$$P(C_i(x_i) \text{ contains } \theta_i \mid \theta_i) = 1-\alpha_i$$

for all θ_i (the probability is over the possible data x_i), i.e., $1-\alpha_i$ is the usual frequentist coverage of the confidence procedure in E_i. We will evaluate (2.1) when reporting $1-\alpha_i(x_i) = 1-\alpha_i$ as the error.

Letting $1_C(\theta)$ be the indicator function on the set C (one if $\theta \in C$ and zero otherwise), note that (2.1) can be rewritten

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}[(1-\alpha_i)-1_{C_i(x_i)}(\theta_i)]=0\,. \tag{2.2}$$

Also, $1-\alpha_i = E[1_{C_i(x_i)}(\theta_i) \mid \theta_i]$ (the expectation being over x_i), so $y_i = [1-\alpha_i - 1_{C_i(x_i)}(\theta_i)]$ is a zero mean random variable with variance bounded by 1. Since (2.2) is just the average of the y_i, the law of large numbers applies and concludes that the limit is indeed zero. Thus the E_i do not have to be problems of the same type, the α_i do not have to be the same, and the θ_i need have no relationships for (2.1) to hold.

2.1.2 Unbiasedness

Consider a sequence of different experiments E_i, with different θ_i, to be estimated by unbiased estimates $\hat{\theta}_i$ (so that $E[\hat{\theta}_i \mid \theta_i] = \theta_i$). If the θ_i were to become known, the differences $\hat{\theta}_i - \theta_i$ would then be mean 0 random variables and one could observe that the average differences converge to 0, under mild conditions. Whether or not this is a useful property can be debated, but it is an empirical frequentist property.

2.1.3 Empirical Bayes

Empirical Bayes analysis [21] is defined in an empirical frequentist way. The θ_i are assumed to arise from some unknown distribution $\pi(\cdot)$; for instance they could be independent draws from a $N(\theta_i \mid \xi, \tau^2)$ distribution, with ξ and τ^2 unknown. Data x_i are then observed from distributions with parameters θ_i. Analysis is then typically done with respect to the series of experiments corresponding to the (x_i, θ_i), and often in accordance with the empirical frequentist principle, under the assumption that the θ_i do arise from $\pi(\cdot)$.

2.1.4 Discussion and Interfaces with Bayesianism

The empirical frequentist principle seems compelling to most people. Imagine that a computer program to compute confidence intervals has been developed. Many different people use the program, specifying the $1-\alpha_i$ they want and the experiment E_i they conducted, and the program returns $C_i(x_i)$. Suppose someone discovers that, over many uses, the average reported confidence was 0.90, while only 70% of the confidence intervals actually contained the true θ_i. This would be a very misleading computer program.

We would argue that even Bayesians should accept the empirical frequentist principle. Perhaps the computer program above is a subjective Baysian program that guides users through the prior elicitation process and ultimately produces Bayesian confidence (credible) intervals. Something is very wrong if the reported confidence in repeated use of the program is 0.90, with actual coverage of only 70%. This could arise, for instance, if the subjective prior elicitation part of the program is producing prior distributions that are too concentrated (a well known issue with subjective elicitation, unless considerable care is taken), leading to credible intervals that are too small.

2.2 Type II. Procedural Frequentistism

Procedural frequentist principle. *Statistical procedures should be evaluated according to their frequentist properties, defined as properties of the procedure that would arise from repeated imaginary application of the procedure to a specified problem (model and unknown parameters given).*

2.2.1 Textbook Confidence Intervals

Confidence is often defined in a procedural frequentist way, with the experiment E being fixed, the unknown θ being fixed, and confidence $1-\alpha(\theta)$ being defined as the probability that the confidence set contains θ for imaginary repetitions of the experiment.(We are overusing α in this paper; $\alpha(\theta)$ here is distinct from the earlier $\alpha(x_i)$.) As observed earlier, if one develops confidence sets in the procedural frequentist way and the confidence $1-\alpha$ does not depend on θ, the confidence procedures will also have the empirical frequentist property. When the confidence does depend on θ, it is not uncommon to report $1-\alpha = \inf_\theta(1-\alpha(\theta))$ and such reports can be given a conservative empirical frequentist interpretation, with $\leq$ replacing $=$ in (2.1).

2.2.2 Consistency

A procedure is consistent if it converges to the truth as the sample size $n \to \infty$. (Note that this is distinct from the earlier N, which referred to the sequence of experiments being conducted.) This is a procedural frequentist principle, in that it involves an imaginary sequence of applications of the procedure to a given problem (model), but with growing sample size. There is no natural sense in which this is an empirical frequentist principle; one does not, in reality, continue to repeat the same experiment, but with growing sample sizes.

2.2.3 Type I Error

Consider testing H_0 versus H_1, with a rejection region $\mathcal{R}$ having Type I error $\alpha = P(\mathcal{R} \mid H_0)$. This is clearly a procedural frequentist quantity but we will see in Section 3.2 that it does not satisfy the empirical frequentist principle.

2.2.4 Sequential Endpoint Testing

Consider a sequence of null and alternative hypotheses $\{H_0^1, H_1^1\}$, $\{H_0^2, H_1^2\}$, ..., that are to be tested sequentially; the ordering of the hypotheses is important, and must be pre-specified. For instance H_1^1 could be the hypothesis that a new drug provides pain relief, H_1^2 could be the hypothesis that the same drug reduces blood pressure, and H_1^3 could be the hypothesis that the same drug promotes weight loss. Indeed, this type of example motivated the name sequential endpoint testing, with the three possible effects of the drug being the particular endpoints being studied.

In the simplest version of sequential endpoint testing, the same Type I error, α, is chosen for each hypothesis test. The procedure is to conduct the first test, stopping if H_0^1 is not rejected. If H_0^1 is rejected, one is allowed to perform the second test, stopping or continuing on depending on whether the second test fails to reject or rejects. Continuing on in this fashion, the end result is some sequence (possibly empty) $\{H_0^1, H_0^2, \ldots, H_0^m\}$ of rejected null hypotheses, with $m+1$ being the first time one fails to reject. The interesting procedural frequentist fact [17] is that

$$P(\text{one or more false rejections} \mid H_{i_1}^1, H_{i_2}^2, \ldots) \leq \alpha\,, \quad (2.3)$$

no matter what sequence of hypotheses is true. So, in the drug illustration and if all three tests are rejections, the drug company could claim that the drug is effective for all three purposes, with the probability that the procedure results in one or more incorrect rejections being no more than α. This is what now occurs in the world, with a drug often being labeled as effective for several things, based on sequential endpoint testing.

This at first seems odd to statisticians because it looks similar to traditional multiple testing for which, to obtain an overall level of α for the three tests, one would need to do the individual tests at level $\alpha/3$ (using Bonferonni, for simplicity). In sequential endpoint testing, however, not all tests are necessarily conducted; a test is conducted only if all the preceding tests were rejections, the crucial reason that no Type I error correction is needed. We show, however, in Section 3.2.3 that this procedure does not satisfy the empirical frequentist principle.

2.2.5 Discussion and Interfaces with Bayesianism

The procedural frequentist principle is less compelling than the empirical frequentist principle, in that it involves an imaginary sequence of experiments. The consistency example is one in which many (most) people would find the principle compelling, even though it is only procedural; using a procedure that fails as the data becomes nearly infinite provides a thought experiment that calls the procedure into question. Even Bayesians routinely accept consistency as necessary.

The procedural case for the Type I error, α, is not so compelling: one considers an imaginary sequence of experiments consisting of draws of data from H_0, and notes that the proportion of the time that the data is in $\mathcal{R}$ is α. Since this sequence of experiments is all under the assumption that H_0 is true, it is not obvious that one has learned much about the testing problem; this will be extensively discussed in Section 3.2. Of course, Type I error is a useful quantity for various other computations. In particular, when designing an experiment, Type I error and power are key quantities to consider, even for a Bayesian. (But, once the data are at hand, a Bayesian would not tend to utilize Type I error or power in making an error report.)

Procedural frequentist properties are often used by objective Bayesians to define objective priors. For instance, the confidence set procedural principle is used to define what are called *matching priors*, which are priors that yield posterior credible sets for a real parameter θ that have good frequentist behavior when viewed as a confidence procedure.

A surprising example. One of the earliest and most interesting examples of the matching prior idea was [15], which showed that the $100(1-\alpha)\%$ equal-tailed credible interval (i.e., the interval whose lower endpoint is the $\alpha/2$-quantile of the posterior distribution and whose upper endpoint is the $[1-\alpha/2]$-quantile) has the following rather astonishing procedural frequentist property: as the sample size $n \to \infty$, the

frequentist coverage of the Bayesian credible sets is $1-\alpha$, up to an error of order C/n for some constant C. This is astonishing because achieving frequentist coverage up to an error of C/n is noteworthy (achieving an error of $C/\sqrt{n}$ is easy), and yet Hartigan's result holds for essentially any prior distribution having full support.

The procedural frequentist procedure defined in sequential endpoint testing is incompatible with Bayesian reasoning. In particular, the Bayesian posterior probabilities of the alternative hypotheses satisfy $P(H_1^1 \mid data) \geq P(H_1^1, H_1^2 \mid data) \geq \cdots \geq P(H_1^1, H_1^2, \ldots H_1^m \mid data)$, so that increasing numbers of rejections result in less probability being assigned to all the rejections being correct. Indeed, this seems intuitively clear. If one managed to conduct 101 $\alpha = 0.05$-level tests via the sequential endpoint procedure, with each of the endpoints being very different (as in the earlier drug illustration), would anyone actually be willing to bet that all 100 rejections were correct? That seems scientifically ridiculous. Of course, this outcome has a nearly negligible probability of occurring and, hence, can be compatible with an overall $\alpha = 0.05$. But the possibility of being in this situation sends a warning that the procedural frequentist property here is difficult to interpret. Indeed, even assigning the same error probability to one rejection as to two or three rejections seems highly questionable.

We discussed four procedural frequentist examples in this section. The first, that of textbook presentation of coverage of confidence intervals, actually has an empirical frequentist justification, so there can be no criticism of it at this point (although see Section 2.4). The second example, that of consistency, has no clear empirical frequentist justification, but is almost universally agreed to be important.

The third example, that of Type I error in testing of a single hypothesis, has no clear empirical frequentist justification and is not as universally respected as consistency, but can be an important quantity to know. But the final example, that of sequential endpoint testing, is a highly suspect procedural frequentist procedure.

2.3 Type III. Computationally Frequentist

Computationally frequentist principle. *Statistical procedures should depend on quantities that involve frequentist averages over the sample space.*

2.3.1 P-values

In testing H_0, based on data x, where large values of $T(x)$ discredit H_0, the p-value $P(T(x) \geq T(x_{obs}) \mid H_0)$, where x_{obs} is the actual observation, is a probability on the sample space, so it satisfies the computational frequentist principle. Note, however, that it does not follow the procedural frequentist principle, because one cannot embed it in an imaginary sequence of problems where the p-value has a long-run frequentist interpretation. For instance, one might consider the imaginary experiments of repeatedly drawing data x_j from H_0, computing the p-value $p(x_j)$, rejecting H_0 if $p(x_j) < 0.05$ and then reporting $p(x_j)$ as a frequentist error probability. But the actual Type I frequentist error of this procedure is clearly 0.05, so that reporting p-values will always underestimate the procedural error. We will also see that the p-value badly fails to satisfy the empirical frequentist principle.

The p-value is often called the 'attained significance level,' in that it is the smallest α for which an α-level test would have rejected. One could imagine then running a sequence of imaginary experiments with this α, but this imaginary sequence will have a long-run frequentist interpretation only if α is used as the error probability, not if p-values are used in the new imaginary experiments.

Note that computationally frequentist arguments can be ridiculous. Here is an example, arising from an e-mail we received that was inquiring about the validity of the analysis.

Example 1. Suppose one observes $X \sim Binomial(\theta, 20)$ and is testing $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$, so that large values of x define the tail area for the p-value. The actual observation was $x_{obs} = 4$, the p-value $P(x \geq 4 \mid H_0) = 0.999$ was calculated, and the purported conclusion was that this was overwhelming evidence in favor of H_0. Of course, $x_{obs} = 4$ is actually quite strong evidence against either hypothesis, and the conclusion reached was based on a completely incorrect interpretation of p-values. (A sensible Bayesian analysis suggests that the evidence indeed favors H_0, but only by a factor of roughly 5 to 1.) One can do bad things using any statistical methodology, so the point here is just that a frequentist computation does not guarantee any statistical validity of a conclusion.

2.3.2 Discussion and Interfaces with Bayesianism

Our perspective is that the computationally frequentist principle lacks any force whatsoever. Just because one has computed some kind of average over the data does not mean that the resulting procedure has any value. We are not asserting that any procedure arising this way is useless, but merely saying that the fact that there was some data averaging going on provides no justification by itself.

p-values are interesting in this regard; they are valuable statistics and have a number of important uses, especially if they are properly calibrated; even Bayesians tend to use (calibrated) p-values for model checking. But Bayesians do not view their value as arising from the fact that they involve an average over data but, rather, that they are just a useful statistic.

Computationally frequentist procedures will generally not have a Bayesian interpretation, although sometimes they do through a mathematical quirk. For instance, in one-sided testing, p-values can equal the posterior probability of the null hypotheses for certain improper priors, but this is more a mathematical curiosity than something fundamental (cf, [4]). In two-sided testing, there is usually an extreme difference between p-values and posterior probabilities, a fact first clearly demonstrated in [10].

2.4 Type IV. Conditional Frequentistism

Bayesian analysis is typically phrased as being about "what is to be concluded from the problem and data at hand?" Frequentist analysis is about long-run performance guarantees. These are not necessarily incompatible. Indeed, conditional frequentists strive to achieve both long-run performance and optimal conclusions for the problem and data at hand. A general discussion of conditioning would take us too far afield (see [3] for a review and history), so we content ourselves here with an example.

We return to the confidence set situation to illustrate the issue. The best report would obviously be the oracle report of the indicator function $I_{C(x_i)}(\theta_i)$: one if the confidence interval contains θ_i and zero if it does not. It is thus natural to compare the stated confidence with how close it is to this oracle report, such as with use of the loss function

$$L(1-\alpha_i(x_i), C(x_i), \theta_i) = (1-\alpha_i(x_i) - I_{C(x_i)}(\theta_i))^2 \,. \quad (2.4)$$

One can consider a variety of ensuing expected losses but we simply present a classical example where any perspective makes the answer clear.

Example 2 (from [5])**.** Two observations, x_1 and x_2, are to be taken, where

$$x_j = \begin{cases} \theta+1 & \text{with probability } \frac{1}{2} \\ \theta-1 & \text{with probability } \frac{1}{2} \end{cases} .$$

Consider the frequentist confidence set, for the unknown θ, defined by

$$C(x_1,x_2) = \begin{cases} \text{the point } \{\frac{1}{2}(x_1+x_2)\} & \text{if } x_1 \neq x_2 \\ \text{the point } \{x_1-1\} & \text{if } x_1 = x_2. \end{cases}$$

The (unconditional) frequentist coverage of this confidence procedure can easily be shown to be

$$1-\alpha_U = P(C(x_1,x_2) \text{ contains } \theta \mid \theta) = 0.75.$$

This is not a sensible conclusion, once the data is at hand. To see this, observe that, if $x_1 \neq x_2$, then we know for sure that the average of the observations equals θ, so that the confidence set is then 100% accurate. On the other hand, if $x_1 = x_2$, θ is either the data's common value plus one or their common value minus one and each of these possibilities is equally likely to have occurred.

To obtain sensible frequentist answers here, one must define a conditioning statistic such as $s = |x_1 - x_2|$, which can be thought of as measuring the 'strength of evidence' in the data ($s = 2$ indicating data with maximal evidential content and $s = 0$ being data of minimal evidential content). Then one defines frequentist coverage conditional on the strength of evidence s. For the example, an easy computation shows that this conditional confidence is, for the two distinct cases,

$$\begin{aligned} 1-\alpha_C(s=2) &= P(C(x_1,x_2) \text{ contains } \theta \mid s=2,\theta) \\ &= 1\,, \end{aligned}$$

$$\begin{aligned} 1-\alpha_C(s=0) &= P(C(x_1,x_2) \text{ contains } \theta \mid s=0,\theta) \\ &= \frac{1}{2}. \end{aligned}$$

Conditional frequentist measures are fully frequentist and seem clearly better than unconditional frequentist measures. They have the same unconditional property (e.g., in the example, one will report 100% confidence half the time and 50% confidence half the time, resulting in an 'average' of 75% confidence, as must be the case to satisfy the empirical frequentist principle), yet give much better indications of the accuracy for the data that one has actually encountered.

To see this formally, consider the loss function in (2.4) and the corresponding frequentist risk (expected loss over the data (x_1, x_2) given θ). The risk of the constant error report, $1-\alpha_U = 0.75$, is

$$\begin{aligned} &E[(0.75 - I_{C(x_1,x_2)}(\theta_i))^2 \mid \theta\,] = \\ &\frac{1}{2}(0.75-1)^2 + \frac{1}{4}(0.75-1)^2 + \frac{1}{4}(0.75-0)^2 = \frac{3}{16}\,. \end{aligned}$$

In contrast, the risk of the conditional report, $1-\alpha_c(s)$, has the smaller risk

$$\begin{aligned} &E[(1-\alpha_C(s) - I_{C(x_1,x_2)}(\theta_i))^2 \mid \theta\,] = \\ &\frac{1}{2}(1-1)^2 + \frac{1}{4}(0.5-1)^2 + \frac{1}{4}(0.5-0)^2 = \frac{2}{16}\,. \end{aligned}$$

2.4.1 Discussion and Interfaces with Bayesianism

Finding good conditioning statistics is, in general, very difficult – so much so that the conditional frequentist theory of statistics is quite underdeveloped. Thus the typical approach today for developing conditional frequentist procedures is to develop objective Bayesian procedures (which automatically condition correctly) and show that they have excellent long-run frequentist behavior. The generalized fiducial approach mentioned in the introduction is another promising approach for doing this.

To illustrate this on the two-observation example in the previous section, the natural objective prior is $\pi(\theta) = 1$. Application of Bayes theorem trivially yields that, if $x_1 \neq x_2$, then the posterior distribution for the unknown θ gives probability one to the point $(x_1+x_2)/2$ while, if $x_1 = x_2$, then the posterior distribution gives probability $1/2$ each to the common value of the data plus 1 and the common value minus 1. It is immediate that the objective Bayesian confidence statements for $C(x_1,x_2)$ are 1 and 0.5 for the two cases, respectively, which is the optimal conditional frequentist answer.

The example in this section showed that even satisfaction of the empirical frequentist principle can be highly inadequate from the conditional frequentist perspective. (This could be corrected within the empirical frequentist paradigm by requiring some type of second empirical frequentist property, involving losses such as (2.4), but we do not pursue this.) This will be seen to be even more of a problem for

procedures that satisfy only the procedural frequentist principle, as will be extensively discussed in the next section.

3. HYPOTHESIS TESTING

3.1 Introduction

Hypothesis testing provides a more challenging illustration of the differences between the types of frequentists, and also illustrates the merging of frequentist and objective Bayesian statistics. As this is a pedagogical article, we do not attempt to study the empirical frequentist interpretation of hypothesis testing in general, but rather focus on the very special case in which the sequence of hypothesis tests being conducted is exchangeable, in the sense of having the same Type I error and power and having the same prior probability π_0 of the null hypothesis (when we are incorporating Bayesian concepts).

One situation in which this happens is daily quality control testing of an assembly line, where the repeated quality control checks involve exchangeable tests. Another example is Genome Wide Association Studies (GWAS) where, in each test, the alternative hypothesis is that a particular gene is associated with a particular disease and the null hypothesis is that there is no association; often, little is known about particular gene/disease associations so the tests are treated as exchangeable. The developments in this chapter could be done in much greater generality, with nonexchangeable hypotheses, but the exchangeable situation is sufficient for pedagogical understanding of the main issues.

It will be seen that involvement of π_0 is usually unavoidable for satisfaction of empirical frequentist properties. Sometimes π_0 is known. The quality control testing of an assembly line is one such example, where historical records provide the probabilities that the assembly line is operating correctly or is out of alignment. The prior probability of an association in GWAS is often considered known (cf. [24]), but can also be estimated from the data and becomes effectively known if the number of GWAS tests is huge (as is typical). More generally, in exchangeable multiple testing scenarios, one can learn π_0 as the number of tests grows [11].

When π_0 is not known, one could resort to the 'objective Bayesian' approach of giving each hypothesis equal prior probability, i.e., setting $\pi_0 = 0.5$. This is obviously not completely compelling, but does provide a reasonable default base for exploring the various frequentist principles.

3.2 Testing with Unconditional Error Probabilities

We have already seen that Type I error in testing is a procedural frequentist quantity. Can it also be given an empirical frequentist interpretation? We study this question here for standard hypothesis testing, multiple testing, and sequential endpoint testing.

3.2.1 Standard Hypothesis Testing

Consider the case of exchangeable simple hypothesis testing, with each of the E_i (recall that $\{E_1, E_2, \ldots, E_N\}$ is the sequence of experiments being considered) being a test of $H_0^i : \theta_i = \theta_{0i}$ versus $H_1^i : \theta_i = \theta_{1i}$, with rejection regions $\mathcal{R}_i$ having the same Type I error α and the same power $\beta = P(\mathcal{R}_i \mid H_1^i)$. There are various empirical frequentist properties that can be discussed in testing. For simplicity, we will focus on what could be called the empirical frequentist error probability under rejection, namely

$$\begin{aligned} & \lim_{N\to\infty}\left[\frac{\#\text{ times } H_0^i \text{ is true when rejecting}}{\#\text{ rejections}}\right] \\ = & \lim_{N\to\infty}\left[\frac{\#\text{ times } H_0^i \text{ is true when rejecting}/N}{\#\text{ rejections}/N}\right] \\ = & P(H_0^i, \mathcal{R}_i)/P(\mathcal{R}_i) \\ = & \frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)\beta} \equiv P(H_0^i \mid \mathcal{R}_i)\,, \end{aligned} \tag{3.1}$$

which is the posterior probability that H_0^i is true if one only knows that the test was a rejection. This is the actual error rate achieved in the sequence of experiments and so is the target for our reported error probabilities.

If π_0 is known, simply reporting error probability $\alpha_U = P(H_0^i \mid \mathcal{R}_i)$ clearly satisfies the empirical frequentist principle. Furthermore, this would be the correct error report corresponding to the experiment, before seeing the data. In [23], this was shown to be equal to what he defined as the pFDR. Thus stating pFDR (if π_0 is known) has an empirical frequentist justification. Note that the expected value of the first bracketed quantity above is essentially the regular FDR [2], which has a procedural frequentist justification but not an empirical frequentist interpretation.

If π_0 is not known and one makes the default assumption that $\pi_0 = 0.5$, note that $P(H_0^i \mid \mathcal{R}_i) = \alpha/[\alpha + \beta]$, which is nearly α when α is small and β is near one. Thus, for a highly powered test and if the hypotheses have equal prior probabilities, reporting α as the error probability does have approximate empirical frequentist justification.

The dependence of empirical frequentist error on prior probabilities is circumvented in Neyman-Pearson testing by only evaluating procedural properties of the test, namely α and β individually. The problem with this is that it is not unusual for people to interpret α as a surrogate for the empirical frequentist quantity $P(H_0^i \mid \mathcal{R}_i)$, but the two can obviously be very different, even if $\pi_0 = 1/2$, as shown in Table 1, where $\alpha = 0.05$ is chosen to define the rejection region. Thus, if the power is only 0.5 (as is common in GWAS), the actual error that will arise in rejecting over a series of real experiments is almost twice α and, as the power drops lower, the actual error rises dramatically. This important role of power in achieving empirical frequentist performance is often neglected, in part because it is not usually explained how to utilize power to understand empirical frequentist rejection error. We return to this issue in Section 3.4.

Table 1. Empirical frequentist error, when the prior probabilities of the hypotheses are equal, when $\alpha = 0.05$, and for various values of the power β.

β	1	0.95	0.8	0.5	0.05	0
$P(H_0^i \mid \mathcal{R}_i)$	.0476	0.05	0.0588	0.0909	0.5	1

3.2.2 Multiple Testing

Consider the multiple testing scenario in which E_i consists of performing m independent tests of hypotheses at nominal Type I error α/m (the Bonferroni correction) and power $\beta(m)$. The Type I error (a procedural frequentist quantity) for each E_i is then α, and we again study the extent to which this report has empirical frequentist justification. One could allow the m to vary over the E_i and the β's to vary – between both the E_i and the tests within each E_i – but the answers remain essentially the same. Also, assume for simplicity that each null hypothesis has prior probability $\pi_0 < 1$ of being true.

Consider the situation in which an error is made in E_i if any of the m tests results in an incorrect rejection (often called family-wide error in rejection). Then one natural empirical frequentist quantity to study is

$$\lim_{N \to \infty} \frac{\#\ E_i \text{ that have at least one incorrect rejection}}{\#\ E_i \text{ that have at least one rejection}} = \frac{P(\text{an } E_i \text{ has at least one incorrect rejection})}{P(\text{an } E_i \text{ has at least one rejection})}, \quad (3.2)$$

the false positive rate for family-wide error in rejection. There are other possibilities here, such as looking at all the tests within each E_i and studying the overall number of tests being incorrectly rejected, but utilizing family-wide error is standard.

Lemma 1. *For the multiple testing problem,*

$$\frac{P(\textit{an } E_i \textit{ has at least one incorrect rejection})}{P(\textit{an } E_i \textit{ has at least one rejection})} = \frac{1 - \left[1 - \frac{\pi_0 \alpha}{m}\right]^m}{1 - \left[1 - \frac{\pi_0 \alpha}{m} - (1 - \pi_0)\beta(m)\right]^m}. \quad (3.3)$$

Proof. Note first that, for a single test in E_i, $P(\text{not being an incorrect rejection}) = \pi_0(1 - \alpha/m) + (1 - \pi_0)$. Since E_i consists of m independent such tests, it follows that

$$\begin{aligned} &P(\text{an } E_i \text{ has at least one incorrect rejection}) \\ &= 1 - P(\text{an } E_i \text{ has no incorrect rejections}) \\ &= 1 - \left[\pi_0 \left(1 - \frac{\alpha}{m}\right) + (1 - \pi_0)\right]^m = 1 - \left[1 - \frac{\pi_0 \alpha}{m}\right]^m. \end{aligned}$$

Similarly,

$$\begin{aligned} &P(\text{an } E_i \text{ has at least one rejection}) \\ &= 1 - P(\text{an } E_i \text{ has no rejections}) \\ &= 1 - \left[\pi_0 \left(1 - \frac{\alpha}{m}\right) + (1 - \pi_0)(1 - \beta(m))\right]^m \\ &= 1 - \left[1 - \frac{\pi_0 \alpha}{m} - (1 - \pi_0)\beta(m)\right]^m. \end{aligned}$$

The conclusion follows. □

For large m, the numerator in (3.3) is approximately $1 - e^{-\pi_0 \alpha} \approx \pi_0 \alpha$ for small α. Typically, $\beta(m)$ goes to 1 as m grows, in which case an approximation to the denominator (and always a lower bound) can be shown to be $[1 - \pi_0^m(1 - \alpha)]$. Thus

$$\frac{P(\text{an } E_i \text{ has at least one incorrect rejection})}{P(\text{an } E_i \text{ has at least one rejection})} \approx \frac{\pi_0 \alpha}{1 - \pi_0^m(1 - \alpha)}. \quad (3.4)$$

To study this, it is important to realize that π_0 is often near 1 when m is large. For instance, in [24], m was huge and $\pi_0 = 1 - 10^{-5}$. It is thus useful to consider three types of behavior of π_0, with regards to increasing m.

Case 1. $\pi_0^m \to 0$ as m grows (e.g., $\pi_0 = 0.5$). Then (3.4) clearly converges to $\pi_0 \alpha$ as m grows, so that the multiple testing procedure does satisfy the empirical frequentist principle. Indeed, if one knows π_0, one can report the smaller error $\pi_0 \alpha$ with complete empirical frequentist validity.

Case 2. $\pi_0^m \to c$ $(0 < c < 1)$ as m grows (e.g., $\pi_0 = 1 + \frac{\log(c)}{m}$). Then (3.4) clearly converges to $\alpha/[1 - c(1 - \alpha)] > \alpha$ as m grows (since $\pi_0 \to 1$ in this situation), so that empirical frequentist validity is lacking.

Case 3. $\pi_0^m \to 1$ as m grows (e.g., $\pi_0 = 1 - \frac{c}{m^2}$). Then (3.4) clearly converges to 1 as m grows (since, again $\pi_0 \to 1$), so that the empirical frequentist performance is as bad as it can be.

3.2.3 Sequential Endpoint Testing

We return to the sequential endpoint testing example, and evaluate it from the empirical frequentist perspective. To keep matters simple, the only case that will be considered is that in which each endpoint test is conducted with the same Type I error α and power β, and the prior probability of each null hypotheses is π_0. Note that E_i is again a possible sequence of individual tests; thus E_{99} could be a sequence in which H_0^1 is rejected, H_0^2 is rejected, and H_0^3 is not rejected. (Recall that the only possible outcomes are of this type: a sequence of rejections followed by an acceptance.)

Of interest is again the actual empirical frequentist rejection error rate among the sequences that contain at least one rejection, namely the quantity (3.2).

Lemma 2. *For the sequential endpoint testing problem and if an infinite sequence of tests is available,*

$$\frac{P(\text{an } E_i \text{ has at least one incorrect rejection})}{P(\text{an } E_i \text{ has at least one rejection})} = \frac{\pi_0\alpha}{[\pi_0\alpha + (1-\pi_0)\beta][1-(1-\pi_0)\beta]}\,. \tag{3.5}$$

Proof. Here, $P(\text{an } E_i \text{ has at least one rejection}) = \pi_0\alpha + (1-\pi_0)\beta$, namely the probability that the first test in a sequence is a rejection; what happens subsequently does not change the fact that it is a sequence with a rejection. Recognizing that the possible ways of having an incorrect rejection are to have an incorrect rejection at the first test, which has probability $\pi_0\alpha$; or to have a correct rejection at the first test and an incorrect rejection at the second test, which has probability $(1-\pi_0)\beta \times \pi_0\alpha$; or to have two correct rejections followed by an incorrect rejection, etc., it follows that

$$\begin{aligned} & P(\text{an } E_i \text{ has at least one incorrect rejection}) \\ = \;& \pi_0\alpha + (1-\pi_0)\beta \times \pi_0\alpha + (1-\pi_0)^2\beta^2 \times \pi_0\alpha + \cdots \\ = \;& \pi_0\alpha[1 + (1-\pi_0)\beta + (1-\pi_0)^2\beta^2 + \cdots] \\ = \;& \frac{\pi_0\alpha}{1-(1-\pi_0)\beta}\,. \end{aligned}$$

The conclusion follows. □

Calculus allows computation of the minimum of (3.5) over β, resulting in the inequality

$$\frac{P(\text{an } E_i \text{ has at least one incorrect rejection})}{P(\text{an } E_i \text{ has at least one rejection})} \geq \frac{4\pi_0\alpha}{(1+\pi_0\alpha)^2}\,. \tag{3.6}$$

This lower bound can be shown to always exceed α, for small α, when $\pi_0 > \frac{1}{4} + \frac{\alpha}{8}$, so that anytime the null hypotheses have even modest probability of being true, sequential endpoint testing will not satisfy the empirical frequentist principle when measured by (3.5).

For the objective choice $\pi_0 = 1/2$ and α small, the above bound is approximately 2α and so stating that the error is α understates the error by a factor of 2. Even reporting 2α as the error does not satisfy the empirical frequentist principle because the inequality above is in the anti-conservative direction.

Similar analysis for sequential endpoint testing consisting of just m steps can be performed and yields lower bounds for the empirical frequentist rejection error (when $\pi_0 = 1/2$ and α is small) of $2(1-2^{-m})\,\alpha$. For instance, if $m = 2$, this is $(1.5)\,\alpha$, which is 50% larger than α. The clear indication is that, even though sequential endpoint testing does not get penalized in terms of Type I error for using α as the rejection level for each test, there is a penalty in terms of empirical frequentist rejection error.

3.3 Testing with Data Dependent Error Probabilities

3.3.1 Introduction

Again, we only consider the case of exchangeable simple hypothesis testing, with each of the E_i being a test of $H_0^i : \theta_i = \theta_{0i}$ versus $H_1^i : \theta_i = \theta_{1i}$, with rejection regions $\mathcal{R}_i$ having Type I error α and power $\beta = P(\mathcal{R}_i \mid H_1^i)$, and π_0 being the prior probability of H_0^i. There are various possible choices for data-dependent error probabilities α_i. Instead of working with the data, it is convenient to work with the p-values p_i (against the null hypotheses), and write $\alpha_i(p_i)$ as the reported error probability upon rejecting in E_i. (The p_i are only being used as convenient statistics here.) Recall that the target is the actual empirical frequentist error probability $P(H_0^i \mid \mathcal{R}_i)$ in (3.1), so the ideal is for the $\alpha_i(p_i)$ to satisfy

$$\begin{aligned} \lim_{N^*\to\infty} \frac{1}{N^*} \sum_{i=1}^{N^*} \alpha_i(p_i) &= P(H_0^i \mid \mathcal{R}_i) \\ &= \frac{\pi_0\alpha}{\pi_0\alpha + (1-\pi_0)\beta}\,, \end{aligned}$$

where N^* is the number of rejections and the average is over the $\alpha_i(p_i)$ in the rejections.

3.3.2 The Basic Empirical Frequentist Identity

Under the null hypotheses, the p_i have a uniform density on $(0,1)$ (assuming they are proper p-values). Let $f_1(p)$ denote the density of the p_i under the alternative hypotheses, the density being common across the E_i because of the exchangeability assumption. The following lemma follows directly.

Lemma 3. *If $\alpha_i(p_i) = \alpha(p_i)$ for some function $\alpha(\cdot)$ and recalling that we are only considering the series of, say, N^* rejections (i.e., $0 \leq p_i \leq \alpha$),*

$$\begin{aligned} \lim_{N^*\to\infty} \frac{1}{N^*} \sum_{i=1}^{N^*} \alpha_i(p_i) &= E[\alpha(p) \mid 0 \leq p \leq \alpha] \\ = \frac{1}{[\pi_0\alpha + (1-\pi_0)\beta]} \int_0^\alpha & \alpha(p)[\pi_0 + (1-\pi_0)f_1(p)]dp\,. \end{aligned} \tag{3.7}$$

This suggests an obvious data-dependent error probability report when π_0 is known, namely

$$\alpha_B(p_i) = \frac{\pi_0}{\pi_0 + (1-\pi_0)f_1(p_i)}\,. \tag{3.8}$$

For this choice, the right hand side of (3.7) clearly equals $P(H_0^i \mid \mathcal{R}_i)$, achieving exact empirical frequentist justification. In addition to this justification, these reported error probabilities have the highly desirable property of being data-dependent, with the reported error probability decreasing as the p-value decreases. As discussed in the conditional

Table 2. Table entries give the right hand side of (3.11) for the three discussed choices of reported errors, so that the indicated reported error satisfies the (conservative) empirical frequentist principle if π_0 is smaller than this bound.

n	α	Bound on π_0 for $\alpha_C(p_i) = \frac{1}{1+f_1(p_i)}$	Bound on π_0 for $\alpha_O(p_i) = \frac{-ep_i \log p_i}{1-ep_i \log p_i}$	Bound on π_0 for $\alpha_P(p_i) = p_i$
1	0.159	0.5	0.566	0.155
2	0.0787	0.5	0.523	0.0987
4	0.0228	0.5	0.369	0.0388
9	0.0013	0.5	0.0737	0.0034

Table 3. Values of R, from (3.10), when $\pi_0 = 1/2$, for the three discussed choices of reported errors.

n	α	R, when $\pi_0 = 0.5$, for $\alpha_C(p_i) = \frac{1}{1+f_1(p_i)}$	R, when $\pi_0 = 0.5$, for $\alpha_O(p_i) = \frac{-ep_i \log p_i}{1-ep_i \log p_i}$	R, when $\pi_0 = 0.5$, for $\alpha_P(p_i) = p_i$
1	0.159	1	1.21	0.248
2	0.0787	1	1.07	0.144
4	0.0228	1	0.635	0.0513
9	0.0013	1	0.091	0.00403

frequentist section, this is thus a much better frequentist report than $P(H_0^i \mid \mathcal{R}_i)$.

When π_0 is known, there is thus no frequentist controversy: report $P(H_0^i \mid \mathcal{R}_i)$ as the pre-experimental error probability under rejection but, upon observing the data, report $\alpha_B(p_i)$. Note that $\alpha_B(p_i) = P(H_0^i \mid p_i)$, i.e., is the posterior probability of the null hypothesis, given the data. The fact that the Bayesian error probability here is also the optimal empirical frequentist error probability was noted and discussed in [8].

When the alternative hypothesis is not simple (as assumed here), then $f_1(p)$ will also not be known. A method for dealing with this is discussed in the next section.

3.3.3 The Empirical Frequentist Performance of Common Error Reports

When π_0 is unknown, the following are commonly considered 'objective' conditional error reports.

Option 1. $\alpha_C(p_i) = 1/[1 + f_1(p_i)]$, the conditional frequentist Type I error considered in [8] (also the posterior probability of H_0 when $\pi_0 = 1/2$). This would be the optimal conditional error probability to report (from the empirical frequentist perspective) if $\pi_0 = 1/2$, but is not optimal otherwise (nor is it available if one does not know $f_1(p)$, as is common when the alternative hypothesis is composite).

Option 2. $\alpha_O(p_i) = -ep_i \log p_i/[1 - ep_i \log p_i]$ (e is the natural number), proposed in [25] and further motivated in [22] as a bound on the objective Bayes error probability, in the sense that

$$\frac{1}{1 + f_1(p)} > \frac{-ep \log p}{1 - ep \log p} \tag{3.9}$$

for almost any reasonable $f_1(p)$. The reason for developing this bound was to avoid the need to determine $f_1(p)$ for composite alternative hypotheses.

Option 3. $\alpha_P(p_i) = p_i$, i.e., report the p-value as the error probability.

The empirical frequentist performance of these data-dependent error probabilities will be studied by considering the ratio

$$\begin{aligned} R &= \frac{\text{average reported error}}{\text{average actual error}} \\ &= \frac{\int_0^\alpha \alpha(p)[\pi_0 + (1-\pi_0)f_1(p)]dp/[\pi_0\alpha + (1-\pi_0)\beta]}{\pi_0\alpha/[\pi_0\alpha + (1-\pi_0)\beta]} \\ &= \frac{1}{\alpha}\left[\int_0^\alpha \alpha(p)\, dp + \left(\frac{1}{\pi_0} - 1\right)\int_0^\alpha \alpha(p) f_1(p)]dp\right]. \end{aligned} \tag{3.10}$$

The (conservative) empirical frequentist principle is satisfied if $R \geq 1$ (the average reported error is not less than the average actual error), which will be true if

$$\pi_0 \leq \left(\frac{\alpha - \int_0^\alpha \alpha(p)\, dp}{\int_0^\alpha \alpha(p) f_1(p)]dp} + 1\right)^{-1}. \tag{3.11}$$

Example 3. Suppose the data is i.i.d. normal with mean θ and variance 1 and the tests are of $H_0 : \theta = -1$ versus $H_1 : \theta = 1$, with rejection region $\bar{x} > 0$. For various sample sizes n, Table 2 gives the right hand side of (3.11) for the three choices of $\alpha(p_i)$, while Table 3 gives the corresponding ratios R for the objective $\pi_0 = 0.5$. Note that, here, $\alpha = \Phi(-\sqrt{n})$, $\beta = 1 - \alpha$, $p = 1 - \Phi(\sqrt{n}[\bar{x} + 1])$, and $1/[1 + f_1(p)] = 1/[1 + e^{2n\bar{x}}]$, where Φ is the standard normal cdf.

From Tables 2 and 3, it is clear that reporting the p-value as the error probability is terrible according to the empirical frequentist principle; it is reasonable only when the (unknown) π_0 is very small. And the underreporting of error for the 'objective' $\pi_0 = 0.5$ is dramatic.

The conditional frequentist (objective Bayesian) report $\alpha_C(p_i) = \frac{1}{1+f_1(p_i)}$ is clearly very reasonable, needing only $\pi_0 \leq 0.5$ to have (conservative) empirical frequentist justification. As argued in [1], it is often the case that $\pi_0 > 0.5$, but then one should be performing a subjective Bayesian analysis.

Reporting $\alpha_O(p_i) = -ep_i \log p_i / [1 - ep_i \log p_i]$ is clearly considerably better than reporting the p-value in terms of the empirical frequentist principle, becoming much too small only for very small p-values. It can be shown that p is a factor of at least 3.85 smaller than $\alpha_O(p)$ when $p < 0.1$, so reporting p is almost 4 times worse than reporting $\alpha_O(p_i)$. For the case of simple hypothesis testing considered here, one could use the superior $\alpha_C(p_i) = 1/[1 + f_1(p_i)]$ with no additional computational cost but, for more general hypothesis testing problems, it can be difficult to compute the objective Bayesian error probability, while computing $\alpha_O(p_i)$ is as easy as computing the p-value.

It is surprising that $\alpha_O(p_i)$ has $R > 1$ when $n = 1$ and $n = 2$, implying that the inequality in (3.9) can fail for larger p-values. The inequality was established under a certain condition on the hazard rate corresponding to $f_1(p)$, and this condition is apparently violated for the simple hypothesis example considered here, when $n = 1$ and $n = 2$. In more general composite hypothesis testing problems arising in practice, the inequality does seem to hold and, interestingly, $\alpha_O(p_i)$ seems to often be quite close to $\alpha_C(p_i)$, as shown in empirical studies from [16] (see also [1]). Thus, general use of $\alpha_O(p_i)$ seems justified, even if it does not always strictly satisfy the empirical frequentist principle.

3.3.4 Data Dependent Procedural Frequentism

One might consider a data dependent version of procedural frequentism. For instance, one could propose evaluating data dependent Type I errors $\alpha_i(p_i) = \alpha(p_i)$ (for some function $\alpha(\cdot)$) by looking at an imaginary sequence of N^{**} rejections under the null hypothesis, and ask that

$$\begin{aligned} &\lim_{N^{**}\to\infty} \frac{1}{N^{**}} \sum_{i=1}^{N^{**}} \alpha(p_i) = E[\alpha(p) \mid H_0, \text{rejection}] \\ = \quad &\frac{1}{\alpha} \int_0^{\alpha} \alpha(p) dp = \alpha \,. \end{aligned} \tag{3.12}$$

One might then claim that reporting $\alpha(p_i)$ is as frequentist as reporting α. As an example, choosing $\alpha(p_i) = 2p_i$ yields $\frac{1}{\alpha}\int_0^{\alpha} 2p dp = \alpha$, so one might assert that reporting twice the p-value is as much a procedural frequentist procedure as reporting α.

This is not, however, a logical conclusion. α is the procedural frequentist property of the test, and the $\alpha(p_i)$ have no real meaning in terms of procedural frequentism.

It is, however, possible to develop data dependent procedural tests through conditioning. Indeed, [8] considers testing conditional on the statistic $S = \max\{p_0, p_1\}$, where p_0 is the p-value under H_0 and p_1 is the p-value under H_1. They show that the conditional Type I error, given S, is, for appropriate rejection regions $\mathcal{R}$, given by $\alpha(S) = P(\mathcal{R} \mid H_0, S) = \frac{1}{1+f_1(p_0)}$. This is a real procedural frequentist quantity, having the interpretation as the Type I error arising in a long series of experiments under H_0, where the data is compatible with the specified S. Noting that $\alpha(S)$ is always much bigger than $2p_i$ further reinforces the notion that satisfaction of (3.12) does not provide any procedural frequentist validity.

3.4 Testing with Odds

3.4.1 Pre-experimental odds

Recalling that π_0 is the prior probability of H_0^i, Bayes theorem gives

$$P(H_0^i \mid \mathcal{R}_i) = \frac{\pi_0 \alpha}{\pi_0 \alpha + (1-\pi_0)\beta} = \frac{\pi_0}{\pi_0 + (1-\pi_0)\frac{\beta}{\alpha}} \,, \tag{3.13}$$

which is commonly rewritten in terms of odds as

$$\begin{gathered} \frac{P(H_1^i|\mathcal{R}_i)}{P(H_0^i|\mathcal{R}_i)} = \frac{(1-\pi_0)}{\pi_0} \times \frac{\beta}{\alpha} \quad \text{or} \\ \text{pre-experimental odds} = \\ \text{prior odds} \times \text{experimental odds} \,, \end{gathered} \tag{3.14}$$

using the terminology in [1]. The pre-experimental odds have the very nice interpretation as the odds that a rejection, arising from the experiment, is correct to incorrect (often also called the odds of a true positive to a false positive). The big advantage of expressing things in terms of odds is that the prior odds separate out, so that those who do not wish to involve prior probabilities can focus on the experimental odds.

In classical statistics, it is left unstated as to how one should combine α and β to make inferences. Combining them through error probabilities, as in Section 3.2, is one possibility, but this mixes α and β up with the prior probabilities of hypotheses; (3.14) makes the sharper statement that inferences should depend on α and β only through the ratio β/α.

Consider (possibly data-dependent) reports $O_i(p_i)$ of the odds of having a correct rejection to an incorrect rejection in experiment E_i. A natural empirical frequentist principle would then be to satisfy, averaging over all N^* rejections,

$$\begin{aligned} &\lim_{N^*\to\infty} \frac{1}{N^*} \sum_{i=1}^{N^*} O_i(p_i) \\ = \quad &\lim_{N^*\to\infty} \frac{\#\text{ true rejections}}{\#\text{ false rejections}} \\ = \quad &\frac{(1-\pi_0)}{\pi_0} \times \frac{\beta}{\alpha} \,. \end{aligned} \tag{3.15}$$

With error probabilities it was natural to evaluate their long run performance by arithmetic averaging, but this is not so

natural with reported odds. Indeed, using either geometric averaging or arithmetic averaging of log odds in (3.15) may be more reasonable.

Note that, if the prior odds are known, the unconditional choice $O_i = (1-\pi_0)\beta/(\pi_0\alpha)$ trivially satisfies (3.15), which is strong empirical frequentist justification for the choice. (This would also be trivially true under geometric averaging.) If the prior odds are unknown, one can, at least, provide a procedural frequentist justification for β/α by considering an imaginary sequence of tests in which the prior odds are fixed at some specified value O^* (e.g., the objective choice $O^* = 1$), and then saying that $O_i = O^*\beta/\alpha$ will satisfy (3.15) for this imaginary sequence.

3.4.2 Data Dependent odds

(3.13) is a version of Bayes theorem applied pre-experimentally, depending only on $\mathcal{R}_i$. The post-experimental odds version of Bayes theorem is

$$\frac{P(H_1^i|p_i)}{P(H_0^i|p_i)} = \frac{(1-\pi_0)}{\pi_0} \times B_{10}(p_i) \quad \text{or} \tag{3.16}$$
$$\text{posterior odds of } H_1^i \text{ to } H_0^i =$$
$$\text{prior odds} \times \text{Bayes factor of } H_1^i \text{ to } H_0^i\,,$$

where, for our testing problem, $B_{10}(p_i) = f_1(p_i)/1$ (the density of the statistic p_i under the alternative hypothesis divided by the density under the null hypothesis).

Turning to the empirical frequentist performance of reporting $B_{10}(p_i)$, computation yields

$$\begin{aligned} &\lim_{N^*\to\infty} \frac{1}{N^*}\sum_{i=1}^{N^*} B_{10}(p_i) \\ &= \int_0^\alpha f_1(p)\frac{[\pi_0 + (1-\pi_0)f_1(p)]}{(\pi_0\alpha + (1-\pi_0)\beta)}dp \\ &= \frac{\pi_0\beta + \int_0^\alpha (1-\pi_0)f_1^2(p)dp}{(\pi_0\alpha + (1-\pi_0)\beta)} \geq \frac{\beta}{\alpha}\,, \end{aligned} \tag{3.17}$$

the last step following from Jensen's inequality, since

$$\int_0^\alpha f_1^2(p)\frac{1}{\alpha}dp \geq \left[\int_0^\alpha f_1(p)\frac{1}{\alpha}dp\right]^2 = \frac{\beta^2}{\alpha^2}\,.$$

If the prior odds O_i are known, the conditional report would be $O_iB_{10}(p_i)$. Thus (3.17) implies that these reports do not have an empirical frequentist justification (the target being $O_i\beta/\alpha$). This is still useful as a bound, however: the odds in favor of H_1^i cannot be larger than $O_iB_{10}(p_i)$.

Algebra shows that (see (3.8)) $P(H_0^i \mid p_i) = \alpha_B(p_i) = 1/(1+O_iB_{10}(p_i))$, and we saw in Section 3.3.2 that this is the optimal empirical frequentist error probability. Thus, if we had defined empirical frequentist performance of posterior odds by averaging the $1/(1+O_iB_{10}(p_i))$, the posterior odds approach would also be optimal. This would be a rather strange way to average odds, however.

One could have, instead, stated the odds of H_0^i to H_1^i. The relevant overall frequentist quantity would then have been α/β, while the Bayes factor would be $B_{01}(p_i) = 1/f_1(p_i)$. Now the empirical frequentist property would be

$$\begin{aligned} &\lim_{N^*\to\infty} \frac{1}{N^*}\sum_{i=1}^{N^*} B_{01}(p_i) \\ &= \int_0^\alpha \frac{1}{f_1(p)}\frac{[\pi_0 + (1-\pi_0)f_1(p)]}{(\pi_0\alpha + (1-\pi_0)\beta)}dp \\ &= \frac{(1-\pi_0)\alpha + \int_0^\alpha [\pi_0/f_1(p)]dp}{(\pi_0\alpha + (1-\pi_0)\beta)} \geq \frac{\alpha}{\beta}\,, \end{aligned} \tag{3.18}$$

the last step again following from Jensen's inequality, since

$$\int_0^\alpha \frac{1}{f_1(p)}\frac{1}{\alpha}dp \geq \frac{1}{\int_0^\alpha f_1(p)\frac{1}{\alpha}dp} = \frac{\alpha}{\beta}\,.$$

So, from an empirical frequentist perspective, one is now overstating the evidence in favor of H_0^i, which could be viewed as being conservative.

Finally, we consider an argument given in [1] concerning the data dependent reports $B_{10}(p_i)$. Averaging these over an imaginary sequence N^{**} of rejected true hypotheses H_0^i yields

$$\begin{aligned} &\lim_{N^{**}\to\infty} \frac{1}{N^{**}}\sum_{i=1}^{N^{**}} B_{10}(p_i) = E[B_{01}(p_i) \mid \mathcal{R}_i, H_0^i] \\ &= E[f_1(p) \mid \mathcal{R}_i, H_0^i] = \int_0^\alpha f_1(p)\frac{1}{\alpha}\,dp = \frac{\beta}{\alpha}\,. \end{aligned} \tag{3.19}$$

Thus it was claimed, in [1], that reporting the $B_{10}(p_i)$ under the H_0^i has the same long run procedural frequentist justification as reporting β/α. But β/α is the procedural frequentist quantity here and it is not clear that the $B_{10}(p_i)$ have any such justification (as was the case for the related interpretation of (3.12)). Recall, however, that $B_{10}(p_i)$ did have partial empirical frequentist justification.

3.4.3 Discussion and Interfaces with Bayesianism

Our conclusion about hypothesis testing is that, if the prior probabilities of the hypotheses are known, estimable or given (as in the objective choice of 1/2 each), then reporting $\alpha_B(p_i)$ is the optimal empirical frequentist error probability (also the optimal Bayesian error probability), because it exactly satisfies the empirical frequentist property, while being fully data-dependent. If prior probabilities are unknown and one is not willing to make the objectivity assumption, the situation is less clear, with the only compelling conclusion being that reporting the p-value as the error probability is terrible from the empirical frequentist perspective.

This lack of clarity, when prior probabilities are unknown, seems to argue for focusing on odds, rather than error probabilities, because one can then clearly separate prior odds and experimental odds. Unfortunately, β/α only has a nice

empirical frequentist interpretation when the prior odds are known, although it always has a procedural frequentist interpretation. The Bayes factor $B_{10}(p_i)$ does not exactly satisfy the empirical frequentist principle, even when the prior odds are known. So, based on frequentist reasoning alone, the situation with odds is murky. However, we could have, instead, averaged the $1/(1+O_iB_{10}(p_i))$, and then the posterior odds would have been the optimal empirical frequentist report.

This section only considered testing of simple hypotheses. References where these issues are discussed in more complicated testing scenarios, from a conditional frequentist perspective, include [6, 7], [9], and [1].

ACKNOWLEDGEMENTS

I thank Stefano Cabras, Luis Pericchi and referees and editors for very helpful corrections and comments.

FUNDING

National Science Foundation grant 1821289 and National Institute of Health grant 1P41EB028744-01A1.

Accepted 18 July 2022

REFERENCES

[1] Bayarri, M., Benjamin, D. J., Berger, J. O. and Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology* **72** 90–103. https://doi.org/10.1016/j.jmp.2015.12.007. MR3506028

[2] Benjamini, Y. and Hochberg, Y. (1995). 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. MR1325392

[3] Berger, J. O. (2014). Conditioning is the issue. *Past, Present, and Future of Statistical Science* 253.

[4] Berger, J. O. and Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association* **94**(446) 542–554. https://doi.org/10.2307/2670175. MR1702325

[5] Berger, J. O. and Wolpert, R. L. (1988). The likelihood principle. Institute of Mathematical Statistics. MR0773665

[6] Berger, J. O., Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science* **12**(3) 133–160. https://doi.org/10.1214/ss/1030037904. MR1617518

[7] Berger, J. O., Boukai, B. and Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika* **86**(1) 79–92. https://doi.org/10.1093/biomet/86.1.79. MR1688073

[8] Berger, J. O., Brown, L. D. and Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics* 1787–1807. https://doi.org/10.1214/aos/1176325757. MR1329168

[9] Dass, S. C. and Berger, J. O. (2003). Unified conditional frequentist and Bayesian testing of composite hypotheses. *Scandinavian Journal of Statistics* **30**(1) 193–210. https://doi.org/10.1111/1467-9469.00326. MR1965102

[10] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70**(3).

[11] Efron, B. (2012) *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* **1**. Cambridge University Press. https://doi.org/10.1017/CBO9780511761362. MR2724758

[12] Fisher, R. A. (1956). Statistical methods and scientific inference. MR0076233

[13] Good, I. J. (1983) *Good thinking: The foundations of probability and its applications*. U of Minnesota Press. MR0723501

[14] Hannig, J., Iyer, H., Lai, R. C. and Lee, T. C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association* **111**(515) 1346–1361. https://doi.org/10.1080/01621459.2016.1165102. MR3561954

[15] Hartigan, J. (1966). Note on the confidence-prior of Welch and Peers. *Journal of the Royal Statistical Society: Series B (Methodological)* **28**(1) 55–56. MR0195194

[16] Ioannidis, J. P. (2008). Effect of formal statistical significance on the credibility of observational associations. *American Journal of Epidemiology* **168**(4) 374–383.

[17] Maurer, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypothesis. *Biomed. Chem.-Pharm. Ind.* **6** 3–18.

[18] Mayo, D. G. (1996) *Error and the growth of experimental knowledge*. University of Chicago Press.

[19] Mayo, D. G. (2018). Statistical inference as severe testing. *Cambridge, UK: Cambridge Univ. Press Access provided by Katholieke Universiteit Leuven-KU Leuven on* **10**(25) 21.

[20] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* 97–131. https://doi.org/10.1007/BF00485695. MR0652325

[21] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics* **35**(1) 1–20. https://doi.org/10.1214/aoms/1177703729. MR0163407

[22] Sellke, T., Bayarri, M. and Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician* **55**(1) 62–71. https://doi.org/10.1198/000313001300339950. MR1818723

[23] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31**(6) 2013–2035. https://doi.org/10.1214/aos/1074290335. MR2036398

[24] Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145) 661–678.

[25] Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society: series B (Methodological)* **55**(2) 317–341. MR1224399

[26] Xie, M. q. g. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review* **81**(1) 3–39. https://doi.org/10.1111/insr.12000. MR3047496

James Berger. Department of Statistical Science, Duke University, USA. E-mail address: berger@duke.edu