Bayesian posteriors for arbitrarily rare events

Drew Fudenberg[a,1], Kevin He[b,1], and Lorens A. Imhof[c,d,1]

[a]Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Department of Economics, Harvard University, Cambridge, MA 02138; [c]Department of Statistics, Bonn University, 53113 Bonn, Germany; and [d]Hausdorff Center for Mathematics, Bonn University, 53113 Bonn, Germany

Contributed by Drew Fudenberg, March 27, 2017 (sent for review November 14, 2016; reviewed by Keisuke Hirano, Demian Pouzo, and Bruno Strulovici)

We study how much data a Bayesian observer needs to correctly infer the relative likelihoods of two events when both events are arbitrarily rare. Each period, either a blue die or a red die is tossed. The two dice land on side 1 with unknown probabilities p_1 and q_1, which can be arbitrarily low. Given a data-generating process where $p_1 \geq cq_1$, we are interested in how much data are required to guarantee that with high probability the observer's Bayesian posterior mean for p_1 exceeds $(1-\delta)c$ times that for q_1. If the prior densities for the two dice are positive on the interior of the parameter space and behave like power functions at the boundary, then for every $\epsilon > 0$, there exists a finite N so that the observer obtains such an inference after n periods with probability at least $1-\epsilon$ whenever $np_1 \geq N$. The condition on n and p_1 is the best possible. The result can fail if one of the prior densities converges to zero exponentially fast at the boundary.

rare event | Bayes etimate | uniform consistency | multinomial distribution | signaling game

Suppose a physician is deciding between a routine surgery and a newly approved drug for her patient. Either treatment can, in rare cases, lead to a life-threatening complication. She adopts a Bayesian approach to estimate the respective probability of complication, as is common among practitioners in medicine when dealing with rare events; see, for example, refs. 1 and 2 on the "zero-numerator problem." She reads the medical literature to learn about n patient outcomes associated with the two treatments and chooses the new drug if and only if her posterior mean regarding the probability of complication due to the drug is lower than $(1-\delta)$ times that of the surgery. As the true probability of complication becomes small for both treatments, how quickly does n need to increase to ensure that the physician will correctly choose surgery with probability at least $1-\epsilon$ when surgery is in fact the safer option?

Phrased more generally, we study how much data are required for the Bayesian posterior means on two probabilities to respect an inequality between them in the data-generating process, where these true probabilities may be arbitrarily small. Each period, one of two dice, blue or red, is chosen to be tossed. The choices can be deterministic or random, but have to be independent of past outcomes. The blue and red dice land on side k with unknown probabilities p_k and q_k, and the outcomes of the tosses are independent of past outcomes. Say that the posterior beliefs of a Bayesian observer satisfy (c,δ) monotonicity for side $\bar{k}$ if his posterior mean for $p_{\bar{k}}$ exceeds $(1-\delta)c$ times that for $q_{\bar{k}}$ whenever the true probabilities are such that $p_{\bar{k}} \geq cq_{\bar{k}}$. We assume the prior densities are continuous and positive on the interior of the probability simplex and behave like power functions at the boundary. Then we show that, under a mild condition on the frequencies of the chosen colors, for every $\epsilon > 0$, there exists a finite N so that the observer holds a (c,δ)-monotonic belief after n periods with probability at least $1-\epsilon$ whenever $np_{\bar{k}} \geq N$. This condition means that the expected number of times the blue die lands on side $\bar{k}$ must exceed a constant that is independent of the true parameter. Examples show that the sample size condition is the best possible and that the result can fail if one of the prior densities converges to zero exponentially fast at the boundary. A crucial aspect of our problem is the behavior of estimates when the true parameter value approaches the boundary of the parameter space, a situation that is rarely studied in a Bayesian context.

Suppose that in every period, the blue die is chosen with the same probability and that outcome $\bar{k}$ is more likely under the blue die than under the red one. Then, under our conditions, an observer who sees outcome $\bar{k}$ but not the die color is very likely to assign a posterior odds ratio to blue vs. red that is not much below the prior odds ratio. That is, the observer is unlikely to update her beliefs in the wrong direction. This corollary is used in ref. 3 to provide a learning-based foundation for equilibrium refinements in signaling games.

The best related result known so far is a consequence of the uniform consistency result of Diaconis and Freedman in ref. 4. Their result leads to the desired conclusion only under the stronger condition that the sample size is so large that the expected number of times the blue die lands on side k exceeds a threshold proportional to $1/p_{\bar{k}}$. That is, the threshold obtained from their result explodes as $p_{\bar{k}}$ approaches zero.

Our improvement of the sample size condition is made possible by a pair of inequalities that relate the Bayes estimates to observed frequencies. Like the bounds of ref. 4, the inequalities apply to all sample sequences without exceptional null sets and they do not involve true parameter values. Our result is related to a recent result of ref. 5, which shows that, under some conditions, the posterior distribution converges faster when the true parameter is on the boundary. Our result is also related to ref. 6, which considers a half space not containing the maximum-likelihood estimate of the true parameter and studies how quickly the posterior probability assigned to the half space converges to zero.

Bayes Estimates for Multinomial Probabilities

We first consider the simpler problem of estimating for a single K-sided die the probabilities of landing on the various sides. Suppose the die is tossed independently n times. Let X_k^n denote the number of times the die lands on side k. Then

Significance

Many decision problems in contexts ranging from drug safety tests to game-theoretic learning models require Bayesian comparisons between the likelihoods of two events. When both events are arbitrarily rare, a large data set is needed to reach the correct decision with high probability. The best result in previous work requires the data size to grow so quickly with rarity that the expectation of the number of observations of the rare event explodes. We show for a large class of priors that it is enough that this expectation exceeds a prior-dependent constant. However, without some restrictions on the prior the result fails, and our condition on the data size is the weakest possible.

Author contributions: D.F., K.H., and L.A.I. designed research, performed research, and wrote the paper.

Reviewers: K.H., Pennsylvania State University; D.P., University of California, Berkeley; and B.S., Northwestern University.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: drew.fudenberg@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618780114/-/DCSupplemental.

$X^n = (X_1^n, \ldots, X_K^n)$ has a multinomial distribution with parameter $n \in \mathbb{N}$ and unknown parameter $p = (p_1, \ldots, p_K) \in \Delta$, where $\mathbb{N}$ is the set of positive integers and $\Delta = \{p \in [0,1]^K : p_1 + \cdots + p_K = 1\}$. Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Let π be a prior density on Δ with respect to the Lebesgue measure λ on Δ, normalized by $\lambda(\Delta) = 1/(K-1)!$. Let $\pi(\cdot|X^n)$ be the posterior density after observing X^n.

Motivated by applications where some of the p_k can be arbitrarily small, we are interested in whether the relative error of the Bayes estimator $\hat{p}_k(X^n) = \int p_k \pi(p|X^n) d\lambda(p)$ is small with probability close to 1, uniformly on large subsets of Δ. Specifically, given $k \in \{1, \ldots, K\}$ and $\epsilon > 0$, we seek conditions on n and p and the prior, so that

$$\mathbb{P}_p(|\hat{p}_k(X^n) - p_k| < p_k\epsilon) \geq 1 - \epsilon. \quad \textbf{[1]}$$

A subscript on $\mathbb{P}$ or $\mathbb{E}$ indicates the parameter value under which the probability or expectation is to be taken.

For a wide class of priors, we show in *Theorem 1* that there is a constant N that is independent of the unknown parameter so that **[1]** holds whenever $\mathbb{E}_p(X_k^n) \geq N$. Denote the interior of Δ by int Δ.

Condition $\mathcal{P}$: We say that a density π on Δ satisfies *Condition* $\mathcal{P}(\alpha)$, where $\alpha = (\alpha_1, \ldots, \alpha_K) \in (0,\infty)^K$, if

$$\frac{\pi(p)}{\prod_{k=1}^K p_k^{\alpha_k - 1}}$$

is uniformly continuous and bounded away from zero on int Δ. We say that π satisfies *Condition* $\mathcal{P}$ if there exists $\alpha \in (0,\infty)^K$ so that π satisfies *Condition* $\mathcal{P}(\alpha)$.

For example, if $K = 2$, then π satisfies *Condition* $\mathcal{P}(\alpha)$ if and only if π is positive and continuous on int Δ and the limit $\lim_{p_k \to 0} \pi(p)/p_k^{\alpha_k - 1}$ exists and is positive for $k = 1, 2$. For every $K \geq 2$, every Dirichlet distribution has a density that satisfies *Condition* $\mathcal{P}$. Note that *Condition* $\mathcal{P}$ does not require that the density is bounded away from zero and infinity at the boundary. The present assumption on the behavior at the boundary is similar to Assumption P of ref. 5.

Theorem 1. *Suppose π satisfies Condition $\mathcal{P}$. Then for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ so that*

$$\mathbb{P}_p(|\hat{p}_k(X^n) - p_k| \geq p_k\epsilon) \leq \epsilon \quad \textbf{[2]}$$

if $np_k \geq N$.

The proofs of the results in this section are given in *SI Appendix*.

The proof of *Theorem 1* uses bounds on the posterior means given in *Proposition 1* below. These bounds imply that there is an $N \in \mathbb{N}$ so that if $np_k \geq N$ and the maximum-likelihood estimator $\frac{1}{n}X_k^n$ is close to p_k, then $|\hat{p}_k(X^n) - p_k| < p_k\epsilon$. It follows from Chernoff's inequality that the probability that $\frac{1}{n}X_k^n$ is not close to p_k is at most ϵ.

Inequality **2** shows a higher accuracy of the Bayes estimator $\hat{p}_k(X^n)$ when the true parameter p_k approaches 0. To explain this fact in a special case suppose that $K = 2$ and the prior is the uniform distribution. Then $\hat{p}_k(X^n) = (X_k^n + 1)/(n + 2)$ and the mean squared error of $\hat{p}_k(X^n)$ is $[np_k(1-p_k) + (1-2p_k)^2]/(n+2)^2$, which converges to 0 like $\frac{1}{n}$ when $p_k \in (0,1)$ is fixed and like $\frac{1}{n^2}$ when $p_k = \frac{1}{n}$. Moreover, by Markov's inequality, the probability in **[2]** is less than $(np_k + 1)/(n^2 p_k^2 \epsilon^2)$, so that in this case we can choose $N = 2/\epsilon^3$. In general, we do not have an explicit expression for the threshold N, but in *Remark 2* we discuss the properties of the prior that have an impact on the N we construct in the proof.

Condition $\mathcal{P}$ allows the prior density to converge to zero at the boundary of Δ like a power function with an arbitrarily large exponent. The following example shows that the conclusion of *Theorem 1* fails to hold for a prior density that converges to 0 exponentially fast.

Example 1: Let $K = 2$, $\pi(p) \propto e^{-1/p_1}$, and $\delta > 0$. Then for every $N \in \mathbb{N}$, there exist $p \in \Delta$ and $n \in \mathbb{N}$ with $n^{\frac{1}{2}+\delta} p_1 \geq N$ so that

$$\mathbb{P}_p(|\hat{p}_1(X^n) - p_1| > p_1) = 1.$$

The idea behind this example is that the prior assigns very little mass near the boundary point where $p_1 = 0$, so if the true parameter p_1 is small, the observer needs a tremendous amount of data to be convinced that p_1 is in fact small. The prior density in our example converges to 0 at an exponential rate as $p_1 \to 0$, and it turns out that the amount of data needed so that $\hat{p}_1(X^n)/p_1$ is close to 1 grows quadratically in $1/p_1$. For every fixed $N \in \mathbb{N}$ and $\delta > 0$, the pairs (n, p_1) satisfying the relation $n^{\frac{1}{2}+\delta} p_1 = N$ involve a subquadratic growth rate of n with respect to $1/p_1$. So we can always pick a small enough p_1 such that the corresponding data size n is insufficient.

The next example shows that the sample size condition of *Theorem 1*, $np_k \geq N$, cannot be replaced by a weaker condition of the form $\zeta(n)p_k \geq N$ for some function ζ with $\limsup_{n\to\infty} \zeta(n)/n = \infty$. Put differently, the set of p for which **[2]** can be proved cannot be enlarged to a set of the form $\{p : p_k \geq \phi_\epsilon(n)\}$ with $\phi_\epsilon(n) = o(1/n)$.

Example 2: Suppose π satisfies *Condition* $\mathcal{P}$. Let $\zeta : \mathbb{N} \to (0,\infty)$ be so that $\limsup_{n\to\infty} \zeta(n)/n = \infty$. Then for every $N \in \mathbb{N}$, there exist $p \in \Delta$ and $n \in \mathbb{N}$ with $\zeta(n)p_1 \geq N$ so that

$$\mathbb{P}_p(|\hat{p}_1(X^n) - p_1| > p_1) = 1.$$

The following proposition gives fairly sharp bounds on the posterior means under the assumption that the prior density satisfies *Condition* $\mathcal{P}$. The result is purely deterministic and applies to all possible sample sequences. The bounds are of interest in their own right and also play a crucial role in the proofs of *Theorems 1* and *2*.

Proposition 1. *Suppose π satisfies Condition $\mathcal{P}(\alpha)$. Then for every $\epsilon > 0$, there exists a constant $\gamma > 0$ such that*

$$(1-\epsilon)\frac{n_k + \alpha_k}{n + \gamma} \leq \frac{\int p_k \left(\prod_{i=1}^K p_i^{n_i}\right) \pi(p)\, d\lambda(p)}{\int \left(\prod_{i=1}^K p_i^{n_i}\right) \pi(p)\, d\lambda(p)} \leq (1+\epsilon)\frac{n_k + \gamma}{n + \gamma} \quad \textbf{[3]}$$

for $k = 1, \ldots, K$ and all $n, n_1, \ldots, n_K \in \mathbb{N}_0$ with $\sum_{i=1}^K n_i = n$.

Remark 1: If π is the density of a Dirichlet distribution with parameter $\alpha \in (0,\infty)^K$, then the inequalities in **[3]** hold with $\epsilon = 0$ and $\gamma = \sum_{k=1}^K \alpha_k$, and the inequality on the left-hand side is an equality. If π is the density of a mixture of Dirichlet distributions and the support of the mixing distribution is included in the interval $[a, A]^K$, $0 \leq a \leq A < \infty$, then for all k and $n_1, \ldots, n_K$ with $\sum_{i=1}^K n_i = n$,

$$\frac{n_k + a}{n + KA} \leq \frac{\int p_k \left(\prod_{i=1}^K p_i^{n_i}\right) \pi(p)\, d\lambda(p)}{\int \left(\prod_{i=1}^K p_i^{n_i}\right) \pi(p)\, d\lambda(p)} \leq \frac{n_k + A}{n + Ka}. \quad \textbf{[4]}$$

The proofs of our main results, *Theorems 1* and *2*, apply to all priors whose densities satisfy inequalities **3** or **4**. In particular, the conclusions of these theorems and of their corollaries hold if the prior distribution is a mixture of Dirichlet distributions and the support of the mixing distribution is bounded.

Remark 2: *Condition* $\mathcal{P}(\alpha)$ implies that the function $\pi(p)/\prod_{k=1}^K p_k^{\alpha_k - 1}$, $p \in$ intΔ, can be extended to a continuous function $\tilde{\pi}(p)$ on Δ. The proof of *Proposition 1* relies on the fact that $\tilde{\pi}$ can be uniformly approximated by Bernstein polynomials. An inspection of the proof shows that the constant γ in **[3]** can be taken to

be $m + \sum_{k=1}^{K} \alpha_k$, where m is so large that h_m, the mth-degree Bernstein polynomial of $\tilde{\pi}$, satisfies

$$\max\{|h_m(p) - \tilde{\pi}(p)| : p \in \Delta\} \leq \frac{\min\{\tilde{\pi}(p) : p \in \Delta\}}{1 + 2\epsilon^{-1}}.$$

Hence, in addition to a small value of ϵ, the following properties of the density π result in a large value of γ: (*i*) if $\sum_{k=1}^{K} \alpha_k$ is large, (*ii*) if π is a "rough" function so that $\tilde{\pi}$ is hard to approximate and m needs to be large, and (*iii*) if $\tilde{\pi}$ is close to 0 somewhere. The threshold N in *Theorem 1* depends on the prior through the constant γ from *Proposition 1* and the properties of π just described will also lead to a large value of N.

In particular, $N \to \infty$ if $\sum_{k=1}^{K} \alpha_k \to \infty$. For example, consider a sequence of priors $\pi^{(j)}$ for $K = 2$, where $\pi^{(j)}$ is the density of the Dirichlet distribution with parameter $(j, 1)$, so that $\pi^{(j)}$ satisfies *Condition* $\mathcal{P}(\alpha)$ with $\alpha_1 = j$. As $j \to \infty$, $\pi^{(j)}$ converges faster and faster to 0 as $p_1 \to 0$, although never as fast as in *Example 1*, where no finite N can satisfy the conclusion of *Theorem 1*. If $n = 4j$ and $p_1 = \frac{1}{12}$, then under $\pi^{(j)}$, $\hat{p}_1(X^n) = (X_1^n + j)/(n + j + 1) \geq 2p_1$, so for every $\epsilon \in (0, 1)$, the probability in *Theorem 1* is 1. Thus, the smallest N for which the conclusion holds must exceed $4j \times \frac{1}{12} = \frac{j}{3}$.

Remark 3: Using results on the degree of approximation by Bernstein polynomials, one may compute explicit values for the constants γ in *Proposition 1* and N in *Theorem 1*. Details are given in in *SI Appendix*, *Remarks 3′* and *3″*.

Remark 4: Suppose $K > 2$ and the statistician is interested in only one of the probabilities p_k, say $p_{\bar{k}}$. Then, instead of using $\hat{p}_{\bar{k}}(X^n)$, he may first reduce the original $(K-1)$-dimensional estimation problem to the problem of estimating the one-dimensional parameter $(p_{\bar{k}}, \sum_{k \neq \bar{k}} p_k)$ of the Dirichlet distribution of $(X_{\bar{k}}^n, \sum_{k \neq \bar{k}} X_k^n)$. He will then distinguish only whether or not the die lands on side $\bar{k}$ and will use the induced one-dimensional prior distribution for the parameter of interest. If the original prior is a Dirichlet distribution on Δ, both approaches lead to the same Bayes estimators for $p_{\bar{k}}$, but in general, they do not. *SI Appendix*, *Proposition 2* shows that whenever the original density π satisfies *Condition* $\mathcal{P}$, then the induced density satisfies *Condition* $\mathcal{P}$ as well. However, it may happen that the induced density satisfies *Condition* $\mathcal{P}$ even though the original density does not. For example, if $K = 3$ and $\pi(p) \propto e^{-1/p_1} + p_2$, then π does not satisfy *Condition* $\mathcal{P}$, but for each $\bar{k} = 1, 2, 3$, the induced density does.

Comparison of Two Multinomial Distributions

Here we consider two dice, blue and red, each with $K \geq 2$ sides. In every period, a die is chosen. We first consider the case where the choice is deterministic and fixed in advance. We later allow the choice to be random. The chosen die is tossed and lands on the kth side according to the unknown probability distributions $p = (p_1, \ldots, p_K)$ and $q = (q_1, \ldots, q_K)$ for the blue and the red die, respectively. The outcome of the toss is independent of past outcomes. The parameter space of the problem is Δ^2. The observer's prior is represented by a product density $\pi(p)\varrho(q)$ over Δ^2; that is, he regards the parameters p and q as realizations of independent random vectors.

Let X^n be a random vector that describes the outcomes, i.e., colors and sides, of the first n tosses. Let b_n denote the number of times the blue die is tossed in the first n periods. Let $\pi(\cdot|X^n)$ and $\varrho(\cdot|X^n)$ be the posterior densities for the blue and the red die after observing X^n. Let $\hat{p}_k(X^n) = \int p_k \pi(p|X^n) d\lambda(p)$ and $\hat{q}_k(X^n) = \int q_k \varrho(q|X^n) d\lambda(q)$. The product form of the prior density ensures that the marginal posterior distribution for either die is completely determined by the observations on that die and the marginal prior for that die.

We study the following problem. Fix a side $\bar{k} \in \{1, \ldots, K\}$ and a constant $c \in (0, \infty)$. Consider a family of environments, each characterized by a data-generating parameter vector $\vartheta = (p, q) \in \Delta^2$ and an observation length n. In each environment, we have $p_{\bar{k}} \geq cq_{\bar{k}}$, and we are interested in whether the Bayes estimators reflect this inequality. In general, one cannot expect that the probability that $\hat{p}_{\bar{k}}(X^n) \geq c\hat{q}_{\bar{k}}(X^n)$ is much higher than $\frac{1}{2}$ when $p_{\bar{k}} = cq_{\bar{k}}$. We therefore ask whether in all of the environments, the observer has a high probability that $\hat{p}_{\bar{k}}(X^n) \geq c(1 - \delta)\hat{q}_{\bar{k}}(X^n)$ for a given constant $\delta \in (0, 1)$.

Clearly, as $p_{\bar{k}}$ approaches 0, we will need a larger observation length n for the data to overwhelm the prior. But how fast must n grow relative to $p_{\bar{k}}$? Applying the uniform consistency result of ref. 4 to each Bayes estimator separately leads to the condition that n must be so large that the expected number of times the blue die lands on side $\bar{k}$, that is, $b_n p_{\bar{k}}$, exceeds a threshold that explodes when $p_{\bar{k}}$ approaches zero. The following theorem shows that there is a threshold that is independent of p, provided the prior densities satisfy *Condition* $\mathcal{P}$.

Theorem 2. *Suppose that π and ϱ satisfy Condition $\mathcal{P}$. Let $\bar{k} \in \{1, \ldots, K\}$, $c \in (0, \infty)$, and $\epsilon, \delta, \eta \in (0, 1)$. Then there exists $N \in \mathbb{N}$ so that for every deterministic sequence of choices of the dice to be tossed,*

$$\mathbb{P}_\vartheta(\hat{p}_{\bar{k}}(X^n) \geq c(1 - \delta)\hat{q}_{\bar{k}}(X^n)) \geq 1 - \epsilon \qquad [5]$$

for all $\vartheta = (p, q) \in \Delta^2$ with $p_{\bar{k}} \geq cq_{\bar{k}}$ and all $n \in \mathbb{N}$ with $b_n p_{\bar{k}} \geq N$ and $b_n/n \leq 1 - \eta$.

We prove *Theorem 2* in the next section.

Note that the only constraints on the sample size here are that the product of b_n with $p_{\bar{k}}$ be sufficiently large and the proportion of periods in which the red die is chosen be not too small. However, $p_{\bar{k}}$ and $q_{\bar{k}}$ can be arbitrarily small. This is useful in analyzing situations where the data-generating process contains rare events.

In the language of hypothesis testing, *Theorem 2* says that under the stated condition on the prior, the test that rejects the null hypothesis $p_{\bar{k}} \geq cq_{\bar{k}}$ if and only if $\hat{p}_{\bar{k}}(X^n) < c(1 - \delta)\hat{q}_{\bar{k}}(X^n)$ has a type I error probability of at most ϵ provided $p_{\bar{k}} \geq N/b_n$ (and $b_n/n < 1 - \eta$). For every n, the bound on the error probability holds uniformly on the specified parameter set. Note that such a bound cannot be obtained for a test that rejects the hypothesis whenever $\hat{p}_{\bar{k}}(X^n) < c\hat{q}_{\bar{k}}(X^n)$.

We now turn to the case where the dice are randomly chosen. The probability of choosing the blue die need not be constant over time but must not depend on the unknown parameter ϑ. Let the random variable B_n denote the number of times the blue die is tossed in the first n periods.

Corollary 1. *Suppose that π and ϱ satisfy Condition $\mathcal{P}$. Let $\bar{k} \in \{1, \ldots, K\}$, $c \in (0, \infty)$, and $\epsilon, \delta \in (0, 1)$. Suppose that in every period, the die to be tossed is chosen at random, independent of the past, and that*

$$\liminf_{n \to \infty} \frac{\mathbb{E}(B_n)}{n} > 0, \qquad \limsup_{n \to \infty} \frac{\mathbb{E}(B_n)}{n} < 1. \qquad [6]$$

Then there exists $N \in \mathbb{N}$ so that

$$\mathbb{P}_\vartheta(\hat{p}_{\bar{k}}(X^n) \geq c(1 - \delta)\hat{q}_{\bar{k}}(X^n)) \geq 1 - \epsilon \qquad [7]$$

for all $\vartheta = (p, q) \in \Delta^2$ with $p_{\bar{k}} \geq cq_{\bar{k}}$ and all $n \in \mathbb{N}$ with $np_{\bar{k}} \geq N$.

The proof of *Corollary 1* is given at the end of the next section.

In the decision problem described at the beginning, *Theorem 2* and *Corollary 1* ensure that whenever surgery is the safer option, the probability that the physician actually chooses surgery is at least $1 - \epsilon$ unless the probability of complication due to the drug is smaller than N/n. Except for this last condition, the bound $1 - \epsilon$ holds uniformly over all possible parameters.

In the rest of this section we assume that in every period the blue die is chosen at random with the same probability μ_B. The value of μ_B need not be known; we assume only that $0<\mu_B<1$, so that condition **6** is met.

The following example shows that the conditions on the prior densities cannot be omitted from *Corollary 1*.

Example 3: Suppose $K=2$ and $0<\mu_B<1$. Suppose π satisfies *Condition* $\mathcal{P}$ and $\varrho(q)\propto e^{-1/q_1}$. Let $c>0$. Then for every $N\in\mathbb{N}$, there exist $\vartheta=(p,q)\in\Delta^2$ with $p_1\geq cq_1$ and $n\in\mathbb{N}$ with $np_1\geq N$ so that

$$\mathbb{P}_\vartheta\left(\hat{p}_1(X^n)<\frac{c}{2}\hat{q}_1(X^n)\right)>\frac{1}{2}.$$

The next example shows that the sample size condition of *Corollary 1*, $np_{\bar{k}}\geq N$, is the best possible for small $p_{\bar{k}}$. It cannot be replaced by a weaker condition of the form $n\zeta(p_{\bar{k}})\geq N$ for some function ζ with $\lim_{t\to 0+}\zeta(t)/t=\infty$. In particular, taking ζ to be a constant function shows that there does not exist $N\in\mathbb{N}$ so that $n\geq N$ implies that **[7]** holds uniformly for all ϑ with $p_k\geq cq_k$.

Example 4: Suppose that $0<\mu_B<1$ and that π and ϱ satisfy *Condition* $\mathcal{P}$. Let $c>0$. Let ζ be a nonnegative function on $[0,1]$ with $\lim_{t\to 0+}\zeta(t)/t=\infty$. Then there exists $\epsilon_0>0$ so that for every $N\in\mathbb{N}$, there exist $\vartheta=(p,q)\in\Delta^2$ with $p_1\geq cq_1$ and $n\in\mathbb{N}$ with $n\zeta(p_1)\geq N$ so that

$$\mathbb{P}_\vartheta\left(\hat{p}_1(X^n)<\frac{c}{2}\hat{q}_1(X^n)\right)>\epsilon_0.$$

Examples 3 and *4* are proved in *SI Appendix*.

Suppose that after data X^n the observer was told that the next outcome was $\bar{k}$ but not which die was used. Then Bayes' rule implies the posterior odds ratio for "blue" relative to "red" is

$$\frac{\mu_B\int p_{\bar{k}}\pi(p|X^n)\,d\lambda(p)}{\mu_R\int q_{\bar{k}}\varrho(q|X^n)\,d\lambda(q)},$$

where $\mu_R=1-\mu_B$.

Corollary 2. *Suppose that π and ϱ satisfy Condition $\mathcal{P}$. Then there exists $N\in\mathbb{N}$ such that whenever $p_{\bar{k}}\geq q_{\bar{k}}$ and $np_{\bar{k}}\geq N$, there is probability at least $1-\epsilon$ that the posterior odds ratio of blue relative to red exceeds $(1-\epsilon)\cdot\frac{\mu_B}{\mu_R}$ when the $(n+1)$th die lands on side $\bar{k}$.*

Corollary 2 is used by Fudenberg and He in ref. 3, who provide a learning-based foundation for equilibrium refinements in signaling games. They consider a sequence of learning environments, each containing populations of blue senders, red senders, and receivers. Senders are randomly matched with receivers each period and communicate using one of K messages. There is some special message $\bar{k}$, whose probability of being sent by blue senders always exceeds the probability of being sent by red senders in each environment. Suppose the common prior of the receivers satisfies *Condition* $\mathcal{P}$ and, in every environment, there are enough periods that the expected total observations of blue sender playing $\bar{k}$ exceed a constant. Then at the end of every environment, by *Corollary 2* all but ϵ fraction of the receivers will assign a posterior odds ratio for the color of the sender not much less than the prior odds ratio of red vs. blue, if they were to observe another instance of $\bar{k}$ sent by an unknown sender, regardless of how rarely the message $\bar{k}$ is observed. A leading case of receiver prior satisfying *Condition* $\mathcal{P}$ is fictitious play, the most commonly used model of learning in games, which corresponds to Bayesian updating from a Dirichlet prior, but *Corollary 2* shows that the Dirichlet restriction can be substantially relaxed.

Proofs of *Theorem 2* and *Corollary 1*

We begin with two auxiliary results needed in the proof of *Theorem 2*. *Lemma 1* is a large deviation estimate that gives a bound on the probability that the frequency of side $\bar{k}$ in the tosses of the red die exceeds an affine function of the frequency of side $\bar{k}$ in the tosses of the blue die. *Lemma 2* implies that, with probability close to 1, the number of times the blue die lands on side $\bar{k}$ exceeds a given number when $b_np_{\bar{k}}$ is sufficiently large. The proofs of *Lemmas 1* and *2* are in *SI Appendix*.

Lemma 1. *Let S_n be a binomial random variable with parameters n and p, and let T_m be a binomial random variable with parameters m and q. Let $0<c'<c$ and $d>0$. Suppose S_n and T_m are independent, and $p\geq cq$. Then*

$$\mathbb{P}\left(\frac{T_m}{m}\geq\frac{1}{c'}\frac{S_n}{n}+\frac{d}{n\wedge m}\right)\leq\left(\frac{c'}{c}\right)^{c'd/(c'+1)}.$$

Lemma 2. *Let $M<\infty$ and $\epsilon>0$. Then there exists $N\in\mathbb{N}$ so that if S_n is a binomial random variable with parameters n and p and $np\geq N$, then*

$$\mathbb{P}_p(S_n\leq M)\leq\epsilon.$$

Proof of Theorem 2: Let $r_n=n-b_n$ be the number of times the red die is tossed in the first n periods. Let Y_n and Z_n be the respective number of times the blue and the red die land on side $\bar{k}$. Choose $\beta>0$ and $c'\in(0,c)$ so that

$$\frac{1-\beta}{(1+\beta)(1-\delta)}>\frac{c}{c'}+\delta. \qquad \textbf{[8]}$$

By *Proposition 1*, there exists $\gamma>0$ so that for every $n\in\mathbb{N}$,

$$\hat{p}_{\bar{k}}(X^n)\geq\phi(b_n,Y_n),\qquad \hat{q}_{\bar{k}}(X^n)\leq\psi(r_n,Z_n), \qquad \textbf{[9]}$$

where

$$\phi(b,y)=(1-\beta)\frac{y}{b+\gamma},\qquad \psi(r,z)=(1+\beta)\frac{z+\gamma}{r}.$$

Let $d>0$ be so that the bound in *Lemma 1* satisfies $(c'/c)^{c'd/(c'+1)}\leq\frac{\epsilon}{2}$.

We now show that for all $b,r\in\mathbb{N}$, $y=0,\dots,b$, and $z=0,\dots,r$, the inequalities

$$\frac{z}{r}<\frac{1}{c'}\frac{y}{b}+\frac{d}{b\wedge r},\quad \frac{2c\gamma}{c'\delta}<b<\frac{r}{\eta},\quad y>M:=\frac{3c(d+\gamma)}{\delta\eta} \qquad \textbf{[10]}$$

imply that

$$\phi(b,y)>c(1-\delta)\psi(r,z). \qquad \textbf{[11]}$$

It follows from the first and the third inequality in **[10]** that

$$\begin{aligned}\psi(r,z)&<\psi\left(r,\frac{ry}{c'b}+\frac{rd}{b\wedge r}\right)\\&=(1+\beta)\left(\frac{y}{c'b}+\frac{d}{b\wedge r}+\frac{\gamma}{r}\right)\\&\leq(1+\beta)\left(\frac{y}{c'b}+\frac{\delta M}{3bc}\right).\end{aligned}$$

Applying this result, inequality **8**, twice the second, and finally the fourth inequality in **[10]** we get

$$\begin{aligned}&\frac{\phi(b,y)-c(1-\delta)\psi(r,z)}{(1-\delta)(1+\beta)}\\&>\frac{y}{b+\gamma}\left(\frac{1-\beta}{(1-\delta)(1+\beta)}-\frac{c}{c'}-\frac{c\gamma}{c'b}\right)-\frac{\delta M}{3b}\\&\geq\frac{y}{b+\gamma}\left(\delta-\frac{\delta}{2}\right)-\frac{\delta M}{3b}\\&\geq\frac{2}{3b}\frac{\delta}{2}M-\frac{\delta M}{3b}=0,\end{aligned}$$

proving **[11]**.

Let $\mathcal{N}=\{n\in\mathbb{N}:\ b_n/n\le 1-\eta\}$, $N_1=\lceil 2c\gamma/(c'\delta)\rceil$, and for every $n\in\mathcal{N}$ with $b_n\ge N_1$ define the events

$$F_n=\left\{\frac{Z_n}{r_n}<\frac{1}{c'}\frac{Y_n}{b_n}+\frac{d}{b_n\wedge r_n}\right\},\qquad G_n=\{Y_n>M\}.$$

For all $n\in\mathcal{N}$, $b_n<r_n/\eta$. Thus, if $n\in\mathcal{N}$ and $b_n\ge N_1$, the implication **[10]** $\Rightarrow$ **[11]** yields that

$$F_n\cap G_n\subset\{\phi(b_n,Y_n)>c(1-\delta)\psi(r_n,Z_n)\}.$$

Therefore, by inequalities **9**,

$$F_n\cap G_n\subset\{\hat{p}_{\bar{k}}(X^n)\ge c(1-\delta)\hat{q}_{\bar{k}}(X^n)\}.$$

It follows from *Lemma 1* and the definition of d that for every $\vartheta=(p,q)$ with $p_{\bar{k}}\ge cq_{\bar{k}}$, $\mathbb{P}_\vartheta(F_n^c)\le\frac{\epsilon}{2}$. By *Lemma 2*, there exists $N_2\in\mathbb{N}$ so that $\mathbb{P}_\vartheta(G_n^c)\le\frac{\epsilon}{2}$ for all n with $b_np_{\bar{k}}\ge N_2$. Thus, if $p_{\bar{k}}\ge cq_{\bar{k}}$, $n\in\mathcal{N}$, and $b_np_{\bar{k}}\ge N:=\max(N_1,N_2)$, then

$$\mathbb{P}_\vartheta(\hat{p}_{\bar{k}}(X^n)\ge c(1-\delta)\hat{q}_{\bar{k}}(X^n))\ge 1-\mathbb{P}_\vartheta(F_n^c)-\mathbb{P}_\vartheta(G_n^c)\ge 1-\epsilon.$$

Note that N does not depend on the sequence of the choices of the dice. □

Remark 5: If $K=2$, then for every $n\ge 1$ and every fixed number of times the red die is chosen in the first n periods, the Bayes estimate of $q_{\bar{k}}$ can be shown to be an increasing function of the number of times the red die lands on side $\bar{k}$. This fact can be combined with *Theorem 1* to give an alternative proof of *Theorem 2* for the case $K=2$. The monotonicity result does not hold for $K>2$ and our *Proof of Theorem 2* does not use *Theorem 1*.

Proof of Corollary 1: By Chebyshev's inequality, $[B_n-\mathbb{E}(B_n)]/n$ converges in probability to 0. Thus, by condition **6**, there exists $\eta>0$ and $N_1\in\mathbb{N}$ so that the event $F_n=\{\eta\le B_n/n\le 1-\eta\}$ has probability $\mathbb{P}(F_n)\ge 1-\frac{\epsilon}{2}$ for all $n\ge N_1$. By *Theorem 2*, there exists $N_2\in\mathbb{N}$ so that

$$\mathbb{P}_\vartheta(\hat{p}_{\bar{k}}(X^n)\ge c(1-\delta)\hat{q}_{\bar{k}}(X^n)|B_n=b_n)\ge 1-\frac{\epsilon}{2}$$

for all $\vartheta=(p,q)\in\Delta^2$ with $p_{\bar{k}}\ge cq_{\bar{k}}$ and all $n\in\mathbb{N}$ and $b_n\in\{1,\dots,n\}$ with $\mathbb{P}(B_n=b_n)>0$ and $b_np_{\bar{k}}\ge N_2$ and $b_n/n\le 1-\eta$. Let $N=\max(N_1,\lceil N_2/\eta\rceil)$. Then for every ϑ with $p_{\bar{k}}\ge cq_{\bar{k}}$ and every $n\in\mathbb{N}$ with $np_{\bar{k}}\ge N$, $F_n\subset\{B_np_{\bar{k}}\ge N_2, B_n/n\le 1-\eta\}$, so that

$$\mathbb{P}_\vartheta(\hat{p}_{\bar{k}}(X^n)\ge c(1-\delta)\hat{q}_{\bar{k}}(X^n)|F_n)\ge 1-\frac{\epsilon}{2},$$

which implies **[7]** because $\mathbb{P}(F_n)\ge 1-\frac{\epsilon}{2}$. □

ACKNOWLEDGMENTS. We thank three referees for many useful suggestions. We thank Gary Chamberlain, Martin Cripps, Ignacio Esponda, and Muhamet Yildiz for helpful conversations. This research is supported by National Science Foundation Grant SES 1558205.

1. US Food and Drug Administration (2000) *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials* (FDA, Rockville, MD), Tech Rep 2006D–0191.
2. Thompson LA (2014) *Bayesian Methods for Making Inferences about Rare Diseases in Pediatric Populations. Presentation at the Food and Drug Administration* (FDA, Rockville, MD).
3. Fudenberg D, He K (2017) Type-compatible equilibria in signalling games. arXiv: 1702.01819.
4. Diaconis P, Freedman D (1990) On the uniform consistency of Bayes estimates for multinomial probabilities. *Ann Stat* 18:1317–1327.
5. Bochkina NA, Green PJ (2014) The Bernstein–von Mises theorem and nonregular models. *Ann Stat* 42:1850–1878.
6. Dudley R, Haughton D (2002) Asymptotic normality with small relative errors of posterior probabilities of half-spaces. *Ann Stat* 30:1311–1344.