OPEN ACCESS

University of BRISTOL

University of Bristol - Explore Bristol Research

General rights

Meta-Analysis of Mid-*p*-Values: Some New Results based on the Convex Order

Patrick Rubin-Delanchy, Nicholas A. Heard & Daniel J. Lawson

Submit your article to this journal ↗

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

Meta-Analysis of Mid-*p*-Values: Some New Results based on the Convex Order

Patrick Rubin-Delanchy [a], Nicholas A. Heard [b], and Daniel J. Lawson [c]

[a]School of Mathematics, University of Bristol, Heilbronn Institute for Mathematical Research, Bristol, United Kingdom; [b]Department of Mathematics, Imperial College London, Heilbronn Institute for Mathematical Research, London, United Kingdom; [c]Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Bristol, United Kingdom

ABSTRACT

The mid-*p*-value is a proposed improvement on the ordinary *p*-value for the case where the test statistic is partially or completely discrete. In this case, the ordinary *p*-value is conservative, meaning that its null distribution is larger than a uniform distribution on the unit interval, in the usual stochastic order. The mid-*p*-value is not conservative. However, its null distribution is dominated by the uniform distribution in a different stochastic order, called the convex order. The property leads us to discover some new finite-sample and asymptotic bounds on functions of mid-*p*-values, which can be used to combine results from different hypothesis tests conservatively, yet more powerfully, using mid-*p*-values rather than *p*-values. Our methodology is demonstrated on real data from a cyber-security application.

1. Introduction

Let T be a real-valued test statistic, with probability measure P_0 under the null hypothesis, denoted H_0. Let X be a uniform random variable on the unit interval that is independent of T under P_0. X is a randomization device which is in practice usually generated by a computer.

We consider the (one-sided) *p*-value,

$$P = P_0(T^* \geq T), \qquad (1)$$

the mid-*p*-value (Lancaster 1952),

$$Q = \frac{1}{2}P_0(T^* \geq T) + \frac{1}{2}P_0(T^* > T), \qquad (2)$$

and the randomized *p*-value,

$$R = XP_0(T^* \geq T) + (1 - X)P_0(T^* > T), \qquad (3)$$

where T^* is a hypothetical independent replicate of T under P_0. If T is absolutely continuous under H_0, then the three quantities are equal and distributed uniformly on the unit interval. More generally, that is, if discrete components are possible, the three are different. Two main factors, one obvious and one more subtle, make this a very common occurrence. First, T is discrete if it is a function of discrete data, for example, a contingency table, categorical data, or a presence/absence event. Second, discrete test statistics often occur as a result of conditioning, as in the permutation test or Kendall's tau test (Sheskin 2003). Partially discrete tests occur, for example, as a result of censoring.

When P, Q, and R are not equal, it is a question which to choose. The ordinary *p*-value is often preferred in relatively strict hypothesis testing conditions, for example, in clinical trials, where the probability of rejecting the null hypothesis must not exceed the nominal level (often 5%). The randomized *p*-value has some theoretical advantages, for example, the nominal level of the test is met exactly. However, to quote one of its earliest proponents, "most people will find repugnant the idea of adding yet another random element to a result which is already subject to the errors of random sampling" (Stevens 1950). Randomized *p*-values also fail Birnbaum's admissibility criterion (Birnbaum 1954). Note that we can also work with an unrealized version of the randomized *p*-value, known as the *fuzzy* or *abstract* *p*-value (Geyer and Meeden 2005), and either stop there—leaving interpretation to the decision-maker—or propagate uncertainty through to any subsequent analysis, for example, multiple-testing (Kulinskaya and Lewin 2009; Habiger 2015).

Although it can allow breaches of the nominal level, the mid-*p*-value is often deemed to better represent evidence against the null hypothesis than the ordinary or randomized *p*-values. Justifications are not just heuristic as, for example, the mid-*p*-value can arise as a Rao–Blackwellization of the randomized *p*-value corresponding to the uniformly most powerful test (Wells 2010), as an optimal estimate of the H_0 versus H_1 truth indicator under squared loss (Hwang and Yang 2001), or from asymptotic Bayesian arguments (Routledge 1994). Performance has also been demonstrated in applications, for example, in the context of healthcare monitoring (Spiegelhalter et al. 2012) (an article read before the Royal Statistical Society), genetics (Graffelman and Moreno 2013), a wealth of examples involving contingency tables (Lydersen, Fagerland, and Laake 2009), and more. Our own interest stems from cyber-security applications, and a motivating example is given in Section 3. Most arguments for using the mid-*p*-value in hypothesis testing scenarios also

work for confidence intervals. Here, using the mid-p-value over the p-value can result in a smaller interval, with a closer-to-nominal coverage probability (Berry and Armitage 1995; Fagerland, Lydersen, and Laake 2015).

In this article, we are able to make further mathematical progress on the mid-p-value by using a stochastic order known as the *convex order*. The problem we focus on is meta-analysis, that is, combining evidence from different hypothesis tests into one, global measure of significance. In several scenarios analyzed, the use of the ordinary p-value leads to suboptimal, and even spurious results. New bounds for some commonly used methods for combining ordinary p-values are derived for mid-p-values. These allow large gains in power over using ordinary p-values, while, unlike any previous study based on mid-p-values, the false positive rate is still controlled exactly (albeit conservatively).

The remainder of this article is structured as follows. In Section 2, we summarize our main results. Section 3 gives a cyber-security application where, using mid-p-values, we are able to detect a cyber-attack that would likely fall under the radar if only ordinary p-values were used. Section 4 elaborates on the results of Section 3, with improved (although more complicated) bounds, simulations, and discussion. Section 5 concludes. All proofs are relegated to the Appendix.

2. Main Results

This section summarizes the main ideas and findings of the article. Let U denote a uniform random variable on the unit interval, with expectation operator E, and let E_0 denote expectation with respect to P_0. Under the null hypothesis, it is well known, see, for example, Casella and Berger (2002), that P dominates U in the *usual stochastic order*, denoted $P \geq_{st} U$. One way to write this is

$$E_0\{f(P)\} \geq E\{f(U)\}, \qquad (4)$$

for any nondecreasing function f, whenever the expectations exist (Shaked and Shanthikumar 2007). It is also well known, and in fact true by design, that R is uniformly distributed under the null hypothesis, denoted $R =_{st} U$. On the other hand, it is not widely known that, under the null hypothesis, Q is dominated by U in the *convex order*, denoted $Q \leq_{cx} U$. One way to write this is (Shaked and Shanthikumar 2007, chap. 3)

$$E_0\{h(Q)\} \leq E\{h(U)\}, \qquad (5)$$

for any *convex* function h, whenever the expectations exist. We have used the qualifier "widely," because an effective equivalent of Equation (5) can be found in Hwang and Yang (2001). However, even there, Equation (5) is not recognized as a major stochastic order, meaning that some of its importance is missed.

In particular, we now present three concrete, new results, made possible by the literature on the convex order. Each provides a method for combining mid-p-values conservatively, the first two in finite samples and the last asymptotically. Details and improved (but more complicated) bounds are given in Section 4. In what follows, $Q_1, \ldots, Q_n$ denote independent (but not necessarily identically distributed) mid-p-values, with joint probability measure denoted $\tilde{P}_0$ under the null hypothesis.

Let $\bar{Q}_n = n^{-1} \sum_{i=1}^n Q_i$ denote the average mid-p-value. For $t \geq 0$,

$$\tilde{P}_0 \left(1/2 - \bar{Q}_n \geq t\right) \leq \exp(-6nt^2). \qquad (6)$$

Note that, first, no knowledge of the specific individual mid-p-value distributions is required. Second, Hoeffding's inequality (Hoeffding 1963), which would be available more generally, gives the larger bound $\exp(-2nt^2)$ (the cubic root).

Let $F_n = -2 \sum_{i=1}^n \log(Q_i)$, known as Fisher's statistic (Fisher 1934) and the most popular method for combining p-values. In the continuous case, it is well-known that F_n has a chi-square distribution with $2n$ degrees of freedom under H_0. For $t \geq 2n$,

$$\tilde{P}_0(F_n \geq t) \leq \exp\{n - t/2 - n\log(2n/t)\}. \qquad (7)$$

Finally, assume additionally that $Q_1, \ldots, Q_n$ are identically distributed. Then applying Fisher's method as usual, that is, treating the mid-p-values as if they were ordinary p-values and using the chi-square tail, is asymptotically conservative as $n \to \infty$.

3. Example: Network Intrusion Detection

The perceived importance of cyber-security research has risen dramatically in recent years, particularly after several well-publicized events in 2016 and 2017. In this area, anomaly detection over very high volumes and rates of network data is a key statistical problem (Adams and Heard 2016). In our experience of the field, discrete data, whether they be presence/absence events, counts or categorical data, are the norm rather than the exception. We will demonstrate the value of our article's contributions in a network intrusion detection problem.

Figure 1 shows publically available authentication data covering 58 days on the Los Alamos National Laboratory computer network (Kent 2016). Nodes in the graph are computers, and an edge indicates that there was at least one connection from one computer to the other, resulting in a graph with $m \approx 18,000$ nodes and ~400,000 directed edges. An exciting opportunity offered by this data resource is that it contains an actual cyber-attack: or, to be precise, records of penetration testing activity conducted by a "red-team." One of the four computers used for the attack (the highest degree of the four, ID "C17693," with 296 out of 534 edges labeled as nefarious) is highlighted in red on the left, with its connections highlighted in pink on the right.

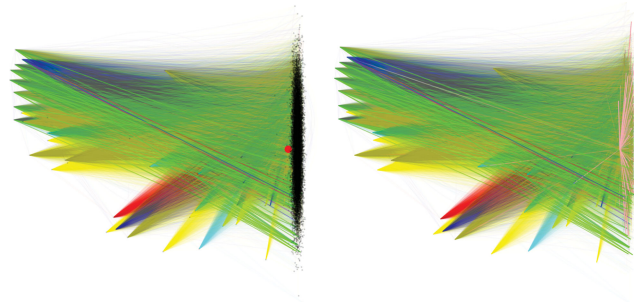**Figure 1.** Authentication data: Full network of connections comprising ~18,000 nodes and ~400,000 directed edges. Edges are colored by authentication type. On the left, nodes are shown as black points, with node ID "C17693" highlighted in red (and larger). On the right, the points are hidden to better see the connections made by node ID "C17693," which are now highlighted in pink.

Earlier work on network intrusion has suggested that the occurrence of *new edges* on the network can be (weakly) indicative of nefarious behavior (Neil et al. 2013; Neil 2015). Looking at the outward connections from a given computer, in particular, those which at the other end involve a computer otherwise receiving relatively few new connections present special interest. Because the first day of data has no red-team activity, we use this day to learn a rate λ_j, $j = 1, \ldots, m$ at which each computer receives new connections, assuming, admittedly unrealistically, that the times are right-censored independent and identically distributed exponential random variables. For every computer on the network, the set of outward new connections made over the remainder of the observation period [1, 58] is scored according to this model. The test-statistic

$$T_{ij} = \begin{cases} 57 & \text{if no connection occurs from } i \text{ to } j, \\ \tau - 1 & \text{if a new connection from } i \text{ to } j \text{ occurs} \\ & \qquad\qquad \text{at time } \tau \geq 1, \end{cases}$$

is considered for every directed pair (i, j) not occurring as an edge on the first day, so that each node i has associated with it a collection of test statistics $T_{i\cdot}$, which are partially discrete, with a point mass at 57.

For regularization purposes, the rates λ_j, $j = 1, \ldots, m$ are assumed a priori to follow a Gamma distribution matching the mean and variance of the empirical rates computed for each $j = 1, \ldots, m$ over the full period of 58 days. The use of this prior implies that before censoring T_{ij} has a Gamma-Exponential (also called Lomax) predictive distribution, which is used to compute the collection of ordinary, mid, and randomized p-values $P_{i\cdot}$, $Q_{i\cdot}$, $R_{i\cdot}$ corresponding to the outward connections of each node $i = 1, \ldots, m$. Mathematical details about the calculations above are in the Appendix.

Since we are interested in the *ranking* of computer ID "C17693" among the other $\sim$18,000 computers, as well as its p-value, it makes sense to extend the ranges of the bounds (6) and (7) as follows:

$$\tilde{P}_0 \left(1/2 - \bar{Q}_n \geq t \right) \leq \exp\{-6\,\mathrm{sgn}(t)nt^2\}, \qquad t \in \mathbb{R}, \qquad (8)$$

$$\tilde{P}_0(F_n \geq t) \leq \exp[\mathrm{sgn}(t - 2n)\{n - t/2 - n\log(2n/t)\}], \quad t > 0, \qquad (9)$$

which preserves monotonicity, and remains valid because larger values than unity are returned outside the old ranges. Our options are:

1. to compute the average ordinary, mid, and randomized p-values, and obtain a significance level using bound (8). Computer ID "C17693" then ranks as 8th (p-value ≈ 1), 8th (p-value $\approx 10^{-7}$), and 9th (p-value $\approx 10^{-7}$) most anomalous of the $\sim$18,000 computers, respectively.

2. to compute Fisher's statistic for the ordinary, mid, and randomized p-values, and obtain a global significance level using bound (9) for the second case, and the chi-square tail otherwise. Computer ID "C17693" now ranks joint 8118th (p-value ≈ 1), 2nd (p-value ≈ 1), and 9th (p-value $\approx 10^{-43}$), respectively.

3. to assume an asymptotic regime and use the chi-square tail for the Fisher-with-mid-p-values statistic instead. Computer ID "C17693" then ranks 8th (p-value ≈ 1).

As rankings go, therefore, the mid-p-value is never beaten, with computer ID "C17693" coming in the top 10 every time and coming second once. The most obvious approach of using Fisher's method with ordinary p-values fails completely. As for the other three red-team computers: using the best performing method, that is, Fisher's statistic with mid-p-values and bound (9), where Computer ID "C17693" comes second, their ranks are 384th (ID "C18025"), 550th (ID "C19932"), and 1079th (ID "C22409").

4. Meta-Analysis of Mid-*p*-Values: Further Details

This section elaborates on the results of Section 2. We say that a random variable (and its measure and distribution function) is *subuniform* if it is less variable than a uniform random variable, U, in the convex order.

To see why the mid-p-value is sub-uniform, notice that $Q = \mathrm{E}_0(R \mid T)$. By Jensen's inequality, for any convex function h,

$$\begin{aligned} \mathrm{E}_0\{h(Q)\} = \mathrm{E}_0[h\{\mathrm{E}_0(R \mid T)\}] &\leq \mathrm{E}_0[\mathrm{E}_0\{h(R) \mid T\}] \\ &= \mathrm{E}_0\{h(R)\} = \mathrm{E}\{h(U)\}, \end{aligned} \qquad (10)$$

whenever the expectations exist, since $R =_{st} U$. Remember that we do not claim this result is new, see, for example, Hwang and Yang (2001), but rather the idea to exploit the literature on the convex order.

To formalize the meta-analysis framework, let $T_1, \ldots, T_n$ be a sequence of independent test statistics. We consider a joint null hypothesis, $\tilde{H}_0$, under which $T_1, \ldots, T_n$ have probability measure $P_0^{(1)}, \ldots, P_0^{(n)}$, respectively. The p-values, P_i, mid-p-values, Q_i, and randomized p-values, R_i, are obtained by replacing P_0 with $P_0^{(i)}$ in (1), (2), and (3), respectively. In the case of the randomized p-value, an independent uniform variable, X_i, is generated each time. $\tilde{P}_0$ denotes the implied joint probability measure of the statistics under $\tilde{H}_0$. The focus of this section is on testing the joint null hypothesis $\tilde{H}_0$. Probability bounds that follow often have the form $\tilde{P}_0\{f(Q_1, \ldots, Q_n) \geq t\} \leq b_n(t)$. If the observed mid-$p$-values are $q_1, \ldots, q_n$ and level of the test is α (e.g., 5%), then a procedure that rejects when $b_n\{f(q_1, \ldots, q_n)\} \leq \alpha$ is conservative: the probability of rejecting the null hypothesis $\tilde{H}_0$, if it holds, does not exceed α.

4.1. Sums of Mid-*p*-Values

An early advocate of mid-p-values, Barnard (1989, 1990) proposed to combine test results from different contingency tables by taking the sum of standardized mid-p-values. His exposition relies on some approximations. Our results make exact inference possible.

We begin with a bound on the sum of independent mid-p-values. This bound bears an interesting resemblance to Hoeffding's inequality (Hoeffding 1963). It will later be extended to be relevant to Barnard's analysis.

Theorem 1. Let $X_1, \ldots, X_n$ denote n independent sub-uniform random variables with mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then, for $0 \leq t \leq 1/2$,

$$P\left(1/2 - \bar{X}_n \geq t\right) \leq \min_{h \geq 0}\left\{2e^{-ht}\sinh(h/2)/h\right\}^n, \qquad (11)$$

$$\leq \exp(-12nt^2)\{\sinh(6t)/(6t)\}^n, \quad (12)$$
$$\leq \exp(-6nt^2). \quad (13)$$

A subuniform random variable has expectation 1/2 and is bounded between 0 and 1. Hoeffding's inequality would therefore give us $P(1/2 - \bar{X}_n \geq t) \leq \exp(-2nt^2)$ for $0 \leq t \leq 1/2$, the cubic root. Our improvement is substantial, for example, suppose we observe an average of 0.4 from $n = 100$ mid-p-values. This is very significant: $\tilde{P}_0(1/2 - \bar{Q}_n \geq 0.1) \leq 0.0025$ using (13). However, we would only find $\tilde{P}_0(1/2 - \bar{Q}_n \geq 0.1) \leq 0.14$ using Hoeffding's inequality.

Instead of summing the mid-p-values directly, Barnard (1990) actually considered sums of the standardized statistics

$$D_i = (1/2 - Q_i)/\sigma_i,$$

where σ_i is the standard deviation of Q_i under $\tilde{H}_0$. The upper tail probability of the sum is then estimated by Gaussian approximation. In the purely discrete case, Barnard shows that $\sigma_i = \{(1 - s_i)/12\}^{1/2}$ where

$$s_i = \sum_{t \in S_i} \left\{ P_0^{(i)}(T_i = t) \right\}^3,$$

and S_i is the (countable) support of Q_i. Instead of appealing to the Gaussian approximation, the convex order allows us to find an exact bound.

Lemma 1. Let $X_1, \ldots, X_n$ denote n independent subuniform random variables with standard deviations $\sigma_1, \ldots, \sigma_n$, respectively, and let

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} (1/2 - X_i)/\sigma_i.$$

Then, for $t \geq 0$,

$$P(\bar{Y}_n \geq t) \leq \min_{h \geq 0} \left(\prod_{i=1}^{n} \exp[-h\{t + 1/(2\sigma_i)\}] \right.$$
$$\left. \times \left\{ \frac{e^{h/\sigma_i} - 1s}{h/\sigma_i} + h^2 \left(\frac{1}{2} - \frac{1}{24\sigma_i^2} \right) \right\} \right), \quad (14)$$
$$\leq \exp\{-6n(\bar{\sigma}t)^2\}, \quad (15)$$

where $\bar{\sigma} = (\prod \sigma_i)^{1/n}$ is the geometric mean of the standard deviations.

In practice, the bound (14), which is an important improvement over (15), can be found numerically by minimizing over h. Of course, even if the optimum cannot be determined exactly the obtained bound still holds, because the tail area is simply over-estimated.

To illustrate how the bound (14) performs in practice, we now revisit Barnard's example (Barnard 1990, p. 606). The first experiment he considers yields $Q_1 = 1/7$, $s_1 = 9002/42^3$, $D_1 = 1.32$. The second yields $Q_2 = 1/9$, $s_2 = 141/729$, $D_2 = 1.5$. Since the sum divided by $\sqrt{2}$ is almost two, that is, two standard deviations away, he finds "serious evidence" against the null hypothesis. Lemma 1 gives $\tilde{P}_0(D_1 + D_2 \geq 1.32 + 1.5) \leq 0.12$, providing some evidence in favor of the alternative, but not significant at, say, the 5% level. On the other hand, evidence would start to become compelling if we were to observe the second result again, $Q_3 = 1/9$, $s_3 = 141/729$, $D_3 = 1.5$; Lemma 1 then finds $\tilde{P}_0(D_1 + D_2 + D_3 \geq 1.32 + 1.5 + 1.5) \leq 0.036$.

4.2. Products of Mid-p-Values (Fisher's Method)

Fisher's method (Fisher 1934) is the most popular way of combining p-values. As is well-known, under $\tilde{H}_0$, the statistic $-2 \sum_{i=1}^{n} \log(P_i)$ has a chi-square distribution with $2n$ degrees of freedom if P_i are absolutely continuous. Therefore, the p-value of the combined test is $P^\dagger = S_{2n}\{-2 \sum_{i=1}^{n} \log(P_i)\}$, where S_k is the survival function of a chi-square distribution with k degrees of freedom. This results in an exact procedure when P_i are absolutely continuous, and a conservative one otherwise, that is, $P^\dagger \geq_{st} U$ under $\tilde{H}_0$.

Our next result allows us to use the mid-p-values $Q_1, \ldots, Q_n$ in place of $P_1, \ldots, P_n$ while retaining a conservative procedure. We were able to derive three probability bounds. None beats the other two uniformly for all n and all significance levels (see Figure 2), but the last is often the winner, hence the simpler statement of Section 2.

Theorem 2. Let $X_1, \ldots, X_n$ be a sequence of independent subuniform random variables. Then for $x \geq 2n$,

$$P\left(-2 \sum_{i=1}^{n} \log(X_i) \geq x \right)$$
$$\leq \min \left[S_{2m}(x - 2n \log 2), \, n \big/ \left[n + \{(x - 2n)/2\}^2 \right], \right.$$
$$\left. \exp\{n - x/2 - n \log(2n/x)\} \right] = u_n(x).$$
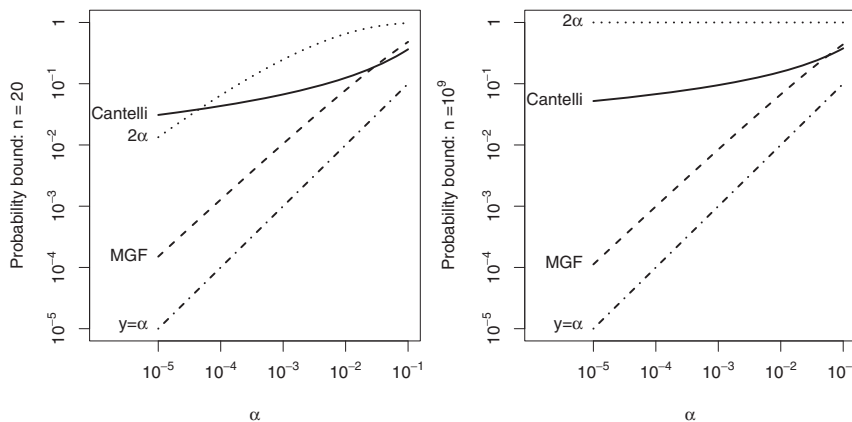
Figure 2. Comparison of the probability bounds given by Theorem 2 for Fisher's method using mid-p-values. Theorem 2 gives explicit formulas for 2α, Cantelli and MGF, in that order. Both axes are on the logarithmic scale.
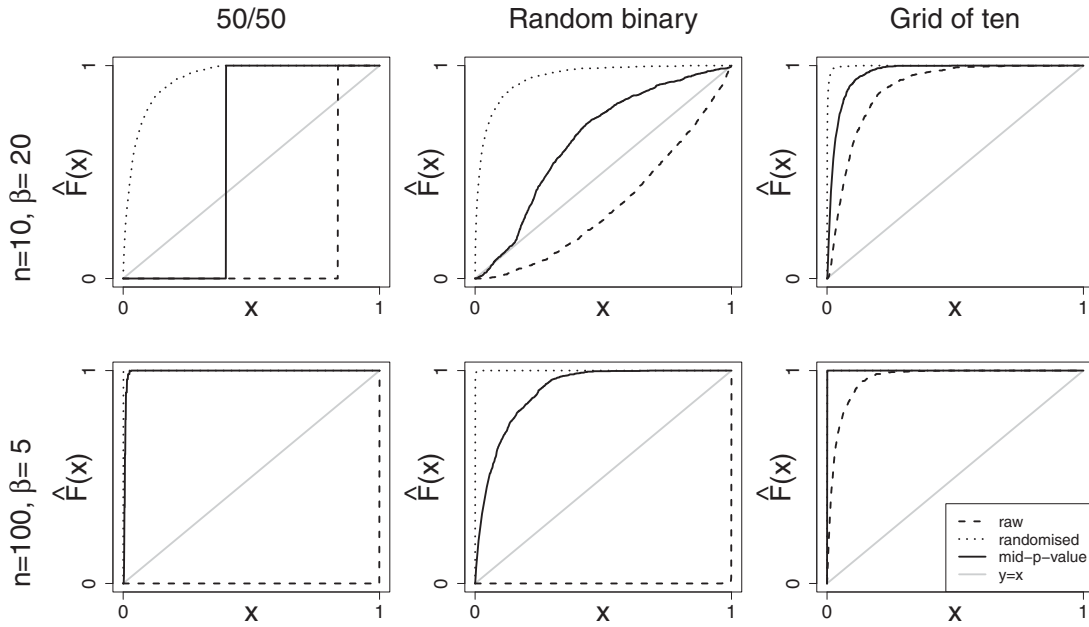
Figure 3. Fisher's method with discrete p-values. Empirical distribution functions of Fisher's combined p-value under different conditions. 50/50: Each p-value is equal to 1/2 or 1 (with probability 1/2 each under $\tilde{H}_0$). Random binary: Each p-value is equal to p or 1 (with probability p and $1 - p$, respectively, under $\tilde{H}_0$). p is drawn uniformly on [0, 1] (independently of whether $\tilde{H}_0$ or $\tilde{H}_1$ holds). Grid of 10: Each p-value is drawn from 1/10, 2/10 ..., 1 (with probability 1/10 each under $\tilde{H}_0$). $n = 10$, $\beta = 20$: 10 p-values from a left-censored Beta(1, 20) distribution. $n = 100$, $\beta = 5$: 100 p-values from a left-censored Beta(1, 5) distribution. Dotted line: Randomized p-values. Solid line: Mid-p-value. Dashed line: Standard p-values. Further details in the main text.

The first uses $P(X_i \leq \alpha) \leq 2\alpha$ for $\alpha \geq 0$, obvious in the case of a mid-p-value, but actually true of any subuniform random variable (Meng 1994). The second uses bounds on the mean and variance of $-\log(X_i)$ (given in Lemma 2, in the Appendix) and then applies the Chebyshev-Cantelli inequality. The third is based on a bound on the moment generating function of $-\log(X_i)$. Derivation details are in the Appendix.

For a given n and $\alpha \in (0, 1]$, let $t_{\alpha,n}$ denote the critical value of Fisher's statistic, that is, $t_{\alpha,n}$ satisfies $S_{2n}(t_{\alpha,n}) = \alpha$. Figure 2 presents the behavior of the different bounds for different n (20 on the left and 1 billion on the right) and α. The curves show the bound given by each formula at different α (which can be interpreted as "canonical levels"), that is, inputting $x = t_{\alpha,n}$ in Theorem 2, as α ranges from 10^{-5} to 0.1. For low α, the bound based on the moment generating function, marked MGF, is by far superior.

Let $Q^\dagger = u_n\{-2\sum_{i=1}^n \log(Q_i)\}$. Then $Q^\dagger$ is again conservative, that is, $Q^\dagger \geq_{st} U$ under $\tilde{H}_0$. Both $P^\dagger$ and $Q^\dagger$ are valid p-values. Clearly, if the underlying p-values are continuous, then the standard $P^\dagger$ is superior (in fact, deterministically smaller). However, $Q^\dagger$ seems to be substantially more powerful in a wide range of discrete cases. This is demonstrated by simulation in Section 4.3.

Finally, we find this interesting asymptotic result.

Theorem 3 (Fisher's method is asymptotically conservative). Let $X_1, X_2, \ldots$ denote independent and identically distributed subuniform random variables. For any $\alpha \in (0, 1]$, there exists $N \in \mathbb{N}$ such that

$$P\left(-2\sum_{i=1}^n \log(X_i) \geq t_{\alpha,n}\right) \leq \alpha,$$

for any $n \geq N$.

Hence, we can dispense with any correction entirely if n is large enough and the Q_i are identically distributed. A formal proof is given in the Appendix. Since $E\{-\log(X_i)\} \leq E\{-\log(U)\}$, from the definition of the convex order, a direct application of the law of large numbers gets us most of way, except for the possibility $E\{-\log(X_i)\} = E\{-\log(U)\}$. In fact, this exception is no problem because, perhaps surprisingly, it implies that the X_i are uniform, using Shaked and Shanthikumar (2007, Theorem 3.A.43).

4.3. Simulations

To illustrate the potential improvement of employing Fisher's method with mid-p-values, using the bound (7), over the traditional approach of using ordinary p-values and the chi-square tail, we considered p-values from three types of support. In the first column of Figure 3, each p-value P_i can only take one of two values, 1/2 and 1. We therefore have $Q_i = 0.25$ if $P_i = 1/2$ and $Q_i = 0.75$ if $P_i = 1$. Under the null hypothesis, $P_0^{(i)}(P_i = 1/2) = P_0^{(i)}(P_i = 1) = 1/2$. In the second column, each p-value P_i is supported on the pair $\{p_i, 1\}$, where p_i were drawn uniformly on the unit interval but are subsequently treated as fixed known values. We have $Q_i = p_i/2$ if $P_i = p_i$ and $Q_i = (1 + p_i)/2$ otherwise. Under the null hypothesis, we have $P_0^{(i)}(P_i = p_i) = 1 - P_0^{(i)}(P_i = 1) = p_i$, for each i. Finally, in the third column each p-value P_i takes one of 10 values, $1/10, 2/10, \ldots, 1$, and therefore $Q_i = P_i - 1/20$. Under the null hypothesis, $P_0^{(i)}(P_i = j/10) = 1/10$, for $j = 1, \ldots, 10$. The rows represent two different alternatives and sample sizes. In both cases, the P_i are generated by left-censoring a sequence of independent and identically distributed Beta variables, $B_1, \ldots, B_n$, that is, P_i is the smallest supported value larger than B_i. In the first scenario, the dataset is small ($n = 10$), but the

signal is strong (a Beta distribution with parameters 1 and 20). In the second, the dataset is larger ($n = 100$) but the signal is made weaker accordingly (a Beta distribution with parameters 1 and 5). Comparing just the solid and dashed lines first, we see that $Q^\dagger$ always outperforms $P^\dagger$ substantially, and sometimes overwhelmingly. In the bottom-left corner, for example, we have a situation where, at a false positive rate set to 5% say, the test $Q^\dagger$ would detect the effect with probability close to one whereas with $P^\dagger$ the probability would be close to zero.

As a final possibility, consider $R^\dagger = S_{2n}\{-2 \sum_{i=1}^{n} \log(R_i)\}$. A disappointment is that this randomized version, the dotted line in Figure 3, tends to outperform even the mid-p-values, and by a substantial margin. On the other hand, as pointed out in the introduction, the randomized p-value has some important philosophical disadvantages, and did not perform better in our real data example.

5. Conclusion

The convex order provides a formal platform for the treatment and interpretation of mid-p-values. This article used mathematical results from this literature to combine mid-p-values, which are not conservative individually, into an overall significance level that is conservative. As shown in real data and simulations, the gains in power can be substantial.

Whereas the focus of this article was on meta-analysis, another canonical problem is multiple testing, where the task is to subselect from or adjust a set of p-values, for example, subject to a maximum false discovery rate (Benjamini and Hochberg 1995). The case of discrete data has been analyzed in a number of articles, including Kulinskaya and Lewin (2009); Habiger and Pena (2011); Liang (2016); Habiger (2015). A promising (but ostensibly harder) avenue of research would be to investigate the use of the convex order in this problem.

Appendix: Mathematical Details and Proofs

A1. Section 3: Mathematical Details

First, we calculate the empirical rates

$$r_j = \frac{\text{number of new connections to } j \text{ over } [0, 58]}{(\text{number of nodes - 1}) \times 58},$$

for $j = 1, \ldots, m$, and then set $\alpha = \{\text{mean}(r_j)\}^2/\text{var}(r_j)$, $\beta = \text{mean}(r_j)/\text{var}(r_j)$, so that a Gamma distribution with parameters α and β has mean and variance equal to the empirical mean and variance of r_j, respectively. This distribution is used as the prior for each rate λ_j. After a day of observation, the posterior distribution for λ_j is also Gamma, with parameters

$$\alpha_j = \alpha + \#\{\text{new connections to } j \text{ over } [0, 1)\}; \quad \beta_j = \beta + \sum \tau_{ij},$$

where

$$\tau_{ij} = \begin{cases} 1 & \text{if no connection occurs from } i \text{ to } j \text{ in } [0, 1), \\ \tau & \text{if a new connection from } i \text{ to } j \text{ occurs at time } \tau < 1. \end{cases}$$

For each i, restrict j to the indices of nodes that did not receive a connection from i over $[0, 1)$. Conditional on the observation period $[0, 1)$, each

statistic T_{ij} has a probability measure, denoted $\mathrm{P}_0^{(ij)}$, with a point mass at 57 and an absolutely continuous distribution over $[0, 57)$ given by

$$\mathrm{P}_0^{(ij)}(T_{ij} \le t) = 1 - \int_0^\infty e^{-\lambda_j t} \cdot \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \lambda_j^{\alpha_j - 1} e^{-\beta_j \lambda_j} d\lambda_j$$

$$= 1 - \left(1 + \frac{t}{\beta_j}\right)^{-\alpha_j}, \quad \text{for } t \in [0, 57),$$

$$\mathrm{P}_0^{(ij)}(T_{ij} = 57) = (1 + 57/\beta_j)^{-\alpha_j}.$$

Since low values of T_{ij} are of interest, the p-value and mid-p-value associated with T_{ij} are computed from lower-tailed versions of (1) and (2) by substituting $\mathrm{P}_0^{(ij)}$ in for P_0, giving

$$P_{ij} = \begin{cases} 1, & T_{ij} = 57, \\ 1 - (1 - T_{ij}/\beta_j)^{-\alpha_j}, & T_{ij} < 57, \end{cases}$$

and

$$Q_{ij} = \begin{cases} 1/2 + \{1 - (1 - 57/\beta_j)^{-\alpha_j}\}/2, & T_{ij} = 57, \\ 1 - (1 - T_{ij}/\beta_j)^{-\alpha_j}, & T_{ij} < 57, \end{cases}$$

respectively.

Proofs

Proof of Theorem 1. Since $1 - X$ is subuniform if and only if X is subuniform, it is sufficient to prove the bounds in (11), (12), and (13) hold for $\mathrm{P}(\bar{X}_n - 1/2 \ge t)$. Since $\exp(xh)$ is a convex function in x for any h, the convex order gives us $\mathrm{E}\{\exp(hX_i)\} \le \mathrm{E}\{\exp(hU)\} = (e^h - 1)/h$. Therefore, for any $h \ge 0$,

$$\mathrm{P}\left(\bar{X}_n - 1/2 \ge t\right) = \mathrm{P}\left[\exp\left(\sum_{i=1}^n hX_i\right) \ge \exp\{nh(t + 1/2)\}\right],$$

$$\le \exp\{-nh(t + 1/2)\}\mathrm{E}\left\{\exp\left(\sum_{i=1}^n hX_i\right)\right\},$$

$$\le \exp\{-nh(t + 1/2)\}\{(e^h - 1)/h\}^n$$

$$= \left\{2e^{-ht}\sinh(h/2)/h\right\}^n,$$

where the second line follows from Markov's inequality. The choice $h = 12t$ (motivated by an analysis of the Taylor expansion in h at 0) leads to

$$\mathrm{P}\left(\bar{X}_n - 1/2 \ge t\right) \le \exp(-12nt^2)\{\sinh(6t)/(6t)\}^n$$

$$\le \exp(-6nt^2)\left\{e^{-6t}\sinh(6t)/(6t)\right\}^n \le \exp(-6nt^2),$$

using the fact that $e^{-x}\sinh(x)/x = (1 - e^{-2x})/(2x)$ is one at $x = 0$ (using l'Hospital's rule) and decreasing. $\square$

Proof of Lemma 1. Again, we will prove the bound holds for $W_n = n^{-1}\sum(X_i - 1/2)/\sigma_i$, so that the theorem holds by symmetry. For any $h \ge 0$,

$$\mathrm{E}\{\exp(hX_i/\sigma_i)\} = 1 + \mathrm{E}(hX_i/\sigma_i) + \mathrm{E}\left\{(hX_i/\sigma_i)^2\right\}/2 + \cdots$$

$$= 1 + \mathrm{E}(hU/\sigma_i) + h^2\left(\frac{1}{2} + \frac{1}{8\sigma_i^2}\right) + \cdots$$

$$\le \mathrm{E}\{\exp(hU/\sigma_i)\} + h^2\left(\frac{1}{2} + \frac{1}{8\sigma_i^2} - \frac{1}{6\sigma_i^2}\right),$$

because $E\{(hX_i/\sigma_i)^n\} \leq E\{(hU/\sigma_i)^n\}$ for $n \geq 3$, by the convex order, and $E\{(U/\sigma_i)^2\}/2 = 1/(6\sigma_i^2)$. Therefore,

$$P(W_n \geq t) = P\left[\exp\left\{\sum_{i=1}^n h(X_i - 1/2)/\sigma_i\right\} \geq e^{hnt}\right],$$

$$\leq e^{-hnt}E\left[\exp\left\{\sum_{i=1}^n h(X_i - 1/2)/\sigma_i\right\}\right],$$

$$= \prod_{i=1}^n \exp[-h\{t + 1/(2\sigma_i)\}]\left\{\frac{e^{h/\sigma_i} - 1}{h/\sigma_i} + h^2\left(\frac{1}{2} - \frac{1}{24\sigma_i^2}\right)\right\},$$

proving that (14) holds. Next, since $\sigma_i^2 \leq 1/12$,

$$P(W_n \geq t) \leq \prod_{i=1}^n \exp[-h\{t + 1/(2\sigma_i)\}]\left(\frac{e^{h/\sigma_i} - 1}{h/\sigma_i}\right)$$

$$= \left(2e^{-ht}\left[\prod_{i=1}^n \sin h\{h/(2\sigma_i)\}\right]^{1/n} \Big/ (h/\bar\sigma)\right)^n$$

$$\leq \left\{2e^{-ht}\sin h(h/(2\bar\sigma))/(h/\bar\sigma)\right\}^n,$$

using the fact that the function sinh is geometrically convex on $[0, \infty)$ (Niculescu 2000). We proceed as in the proof of Theorem 1, choosing $h = 12\bar\sigma t$. □

The proofs of Theorems 2 and 3 both need the following result.

Lemma 2. Let X be a sub-uniform random variable. Then either (i) X is uniform on $[0, 1]$ or (ii)

$$E\{-\log(X)\} < E\{-\log(U)\} = 1; \quad \text{var}\{-\log(X)\} < \text{var}\{-\log(U)\} = 1,$$

where U is a uniform random variable on $[0, 1]$.

Proof. Shaked and Shanthikumar (2007, Theorem 3.A.43) provide the following theorem. If $X \leq_{cx} Y$ and for some strictly convex function h we have $E\{h(X)\} = E\{h(Y)\}$ then X is distributed as Y. The function $-\log(x)$ is strictly convex, therefore either X is uniform or $E\{-\log(X)\} < E\{-\log(U)\}$. If the latter is true, then

$$\text{var}\{-\log(X)\} = E[-\log(X) - E\{-\log(X)\}]^2$$
$$< E[-\log(X) - E\{-\log(U)\}]^2$$
$$\leq E\{\log(U) + 1\}^2$$
$$= \text{var}\{-\log(U)\}.$$

In the second line, the fact that the expected squared distance from the mean is smaller than from any other point is used, and in the fourth we used the fact that $(\log(x) + 1)^2$ is convex. □

Proof of Theorem 2. Let $G_n = -2\sum\log(X_i)$. Now, $U_i/2 \leq_{st} X_i$, for $i = 1, \ldots, n$, where $U_1, \ldots, U_n$ are independent uniform random variables on $[0, 1]$. This implies $-\log(X_i) \leq_{st} -\log(U_i/2)$. Because the usual stochastic order is closed under convolution (Shaked and Shanthikumar 2007, Theorem 1.A.3), we have $G_n \leq_{st} -2\sum\log(U_i) + 2n\log 2$. The sum $-2\sum\log(U_i)$ has a chi-square distribution with $2n$ degrees of freedom, proving the first bound.

Lemma 2 implies $E(G_n) \leq 2n$ and $\text{var}(G_n) \leq 4n$. Therefore, using Cantelli's inequality,

$$P[G_n \geq x] \leq \text{var}(G_n)/\left[\text{var}(G_n) + \{x - E(G_n)\}^2\right]$$
$$\leq \text{var}(G_n)/\left[\text{var}(G_n) + \{x - 2n\}^2\right]$$
$$\leq n/\left[n + \{(x - 2n)/2\}^2\right],$$

for $x \geq 2n$. This proves the second bound. Finally, the moment generating function of G_n is $E\{\exp(tG_n)\} = \prod E(X_i^{-2t})$ for $t \geq 0$. For $t \in [0, 1/2)$ each

$E(X_i^{-2t}) \leq E(U^{-2t}) = (1 - 2t)^{-1}$ since x^{-2t} is a convex function in x for $x \in [0, 1]$. Using Markov's inequality,

$$P(G_n \geq x) = P\{\exp(tG_n) \geq \exp(tx)\}$$
$$\leq \exp(-tx)E\{\exp(tG_n)\}$$
$$\leq \exp(-tx - n\log(1 - 2t)),$$

for $t \in [0, 1/2)$. The minimum of this function is at $t = 1/2 - n/x$, giving the third bound. □

Proof of Theorem 3. Let $V_i = -2\log(X_i)$, $\mu_V = E(V_i)$, $W_i = -2\log(U_i)$, where $U_1, \ldots, U_n$ are independent uniform random variables on $[0, 1]$, and $\mu_W = E(W_i)$. If $\mu_V = \mu_W$ then by Lemma 2 the X_i are uniform on $[0, 1]$ and we are done. The statement is also true if $\alpha = 1$. Therefore assume $\mu_V < \mu_W, \alpha \in (0, 1)$ and let $t \in (\mu_V, \mu_W)$. By the weak law of large numbers there exists an $N' \in \mathbb{N}$ such that, for $n \geq N'$,

$$P\left(\sum_{i=1}^n W_i \geq nt\right) \geq \alpha,$$

so that $t_{\alpha,n} \geq nt$. Therefore, for $n \geq N'$,

$$P\left(\sum_{i=1}^n V_i \geq t_{\alpha,n}\right) \leq P\left(\sum_{i=1}^n V_i \geq nt\right).$$

Again by the law of large numbers, the right-hand side tends to zero. Hence, there exists an $N \geq N'$ such that it is bounded by α for $n \geq N$. □

ORCID

Patrick Rubin-Delanchy http://orcid.org/0000-0003-0577-0795
Nicholas A. Heard http://orcid.org/0000-0002-8767-0810
Daniel J. Lawson http://orcid.org/0000-0002-5311-6213

References

Adams, N., and Heard, N. (2016), *Dynamic Networks and Cyber-Security* (Vol. 1), London, UK: World Scientific. [2]

Barnard, G. (1989), "On Alleged Gains in Power from Lower p-Values," *Statistics in Medicine*, 8, 1469–1477. [3]

—— (1990), "Must Clinical Trials be Large? The Interpretation of p-values and The Combination of Test Results," *Statistics in Medicine*, 9, 601–614. [3,4]

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, 57, 289–300. [6]

Berry, G., and Armitage, P. (1995), "Mid-P Confidence Intervals: A Brief Review," *The Statistician*, 44, 417–423. [2]

Birnbaum, A. (1954), "Combining Independent Tests of Significance," *Journal of the American Statistical Association*, 49, 559–574. [1]

Casella, G., and Berger, R. L. (2002), *Statistical Inference* (Vol. 2), Pacific Grove, CA: Duxbury. [2]

Fagerland, M. W., Lydersen, S., and Laake, P. (2015), "Recommended Confidence Intervals for Two Independent Binomial Proportions," *Statistical Methods in Medical Research*, 24, 224–254. [2]

Fisher, R. A. (1934), *Statistical Methods for Research Workers* (4th ed.), Edinburgh: Oliver & Boyd. [2,4]

Geyer, C. J., and Meeden, G. D. (2005), "Fuzzy and Randomized Confidence Intervals and p-Values," *Statistical Science*, 20, 358–366. [1]

Graffelman, J., and Moreno, V. (2013), "The Mid p-Value in Exact Tests for Hardy-Weinberg Equilibrium," *Statistical Applications in Genetics and Molecular Biology*, 12, 433–448. [1]

Habiger, J. D. (2015), "Multiple Test Functions and Adjusted p-Values for Test Statistics with Discrete Distributions," *Journal of Statistical Planning and Inference*, 167, 1–13. [1,6]

Habiger, J. D., and Pena, E. A. (2011), "Randomised p-Values and Nonparametric Procedures in Multiple Testing," *Journal of Nonparametric Statistics*, 23, 583–604. [6]

Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30. [2,3]

Hwang, J. G., and Yang, M.-C. (2001), "An Optimality Theory for Mid p-Values in 2 x 2 Contingency Tables," *Statistica Sinica*, 11, 807–826. [1,2,3]

Kent, A. D. (2016), "Cybersecurity Data Sources for Dynamic Network Research," in *Dynamic Networks and Cybersecurity*, eds. N. Adams and N. Heard, London, UK: World Scientific, pp. 37–66. [2]

Kulinskaya, E., and Lewin, A. (2009), "On Fuzzy Familywise Error Rate and False Discovery Rate Procedures for Discrete Distributions," *Biometrika*, 96, 201–211. [1,6]

Lancaster, H. (1952), "Statistical Control of Counting Experiments," *Biometrika*, 39, 419–422. [1]

Liang, K. (2016), "False Discovery Rate Estimation for Large-Scale Homogeneous Discrete p-Values," *Biometrics*, 72, 639–648. [6]

Lydersen, S., Fagerland, M. W., and Laake, P. (2009), "Recommended Tests for Association in 2 × 2 Tables," *Statistics in Medicine*, 28, 1159–1175. [1]

Meng, X.-L. (1994), "Posterior Predictive p-Values," *The Annals of Statistics*, 22, 1142–1160. [5]

Neil, J., Uphoff, B., Hash, C., and Storlie, C. (2013), "Towards Improved Detection of Attackers in Computer Networks: New Edges, Fast Updating, and Host Agents," in *6th International Symposium on Resilient Control Systems (ISRCS), 2013*, IEEE, pp. 218–224. [3]

Neil, J. C. (2015), "Using New Edges for Anomaly Detection in Computer Networks," US Patent 9,038,180. [3]

Niculescu, C. P. (2000), "Convexity According to the Geometric Mean," *Journal of Mathematical Inequalities*, 3, 155–167. [7]

Routledge, R. (1994), "Practicing Safe Statistics with the mid-p," *Canadian Journal of Statistics*, 22, 103–110. [1]

Shaked, M., and Shanthikumar, J. G. (2007), *Stochastic Orders*, New York: Springer. [2,5,7]

Sheskin, D. J. (2003), *Handbook of Parametric and Nonparametric Statistical Procedures*, Boca Raton, FL: Chapman and Hall/CRC Press. [1]

Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., and Grigg, O. (2012), "Statistical Methods for Healthcare Regulation: Rating, Screening and Surveillance," *Journal of the Royal Statistical Society*, Series A, 175, 1–47. [1]

Stevens, W. (1950), "Fiducial Limits of the Parameter of a Discontinuous Distribution," *Biometrika*, 37, 117–129. [1]

Wells, M. T. (2010), "Optimality Results for mid p–values," in *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, eds. J. O. Berger, T. T. Cai, and I. M. Johnstone, Beachwood, OH: Institute of Mathematical Statistics, pp. 184–198. [1]