Biometrics 71, 146–156
March 2015

DOI: 10.1111/biom.12231

Combining One-Sample Confidence Procedures for Inference in the Two-Sample Case

Michael P. Fay,* Michael A. Proschan, and Erica Brittain

National Institute of Allergy and Infectious Diseases, 6700B Rockledge Dr. MSC 7630, Bethesda, Maryland, 20892-7630, U.S.A.
**email:* mfay@niaid.nih.gov

Summary. We present a simple general method for combining two one-sample confidence procedures to obtain inferences in the two-sample problem. Some applications give striking connections to established methods; for example, combining exact binomial confidence procedures gives new confidence intervals on the difference or ratio of proportions that match inferences using Fisher's exact test, and numeric studies show the associated confidence intervals bound the type I error rate. Combining exact one-sample Poisson confidence procedures recreates standard confidence intervals on the ratio, and introduces new ones for the difference. Combining confidence procedures associated with one-sample t-tests recreates the Behrens–Fisher intervals. Other applications provide new confidence intervals with fewer assumptions than previously needed. For example, the method creates new confidence intervals on the difference in medians that do not require shift and continuity assumptions. We create a new confidence interval for the difference between two survival distributions at a fixed time point when there is independent censoring by combining the recently developed beta product confidence procedure for each single sample. The resulting interval is designed to guarantee coverage regardless of sample size or censoring distribution, and produces equivalent inferences to Fisher's exact test when there is no censoring. We show theoretically that when combining intervals asymptotically equivalent to normal intervals, our method has asymptotically accurate coverage. Importantly, all situations studied suggest guaranteed nominal coverage for our new interval whenever the original confidence procedures themselves guarantee coverage.

Key words: Behrens–Fisher problem; Confidence distributions; Difference in medians; Exact confidence interval; Fisher's exact test; Kaplan–Meier estimator.

1. Introduction

We propose a simple procedure to create a confidence interval (CI) for certain functions (e.g., the difference or the ratio) of two scalar parameters from each of two independent samples. The procedure only requires nested confidence intervals from the independent samples and certain monotonicity constraints on the function, and can be applied quite generally. We call our new CIs "melded confidence intervals" since they meld together the CIs from each of the two-samples. In this article we focus on melded CIs that are created from two one-sample CIs with guaranteed nominal coverage, and we conjecture that the resulting CIs themselves guarantee coverage. This conjecture is supported by simulated, numerical, and mathematical results.

The melded CI method is closely related to methods that have expanded or modified fiducial inference yet focus on frequentist properties, such as confidence structures (Balch, 2012), generalized fiducial inference (Hannig, 2009), inferential models (Martin and Liu, 2013), or confidence distributions (Xie and Singh, 2013). The melded CIs are much simpler to describe than the first three methods mentioned, and, unlike confidence distributions, can be applied to small sample discrete problems.

Fiducial inference is no longer part of mainstream statistics (for more background see Pedersen, 1978; Zabell, 1992; Hannig, 2009); nevertheless, it will be helpful to briefly describe some examples of fiducial inference and some of its shortcomings to show how the melded CIs relate to it and avoid those shortcomings. Unlike frequentist inference where parameters are fixed, or Bayesian inference where parameters are random, fiducial inference is not clearly in either camp, and hence has been the source of much confusion. Fiducial inference is a way of conditioning on the data and getting a distribution on the parameter without using a prior distribution. For example, if x is an observation drawn from a normal distribution with mean μ and variance 1 (i.e., $N(\mu, 1)$), then the corresponding fiducial distribution for μ is $N(x, 1)$. The middle 95% of that fiducial distribution is the usual 95% confidence interval for μ. The problem is that the fiducial distribution cannot be used to get confidence intervals on non-monotonic transformations of the parameter. For example, using the fiducial distribution for μ of $N(x, 1)$ as above, the corresponding distribution for μ^2 is a non-central chi square. A fiducial approach to creating in a one-sided 95% lower confidence limit for μ^2 is to take the 5% percentile of that non-central chi square distribution, but this does not work well; when $\mu = 0.1$, the coverage is only about 66% (see Pedersen, 1978, pp. 153–155). Another complication is that for discrete data such as a binomial observation, there are two fiducial distributions associated with the parameter, one can be used for obtaining the lower confidence limit and one for the upper limit (Stevens, 1950).

The melded CI approach is similar to the fiducial approach in that without using priors we associate probability distributions with parameters after conditioning on the data. But those probability distributions are only tools used to obtain the melded CIs and need not be interpreted as fiducial probabilities; all statistical theory in this article is firmly frequentist. The melded CI approach avoids the problems of fiducial inference two ways. First, we create distributions for parameters using only nested (defined in Section 2) one-sample confidence intervals, whose theory is well developed and understood. This seamlessly creates either one distribution (e.g., in the normal case) or two distributions (e.g., in the binomial case) as needed. Second, we limit the application to functions of the parameters that meet some monotonicity constraints, so that when the one-sample CIs have guaranteed coverage, the resulting melded CIs appear to also have guaranteed coverage.

Besides motivating some classical CIs and creating new CIs for these canonical examples, the melded CI method is a very general tool that can easily be used in essentially any complex two-sample inference setting, as long as there is an established approach for computing confidence intervals for a single sample. As an example of a new CI consider the difference in medians. Existing methods require continuity or shift assumptions (the Hodges and Lehmann (1963) intervals) or large samples (the nonparametric bootstrap). A melded CI for this situation inverts the sign test, and requires none of those assumptions. Simulations show that, unlike the Hodges and Lehmann (1963) intervals or nonparametric bootstrap intervals, the melded confidence intervals retain nominal coverage in all cases studied including discrete cases, non-shift cases, and small sample cases.

Here is an outline of the article. First, we define the procedure in Section 2. Then, we motivate the melded CIs for a simple example in Section 3, giving intuition for why it appears to retain the type I error rate. Section 4 gives more general mathematical results. The heart of the article shows the applications, with connections to well-known tests for simple cases and new tests and confidence intervals for less simple cases. In Sections 5–7, we discuss the melded CIs applied to the normal, binomial, and Poisson problems, respectively. In Section 8 we study a nonparametric melded CI for the difference in medians. In Section 9 we explore the application to the difference in survival distributions. In Section 10 we discuss the relationship with the confidence distribution approach, and we end with a short discussion.

2. The Melded Confidence Interval Procedure

Suppose we have two independent samples, where for the ith sample, $\mathbf{x}_i$ is the data vector, and $\mathbf{X}_i$ is the associated random variable whose distribution depends on a scalar parameter θ_i and possibly other nuisance parameters. Let the nested $100q\%$ one-sided lower and upper confidence limits for θ_i be $L_{\theta_i}(\mathbf{x}_i, q)$ and $U_{\theta_i}(\mathbf{x}_i, q)$, respectively. By nested we mean that if $q_1 < q_2$ then $L_{\theta_i}(\mathbf{x}_i, q_2) \leq L_{\theta_i}(\mathbf{x}_i, q_1)$ and $U_{\theta_i}(\mathbf{x}_i, q_1) \leq U_{\theta_i}(\mathbf{x}_i, q_2)$. We limit our application to functions of the parameters, written $g(\theta_1, \theta_2)$, that are, loosely speaking, decreasing in θ_1 among all allowable values of θ_2 and increasing in θ_2 among all allowable values of θ_1 (see Supplementary Material Section A for a precise statement of the monotonicity constraints). We consider three examples for $g(\cdot, \cdot)$ in this article: the difference, $g(\theta_1, \theta_2) = \theta_2 - \theta_1$; the ratio, $g(\theta_1, \theta_2) = \theta_2/\theta_1$, which can be used if the parameter space for θ_i is $\theta_i \geq 0$ for $i = 1, 2$; and the odds ratio, $g(\theta_1, \theta_2) = \left\{\theta_2(1-\theta_1)\right\} / \left\{\theta_1(1-\theta_2)\right\}$, which can be used if the parameter space for θ_i is $0 \leq \theta_i \leq 1$ for $i = 1, 2$. Note this is a crucial restriction because the coverage properties of the method may not hold if the monotonicity constraints on g are violated (see Section 10). Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]$. Then the $100(1-\alpha)\%$ lower and upper one-sided melded confidence limits for $\beta = g(\theta_1, \theta_2)$ are

$$L_\beta(\mathbf{x}, 1-\alpha) = \text{the } \alpha\text{th quantile of } g\{U_{\theta_1}(\mathbf{x}_1, A), L_{\theta_2}(\mathbf{x}_2, B)\}, \tag{1}$$

and

$$U_\beta(\mathbf{x}, 1-\alpha) \\ = \text{the } (1-\alpha)\text{th quantile of } g\{L_{\theta_1}(\mathbf{x}_1, A), U_{\theta_2}(\mathbf{x}_2, B)\}, \tag{2}$$

where here and throughout the article, A and B are independent and uniform random variables. The melded CIs can be calculated by Monte Carlo simulation or numeric integration. For the examples in this article, we used numeric integration. See Section 6 for a worked example.

We can invert the confidence intervals to give p-values associated with the corresponding series of hypothesis tests. For example, $p_L(\mathbf{x}, \beta_0) = \inf\left\{p : L_\beta(\mathbf{x}, 1-p) > \beta_0\right\}$ is the corresponding p-value for testing the null hypothesis $H_0 : g(\theta_1, \theta_2) \leq \beta_0$. The p-values have a simple form when $g(\theta_1, \theta_2) = \beta_0$ implies $\theta_1 = \theta_2$ (see Web Appendix B):

$$p_L(\mathbf{x}, \beta_0) = P_{A,B}\left[L_{\theta_2}(\mathbf{x}_2, B) \leq U_{\theta_1}(\mathbf{x}_1, A)\right], \tag{3}$$

and, for testing $H_0 : g(\theta_1, \theta_2) \geq \beta_0$,

$$p_U(\mathbf{x}, \beta_0) = P_{A,B}\left[U_{\theta_2}(\mathbf{x}_2, B) \geq L_{\theta_1}(\mathbf{x}_1, A)\right]. \tag{4}$$

3. Motivation

We motivate melded confidence intervals using the example of calculating the upper 64% one-sided confidence limit for the difference in two proportions, $\beta = \theta_2 - \theta_1$. We use the 64% confidence interval because the graphs will be easier to interpret, but the ideas are analogous for more standard levels. Suppose we observe $x_1 = 4$ out of $n_1 = 11$ positive responses in group 1 and $x_2 = 13$ out of $n_2 = 24$ in group 2. Then the difference in sample proportions is $13/24 - 4/11 = 0.542 - 0.364 = 0.178$. A very simple 64% confidence interval has upper limit, $U_\beta([4, 13], 0.64) = U_{\theta_2}(13, 0.8) - L_{\theta_1}(4, 0.8) = 0.644 - 0.217 = 0.427$, where U_{θ_2} and L_{θ_1} are one-sided Clopper–Pearson limits. This CI is at least level 0.64, because by the nestedness property of the CIs for the θ_i we can write,

$$\begin{aligned} &P\left[\theta_2 - \theta_1 \leq U_{\theta_2}(X_2, 0.8) - L_{\theta_1}(X_1, 0.8)\right] \\ &\quad \geq P\left[\left\{\theta_2 \leq U_{\theta_2}(X_2, 0.8)\right\} \text{ and } \left\{L_{\theta_1}(X_1, 0.8) \leq \theta_1\right\}\right] \\ &\quad = P\left[\theta_2 \leq U_{\theta_2}(X_2, 0.8)\right] P\left[L_{\theta_1}(X_1, 0.8) \leq \theta_1\right] \\ &\quad \geq (0.8)(0.8) = 0.64. \end{aligned}$$

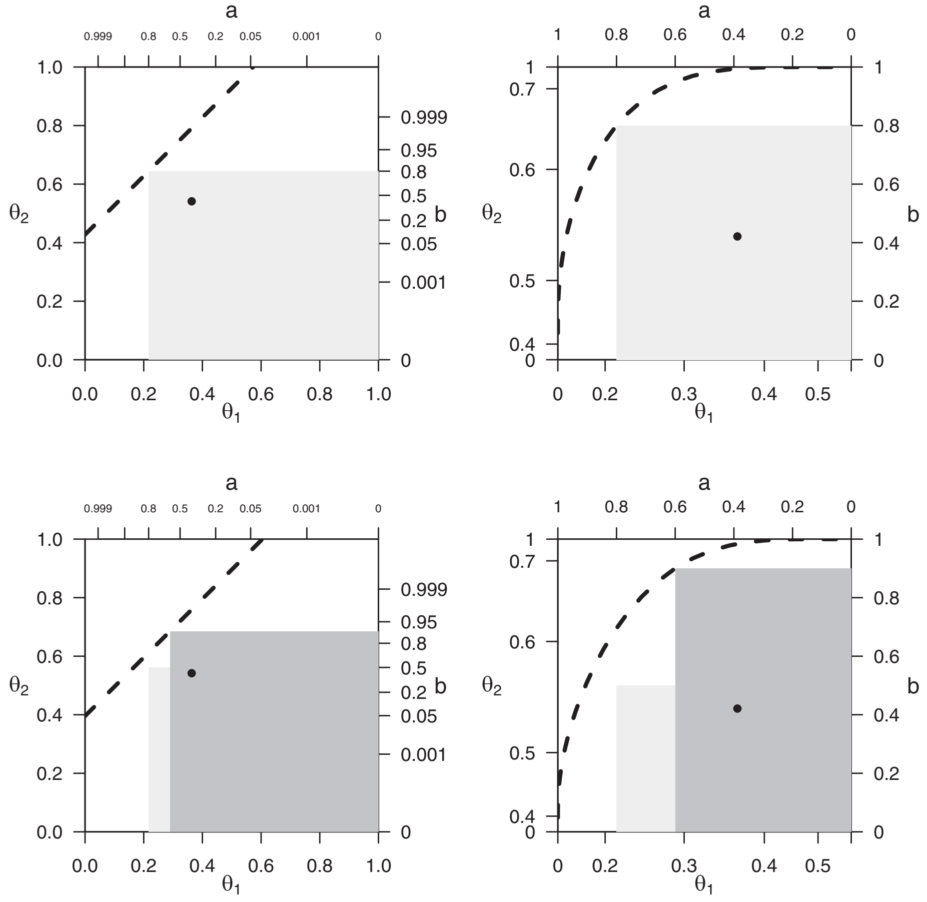

Figure 1. Plots of simple 64% upper one-sided confidence limits for $\theta_2 - \theta_1$ with sample proportions $\hat{\theta}_1 = 4/11$ and $\hat{\theta}_2 = 13/24$. Top graphs depict $U_{\theta_2}(0.8) - L_{\theta_1}(0.8)$. The bottom graphs depict the CI constructed by combining two rectangles. The left graphs are plotted in the θ_1 versus θ_2 space with the associated levels for the lower limit levels (a) given on the top and the upper limit levels (b) given on the right. The right graphs are plotted on the a versus b space with the θ_1 and θ_2 axes adjusted accordingly. The dotted lines represent the level curve $\theta_2 - \theta_1 = 0.427$ (top) or 0.396 (bottom), the upper one-sided confidence limit for $\theta_2 - \theta_1$, and the points represent the sample proportions. The right gray areas are 0.64, and pictorially represent the nominal level.

This CI for β is illustrated in the upper quadrants of Figure 1. The level curve $\theta_2 - \theta_1 = U_\beta([4, 13], 0.64) = 0.427$ is represented by the dotted lines, and the gray areas represent the set, $\left\{\theta_2 \leq U_{\theta_2}(13, 0.8) \text{ and } \quad L_{\theta_1}(4, 0.8) \leq \theta_1\right\}$. The left graph is represented in the $(\theta_1 \times \theta_2)$-space with the corresponding lower limit levels (a) provided on the top, and the upper limit levels (b) provided on the right. The right graph is represented in the $(a \times b)$-space with the corresponding θ values displayed on the left and bottom. The area of the gray rectangle in the right graph is 0.64 representing the nominal level.

To obtain a lower U_β value, we can combine two gray rectangles as depicted in the lower graphs of Figure 1. Let $U_\beta(\mathbf{x}, 0.64) = \max_i \left\{U_{\theta_2}(x_2, b_i) - L_{\theta_1}(x_1, a_i)\right\}$, where $0 < a_1 < a_2 < 1$ and $1 > b_1 > b_2 > 0$. For U_β in this form, the coverage is at least $q_{\text{nom}} = a_1 b_1 + a_2 b_2 - a_1 b_2$ (i.e., the area of the gray regions in the right bottom graph). A formal statement of this is given in Theorem 1 (Section 4). For the lower graphs of Figure 1 we use $a_1 = 0.6$, $a_2 = 0.8$, $b_1 = 0.9$ and $b_2 = 0.5$, so that $q_{\text{nom}} = 0.64$. For this confidence limit, $U_\beta = 0.396$, which is smaller than the value of 0.427 of the upper graphs. Notice the lighter rectangles in the bottom quadrants of Figure 1 do not touch the dotted line at the corner, so there is room for improvement. That is, if we extend the lighter rectangle to the left, we can reduce the height of the darker rectangle; we can then shift the level curve $\theta_2 - \theta_1 = 0.396$ to the "southeast" (i.e., $\theta_2 - \theta_1 = c$, where $c < 0.396$), producing a narrower confidence interval.

We can continue adding more rectangles, but in a smarter way such that the corners of the rectangles touch the dotted line at the U_β value. For example, suppose we posit a value for U_β and values for $0 = a_0 < a_1 < a_2 < a_3 < \cdots < a_k$. Then as long as U_β is not too small or the a_i values are not too close to 0 or 1, we can solve for the b_i values such that $U_\beta = U_{\theta_2}(x_2, b_i) - L_{\theta_1}(x_1, a_i)$. The nominal level, q_{nom}, is the gray area in the right panels of Figure 2. Theorem 1 in Section 4 shows that $q_{\text{nom}} = \sum_{i=1}^{k} (a_i b_i - a_{i-1} b_i)$, and that the CIs achieve at least that nominal level of coverage. In Figure 2 we do this by positing U_β values of 0.30. For the top graphs we

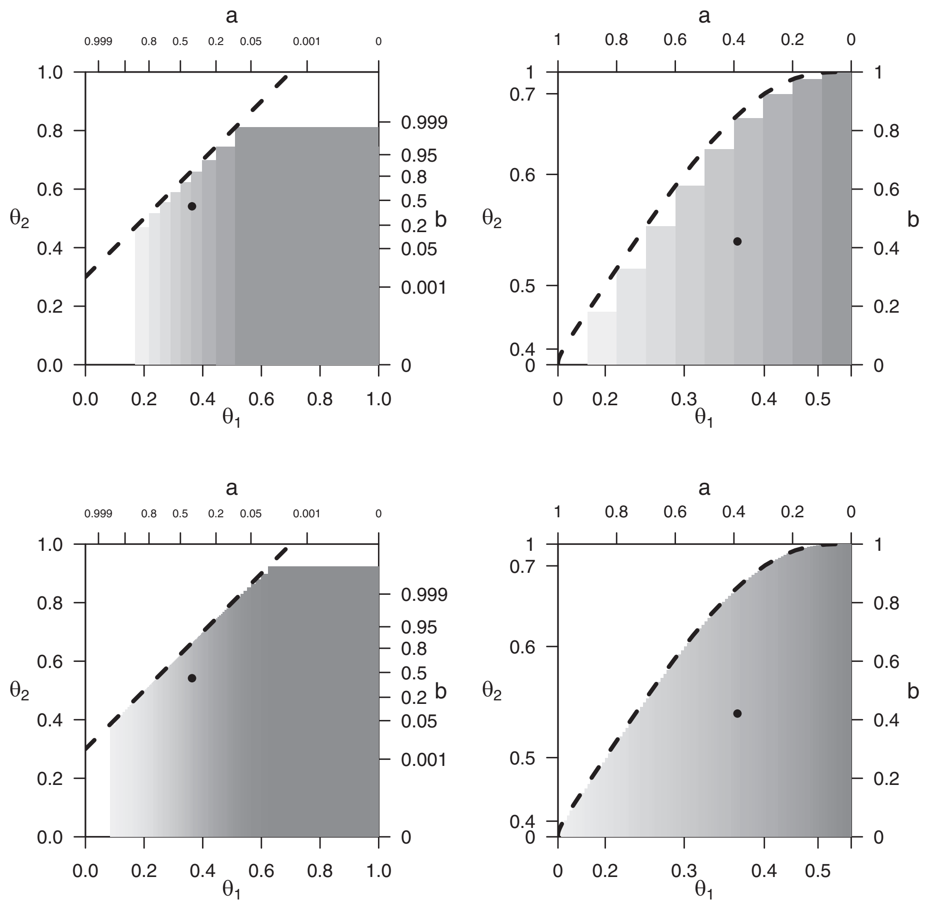

Figure 2. Plots of 64% upper one-sided confidence limits for $\theta_2 - \theta_1$ with sample proportions $\hat{\theta}_1 = 4/11$ and $\hat{\theta}_2 = 13/24$. Top graphs depict use 9 rectangles ($a = 0.1, 0.2, \ldots, 0.9$), while the bottom graphs use 98 rectangles ($a = 0.02, 0.03, .04, \ldots, 0.99$). The associated b values are chosen so that $U_{\theta_2}(b) - L_{\theta_1}(a)$ equals 0.30. The right gray areas represent the nominal level and are 0.606 (top) and 0.654 (bottom). As with Figure 1, the left graphs are plotted in the θ_1 versus θ_2 space with the associated levels for the lower limit levels (a) given on the top and the on the upper limit levels (b) given on the right. The right graphs are plotted on the a versus b space with the θ_1 and θ_2 axes adjusted accordingly. The dotted lines represent the upper one-sided confidence limit for $\theta_2 - \theta_1$ and the points represent the sample proportions.

use 9 rectangles and get $q_{\text{nom}} = 0.606$, which is less than our target of 0.64. But if we increase the number of rectangles to 98 (bottom graphs), we get $q_{\text{nom}} = 0.654 > 0.64$.

The panel in the lower right of Figure 2 shows that there is now little room for improvement, since there is not much white space below and to the right of the dotted line, and 0.654 is close to the nominal level of 0.64. The melded CIs are equivalent to finding the dotted line, and its corresponding U_β, such that the area under the dotted curve on the a versus b plot is exactly 0.64. For this example, this value is $U_\beta = 0.292$, much improved over the original 0.427. We next provide these statements in a more general way (i.e., allowing other functions besides the difference, and not just the binomial case), but there are essentially no new conceptual ideas needed for applying the method more generally.

4. Some General Theorems

We now gather the motivating ideas into two general theorems and propose another about power. The theorems are for the one-sided upper interval; the one-sided lower is analogous and is not presented.

Theorem 1. *Define $g(\cdot, \cdot)$ with monotonicity constraints as in Section 2. Let $0 = a_0 < a_1 < a_2 < \cdots < a_k < 1$, and $1 > b_1 > b_2 > \cdots > b_k > 0$, and $q_{\text{nom}} = \sum_{i=1}^{k} (a_i - a_{i-1}) b_i$, and*

$$u(\mathbf{X}, \mathbf{a}, \mathbf{b}) = \max \Big\{ g(s,t) : L_{\theta_1}(\mathbf{X}_1, a_i) \leq s \text{ and } t \leq U_{\theta_2}(\mathbf{X}_2, b_i), \text{ for } i = 1, \ldots, k \Big\}.$$

Then

$$P_{\mathbf{X}} \left[g(\theta_1, \theta_2) \leq u(\mathbf{X}, \mathbf{a}, \mathbf{b}) \right] \geq q_{\text{nom}}. \tag{5}$$

The theorem is proven in Web Appendix C.

We relate this theorem to the melded CIs by the following.

THEOREM 2. *For each data vector, $\mathbf{x}$, the value $U_\beta(\mathbf{x}, q)$ of equation 2 gives the infimum value of $u(\mathbf{x}, \mathbf{a}, \mathbf{b})$ such that $q_{\text{nom}} \geq q$ over all possible vectors $\mathbf{a}$ and $\mathbf{b}$ as defined in Theorem 1.*

The theorem is proven in Web Appendix D.

Theorems 1 and 2 suggest that the melded CIs guarantee coverage when each of the single sample CIs guarantee coverage. That conjecture has not been rigorously proven. Although Theorem 1 holds for any fixed $\mathbf{a}$ and $\mathbf{b}$, the values of $\mathbf{a}$ and $\mathbf{b}$ that give the infimum value of $u(\mathbf{x}, \mathbf{a}, \mathbf{b})$ in Theorem 2 depend on $\mathbf{x}$. So to rigorously show guaranteed coverage, we need to show an inequality analogous to expression 5, except allowing $\mathbf{a}$ and $\mathbf{b}$ to depend on $\mathbf{X}$. Despite this lack of rigor, in every example studied in the article, the evidence fully supports the conjecture.

For any confidence interval or series of hypothesis tests, we want not just guaranteed coverage and controlled type I error rates, but tight CIs and powerful tests. To show that the melded CIs are a good strategy in this respect, we turn to the case when each of the individual CIs that are melded together are asymptotically equivalent to standard normal theory confidence intervals.

THEOREM 3. *Let $L^Z_{\theta_i}(\mathbf{X}_i, 1-q) = U^Z_{\theta_i}(\mathbf{X}_i, q) = \hat{\theta}_i(\mathbf{X}_i) + \frac{\hat{\sigma}_i(\mathbf{X}_i)}{\sqrt{n_i}}\Phi^{-1}(q)$ be asymptotically normal, that is, $\sqrt{n_i}\{\hat{\theta}_i(\mathbf{X}_i) - \theta_i\} \to N(0, \sigma_i^2)$, and assume that $\hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}_1$, and $\hat{\sigma}_2$ converge almost surely to $\theta_1, \theta_2, \sigma_1$, and σ_2, respectively. Suppose that*

$$\sqrt{n_i}\left\{L_{\theta_i}(\mathbf{X}_i, 1-\alpha) - L^Z_{\theta_i}(\mathbf{X}_i, 1-\alpha)\right\} \overset{a.s.}{\to} 0 \text{ and}$$

$$\sqrt{n_i}\left\{U_{\theta_i}(\mathbf{X}_i, 1-\alpha) - U^Z_{\theta_i}(\mathbf{X}_i, 1-\alpha)\right\} \overset{a.s.}{\to} 0.$$

If g has continuous partial derivatives, then the melded CIs using L_{θ_i} and U_{θ_i} have asymptotically accurate coverage probabilities and are asymptotically equivalent to applying the delta method on the function $g(\hat{\theta}_1, \hat{\theta}_2)$; that is, treating $g(\hat{\theta}_1, \hat{\theta}_2)$ as asymptotically normal with mean $g(\theta_1, \theta_2)$ and variance

$$\left(\left.\frac{\partial g(s,t)}{\partial s}\right|_{s=\theta_1, t=\theta_2}\right)^2 \left(\frac{\sigma_1^2}{n_1}\right) + \left(\left.\frac{\partial g(s,t)}{\partial t}\right|_{s=\theta_1, t=\theta_2}\right)^2 \left(\frac{\sigma_2^2}{n_2}\right).$$

We can apply Theorem 3 to situations for which the one-sample intervals are asymptotically normal, such as the binomial case of Section 5. For a proof of the theorem and how it applies to the binomial case, see Web Appendix E.

5. Normal Case

Let $\mathbf{X}_i = [Y_{i1}, \ldots, Y_{in_i}]$ be independently distributed $N(\theta_i, \sigma_i^2)$ for $i = 1, 2$. Let $\bar{y}_i$ and s_i^2 be the usual sample mean and unbiased variance estimate from the ith group. Consider first the case with known variances. Then $L_{\theta_i}(\mathbf{x}_i, q) = \bar{y}_i - \frac{\sigma_i}{\sqrt{n_i}}\Phi^{-1}(q)$. Thus, $L_{\theta_i}(\mathbf{x}_i, A) \sim N\left(\bar{y}_i, \frac{\sigma_i^2}{n_i}\right)$ and $U_{\theta_i}(\mathbf{x}_i, A) \sim N\left(\bar{y}_i, \frac{\sigma_i^2}{n_i}\right)$ as well. When $g(\theta_1, \theta_2) = \theta_2 - \theta_1$ then $g(L_{\theta_1}(\mathbf{x}_1, A), U_{\theta_2}(\mathbf{x}_2, B)) \sim N\left(\bar{y}_2 - \bar{y}_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ and $U_\beta(\mathbf{x}, 1-\alpha) = \bar{y}_2 - \bar{y}_1 + \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, which are equivalent to the CIs that match the uniformly most powerful (UMP) one-sided tests (see e.g., Lehmann and Romano, 2005, p. 90).

Now suppose the σ_i^2 are unknown and not assumed equal. Then the usual one-sample t-based confidence intervals at levels a and b give,

$$L_{\theta_1}(\mathbf{x}_1, a) = \bar{y}_1 - \frac{s_1}{\sqrt{n_1}}t^{-1}_{n_1-1}(a) \text{ and } U_{\theta_2}(\mathbf{x}_2, b) = \bar{y}_2 + \frac{s_2}{\sqrt{n_2}}t^{-1}_{n_2-1}(b) \tag{6}$$

where $t_d^{-1}(q)$ is the qth quantile of a central t-distribution with d degrees of freedom. By the probability integral transform and with A and B uniform, $T_1 = t^{-1}_{n_1-1}(A)$ and $T_2 = t^{-1}_{n_2-1}(B)$ are t random variables. Then the $100(1-2a)$ melded CI for $\theta_2 - \theta_1$ is the ath and $(1-a)$th quantiles of $g\left(L_{\theta_1}(\mathbf{x}_1, A), U_{\theta_2}(\mathbf{x}_2, B)\right)$. This is equivalent to the Behrens–Fisher interval (Fisher, 1935). Thus, a test based on that interval would give significance whenever the Behrens–Fisher solution declared the two means significantly different. The coverage of the Behrens–Fisher interval is generally not exactly equal its nominal value, but is thought to be conservative. Robinson (1976) conjectured and supported through extensive calculations that the test based on the Behrens–Fisher solution retains the type I error rate. As far as we are aware (see also Lehmann and Romano, 2005, p. 415), the first proof of this retention of the type I error rate was Balch (2012), which used Dempster–Shafer evidence theory (see e.g., Yager and Liu, 2008) and his newly developed confidence structures. (Note: the rigor of Balch's proof may be similar to the relationship of Theorem 1 and 2 to our conjecture, because the A in Balch's Confidence-Mapping Lemma would typically depend on the data, and this is not explicitly accounted for in Balch's proof.) Theorems 1 and 2 of this article provide additional support for this claim.

6. Binomial Case

Suppose $X_i \sim \text{Binomial}(n_i, \theta_i)$. Then using the usual exact (i.e., guaranteed coverage for all values of θ_i, but possibly conservative for many values of θ_i) one-sided intervals for a binomial (Clopper and Pearson, 1934), we have $L_{\theta_i}(x_i, A) \sim \text{Beta}(x_i, n_i - x_i + 1)$ and $U_{\theta_i}(x_i, B) \sim \text{Beta}(x_i + 1, n_i - x_i)$, where A and B are uniform, and for notational convenience we extend the definition of the beta distribution to include point mass distributions at the limits, so $\text{Beta}(0, j)$ is a point mass at 0 and $\text{Beta}(i, 0)$ is a point mass at 1 for $i, j > 0$. We can obtain new exact confidence intervals for the difference: $\theta_2 - \theta_1$, the ratio: θ_2/θ_1, or the odds ratio: $\left\{\theta_2(1-\theta_1)\right\} / \left\{\theta_1(1-\theta_2)\right\}$ by choosing the appropriate g.

We illustrate the calculations for the data in Section 3 and the difference, $g(\theta_1, \theta_2) = \theta_2 - \theta_1$, but using a more conventional confidence limit of 95%. Recall that $\hat{\theta}_1 = 4/11$ and $\hat{\theta}_2 = 13/24$. First, we run the Monte Carlo calculation, with $m = 10^6$ replications. For the lower limit, we use the kth largest ($k = 0.025m$) out of m pseudo-random samples of $T_{L2} - T_{U1}$, where $T_{L2} \sim \text{Beta}(13, 12)$ and $T_{U1} \sim \text{Beta}(5, 7)$, giving -0.2322. Similarly, for the upper limit we use the

$(m-k+1)$th (see Efron and Tibshirani, 1993, p. 160) largest out of m pseudo-random samples of $T_{U2}-T_{L1}$, where $T_{U2} \sim \text{Beta}(14,11)$ and $T_{L1} \sim \text{Beta}(4,8)$, giving 0.5263. The one-sided p-values are the proportion of the T_{L2} that are less than T_{U1}, giving $p_L(0)=0.2703$, and the proportion of the T_{U2} that are greater than T_{L1}, giving $p_U(0)=0.9114$. Alternatively, we could use the numeric integration calculation. Using the relationship between one-sided p-values and confidence limits,

$$p_L(\beta_0) = P_{A,B}\left[L_{\theta_2}(B) - U_{\theta_1}(A) \le \beta_0\right] = P\left[T_{L2} \le \beta_0 + T_{U1}\right]$$
$$= \int_0^1 F_{L2}(t+\beta_0) f_{U1}(t)\,\mathrm{d}t,$$

where F_{L2} is the cumulative distribution of T_{L2}, and f_{U1} is the density function of T_{U1}. Then using a root solving function, we find the value of β_0 such that $p_L(\beta_0)=0.025$, giving -0.2321. Analogously, we solve $p_U(\beta_0)=0.975$ for β_0 using numeric integration to get 0.5262. Using numeric integration, the one-sided p-values for testing $\beta_0=0$ are $p_L(0)=0.2706$ and $p_U(0)=0.9110$, giving a two-sided p-value of $2p_L(0)=0.541$.

Note that the associated p-values for testing the one-sided equality of the θ_i (i.e., $p_L(0)$ and $p_U(0)$ for the difference, or generally equations 3 or 4) are equivalent to the one-sided p-values using Fisher's exact test. This equivalence has been shown in the context of the Bayesian analysis of a 2×2 table by Altham (1969). Because of this equivalence the type I error rate is bounded at the nominal level when testing one-sided tests that $\theta_1 \le \theta_2$ or $\theta_1 \ge \theta_2$.

In order to test the coverage, we performed extensive numerical calculations. For any fixed n_1 and n_2 we calculated all the possible melded upper 95% confidence limits. Then, using those upper limits, we calculated the coverage for all $(101)^2$ values of θ_1 and θ_2 in $\{0, 1/100, 2/100, \ldots, 1\}$. We repeated this calculation for all $n_1, n_2 \in \{1, 2, \ldots, 100\}$. We repeated these steps for the differences, ratios, and odds ratios. We found that the coverage was always at least 95%. Because of the symmetrical nature of the problem, this implies 95% coverage for the lower limits as well. Thus, it appears that these melded confidence intervals guarantee coverage.

This problem is the widely studied 2×2 table with one margin (namely, the sample sizes) fixed. There is no consensus on the best inferential method for this situation. Some argue that conditioning is merited (Yates, 1984, see discussion), but others argue that the unconditional test is preferred because it is generally more powerful in this case (Lydersen, Fagerland, and Laake, 2009). An issue is that *if* you fix the significance level, then the discreteness of the conditional distribution will typically make the conditional inferences less powerful than the unconditional ones. Some argue for conditioning by noting that fixing the significance level is not needed or scientific (Upton, 1992), or that we condition on the closely related Poisson problems without controversy (Little, 1989). If we remove the discreteness problem by the impractical use of randomization, then a conditional test, the randomized version of Fisher's exact test, is the uniformly most powerful unbiased test (see Lehmann and Romano, 2005, p.127).

Since our melded confidence limits match the one-sided Fisher's exact test as mentioned above, the melded confidence limits allow conditional-like inferences for the difference and ratio, whereas previously they have only been available in practice for the odds ratio. Additionally, the melded CIs are much faster to calculate than the unconditional intervals because the unconditional intervals require searching over the space of the nuisance parameter (see, e.g., Chan and Zhang, 1999).

7. Poisson Case

Suppose $X_i \sim \text{Poisson}(n_i\theta_i)$ for $i=1,2$, where the mean $n_i\theta_i$ is the rate, θ_i, times the time at risk, n_i. Suppose we are interested in testing $H_0 : g(\theta_1,\theta_2) = \frac{\theta_2}{\theta_1} \le r$. As with the binomial case, the UMPU test is a randomized one, and practically, we use a non-randomized version of it. In this practical test, we condition on X_1+X_2; then when $\theta_2 = r\theta_1$ we have $X_1 \mid X_1 + X_2 \sim \text{Binomial}\left(X_1+X_2, \frac{n_1}{n_1+rn_2}\right)$ (see, e.g., Lehmann and Romano, 2005). We reject when X_1 is large and the p-value is

$$p_b(x_1,x_2)$$
$$= \sum_{i=x_1}^{x_1+x_2} \binom{x_1+x_2}{i} \left(\frac{n_1}{n_1+rn_2}\right)^i \left(1-\frac{n_1}{n_1+rn_2}\right)^{x_1+x_2-i}.$$

We show in Web Appendix F that p_b is equivalent to the melded p-value based on the standard one-sample exact Poisson intervals (Garwood, 1936).

The advantage of the melded intervals is that we may get intervals for the difference in the θ_i. The difference may be more important for measuring public health implications of interventions, since it can be translated into how many lives are affected (see, e.g., Chan and Wang, 2009). For example, halving the relative risk from a baseline disease rate of 2% is very different from a public health perspective than if the baseline rate is 20%. Conversely, changing the risk by decreasing the rate of a disease by 1% affects a similar number of people regardless of whether the baseline risk is 2% or 20%. As with the binomial case, the unconditional exact method is much more difficult to calculate because one needs to search over the nuisance parameter space. There are approximate and quasi-exact methods available (Chan and Wang, 2009), but no conditional exact method. Because of Theorems 1 and 2, we suspect that the melded CIs retain nominal coverage, and they are easy to calculate. Full exploration of that option and the comparison with the best competitor is left to future work.

8. Difference in Medians

Several methods have been proposed for CIs on the difference in medians from two-samples for non-censored responses. First, assuming that the two distributions represent continuous responses and differ only by a location shift, then the method of Hodges and Lehmann (1963) provides CIs on the difference in medians that guarantee coverage. However, the Hodges–Lehmann CIs can have far less than nominal coverage if either assumption does not hold, as will be shown. Second, the nonparametric bootstrap is valid asymptotically and does not require the shift assumption (Efron and Tibshirani, 1993). Other asymptotic methods require the continuity assumption and allow different types of censoring and will not be discussed further (Su and Wei, 1993; Kosorok, 1999).

Table 1

Simulated coverage for nominal 95% confidence intervals for difference in medians. The five scenarios are described in the text, but briefly: Normals is a null case with both groups standard normal, Figures 3a–c are mixtures of normals denoting a shift (Figure 3a) or asymmetric mixtures (Figures 3b and c), and Poissons are Poisson with means 2.6 and 2.7. Bolded values are significantly less than the nominal 95%.

Description	n per group	Percent coverage H–L	Percent coverage bootstrap	Percent coverage melded	Ratio of median CI lengths (melded/H–L)	Ratio of median CI lengths (melded/bootstrap)
Normals	20	95.1	96.7	99.2	1.51	1.27
	100	95.1	96.4	97.6	1.35	1.09
Figure 3a	20	94.9	96.4	99.0	2.44	1.20
	100	95.2	96.6	97.8	4.18	1.09
Figure 3b	20	**47.4**	**92.4**	97.5	2.31	1.20
	100	**0.6**	95.0	96.7	3.69	1.09
Figure 3c	20	**74.7**	95.7	98.8	1.42	1.26
	100	**17.0**	95.8	97.3	1.08	1.10
Poissons	20	**87.5**	**90.3**	100.0	2.50	2.00
	100	**57.0**	**91.8**	100.0	4.00	2.00

We create melded CIs for this situation, using single sample CIs derived by inverting the sign test (see, e.g., Slud, Byar, and Green, 1984). In Web Appendix G, we derive the lower (equation 15) and upper (equation 16) one-sided confidence limit functions that guarantee coverage even for discrete distributions. These one-sample CI functions may return either $-\infty$ (for the lower limit) or ∞ (for the upper limit), and the melded CI may give $(-\infty, \infty)$ if the sample size is too small. For example, in the continuous case with equal sample sizes in the two groups, we need at least 7 in each group to get finite 95% CIs. This is more restrictive than the Hodges–Lehmann procedure, which requires at least 4 in each group for that situation to get finite 95% CIs.

We compare the melded CIs to the Hodges-Lehman CIs and nonparametric percentile bootstrap CIs that uses 2000 replications. We simulate five scenarios, with 10,000 data sets for each scenario with $n_i = 20$ or $n_i = 100$ in each sample. Let F_i be the distributions for group i, $i = 1, 2$, and let $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2. The five scenarios are:

Normals: null case, $F_1 = F_2 = N(0, 1)$;
Figure 3a: shift case, $F_1 = 0.55 \cdot N(0, 1) + 0.45 \cdot N(-5, 1)$ (median=-1.335) and $F_2 = 0.55 \cdot N(2, 1) + 0.45 \cdot N(-3, 1)$ (median=0.665);
Figure 3b: asymmetric continuous case 1, $F_1 = 0.55 \cdot N(0, 1) + 0.45 \cdot N(-5, 1)$ (median=-1.335) and $F_2 = 0.55 \cdot N(0, 1) + 0.45 \cdot N(5, 1)$ (median=1.335);
Figure 3c: asymmetric continuous case 2, $F_1 = 0.5 \cdot N(-2, 1) + 0.5 \cdot N(2, 9)$ (median= -1.000) and $F_2 = 0.5 \cdot N(-2, 9) + 0.5 \cdot N(2, 1)$ (median=1.000);
Poissons: discrete and asymmetric case, F_1 is Poisson with mean 2.6 (median=2), and F_2 is Poisson with mean 2.7 (median=3).

The simulation results are given in Table 1. When the continuous and shift assumptions are met (*Normals* and Figure 3a), the Hodges-Lehman CIs have coverage near the nominal 95%, while in the other cases where those assumptions are not met, the Hodges–Lehmann CI has poor coverage that becomes worse as the sample size increases. In those latter cases, because the assumptions do not hold, the Hodges–Lehmann CIs on the shift are not measuring the difference in medians, and applying the method in those scenarios would lead to incorrect confidence intervals. The bootstrap does reasonably well in most situations, but does not appear to have proper coverage for Figure 3b (when $n_i = 20$ per group) and the *Poissons* case (even when $n_i = 100$). Generally, the melded CIs have wider CIs than the Hodges–Lehmann and the bootstrap CIs,

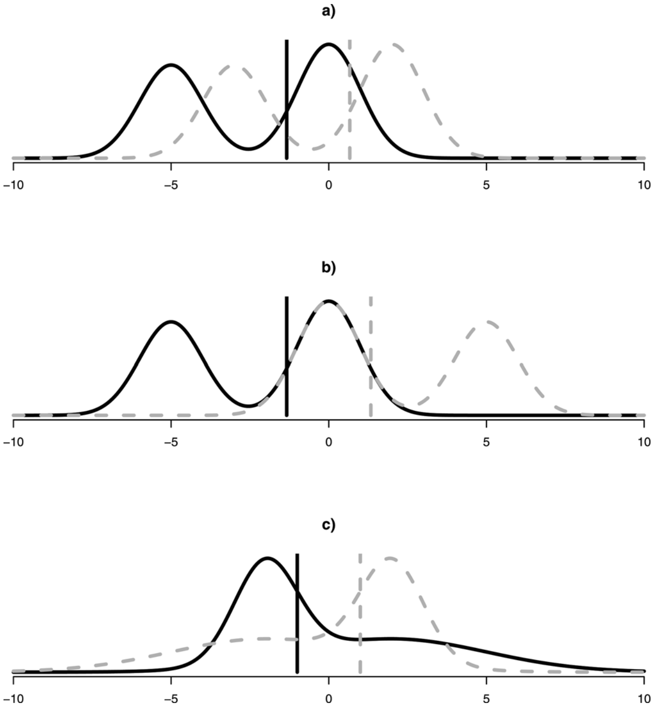

Figure 3. Mixture of Normal distributions for median simulations. Sample 1 is black solid, sample 2 is gray dotted, vertical lines are medians.

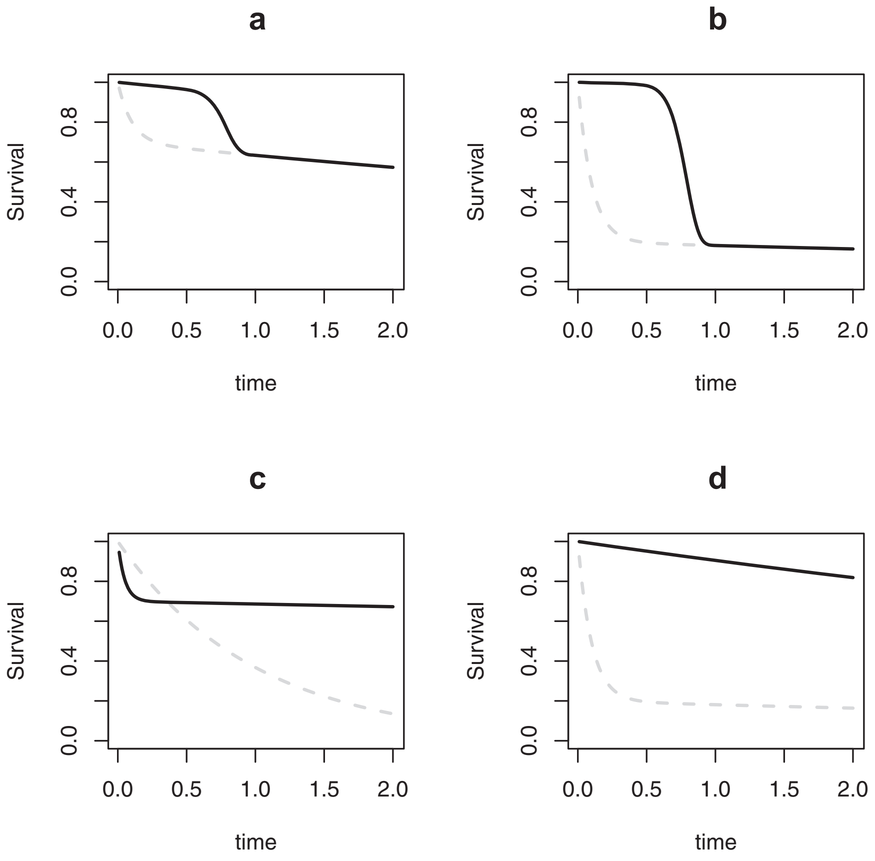

Figure 4. Survival distributions for simulations, control arm is dotted gray, treatment arm is solid black. The survival distributions are compared at time 1.0.

but this wideness ensures simulated coverage of at least the nominal level for all scenarios studied. Thus, if the priority is guaranteed coverage regardless of sample size or distributional assumptions, then the melded CIs are recommended.

9. Inferences Between Survival Distributions at a Fixed Time for Right Censored Data

The logrank test or weighted logrank tests are popular for testing for differences in survival distributions because of their good power under proportional hazards models or accelerated failure time (AFT) models (see, e.g., Kalbfleisch and Prentice, 2002, Chapter 7). In some situations those models do not fit the data well. For example, an aggressive new treatment may lead to substantial mortality immediately after initiation, but can increase survival compared to the standard treatment if the patient survives the first few weeks of treatment (see, e.g., Figure 4c). In this case the AFT models do not fit, and a more useful and relevant test is a difference comparison in survival after a fixed amount of time, for example, 1 year after randomization to treatment. Another example is plotted in Figure 4a. Suppose 30% of individuals have a serious version of a disease and die within the first year, while the other 70% have a less serious version and survive longer. Suppose a new treatment prolongs the life of those with the serious version for a short period (less than a year) but does not change that of the others. The logrank test may show significance for the new treatment, but the new treatment is not really curing patients for the long term. A better test may be to test for significant differences at 1 year.

Klein et al. (2007) studied several two-sample tests for comparing survival estimates at a fixed time. They concluded that the test based on the normal approximation on the complementary log log (CLL) transformation of the Kaplan–Meier survival estimator, estimating the variance for each sample with Greenwood's formula and the delta method (see equation 3 of that article) was generally the best at retaining the Type I error rate. We call this the CLL test, and it can fail to retain the type I error rate with small samples and/or heavy censoring. Further, the CLL test cannot be used if the Kaplan–Meier estimator for either one of the groups is equal to 0 or 1 at the fixed test time.

Thus, if we are interested in comparing survival at a fixed time in a small sample case with heavy censoring, the CLL may not be a good test, especially if a conservative procedure is desired, such as in a regulatory setting. As an alternative, we can use the melded confidence intervals based on the beta product confidence procedure (BPCP) for each survival distribution (Fay, Brittain, and Proschan, 2013). The BPCP was designed to guarantee central coverage for survival at a fixed point, and bounds the type I error rate in situations

Table 2

Simulated Percent that Reject $S_0(1) = S_1(1)$ at the one-sided 2.5% level. Survival models a and b are null models so simulated percent should be 2.5%, and bolded values for those models are significantly larger than 2.5% (at the two-sided 0.05 level). Models described by Figure 4 and Web Appendix H.

n_i	Model	Censoring	Meld low	Meld high	CLL low	CLL high	CLL, % undefined
50	a	Moderate	0.28	0.43	2.69	2.12	0.00
50	a	Heavy	0.01	0.00	2.07	**9.97**	3.34
50	b	Moderate	0.23	0.68	1.78	2.50	2.13
50	b	Heavy	0.00	0.19	0.00	**9.64**	34.92
50	c	Moderate	0.00	51.78	0.00	83.31	0.11
50	c	Heavy	0.00	0.68	0.00	58.56	6.71
50	d	Moderate	0.00	99.96	0.00	97.94	2.06
50	d	Heavy	0.00	41.07	0.00	92.15	4.67
100	a	Moderate	0.63	0.76	**2.94**	2.56	0.00
100	a	Heavy	0.01	0.04	**3.01**	**7.04**	0.51
100	b	Moderate	0.51	0.78	2.74	2.51	0.08
100	b	Heavy	0.00	0.29	0.05	**5.20**	19.81
100	c	Moderate	0.00	93.58	0.00	98.47	0.00
100	c	Heavy	0.00	6.85	0.00	86.08	3.19
100	d	Moderate	0.00	100.00	0.00	99.96	0.04
100	d	Heavy	0.00	84.47	0.00	97.97	0.87

where alternative CIs (including the bootstrap) fail to do so. Thus, the BPCP is a good choice when guaranteed coverage is important with small samples. Further, it has no requirement that the Kaplan–Meier estimators for each sample be between 0 and 1. Using the method of moments implementation of the BPCP, we can create the random variables associated with the BPCP limits using beta distributions. Because the BPCP reduces to the Clopper–Pearson intervals when there is no censoring, the melded confidence limits in this case reduce to the Fisher's exact test when testing the equality of the survival distributions (see Section 6).

For the simulations, we model the failure times of the two groups as mixture distributions. We consider 4 different pairs of mixture distributions, with survival distributions given by Figure 4, each with either moderate or heavy censoring. We let the number in each group (n_i) be 50 or 100. We simulate 10,000 data sets per condition. Details of the simulation are given in Web Appendix H.

The results are given in Table 2. We divide up the missed coverage into low (test arm has lower survival than control arm) and high (test has higher survival). We use 95% confidence limits so we expect 2.5% error for each side at the nominal level. Data sets with the CLL tests undefined (due to the Kaplan–Meier from either group being equal to 0 or 1 when time is 1) were considered non-rejections in the table. With heavy censoring for model b, there is 33.8% undefined. In most other cases with $n_i = 50$ there is 2–4% undefined. So this is a major practical disadvantage in a clinical trial, where the primary test should be specified in advance. The coverage under null hypotheses (models a and b) can be very inflated with heavy censoring, with upper error about 10% instead of the nominal 2.5%.

The melded CI method estimated type I error rate is substantially smaller than the nominal type I error rate in all cases simulated. This leads to reduced power compared to the CLL (see model c). This is because with heavy censoring there are very few observations at risk at time=1, so that in each arm the BPCP confidence limits can get very conservative, and conservativeness of the CI associated with each arm naturally propagates to conservativeness of the melded CIs. Conversely, the anti-conservativeness of the CIs based on the asymptotic normal approximation of the complementary log–log transformation in this heavy censoring situation leads to the anti-conservativeness of the CLL tests.

This is only a preliminary assessment of how the melded confidence interval performs. It appears to be a promising approach when a conservative test is required, although it clearly has low power for some parameters we considered. This may reflect the fact that inference at a fixed time point is fraught with difficulty when there is considerable censoring. Finally, note that because the beta product confidence procedure on the survival distribution may be inverted to get one-sample confidence intervals on the median with right censoring (see Fay et al., 2013, Section 6.1), we can use this procedure to get melded CIs for the difference in medians with right censoring. This once again illustrates the flexibility of the melded confidence interval approach.

10. Connections to the Confidence Distribution Method

The confidence distribution (CD) is a frequentist distributional estimator of a parameter. For example, consider the case where we have an exact one-sided CI for continuous data, where the coverage associated with the one-sided upper confidence limit, $U_{\theta_i}(\mathbf{X}_i, q)$, is q for all $q \in (0, 1)$. In this case $T_{Ui} \equiv U_{\theta_i}(\mathbf{x}_i, A)$ is a CD random variable, and its cumulative distribution function at t is $H_U(\mathbf{x}_i, t) \equiv U^{-1}_{\theta_i}(\mathbf{x}_i, t)$. We call $H_U(\mathbf{x}_i, \cdot)$ the CD. It can be used similarly to other distribution estimators like the bootstrap or the posterior distribution. For example, the middle $100q$% of the distribution is a $100q$% central confidence interval for any q. This application is circular, since we derive the CD from the CI process, then use the CD to get back the CIs. The modern definition of a CD avoids the circular reasoning by defining a CD as a function having two properties: (i) for each $\mathbf{x}_i$, $H(\mathbf{x}_i, t)$ is a cumulative distribution for T_i, and (ii) at the true value θ_i, $H(\mathbf{X}_i, \theta_i)$ is uniform. Importantly, the CD has other uses besides estimating CIs, like estimating the parameter itself or combining information on a parameter from independent samples (see Xie and Singh, 2013; Yang et al., 2014). The latter application is similar to what the melded CI for the continuous case does, it takes the CD random variable for θ_1 and melds it with the CD random variable for θ_2 to create a CD-like random variable for β that is used to create a CI for β.

For the continuous case with exact CIs, we could have equivalently defined the CD random variable as $T_{Li} \equiv L_{\theta_i}(\mathbf{x}_i, 1-A)$ with distribution $H_{Li}(\mathbf{x}_i, t) = 1 - L_{\theta_i}^{-1}(\mathbf{x}_i, t)$, since $H_U(\mathbf{x}_i, t) = H_L(\mathbf{x}_i, t) \equiv H(\mathbf{x}_i, t)$ so that $T_{Ui} = T_{Li} \equiv T_i$. Unfortunately, applying the CD method to discrete data is not straightforward because for CIs with guaranteed coverage we generally have $H_U(\mathbf{x}_i, t) \neq H_L(\mathbf{x}_i, t)$ and there is not one clear distribution to define as the CD. Further, for discrete data $H(\mathbf{X}_i, \theta_i)$ cannot be a uniform distribution.

The melded CIs are one way to generalize the CD approach to handle two-sample discrete small sample situations (for another approximate way see Hannig and Xie, 2012). We can generalize by defining the upper and lower CDs as $H_U(\mathbf{x}_i, t)$ and $H_L(\mathbf{x}_i, t)$, respectively, representing the cumulative distributions of T_{Ui} and T_{Li} evaluated at t. For each $\mathbf{x}_i$, the upper and lower CDs are each a cumulative distribution, and at the true value of θ, $H_U(\mathbf{X}_i, \theta) \overset{sto}{\leq} A \overset{sto}{\leq} H_L(\mathbf{X}_i, \theta)$, where A is a uniform random variable, and $Y_1 \overset{sto}{\leq} Y_2$ implies $Pr[Y_1 \leq t] \geq Pr[Y_2 \leq t]$ for all t. When we create the upper melded confidence limit for $g(\theta_1, \theta_2)$, we use the lower CD for θ_1 and the upper CD for θ_2 in $g(\cdot, \cdot)$ to lead to more conservative coverage.

To ensure valid confidence intervals from CDs on functions of parameters in the continuous case, we require monotonicity (see Xie and Singh, 2013, p.14). In a similar way, for melded CIs we require $g(\theta_1, \theta_2)$ to meet the monotonicity constraints (see Web Appendix A). If $g(\cdot, \cdot)$ does not meet these monotonicity constraints, then the resulting interval may not guarantee coverage. For example, the ratio of two parameters is not monotonic in the denominator if the parameter in the numerator is non-zero and the denominator crosses zero. For the ratio of normals, if one performs the melded confidence interval method despite the assumption violation on $g(\cdot, \cdot)$, then an anonymous reviewer of this article has shown that the coverage can be either conservative or anti-conservative (see also Xie and Singh, 2013, Example 6).

11. Discussion

We have proposed a simple confidence interval procedure for inferences in the two-sample problem. Our melded CI can be interpreted as a generalization of the confidence distribution approach. We take frequentist distributional estimators of single-sample parameters and combine them using a monotonic function to create two-sample confidence intervals that appear to guarantee coverage.

Although we are unable to rigorously prove that the melded CI method controls error rates (see discussion after Theorem 2 in Section 4), several lines of additional argument suggest that it does. The first is the remarkable fact that it reproduces accepted tests and intervals in many settings examined (binomial, normal, Poisson, Behrens–Fisher problem). Second, extensive numerical calculations in the binomial case and further simulations in two other situations (difference in medians and difference in survival distributions) failed to find any situation where the melded CIs had less than nominal coverage.

In addition to reproducing some well-accepted tests and confidence intervals, the melding method has yielded new intervals. For example, in the binomial case, it gives confidence intervals for the relative risk and risk difference that match the inferences from Fisher's exact test. Previously, such intervals were readily available only for the odds ratio. Thus, the new CI for the risk difference could be used as the primary analysis in the regulatory setting where risk difference is traditionally used, such as a new antibiotic is being compared to an existing one in a non-inferiority trial. The melded CI method is so general, it can easily generate a two-sample procedure in any setting where there is a one-sample procedure that guarantees coverage, as illustrated in the methods presented in Sections 8 and 9. We briefly mention two more possible applications. Consider a randomized clinical trial measuring the effect of treatment compared to placebo, and suppose there is an accepted confidence procedure for that treatment effect. If one wants to determine if there is a difference in treatment effects between two subgroups (say between men and women), then the melded CI approach can answer that question. Next consider two trials that measure vaccine efficacy for two different vaccines both designed to protect against the same disease. Each trial estimates vaccine efficacy by comparing the ratio of infection rates of the vaccinated to the unvaccinated. If the trials are done on similar populations and both use the same control vaccine, a comparison of the two vaccine efficacies can be done using melded CIs. So we see that the potential for developing useful new two-sample tests and intervals from exact one-sample procedures makes the melded confidence interval approach an appealing addition to the applied statistician's toolkit.

12. Supplementary Materials and Software

Web Appendices referenced in Sections 2, 4, 7, 8, 9, and 10 are available with this paper at the *Biometrics* website on Wiley Online Library. Additionally, the R scripts used in the simulations are available at the Biometrics website on Wiley Online Library. Two R packages, exact2x2 and bpcp, are available on CRAN (`http://cran.r-project.org/`) and have functions to calculate the binomial melded CIs (binomMeld.test in exact2x2), the difference in medians melded CIs (mdiffmedian.test in bpcp), and the difference in survival distribution melded CIs (bpcp2samp in bpcp).

Acknowledgements

We thank Dean Follmann the anonymous reviewers of the article for helpful suggestions that improved the article. The numerical calculations and simulations were run using the

Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

References

Altham, P. M. (1969). Exact Bayesian analysis of a 2× 2 contingency table, and fisher's" exact" significance test. *Journal of the Royal Statistical Society, Series B (Methodological)* **31**, 261–269.

Balch, M. (2012). Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning* **53**, 1003–1019.

Chan, I. S. and Wang, W. W. (2009). On analysis of the difference of two exposure-adjusted poisson rates with stratification: From asymptotic to exact approaches. *Statistics in Biosciences* **1**, 65–79.

Chan, I. S. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55**, 1202–1209.

Clopper, C. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Vol. 57. New York: CRC Press.

Fay, M. P., Brittain, E. H., and Proschan, M. A. (2013). Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics* **14**, 723–736.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics* **6**, 391–398.

Garwood, F. (1936). Fiducial limits for the poisson distribution. *Biometrika* **28**, 437–442.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica* **19**, 491.

Hannig, J. and Xie, M. (2012). A not on Dempster–Shafer recombination of confidence distributions. *Electronic Journal of Statistics* **6**, 1943–1966.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics* **34**, 598–611.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. New York: Wiley.

Klein, J., Logan, B., Harhoff, M., and Andersen, P. (2007). Analyzing survival curves at a fixed point in time. *Statistics in Medicine* **26**, 4505–4519.

Kosorok, M. R. (1999). Two-sample quantile tests under general conditions. *Biometrika* **86**, 909–921.

Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*, 3rd edition. New York: Springer.

Little, R. J. (1989). Testing the equality of two independent binomial proportions. *The American Statistician* **43**, 283–288.

Lydersen, S., Fagerland, M. W., and Laake, P. (2009). Recommended tests for association in 2× 2 tables. *Statistics in Medicine* **28**, 1159–1175.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association* **108**, 301–313 (correction 108(503):1138–1139).

Pedersen, J. (1978). Fiducial inference. *International Statistical Review* **46**, 147–170.

Robinson, G. (1976). Properties of Students t and of the Behrens–Fisher solution to the two means problem. *The Annals of Statistics* **4**, 963–971.

Slud, E. V., Byar, D. P., and Green, S. B. (1984). A comparison of reflected versus test-based confidence intervals for the median survival time, based on censored data. *Biometrics* **40**, 587–600.

Stevens, W. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37**, 117–129.

Su, J. Q. and Wei, L. (1993). Nonparametric estimation for the difference or ratio of median failure times. *Biometrics* **49**, 603–607.

Upton, G. J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **155**, 395–402.

Xie, M.-g. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review* **81**, 3–77.

Yager, R. and Liu, L. e. (2008). *Classic Works of the Dempster-Shafer Theory of Belief Functions*, Vol. 219. New York: Springer.

Yang, G., Liu, D., Liu, R. Y., Xie, M., and Hoaglin, D. C. (2014). Efficient network meta-analysis: A confidence distribution approach. *Statistical Methodology* **20**, 105–125.

Yates, F. (1984). Tests of significance for 2× 2 contingency tables. *Journal of the Royal Statistical Society, Series A (General)* **147**, 426–463.

Zabell, S. (1992). R. a. fisher and the fiducial argument. *Statistical Science* **7**, 369–387.

Received January 2014. *Revised July* 2014.
Accepted August 2014.

Web-Based Supplementary Materials for Combining One-Sample Confidence Procedures for Inference in the Two-Sample Case

MICHAEL P. FAY, MICHAEL A. PROSCHAN, ERICA BRITTAIN
National Institute of Allergy and Infectious Diseases,
Bethesda, MD, 20892-7630, USA
correspondence to M.P. Fay: mfay@niaid.nih.gov

Summary

This supplement provides additional mathematical and simulation design details for the paper. Equation numbers and Figure numbers continue from the main paper.

Web Appendix A Monotonicity Constraints on g

Let the possible values of θ_i be Θ_i and be contiguous, so that we can write the set Θ_i in one of four ways,

$$\begin{array}{ll} (S1) & \Theta_i = [\theta_{i,min}, \theta_{i,max}] \\ (S2) & \Theta_i = (-\infty, \theta_{i,max}] \\ (S3) & \Theta_i = [\theta_{i,min}, \infty) \\ (S4) & \Theta_i = (-\infty, \infty) \, . \end{array}$$

When both Θ_1 and Θ_2 can be written as $(S1)$, then the monotonicity constraints on $g(\cdot,\cdot)$ may be written as:

$$\begin{array}{lll} (C1) & g(s_2,t) \le g(s_1,t) & \text{for } s_1 < s_2,\ s_i \in (\theta_{1,min}, \theta_{1,max}], \text{ and } t \in [\theta_{2,min}, \theta_{2,max}) \\ (C2) & g(s,t_1) \le g(s,t_2) & \text{for } t_1 < t_2,\ s \in (\theta_{1,min}, \theta_{1,max}], \text{ and } t_i \in [\theta_{2,min}, \theta_{2,max}) \\ (C3) & g(s_2,t) \le g(s_1,t) & \text{for } s_1 < s_2,\ s_i \in [\theta_{1,min}, \theta_{1,max}), \text{ and } t \in (\theta_{2,min}, \theta_{2,max}] \\ (C4) & g(s,t_1) \le g(s,t_2) & \text{for } t_1 < t_2,\ s \in [\theta_{1,min}, \theta_{1,max}), \text{ and } t_i \in (\theta_{2,min}, \theta_{2,max}] \end{array}$$

where we interpret $\infty \le \infty$ and $-\infty \le -\infty$ as true. The constraint conditions (C1) and (C2) are needed for the lower limit of β, and the constraint conditions (C3) and (C4) are needed for the upper limit of β. The function $g(\theta_1, \theta_2) = \theta_2/\theta_1$ with $x/0$ defined as ∞ for $x > 0$, meets the constraints when $\Theta_i = [0,1]$. If we had written the constraints with $s, s_1, s_2 \in \Theta_1$ and $t, t_1, t_2 \in \Theta_2$ for all conditions, then there would be problems of needing to define $0/0$ two different ways.

If we allow $\theta_{i,min} = -\infty$ with the understanding that $[\theta_{i,min}, x)$ is interpreted as $(-\infty, x)$ for any x, and analogously, allow $\theta_{i,max} = \infty$, then the 4 constraint conditions will hold for any of the 4 set representations (S1,S2,S3, or S4) for the Θ_i. Note that importantly, the function $g(\theta_1, \theta_2) = \theta_2/\theta_1$ with $x/0$ defined as ∞ for $x > 0$ does not meet the constraint conditions when $\Theta_i = (-\infty, \infty)$.

Non-contiguous sets for Θ_i may be possible, but we do not consider any such situations in this paper.

Web Appendix B Relationship of CIs to p-values

Using expectation and indicator functions, we have

$$\begin{aligned} p_L(\beta_0) &= P_{A,B}\left[L_{\theta_2}(B) \le U_{\theta_1}(A)\right] \\ &= E_{A,B}\left[I\left\{L_{\theta_2}(B) \le U_{\theta_1}(A)\right\}\right] \\ &= E_{A,B}\left[I\left\{g\left[U_{\theta_1}(A), L_{\theta_2}(B)\right] \le \beta_0\right\}\right], \end{aligned}$$

where the last step comes from the monotonicity assumptions on $g(\cdot,\cdot)$ and the assumption that $g(t,t) = \beta_0$ for all t. The last expression equals p if $L_\beta(1-p) = \beta_0$, so this implies that $L_\beta\left\{1 - p_L(\beta_0)\right\} = \beta_0$ and because of the monotonicity assumptions on the $g(\cdot,\cdot)$ and the nestedness assumptions we have $L_\beta\left\{1-p_1\right\} \leq L_\beta\left\{1-p_2\right\}$ when $p_1 < p_2$. So

$$p_L(\beta_0) = \left\{\sup_p L_\beta(1-p) \leq \beta_0\right\}.$$

A similar relationship holds for the upper limit of β.

Web Appendix C Proof of Theorem 1

Let $P_1(a) = P_{\mathbf{X}_1}\left[L_{\theta_1}(\mathbf{X}_1, a) \leq \theta_1\right]$ and $P_2(b) = P_{\mathbf{X}_2}\left[\theta_2 \leq U_{\theta_2}(\mathbf{X}_2, b)\right]$. Let $P_1(a) = a + \epsilon_1(a)$ and $P_2(b) = b + \epsilon_2(b)$, and by the exactness properties $\epsilon_1(a) \geq 0$ and $\epsilon_2(b) \geq 0$. Further, assume $P_1(0) = 0$, so $\epsilon_1(0) = 0$.

Let

$$\Psi = \left\{g(s,t) : L_{\theta_1}(\mathbf{X}_1, a_i) \leq s \text{ and } t \leq U_{\theta_2}(\mathbf{X}_2, b_i), \text{ for } i = 1,\ldots,k\right\}.$$

Then Ψ can be written as the union of k disjoint sets, $\Psi_1,\ldots,\Psi_k$,

$$\Psi_i = \left\{g(s,t) : L_{\theta_1}(\mathbf{X}_1, a_{i-1}) < s \leq L_{\theta_1}(\mathbf{X}_1, a_i) \text{ and } t \leq U_{\theta_2}(\mathbf{X}_2, b_i)\right\}$$

for $i = 1,\ldots,k$, with $L_{\theta_1}(\mathbf{X}_1, 0) \equiv -\infty$. By independence

$$P_{\mathbf{X}}\left[g(\theta_1,\theta_2) \in \Psi_i\right] = \left\{P_1(a_i) - P_1(a_{i-1})\right\} P_2(b_i).$$

Letting $Q = P_{\mathbf{X}}\left[g(\theta_1,\theta_2) \in \Psi\left(\mathbf{X},\mathbf{a},\mathbf{b}\right)\right]$, we have

$$\begin{aligned}
Q &= \sum_{i=1}^{k}\left\{P_1(a_i) - P_1(a_{i-1})\right\} P_2(b_i) \\
&= \sum_{i=1}^{k}\left\{a_i + \epsilon_1(a_i) - a_{i-1} - \epsilon_1(a_{i-1})\right\}\left\{b_i + \epsilon_2(b_i)\right\} \\
&= \sum_{i=1}^{k}\left\{a_i - a_{i-1}\right\} b_i + \sum_{i=1}^{k}\left\{\epsilon_1(a_i) - \epsilon_1(a_{i-1})\right\} b_i + \\
&\qquad \sum_{i=1}^{k}\left\{P_1(a_i) - P_1(a_{i-1})\right\}\epsilon_2(b_i) \\
&= q_{nom} + \sum_{i=1}^{k}\epsilon_1(a_i)b_i - \sum_{i=1}^{k-1}\epsilon_1(a_i)b_{i+1} + \sum_{i=1}^{k}\left\{P_1(a_i) - P_1(a_{i-1})\right\}\epsilon_2(b_i) \\
&= q_{nom} + \epsilon_1(a_k)b_k + \sum_{i=1}^{k-1}\epsilon_1(a_i)(b_i - b_{i+1}) + \sum_{i=1}^{k}\left\{P_1(a_i) - P_1(a_{i-1})\right\}\epsilon_2(b_i)
\end{aligned}$$

By the nestedness condition, $P_1(a_i) - P_1(a_{i-1}) \geq 0$, and by definition, $b_i - b_{i+1} > 0$, so all the terms in the final equation are non-negative, and $Q \geq q_{nom}$. Finally, $P_{\mathbf{X}}\left[g(\theta_1,\theta_2) \leq u(\mathbf{X},\mathbf{a},\mathbf{b})\right] = P_{\mathbf{X}}\left[g(\theta_1,\theta_2) \leq \max\Psi\left(\mathbf{X},\mathbf{a},\mathbf{b}\right)\right] \geq Q$.

Web Appendix D Proof of Theorem 2

Let W_q be the set $\{u(\mathbf{X},\mathbf{a},\mathbf{b}) :\ \mathbf{a}$ and $\mathbf{b}$ ranging over all finite dimensional vectors of ordered probabilities such that $\sum(a_i - a_{i-1})b_i \geq q\}$. We will prove that $U_\beta(q) = \inf(W_q)$. The proof has 2 parts: 1) that $U_\beta(q)$ is a lower bound on W_q, and 2) that there is no greater lower bound on W_q than $U_\beta(q)$.

Part 1: $U_\beta(q)$ is a lower bound on W_q. The area below the level curve $g(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)) = U_\beta(q)$ is q. For any $c < U_\beta(q)$, the area below the level curve $g(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)) = c$ is strictly less than q by the monotonicity properties on g. It follows that the sums of areas of any rectangles contained within the region $g(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)) \leq c$ (upper panel of Figure 5) must also be strictly less than q. Therefore, for no value $c < U_\beta(q)$ can there exist finite dimensional vectors of ordered probabilities $\mathbf{a}$ and $\mathbf{b}$ such that each rectangle is contained in the region $g(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)) \leq c$ and $\sum(a_i - a_{i-1})b_i \geq q$. In other words, $U_\beta(q)$ is a lower bound on W_q.

Part 2: There is no greater lower bound on W_q than $U_\beta(q)$. Let $c > U_\beta(q)$. Then the area below the level curve $g(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)) = c$ is $r > q$. For each a, the function $h(a) = \sup\{b : \; g(L_{\theta_1}(a), U_{\theta_2}(b)) \leq c\}$ is a Riemann integrable function because h is a monotone function of a. Also, $\int_0^1 h(a)da$ is the area under the level curve $g(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)) = c$, namely $r > q$. It follows from the fact that h is Riemann integrable that for any $\epsilon > 0$, there is a partition $a_1 < a_2 < \ldots < a_k$ such that $\sum(a_i - a_{i-1})h(a_{i-1}) \geq r - \epsilon$ (bottom panel of Figure 5). If we choose $\epsilon = (r-q)/2$, for example, then $\sum(a_i - a_{i-1})h(a_{i-1}) \geq q$. In other words, c is not a lower bound on W_q. Therefore, no value larger than $U_\beta(q)$ can be a lower bound on W_q, completing the proof.

Web Appendix E Proof of Theorem 3 and Application to Binomial Case

Web Appendix E.1 Proof of Theorem 3

For notational ease assume that $n_1 = n_2 = n$. We relax this assumption at the end. Assume that all random variables other than the auxilliary variables A and B are defined on a probability space $(\Omega, \mathcal{F}, P)$. Denote the random variables from the two samples as $\mathbf{X}^{(n)}(\omega) = [\mathbf{X}_1^{(n)}(\omega), \mathbf{X}_2^{(n)}(\omega)]$. The probability space $[0,1] \times [0,1]$ equipped with Lebesgue measure on R^2 governs the generation of A, B. We can define all random variables on the same product space $\Omega \times [0,1] \times [0,1]$ using product measure, but it is sometimes helpful to separate them. In particular, we sometimes fix ω, which fixes $\mathbf{X}^{(n)}(\omega)$, and focus on the resulting sequence of random variables on the (A, B) probability space. For notational ease, we sometimes suppress ω, writing $L_{\theta_i}(A)$ for $L_{\theta_i}(\mathbf{X}_i^{(n)}(\omega), A)$, for example. Let $U_{\beta,n}(q) \equiv U_\beta(\mathbf{X}^{(n)}(\omega), q)$ be the $100q\%$ upper melded confidence limit formed from L_{θ_i} and U_{θ_i}.

By definition, $U_{\beta,n}(q)$ is the solution to

$$P_{A,B}\left[g\left\{U_{\theta_1}(A), L_{\theta_2}(B)\right\} \leq U_{\beta,n}(q)\right] = q.$$

The notation $P_{A,B}$ emphasizes that ω is fixed and A and B are random.

We now use the multivariate Taylor theorem (see e.g., Serfling, 1980, p.44). Let

$$g^{(1,0)}(\theta_1, \theta_2) = \left(\frac{\partial g(s,t)}{\partial s}\right)_{s=\theta_1, t=\theta_2} \text{ and } g^{(0,1)}(\theta_1, \theta_2) = \left(\frac{\partial g(s,t)}{\partial t}\right)_{s=\theta_1, t=\theta_2}$$

Expand $g\left\{U_{\theta_1}(A), L_{\theta_2}(B)\right\}$ about $g\left\{U^Z_{\theta_1}(A), L^Z_{\theta_2}(B)\right\}$ to get

$$\begin{aligned}&P_{A,B}\left[g\left\{U^Z_{\theta_1}(A), L^Z_{\theta_2}(B)\right\} + R_n(A,B) \leq U_{\beta,n}(q)\right] = q, \text{ where}\\ &R_n(A,B) = g^{(1,0)}(s_n, t_n)\left\{U_{\theta_1}(A) - U^Z_{\theta_1}(A)\right\} + g^{(0,1)}(s_n, t_n)\left\{L_{\theta_2}(B) - L^Z_{\theta_2}(B)\right\},\end{aligned}$$

and (s_n, t_n) lies on the line segment joining $\{U_{\theta_1}(A), L_{\theta_2}(B)\}$ and $\left\{U^Z_{\theta_1}(A), L^Z_{\theta_2}(B)\right\}$.

For fixed (ω, A, B) outside a set of probability 0, the following hold:

1. $U^Z_{\theta_1}(A) \to \theta_1$,
2. $L^Z_{\theta_2}(B) \to \theta_2$,
3. $\sqrt{n}\left\{U_{\theta_1}(A) - U^Z_{\theta_1}(A)\right\} \to 0$, and
4. $\sqrt{n}\left\{L_{\theta_2}(B) - L^Z_{\theta_2}(B)\right\} \to 0$.

Therefore, again for for fixed (ω, A, B), $g^{(1,0)}(s_n, t_n) \to g^{(1,0)}(\theta_1, \theta_2)$. Similarly, $g^{(0,1)}(s_n, t_n) \to g^{(0,1)}(\theta_1, \theta_2)$. This implies that

$$\sqrt{n}R_n(A,B) \to 0 \text{ for fixed } (\omega, A, B). \tag{7}$$

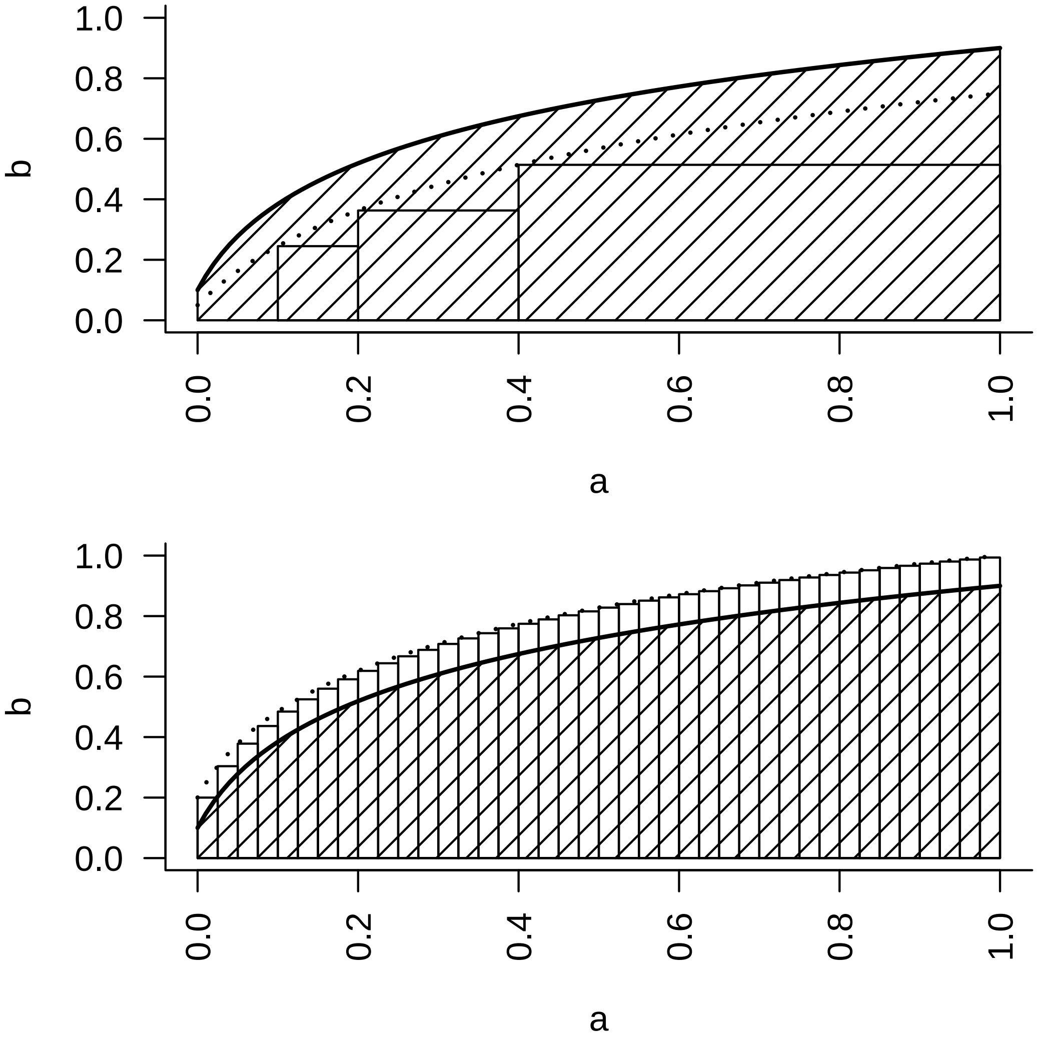

Figure 5: Figure to accompany Proof of Theorem 2, upper for part 1, lower for part 2. Solid lines represent the level curve $g\left(L_{\theta_1}(\cdot), U_{\theta_2}(\cdot)\right) = U_\beta(q)$.

Use the multivariate Taylor theorem again, but now expand

$$g\left\{U^Z_{\theta_1}(A), L^Z_{\theta_2}(B)\right\} = g\left\{\hat{\theta}_1 + \Phi^{-1}(1-A)\frac{\hat{\sigma}_1}{\sqrt{n}}, \hat{\theta}_2 - \Phi^{-1}(1-B)\frac{\hat{\sigma}_2}{\sqrt{n}}\right\}$$

about $g(\hat{\theta}_1, \hat{\theta}_2)$:

$$\begin{aligned} &P_{A,B}\left[g\left(\hat{\theta}_1, \hat{\theta}_2\right) + R^*_n(A,B) + R_n(A,B) \leq U_{\beta,n}(q)\right] = q, \text{ where} \\ &R^*_n(A,B) = g^{(1,0)}(v_n, w_n)\left\{\Phi^{-1}(1-A)\frac{\hat{\sigma}_1}{\sqrt{n}}\right\} + g^{(0,1)}(v_n, w_n)\left\{-\Phi^{-1}(1-B)\frac{\hat{\sigma}_2}{\sqrt{n}}\right\}, \end{aligned}$$

and (v_n, w_n) lies on the line segment joining $\left(U^Z_{\theta_1}(A), L^Z_{\theta_2}(B)\right)$ and $\left(\hat{\theta}_1, \hat{\theta}_2\right)$.

Now let

$$\tau = \tau(\theta_1, \theta_2, \sigma_1, \sigma_2) = \sqrt{\left\{g^{(1,0)}(\theta_1, \theta_2)\right\}^2 \sigma_1^2 + \left\{g^{(0,1)}(\theta_1, \theta_2)\right\}^2 \sigma_2^2} \text{ and } \hat{\tau} = \tau(\hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}_1, \hat{\sigma}_2).$$

Rearrange terms and divide both sides of the expression within the $P_{A,B}$ statement by $\hat{\tau}/\sqrt{n}$ to get

$$\begin{aligned} P_{A,B}\Bigg[\frac{g^{(1,0)}(v_n, w_n)\Phi^{-1}(1-A)\hat{\sigma}_1}{\hat{\tau}} \quad &- \quad \frac{g^{(0,1)}(v_n, w_n)\Phi^{-1}(1-B)\hat{\sigma}_2}{\hat{\tau}} + \frac{\sqrt{n}R_n(A,B)}{\hat{\tau}} \leq \\ &\frac{\sqrt{n}\left\{U_{\beta,n}(q) - g\left(\hat{\theta}_1, \hat{\theta}_2\right)\right\}}{\hat{\tau}}\Bigg] = q. \end{aligned} \tag{8}$$

By expression 7, for fixed (ω, A, B), $\frac{\sqrt{n}R_n(A,B)}{\hat{\tau}} \to 0$. We can show that $g^{(1,0)}(v_n, w_n) \to g^{(1,0)}(\theta_1, \theta_2)$ and $g^{(0,1)}(v_n, w_n) \to g^{(0,1)}(\theta_1, \theta_2)$.

Denote the left-hand-side of the inequality within $P_{A,B}$ as $\psi_{\omega,n}(A,B)$. For fixed ω, the sequence of random variables $\psi_{\omega,n}(A,B)$ on the (A,B) probability space converges almost surely with respect to two dimensional Lebesgue measure to

$$\frac{g^{(1,0)}(\theta_1, \theta_2)\sigma_1\Phi^{-1}(1-A) - g^{(0,1)}(\theta_1, \theta_2)\sigma_2\Phi^{-1}(1-B)}{\tau}, \tag{9}$$

another random variable on the (A,B) probability space. This implies that the distribution $F_n(x)$ of $\psi_{\omega,n}$ (holding ω fixed) converges to the distribution of (9), namely the standard normal distribution function $\Phi(x)$. Moreover, because the limiting distribution is continuous, $F_n(x)$ converges to $\Phi(x)$ uniformly in x by Polya's theorem (see e.g., Serfling, 1980). Because (8) holds for every n, it must be the case that the right side of the inequality must converge to $\Phi^{-1}(q)$, i.e., for fixed ω,

$$\frac{\sqrt{n}\left\{U_{\beta,n}(q) - g\left(\hat{\theta}_1, \hat{\theta}_2\right)\right\}}{\tau} \to \Phi^{-1}(q).$$

In other words, the melded confidence interval is asymptotically equivalent to treating $(\hat{\theta}_1, \hat{\theta}_2)$ as independent normals with means (θ_1, θ_2) and variances $(\sigma_1^2/n, \sigma_2^2/n)$, and then applying the delta method (see e.g., Lehmann, 1999) to $g(\hat{\theta}_1, \hat{\theta}_2)$. Finally, we can relax the assumption that $n_1 = n_2$, as long as we assume that $n_1/(n_1+n_2) \to \lambda$ with $0 < \lambda < 1$. In this case, the proof is very similar but just more notationally cumbersome and hence will not be shown.

Web Appendix E.2 Applying Theorem 3 to the Binomial Case

Consider the binomial case, with for a single sample, $X \sim Binomial(n, \theta)$ (we drop the subscripts for the ith sample in this section). We show the asymptotic equivalence of the upper confidence limit to the usual normal theory upper limit. The result for the lower limit is analogous. Let the upper Clopper-Pearson confidence limit be $U_n(x) = U_\theta(x, 1-\alpha)$ and the usual normal theory upper confidence limit be

$$U_n^Z(x) = U_\theta^Z(x, 1-\alpha) \quad = \quad \frac{x}{n} + \Phi^{-1}(1-\alpha)\sqrt{\frac{\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)}{n}}.$$

By definition, we have

$$\alpha \quad = \quad \sum_{k=0}^{x} \left(\begin{array}{c} n \\ k \end{array} \right) U_n(x)^k \left(1 - U_n(x)\right)^{n-k} \tag{10}$$

Let $R_n(x) = U_n(x) - U_n^Z(x)$. Because $U_n(x) \overset{a.s.}{\to} \theta$ and $U_n^Z(x) \overset{a.s.}{\to} \theta$, $R_n(x) \overset{a.s.}{\to} 0$.

From equation 10 and the Berry-Essen Theorem (see e.g., Lehmann, 1999, p. 78),

$$\alpha = \Phi\left(\frac{x - nU_n(x)}{\sqrt{nU_n(x)(1-U_n(x))}}\right) + R_n^*, \tag{11}$$

where (using $C = 1$ for the constant in the Berry-Essen theorem)

$$|R_n^*| \leq \frac{\mu_3(U_n(x))}{\sqrt{n}\left(\sqrt{U_n(x)(1-U_n(x))}\right)^3},$$

where for a given p, $\mu_3(p) = E(|X-p|^3)$. Because $U_n(x) \overset{a.s.}{\to} \theta$, $R_n^* \overset{a.s.}{\to} 0$. Now substitute $U_n(x) = U_n^Z(x) + R_n(x)$ into equation 11:

$$\begin{aligned}\alpha \quad &= \quad \Phi\left(\frac{x - x - \Phi^{-1}(1-\alpha)\sqrt{n\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)} - nR_n(x)}{\sqrt{nU_n(x)(1-U_n(x))}}\right) + R_n^*, \\ &= \quad \Phi\left(\frac{-\Phi^{-1}(1-\alpha)\sqrt{\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)}}{\sqrt{U_n(x)(1-U_n(x))}} - \frac{\sqrt{n}R_n(x)}{\sqrt{U_n(x)(1-U_n(x))}}\right) + R_n^*, \end{aligned} \tag{12}$$

From equation 12, because $R_n^* \overset{a.s.}{\to} 0$, as $n \to \infty$,

$$\frac{-\Phi^{-1}(1-\alpha)\sqrt{\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)}}{\sqrt{U_n(x)(1-U_n(x))}} - \frac{\sqrt{n}R_n(x)}{\sqrt{U_n(x)(1-U_n(x))}} \overset{a.s.}{\to} -\Phi^{-1}(1-\alpha)$$

Since

$$\frac{\sqrt{\left(\frac{x}{n}\right)\left(1-\frac{x}{n}\right)}}{\sqrt{U_n(x)(1-U_n(x))}} \overset{a.s.}{\to} 1$$

and $\sqrt{U_n(x)(1-U_n(x))} \overset{a.s.}{\to} \sqrt{\theta(1-\theta)}$, in this case we must have $\sqrt{n}R_n(x) \overset{a.s.}{\to} 0$.

Web Appendix F Two Poisson Variables

Note that (see Casella and Berger, 2002, p. 100)

$$1 - F(x-1; n\theta) = G\left(\theta; x, \frac{1}{n}\right) \tag{13}$$

where F is the Poisson cdf with mean $n\theta$ and G is the gamma cdf with parameters x and $1/n$ (i.e., with mean x/n and variance x/n^2). We extend the definition of the gamma cdf so that $G(t;0,b) = I(t>0)$, i.e., the distribution represents a point mass at zero. So $L_{\theta_1}(x_1, A) \sim G(\theta_1; x_1, 1/n_1)$. Similarly, using equation 13, we get $Pr[X_2 \le x_2; n_2\theta_2] = 1 - G\left(\theta; x_2+1, \frac{1}{n_2}\right)$, so that $U_{\theta_2}(x_2, B) \sim G\left(\theta; x_2+1, \frac{1}{n_2}\right)$. The melded p-value is

$$p_U(x_1,x_2) = Pr\left[T_1 < \frac{T_2}{r} \middle| T_1 \sim G\left(T_1; x_1, \frac{1}{n_1}\right), T_2 \sim G\left(T_2; x_2+1, 1/n_2\right)\right].$$

We now show $p_b(x_1,x_2) = p_U(x_1,x_2)$. First, consider the case where $x_1 = 0$. In this case, we can see by inspection that $p_b(0,x_2) = p_U(0,x_2) = 1$. Now consider when $x_1 > 0$.

$$\begin{aligned}
p_U(x_1,x_2) &= \int_0^\infty G(t;x_1,1/n_1)\,r g(tr;x_2+1,1/n_2)dt \\
&\quad \text{where } g(t;a,b) \text{ is a gamma pdf with parameters } a \text{ and } b \\
&= \int_0^\infty \left\{1 - F(x_1-1;n_1 t)\right\} r g(tr;x_2+1,1/n_2)dt \\
&\quad \text{by equation 13} \\
&= 1 - \int_0^\infty \left\{\sum_{x=0}^{x_1-1} \frac{(n_1 t)^x e^{-n_1 t}}{x!}\right\} \frac{1}{\Gamma(x_2+1)(1/n_2)^{x_2+1}} t^{x_2} r^{x_2+1} e^{-t n_2^*} dt,
\end{aligned}$$

where $n_2^* = r n_2$. So,

$$\begin{aligned}
p_U(x_1,x_2) &= 1 - \sum_{x=0}^{x_1-1} \frac{n_1^x n_2^{*x_2+1}}{x! x_2!} \int_0^\infty t^{x+x_2} e^{-t(n_1+n_2^*)} dt \\
&= 1 - \sum_{x=0}^{x_1-1} \frac{n_1^x n_2^{*x_2+1}(x+x_2)!}{x! x_2! (n_1+n_2^*)^{x+x_2+1}} \int_0^\infty \frac{1}{\Gamma(x+x_2+1)\left(\frac{1}{n_1+n_2^*}\right)^{x+x_2+1}} t^{x+x_2} e^{-t(n_1+n_2^*)} dt \\
&= 1 - \sum_{x=0}^{x_1-1} \frac{n_1^x n_2^{*x_2+1}(x+x_2)!}{(n_1+n_2^*)^{x+x_2+1} x! x_2!} \int_0^\infty g(t; x+x_2+1, \frac{1}{n_1+n_2^*}) dt \\
&= 1 - \sum_{x=0}^{x_1-1} \binom{x+x_2}{x} \left(\frac{n_2^*}{n_1+n_2^*}\right)^{x_2+1} \left(\frac{n_1}{n_1+n_2^*}\right)^x \\
&= 1 - F_{NB}\left(x_1-1; x_2+1, \frac{n_2^*}{n_1+n_2^*}\right)
\end{aligned}$$

where $F_{NB}(x;r,p)$ is the cdf of a negative binomial distribution. Continuing,

$$p_U(x_1,x_2) = F_{Bin}\left(x_2; x_1+x_2, \frac{n_2^*}{n_1+n_2^*}\right),$$

where F_{Bin} is the binomial cdf (see e.g., Casella and Berger, 2002, p. 130 for the NB-Bin relationship). Finally,

$$p_U(x_1,x_2) = 1 - F_{Bin}\left(x_1-1; x_1+x_2, \frac{n_1}{n_1+n_2^*}\right) = p_b(x_1,x_2).$$

Web Appendix G One-Sample Median Confidence Procedure Allowing for Ties

Suppose[1] $Y_1, Y_2, \ldots, Y_n$ are independent and all distributed according to the distribution F. A median is defined as a value θ such that $F(\theta) \equiv P[Y_j \le \theta] \ge \frac{1}{2}$ and $\bar{F}(\theta) \equiv P[Y_j \ge \theta] \ge \frac{1}{2}$. For a continuous distribution and many discrete distributions

[1] This section has been corrected from the original submission.

the median is unique, but for a discrete distribution where a median θ has $P[Y_j \leq \theta] = F(\theta) = \frac{1}{2}$ or $P[Y_j \geq \theta] \equiv \bar{F}(\theta) = \frac{1}{2}$ then the set of medians is an interval. For example, if the sample space of Y_j is $\{1, 2, 3, 4\}$ with an equal probability of $\frac{1}{4}$ on each value, then the set of medians is $[2, 3]$. Let the set of medians be denoted $[\theta_{min}, \theta_{max}]$, where if the median is unique then $\theta = \theta_{min} = \theta_{max}$. We define a valid lower one-sided confidence limit for the set of medians as $L(\mathbf{y}, q)$ such that $P[L(\mathbf{Y}, q) \leq \theta] \geq q$ for all θ that are medians. This is equivalent to $P[L(\mathbf{Y}, q) \leq \theta_{min}] \geq q$. Similarly, define a valid upper one-sided confidence limit for the set of medians as $U(\mathbf{y}, q)$ such that $P[U(\mathbf{Y}, q) \geq \theta] \geq q$ for all θ that are medians, implying that $P[U(\mathbf{Y}, q) \geq \theta_{max}] \geq q$.

A one-sided test of interest is

$$\begin{aligned} H_0 : \theta_{min} &\leq \theta_0 \\ H_1 : \theta_{min} &> \theta_0, \end{aligned}$$

and the alternative represents the state where all medians are greater than θ_0. So inverting that hypothesis test, we get the one-sided confidence set bounded below by $L(\mathbf{y}, 1 - \alpha)$, which is the set of θ that fail to reject the null hypothesis at level α.

Let $W_n(\theta) = \sum_{i=1}^{n} I\{Y_i \leq \theta\}$, and for any median, θ, we have $F(\theta) \geq 0.5$. Then $W_n(\theta) \sim Binomial(n, F(\theta))$. Let $W_n^* \sim Binomial(n, 0.5)$. Then for any w,

$$P[W_n(\theta) \leq w] \leq P[W_n^* \leq w]. \tag{14}$$

So a one-sided test bounds the type I error rate at level α, and rejects when $w_n(\theta_0) \leq c_n(\alpha)$, where $w_n(\theta_0) = \sum_{i=1}^{n} I\{y_i \leq \theta_0\}$ is the realized value of $W_n(\theta_0)$, and $c_n(\alpha)$ is defined such that

$$\begin{aligned} & P[W_n^* \leq c_n(\alpha)] \leq \alpha < P[W_n^* \leq c_n(\alpha) + 1] \\ = \; & F_{Bin}[c_n(\alpha); n, 0.5] \leq \alpha < F_{Bin}[c_n(\alpha) + 1; n, 0.5]. \end{aligned}$$

A $q = 1 - \alpha$ confidence set is the set of θ_0 that fail to reject. Let $t_1 < t_2 < \cdots < t_k$ be the unique ordered values of $Y_1, \ldots, Y_n$. So, letting $w_n(t_0) = 0$, the lower limit is

$$L_\theta(1 - \alpha) = \begin{cases} -\infty & \text{if } c_n(\alpha) < 0 \\ t_j & \text{if } w_n(t_{j-1}) \leq c_n(\alpha) < w_n(t_j) \text{ for } j = 1, 2, \ldots, k \end{cases}$$

So applying the $F_{Bin}(\cdot; n, 0.5)$ function to both sides of the conditions, and using the definition of $c_n(\alpha)$, we can write $L_\theta(1-a)$ as

$$L_\theta(1 - a) = \begin{cases} -\infty & \text{if } 0 \leq a < F_0 \\ t_1 & \text{if } F_0 \leq a < F_1 \\ \vdots & \vdots \\ t_j & \text{if } F_{j-1} \leq a < F_j \\ \vdots & \vdots \\ t_k & \text{if } F_{k-1} \leq a < F_k = 1 \end{cases} \tag{15}$$

where $F_j = F_{Bin}[w_n(t_j); n, 0.5]$, and $F_0 = F_{Bin}(0; n, 0.5)$.

The upper limit is analogous. The one-sided test of interest is now,

$$\begin{aligned} H_0 : \theta_{max} &\geq \theta_0 \\ H_1 : \theta_{max} &< \theta_0, \end{aligned}$$

and the alternative represents the state where all medians are less than θ_0. So inverting that hypothesis test, we get the one-sided confidence set bounded above by $U(\mathbf{y}, 1 - \alpha)$. Let $\bar{W}_n(\theta) = \sum_{i=1}^{n} I\{Y_i \geq \theta\}$. Then $\bar{W}_n(\theta) \sim Binomial(n, \bar{F}(\theta))$. For any median, θ, we have $\bar{F}(\theta) \geq 0.5$, so that for any w,

$$P[\bar{W}_n(\theta) \geq w] \leq P[W_n^* \geq w], \tag{16}$$

where as before $W_n^* \sim Binom(n, 0.5)$. So a one-sided test that bounds the type I error rate at level α, rejects when $\sum_{i=1}^n I(y_i \geq \theta_0) = \bar{w}_n(\theta_0) \leq c_n(\alpha)$. The upper $q = 1 - \alpha$ confidence limit is the largest θ_0 that fails to reject. In other words,

$$U_\theta(1-\alpha) \quad = \quad \begin{cases} \infty & \text{if } c_n(\alpha) < 0 \\ t_j & \text{if } \bar{w}_n(t_{j+1}) \leq c_n(\alpha) < \bar{w}_n(t_j) \text{ for } j = k, k-1, , \ldots, 1 \end{cases}$$

where $\bar{w}_n(t_{k+1}) \equiv 0$. Now we apply the function $F(\cdot; n, 0.5)$ on the conditions. Using the definition of $\bar{c}_n(\alpha)$, this gives,

$$U_\theta(q) \quad = \quad \begin{cases} \infty & \text{if } \alpha < F_{k+1}^* \\ t_j & \text{if } F_{j+1}^* \leq \alpha < F_j^* \text{ for } j = k, k-1, \ldots, 1 \end{cases} \tag{17}$$

where $F_j^* = F(\bar{w}_n(t_j))$ and $F_{k+1}^* = F(\bar{w}_n(t_{k+1})) = F_0$.

Web Appendix H Survival Simulation Details

For Figure 4a, the distribution for the control arm (dotted gray) is a mixture of two exponentials (one with mean 0.1 and the other with mean 10), while the new treatment arm (solid black) is a mixture of a Weibull (with shape 10 and scale 0.8) and an exponential (with mean 10). For each arm, we randomly choose an indicator with probability of $p_a = 30\%$ of having the first of the mixtures. The distributions for Figure 4b are the same except $p_a = 80\%$. For Figures 4a and 4b, the one year survival times for both arms are equal to the nearest 0.0001, and equal to $S_0(1) = S_1(1) = 0.6334$ (Figure 4a) or $S_0(1) = S_1(1) = 0.1810$ (Figure 4b). For Figure 4c, the control arm is exponential with mean 1, while the treatment arm is a mixture, 30% Weibull (shape 1, scale 0.05) and 70% exponential with mean 50. For Figure 4d, the control arm is the same as for the control in Figure 4b, while the treatment arm is a mixture, 80% Weibull (shape 1, scale 10) and 20% exponential with mean 10. The moderate censoring has all subjects exposed to independent uniform censoring on 0.5 to 1.2, while the heavy censoring randomly chooses 90% of the subjects to have uniform censoring on 0.5 to 0.8 and the others to have uniform censoring on 0.8 to 1.2.

References

Casella, G. and Berger, R. L. (2002). *Statistical Inference, second edition.* Duxbury Press.

Lehmann, E. (1999). *Elements of Large Sample Theory.* Springer.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics.* Wiley, New York.