Alternative Study Designs and Nonparametric Statistical Methods for Adaptive Management Studies of Invasive Plants

James N. McNair, Daniel Frobish, Emma K. Rice, and Ryan A. Thum

Supplementary Appendix S1: Statistical Details and Examples

This document provides additional details regarding the statistical methods presented in the main text, including outlines of the underlying statistical theory as well as empirical examples using real data. Question numbers are the same as in the main text, with the exception of three "sub-questions" (Questions 8a, 8b, and 8c) in this appendix that are extensions of Questions 8 but, to shorten the main text, are not addressed there.

1 Management Experiments Using Marked Plants

1.1 Assessing the efficacy of a single management treatment on a single species

Question 1: What is the PET and 95% confidence interval for a given test treatment?

The goal here is to estimate the PET and its 95% confidence interval for a particular species, treatment type, and treatment level.

Assumptions and notation: We assume there are n marked plants labeled $1, 2, \ldots, n$. A test treatment is applied in a uniform manner to each of the n plants, which then respond independently. The treatment outcome for each plant is a Bernoulli random variable, with treatment being effective with probability p and ineffective with probability $1 - p$. Let random variable N_1 be the number of plants for which treatment is effective (the "successes"), and let $N_2 := n - N_1$ be the number for which treatment is ineffective (the "failures"). Then N_1 is binomially distributed with parameters n and p, expected value np, and variance $np(1 - p)$. Moreover, for n sufficiently large, N_1 is approximately normally distributed with mean np and variance $np(1 - p)$.

Statistical method: Let n_1 and n_2 denote the observed values of N_1 and N_2 in a particular management experiment. To answer Question 1 adequately, we require estimators for binomial parameter p (with n, n_1, and n_2 known) and its upper and lower $100(1 - \alpha)\%$ confidence limits, where α is the chosen significance level (usually 0.05). The maximum likelihood estimator $\hat{p}$ for PET parameter p is given by

$$\hat{p} = N_1/n \tag{1}$$

(e.g., Agresti, 2013, p. 10).

Numerous methods have been proposed for estimating confidence intervals for a binomial success probability. The simplest is the Wald confidence interval, which Brown et al. (2001) call the "standard interval" because it is the most common interval presented in introductory statistics textbooks. Unfortunately, it produces a confidence interval that on average corresponds to a true confidence level well below the nominal

$1 - \alpha$ level unless n is very large (much greater than 100) and therefore should not be used (Agresti, 2013; Agresti & Coull, 1998; Brown et al., 2001). Based on simulation studies (Agresti & Coull, 1998; Brown et al., 2001), two methods that are almost as simple as the Wald interval but perform much better are the Wilson interval and the Agresti-Coull interval, with the former performing slightly better than the latter except when p is very close to 0 or 1. The $100(1 - \alpha)\%$ Wilson confidence interval is given by

$$\tilde{p} \pm \frac{z_{\alpha/2}}{\tilde{n}} \sqrt{N_1(1 - N_1/n) + z^2_{\alpha/2}/4}, \tag{2}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution and

$$\tilde{n} = n + z^2_{\alpha/2}, \quad \tilde{p} = (N_1 + z^2_{\alpha/2}/2)/\tilde{n}$$

(Brown et al., 2001; Agresti, 2013: pp. 14–15). The $100(1 - \alpha)\%$ Agresti-Coull confidence interval is given by

$$\tilde{p} \pm z_{\alpha/2} \sqrt{\tilde{p}(1 - \tilde{p})/\tilde{n}} \tag{3}$$

(Brown et al., 2001; Agresti, 2013: pp. 32–33). As usual, the upper and lower limits of both of these intervals are random variables; statistical software reports an estimate of each interval obtained by replacing random variable N_1 by observed value n_1 in Eqs. (2) and (3).

R functions: Wilson and Agresti-Coull confidence intervals can be computed with the `binom.confint()` function in R package `binom` (Dorai-Raj, 2022), which also has function `binom.coverage()` for computing the coverage probability (the proportion of a large number of simulated trials for which the estimated confidence interval includes the true value of p).

Example: Preventing post-removal resprouting by Norway maple saplings. As part of an adaptive management project conducted by one of the authors (JNM) to assist managers of the extensive natural lands of the Fairmount Park System in Philadelphia, Pennsylvania (USA), a field experiment was conducted to assess two methods of removing invasive Norway maple saplings (cutting down, and cutting down followed by stump application of the herbicide, triclopyr) from urban forests in the park system. Each removal method was randomly assigned to 60 marked saplings. Saplings were cut down during the growing season, and stumps in the triclopyr group were immediately treated. All stumps were checked one year later to determine the number in each treatment group for which the removal method had been effective, meaning that no visible resprouting from the stump or roots had occurred. The results are shown in Table 1, along with maximum likelihood estimates n_1/n of p and the Wilson and Agresti-Coull 95% confidence limits as computed by the `binom.confint()` function in R. Note that regardless of which type of confidence interval is used, the estimated PET for the Cut + Triclopyr treatment is greater than that for the Cut treatment, and the 95% confidence intervals for the two treatments do not overlap.

Table 1. Estimates and 95% Wilson and Agresti-Coull confidence intervals (CIs) for PETs for two different treatments in the Norway maple management experiment.

Treatment	n_1	n	PET	CI type	Lower limit	Upper limit
Cut	12	60	0.200	Wilson	0.118	0.318
				Agresti-Coull	0.117	0.319
Cut + Triclopyr	58	60	0.967	Wilson	0.886	0.991
				Agresti-Coull	0.880	0.997

Question 2: Does the PET for a given test treatment exceed prescribed management threshold $p^\star$?

The goal now is to determine whether there is strong evidence that management objective PET $> p^\star$ has been achieved for a particular species, treatment type, and treatment level, where $p^\star$ is a threshold PET to be exceeded.

Statistical method: The null hypothesis here is H_0: $p - p^\star = 0$, and the relevant alternative is the one-sided hypothesis H_1: $p - p^\star > 0$. The main tests are a large-sample test for proportions and, for small samples, the exact binomial test and the mid-P binomial test. Simulation studies show that the exact binomial test is unduly conservative; the false-positive error rate of the mid-P binomial test is closer to the nominal α.

Assumptions and notation: The assumptions here are the same as for Question 1.

Statistical method: The maximum likelihood estimator $\hat{p}$ for p is again given by Eq. (1). The null hypothesis to be tested is H_0: $p - p^\star = 0$, and since $p^\star$ is a management threshold to be exceeded, the relevant alternative hypothesis is H_1: $p - p^\star > 0$. Under H_0, random variable $N_1 - np^\star$ has mean 0 and variance $np^\star(1 - p^\star)$. For sufficiently large n, then, the standardized random variable V defined by

$$V = \frac{N_1 - np^\star}{\sqrt{np^\star(1 - p^\star)}} \tag{4}$$

will be approximately normally distributed with mean 0 and variance 1 (Hollander et al., 2014: p. 13). There is uncertainty as to exactly what "sufficiently large" means; by analogy with methods for comparing two success probabilities discussed in section 1.2 below, we may interpret it to mean that the expected number of "successes" $np^\star$ and the expected number of "failures" $n(1 - p^\star)$ under H_0 both should be at least 5. The P-value for H_0 is simply the probability that $Z > v$, where Z is a standard normal random variable and v is the observed value of V. We reject H_0 if and only if $P \leq \alpha$, where α is the chosen significance level.

If the expected number of successes or failures is less than 5, or if one wishes to remove uncertainty regarding adequacy of the normal approximation underlying the large-sample test, some form of binomial test can be used (Agresti, 2013: pp. 16–17). For the exact binomial test, the P-value is the probability of obtaining a value of N_1 as extreme or more so than the observed value if H_0 were true. Thus, it is the probability that $N_1 \geq n_1$ under H_0, where N_1 has a binomial distribution with parameters n and $p^\star$ (equivalently, it is the probability that $N_1 > n_1 - 1$, which is more convenient to compute with the complementary distribution function). The exact binomial test, however, is unduly conservative in the sense that its actual false-positive (Type-1) error rate tends to be meaningfully less than the nominal rate α. The mid-P version of this test (which we will call the mid-P binomial test) is less conservative, yielding a false-positive error rate closer to the nominal α level, though there is no guarantee it will not slightly exceed α (Agresti, 2013: p. 17). The P-value for the mid-P test is simply the P-value for the exact binomial test minus one-half the probability of the event $\{N_1 = n_1\}$ under H_0. That is,

$$\text{mid-}P\ P\text{-value} = \Pr\{N_1 > n_1 - 1 \,|\, n, p^\star\} - \tfrac{1}{2}\Pr\{N_1 = n_1 \,|\, n, p^\star\}. \tag{5}$$

R functions: The large-sample test referred to can be performed with the `prop.test()` function in R, or simply by using the `pnorm()` function. For small samples, an exact binomial test can be performed with R's `pbinom()` or `binom.test()` function. The resulting P-value can be adjusted downward to obtain the mid-P P-value using R's `dbinom()`function (see section 1.2 of SI1). All of these functions are included in the `stats` package (R Core Team, 2023), which is part of R's standard library.

Example: Preventing post-removal resprouting by Norway maple saplings (continued). For the Norway maple example, suppose (arbitrarily) that a method for cutting down saplings is desired for which the probability of effective treatment is greater than $p^\star = 0.8$. The data in Table 1 show that the expected numbers

of "successes" and "failures" for both experimental groups are $np^\star = 48$ and $n(1 - p^\star) = 12$. Since both of these expected numbers exceed 5, the large-sample test based on V should be adequate. For purposes of illustration, we apply the large-sample test (using the `prop.test()` function in R) as well as the exact and mid-P binomial tests (using the `binom.test()` and `dbinom()` functions in R). The results are shown in Table 2. Regardless of which test is used, there is strong evidence that the success probability exceeds the management threshold of 0.8 for the Cut + Triclopyr treatment but not for the Cut treatment.

Table 2. Large-sample, exact binomial, and mid-P binomial tests of the null hypothesis H_0: $p - p^\star = 0$ versus alternative hypothesis H_1: $p - p^\star > 0$ for the Norway maple sapling data presented in Table 1, with $p^\star = 0.8$. Both unadjusted (Raw) and Holm-adjusted (Adj.) P-values are shown.

		Large-sample test		Exact binomial test		Mid-P binomial test	
Treatment	n_1/n	Raw P-val.	Adj. P-val.	Raw P-val.	Adj. P-val.	Raw P-val.	Adj. P-val.
Cut	0.200	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Cut + Triclopyr	0.967	0.0006	0.0012	0.0002	0.0004	0.0001	0.0002

1.2 Comparing treatment efficacy in pairs of treatments or species

Question 3: Do PETs p_A and p_B for plant groups A and B differ?

Here the goal is to determine whether there is strong evidence that the PETs for two groups of plants differ. In the most common situation, the two groups are plants of a target invasive species that receive different management treatments, one of which might be the null treatment. The null hypothesis is that there is no difference; the alternative hypothesis can be either two-sided or one-sided, depending on whether both or only one alternative is of interest.

Assumptions and notation: The assumptions and notation are similar to those for Questions 1 and 2, except that two treatment groups are assessed simultaneously. (If there are more than two groups, they are assessed pairwise and the P-values are Holm-adjusted.) Let the two groups be denoted A and B. We assume there are n_A individuals in group A (labeled $1, 2, \ldots, n_A$) and n_B individuals in group B (labeled $1, 2, \ldots, n_B$), with the total number of individuals being $n := n_A + n_B$. Each group receives a different management treatment, which is applied in a uniform manner to each plant and to which plants respond independently. The treatment outcome for each plant is a Bernoulli random variable. For plants in group i ($i = A, B$), treatment is effective with probability p_i and ineffective with probability $1 - p_i$. Let random variable N_{i1} be the number of effectively treated plants in group i, let $N_{i2} := n_i - N_{i1}$ be the number of ineffectively treated plants in group i, and let $N_i = N_{i1} + N_{i2}$ (Table 3). Then N_{i1} is binomially distributed with parameters n_i and p_i, expected value $n_i p_i$, and variance $n_i p_i(1 - p_i)$. Moreover, for n_i sufficiently large, N_{i1} is approximately normally distributed with mean $n_i p_i$ and variance $n_i p_i(1 - p_i)$.

Table 3. Outcome table with two treatments and a binary treatment effect. Row sums are fixed, but column sums for the numbers of plants treated effectively and ineffectively are not.

Treatment	Plants treated effectively	Plants treated ineffectively	Row sum
A	N_{A1}	$n_A - N_{A1}$	n_A
B	N_{B1}	$n_B - N_{B1}$	n_B
Column sum:	N_1	$n - N_1$	n

Statistical method: The maximum likelihood estimator $\hat{p}_i$ for p_i ($i = \mathrm{A}, \mathrm{B}$) is

$$\hat{p}_i = N_{i1}/n_i. \tag{6}$$

We are interested in testing the null hypothesis H_0: $p_\mathrm{A} - p_\mathrm{B} = 0$. Under H_0, $p_\mathrm{A} = p_\mathrm{B} = p$. The maximum likelihood estimator for p is

$$\hat{p} = (N_{\mathrm{A}1} + N_{\mathrm{B}1})/(n_\mathrm{A} + n_\mathrm{B}) = N_1/n, \tag{7}$$

and random variable $\hat{p}_\mathrm{A} - \hat{p}_\mathrm{B}$ has expected value 0 and variance $p(1-p)(1/n_\mathrm{A} + 1/n_\mathrm{B})$. For sufficiently large n_A and n_B, $\hat{p}$ will be a good estimate of p. The variance of $\hat{p}_\mathrm{A} - \hat{p}_\mathrm{B}$ will then be approximately $\hat{p}(1-\hat{p})(1/n_\mathrm{A} + 1/n_\mathrm{B})$, and the random variable V defined by

$$V = \frac{\hat{p}_\mathrm{A} - \hat{p}_\mathrm{B}}{\sqrt{\hat{p}(1-\hat{p})(1/n_\mathrm{A} + 1/n_\mathrm{B})}} \tag{8}$$

will have an approximate standard normal distribution (Agresti 2013: p. 78, Hollander et al. 2014: pp. 497, 505). V can therefore be used to test the null hypothesis H_0 against a one- or two-sided alternative hypothesis H_1. Whenever n_A and n_B are large enough for V to be approximately normally distributed, V^2 will be distributed approximately as a chi-squared random variable with 1 degree of freedom. This fact provides another way to test H_0 against the two-sided H_1, but use of V is preferable when a one-sided H_1 is of interest.

For the one-sided alternative hypothesis H_1: $p_\mathrm{A} - p_\mathrm{B} > 0$, we compute the P-value for H_0 as the probability that $Z \geq v$, where Z is a standard normal random variable and v is the observed value of V. For the one-sided alternative hypothesis H_1: $p_\mathrm{A} - p_\mathrm{B} < 0$, the P-value is the probability that $Z \leq v$, and for the two-sided alternative hypothesis H_1: $p_\mathrm{A} - p_\mathrm{B} \neq 0$, the P-value is twice the probability that $Z \geq |v|$. In each case, we reject H_0 in favor of H_1 if and only if P-value $< \alpha$, where α is the significance level.

To determine whether n_A and n_B are large enough for V to be approximately normal (or for V^2 to be approximately chi-squared), estimates e_{ij} of the expected values of successes (plants for which treatment was effective) and failures are computed as

$$\begin{aligned} e_{\mathrm{A}1} &= n_\mathrm{A}\hat{p}(n_1),\ e_{\mathrm{A}2} = n_\mathrm{A}(1-\hat{p}(n_1)), \\ e_{\mathrm{B}1} &= n_\mathrm{B}\hat{p}(n_1),\ e_{\mathrm{B}2} = n_\mathrm{B}(1-\hat{p}(n_1)), \end{aligned} \tag{9}$$

where $\hat{p}(n_1)$ is the estimate of p obtained by replacing random variable N_1 in the maximum likelihood estimator $\hat{p}$ by its observed value n_1; that is,

$$\hat{p}(n_1) = n_1/n. \tag{10}$$

A simple rule of thumb for assessing adequacy of the normal approximation for V and the chi-squared approximation for V^2 is that

$$e_{ij} \geq 5 \ \text{ for all } \ i, j \tag{11}$$

(Hollander et al., 2014: p. 505). If any e_{ij} is less than 5, or simply to remove uncertainty regarding adequacy of the normal approximation, a small-sample test should be used instead (Agresti, 2013: pp. 93–94).

The small-sample test traditionally used for such data is Fisher's exact test, also known more informatively as the exact conditional binomial test. It assumes that both row and column sums are fixed, which is inappropriate for experiments of the type considered here because only the row sums are fixed (Lydersen et al., 2009; Agresti, 2013: p. 93). Simulation studies demonstrate that the exact conditional binomial test tends to produce P-values well above the correct unconditional P-value, and false-positive error rates well below the nominal rate α (Lydersen et al., 2009).

A more-accurate P-value and false-positive error rate can easily be obtained by employing the mid-P version of the exact conditional binomial test (Lydersen et al., 2009; Agresti, 2013) or by employing various versions and modifications of Barnard's exact unconditional binomial test. Exact unconditional binomial tests treat only the row sums as fixed, which is correct for experiments of the type considered here (in this traditional terminology, "unconditional" refers only to the column sums). Simulation studies show that these tests are also more powerful than the exact conditional test for 2×2 outcome tables (e.g., Attwood et al., 2022). Both the mid-P version of the exact conditional test and the exact unconditional tests perform better than Fisher's exact conditional test. Based on a simulation study, Lydersen et al. (2009) conclude: "Our general recommendation is not to condition on any marginals not fixed by design... The traditional Fisher's exact test should practically never be used" (the only exception being the rare study where both marginals are in fact fixed by design). We prefer Barnard's exact unconditional test over the mid-P version of the conditional test, mainly because its assumed experimental design is correct and the computational demands of Barnard's test that formerly limited its use are no longer a concern in the small-sample case.

R functions: R function `prop.test()` in the `stats` package tests H_0 using a large-sample test based on Z^2 for a two-sided H_1 and based on Z for a one-sided H_1. It checks the sample-size condition in Eq. (11) of the main text and reports a warning if it is violated. Function `prop.test()` also computes the Newcombe confidence interval for $p_A - p_B$. Various versions of the exact unconditional binomial test can be performed using the `exact.test()` function in R package `Exact` (Calhoun, 2022); the package author recommends using Barnard's original version of the test when computationally feasible, as specified by option `method="CAM"`. This function also has an option to compute a confidence interval for $p_A - p_B$.

Example: Preventing post-removal resprouting by Norway maple saplings (continued). Using the Norway maple data shown in Table 1, we can test the null hypothesis H_0: $p_{CT} - p_C = 0$ for the two removal methods (CT: cut down, then treat stump with triclopyr; C: cut down, with no follow-up triclopyr treatment) using the `prop.test()` function in R (an R snippet for this example is included in the supplemental information). We find that $\hat{p}_{CT}(n_1) - \hat{p}_C(n_1) = 0.767$ and $v = 8.518$. We also find that $e_{A1} = e_{B1} = 35$ and $e_{A2} = e_{B2} = 25$. All e_{ij} are much greater than 5, so the large-sample test appears appropriate. Since Norway maple is known to resprout readily from untreated stumps when cut down, and since stump application of triclopyr is expected to either increase the probability of preventing resprouts or have no effect, the most appropriate alternative hypothesis is H_1: $p_{CT} - p_C > 0$. Using R's `prop.test()` function, the reported P-value is approximately 8.1×10^{-18}, well below the usual significance level of $\alpha = 0.05$ (in publications, a P-value this small would normally be reported as, for example, $P < 0.001$). Thus, there is strong evidence that applying triclopyr to stumps after cutting down saplings increases the proportion of stumps that do not resprout, compared to cutting down saplings without applying triclopyr.

For purposes of illustration, we test the same null and alternative hypotheses with Barnard's exact unconditional binomial test, using function `exact.test()` in R package `Exact`. We find that $\hat{p}_{CT}(n_1) - \hat{p}_C(n_1) = 0.767$ and the reported P-value is approximately 7.0×10^{-20}. Thus, this test also provides strong evidence that applying triclopyr to stumps after cutting down saplings increases the proportion of stumps that do not resprout, compared to cutting down saplings without applying triclopyr.

Question 4: What is PET difference $p_A - p_B$ and its 95% confidence interval for plant groups A and B?

The goal in this case is to estimate the difference between the PETs for groups A and B, and to estimate the 95% confidence interval for the difference.

Assumptions and notation: Except where noted below, the assumptions and notation are the same as for Question 3.

Statistical method: The maximum likelihood estimator for p_i (i = A, B) is given by Eq. (6). An estimator for the difference $p_A - p_B$ between the PETs for treatment groups A and B is $\hat{p}_A - \hat{p}_B$. As in the case of a single proportion, the simple Wald confidence interval for $p_A - p_B$ has poor statistical properties and should not be used (Agresti & Caffo, 2000; Fagerland et al., 2013; Agresti, 2013). Simulation studies show that two types of simple confidence interval—the Newcombe hybrid score and Agresti-Caffo intervals—perform much better.

For the Newcombe hybrid score interval, let $(\lambda_i(N_{i1}), \upsilon_i(N_{i1}))$ be the $100(1-\alpha)\%$ Wilson confidence interval for p_i, as defined above. Then the $100(1-\alpha)\%$ Newcombe hybrid score interval for $p_A - p_B$ is given by

$$(\hat{p}_A - \hat{p}_B - z_{\alpha/2}S_L,\ \hat{p}_A - \hat{p}_B + z_{\alpha/2}S_U), \tag{12}$$

where

$$S_L = \sqrt{\lambda_A(1-\lambda_A)/n_A + \upsilon_B(1-\upsilon_B)/n_B},\ S_U = \sqrt{\upsilon_A(1-\upsilon_A)/n_A + \lambda_B(1-\lambda_B)/n_B}. \tag{13}$$

The $100(1-\alpha)\%$ Agresti-Caffo interval is given by

$$\check{p}_A - \check{p}_B \pm z_{\alpha/2}\sqrt{\check{p}_A(1-\check{p}_A)/\check{n}_A + \check{p}_B(1-\check{p}_B)/\check{n}_B}, \tag{14}$$

where

$$\check{n}_i = n_i + 2,\ \check{p}_i = (N_{i1}+1)/\check{n}_i. \tag{15}$$

As usual, the upper and lower limits of these confidence intervals are random variables, and statistical software reports a sample estimate in which random variables N_{i1} are replaced by observed values n_{i1}.

For small samples, the Newcombe hybrid score and Agresti-Caffo intervals should not be considered reliable. Instead, an exact unconditional confidence interval should be employed. Several of these intervals and software for estimating them are discussed by Fagerland et al. (2015). As is true of Barnard's test and its modifications, estimation of exact unconditional intervals is computationally intensive—even more so than determining a P-value for testing the null hypothesis, $p_A - p_B = 0$.

R functions: Newcombe hybrid score and Agresti-Caffo intervals confidence interval for $p_A - p_B$ can be computed using the function `pairwiseCI()` in R package `pairwiseCI`. The Newcombe confidence interval can also be computed using the `prop.test()` function in R's `stats` package. For small samples, the `exact.test()` function in the `Exact` package mentioned above in discussing Question 3 has an option to compute an estimated confidence interval for $p_A - p_B$. An alternative that seems to be computationally more efficient is the `BinomCI()` function in the `ExactCIdiff` package (Shan and Wang, 2022), which estimates an exact unconditional confidence interval due to Wang (2010). Neither of these R functions is included in the review by Fagerland et al. (2013). We illustrate use of `BinomCI()` in the following example.

Example: Preventing post-removal resprouting by Norway maple saplings (continued). Estimates of the PET difference $p_{CT} - p_C$ and its 95% confidence interval for the Norway maple data shown in Table 1 were computed using R's `pairwiseCI()` function (Newcombe and Agresti-Caffo intervals) and `BinomCI()` function (Wang interval). Saplings in group CT were cut down and the resulting stumps were immediately treated with triclopyr; saplings in group C were simply cut down. The results are shown in Table 4. Note that all three types of confidence interval provide strong evidence that stump application of triclopyr immediately after cutting a dappling down increases the probability that resprouting will not occur, as compared to simply cutting the sampling down. This conclusion is also supported by the method used in the previous example to answer Question 3. The key advantage of the methods used to answer Question 4 is that, in addition to providing an estimate of the PET difference and a conclusion as to whether it is significantly different from zero, they also provide upper and lower boundaries within which the true PET difference is likely to lie.

Table 4. Estimates and 95% Newcombe, Agresti-Caffo, and Wang confidence intervals for the difference $p_{CT} - p_C$ between PETs for two different management treatments (CT: Cut down, then treat stump with triclopyr; C: Cut down only) in the Norway maple management experiment.

Confidence Interval Type	$p_{CT} - p_C$	Lower limit	Upper limit
Newcombe	0.767	0.624	0.852
Agresti-Caffo	0.767	0.627	0.857
Wang	0.767	0.635	0.860

2 Management Experiments Using Point Intercept Surveys

2.1 Assessing the efficacy of a single management treatment on a single population

2.1.1 Methods based on binary data

Question 7: What is PDS change $P'_1 - P_1$ and its 95% confidence interval for pre- and post-treatment PDS parameters P'_1 and P_1?

The goal here is to estimate the change in the PDS before and after treatment, as well as the 95% confidence interval for the change. Unlike the previous two questions, the statistical analysis now utilizes information from both the pre-treatment and the post-treatment surveys.

Assumptions and notation: All survey points are assessed both before and after the management treatment is applied. During each assessment, each survey point is in one of two mutually exclusive states: 1 (desirable, from a management perspective) or 2 (undesirable). In practice, the desirable and undesirable states often are absence and presence of live plants of the target invasive species.

Let n denote the (fixed) number of survey points, and let random variables X_k and X'_k denote the states of survey point k before and after treatment. The possible values of the pair of random variables (X_k, X'_k) are $(1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$. Let P_{ij} denote the probability that $X_k = i$ and $X'_k = j$, with the possible values of i being 1 and 2 and similarly for j, and let P_i and P'_i denote the probabilities that $X_k = i$ and that $X'_k = i$. We assume that P_{ij} is the same for all survey points, and similarly for P_i and P'_i. Because the same survey points are used in both surveys, X_k and X'_k are not independent, meaning that P_{ij} is not equivalent to the product $P_i P'_j$. Let $p_{j|i}$ denote the probability that $X'_k = j$, given that $X_k = i$, where $p_{j|i}$ is assumed to be the same for all survey points k. Then

$$P_{ij} = P_i p_{j|i} \tag{16}$$

(here we are using the basic concepts of independence and conditional probability; for a lucid but rigorous elementary account, see Hoel et al., 1971: sections 1.4, 1.5). We also have $P_1 + P_2 = 1$, $P'_1 + P'_2 = 1$, and $p_{1|i} + p_{2|i} = 1$ for $i = 1, 2$.

Let random variable N_{ij} denote the number of survey points for which $X_k = i$ and $X'_k = j$ (Table 5). Then the total number n of survey points is given by $n = \Sigma_{i,j} N_{ij}$. Random variable N_{ij} has a multinomial distribution with parameters n and P_{ij} (analogy: draw n balls with replacement from a bag containing balls of colors C_{11}, C_{12}, C_{21}, and C_{22} in proportions P_{11}, P_{12}, P_{21}, and P_{22}; the resulting number of balls of the different colors will be N_{11}, N_{12}, N_{21}, and N_{22}, with $N_{11} + N_{12} + N_{21} + N_{22} = n$). Let the number of survey points for which $X_k = i$ be $N_i = N_{i1} + N_{i2}$, and let the number of survey points for which $X'_k = i$ be $N'_i = N_{1i} + N_{2i}$. One possible management goal is to increase the number (equivalently, the proportion) of survey points that are in desirable state 1. This event will occur if and only if $N'_1 > N_1$, which in turn will occur if and only if $N_{21} > N_{12}$ (the values of N_{11} and N_{22} are irrelevant, because survey points that

retain their state cannot cause a change in the number of survey points in the desirable state). Similarly, since $P_i = P_{i1} + P_{i2}$ and $P'_i = P_{1i} + P_{2i}$, it follows that $P'_1 - P_1 = P_{21} - P_{12}$, so that $P'_1 > P_1$ if and only if $P_{21} > P_{12}$. For this reason, statistical tests of hypotheses concerning whether this management goal has been achieved are usually based on random variables N_{12} and N_{21} alone.

Table 5. Outcome table for McNemar's and related tests for assessing efficacy of a management method using point intercept data, where the management state is either desirable (e.g., invasive plant is absent) or undesirable (e.g., invasive plant is present). X: Management state of a survey point before treatment, X': management state of a survey point after treatment, N_{ij}: number of survey points in state i before treatment and state j after treatment, N_i: number of survey points in state i before treatment, N'_j: number of survey points in state j after treatment.

	State After Treatment		
State Before Treatment	$X' = 1$ (Desirable)	$X' = 2$ (Undesirable)	Row Sum
$X = 1$ (Desirable)	N_{11}	N_{12}	N_1
$X = 2$ (Undesirable)	N_{21}	N_{22}	N_2
Column Sum	N'_1	N'_2	n

Statistical method: Here we are interested in obtaining an estimate of the difference $P'_1 - P_1$ and its $100(1 - \alpha)\%$ confidence interval. Recalling that $P'_1 - P_1 = P_{21} - P_{12}$ and that the N_{ij} are multinomially distributed with parameters n and P_{ij}, the maximum likelihood estimators for P_{12} and P_{21} are given by

$$\hat{P}_{12} = N_{12}/n, \quad \hat{P}_{21} = N_{21}/n. \tag{17}$$

Therefore, $\hat{P}_{21} - \hat{P}_{12} = N_{21}/n - N_{12}/n$ is an estimator for $P_{21} - P_{12}$ and hence for $P'_1 - P_1$. That is,

$$P'_1 - P_1 \approx \hat{P}_{21} - \hat{P}_{12} = (N_{21} - N_{12})/n. \tag{18}$$

As in the case of two independent binomial probabilities, which we considered above, the simple Wald confidence interval based on the asymptotic normality of $\hat{P}_{21} - \hat{P}_{12}$ has poor statistical properties and should not be used. Agresti & Min (2005) propose the following better but still simple confidence interval for $P'_1 - P_1$:

$$\frac{N_{21} - N_{12}}{n+2} \pm \frac{z_{\alpha/2}}{n+2} \sqrt{N_{21} + N_{12} + 1 - \frac{(N_{21} - N_{12})^2}{n+2}}. \tag{19}$$

Note that the midpoint of the estimated confidence interval is somewhat less than the estimator for $P'_1 - P_1$ stated in Eq. (18).

R functions: Function `diffpropci.mp()` in R package `PropCIs` (Scherer, 2018) computes the Agresti-Min adjusted Wald confidence interval given by Eq. (19) of the main text. Function `scoreci.mp` in the same package computes a score confidence interval for $P'_2 - P_2$ due to Tango (1998).

Example: Control of an invasive baby's breath population. Two of the authors (EKR and JNM) conducted an adaptive management study of the invasive plant, baby's breath (*Gypsophila paniculata*), in a coastal dune habitat in northwestern Michigan (USA) (Rice et al., 2020). As part of the study, the efficacy of a commonly-used management protocol (foliar application of the herbicide, glyphosate, using backpack sprayers or, when weather conditions prevent use of glyphosate, manual removal of plants) was assessed using point intercept surveys conducted in two years (2017 and 2018). The pre-treatment survey was conducted in May of 2017 and the post-treatment survey in May of 2018. The survey grid and sampling method used in this study are described in section 2.2 of the main text, where we note that a quantitative abundance

estimate was made in the vicinity of each survey point. These estimates can easily be converted to presence-absence (binary) data, which is the form in which we will treat them in this example. Specifically, we will consider presence-absence data for two sampling zones; one (Exp2) received the test treatment and the other (Ref) was a reference area. The observed values of N_{ij} and n for the two zones are displayed in Table 6.

Table 6. Observed values n_{ij} and n for baby's breath presence-absence data from zones Exp2 and Ref in 2017 and 2018. State 1 is absence of the plant, state 2 is presence. Also shown are values of n^*, which are used in the example for Question 8.

Zone	n_{11}	n_{12}	n_{21}	n_{22}	n	n^*
Exp2	9	0	11	46	66	11
Ref	8	3	1	48	60	4

Eqs. (18) and (19) of the main text and the values in Table 6 above were used to calculate estimates of $P_1' - P_1$ and their corresponding 95% Agresti-Min confidence intervals, which are shown in Table 7. Both estimates for $P_1' - P_1$ are roughly 0.16 for the Exp2 zone, and the corresponding Agresti-Min 95% confidence interval does not include zero. It follows that there is strong evidence that the proportion of the Exp2 zone that was free of baby's breath decreased between 2017 and 2018, following the treatment in 2017. By contrast, the 95% confidence interval for the reference area does include zero, so there is no strong evidence for a change in the proportion of the reference area that was free of baby's breath. Taken together, these two results are evidence that the test treatment was effective.

Table 7. Estimates and 95% Agresti-Min confidence intervals for the post-treatment change $P_1' - P_1$ in PDS for baby's breath presence-absence data from zones Exp2 and Ref. The two estimates of $P_1' - P_1$ are the difference of maximum likelihood estimates for P_1' (after treatment) and P (before treatment) and the midpoint of the confidence interval.

	$P_1' - P_1$		95% Confidence Interval	
Zone	Max. Likelihood	CI Midpoint	Lower Limit	Upper Limit
Exp2	0.167	0.162	0.070	0.254
Ref	−0.033	−0.032	−0.103	0.038

Question 8: Is $P_1' - P_1 > 0$, meaning that the PDS increased following treatment?

The goal here is simply to determine whether there is strong evidence that the PDS increased following treatment. The answer to this question provides no information regarding the magnitude of a detected increase. In particular, it is important to remember that the size of the P-value is not a measure of the magnitude of increase, but only a measure of the strength of evidence that an increase occurred. Question 7 deals with the issue of how large the increase (if any) was, which usually is more important.

Assumptions and notation: The assumptions and notation are the same as for Question 7, except that nearly all available statistical tests for answering Question 8 impose the additional condition that observed sum $n^* = n_{12} + n_{21}$ is fixed, where n_{ij} denotes the observed value of random variable N_{ij} (the only test we are aware of the does not impose this condition is the computationally-complex exact unconditional test attributed to Suissa and Shuster (1991) and assessed by Fagerland et al. (2013)). Let N_{12}^* and N_{21}^* denote

N_{12} and N_{21} conditioned on n^*. Then $N^*_{12} = n^* - N^*_{21}$. Also, N^*_{12} is binomially distributed with parameters n^* and P^*_{12}, where P^*_{12} is the probability that any particular survey point undergoes state transition $1 \rightarrow 2$, conditional on the transition being either $1 \rightarrow 2$ or $2 \rightarrow 1$, and is given by

$$P^*_{12} = \frac{P_{12}}{P_{12} + P_{21}} = \frac{1}{1 + P_{21}/P_{12}}. \tag{20}$$

N^*_{12} and N^*_{21} have expected values $n^* P^*_{12}$ and $n^* P^*_{21}$ and common variance $n^* P^*_{12}(1 - P^*_{12})$, where $P^*_{21} = 1 - P^*_{12}$.

Statistical method: For this question, we are interested in testing the null hypothesis that $P'_1 - P_1 = 0$ against the one-sided alternative hypothesis $P'_1 - P_1 > 0$. As shown above, the sign of $P'_1 - P_1$ is necessarily the same as the sign of $P^*_{21} - P^*_{12}$, so testing H_0: $P'_1 - P_1 = 0$ against H_1: $P'_1 - P_1 > 0$ is equivalent to testing H_0: $P^*_{21} - P^*_{12} = 0$ against H_1: $P^*_{21} - P^*_{12} > 0$. A score test closely related to McNemar's test is used for this purpose.

It is important to note that the chi-squared test recommended by Madsen (1999), Parsons (2001), and Hauxwell et al. (2010), and used by Mikulyuk et al. (2010) as the basis for a power analysis, is not appropriate for comparing pre- and post-treatment point intercept data if the same sampling grid is used for both surveys, because the pre and post observations for each survey point are paired and therefore cannot be assumed to be statistically independent. McNemar's test and the closely related score test are designed specifically for this case and have been used by, for example, Wersal et al. (2006, 2010), Madsen et al. (2006, 2008), and Rice et al. (2020).

Because $P^*_{21} + P^*_{12} = 1$, testing H_0: $P^*_{21} - P^*_{12} = 0$ against H_1: $P^*_{21} - P^*_{12} > 0$ is equivalent to testing H_0: $P^*_{21} - 1/2 = 0$ against H_1: $P^*_{21} - 1/2 > 0$. The score test (Agresti, 2013: p. 416) is based on the standardized random variable V given by

$$V = \frac{N^*_{21} - N^*_{12}}{\sqrt{N^*_{21} + N^*_{12}}}. \tag{21}$$

When n^* is sufficiently large ($n^* > 10$, according to Agresti, 2013: p. 416), the distribution of V under the null hypothesis approximates a standard normal distribution. (The test statistic for McNemar's classical asymptotic test is V^2, which is approximately chi-squared with one degree of freedom for large n^* and is appropriate for testing the null hypothesis against the two-sided alternative.) In this typical case, the P-value under null hypothesis $P^*_{21} - P^*_{12} = 0$ with one-sided alternative hypothesis $P^*_{21} - P^*_{12} > 0$ is the probability that $Z > v$, where Z is a standard normal random variable and v is the observed value of random variable V in Eq. (21). Numerical studies by Fagerland et al. (2013) show that this asymptotic test has the highest statistical power of the five commonly-used tests they assessed but tends to produce false-positive (type-I) error rates slightly greater than the nominal α. These authors also show that the continuity correction often used with McNemar's asymptotic test (the default option for R function `mcnemar.test()`) should not be used, because it usually results in false-positive error rates that are much less than the nominal α.

Two alternative tests that can easily be performed with standard statistical software and do not require large n^* are the exact conditional test and the mid-P test. Both use the fact that under H_0, N^*_{21} is binomially distributed with parameters n^* and $P^*_{21} = 1/2$. For the exact conditional test, the relevant P-value for testing the null hypothesis against our one-sided alternative hypothesis is the probability that the value of random variable N^*_{21} would be as large as, or larger than, observed value n_{21} if the null hypothesis were true. Thus, it is the probability that $N^*_{21} \geq n_{21}$ (equivalently, $N^*_{21} > n_{21} - 1$) with n^* fixed at the observed value and $P^*_{21} = 1/2$. Fagerland et al. (2013) show that this test is overly conservative, often producing false-positive error rates less than half the nominal α. The mid-P McNemar test usually produces a false-positive error

rate that is much closer to the nominal α without exceeding it, while achieving power nearly as high as that of the asymptotic test (Fagerland et al., 2013). This test is performed by subtracting half the binomial probability of event $\{N_{21} = n_{21}\}$ from the one-sided P-value produced by the exact conditional test; that is,

$$\text{mid-}P\ P\text{-value} = \Pr\left(N_{21} \geq n_{21} \mid n^*, 0.5\right) - \tfrac{1}{2}\Pr\left(N_{21} = n_{21} \mid n^*, 0.5\right), \tag{22}$$

where the probabilities are binomial with parameters n^* and $P^*_{21} = 0.5$.

Based on an extensive numerical study, Fagerland et al. (2013) recommend not using the asymptotic McNemar test with continuity correction or the exact conditional test. They summarize their conclusions regarding the other three tests they assessed as follows: "The easy-to-calculate mid-P test is an excellent alternative to the complex exact unconditional test. Both can be recommended for use in any situation. We also recommend the asymptotic test if small but frequent violations of the nominal level is acceptable."

R functions: R's `mcnemar.test()` performs the classical McNemar asymptotic test of the null hypothesis against the two-sided alternative. Rather than dealing with the problem of how to utilize this function to produce a P-value for the appropriate one-sided alternative hypothesis, we recommend performing the test more transparently by referring score statistic V to a standard normal distribution using R function `pnorm()`. The exact conditional test is easy to perform using R function `pbinom()` to calculate $\Pr(N_{12} > n_{12} - 1 \mid n^*, 0.5)$; Hollander et al. (2014: p. 507) give an example. The P-value for the mid-P test also can easily be computed using `pbinom()` and Eq. (22) of the main text by first computing the exact conditional P-value $\Pr(N_{12} > n_{12} - 1 \mid n^*, 0.5)$ and then adjusting it by subtracting $\Pr(N_{12} = n_{12} \mid n^*, 0.5)/2$.

Example: Control of an invasive baby's breath population (continued). Table 6 contains all the data required for applying the mid-P McNemar and exact conditional tests to the baby's breath presence-absence data. As we have already noted, Fagerland et al. (2013) show that the mid-P McNemar test is clearly superior to the exact conditional test, but we nevertheless include both in this example to illustrate the difference in the resulting P-values. Table 8 shows the P-values for tests of the null hypothesis H_0: $P'_1 - P_1 = 0$ versus the one-sided alternative hypothesis H_0: $P'_1 - P_1 > 0$ for both tests. Note that there is very strong evidence in favor of the alternative hypothesis in both cases, but that the P-values for the exact conditional test are larger than those for the mid-P test. This difference does not change the conclusion in the present example but could do so in cases where the P-values were closer to the chosen significance level.

Table 8. P-values for tests of the null hypothesis $P'_1 - P_1 = 0$ versus the alternative hypothesis $P'_1 - P_1 > 0$ for baby's breath presence-absence data from zones Exp2 and Ref. Results are shown for both the mid-P McNemar test and the exact conditional test.

	P-value	
Zone	Mid-P McNemar	Exact Conditional
Exp2	0.0002	0.0005
Ref	0.8125	0.9375

Additional remarks. From a management perspective, it is important to know whether $P'_1 > P_1$, because it provides evidence as to whether the relative frequency of the invasive plant decreased. But what does knowing that this frequency decreased really tell us about the effectiveness of the management treatment employed? This question may seem strange, so we provide two simple examples to show why it is important to consider.

Suppose the desirable management state is absence of the focal invasive plant and the undesirable state

is presence. From Eqs. (16) and (20), we have

$$P_{21}^{*} = \frac{1}{1 + (P_1/P_2) \cdot (p_{2|1}/p_{1|2})}. \tag{23}$$

Now suppose $p_{2|1} = p_{1|2}$, meaning that it is just as likely that the invasive plant will invade a survey point where it initially was not present as it is that the plant will be extirpated from a survey point where it initially was present. Then the sign of $P_{21}^{*} - 1/2$ is determined entirely by the sign of $P_2 - P_1$; i.e., by whether the initial proportion of survey points occupied by the invasive plant is greater or less than the initial proportion not occupied by it. But this has nothing to do with the effectiveness of the management treatment applied in the present field experiment; it simply reflects the fact that if there initially are more (or fewer) survey points where the invasive plant is present than where it is absent, there will be more (or fewer) chances for extirpation to occur than for colonization to occur.

Alternatively, suppose $P_2 = P_1$, so that each survey point is equally likely to be in the desirable or undesirable state before the management treatment is applied. Then the sign of $P_{21}^{*} - 1/2$ is determined entirely by the sign of $p_{1|2} - p_{2|1}$; i.e., by whether whether it is more (or less) likely that the invasive plant will be extirpated from a survey point where it is initially present or will invade a survey point where it is initially absent. This condition is directly related to the apparent effectiveness of the management treatment in locally extirpating the invasive plant relative to the apparent effectiveness of the plant in invading new locations. (We say "apparent" because, depending on properties of the plant and the method used to relocate survey points for the post-treatment survey, some of the observed differences in states of survey points before and after treatment may reflect measurement error.)

These examples show that, while it is important to know whether the probability of a survey point being occupied by the invasive plant decreased following treatment, the change in this probability typically confounds the influences of two distinct factors, one of which is related to the effectiveness of management and the other not. It therefore seems worthwhile to supplement the assessment by focusing on the first factor, which is the relative sizes of the probabilities of apparent local extirpation ($p_{1|2}$) and colonization ($p_{2|1}$). We address this issue in the following three research questions. In all three questions, we assume the desirable management state is absence of live plants of the invasive species and the undesirable state is presence.

Question 8a: What are the probabilities of apparent local extirpation ($p_{1|2}$) and apparent local colonization ($p_{2|1}$) and their 95% confidence intervals?

We can answer this question using the same statistical methods we used for Question 1. Let n_1 and n_2 be the known number of survey points in state 1 (invasive plant is absent) and state 2 (invasive plant is present) before treatment, where

$$n_1 = n_{11} + n_{12}, \quad n_2 = n_{21} + n_{22}. \tag{24}$$

We assume that after the management treatment is applied, each survey point that was in state i before treatment is in state j with probability $p_{j|i}$, where all state transitions occur independently (note: transitions $1 \to 1$ and $2 \to 2$ are allowed). Of the n_i survey points in state i before treatment, the number in state j after treatment is a binomial random variable with parameters n_i and $p_{j|i}$. The estimators for $p_{j|i}$ and its $100(1 - \alpha)\%$ confidence interval are therefore the same as for Question 1 except that n_i replaces n, $p_{j|i}$ replaces p, and N_{ij} replaces N_1. The maximum likelihood estimator of $p_{j|i}$ is given by

$$\hat{p}_{j|i} = N_{ij}/n_i, \tag{25}$$

where n_i is given by Eq. (24).

Examples. An example using baby's breath data is presented in section 1.2.1 of the Supporting Information; an R program that analyzes a set of appropriate simulated data (available online in a separate spreadsheet file) is presented in section 2.2.1.

Question 8b: Is the probability $p_{1|2}$ of apparent local extirpation greater than the probability $p_{2|1}$ of apparent local colonization?

We can answer this question using the same statistical methods as for Question 3, with changes in notation similar to those made in answering Question 8a. Two groups are again being compared, but the groups are now the sets of survey points that were in the two different management states (1 and 2) before treatment instead of sets of plants that received two different treatments. Treatment outcomes are again binary, but "success" now corresponds to a change in management state of a survey point following treatment ($1 \rightarrow 2$ or $2 \rightarrow 1$) and "failure" now corresponds to the management state remaining unchanged. Of the n_i (known) survey points in state (or group) i before treatment, the number in state $j \neq i$ after treatment is a binomial random variable with parameters n_i and $p_{j|i}$, as in Question 8a. Now, however, we wish to determine whether there is strong evidence that the "success" probabilities $p_{1|2}$ and $p_{2|1}$ for the two groups differ. The outcome table and notation for the statistical method are presented in Table 9. We can apply the statistical methods used in Question 3 with groups 1 and 2 replacing treatments A and B, n_1 and n_2 replacing n_A and n_B, and $p_{2|1}$ and $p_{1|2}$ replacing p_A and p_B. The null hypothesis will always be H_0: $p_{1|2} - p_{2|1} = 0$, while the relevant alternative hypothesis typically will be the one-sided hypothesis H_1: $p_{1|2} - p_{2|1} > 0$, since we are interested in knowing specifically whether there is strong evidence that the test treatment increases the relative frequency of transitions from management state 2 to state 1.

Table 9. Outcome table for Question 8b. The possible changes in management state are $2 \rightarrow 1$ (for survey points in state 2 before treatment) and $1 \rightarrow 2$ (for survey points in state 1 before treatment). X: state of survey point before treatment, N_{ij}: number of survey points in state i before treatment and state $j \neq i$ after treatment, n_k: fixed number of survey points in group k (k = A, B) before treatment, N_Δ: total number of survey points that changed state, n: total number of survey points.

Group	State Change	Occurred	Did Not Occur	Row Sum
A ($X = 2$)	$2 \rightarrow 1$	N_{21}	$n_A - N_{21}$	n_A
B ($X = 1$)	$1 \rightarrow 2$	N_{12}	$n_B - N_{12}$	n_B
	Column Sum:	N_Δ	$n - N_\Delta$	n

Examples. An example using baby's breath data is presented in section 1.2.1 of the Supporting Information; an R program that analyzes a set of appropriate simulated data (available online in a separate spreadsheet file) is presented in section 2.2.1.

Question 8c: What is the difference $p_{1|2} - p_{2|1}$ and its 95% confidence interval for the probabilities of apparent local extirpation and apparent local colonization?

We can answer this question using the same statistical methods we used for Question 4, with the same changes in notation as for Question 8b. The central management issue is whether the ability of a test treatment to extirpate an invasive species from locations where it was present before treatment is greater than the ability of the invader to expand into locations where it was not present before treatment.

Examples. An example using baby's breath data is presented in section 1.2.1 of the Supporting Information; an R program that analyzes a set of appropriate simulated data (available online in a separate spreadsheet file) is presented in section 2.2.1.

2.1.2 Methods based on quantitative data

Question 9: What are the mean local densities μ_i and μ_i' in group i and their 95% confidence intervals before and after treatment?

The goal here is simply to characterize the mean density of an invasive plant species in the restoration or reference area before and after treatment, focusing on one combination of plant species, treatment type, and treatment level at a time.

R functions: The function `meanCI()` in R package `MKinfer` (Kohl, 2023) is a flexible and convenient function for constructing bootstrap confidence intervals for the mean density (or abundance) of an invasive plant. Its `bootci.type` argument provides options for computing several different types of bootstrap confidence intervals, which are explained by Efron and Tibshirani (1998: chapter 14). To use this function, we suggest setting argument `boot=TRUE` and accepting the default choice for the others. This will produce an estimate of the mean density and five different types of bootstrap confidence intervals (viewing all of these is interesting and may stimulate you to read chapter 14 of Efron and Tibshirani (1998)), of which we suggest using the BC_a (Bias-Corrected and *a*ccelerated) interval.

Example: Control of an invasive baby's breath population (continued). Continuing with the baby's breath example, we now employ quantitative density data instead of the presence-absence data used in section 4.1.1 of the main text. Recall that zone Exp2 received the test treatment in 2017 (only), *after* the 2017 point intercept survey had been conducted, and that Ref was the reference area. Bootstrap estimates of mean density (plants m^{-2}) and corresponding BC_a 95% confidence interval for zones Exp2 and Ref in 2017 (pre-treatment) and 2018 (post-treatment) are shown in Table 10. Note that these results suggest a substantial decrease in mean density in Exp2 between 2017 and 2018 but do not suggest a decrease in the reference area.

Table 10. Bootstrap estimates of mean baby's breath density and corresponding BC_a 95% confidence interval (CI) in zones Exp2 and Ref for years 2017 and 2018. Density units: plants m^{-2}.

			BC_a CI	
Zone	Year	Mean	Lower	Upper
Exp2	2017	2.187	1.705	2.862
	2018	0.397	0.258	0.699
Ref	2017	1.789	1.362	2.378
	2018	1.570	1.150	2.239

Question 10: Is the mean local density μ_i' in group i after treatment less than prescribed management threshold $\mu^\star$?

The goal in this case is to test the null hypothesis that $\mu_i' - \mu^\star = 0$ against the one-sided alternative hypothesis $\mu_i' - \mu^\star < 0$, where $\mu^\star$ is an appropriately low prescribed density of the target invasive plant. Rejection of

the null hypothesis at a confidence level of, say, 0.95 provides strong evidence that the management goal was achieved.

R functions: Useful R packages for bootstrap methods include `boot` (Canty & Ripley, 2021; originally created for the book by Davison and Hinkley (1997) and now part of the standard distribution of R), `bootstrap` (Tibshirani & Leisch, 2019; original S version created for the book by Efron and Tibshirani (1998)), and the more-recent `MKinfer` package, which also includes functions for permutation tests. We find the bootstrap and permutation functions in the `MKinfer` package particularly well designed and easy to use. The `boot` and `bootstrap` packages provide greater flexibility for bootstrap methods but, in our opinion, require more expertise to use properly. For this reason, we will restrict attention to the bootstrap and permutation test functions in the `MKinfer` package in this review.

Example: Control of an invasive baby's breath population (continued). In the baby's breath example, let us consider two possible threshold densities, both completely arbitrary: $\mu^\star = 1.0$ and $\mu^\star = 0.5$ plants m^{-2}. The test treatment was applied to zone Exp2 in 2017; zone Ref is the reference area. We consider the densities measured in both zones in 2018. Table 11 shows the bootstrap t test results. The reported P-values are not adjusted to account for multiple comparisons in this case, because in practice, only the hypothesis test for zone Exp2 and for a single threshold density would be of interest for judging management success. Note that the results provide strong evidence that the 2018 baby's breath density was below the threshold of 1.0 plants m^{-2} in zone Exp2 but does not provide strong evidence that it was below the alternative threshold of 0.5 plants m^{-2}. There is no evidence that either threshold was achieved in the reference area.

Table 11. Tests of null hypothesis H_0: $\mu_i' - \mu^\star = 0$ versus alternative hypothesis $\mu_i' - \mu^\star < 0$ for the 2018 baby's breath point intercept data from zones Exp2 (which received the test treatment) and Ref (the reference area). μ_i' is the mean density (plants m^{-2}) a year after the test treatment was applied, while $\mu^\star$ is the prescribed threshold density. For purposes of illustration, two different threshold densities are considered.

Zone	Mean Density	Target Density	P-value
Exp2	2.187	1.0	0.004
		0.5	0.241
Ref	1.789	1.0	0.995
		0.5	1.000

Question 11: Is $\mu_i' - \mu_i < 0$, meaning that the mean local density in plant group i decreased following treatment?

The goal in this case is simply to determine whether there is strong evidence that the mean local density decreased following treatment. The null hypothesis is H_0: $\mu_i' - \mu_i = 0$ and the appropriate alternative hypothesis is H_0: $\mu_i' - \mu_i < 0$. In testing the null hypothesis, it is necessary to account for the fact that the point intercept data consist of matched pairs, since they were collected at the same survey points.

R functions: Functions `perm.t.test()` and `boot.t.test()` from R package `MKinfer` can again be used. There are now two sets of local density data: pre-treatment and post-treatment, each as a data vector. Both functions have a `paired` argument, which should be assigned the value `TRUE`. Both functions also have an `alternative` argument, which should be assigned the value `"less"` if the first of the two data arguments is the one whose density is asserted to be less in the alternative hypothesis. For Question 11, we are interested in the reported P-values.

Example: Control of an invasive baby's breath population (continued). Returning to the baby's breath example yet again, the pre-treatment data are those for 2017 and the post-treatment data are those for 2018. For Question 11, the hypothesis tests for both zones Exp2 and Ref are relevant, because we want to know whether we can attribute any density decrease detected in Exp2 to the test treatment instead of some region-wide decrease in density due to some other factor. We therefore used R's `p.adjust()` function with Holm's method (the default) to adjust the P-values separately for each type of t test to account for the multiple comparisons (only the smaller of the two P-values will change with this method). The results are shown in Table 12. Note that there is strong evidence that the local density of baby's breath decreased in both zones, though the evidence clearly is stronger for zone Exp2. Of course, strength of evidence that a change occurred does not provide evidence regarding the magnitude of change. The estimates of difference $\mu_i' - \mu_i$ for groups Exp2 and Ref in Table 11 suggest a substantially larger decrease occurred in Exp2 than in the reference area. In section 4.2 of the main text, we consider statistical methods that allow one to determine whether there is strong evidence that local density in fact decreased more in Exp2 than in the reference area; if so, there is strong evidence that the test treatment was effective, even though local density decreased in the reference area.

Table 12. Tests of null hypothesis H_0: $\mu_i' - \mu_i = 0$ versus alternative hypothesis $\mu_i' - \mu_i < 0$ for the 2017 and 2018 baby's breath point intercept data from zones Exp2 (which received the test treatment) and Ref (the reference area). μ_i' is the mean local density (plants m^{-2}) a year after the test treatment was applied, while μ_i is the mean local density shortly before the test treatment was applied. P-values were adjusted by Holm's method. Estimate: maximum likelihood estimate of difference $\mu_i' - \mu_i$.

			P-value	
Zone	Estimate	t Test Type	Unadjusted	Adjusted
Exp2	−1.790	Bootstrap	$< 2.2 \times 10^{-16}$	$< 4.4 \times 10^{-16}$
		Permutation	$< 2.2 \times 10^{-16}$	$< 4.4 \times 10^{-16}$
Ref	−0.219	Bootstrap	0.036	0.036
		Permutation	0.028	0.028

Question 12: What is the change $\mu_i' - \mu_i$ and its 95% confidence interval for mean local density in plant group i following treatment?

The goal in this case is to estimate the change in mean density before and after treatment and to estimate the 95% confidence interval for the difference. The change will be negative if mean density decreased.

R functions: Functions `perm.t.test()` and `boot.t.test()` from R package `MKinfer` can be used yet again. These functions produce P-values as well as confidence intervals, so a single R program can conveniently be written to generate both types of result. We are now interested in the confidence intervals. We note that both functions can produce two different types of confidence interval: one-sided and two-sided. These are obtained by assigning values `"less"` and `"two.sided"` to function argument `alternative`. The one-sided interval is useful for estimating only a reasonable upper bound on the true change in density and provides no information about how far below zero the true change is likely to be. By contrast, the traditional two-sided interval is useful for estimating reasonable upper as well as lower bounds on the true change in density and thus provides a reasonable indication of how much uncertainty surrounds the maximum likelihood estimate.

Example: Control of an invasive baby's breath population (continued). We may utilize the baby's

breath data again to illustrate the two types of confidence intervals. We wrote a single R program that used functions `perm.t.test()` and `boot.t.test()` to generate the *P*-values displayed above in Table 12 and the confidence intervals displayed below in Table 13. As noted in discussing Question 11, the two sets of density data (for 2017 and 2018) consist of matched pairs, so function argument `paired` was set to `TRUE`. Note that the one-sided confidence intervals have infinite width, but their upper limits are lower than those of the two-sided intervals. All upper limits of the one-sided intervals are negative, providing strong evidence that mean density decreased in both the Exp2 and Ref zones, consistent with results of the hypothesis tests in Table 12.

Table 13. One-sided and two-sided 95% confidence intervals for density difference $\mu_i' - \mu_i$ for the 2017 and 2018 baby's breath point intercept data from zones Exp2 (which received the test treatment) and Ref (the reference area). μ_i' is the mean local density (plants m^{-2}) a year after the test treatment was applied, while μ_i is the mean local density shortly before the test treatment was applied. Estimate: maximum likelihood estimate of difference $\mu_i' - \mu_i$.

			One-sided 95% CI		Two-sided 95% CI	
Zone	Estimate	Method	Lower	Upper	Lower	Upper
Exp2	−1.790	Bootstrap	$-\infty$	−1.378	−2.292	−1.335
		Permutation	$-\infty$	−1.438	−2.279	−1.290
Ref	−0.219	Bootstrap	$-\infty$	−0.034	−0.435	0.005
		Permutation	$-\infty$	−0.031	−0.438	−0.003

2.2 Comparing treatment efficacy in pairs of treatments or species

Question 13: Is the mean decrease $\mu_A - \mu_A'$ in local density in plant group A before and after treatment greater (or different, or less) than the mean decrease $\mu_B - \mu_B'$ in local density in plant group B?

R functions: Functions `perm.t.test()` and `boot.t.test()` from R package `MKinfer` can be used to answer Question 13. The two data vectors are the density changes for group A matched pairs and group B matched pairs. The function argument `paired` should be set to `FALSE` (the default) and the argument `alternative` should be set to either `"two.sided"`, `"less"`, or `"greater"`, as appropriate.

Example: Control of an invasive baby's breath population (continued). Recall that results for the examples in Questions 11 and 12 provide strong evidence that local density of baby's breath decreased in restoration zone Exp2 after receiving the test treatment, but that there was also strong evidence that local density decreased in reference area Ref, as well. Also recall that the estimated decreases in density in the two zones, as well as the corresponding confidence intervals, suggest that the magnitude of decrease was greater in Exp2 than in Ref. Let us now test this suggestion rigorously.

Each entry *i* in the 2017 vector of local densities for zone Exp2 corresponds to the same survey point as entry *i* in the 2018 vector, so data in the two vectors are matched pairs. The same is true for the 2017 and 2019 data vectors for zone Ref. Subtracting the 2017 density vector from the 2018 density vector for zone Exp2, we obtain a single vector of decreases for zone Exp2. Doing the same for the reference area, we obtain a single vector of decreases for zone Ref. Note that positive values of the calculated differences imply decreases in density between 2017 and 2018. Next, we assign the decrease vectors for Exp2 and Ref (in that order) to the first two (`x` and `y`) arguments of the `boot.t.test()` and `perm.t.test()` functions. We retain the default values for all other arguments except that argument `alternative` is set to `"greater"` so the null hypothesis of equal decrease in Exp2 and Ref will be tested against the alternative hypothesis

that the magnitude of decrease in density was greater in zone Exp2 than in zone Ref.

The results show that the difference between the estimated decrease in Exp2 and the estimated decrease in Ref is 1.572 plants m^{-2}. The *P*-value for the test of the hull hypothesis that the two decreases were equal against the alternative hypothesis that the decrease was greater in Exp2 than in Ref is reported by both `boot.t.test()` and `perm.t.test()` as $< 2.2 \times 10^{-16}$, so we may confidently reject the null hypothesis. The results therefore provide strong evidence that the mean local density of baby's breath decreased more in zone Exp2 than in zone Ref. This finding constitutes strong evidence that the test treatment was effective, even though a statistically significant decrease in mean local density was detected in the reference area.

3 References

Agresti, A. 2013. Categorical Data Analysis, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.

Agresti, A. and Caffo, B. 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. The American Statistician 54: 280–288.

Agresti, A. and Coull, B.A. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician 52: 119–126.

Agresti, A. and Min, Y. 2005. Simple improved confidence intervals for comparing matched proportions. Statistics in Medicine 24: 729–740.

Attwood, K., Park, S., and Hutson, A.D. 2022. Practical and robust test for comparing binomial proportions in the randomized phase II setting. Pharmaceutical Statistics 21: 361–371.

Bivand, R.S., Pebesma, E., and Goméz-Rubio, V. 2013. Applied Spatial Data Analysis with R, 2nd ed. New York: Springer.

Bjørnstad, O.N. 2022. ncf: Spatial Covariance Functions. R package version 1.3-2. https://CRAN.R-project.org/package=ncf.

Brown, L.D., Cai, T.T., and DasGupta, A. 2001. Interval estimation for a binomial proportion. Statistical Science 16: 101–133.

Calhoun, P. 2022. Exact: Unconditional Exact Test. R package version 1.3-2. https://CRAN.R-project.org/package=ncf,

Canty, A. and Ripley, B.D. 2021. boot. R package version 4.2.0. https://CRAN.R-project.org/package=boot.

Davison, A.C. and Hinkley, D.V. 1997. Bootstrap Methods and Their Application. Cambridge: Cambridge University Press.

Dorai-Raj, S. 2022. binom: Binomial Confidence Intervals for Several Parameterizations. R package version 1.1-1.1. https://CRAN.R-project.org/package=binom.

Efron, B. and Tibshirani, R.J. 1993. An Introduction to the Bootstrap. Hoboken, NJ: John Wiley & Sons, Inc.

Fagerland, M.W., Lydersen, S., and Laake, P. 2013. The McNemar test for binary matched-pairs data: mid-*p* and asymptotic are better than exact conditional. BMC Medical Research Methodology 13: 1–8.

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.

Hauxwell, J., Knight, S., Wagner, K., Mikulyuk, A., Nault, M., Porzky, M., and Chase, S. 2010. Recommended Baseline Monitoring of Aquatic Plants in Wisconsin: Sampling Design, Field and Laboratory Procedures, Data Entry and Analysis, and Applications. Wisconsin Department of Natural Resources, Madison, WI. PUB SS-1068.

Hollander, M., Wolfe, D.A., and Chicken, E. 2014. Nonparametric Statistical Methods, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.

Kohl, M. 2023. MKinfer: Inferential Statistics. R package version 1.1. https://www.stamats.de.

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

Madsen, J.D. 1999. Point intercept and line intercept methods for aquatic plant management. Aquatic Plant Control Technical Note MI-02. U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

Madsen, J.D., Wersal, R.M., Tyler, M., and Gerard, P.D. 2006. The distribution and abundance of aquatic macrophytes in Swan Lake and Middle Lake, Minnesota. Journal of Freshwater Ecology 21: 421–429.

Madsen, J.D., Stewart, R.M., Getsinger, K.D., Johnson, R.L., and Wersal, R.M. 2008. Aquatic plant communities in Waneta Lake and Lamoka Lake, New York. Northeastern Naturalist 15: 97–110.

Mikulyuk, A., Hauxwell, J., Rasmussen, P., Knight, S., Wagner, K. I., Nault, M. E., and Ridgely, D. 2010. Testing a methodology for assessing plant communities in temperate inland lakes. Lake and Reservoir Management 26: 54–62.

Parsons, J.K. 2001. Aquatic plant sampling protocols. Washington State Department of Ecology, Environmental Assessment Program. https://apps.ecology.wa.gov/publications/documents/0103017.pdf

R Core Team. 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rice, E.K., Leimbach-Maus, H., Partridge, C., and McNair, J. N. 2020. Assessment of invasive *Gypsophila paniculata* control methods in the northwest Michigan dunes. Invasive Plant Science and Management 13: 94–101.

Shan, G. and Wang, W. 2022. ExactCIdiff: Inductive Confidence Intervals for the Difference Between Two Proportions. R package version 2.1. https://CRAN.R-project.org/package=ExactCIdiff.

Tango, T. 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Statistics in Medicine 17: 891–908.

Tibshirani, R. and Leisch, F. 2019. bootstrap. Functions for the Book "An Introduction to the Bootstrap". R package version 2019.6. https://gitlab.com/scottkosty/bootstrap.

Wang, W. 2010. On construction of the smallest one-sided confidence interval for the difference of two proportions. The Annals of Statistics 38: 1227–1243.

Wersal, R.M., Madsen, J.D., McMillan, B.R., and Gerard, P.D. 2006. Environmental factors affecting biomass and distribution of *Stuckenia pectinata* in the Heron Lake System, Minnesota, USA. Wetlands 26: 313–321.

Wersal, R.M., Madsen, J.D., Woolf, T.E., and Eckberg, N. 2010. Assessment of herbicide efficacy on Eurasian watermilfoil and impacts to the native submersed plant community in Hayden Lake, Idaho, USA. Journal of Aquatic Plant Management 48: 5–11.