The Stata Journal (2023)
23, Number 1, pp. 24–52 DOI: 10.1177/1536867X231161971

artbin: Extended sample size for randomized trials with binary outcomes

Ella Marley-Zagar
MRC Clinical Trials Unit
University College London
London, U.K.
e.marley-zagar@ucl.ac.uk

Ian R. White
MRC Clinical Trials Unit
University College London
London, U.K.
ian.white@ucl.ac.uk

Patrick Royston
MRC Clinical Trials Unit
University College London
London, U.K.
j.royston@ucl.ac.uk

Friederike M.-S. Barthel
PRA / ICON PLC Germany
Mannheim, Germany
sophie@fm-sbarthel.de

Mahesh K. B. Parmar
MRC Clinical Trials Unit
University College London
London, U.K.
mp@ctu.mrc.ac.uk

Abdel G. Babiker
MRC Clinical Trials Unit
University College London
London, U.K.
a.babiker@ucl.ac.uk

Abstract. We describe the command `artbin`, which offers various new facilities for the calculation of sample size for binary outcome variables that are not otherwise available in Stata. While `artbin` has been available since 2004, it has not been previously described in the *Stata Journal*. `artbin` has been recently updated to include new options for different statistical tests, methods and study designs, improved syntax, and better handling of noninferiority trials. In this article, we describe the updated version of `artbin` and detail the various formulas used within `artbin` in different settings.

Keywords: st0013_3, artbin, sample size, power, binary outcome, randomized clinical trial, superiority trial, noninferiority trial

1 Introduction

Sample-size calculation is essential in the design of a randomized clinical trial to ensure that there is adequate power to evaluate treatment. It is also used in the design of randomized experiments in other fields such as education, international development (Attanasio, Kugler, and Meghi 2011), and criminology (Braga et al. 1999). It can also be used in the design of nonrandomized comparative studies (Quigley et al. 2019).

In Stata, several standard sample-size calculations are available in the inbuilt `power` family. More-advanced sample-size calculations are provided in the Analysis of Resources for Trials (ART) package (Barthel, Royston, and Babiker 2005; Barthel et al.

2006; Royston and Barthel 2010). ART is primarily aimed at trials with a time-to-event outcome, but it also includes the command `artbin` for trials with a binary outcome. `artbin` differs from the official `power` command by allowing many statistical tests, such as score, Wald, conditional, and trend across K groups, and by offering calculations under local or distant alternatives with or without continuity correction.

The calculations in `artbin` are based on a set of anticipated probabilities of the binary outcome, one in each treatment group. If the unknown probabilities of the binary outcome equal the anticipated probabilities, then `artbin` tells us the power achieved for a specified sample size or the sample size required to achieve the specified power.

The basic idea of sample-size calculation with a binary outcome is well known. We define the power $1-\beta$ to be the probability of rejecting the null hypothesis at the two-sided α level of significance.

In a two-group superiority trial, the null hypothesis is that the outcome probabilities in the two groups are equal and the alternative hypothesis is that they take the unequal anticipated probabilities π_1^a and π_2^a. If the trial has equal sample sizes n in each group, then a popular formula for the total sample size required is

$$2n = 2\frac{\left\{z_{1-\alpha/2}\sqrt{2\overline{\pi}^a(1-\overline{\pi}^a)} + z_{1-\beta}\sqrt{\pi_1^a(1-\pi_1^a)+\pi_2^a(1-\pi_2^a)}\right\}^2}{(\pi_2^a-\pi_1^a)^2}$$

where $z_c = \Phi^{-1}(c)$ is the standard normal deviate and $\overline{\pi}^a = (\pi_1^a + \pi_2^a)/2$ (Julious and Campbell 2012). Extensions are well known for unequal sample sizes.

However, several complications arise that are tackled by `artbin`. Some trials have more than two groups, and in these cases we may test for trend across the groups or for heterogeneity between the groups. There are variants of the sample-size formulas for different versions of the test applied to the data (for example, Pearson's χ^2 or Wald), and there are "local" variants that are valid only when the treatment effect is small. A loss to follow-up option is useful for the replication of sample-size calculations, as advocated by Clark, Berger, and Mansmann (2013).

Further, some two-group trials are noninferiority trials, in which the null hypothesis is that the experimental treatment is no worse than the control treatment by a prespecified amount m, termed the margin. They are used when the experimental treatment is not expected to be superior, but they do have other benefits, such as being cheaper, less toxic, or easier to administer, for example. Substantial-superiority trials are now increasingly used, especially in vaccine trials, where the null hypothesis states that the experimental treatment is better than the control treatment by at least m (see Krause et al. [2020]).

The latest upgrade of `artbin` substantially improves the original version released in 2004. The option to specify a margin for noninferiority or substantial-superiority trials has been included to enable sample-size and power calculations for more-complex two-group trials. New options for statistical tests and methods are now available, such as the Wald test, which is commonly used for sample-size calculation in noninferiority trials in medicine. The syntax and output have been improved, with more options available and clearer output. `artbin` does not require the anticipated event probabilities to be the same in the two groups for noninferiority or substantial-superiority trials, unlike any other software packages currently available in Stata. Previous users of `artbin` will need to alter existing `artbin` code to accommodate the changes. Please see the description of what has changed (appendix 1) for further details.

This article has three aims. First, it clearly lays out the scope of the `artbin` package and its dialog boxes and exemplifies its use. Second, it describes the updates made. Third, it clarifies the formulas used.

The article comprises a description of the new syntax (section 3.1), illustrative examples (section 3), a description of the updated menus and dialogs (section 4), details of the methods used (section 5), a description of how the software has been tested (section 6), and conclusions (section 7).

2 The artbin command

2.1 Syntax

```
artbin, pr(numlist) [margin(#)
   [unfavourable|unfavorable|favourable|favorable] [power(#)|n(#)]
   aratios(aratio_list) ltfu(#) alpha(#) onesided trend doses(dose_list)
   condit wald ccorrect local noround force]
```

`artbin` calculates the power or total sample size for various tests comparing K anticipated probabilities. Power is calculated if `n()` is specified; otherwise, total sample size is estimated. `artbin` can be used in designing superiority, noninferiority, and substantial-superiority trials.

`artbin` makes comparisons on the scale of difference in probabilities. The results on other scales, such as odds ratios, will be very similar for superiority trials but potentially very different for noninferiority and substantial-superiority trials (Quartagno et al. 2020).

In a multigroup trial, `artbin` is based on a test of the global null hypothesis that the probabilities are equal in all groups. The alternative hypothesis is that there is a difference between two or more of the groups.

In a two-group superiority trial, `artbin` is based on a test of the null hypothesis that the probabilities in the two groups are equal. The alternative hypothesis is that they take unequal values, such that the experimental treatment is better than the control treatment.

In a noninferiority trial, `artbin` is based on a test of the null hypothesis that the experimental treatment is worse than the control treatment by at least a prespecified amount, termed the margin. `artbin` supports the design of more-complex noninferiority trials in which π_1^a and π_2^a are unequal. Substantial-superiority trials are increasingly used; here the null hypothesis is that the experimental treatment is better than the control treatment by the margin at most.

To minimize the risk of error in two-group trials, the user is advised to identify whether the trial outcome is `favorable` or `unfavorable`. By default, `artbin` infers favorability status from the `pr()` and `margin()` options. If $\pi_2^a > \pi_1^a + \texttt{margin()}$, the outcome is assumed to be favorable; otherwise, it is assumed to be unfavorable.

2.2 Options

`pr(`*#1* ... *#K*`)` specifies the anticipated outcome probabilities in the groups that will be compared. *#1* is the anticipated probability in the control group (π_1^a), and *#2*, ..., *#K* are the anticipated probabilities in the treatment groups (π_2^a, ..., π_K^a). `pr()` is required.

`margin(`*#*`)` is used with two-group trials and must be specified if a noninferiority or substantial-superiority trial is being designed. The default is `margin(0)`, denoting a superiority trial. If the event of interest is unfavorable, the null hypothesis for all of these designs is $\pi_2 - \pi_1 \geq m$, where m is the prespecified margin. The alternative hypothesis is $\pi_2 - \pi_1 < m$. $m > 0$ denotes a noninferiority trial, whereas $m < 0$ denotes a substantial-superiority trial. On the other hand, if the event of interest is favorable, the above inequalities are reversed. The null hypothesis for all of these designs is then $\pi_2 - \pi_1 \leq m$, and the alternative hypothesis is $\pi_2 - \pi_1 > m$. $m < 0$ denotes a noninferiority trial, while $m > 0$ denotes a substantial-superiority trial. The hypothesized margin for the difference in anticipated probabilities, *#*, must lie between -1 and 1.

`unfavourable` | `unfavorable` or `favourable` | `favorable` are used with two-group trials to specify whether the outcome is unfavorable or favorable. If either option is used, `artbin` checks the assumptions; otherwise, it infers the favorability status. American and English spellings are both allowed.

`power(`*#*`)` specifies the required power of the trial at the `alpha()` significance level and computes the total sample size. `power()` cannot be used with `n()`. The default is `power(0.8)`.

`n(`*#*`)` specifies the total sample size available and computes the corresponding power. `n()` cannot be used with `power()`. The default is to calculate the sample size for power 0.8.

`aratios(`*aratio_list*`)` specifies the allocation ratios. The allocation ratio for group k is *#k*, $k = 1, \ldots, K$; for example, `aratios(1 2)` means that two participants are randomized to the experimental group for each one randomized to the control group. With two groups, `aratios(`*#*`)` is taken to mean `aratios(1` *#*`)`. The default is equal allocation to all groups.

`ltfu(`*#*`)` assumes a proportional loss to follow-up of #, where # is a number between 0 and 1. The total sample size is divided by 1−# before rounding. The default is `ltfu(0)`, meaning no loss to follow-up.

`alpha(`*#*`)` specifies that the trial will be analyzed using a significance test with level #. That is, # is the type 1 error probability. The default is `alpha(0.05)`.

`onesided` is used for two-group trials and for trend tests in multigroup trials. It specifies that the significance level given by `alpha()` is one sided. Otherwise, the value of `alpha()` is halved to give a one-sided significance level. Thus, for example, `alpha(0.05)` is exactly the same as `alpha(0.025) onesided`.

`artbin` always assumes that a two-group trial or a trend test in a multigroup trial will be analyzed using a one-sided alternative, regardless of whether the alpha level was specified as one sided or two sided. `artbin`, therefore, uses a slightly different definition of power from the `power` command: when a two-tailed test is performed, `power` reports the probability of rejecting the null hypothesis in either direction, whereas `artbin` only considers rejecting the null hypothesis in the direction of interest.

`artbin` assumes that multigroup trials will be analyzed using a two-sided alternative, so `onesided` is not allowed with multigroup trials unless `trend` or `doses()` is specified (see below).

`trend` is used for trials with more than two groups and specifies that the trial will be analyzed using a linear trend test. The default is a test for any difference between the groups. See also `doses()`.

`doses(`*dose_list*`)` is used for trials with more than two groups and specifies "doses" or other quantitative measures for a dose–response (linear trend) test. `doses()` implies `trend`. `doses(`*#1 #2* ... *#r*`)` assigns doses for groups $1, \dots, r$. If $r < K$ (the total number of groups), the dose is assumed equal to *#r* for groups $r+1, r+2, \dots, K$. If `trend` is specified without `doses()`, then the default is `doses(1 2` ... *K*`)`. `doses()` is not permitted for a two-group trial.

`condit` specifies that the trial will be analyzed using Peto's conditional test. This test conditions on the total number of events observed and is based on Peto's local approximation to the log odds-ratio. This option is also likely to be a good approximation with other conditional tests. The default is the usual Pearson χ^2 test. `condit` is not available for noninferiority and super-superiority trials. `condit` cannot be used with `wald`, because only one test type is allowed. `condit` implies `local`. The `ccorrect` option is not available with `condit`.

`wald` specifies that the trial will be analyzed using the Wald test. The default is the usual Pearson χ^2 test. `wald` cannot be used with `condit`, because only one test type is allowed. The Wald test inherently allows for distant alternatives, so `wald` and `local` cannot be used together.

`ccorrect` specifies that the trial will be analyzed with a continuity correction. `ccorrect` is not available with `condit`. The default is no continuity correction.

`local` specifies that the calculation should use the variance of the difference in proportions only under the null. This approximation is valid when the treatment effect is small. The default uses the variance of the difference in proportions both under the null and under the alternative hypothesis. The local method is not recommended and is only included to allow comparisons with other software. The Wald test inherently allows for distant alternatives, so `wald` and `local` cannot be used together.

`noround` prevents rounding of the calculated sample size in each group up to the nearest integer. The default is to round.

`force` can be used with two-group studies to override the program's inference of the `favorable` or `unfavorable` outcome type. This may be needed, for example, when designing an observational study with a harmful risk factor; the favorability types would be reversed and the `force` option applied.

3 Examples

3.1 Binary outcome and comparison with published sample size

We reproduce the sample-size calculation in Pocock (1983) for a two-group superiority trial comparing the efficacy of therapeutic doses of Anturan in patients after a myocardial infarction with the placebo standard treatment. The primary outcome was death from any cause within one year of first treatment. The control (placebo) group was anticipated to have a 10% probability of death within one year and the Anturan treatment group a 5% probability, with the trial powered at 90%. The patient outcome was binary: either failure (death in a year) or success (survival). The published sample size was 578 patients per group (1,156 patients in total).

In the below `artbin` example, we do not specify in the syntax whether the outcome is favorable or unfavorable; rather, we let the program infer it. The aim of a clinical trial is always to improve patient outcome. Therefore, because the experimental-group anticipated probability ($\pi_2^a = 0.05$) is less than the control-group anticipated probability ($\pi_1^a = 0.1$), it can be inferred that the outcome is unfavorable (that is, the trial is aiming to reduce the probability of the event occurring, in this case, death).

```
. artbin, pr(0.1 0.05) alpha(0.05) power(0.9) wald
ART - ANALYSIS OF RESOURCES FOR TRIALS (binary version 2.0.1 09june2022)
──────────────────────────────────────────────────────────────────────────────
A sample size program by Abdel Babiker, Patrick Royston, Friederike Barthel,
Ella Marley-Zagar and Ian White
MRC Clinical Trials Unit at UCL, London WC1V 6LJ, UK.
──────────────────────────────────────────────────────────────────────────────
Type of trial                          superiority
Number of groups                       2
Favourable/unfavourable outcome        unfavourable
                                       Inferred by the program
Allocation ratio                       equal group sizes
Statistical test assumed               unconditional comparison of 2
                                        binomial proportions
                                        using the wald test
Local or distant                       distant
Continuity correction                  no
Anticipated event probabilities        0.100  0.050
Alpha                                  0.050 (two-sided)
                                       (taken as .025 one-sided)
Power (designed)                       0.900
Total sample size (calculated)         1156
Sample size per group (calculated)     578 578
Expected total number of events        86.70
──────────────────────────────────────────────────────────────────────────────
```

The `artbin` output table shows the trial setup information, including the study design, statistical tests, and methods used. The hypothesis tests are shown with the calculated sample size and events based on the selected power. A total sample size of 1,156 participants is required, as per the published sample size given by Pocock (1983). The same result is achieved by the command `artbin, pr(0.9 0.95) alpha(0.05) power(0.9) wald`, assuming a favorable outcome (survival) instead. The Wald test is used instead of the default score test because Pocock used the sample estimate in the method of estimating the variance of the difference in proportions under the null hypothesis H_0.

3.2 Binary outcome and comparison with power

We compare the output of `artbin` with the output of Stata's `power` command, which, like `artbin`, uses the score test as the default.

```
. power twoproportions 0.1 0.05, alpha(0.05) power(0.9)
Performing iteration ...
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
H0: p2 = p1  versus  Ha: p2 != p1
Study parameters:
        alpha =    0.0500
        power =    0.9000
        delta =   -0.0500  (difference)
           p1 =    0.1000
           p2 =    0.0500
Estimated sample sizes:
            N =     1,164
  N per group =       582
. artbin, pr(0.1 0.05) alpha(0.05) power(0.9)
ART - ANALYSIS OF RESOURCES FOR TRIALS (binary version 2.0.1 09june2022)
──────────────────────────────────────────────────────────────────────────────
A sample size program by Abdel Babiker, Patrick Royston, Friederike Barthel,
Ella Marley-Zagar and Ian White
MRC Clinical Trials Unit at UCL, London WC1V 6LJ, UK.
──────────────────────────────────────────────────────────────────────────────
Type of trial                          superiority
Number of groups                       2
Favourable/unfavourable outcome        unfavourable
                                       Inferred by the program
Allocation ratio                       equal group sizes
Statistical test assumed               unconditional comparison of 2
                                        binomial proportions
                                        using the score test
Local or distant                       distant
Continuity correction                  no
Anticipated event probabilities        0.100  0.050
Alpha                                  0.050 (two-sided)
                                       (taken as .025 one-sided)
Power (designed)                       0.900
Total sample size (calculated)         1164
Sample size per group (calculated)     582 582
Expected total number of events        87.30
──────────────────────────────────────────────────────────────────────────────
```

Both give a total sample size of 1,164.

3.3 One-sided noninferiority trial

Next we show a one-sided noninferiority trial with the `onesided` option. We anticipate a 90% probability of survival in both the control group and the treatment group, with the null hypothesis that the treatment group is at least 5% less effective than the control.

```
. artbin, pr(0.9 0.9) margin(-0.05) onesided
ART - ANALYSIS OF RESOURCES FOR TRIALS (binary version 2.0.1 09june2022)
------------------------------------------------------------------------------
A sample size program by Abdel Babiker, Patrick Royston, Friederike Barthel,
Ella Marley-Zagar and Ian White
MRC Clinical Trials Unit at UCL, London WC1V 6LJ, UK.
------------------------------------------------------------------------------
Type of trial                         non-inferiority
Number of groups                      2
Favourable/unfavourable outcome       favourable
                                      Inferred by the program
Allocation ratio                      equal group sizes
Statistical test assumed              unconditional comparison of 2
                                       binomial proportions
                                       using the score test
Local or distant                      distant
Continuity correction                 no
Null hypothesis H0:                   H0: pi2 - pi1 <= -.05
Alternative hypothesis H1:            H1: pi2 - pi1 > -.05
Anticipated event probabilities       0.900  0.900
Alpha                                 0.050 (one-sided)
Power (designed)                      0.800
Total sample size (calculated)        914
Sample size per group (calculated)    457 457
Expected total number of events       822.60
------------------------------------------------------------------------------
```

A sample size of 457 is required in each group.

3.4 Superiority trial with multiple groups

Here we demonstrate a superiority trial with more than two groups. Instead of comparing each of the treatment groups with the control group, `artbin` uses a global test to assess if there is any difference among the groups.

```
. artbin, pr(0.1 0.2 0.3 0.4) alpha(0.1) power(0.9)
ART - ANALYSIS OF RESOURCES FOR TRIALS (binary version 2.0.1 09june2022)
-------------------------------------------------------------------------------
A sample size program by Abdel Babiker, Patrick Royston, Friederike Barthel,
Ella Marley-Zagar and Ian White
MRC Clinical Trials Unit at UCL, London WC1V 6LJ, UK.
-------------------------------------------------------------------------------
Type of trial                          superiority
Number of groups                       4
Favourable/unfavourable outcome        not determined
Allocation ratio                       equal group sizes
Statistical test assumed               unconditional comparison of 4
                                        binomial proportions
                                        using the score test
Local or distant                       distant
Continuity correction                  no

Anticipated event probabilities        0.100 0.200 0.300 0.400

Alpha                                  0.100 (two-sided)
Power (designed)                       0.900

Total sample size (calculated)         176

Sample size per group (calculated)     44 44 44 44
Expected total number of events        44.00
-------------------------------------------------------------------------------
```

A sample size of 44 is required in all four groups.

3.5 Complex noninferiority trial in a real-life setting

Finally, we demonstrate a more complex noninferiority design from the STREAM trial. The need for the STREAM trial arose from the increase of multidrug-resistant strains of tuberculosis, especially in countries without robust healthcare systems that were unable to administer treatment over long periods of time. The STREAM trial evaluated a shorter, more intensive treatment for multidrug-resistant tuberculosis compared with the lengthier treatment recommended by the World Health Organization.

A favorable outcome was defined by cultures negative for mycobacterium tuberculosis at 132 weeks and at a previous occasion, with no intervening positive culture or previous unfavorable outcome (Nunn et al. 2019). The sample-size calculation used an anticipated 0.7 probability of a favorable outcome on control (π_1^a) and 0.75 on treatment (π_2^a). Hence, it was assumed that 70% of the participants in the long-regimen group and 75% in the short-regimen group would attain a favorable outcome. A 10-percentage-point noninferiority margin was considered to be an acceptable difference in efficacy, given the shorter treatment duration ($m = -0.1$ defined as $\pi_2 - \pi_1$). It was assumed there were twice as many patients in treatment compared with control. The `wald` test was applied because it is often used in noninferiority trials.

```
. artbin, pr(0.7 0.75) margin(-0.1) power(0.8) aratios(1 2) wald ltfu(0.2)
ART - ANALYSIS OF RESOURCES FOR TRIALS (binary version 2.0.1 09june2022)
------------------------------------------------------------------------------
A sample size program by Abdel Babiker, Patrick Royston, Friederike Barthel,
Ella Marley-Zagar and Ian White
MRC Clinical Trials Unit at UCL, London WC1V 6LJ, UK.
------------------------------------------------------------------------------
Type of trial                         non-inferiority
Number of groups                      2
Favourable/unfavourable outcome       favourable
                                      Inferred by the program
Allocation ratio                      1:2
Statistical test assumed              unconditional comparison of 2
                                       binomial proportions
                                       using the wald test
Local or distant                      distant
Continuity correction                 no
Null hypothesis H0:                   H0: pi2 - pi1 <= -.1
Alternative hypothesis H1:            H1: pi2 - pi1 > -.1
Anticipated event probabilities       0.700  0.750
Alpha                                 0.050 (two-sided)
                                      (taken as .025 one-sided)
Power (designed)                      0.800
Loss to follow up assumed:            20 %
Total sample size (calculated)        399
Sample size per group (calculated)    133 266
Expected total number of events       292.60
------------------------------------------------------------------------------
```

The noninferiority trial required a total sample size of 399 (133 in control and 266 in treatment), assuming 20% of patients were not assessable in primary analysis. When the STREAM trial concluded, it estimated that a shorter, more intensive treatment for multidrug-resistant tuberculosis was only 1% less effective than the lengthier treatment recommended by the World Health Organization and demonstrated significant evidence of noninferiority.

4 Menu and dialogs

All the features in `artbin` are available from the `artbin` menu and associated dialogs. Once the selections have been inputted into the menu box, the associated command line will be displayed in the Review window. If the user would like to generate a do-file to reproduce the calculations, a log file can be opened before executing the commands via the dialog, which will then save the command line.

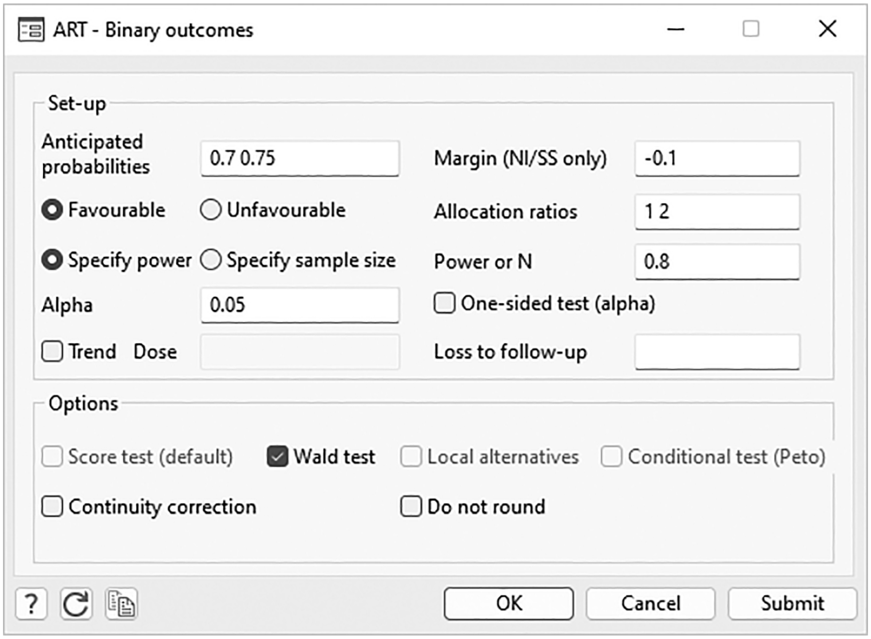
Once the ART package has been installed in Stata, the `artbin` dialog menu can be used. To access the interactive menu, type `artmenu on`, which will cause a new item, ART, to appear on the system menu bar under User. To turn this menu off, type `artmenu off`. ART consists of three programs, namely,

- survival outcomes (corresponding to `artsurv`),
- projection of events and power (corresponding to `artpep`), and
- binary outcomes (corresponding to `artbin`).

`artsurv` and `artpep` are described in Barthel, Royston, and Babiker (2005) and Royston and Barthel (2010), respectively.

Compared with previous versions, new options such as *Margin*, *Favourable* or *Unfavourable*, *Loss to follow-up*, *Score test*, *Wald test*, *Continuity correction*, and *Do not round* have now been included within an updated layout design.

Figure 1 illustrates the dialog box for binary outcomes. The `artbin` dialog box allows the user to input the parameters for the trial setup. Options are deselected based on the user's choices; for example, if the *Conditional test (Peto)* checkbox is selected, then the *Wald test* checkbox will be grayed out.



ART - Binary outcomes

Set-up

Anticipated probabilities: 0.7 0.75

Margin (NI/SS only): -0.1

● Favourable ○ Unfavourable

Allocation ratios: 1 2

● Specify power ○ Specify sample size

Power or N: 0.8

Alpha: 0.05

□ One-sided test (alpha)

□ Trend Dose

Loss to follow-up

Options

□ Score test (default) ☑ Wald test □ Local alternatives □ Conditional test (Peto)

□ Continuity correction □ Do not round

Figure 1. Example of a completed `artbin` menu for binary outcomes

The dialog box output is the same as the output in section 3.5 and corresponds to the inputs shown in the figure 1 menu box. The detailed display enables the user to check that the trial design has been inputted correctly.

5 Methods and formulas

5.1 Notation

Consider the design of a study to compare K independent groups in terms of a binary outcome whose probability of occurrence for an individual in group k is π_k, $k = 1, 2, \ldots, K$. We refer to group 1 as a control group and groups $2, \ldots, K$ as experimental groups.

Let Y_k be the number of events in a sample of size $n_k = r_k N$ from a total sample size N, where r_k is the fraction allocated to group k for $k = 1, 2, \ldots, K$. Then Y_k has the binomial distribution $\text{binom}(n_k, \pi_k)$. Write $\overline{\pi} = \sum_{k=1}^{K} r_k \pi_k$ as the overall outcome probability. Let $Y_{\cdot} = \sum_k Y_k$. The estimated outcome probabilities $\widehat{\pi}_k$ and $\widehat{\overline{\pi}}$ are $\widehat{\pi}_k = Y_k / r_k N$ and $\widehat{\overline{\pi}} = Y_{\cdot}/N = \sum_{k=1}^{K} r_k \widehat{\pi}_k$.

We consider the general case and then the case $K = 2$. For each case, we define a test statistic and derive its distribution under the null and alternative hypotheses (section 5.2). We then apply generic methods to derive sample sizes or powers (section 5.3).

5.2 Summary of test statistics and their distributions

Unconditional methods are based on a score vector $\mathbf{U} = (U_2, \ldots, U_K)'$, where $U_k = \widehat{\pi}_k - \widehat{\overline{\pi}}$. Conditional methods are based on a different score vector $\mathbf{X} = (X_2, \ldots, X_K)'$, where $X_k = Y_k - r_k Y_{\cdot} = r_k N U_k$. Table 1 shows the test statistics and their null and alternative distributions. See appendix 2 for further details of definitions, such as Q, $\mathbf{V}$, $\mathbf{A}$, M, and T. All methods are unconditional unless otherwise stated. The approximate distant method is based on the work of Yuan and Bentler (2010).

Table 1. Summary of test statistics and their distributions

Method	Statistic	Distribution	
		Null	Alternative
K groups, heterogeneity			
Score local	$Q_u = N\mathbf{U}'\widehat{\mathbf{V}}_u^{-1}\mathbf{U}$ $\widehat{\mathbf{V}}_u = N\widehat{\text{var}}\,(\mathbf{U}\|H_0)$	χ^2_{K-1}	$NC\chi^2(K-1,\lambda)$ $\lambda = N\boldsymbol{\mu}'\mathbf{V}_u^{-1}\boldsymbol{\mu}$ $\mu_k = \pi_k^a - \overline{\pi}^a$
Score distant approximate	same	same	$cNC\chi^2(K-1,\gamma)$ Yuan and Bentler (2010) with equations for c,γ (see appendix 2)
Wald	$Q_w = N\mathbf{U}'\widehat{\mathbf{A}}^{-1}\mathbf{U}$ $\widehat{\mathbf{A}} = N\widehat{\text{var}}\,(\mathbf{U}\|H_a)$	χ^2_{K-1}	$NC\chi^2(K-1,\lambda)$ $\lambda = N\boldsymbol{\mu}'\mathbf{A}^{-1}\boldsymbol{\mu}$
Conditional local	$Q_c = \mathbf{X}'\mathbf{V}_c^{-1}\mathbf{X}/M$ $M = \overline{\overline{\pi}}(1-\overline{\overline{\pi}})N^2/(N-1)$ $\mathbf{V}_c = \text{var}\,(\mathbf{X}\|H_0)\,/M$	χ^2_{K-1}	$NC\chi^2(K-1,\lambda)$ $\lambda = M\boldsymbol{\eta}'\mathbf{V}_c\boldsymbol{\eta}$ $\eta_k = \text{logit } \pi_k^a - \text{logit } \pi_1^a$
K groups, trend			
Score local	$T_u = \mathbf{c}'\mathbf{U}$ $c_k = r_k(d_k - d_1)$ where $d_1, d_2, \ldots, d_k$ are doses for groups $1, 2, \ldots, k$	$\mathbf{N}(0, \mathbf{c}'\mathbf{V}_u\mathbf{c}/N)$	$\mathbf{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{V}_u\mathbf{c}/N)$
Score distant	same	same	$\mathbf{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{A}\mathbf{c}/N)$
Wald	same	$\mathbf{N}(0, \mathbf{c}'\mathbf{A}\mathbf{c}/N)$	$\mathbf{N}(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{A}\mathbf{c}/N)$
Conditional local	$T_c = \mathbf{c}'\mathbf{X}/M$	$\mathbf{N}(0, \mathbf{c}'\mathbf{V}_c\mathbf{c}/M)$	$\mathbf{N}(\mathbf{c}'\mathbf{V}_c\boldsymbol{\eta}, \mathbf{c}'\mathbf{V}_c\mathbf{c}/M)$
Two groups, superiority or noninferiority			
All	$T_2 = \widehat{\delta} - m$ $\widehat{\delta} = \widehat{\pi}_2 - \widehat{\pi}_1$ $m =$ margin	$\mathbf{N}(0, V_n/N)$	$\mathbf{N}(\delta - m, V_a/N)$ $V_a = \frac{\pi_1^a(1-\pi_1^a)}{r_1} + \frac{\pi_2^a(1-\pi_2^a)}{r_2}$
In the above, $V_n = \{\widetilde{\pi}_1^a(1-\widetilde{\pi}_1^a)\}/r_1 + \{\widetilde{\pi}_2^a(1-\widetilde{\pi}_2^a)\}/r_2$, where $\widetilde{\pi}_1^a$ and $\widetilde{\pi}_2^a$ are values of π_1^a and π_2^a modified to conform to H_0 in one of the following ways:			
Score distant	Maximum likelihood estimates of π_1 and π_2 constrained to $\delta = m$		
Score local	As score, but replacing V_a with V_n		
Wald	$\widetilde{\pi}_1^a = \pi_1^a$ and $\widetilde{\pi}_2^a = \pi_2^a$ (so $V_n = V_a$)		
Conditional local	Methods for K groups are used (superiority trial only)		

5.3 Summary of methods

5.3.1 K groups, heterogeneity

The following statistics are approximated as χ^2_{K-1} under the null. Let $x_\alpha(m)$ be the $(1-\alpha)100$th percentile of the (central) χ^2 distribution with m degrees of freedom. Then, for a test statistic for which we write S_N to emphasize its dependence on sample size N, power is related to the total sample size N by the equation

$$\text{power} = \Pr\{S_N > x_\alpha(K-1)|H_a\} \tag{1}$$

The distributions under the alternative hypothesis are all of the form cX, where c is a constant depending on N and X is a noncentral χ^2 random variable with $K-1$ degrees of freedom and noncentrality parameter λ depending on N and the anticipated probabilities. Then (1) gives the key equation

$$\text{power} = 1 - F_{K-1,\lambda}\left\{x_\alpha(K-1)/c\right\}$$

where $F_{K-1,\lambda}(x)$ is the cumulative distribution function of the noncentral χ^2 distribution with $K-1$ degrees of freedom and noncentrality parameter λ. We can directly evaluate this for power given N. Solving for N given power involves iterative methods in some cases.

5.3.2 All other cases

These statistics S_N are all approximated as $\mathbf{N}(0,\sigma_0^2/N)$ under H_0 and $\mathbf{N}(\mu_1,\sigma_1^2/N)$ under H_a, where σ_1 depends on the anticipated probabilities. Let z_a denote the $(1-a)100$th percentile of the standard normal distribution, where, for a one-sided test, $a=\alpha$, and for a two-sided test, $a=\alpha/2$. Then (1) gives the key equation

$$\text{power} = \Pr\left(S_N > z_a\sigma_0/\sqrt{N}|H_a\right) = \Phi\left(\frac{\mu_1 - z_a\sigma_0/\sqrt{N}}{\sigma_1/\sqrt{N}}\right)$$

Rearranging, the total sample size to achieve power $1-\beta$ is

$$N = \left(\frac{z_a\sigma_0 + z_\beta\sigma_1}{\mu_1}\right)^2$$

6 Software testing

`artbin` is for use in the design of randomized trials, so we have tested it extensively. The program was modified by Ella Marley-Zagar and tested by Ella Marley-Zagar, Ian R. White, Patrick Royston, and Abdel G. Babiker. We report the testing methods below to verify both the sample-size and the power results. We ran the test scripts with the default variable type (`set type`) as `float` and as `double`.

1. We compared results for noninferiority trials with those given by Julious and Owen (2011), Blackwelder (1982), Pocock (2003), and the online calculator Sealed Envelope (2012). Exact agreement was achieved.

2. We compared results for a superiority binary outcome with those given by Pocock (1983) and the online calculator Sealed Envelope (2012). Exact agreement was achieved.

3. We tested several scenarios including continuity correction results given by `artbin` and those given by the Stata program `power`. The results from both programs were in agreement.

4. We checked the results given by `artbin` using the `margin()` option against Julious and Owen (2011). Exact agreement was achieved.

5. The output of `artbin` was compared with Cytel's software EAST, which is a sophisticated package able to produce sample-size and power calculations for several binary outcomes in clinical trial settings. We achieved perfect agreement in all but a handful of cases where the sample size differed by 1, which we believe is due to the difference in the way the packages round sample size.

6. For the new syntax options, we tested `onesided` for a one-sided test and `ccorrect` to apply a continuity correction.

7. We tested every permutation of two-group and more than two-group and noninferiority, substantial-superiority, and superiority trials with margin, local or distant, conditional or unconditional, trend, and Wald test options to check that the results were as expected and that sample size was increased or decreased accordingly.

8. We checked error messages in several impossible cases to ensure that we obtained error messages as required.

9. We tested the dialog box menu options to verify that the results were as required.

7 Conclusions

We have written `artbin` to include new syntax with additional options, including extensions to the tests and methods offered by previous versions of the software. We have also refreshed the layout of the dialog box for `artbin`, with mutually exclusive options grayed out for clarity. The updated `artbin` program compares well with Stata's `power` program, as well as other commercially available products such as Cytel's EAST and the Sealed Envelope Calculator. One of the main features of `artbin` that sets it apart from the other available software in Stata is the range of trial types, statistical tests, and methods that it offers for sample-size calculation. Notably, Stata's `power` can provide sample size for superiority trials only.

As noted in section 2.2, `artbin` reports power as the probability of rejecting the null hypothesis in the direction of interest, whereas `power` reports the probability of rejecting the null hypothesis in either direction if a two-tailed test is performed. We believe the former is more appropriate for a clinical trial. Technically, this procedure is conservative, but the difference matters only for unrealistically large alpha.

The majority of noninferiority trials are designed so that $\pi_1^a = \pi_2^a$. However, `artbin` allows more flexibility where π_1^a and π_2^a can differ, as in section 3.5. The noninferiority margin is expressed on the risk-difference scale, and the results would be very different for other scales (Quartagno et al. 2020). All calculations in `artbin` are based on the approximation that the difference in proportions is normally distributed (or for the conditional case that the score statistic is normally distributed). This approximation may fail with very small sample sizes, in which case the continuity correction should be used. We suggest using the usual rule for the Pearson χ^2 test, namely, to mistrust the results when any expected cell count is lower than about 5. Concerned users should check the power by simulation.

We have not so far offered advice on which method to use. In our experience, analysts often use the score test for superiority trials and the Wald test for noninferiority trials. For small trials, conditional tests are often used. With small differences in probabilities, all tests give similar results. We recommend avoiding the Wald test when there are large differences in probabilities, and we would never use the `local` option except when comparing results from other programs.

Furthermore, the design of multigroup trials in `artbin` is based on testing the global null hypothesis evaluating if there is a difference between any of the groups. The latter is in contrast to the case of comparing each group with the control. This can, however, be achieved by applying the two-group case; if the familywise error rate is to be controlled, this can be done by dividing alpha by the number of comparisons.

`artbin` has been created to assist the design of clinical trials, but it can also be used in the design of observational studies to explore a protective or harmful factor. The trial and outcome types may need to be reinterpreted; for example, for a harmful risk factor in an observational study, the favorable or unfavorable outcome types would be reversed. This would be an example of when the option `force` would be used. An observational study design to demonstrate a protective factor could be designed in exactly the same way as a trial, but the term *superiority* might be replaced by *benefit*. This is further described in the newly available `artcat`, a Stata program to calculate sample size or power for a two-group trial with an ordered categorical outcome (White et al. 2023).

A useful future extension will be for `artbin` to handle the conditional test for noninferiority or substantial-superiority trials.

8 Acknowledgments

This work was supported by the Medical Research Council Unit Programme number MC_UU_00004/09. We thank Henry Bern and Tim Morris for their very helpful comments and for testing the program.

9 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 23-1
. net install st0013_3      (to install program files, if available)
. net get st0013_3          (to install ancillary files, if available)
```

The artbin command also is available on the Statistical Software Components archive and can be installed directly in Stata with the command

```
. ssc install art
```

All the code we used for testing and the output testing files are included in the package. The files are also available along with the program itself on the GitHub repository https://github.com/UCL/artbin.

10 References

Attanasio, O., A. D. Kugler, and C. Meghi. 2011. Subsidizing vocational training for disadvantaged youth in Colombia: Evidence from a randomized trial. *American Economic Journal: Applied Economics* 3: 188–220. https://doi.org/10.1257/app.3.3.188.

Barthel, F. M.-S., A. Babiker, P. Royston, and M. K. B. Parmar. 2006. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Statistics in Medicine* 25: 2521–2542. https://doi.org/10.1002/sim.2517.

Barthel, F. M.-S., P. Royston, and A. Babiker. 2005. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome: Update. *Stata Journal* 5: 123–129. https://doi.org/10.1177/1536867X0500500114.

Blackwelder, W. C. 1982. "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials* 3: 345–353. https://doi.org/10.1016/0197-2456(82)90024-1.

Box, G. E. P. 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* 25: 290–302. https://doi.org/10.1214/aoms/1177728786.

Braga, A. A., D. L. Weisburd, E. J. Waring, L. G. Mazerolle, W. Spelman, and F. Gajewski. 1999. Problem-oriented policing in violent crime places: A randomized controlled experiment. *Criminology* 37: 541–580. https://doi.org/10.1111/j.1745-9125.1999.tb00496.x.

Clark, T., U. Berger, and U. Mansmann. 2013. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: Review. *BMJ* 346: f1135. https://doi.org/10.1136/bmj.f1135.

Farrington, C., and G. Manning. 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* 9: 1447–1454. https://doi.org/10.1002/sim.4780091208.

Fleiss, J. L., A. Tytun, and H. K. Ury. 1980. A simple approximation for calculating sample sizes for comparing independent proportions. *International Biometric Society* 36: 343–346. https://doi.org/10.2307/2529990.

Julious, S. A., and M. J. Campbell. 2012. Tutorial in biostatistics: Sample sizes for parallel group clinical trials with binary data. *Statistics in Medicine* 31: 2904–2936. https://doi.org/10.1002/sim.5381.

Julious, S. A., and R. J. Owen. 2011. A comparison of methods for sample size estimation for non-inferiority studies with binary outcomes. *Statistical Methods in Medical Research* 20: 595–612. https://doi.org/10.1177/0962280210378945.

Krause, P., T. R. Fleming, I. Longini, A. M. Henao-Restrepo, and R. Peto. 2020. COVID-19 vaccine trials should seek worthwhile efficacy. *Lancet* 396: 741–743. https://doi.org/10.1016/S0140-6736(20)31821-3.

Mathai, A., and S. Provost. 1992. *Quadratic Forms in Random Variables: Theory and Applications*. New York: Dekker.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall/CRC.

Nunn, A. J., P. P. J. Phillips, S. K. Meredith, C.-Y. Chiang, F. Conradie, D. Dalai, A. van Deun, et al. 2019. A trial of a shorter regimen for rifampin-resistant tuberculosis. *New England Journal of Medicine* 380: 1201–1213. https://doi.org/10.1056/NEJMoa1811867.

Pocock, S. J. 1983. *Clinical Trials: A Practical Approach*. Chichester, U.K.: Wiley.

———. 2003. The pros and cons of noninferiority trials. *Fundamental and Clinical Pharmacology* 17: 483–490. https://doi.org/10.1046/j.1472-8206.2003.00162.x.

Quartagno, M., A. S. Walker, A. G. Babiker, R. M. Turner, M. K. B. Parmar, A. Copas, and I. R. White. 2020. Handling an uncertain control group event risk in non-inferiority trials: Non-inferiority frontiers and the power-stabilising transformation. *Trials* 21: 145. https://doi.org/10.1186/s13063-020-4070-4.

Quigley, J. M., J. C. Thompson, N. J. Halfpenny, and D. A. Scott. 2019. Critical appraisal of nonrandomized studies—A review of recommended and commonly used tools. *Journal of Evaluation in Clinical Practice* 25: 44–52. https://doi.org/10.1111/jep.12889.

Rencher, A. C., and G. B. Schaalje. 2008. *Linear Models in Statistics*. 2nd ed. Hoboken, NJ: Wiley.

Royston, P., and F. M.-S. Barthel. 2010. Projection of power and events in clinical trials with a time-to-event outcome. *Stata Journal* 10: 386–394. https://doi.org/10.1177/1536867X1001000306.

Satterthwaite, F. E. 1941. Synthesis of variance. *Psychometrika* 6: 309–316. https://doi.org/10.1007/BF02288586.

Sealed Envelope. 2012. Power calculator for binary outcome non-inferiority trial. https://www.sealedenvelope.com/power/binary-noninferior/.

Welch, B. L. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29: 350–362. https://doi.org/10.2307/2332010.

White, I. R., E. Marley-Zagar, T. P. Morris, M. K. B. Parmar, P. Royston, and A. G. Babiker. 2023. artcat: Sample-size calculation for an ordered categorical outcome. *Stata Journal* 23: 3–23. https://doi.org/10.1177/1536867X231161934.

Yuan, K.-H., and P. M. Bentler. 2010. Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology* 63: 273–291. https://doi.org/10.1348/000711009X449771.

About the authors

Ella Marley-Zagar is a senior research associate and medical statistician in methodological software at the MRC Clinical Trials Unit in London, U.K. Her interests include developing new software and research into health and the environment, particularly issues affecting lower- and middle-income countries.

Ian White is a professor of statistical methods for medicine at the MRC Clinical Trials Unit in London, U.K., where he coleads programs of design of clinical trials, analysis of clinical trials, and meta-analysis. His research interests include study design, handling missing data and noncompliance in clinical trials, statistical models for meta-analysis, and simulation studies. He is the author of other Stata commands, including `mvmeta`, `network`, and `simsum`.

Patrick Royston is a medical statistician with more than 40 years of experience and a strong interest in biostatistical methods and in statistical computing and algorithms. He works largely in methodological issues in the design and analysis of clinical trials and observational studies. He is currently focusing on alternative outcome measures and tests of treatment effects in trials with a time-to-event outcome and nonproportional hazards, on parametric modeling of survival data, and on novel clinical trial designs.

Sophie Barthel is currently a functional manager of the real world solutions group at PRA/ICON PLC. Her work includes consultancy in clinical research in the areas of clinical trials and real world data. She is a published author of international research papers in statistics and eating disorders and has presented at many international conferences, including several invited presentations.

Mahesh Parmar is a professor of medical statistics and epidemiology and the director of the MRC Clinical Trials Unit at University College London and the Institute of Clinical Trials and Methodology at University College London. The unit he directs is at the forefront of resolving internationally important questions, particularly in infectious diseases, cancer, and more recently neurodegenerative diseases, and it also aims to deliver swifter and more effective translation of scientific research into patient benefits by carrying out challenging and innovative studies and by developing and implementing methodological advances in study design, conduct, and analysis. Examples of his methodological contributions include the development and implementation of the MAMS platform and DURATIONS designs.

Abdel Babiker is a professor of epidemiology and medical statistics at the MRC Clinical Trials Unit at University College London. He works on clinical trials in infectious diseases, including HIV, influenza, and COVID-19, and associated methodology.

Appendix 1: Description of what has changed

Program structure

`artbin` calls the subroutine `art2bin` for all two-group trials, which also allows for substantial-superiority trials. Previously, `art2bin` was called only for noninferiority trials in `artbin`; now it is called for all two-group trials. `art2bin` can be used as a standalone command, but we do not recommend this.

New syntax

Some improvements have been made to `artbin`. The user will need to alter previous coding using `artbin` to accommodate the following changes.

The syntax for `artbin` has been updated to include a `margin()` option for two-group trials. For a noninferiority or substantial-superiority trial, the program will use `pr(p1 p2)` and the new option `margin()`. For example, in the previous version (`artbin version 1.1.2`), the syntax `artbin, pr(.2 .3) ni(1)` would now be specified as `artbin, pr(.2 .2) margin(.1)`. The option `ni()` is now redundant.

Previously, `local` was taken as the default in superiority trials. Now it is distant; the `distant()` option has been replaced by `local` in the syntax. Previous syntax (up to version 1.1.2) will need to be altered so that `artbin, pr(.1 .2) distant(1)` will now be `artbin, pr(.1 .2)` and `artbin, pr(.1 .2) distant(0)` will now be `artbin, pr(.1 .2) local`, for example.

The user may identify whether the outcome is favorable or unfavorable in the context of a two-group trial. With this information plus the margin, the program will then determine the type of trial (that is, noninferiority, substantial superiority, or superiority). If the user does specify favorable or unfavorable, the program will check the assumptions. If not, then the program will infer it. The `force` option can be used to override the program's inference of the favorability status, for example, in the design of observational studies.

The `wald` option has also been included for the Wald test as an alternative to the default score test.

Sample size per group is now reported, and rounding up to the nearest integer is performed per group. A `noround` option has been included for the case when the user does not want `artbin` to round the calculated sample size up to the nearest integer. A loss to follow-up option is now available.

The option `condit` always implies the `local` option because there is no conditional distant option available in `artbin`. If the conditional option is selected, then `local` will be used (instead of the default distant).

The allocation ratio reflects the fact that sample size is now rounded upward in each group rather than overall, and the expected number of events is calculated using the rounded sample size (unless the `noround` option for calculated sample size is used).

Earlier versions of `artbin` required several yes or no options to be specified numerically, for example, `onesided(1)` or `onesided(0)`. In updating the syntax, we have enabled the more standard options, for example, `onesided` and `ccorrect`, but the numerical version of the syntax is retained if the user wishes to use it.

The number of groups is taken as the number of anticipated probabilities in all cases, and the `ngroups()` option is now redundant. The required option `pr()` now takes a numlist instead of a string.

Changes have been made to the output table so that results are presented in the same format whether `art2bin` was called. Now included in the description table are whether the trial is noninferiority, substantial superiority, or superiority; the trial outcome type; the statistical test assumed (including score or Wald); whether local or distant alternatives were used and the hypothesis tests; and whether the continuity correction was used. Minor formatting was also made to the existing allocation ratio, alpha, linear trend output, and version numbering output. Sample size per group is reported, and the returned values have been streamlined to include only results as opposed to user-inputted options.

The text output has been changed from $p0$ and $p1$ to π_1 and π_2. Therefore, the control-group anticipated outcome probability for noninferiority trials is π_1. The corresponding hypotheses tests (included in the output table) are

$$\begin{aligned} H_0 &: \pi_2 - \pi_1 >= / <= m \\ H_a &: \pi_2 - \pi_1 < / > m \end{aligned}$$

The program will now produce error or warning messages for disallowed or uncoded combinations of options, namely,

- noninferiority or substantial-superiority design with conditional test or trend,
- conditional test and nonlocal alternatives,
- conditional test and Wald test,
- Wald test and local alternatives, and
- continuity correction and the conditional case.

Also, an error message will be produced for >2 groups if the user specifies fewer numbers in `aratios()` than in `pr()`.

Appendix 2: Details of methods

Comparison of K anticipated probabilities

The unconditional tests are based on the score vector $\mathbf{U}=(U_2,\ldots,U_K)'$, where $U_k=\widehat{\pi}_k-\overline{\widehat{\pi}}$.

Define within-group variances $s_k=\pi_k(1-\pi_k)$ within group k and $\overline{s}=\sum_{k=1}^K r_k s_k$ overall, and define total variance $s=\overline{\pi}(1-\overline{\pi})$. Under the null hypothesis $H_0:\pi_1=\pi_2=\ldots=\pi_K$, we have $E(U_k|H_0)=0$ and $N\text{cov}(U_k,U_l|H_0)=v_{kl}=s(\delta_{kl}/r_k-1)$. Under the global alternative hypothesis $H_a:\pi_k\neq\pi_l$, for some k,l. Under the anticipated probabilities $\pi_k=\pi_k^a$ for all k, we have $E(U_k|H_a)=\mu_k=\pi_k^a-\overline{\pi}^a$ and $N\text{cov}(U_k,U_l|H_a)=a_{kl}=s_k(\delta_{kl}/r_k-1)-s_l+\overline{s}$.

Let $\boldsymbol{\mu}=(\mu_2,\ldots,\mu_K)'$; $\mathbf{V}_u=(v_{kl})_{k,l=2,\ldots,K}$; $\mathbf{A}=(a_{kl})_{k,l=2,\ldots,K}$; and $\widehat{\mathbf{V}}_u$ and $\widehat{\mathbf{A}}$ be the sample estimates of $\mathbf{V}_u$ and $\mathbf{A}$ obtained by replacing $\{\pi_k;k=1,\ldots,K\}$ by their sample estimates.

We consider first the unconditional score tests for heterogeneity and trend and then the equivalent Wald and conditional tests.

Unconditional score test for K groups

The score test statistic is $Q_u=N\mathbf{U}'\widehat{\mathbf{V}}_u^{-1}\mathbf{U}$. Direct expansion of the quadratic form shows that Q_u is equal to the Pearson statistic

$$Q_u=\sum_{k=1}^K\left(Y_k-r_kN\overline{\widehat{\pi}}\right)^2/\left\{r_kN\overline{\widehat{\pi}}\left(1-\overline{\widehat{\pi}}\right)\right\}$$

Asymptotically, under H_0, $Q_u \sim \chi^2_{K-1}$, a central χ^2 distribution with $K-1$ degrees of freedom. Denoting the $(1-\alpha)100$th percentile of the (central) χ^2 distribution with m degrees of freedom by $x_\alpha(m)$, power is related to the total sample size N by the equation

$$\text{power} = \Pr\{Q_u > x_\alpha(K-1)|H_a\}$$

There is no analytic solution to this equation for $K > 2$. We consider two ways to approximate the asymptotic distribution of Q_u under H_a in terms of a noncentral χ^2 distribution with $K-1$ degrees of freedom and noncentrality parameter λ, whose cumulative distribution function we denote as $F_{K-1,\lambda}(x)$.

Local alternative method. If $\max|\pi^a_i - \pi^a_j|$ is small, the asymptotic distribution of Q_u may be approximated by a noncentral χ^2 with $K-1$ degrees of freedom and noncentrality parameter $\lambda = N\boldsymbol{\mu}'\mathbf{V}_u^{-1}\boldsymbol{\mu} = N\sum_k \mu_k^2 r_k/s$.

Then the key equation is

$$\text{power} = 1 - F_{K-1,\lambda}\left\{x_\alpha(K-1)\right\} \tag{2}$$

which we solve for power given N or for N given power.

Approximate distant method. We instead approximate the distribution of Q_u by that of cX, where c is a constant and X is a noncentral χ^2 random variable with $K-1$ degrees of freedom and noncentrality parameter γ; c and γ both depend on N. Such an approximation for the two-group case was originally proposed by Welch (1938) and further studied by Satterthwaite (1941) and Box (1954). See also Yuan and Bentler (2010). The constant multiple c and the noncentrality parameter γ are calculated by equating the first two moments of Q_u and cX using well-known formulas for the mean and variance of quadratic forms of normal variables (Mathai and Provost 1992; Rencher and Schaalje 2008):

$$\begin{aligned} E(Q_u) &= tr(\mathbf{V}_u^{-1}\mathbf{A}) + N\boldsymbol{\mu}'\mathbf{V}_u^{-1}\boldsymbol{\mu} = c(K-1+\gamma) \\ \text{var}(Q_u) &= 2tr\left\{\left(\mathbf{V}_u^{-1}\mathbf{A}\right)^2\right\} + 4N\boldsymbol{\mu}'\mathbf{V}_u^{-1}\mathbf{A}\mathbf{V}_u^{-1}\boldsymbol{\mu} = 2c^2(K-1+2\gamma) \end{aligned}$$

We then modify (2) and solve the equation

$$\text{power} = 1 - F_{K-1,\gamma}\left\{x_\alpha(K-1)/c\right\}$$

Wald test. The Wald test statistic is $Q_w = N\mathbf{U}'\widehat{\mathbf{A}}^{-1}\mathbf{U}$, and the formulas for power and sample size are like those for the score test but with the covariance matrix $\mathbf{V}_u$ replaced by $\mathbf{A}$. Thus, the asymptotic distribution of Q_w is noncentral χ^2 with $K-1$ degrees of freedom and noncentrality parameter $\lambda = N\boldsymbol{\mu}'\mathbf{V}_u^{-1}\boldsymbol{\mu}$.

Unconditional score test for trend

For dose–response, the test for trend with dose scores $d_1, \ldots, d_K$ is based on the statistic $T_u = \sum_{k=1}^K r_k d_k U_k = \sum_{k=1}^K c_k U_k$, where $c_k = r_k(d_k - b)$ and b is an arbitrary constant (because $\sum_{k=1}^K r_k U_k = 0$). Taking $b = d_1$, we have $T = \sum_{k=2}^K c_k U_k = \mathbf{c}'\mathbf{U}$, where $\mathbf{c} = (c_2, \ldots, c_K)'$. The mean and variance of T under the null and alternative hypotheses are

$$E(T_u|H_0) = 0; \, \mathrm{var}(T_u|H_0) = \mathbf{c}'\mathbf{V}_u\mathbf{c}/N$$
$$E(T_u|H_a) = \mathbf{c}'\boldsymbol{\mu}; \, \mathrm{var}(T_u|H_a) = \mathbf{c}'\mathbf{A}\mathbf{c}/N$$

Let z_a denote the $(1-a)$100th percentile of the standard normal distribution. For a one-sided test, let $a = \alpha$, and for a two-sided test, let $a = \alpha/2$. The total sample size to achieve power $1-\beta$ for a distant test is

$$N = \left(\frac{z_a\sqrt{\mathbf{c}'\mathbf{V}_u\mathbf{c}} + z_\beta\sqrt{\mathbf{c}'\mathbf{A}\mathbf{c}}}{\mathbf{c}'\boldsymbol{\mu}} \right)^2$$

For a local test, $\mathbf{A}$ is replaced by $\mathbf{V}_u$. Conversely, for a Wald test, $\mathbf{V}_u$ is replaced by $\mathbf{A}$.

Conditional test

Some analyses condition on the margins of the contingency table of outcome by treatment group, for example, Fisher's exact test. For such analyses, a conditional calculation is preferred. As noted in the main text, this uses a different score vector $\mathbf{X} = (X_2, \ldots, X_K)'$, where $X_k = Y_k - r_k Y_.$ and $Y_. = \sum_{k=1}^K Y_k$ is the total number of events.

Let $\eta_k = \log\{\pi_k/(1-\pi_k)\} - \log\{\pi_1/(1-\pi_1)\}$ denote the log odds-ratio for the occurrence of the event in group k relative to group 1, $\boldsymbol{\eta} = (\eta_2, \ldots, \eta_K)'$. Conditional on $Y_. = y_.$, $\mathbf{Y} = (Y_2, \ldots, Y_K)'$ has a multivariate noncentral hypergeometric distribution with support

$$D = \left\{ (y_2, \ldots, y_K) : 0 \leq y_k \leq n_k, 0 \leq y_. - \sum_{k=2}^K y_k \leq n_1 \right\}$$

and probability function

$$f(y_2, \ldots, y_K) = \frac{1}{P} \prod_{k=1}^K \binom{n_k}{y_k} \exp(y_k \eta_k) \tag{3}$$

where $y_1 = y_. - \sum_{k=2}^K y_k$, $\eta_1 = 0$, and

$$P = \sum_{\{(y_2,\ldots,y_K)\epsilon D\}} \prod_{k=1}^K \binom{n_k}{y_k} \exp(y_k \eta_k)$$

Denote by $e(\boldsymbol{\eta})$ and $\mathbf{C}(\boldsymbol{\eta})$ the conditional mean and covariance matrix of $\mathbf{Y}$. Differentiating the log of $f(y_2,\ldots,y_K)$ in (3) with respect to $\boldsymbol{\eta}$ yields the score (gradient vector of the log-likelihood function) for $\boldsymbol{\eta}$

$$\mathbf{S}(\boldsymbol{\eta}) \overset{\text{def}}{=} \partial \log\{f(\mathbf{y})\}/\partial\boldsymbol{\eta}' = \mathbf{y} - E(\mathbf{Y}|y_{\cdot},\boldsymbol{\eta})$$

the observed minus the (conditionally) expected number of events in groups $2,\ldots,K$ given $y_{\cdot},\boldsymbol{\eta}$. Under the null hypothesis $H_0:\boldsymbol{\eta}=\mathbf{0}$, and conditional on $y_{\cdot}$, $\mathbf{Y}$ has a (central) hypergeometric distribution with the elements of the mean vector and covariance matrix

$$\begin{aligned} e_k(0) &\overset{\text{def}}{=} E(Y_k|y_{\cdot},\boldsymbol{\eta}=\mathbf{0}) = r_k y_{\cdot} \\ \mathbf{C}(0)_{kl} &\overset{\text{def}}{=} \text{cov}(Y_k,Y_l|y_{\cdot},\boldsymbol{\eta}=\mathbf{0}) = Mv_{kl} \end{aligned} \tag{4}$$

where

$$M = y_{\cdot}(N-y_{\cdot})/(N-1) \tag{5}$$

and $v_{kl} = r_k(\delta_{kl}-r_l)$ (McCullagh and Nelder 1989, chap. 7).

The score statistic $Q_c = \mathbf{S}(0)'\mathbf{C}(0)^{-1}\mathbf{S}(0)$ is a quadratic form based on the score vector $\mathbf{S}(0)$ and its covariance matrix $\mathbf{C}(0)$ under the null hypothesis. Denote the kth element of $\mathbf{S}(0)$ by $X_k = Y_k - r_k y_{\cdot}$, $\mathbf{X}=(X_2,\ldots,X_K)'$, and let $\mathbf{V}_c$ be the $(K-1)\times(K-1)$ matrix with elements v_{kl}. Using (4), the score statistic can be written as

$$Q_c = (\mathbf{X}/\sqrt{M})'\mathbf{V}_c^{-1}\left(\mathbf{X}/\sqrt{M}\right) = \mathbf{X}'\mathbf{V}_c^{-1}\mathbf{X}/M$$

Under H_0, the asymptotic distribution of Q_c is (central) χ^2 with $K-1$ degrees of freedom. However, there is no simple form for its asymptotic distribution under a general alternative hypothesis $H_a:\boldsymbol{\eta}\neq\mathbf{0}$. We use a local approach. Under a local alternative $[\eta_k \sim O(1/\sqrt{N})$ for $k=2,\ldots,K]$, $\mathbf{S}(\boldsymbol{\eta})$ can be approximated by a linear function using a first-order Taylor expansion about $\boldsymbol{\eta}=\mathbf{0}$,

$$\mathbf{S}(\boldsymbol{\eta}) \doteq \mathbf{S}(0) + \dot{\mathbf{S}}(0)\boldsymbol{\eta} \tag{6}$$

where $\dot{\mathbf{S}}(\boldsymbol{\eta}) = \partial\mathbf{S}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}' = \partial^2\log\{f(y)\}/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'$ is the matrix of second partial derivatives of the log likelihood with respect to $\boldsymbol{\eta}$. Note that $E\{\mathbf{S}(\boldsymbol{\eta})\}=0$ and $\text{cov}\{\mathbf{S}(\boldsymbol{\eta})\} = -E\{\dot{\mathbf{S}}(\boldsymbol{\eta})\}$. Taking the expectation of both sides of (6), we have

$$E(\mathbf{X}|\boldsymbol{\eta}) = E\{\mathbf{S}(0)|\boldsymbol{\eta}\} = M\mathbf{V}_c\boldsymbol{\eta}$$

Now let the anticipated value of $\boldsymbol{\eta}$ be $\boldsymbol{\eta}^a$ with $\eta_k^a = \log\{\pi_k^a/(1-\pi_k^a)\} - \log\{\pi_1^a/(1-\pi_1^a)\}$ for all k. Under a local alternative, the asymptotic distribution of Q_c is noncentral χ^2 with $K-1$ degrees of freedom and noncentrality parameter

$$\lambda = Mq(\boldsymbol{\eta}^a)$$

where $q(\boldsymbol{\eta}) = \boldsymbol{\eta}'\mathbf{V}_c\boldsymbol{\eta}$. We therefore have this equation relating power to M and hence to N:

$$\text{power} = 1 - F_{K-1,\lambda}\{x_\alpha(K-1)/c\} \tag{7}$$

Given N, we can compute M from (5) and hence compute power from (7). To compute N from power, we first use (7) to compute λ. We then solve for N as follows. Asymptotically, $M = T(N-T)/(N-1)$, where $T = E(Y_.) = N\overline{\pi}^a$ is the expected total number of events. It follows that

$$\lambda = T(N-T)q(\boldsymbol{\eta})/(N-1) = T(T/\overline{\pi}^a - T)q(\boldsymbol{\eta})/(T/\overline{\pi}^a - 1) \tag{8}$$

Equation (8) is a quadratic equation in T that can be expressed as

$$(1-\overline{\pi}^a)q(\eta)T^2 - \lambda T + \lambda\overline{\pi}^a = 0$$

The smaller solution is inappropriate, and so

$$T = \left\{\lambda + \sqrt{\lambda^2 - 4q(\boldsymbol{\eta})\lambda\overline{\pi}^a(1-\overline{\pi}^a)}\right\} / \left\{2(1-\overline{\pi}^a)q(\boldsymbol{\eta})\right\}$$

Finally, the total sample size $N = T/\overline{\pi}^a$.

Conditional test for trend

For dose–response, the test for trend with dose scores $d_1, \ldots, d_K$ is based on the statistic $T_c = \mathbf{c}'\mathbf{X}/\sqrt{M} = \sum_{k=1}^{K} c_k X_k/\sqrt{M}$, where as before $\mathbf{c} = (c_1, \ldots, c_K)'$; $c_k = r_k(d_k - d_1)$; and $M = y_.(N - y_.)/(N-1)$. The mean and variance of T_c under the null and alternative hypotheses are

$$E(T_c|H_0) = 0;\ \mathrm{var}(T_c|H_0) = \mathbf{c}'\mathbf{V}_c\mathbf{c}$$

$$E(T_c|H_a) = \sqrt{M}\mathbf{c}'\mathbf{V}_c\boldsymbol{\eta};\ \mathrm{var}(T_c|H_a) = \mathbf{c}'\mathbf{V}_c\mathbf{c}$$

The total sample size to achieve power $1-\beta$ is obtained from

$$M = \left(\frac{z_a\sqrt{\mathbf{c}'\mathbf{V}_c\mathbf{c}} + z_\beta\sqrt{\mathbf{c}'\mathbf{V}_c\mathbf{c}}}{\mathbf{c}'\mathbf{V}_c\boldsymbol{\eta}}\right)^2$$

and equating M to its asymptotic value

$$E(M) = E(Y.)\left\{N - E(Y_.)\right\}/(N-1) = E(Y.)\left\{N - E(Y_.)\right\}/(N-1)$$

and noting that $E(Y.) = N\overline{\pi}^a$ as in the derivation of (8).

Comparing two treatment groups: Noninferiority and substantial superiority

Two-arm studies to assess superiority of an experimental treatment use the formulas given above for K groups. In studies designed to assess noninferiority or substantial superiority of an experimental treatment (group 2) relative to a control treatment (group 1), the aim is to test whether the outcome in two treatment groups differs by more than a prespecified amount, and the single parameter of interest is $\delta = \pi_2 - \pi_1$. If the binary outcome is unfavorable, the null hypothesis for testing noninferiority takes the form H_0: $\delta \leq m$, where m is a prespecified margin and the alternative hypothesis is $H_a : \delta > m$. The null hypothesis is tested at its boundary $H_0 : \delta = m$. As above, let Y_i be the number of events in group i, $\widehat{\pi}_i = Y_i/n_i$, and $n_i = r_i N$, for $i = 1, 2$.

We consider test statistics of the form $T_* = \widehat{\delta} - m = \widehat{\pi}_2 - \widehat{\pi}_1 - m$, whose distribution under the null hypothesis is approximately $\mathbf{N}(0, V_n/N)$, for various definitions of the variance V_n (discussed below). The anticipated distribution of T_* under H_a is $\mathbf{N}(\delta - m, V_a/N)$, where $V_a = \pi_1^a(1 - \pi_1^a)/r_1 + \pi_2^a(1 - \pi_2^a)/r_2$. The sample size for a two-sided test at level α (one-sided test at level $\alpha/2$), power $1 - \beta$, is

$$N = \left(z_\alpha \sqrt{V}_n + z_\beta \sqrt{V_a}\right)^2 / (\delta - m)^2$$

It remains to specify the variance V_n, using the form

$$V_n = \widetilde{\pi}_1^a(1 - \widetilde{\pi}_1^a)/r_1 + \widetilde{\pi}_2^a(1 - \widetilde{\pi}_2^a)/r_2$$

where $\widetilde{\pi}_1^a$ and $\widetilde{\pi}_2^a$ are the values π_1^a and π_2^a modified so that $\pi_2^a - \pi_1^a = m$. They may be computed in several ways (Farrington and Manning 1990):

- Score test (distant): $\widetilde{\pi}_1^a$ and $\widetilde{\pi}_2^a$ are maximum likelihood estimates of π_1 and π_2 constrained to $\delta = m$.

- Score test with local approximation: like the score test, but V_a is set to equal V_n. Unlike in the case of a superiority trial, this approximation is not a simpler calculation than the more appropriate distant calculation, so it should not be used.

- Wald test: $\widetilde{\pi}_1^a = \pi_1^a$ and $\widetilde{\pi}_2^a = \pi_2^a$; equivalently, $V_n = V_a$.

- Score test variant: $\widetilde{\pi}_1^a$ and $\widetilde{\pi}_2^a$ are estimates of π_1^a and π_2^a constrained to $\delta = m$ and $r_1\widetilde{\pi}_1^a + r_2\widetilde{\pi}_2^a = r_1\pi_1^a + r_2\pi_2^a$. These constraints amount to fixing the margins, like the conditional test; however, the score test variant is not a conditional method, because it is based on the risk difference, whereas the conditional test is based on the odds ratio.

  The score test variant is available (but not recommended) by setting the null variance method using the undocumented option `nvmethod(2)`, where `nvmethod(1)` corresponds to the Wald test and `nvmethod(3)` corresponds to the score test. The `nvmethod()` option was used more widely in earlier versions of `artbin`.

Continuity correction

The continuity-corrected sample size is estimated by computing the unadjusted sample size in each group and then inflating these by the factor

$$\frac{1}{4}\left(1 + \sqrt{1 + \frac{2c}{N_{\text{un}}}}\right)^2$$

where N_{un} is the total unadjusted sample size and $c = 1/(r_1 r_2 |\delta - m|)$ (Fleiss, Tytun, and Ury 1980).

The continuity-corrected power is estimated by deflating the given sample size N_{adj} by a factor of

$$1 - \frac{c}{N_{\text{adj}}}\left(1 - \frac{c}{4N_{\text{adj}}}\right)$$

and then using the standard method on the deflated sample size.