THE STATA JOURNAL

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data

Edmond S.-W. Ng
Department of Health Services Research and Policy
London School of Hygiene and Tropical Medicine
London, UK
edmondngsw@googlemail.com

Richard Grieve
Department of Health Services Research and Policy
London School of Hygiene and Tropical Medicine
London, UK

James R. Carpenter
Department of Medical Statistics
London School of Hygiene and Tropical Medicine
London, UK

Abstract. This article describes a new Stata command, `tsb`, for performing a stratified two-stage nonparametric bootstrap resampling procedure for clustered data. Estimates for uncertainty around the point estimate, such as standard error and confidence intervals, are derived from the resultant bootstrap samples. A shrinkage estimator proposed for correcting possible overestimation due to second-stage sampling is implemented as default. Although this command is written with cost effectiveness analyses alongside cluster trials in mind, it is applicable to the analysis of continuous endpoints in cluster trials more generally. The use of this command is exemplified with a case study of a cost effectiveness analysis undertaken alongside a cluster randomized trial. We also report bootstrap confidence interval coverage by using data from a published simulation study.

Keywords: st0288, tsb, tsbceprob, two-stage nonparametric bootstrap, shrinkage correction, clustered data, cost effectiveness, health economics

1 Introduction

The bootstrap method can be used for estimating uncertainty around the point estimate of a statistic of interest (Efron and Tibshirani 1993; Davison and Hinkley 1997). It provides an alternative to statistical methods that rely on normality when such an assumption is implausible and transformation of the original data (to approximate normality) is either problematic or undesirable. A prime example is cost data, which tend to be right skewed. Here models that assume normality can provide inefficient estimates of the mean cost. One approach is to transform costs (for example, by log transformation), but simply back transforming the resultant estimates does not provide

the estimate of interest, the effect of treatment on the arithmetic mean cost (Manning 1998; Briggs et al. 2005; Faddy, Graves, and Pettitt 2009). Nonparametric bootstrap methods are attractive in this context because they avoid making distributional assumptions.

Many health economic evaluations are undertaken together with cluster randomized trials (CRTs), for example, because the intervention (for instance, vaccination programs) is delivered at the level of the group or cluster (Sullivan et al. 2005; Colvin et al. 2006; Wolters et al. 2006; Bachmann et al. 2007; Gomes et al. 2012a). Here the unit of randomization is the cluster (for example, school or hospital) rather than the individual (pupil or patient). Clusters are randomized to one of a number of alternative interventions; individuals within the same cluster all receive the same intervention. Such study design also helps in minimizing the chance of treatment contamination and helps in those cases where individual randomization may not be feasible or may be considered unethical (Donner and Klar 2000). However, the potentially dependent nature of the CRT data may violate the independently and identically distributed assumption on which many standard statistical methods, including the bootstrap, rely (Flynn and Peters 2004; Nixon, Wonderling, and Grieve 2010).

Davison and Hinkley (1997) proposed an extension to the standard bootstrap procedure that recognizes clustering by resampling clusters and then individuals within clusters. We shall refer to this algorithm as the two-stage bootstrap (TSB). This bootstrap algorithm naturally extends to allowing for any correlation between two or more endpoints (such as costs and health outcomes). While this TSB has been applied previously (Bachmann et al. 2007; Flynn and Peters 2005), there are no routinely available commands for implementing the routine. Further, previous studies do not appear to have followed Davison and Hinkley's (1997) original suggestion of using a shrinkage correction for correcting the potential overestimation of variance due to resampling at the second stage, and the original algorithm was only proposed for studies with equal numbers per cluster (balanced designs). The aim of this article is to provide Stata commands for implementing the TSB and to extend the original algorithm to CRTs with unequal numbers per cluster. The commands are of central relevance to cost effectiveness analyses together with CRTs, but the flexibility of the package allows it to extend more generally to CRTs with other continuous endpoints.

Section 2 gives an overview of the extended bootstrap resampling for clustered data with and without a shrinkage correction. Section 3 describes the new suite of commands and their options. Section 4 illustrates the TSB commands by applying them to a cost effectiveness analysis alongside a CRT. Section 5 reports bootstrap confidence interval coverage of TSB by using data from a previously published simulation study. Finally, we finish with a discussion in section 6.

2   TSB resampling and shrinkage correction

2.1   The standard nonparametric bootstrap

In brief, the standard nonparametric bootstrap approach assumes the observed data are a sample, s, drawn from a population with distribution F. A statistic of interest, R, estimated from the observed sample, is given by $\widehat{R}$. Bootstrap samples are generated by resampling with replacement, from the observed sample, B times, resulting in B bootstrap samples. For each resample, the statistic of interest is calculated and denoted by $\widehat{R}_b^*$ for $b = 1$ to B.[1] The bootstrap estimates $\widehat{R}(s)_1^*, \ldots, \widehat{R}(s)_B^*$ provide an empirical distribution of the statistic of interest that can be used to approximate the sampling distribution of the statistic. Measures of uncertainty around the point estimate, such as standard error and confidence intervals (CIs), are then constructed from this empirical distribution. See Efron and Tibshirani (1993) for a comprehensive discussion of bootstrap methods.

For data collected from clusters such as schools, workplaces, and general practices, the independently and identically distributed assumption required for many standard statistical methods, including the bootstrap, may not hold (Liu 1988). The dependent nature of the data can be accounted for by extending the resampling strategy to mimic the way in which the data are sampled from the population so as to preserve the structure of the original data. We now consider alternative ways of modifying the bootstrap routine to address this.

2.2   The "cluster bootstrap"

One way that the bootstrap routine can be modified to recognize the clustering is to resample clusters rather than individuals (Davison and Hinkley 1997). The routine is otherwise as for the standard bootstrap and is readily available to Stata users through the `bootstrap` command with `cluster()` option. However, the cluster bootstrap approach has been found to perform poorly with CI coverage levels below the nominal level (Flynn and Peters 2005).

2.3   The TSB for clustered and correlated data, with particular application to cost effectiveness analysis

A different strategy proposed by Davison and Hinkley (1997) requires the resampling to be performed in two stages (both with replacement). Here clusters are resampled in the first stage, and individuals within the chosen clusters are resampled in the second stage (with replacement in both stages) to obtain a bootstrap sample. The statistic of interest is calculated for the bootstrap sample. This process is repeated B times to form the empirical distribution of the statistic of interest. In the TSB, any correlation

---

1. General recommendations suggest that at least 1,000 replicates are required for bootstrap confidence intervals (Davison and Hinkley 1997; Campbell and Torgerson 1999; Nixon, Wonderling, and Grieve 2010).

between the endpoints can be recognized by resampling jointly those variables that are required for calculating the statistic of interest.

In health economic evaluations, a common statistic of interest is the incremental net monetary benefit (INB), which reports the relative value for money of alternative health care programs (Stinnett and Mullahy 1998). The INB is calculated by estimating the difference between the treatment alternatives in the mean health outcomes, valuing this by the threshold willingness to pay (WTP) for a unit of health gain, and subtracting from this the incremental costs. Methods guidance recommends that measures of uncertainty for statistics such as the INB should recognize the potential correlation between the endpoints, cost and health outcomes (Willan and Briggs 2006).[2]

The statistical uncertainty in the estimated cost effectiveness can be reported by estimating 95% CIs around the INB. Hence, under the TSB approach, the INB can be calculated in each bootstrap replicate, and the 95% CIs estimated from the resultant empirical distribution. Another recommended metric for summarizing the statistical uncertainty surrounding the cost-effectiveness (CE) measure is the cost-effectiveness acceptability curve (CEAC) (Van Hout et al. 1994). The CEAC presents the probability that an intervention is cost effective, given the data, at alternative threshold levels of WTP for a unit of health outcome, λ. See the description section of the new command `tsbceprob` for further details of the estimation of CEACs with TSB.

Either measure can allow for the correlation between costs and health outcomes within the TSB routine by resampling pairs of cost, and effect can be taken from individuals at the second stage. In the new command `tsb`, we implement joint resampling of endpoint variables from the original data. Where stratified sampling is appropriate (for example, data from a CRT comparing two interventions), resampling is performed independently within each stratum.

2.4   TSB with shrinkage correction

Davison and Hinkley (1997) noted that unless the number of clusters and individuals per cluster are both large, this method may overestimate the variance due to resampling at the second stage. Resampling at the second stage is likely to double count the within-cluster variance because the estimates of the cluster means resampled from stage 1 already incorporate both within- and between-cluster variability (Davison and Hinkley 1997; Flynn and Peters 2004; Gomes et al. 2012b). Davison and Hinkley (1997) therefore proposed a shrinkage correction to avoid overestimating the variance. This approach differs from that described above in that rather than resampling clusters and then individuals within the chosen clusters, the algorithm resamples estimates for two distributions directly, namely, Fx for cluster means and Fz for individual deviations or residuals from the cluster means, denoted as z_{ij}, for the ith individual in cluster j.

---

2. INB $= \lambda \times \Delta e - \Delta c$, where Δc and Δe are the incremental cost and health outcome ($\Delta e = \overline{e}_1 - \overline{e}_0$ and $\Delta c = \overline{c}_1 - \overline{c}_0$ with the intervention arm denoted by subscript 1 and 0 for control), and λ the maximum WTP per unit of health outcomes. The variance of the INB is given by $V(\text{INB}) = \lambda^2 V(\Delta e) + V(\Delta c) - 2\lambda \text{cov}(\Delta e, \Delta c)$.

Details of this approach are elaborated in the algorithm below. It includes modifications to the original proposal to allow for unbalanced cluster sizes, which makes the procedure more applicable to real data.

2.5   Algorithm for TSB with shrinkage correction:

1. For $k = 1$ ($k = 1, \ldots, K$ intervention arms; for simplicity sake, subscript k is omitted in the following steps).

2. Calculate shrunken cluster means, $\widehat{x}_j = c\overline{y}_{..} + (1-c)\overline{y}_j$, for $j = 1, \ldots, N_c$, where N_c is the number of clusters in the stratum. See footnote[3] for derivation of the constant c.

3. Calculate the standardized individual-level residuals (from the estimated cluster means), $\widehat{z}_{ij} = (y_{ij} - \overline{y}_j)/\sqrt{(1 - n_*^{-1})}$, where n_* is the cluster size. For unbalanced cluster sizes, n_* is replaced by different measures of "average" cluster size in tsb (see options for tsb).

4. Randomly sample (with replacement) from $\widehat{x}_1, \ldots, \widehat{x}_{N_c}$ to obtain $\widehat{x}_{j'}^*$ for $j' = 1, \ldots, N_c$. j' is a new index for the chosen clusters.

5. Randomly sample (with replacement) from $\widehat{z}_{11}, \ldots, \widehat{z}_{n_{N_c} N_c}$ to obtain $\widehat{z}_{i'j'}^*$ for $i' = 1, \ldots, n_{j'}$, where $n_{j'}$ is the size of the jth cluster chosen in step 4. The index $n_{N_c} N_c$ denotes the last individual in the last cluster in the original data. The number of units to be sampled in this step is dependent on the sum of the sizes of the chosen clusters in step 4. Note also that the cluster membership of the original sample is ignored here.

6. Reconstruct the sample by creating a "synthetic" sample by $y_{ij}^* = \widehat{x}_{j'}^* + \widehat{z}_{i'j'}^*$.

7. Repeat steps 2 to 6 for the remaining $K - 1$ intervention arms; then stack the K synthetic samples to form the bth bootstrap sample.

8. Calculate the statistic of interest R for the bth bootstrap sample to obtain $\widehat{R}_b(y_{ij}^*)$.

9. Repeat steps 1 to 8 B times to obtain the bootstrap estimates, $\widehat{R}_b(y_{ij}^*)$, for $b = 1, \ldots, B$.

---

3. The constant c is given by $(1-c)^2 = (N_{ck}/N_{ck-1}) - SS_w/\{n_j(n_j-1)SS_b\}$ or set to 1 if the right-hand side of the expression is negative. $SS_w = \sum_{j=1}^{N_c} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2$, $SS_b = \sum_{j=1}^{N_c} (\overline{y}_j - \overline{y}_{..})^2$, and the grand mean is given by $\overline{y}_{..} = (\sum_{j=1}^{N_c} n_j)^{-1} \sum_{j=1}^{N_c} \sum_{i=1}^{n_j} y_{ij}$.

The distribution of the $\widehat{R}_b(y_{ij}^*)$'s approximates the empirical distribution of R and can be used for estimating the bootstrap standard error and confidence intervals. Where two or more variables from the original data are required to calculate the statistic of interest, the above resampling is performed jointly for the variables involved. For health economic evaluations, individual costs and effects are resampled jointly: $\widehat{x}_j$ and $\widehat{z}_{ij}$ become matrices of dimensions $(1 \times m)$ and $(n_j \times m)$, where m is the number of variables or endpoints (needed to calculate the statistic of interest), and n_j is the number of individuals in cluster j. For example, $m = 2$ for individual costs and health outcomes for calculating the INB. The algorithm above assumes that stratification between the treatment arms is required, but if that is not the case, then steps 1 and 7 should be omitted.

The original proposal assumed constant cluster size, n_*, across clusters. In our implementation of TSB with shrinkage correction, it has been generalized to allow for unequal cluster sizes by replacing n_* by an "average" cluster size (within the stratum) in step 3 and by acknowledging the variable stratum sizes in step 5, where the second stage resampling takes place. n_* is replaced by three different measures of "average" cluster size in `tsb`: `dk` (see page 9 of Donner and Klar [2000]), `median`, and `mean`. See the next section on the `tsb` command for explanation.

3 Commands

3.1 tsb

Syntax

`tsb` *varlist* $\big[\, if \,\big]$ $\big[\, in \,\big]$, `stats(&f())` `cluster(`*varname*`)` $\big[$ `strata(`*varname*`)`
    `reps(`#`)` `seed(`#`)` `unbal(`*string*`)` `lambda(`*real*`)` `noshrink` `level(`#`)` `nodots` $\big]$

Description

`tsb` performs a two-stage bootstrap sampling procedure on a user-supplied statistic for clustered data. It is implemented as an ado-file, which serves as a wrapper function for invoking a number of Mata functions, to perform TSB with or without shrinkage correction. The uncertainty around the point estimate calculated from the original sample is quantified by the bootstrap standard error and confidence intervals. These estimates are given by applying the Stata programmer's command `bstat` on the bootstrap replicates of the statistic of interest. A detailed discussion of the derivation of bootstrap confidential intervals is beyond the remit of this article. Readers are referred to, for example, Carpenter and Bithell (2000) for a comprehensive review.

Among the nonparametric confidence intervals reported by `bstat` and `bootstrap`, the bias-corrected and accelerated (BCa) confidence interval requires the estimation of an acceleration parameter that adjusts for the skewness in the sampling distribution

(Briggs, Wonderling, and Mooney 1997; Carpenter and Bithell 2000). In our Stata implementation of TSB, the acceleration parameter is calculated by using a stand-alone Mata function before the result is passed to `bstat` for constructing the BCa confidence interval.[4]

Options

`stats(&f())` specifies a user-supplied Mata function, $f()$, for calculating the statistic for bootstrapping. See appendix for examples of such functions. Note that the ampersand (`&`) before $f()$ is required as part of the syntax. `stats()` is required.

`cluster(`*varname*`)` specifies the variable that identifies clusters. `cluster()` is required.

`strata(`*varname*`)` specifies the variable that identifies strata. For an example of a cluster randomized trial, the strata would be levels of a cluster-level treatment or intervention variable. The default is `strata(constant)`.

`reps(`*#*`)` specifies the number of bootstrap replications to be performed. For estimation of the confidence interval, 1,000 or more replications are generally recommended (Efron and Tibshirani 1993; Davison and Hinkley 1997). The default is `reps(1000)`.

`seed(`*#*`)` specifies the random number seed.

`unbal(`*string*`)` specifies the "average" cluster size, n_*. *string* can be `dk` (page 9 of Donner and Klar [2000]), `median`, or `mean`. The default is `unbal(dk)`. These are calculated independently for each stratum specified in strata. For the option `unbal(dk)`, $n_* = \overline{n}_. - \sum_{j=1}^{N_c}(n_j - \overline{n}_.)^2/(N_c - 1)M$, where $\overline{n}_. = M/N_c$ (arithmetic mean cluster size) and $M = \sum_{j=1}^{N_c} n_j$ (total number of individuals in the N_c clusters).

`lambda(`*real*`)` is relevant for cost-effectiveness analysis (CEA) and is the threshold WTP for a unit of health outcome; the user specifies an optional value, *real*, that can be called from within the user-supplied function $f()$, if required.

`noshrink` specifies that the two-stage bootstrap resampling is performed without shrinkage correction. If this option is chosen, instead of cluster means, whole clusters are resampled with replacement in stage 1. In stage 2, individuals within the chosen clusters are then resampled also with replacement. Cluster membership in the original data is respected in this case.

---

4. Our early attempts of implementing TSB through `bootstrap` with the `bca` option suggested that second-stage resampling caused instability in the estimation and on occasions caused Stata to shut down. The instability was due partly to the jackknife sampling procedure used for estimating the acceleration parameter in `bootstrap` (Stata Technical Support 2010). As a result, we decided to write our own computer code for implementing TSB.

level(#) specifies the confidence level, as a percentage, for confidence intervals. The
   default is level(95) or as set by set level.

nodots suppresses display of the replication dots. One dot character is displayed for
   each successful replication.

Saved results

In addition to the standard output given by bstat shown in the Results window, a
number of useful results are stored in Stata matrices e() and r() postestimation. The
statistic of interest calculated from the original sample is stored in e(b). The bootstrap
replicates of the statistic of interest are stored in r(tsb_sam) (with its mean stored in
e(b_bs)).

3.2   tsbceprob

Syntax

tsbceprob *varlist* $\big[$ *if* $\big]$ $\big[$ *in* $\big]$, stats(&nb()) cluster(*varname*)
   strata(*varname*) lambda(*real*) $\big[$reps(#) seed(#) unbal(*string*) noshrink
   level(#) nodots$\big]$

Description

tsbceprob is designed for calculating CE probabilities and has to be used in conjunction
with the Mata function nb(). nb() calculates the net monetary benefit (NMB) for each
level of the cluster-level interventions (or comparators) defined by the strata variable.
NMB_k is calculated as $\lambda \overline{e}_k - \overline{c}_k$, where $\overline{e}_k$ is the arithmetic mean of health outcomes
(second variable in *varlist*), $\overline{c}_k$ the mean of costs (first variable in *varlist*) for the kth
comparator, and λ is the WTP threshold.

For a two-way comparison, the intervention is defined as the most cost effective
if it has a positive INB versus the comparator. For an n-way comparison, the most
cost-effective alternative is that with the highest NMB, where the NMB is calculated
for each comparator by valuing the absolute level of health outcomes by λ. CEACs
can be calculated by estimating the NMB in each bootstrap replicate and reporting
the probability that each intervention is the most cost effective as the proportion of
replicates in which each intervention has a positive INB versus the comparator (two-way
comparison) or the highest NMB (n-way comparison).

In tsbceprob, when exactly the same NMB value is calculated for two or more
comparators in the same replicate (that is, resulting in ties), these comparators are
considered equally cost effective. For example, in a replicate where the first of three
comparators yielded the highest NMB, this would result in a row vector of (1, 0, 0) for
the replicate; if the second and third comparators yielded the same highest NMB, the

indicator vector would become (0, 0.5, 0.5). The same principle applies for two- or n-way comparison. The CE probabilities are then estimated by the column means of a matrix consisting of the row vectors for all replicates.

Options

The options are the same as those for `tsb` except that `strata()` and `lambda()` have to be specified and that `tsbceprob` is designed to be used with the NMB function `nb()` only.

Output

`tsbceprob` returns estimates of the CE probabilities at the given WTP value in the matrix `r(tsb_ceprob)`. The content in `r(tsb_ceprob)` is shown in the Results window. `r(tsb_ceprob)` is of dimension $\{1 \times (K+1)\}$, where elements in the first K columns are the CE probabilities for the K comparators (defined in the strata variable), and the last element is the corresponding λ WTP value used for the estimation. These probabilities can then be used to plot the CEACs against a range of threshold values, λ. An example based on data of a published CRT is given in the next section. See appendix for examples of some user-supplied functions in Mata.

4   Illustrative examples

In the following examples, we use cost-effectiveness data from a published CRT for evaluating alternative interventions for preventing postnatal depression. The PoNDER (psychological interventions for postnatal depression-randomized controlled trial and economic evaluation) study is a UK health technology assessment of alternative interventions compared with usual care for preventing postnatal depression (Morrell et al. 2009). The participating general practices were randomized to usual care (control arm) or one of two interventions, a person-centered approach (PCA) or cognitive-behavioral approach (CBA). We use cost and health-related, quality-of-life data reported at six months for 1,732 patients (70 general practices) with complete information for our illustration. Health effects are calculated by the change (six months from baseline) in the health-related, quality-of-life measure.

4.1   Example 1: Mean cost (mept())

In our first example, we grouped the two interventions (PCA and CBA) into a single intervention group. Individual-level health care costs are typically highly right skewed, as in the case for PoNDER (figure 1).
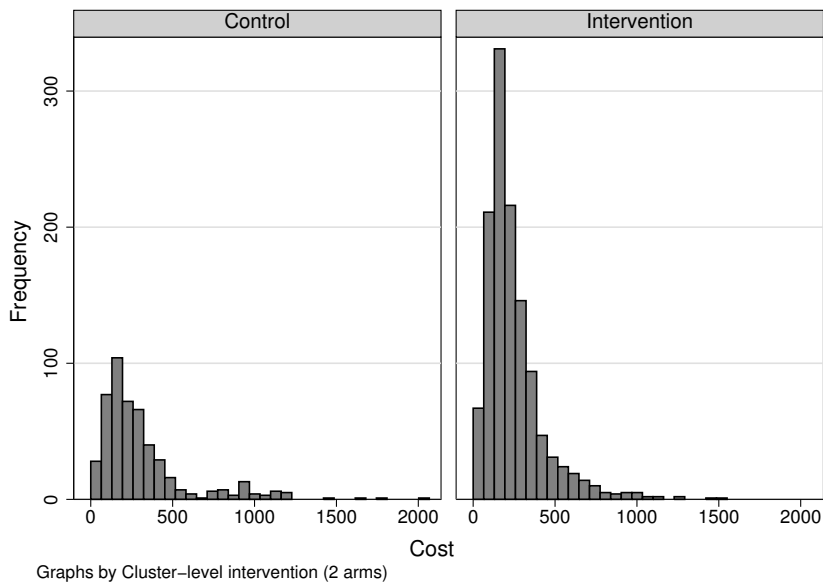
Figure 1. Distributions of individual-level costs (in British pounds) by intervention (PoNDER)

TSB can be used here for estimating the uncertainties around the two point estimates for mean costs for the control and intervention arms. Here we show how the bootstrap standard error and different confidence intervals can be estimated by using `tsb` in conjunction with the user-supplied Mata function `mept()`. `mept()` calculates the mean of the single variable specified in *varlist* for `tsb`. The variables used in the example are `cost` for individual-level costs (measured in British pounds, GBP), `int1` for cluster-level treatment (0 for control, 1 for intervention), and `cluster` for cluster identifier.
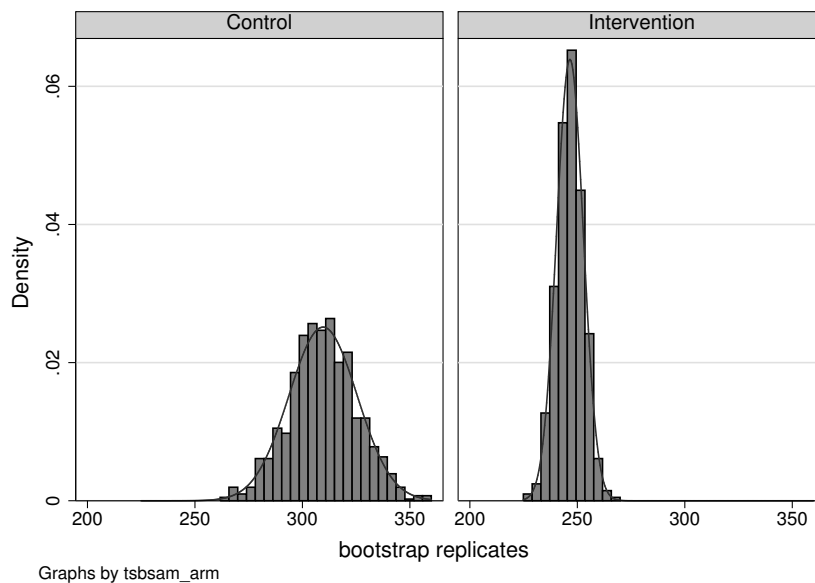
Figure 2. Distributions of bootstrap replicates of mean costs (in British pounds) by intervention arm (PoNDER)

Note: Histograms overlaid with corresponding normal densities.

Despite the high level of skewness in the observed costs, the distributions of the bootstrap replicates of mean costs appear symmetric in both arms (figure 2). As a result, the differences in the limits of the normal approximation and percentile-based confidence intervals are small. For the control arm, the 95% normal CIs are estimated to be 278.8 to 341.0 GBP and 279.9 to 341.8 GBP for the BCa CIs. The TSB sampling took less than 10 seconds to complete for both arms (see `Elapsed time` in the output).[5]

---

5. All examples were performed with Stata/IC 11.2 for Windows (32-bit) on a Dell PC with Xeon(R) 2.93 GHz CPU and 12 GB RAM.

```
. tsb cost if int1==0, stats(&mept()) cluster(cluster) seed(101)

*** User-supplied settings ***
Cluster variable: cluster
Statistic (function): &mept()
Strata variable:  not supplied (assumed constant)

Two-stage bootstrap WITH shrinkage estimator
..............................................        50
..............................................       100

  (output omitted)

..............................................      1000

Elapsed time (mins) = .133
```

Bootstrap results   Number of obs   =   495
                                           Replications   =   1000

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf. Interval]	
bsam	309.8899	−.2805662	15.871883	278.7816	340.9982	(N)
				278.7852	340.61	(P)
				279.0978	341.143	(BC)
				279.9131	341.812	(BCa)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval

Mean of TSB sample of statistic of interest =
  309.60933

. tsb cost if int1==1, stats(&mept()) cluster(cluster) seed(101) nodots

*** User-supplied settings ***
Cluster variable: cluster
Statistic (function): &mept()
Strata variable:  not supplied (assumed constant)

Two-stage bootstrap WITH shrinkage estimator

Elapsed time (mins) =    .15
```

Bootstrap results   Number of obs   =   1237
                                           Replications   =   1000

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf. Interval]	
bsam	246.4919	.0702566	6.2435131	234.2548	258.729	(N)
				234.5496	258.2897	(P)
				234.4035	258.169	(BC)
				234.5774	258.3279	(BCa)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval

Mean of TSB sample of statistic of interest =
  246.56216
```

4.2 Example 2: Incremental net benefit (inb())

At a societal WTP threshold value of 20,000 GBP per unit of effect, the INB calculated from the observed sample was 98.4 GBP (see `Observed Coef.` in the output). The uncertainty around this point estimate is quantified by the bootstrap standard error and confidential intervals. Where the symmetry of the sampling distribution of the statistic of interest is questionable because of, for example, skewed data, the normal-approximation-based confidence interval may be inappropriate, as indicated in Campbell and Torgerson (1999). Gomes et al. (2012b) showed that the BCa confidence intervals based on TSB with shrinkage correction provide good confidence interval coverage (close to nominal level) over a range of challenging data scenarios, including few clusters, unbalanced cluster sizes, and skewed costs, in their simulation study. Here the BCa confidence interval suggests there is a 95% chance that the true INB lies between 33.6 and 173.2 GBP. The TSB sampling estimation took 34 seconds to complete.

```
. tsb cost qalygain, stats(&inb()) cluster(cluster) strata(int1) seed(101)
> lambda(20000)

*** User-supplied settings ***
Cluster variable: cluster
Statistic (function): &inb()
Strata variable: int1
Lambda = 20000

Two-stage bootstrap WITH shrinkage estimator
...............................................          50

  (output omitted)

...............................................          1000

Elapsed time (mins) =  .567

Bootstrap results                           Number of obs      =      1732
                                            Replications       =      1000
```

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf. Interval]	
bsam	98.395852	-1.580874	35.236647	29.33329	167.4584	(N)
				28.34896	166.2569	(P)
				32.92921	172.3587	(BC)
				33.58042	173.1515	(BCa)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval

Mean of TSB sample of statistic of interest =
  96.814978
```

4.3   Example 3: Cost-effectiveness probabilities (tsbceprob()) and CEACs

In example 3, we use `int2` as the cluster-level treatment variable. The interventions defined in this variable are control, PCA, and CBA. The following syntax shows how `tsbceprob` could be used to estimate CE probabilities by embedding it in a `foreach` loop. The resultant CE probabilities estimated for a range of WTP values from 0 to 60,000 GBP (in steps of 5,000) are stored in the Stata matrix `ceprob_mat`. The context of the matrix was then exported into the current Stata dataset by `svmat` for plotting the CEACs (figure 3). Here the treatment variable is `int2` with `0` for control, `1` for CBA, and `2` for PCA. A seed value is used for reproducible results.

```
. capture matrix drop ceprob_mat

. foreach num of numlist 0(5000)60000 {
  2. tsbceprob cost qalygain, stats(&nb()) cluster(cluster) strata(int2)
> reps(1000) unbal(dk) nodots seed(101) lambda(`num´)
  3. matrix ceprob_mat = (nullmat(ceprob_mat)\r(tsb_ceprob))
  4. }

*** User-supplied settings ***
Cluster variable: cluster
Statistic (function): &nb()
Strata variable: int2
Average cluster size: dk
Lambda = 0

Two-stage bootstrap WITH shrinkage estimator

Elapsed time (mins) =   .533

Cost-effective probabilities and WTP value
    0  .92  .08    0


*** User-supplied settings ***
Cluster variable: cluster
Statistic (function): &nb()
Strata variable: int2
Average cluster size: dk
Lambda = 5000

Two-stage bootstrap WITH shrinkage estimator

Elapsed time (mins) =   .533

Cost-effective probabilities and WTP value
    0  .917  .083  5000
  (output omitted )
```

```
. matrix list ceprob_mat
ceprob_mat[13,4]
        c1      c2      c3      c4
r1       0     .92     .08       0
r1       0    .917    .083    5000
r1       0    .872    .128   10000
r1       0    .817    .183   15000
r1       0    .783    .217   20000
r1    .002    .752    .246   25000
r1    .004    .734    .262   30000
r1    .008    .718    .274   35000
r1    .012    .704    .284   40000
r1    .013    .695    .292   45000
r1    .014     .68    .306   50000
r1    .016    .672    .312   55000
r1    .018    .662     .32   60000
. svmat ceprob_mat, names(tsb_ceprob)
. rename tsb_ceprob1 tx_control
. rename tsb_ceprob2 tx_1
. rename tsb_ceprob3 tx_2
. rename tsb_ceprob4 lval
. label variable tx_control "Control"
. label variable tx_1 "CBA"
. label variable tx_2 "PCA"
. label variable lval "Willingness-to-pay threshold ()"
. scatter tx_control tx_1 tx_2 lval, connect(l l l)
> msize(small small small)
> ytitle("Probability cost effective")
> yscale(range(0 1)) ylabel(0(0.2)1)
> xlabel(0(10000)60000)
```

CBA is shown to be the most cost-effective intervention over the entire range of WTP considered according to the CEACs in figure 3. This is followed by PCA and control.
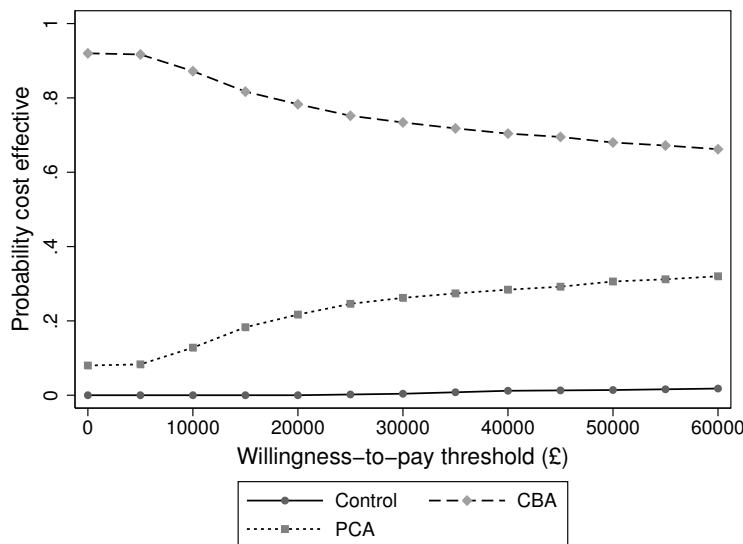
Figure 3. Cost-effectiveness acceptability curves at 6 months (PoNDER)

Note: y axis shows the proportions of bootstrap samples with highest net benefit value of the corresponding intervention.

5   Simulation study

A recent extensive simulation study reported that TSB with shrinkage correction performed as well as mixed models and outperformed robust methods such as seeming unrelated regression and general estimating equation, both with robust variance estimators (Gomes et al. 2012b). Here we apply the new Stata command `tsb` to some of the scenarios from the simulation study. We report the CI coverage of four different CIs given by `bstat`, together with the mean width of the CIs and their lower and upper tail error rate (table 1).

Table 1. Confidence interval coverage of 95% confidence interval reported by `tsb` with and without shrinkage correction on simulated data (true INB = 1,000 GBP)

Confidence interval		Scenario[I]			
		Base case		Challenging	
		Shrinkage correction			
		Without	With	Without	With
Normal	Coverage[III]	0.981	0.944	0.965	0.966
	Mean width	533.8	427.1	3,217.0	3,280.0
	Lower tail error rate	0.009	0.029	0.016	0.016
	Upper tail error rate	0.011	0.028	0.020	0.019
Percentile	Coverage[III]	0.979	0.942	0.960	0.958
	Mean width	533.6	426.5	3,320.0	3,366.0
	Lower tail error rate	0.010	0.028	0.018	0.017
	Upper tail error rate	0.011	0.031	0.023	0.025
Bias-corrected	Coverage[III]	0.983	0.939	0.959	0.956
	Mean width	533.5	426.7	3,185.0	3,251.0
	Lower tail error rate	0.009	0.029	0.018	0.019
	Upper tail error rate	0.009	0.033	0.024	0.026
Bias-corrected and accelerated	Coverage[III]	0.983	0.939	0.954	0.950
	Mean width	533.5	426.7	3,266.0	3,353.0
	Lower tail error rate	0.009	0.029	0.021	0.022
	Upper tail error rate	0.009	0.033	0.026	0.029

Notes: I. Base case: 20 clusters per arm, 50 individuals per cluster, cluster size imbalance ($cv_{\mathrm{imb}} = 0$), intracluster correlation coefficient for costs = 0.01 (effects 0.01), cost skewness ($cv_{cost} = 0.2$ implies no skewness for Gamma distributed costs), individual- (cluster-) level correlation of costs and effects = 0.2 (0); challenging: same as base case but with the following differences: 3 clusters per arm, $cv_{imb} = 1$, and $cv_{cost} = 3$.
II. All seed values were set at 101 and with 1,000 bootstrap replications throughout. Two thousand datasets were simulated for each scenario.
III. The coverage probabilities are each based on 2,000 replicate samples, implying a typical confidence interval width for the coverage probabilities, of $2 \times 1.96 \times \sqrt{0.95 \times (1 - 0.95)/2000} = 0.0191$ (approximately 1.9 percentage points).

In brief, cost and health outcome data were generated from a CRT design assumed to have two randomized arms (intervention and control). Gomes et al. (2012b) simulated the effect of the intervention on mean costs and health outcomes according to a bivariate linear additive model. Each of the two simultaneous equations for costs and health outcomes included a cluster-level mean, a cluster-level incremental effect (for the intervention arm), and an individual deviation from the cluster mean. Individual- and cluster-level costs and health outcomes were allowed to be correlated and, for the base case, followed a bivariate normal distribution. The base case also assumed a balanced design with 50 individuals within each cluster and 20 clusters within each treatment arm.

In a more challenging scenario, Gomes et al. (2012b) assumed that there were only three clusters per arm and that the size of the clusters followed a Gamma distribution with a mean and a coefficient of variation (cv_{imb}) of 1, which resulted in imbalanced cluster sizes. cv_{imb} is obtained by dividing the standard deviation of cluster size by its mean. Individual costs were also assumed to follow a Gamma distribution with varying level of skewness as defined by a cv_{cost} of 3. All simulations used the same true value of 1,000 GBP for the metric of interest, the INB. Two thousand datasets were simulated for each scenario (for more details, see Gomes et al. [2012b]).

Here we find that under both these scenarios, the CIs constructed by using TSB with shrinkage correction all gave CI coverages close to the nominal rate of 0.95. As was anticipated when the TSB was applied without the shrinkage correction, the CIs were too wide, and the CI coverage exceeded nominal levels. Under the more challenging scenario with few clusters, variable cluster sizes, and skewed costs, the bias-corrected and accelerated CI, after application of the shrinkage correction, yielded the coverage rate closest to the nominal.

6   Discussion

In this article, we have provided new Stata commands for implementing and extending the TSB algorithm proposed by Davison and Hinkley (1997). Unlike their original algorithm, our implementation can be applied to the common setting of a CRT with unequal numbers per cluster. We envisage that the suite of commands will be particularly useful in a CEA alongside a cluster trial. In this setting, statistical methods are required to allow for the clustered nature of the data, recognizing that costs and health outcomes may be correlated and that cost data are highly skewed. In general, though, such CEAs fail to address these issues; indeed, a recent review found that around 95% of published CEAs alongside CRTs used inappropriate methods (Gomes et al. 2012a). The TSB tools we provide can help analysts address the challenges of handling clustered data with highly skewed costs that are correlated with health outcomes.

The two-stage nonparametric bootstrap method with shrinkage correction was reported to perform favorably in an extensive simulation study designed to compare the appropriateness of a number of commonly used statistical methods for CEA of CRTs (Gomes et al. 2012b). However, this procedure has not been widely applied in practice among health economists mainly because of a lack of implementation in mainstream and user-friendly software. We hope that making this method available through Stata will help translate research findings into practice in the community.

The TSB approach relies on asymptotic assumptions that may be invalid for small samples (in particular, few cluster-level units for clustered data). Although Gomes et al. (2012b) showed that TSB with shrinkage correction gives good confidence interval coverage even with as few as three clusters per arm, such a result was based on a simulation study with a known data-generating mechanism. For real data with few clusters, the data would have been generated by numerous other factors that would not have been captured by any given data-generating mechanism. Hence, analysts should exercise caution when interpreting the results based on small samples.

Our implementation of TSB generalized Davison and Hinkley's (1997) original proposal by allowing unbalanced cluster sizes in the original data. This should help make our implementation applicable to more realistic data settings where completely balanced clusters may be rare. Finally, although our illustrative examples focus on health economic evaluations, the TSB method is applicable for cluster study design more generally. It should be noted, though, that the application only applied to circumstances with a cluster design with two levels to the data hierarchy. It does not extend to dependencies that may arise in multicenter randomized controlled trials where within a center, individuals are randomized to alternative treatments. Nor does the application extend to a CRT with three or more levels to the data hierarchy (for example, when repeated measures are nested within patients and within clusters).

7 Acknowledgments

8 Funding

9 References

Bachmann, M. O., L. Fairall, A. Clark, and M. Mugford. 2007. Methods for analyzing cost effectiveness data from cluster randomized trials. *Cost Effectiveness and Resource Allocation* 5: 12.

Briggs, A., R. Nixon, S. Dixon, and S. Thompson. 2005. Parametric modelling of cost data: Some simulation evidence. *Health Economics* 14: 421–428.

Briggs, A. H., D. E. Wonderling, and C. Z. Mooney. 1997. Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health Economics* 6: 327–340.

Campbell, M. K., and D. J. Torgerson. 1999. Bootstrapping: Estimating confidence intervals for cost-effectiveness ratios. *QJM: An International Journal of Medicine* 92: 177–182.

Carpenter, J., and J. Bithell. 2000. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19: 1141–1164.

Colvin, M., M. O. Bachmann, R. K. Homan, D. Nsibande, N. M. Nkwanyana, C. Connolly, and E. B. Reuben. 2006. Effectiveness and cost effectiveness of syndromic sexually transmitted infection packages in South African primary care: Cluster randomised trial. *Sexually Transmitted Infections* 82: 290–294.

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

Donner, A., and N. Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Faddy, M., N. Graves, and A. Pettitt. 2009. Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health* 12: 309–314.

Flynn, T. N., and T. J. Peters. 2004. Use of the bootstrap in analysing cost data from cluster randomised trials: Some simulation results. *BMC Health Services Research* 4: 33.

———. 2005. Cluster randomized trials: Another problem for cost-effectiveness ratios. *International Journal of Technology Assessment in Health Care* 21: 403–409.

Gomes, M., R. Grieve, R. Nixon, and W. J. Edmunds. 2012a. Statistical methods for cost-effectiveness analyses that use data from cluster randomized trials: A systematic review and checklist for critical appraisal. *Medical Decision Making* 32: 209–220.

Gomes, M., E. S. Ng, R. Grieve, R. Nixon, J. Carpenter, and S. G. Thompson. 2012b. Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. *Medical Decision Making* 32: 350–361.

Liu, R. Y. 1988. Bootstrap procedures under some non-I.I.D. models. *Annals of Statistics* 16: 1696–1708.

Manning, W. G. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 17: 283–295.

Morrell, C. J., R. Warner, P. Slade, S. Dixon, S. Walters, G. Paley, and T. Brugha. 2009. Psychological interventions for postnatal depression: Cluster randomised trial and economic evaluation. The PoNDER trial. *Health Technology Assessment* 13: 1–176.

Nixon, R. M., D. Wonderling, and R. D. Grieve. 2010. Non-parametric methods for cost-effectiveness analysis: The central limit theorem and the bootstrap compared. *Health Economics* 19: 316–333.

Stata Technical Support. 2010. Personal communication.

Stinnett, A., and J. Mullahy. 1998. Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* 18: S68–S80.

Sullivan, S. D., T. A. Lee, D. K. Blough, J. A. Finkelstein, P. Lozano, T. S. Inui, A. L. Fuhlbrigge, V. J. Carey, E. Wagner, and K. B. Weiss. 2005. A multisite randomized trial of the effects of physician education and organizational change in chronic asthma care: Cost-effectiveness analysis of the Pediatric Asthma Care Patient Outcomes Research Team II (PAC-PORT II). *Archives of Pediatrics and Adolescent Medicine* 159: 428–434.

Van Hout, B. A., M. J. Al, G. S. Gordon, and F. F. H. Rutten. 1994. Costs, effects and C/E-ratios alongside a clinical trial. *Health Economics* 3: 309–319.

Willan, A. R., and A. H. Briggs. 2006. *Statistical Analysis of Cost-Effectiveness Data.* Chichester, UK: Wiley.

Wolters, R., R. Grol, T. Schermer, R. Akkermans, R. Hermens, and M. Wensing. 2006. Improving initial management of lower urinary tract symptoms in primary care: Costs and patient outcomes. *Scandinavian Journal of Urology and Nephrology* 40: 300–306.

About the authors

Edmond Ng was a lecturer in medical statistics at the London School of Hygiene and Tropical Medicine, London, UK. He is now a research statistician of the Clinical Practice Research Datalink group at the Medicines and Healthcare products Regulatory Agency. He has been programming in Mata for over a year and used it for writing the tsb command suite.

Richard Grieve is a reader in health economics at the London School of Hygiene and Tropical Medicine.

James Carpenter is a professor in medical statistics at the London School of Hygiene and Tropical Medicine.

Appendix. Sample Mata functions for use with TSB

The three Mata functions used in our examples are shown below, namely, mept(), inb(), and nb(). These child functions take an input data object, data, from their parent function tsb when they are called by the latter. The data object is of dimension $\{N_k \times (n_{ept} + 2)\}$, where N_k is the total number of observations in stratum k (the subscript k is omitted when resampling is performed without stratification), and n_{ept} is the number of endpoints to be resampled jointly. For CEAs where two endpoints, individual costs and effects, are to be resampled jointly, n_{ept} equals 2. The last two columns in data are reserved for cluster and strata variables (see technical notes below).

Technical notes on TSB

The two rightmost columns of the data object are reserved for the cluster identifier (second column from the right) and strata (rightmost column) variables. For example, `inb()` and `nb()` both assume the rightmost column (fourth column) in data, the strata variable (the object `treat` is used in these functions for the strata variable). Analysts wishing to use `tsb` on other statistics of interest must bear this in mind when writing their own child Mata functions.

The first n_{ept} columns in the data object are the variables (endpoints) specified to be resampled jointly. Where $n_{ept} > 1$, the ordering of the variables in the child Mata function must match those given in *varlist* for `tsb`. For example, the function `inb()` takes the first column of the data object as costs and the second column as health effects. Therefore, these two endpoints must appear in the same order in the `tsb` command (for example, `tsb costs effects, stats(&inb()) ...`).

When the following three sets of syntax are issued in Stata, the corresponding Mata functions will be saved in one's (Stata's) personal directory. If one is unsure where one's personal directory is, it will be displayed when the command `personal` is issued in Stata. Replace "local directory" in the following syntax with one that matches one's local setting so that the Mata functions are stored in the appropriate directory for access.

A.1. mept()—mean endpoint

```
mata:
real scalar function mept(transmorphic matrix data)
{
version 11.2
/* DECLARATIONS */
real matrix cost
real scalar mept
/* Extract data from "data" for calculation */
cost=data[,1]
/* Calculate mean endpoint  */
mept=mean(cost)
return(mept)
}
mata mosave mept(), dir("local directory:\ado\personal") replace
end
```

A.2. inb()—incremental net benefit

```
mata:
real scalar function inb(transmorphic matrix data, |real scalar lambda)
{
version 11.2
/* DECLARATIONS */
real matrix cost, effect, treat
real scalar mc_ctl, mc_tmt, me_ctl, me_tmt
/* DATA CHECKING */
if (cols(data)!=4){
 _error("Function requires an input data matrix of 4 columns wide.")
}
cost=data[,1]
effect=data[,2]
treat=data[,4]
/* SET UNSPECIFIED PARAM VALUES */
if (lambda==.) lambda=20000
/* DATA CHECKING */
/* 1. Stop if treatment var has anything other than 2 levels */
if (rows(uniqrows(treat))<2) {
 _error("Treatment variable has <2 unique values.")
 } else if (rows(uniqrows(treat))>2) {
 _error("Treatment variable has >2 unique values.")
}
/* Calculate INB */
mc_ctl=mean(select(cost,treat:==uniqrows(treat)[1])) /* mean cost for ctl */
mc_tmt=mean(select(cost,treat:==uniqrows(treat)[2])) /* mean cost for tmt */
me_ctl=mean(select(effect,treat:==uniqrows(treat)[1])) /* mean eff for ctl */
me_tmt=mean(select(effect,treat:==uniqrows(treat)[2])) /* mean eff for tmt */
inb=(me_tmt-me_ctl)*lambda - (mc_tmt-mc_ctl)
return(inb)
}
mata mosave inb(), dir("local directory:\ado\personal") replace
end
```

A.3. nb()—net benefit (for calculating cost-effectiveness probabilities)

```
mata:
real matrix function nb(transmorphic matrix data, | real scalar lambda)
{
version 11.2
/* DECLARATIONS */
real matrix cost, effect, treat
real scalar mc_ctl, mc_tmt, me_ctl, me_tmt
/* DATA CHECKING */
if (cols(data)!=4){
 _error("Function requires an input data matrix of 4 columns wide.")
 }
cost=data[,1]
effect=data[,2]
treat=data[,4]
/* SET UNSPECIFIED PARAM VALUES */
if (lambda==.) lambda=20000
nstrata=rows(uniqrows(treat))
nb=J(1,nstrata,.)
for (i=1; i<=nstrata; i++) {
 nb[1,i] = mean(select(effect,treat:==uniqrows(treat)[i]))*lambda -
> mean(select(cost,treat:==uniqrows(treat)[i]))
}
return(nb)
}
mata mosave nb(), dir("local directory:\ado\personal") replace
end
```