Inference in Medical Cost Effectiveness Analysis with Heavy Tail Data.

Eduardo Fé*[1] and Simon Peters[2]

[1] *Department of Social Statistics, University of Manchester, U.K.*
[2] *Department of Economics, University of Manchester, U.K.*

December 19, 2022

Abstract

*Correspondence to: eduardo.fe@manchester.ac.uk

1 Introduction

Cost-effectiveness analysis is a critical step in the evaluation and adoption of medical innovations and policies. National and international health agencies advocate its use, and in some cases, such as the U.K. National Institute for Health and Care Excellence (NICE), recommendations are explicitly issued on the basis of *both* clinical and economic evidence. The Incremental Cost Effectiveness Ratio (ICER) and the associated Incremental Net Benefit (INB) are the two principal nonparametric statistics employed to evaluate the economic efficiency of new medical interventions and technologies[1]. To ensure that estimates of ICER/INB are not due to pure chance, confidence interval and tests statistics need to be calculated. As is generally the case, the exact distributions of the ICER /INB are elusive, because they depend on the unknown probabilistic process that generated the data. Two solutions have been studied to solve this problem: asymptotic approximations (valid as the sample size converges to infinity) and the bootstrap.

When the distribution of the ICER/INB has finite first two moments, the asymptotic distribution of ICER/INB follows from the central limit theorem. The asymptotic distribution of INB, in particular, is a non-degenerate stable distribution with finite moments (a normal distribution), from which approximate quantiles are simple to obtain. Aside from its simplicity, if the underlying data generating process is normal, then asymptotic approximations will yield confidence intervals with good coverage and tests with empirical sizes close to the nominal level. May be because of these reasons, the cost-effectiveness literature, authors have favoured asymptotic approximations (see, for example, Nixon et al., 2010 or recent examples by Ben et al., 2022, Fatoye et al., 2022; Tsiplova et al., 2022). When samples are finite and the underlying data are not normally distributed, however, asymptotic approximations can be crude, in general. This has led researchers to base statistical inference on bootstrap methods (Efron, 1979). When the distribution of a statistic has finite first two moments, and the statistic is a (asymptotic) pivot, bootstrap inference has been shown to be superior to asymptotic inference in finite samples. In a latter section, we replicate these findings, showing that the studentized version of the INB is a pivot and, as such, inference based on the bootstrap results in confidence intervals with superior coverage and tests with less size distortion.

Both asymptotic methods and the standard bootstrap work under the assumption that a given statistic follows a stable distribution with finite first and second moments. In cost-effectiveness analysis, however, this assumption is problematic. As noted by Jones et al. (2014), cost data (and economic data more generally) tends to exhibit heavy tails[2]. This empirical regularity was first remarked by Vilfredo Pareto in the late 1800s. Specifically, he pointed that the proportion of individuals with income exceeding a level u can be described by a scaling distribution, $P(u) \sim Cu^{-k}$, for a constant C and a parameter k known as the *index of stability*

[1] For example, at the time of writing in December 2022, PubMed.gov returns 1539 articles including the term Incremental Net Benefit or Cost Effectiveness Ratio for 2022 alone.

[2] Often accompanied by right-hand skewness. Note that the heaviness of the tail of a distribution determines the rate of decay of probabilities away from the mean, and it does not imply skewness. Similarly, skewed distributions does not necessarily have heavy tails -the log normal and gamma distributions being two examples of this. Extreme skewness is a feature of secondary importance that one would expect to see in variables that have a heavy tail and are non-negative.

and which accounts for the rate of decay away from the center of the distribution[3]. In the cost-effectiveness literature, this is mirrored in the stylised fact that small proportions of individuals tend to account for a very large proportion of medical costs. When a variable has a scaling distribution with heavy tails, either $1 \leq k < 2$ (implying that $E|U^2|) \rightarrow \infty$) or $0 \leq k < 1$ (in which case $E|U| \rightarrow \infty$ as well). Aside from violating the assumptions of the central limit theorem, variables with heavy tails exhibit extreme outlying observations, and these are known to present particular challenges to the standard bootstrap. The bootstrap distribution of a normalised sum of infinite variance random variables tends to a random distribution, and this lead to a failure of standard bootstrap methods (e.g. Athreya, 1987; Hall, 1990).From an empirical perspective, the standard bootstrap is based on an i.i.d. random resampling with replacement scheme. The probability of any element of the original sample not featuring in the bootstrap sample approaches[4] $((n-1)/n)^n \rightarrow 1/e = 0.3679$ as the sample size goes to infinity. Outlying observations will be regularly missed in bootstrap samples, and this will result in biased inferences, through under-representation of the variation in the data.

Several alternative bootstrap designs have been proposed for inference with heavy tails. Politis and Romano (1994) and Romano and Wolf (1999) proposed an m-out-of-n bootstrap based on sub-samples of size $m < n$. This scheme consistently estimates the true asymptotic distribution of a sample mean. However, Hall and Yao (2003) note that subsampling can return very conservative confidence intervals. Further, results in Davidson and Flachaire (2007) and Cornea-Madeira and Davidson (2015) suggest the existence of an optimal m/n, with performance deteriorating as m departs from the optimal choice. These latter papers present semi-/parametric bootstrap methods that improve on resampling designs, but which require structural assumptions about some aspects of the distribution of the statistic.

Cavaliere et al. (2013), propose a wild bootstrap method for inferences about a mean in the infinite variance case. The wild bootstrap (Wu, 1986, Liu, 1988), has been shown to provide refinements for inference in various settings, including linear models with heteroskedastic errors (Davidson and Flachaire, 2008), high-dimensional linear models (Mammen, 1993) consistent nonparametric testing (Hardle and Mammen, 1993), and linear models with clustered errors (Cameron et al., 2008a; Djogbenou et al., 2019). Cavaliere et al. (2013) show that their scheme delivers consistent estimation of the asymptotic distribution of the sample mean, conditional on $\{|X_i - E(X_i)|\}_{i=1}^n$. The latter ensures that resampling preserves the sample extremes and, as a result, it improves the coverage of confidence intervals and reduces the size distortion of tests of hyptheses. Unlike subsampling methods, the wild bootstrap sets $m = n$ so that all information is used to compute the bootstrap distribution and there is not need to find and optimal resampling ratio m/n. Unlike semi/parametric bootstrap methods, researchers do not need to estimate the index of stability k.

In this paper, we study inference about INB under heavy tails, using asymptotic approximations, the standard bootstrap and the wild bootstrap in Cavaliere et al. (2013). However,

[3] Following Pareto's observation, work by Zipf (1949), Mandelbrot (1963), Singh and Maddala (1976) and others provided further theoretical justification for this idea and, in recent times, authors have shown the adequacy of scaling, heavy tail distributions to fit economic data (Kumar, 2017; Jenkins, 2017; Schluter and Trede, 2002).

[4] The probability of not observing element i in a random sampling with replacement from $i = 1...n$ distinct observations is $n - 1/n$, and $e = \lim_{n\rightarrow\infty}((n+1)/n)^n$.

cost-effectiveness analysis in medical research typically follows a randomized experiment, with the randomization distribution being often known a priori. This opens the possibility of implementing randomization inference for the INB ratio. Specifically, knowledge of the randomization distribution enables us to obtain the finite sample distribution of the INB under the null hypothesis $H_0 : INB = inb^*$, for any inb^*. Consequently, one can undertake exact finite sample inferences about the INB, without any assumption beyond our knowledge of the randomization distribution of treatment. Randomization Inference has been shown to provide tests with almost-nominal size and confidence intervals with correct coverage in a wide variety of settings (Cattaneo et al., 2015; Ho and Imai, 2006; Imbens and Rosenbaum, 2005; MacKinnon and Webb, 2020). Below we explore if its promise survives the heavy tail scenario, and evaluate its relative performance against the other inferential methods.

2 Cost-Benefit Statistics.

Consider a randomized trial designed to assess the cost-effectiveness of a new policy, technology or medical innovation. Let $Z_i \in \{0,1\}$, $i = 1, ..., n$ denote person's i treatment status (with 0 denoting assignment to the control group). Assignment is randomized, so that, for any random variable W, $E(W|Z) = E(W)$. Cost and benefit data are collected for each participant. Specifically, let $C_i(z)$, $B_i(z)$ be the potential cost and benefit of person i under assignment $z \in \{0,1\}$. This notation assumes Stable Unit Treatment Value Assumption. Researchers then observe

$$C_i = C_i(1) \cdot Z_i + C_i(0) \cdot (1 - Z_i) \tag{1}$$

$$B_i = B_i(1) \cdot Z_i + B_i(0) \cdot (1 - Z_i) \tag{2}$$

A new technology is considered to be cost-effective if the estimated Incremental Cost Effectiveness Ratio, ICER, satisfies,

$$ICER = \frac{E(C(1) - C(0))}{E(B(1) - B(0))} = \frac{E(C_i|Z_i = 1) - E(C_i|Z_i = 0)}{E(B_i|Z_i = 1) - E(B_i|Z_i = 0)} \leq K \tag{3}$$

where the second equality follows from random assignment of Z. The quantity K is a known constant, and represents the policymaker's willingness to pay for the innovation. This is normally fixed beforehand, by a decision-taker. A potential weakness of the ICER is that denominator is unrestricted, and can thus equal (or approach) zero if an innovation does not yield any benefits. As a result, the ICER could be unstable and elusive for estimation methods. In practice, researchers work with the equivalent Incremental Net Benefit, INB,

$$\begin{aligned} INB &= K \cdot E(B(1) - B(0)) - E(C(1) - C(0)) \\ &= K \cdot [E(B_i|Z_i = 1) - E(B_i|Z_i = 0)] - [E(C_i|Z_i = 1) - E(C_i|Z_i = 0)] \end{aligned} \tag{4}$$

Let $\sigma^2_{W_z} = V(W_z)$ be the variance of W under assignment $z \in \{0,1\}$, which under random assignment equals $V(W_i|Z_i = z)$. Similarly, let $cov(C_i(z), B_i(z))$ be the covariance of cost and benefit under assignment z, which once again equals its conditional version under random assignment. Then, it is straightforward to show that the variance of INB is given by:

$$\Sigma = K^2 \cdot (\sigma^2_{B_1} + \sigma^2_{B_0}) + (\sigma^2_{C_1} + \sigma^2_{C_0}) - 2 \cdot K \cdot (cov(C_1, B_1) + cov(C_0, B_0)). \tag{5}$$

Define the sample averages,

$$\bar{C}_1 = \frac{\sum_{i=1}^{n} C_i \cdot Z_i}{\sum_{i=1}^{n} \cdot Z_i} \text{ and } \bar{C}_0 = \frac{\sum_{i=1}^{n} C_i \cdot (1 - Z_i)}{\sum_{i=1}^{n} \cdot (1 - Z_i)} \tag{6}$$

with $\bar{B}_0$, $\bar{B}_1$ defined similarly. It follows directly from random assignment of Z_i and a law of large numbers that, if $C_i(z), B_i(z)$ have finite first moments, then these sample means estimate their population counterparts, $E(C(z)), E(B(z))$. Similar,

$$\hat{\sigma}^2_{C_1} = \frac{\sum_{i=1}^{N} \left(C_i \cdot Z_i - \bar{C}_1\right)^2}{\sum_{i=1}^{n} \cdot Z_i} \tag{7}$$

is a consistent estimator of $\sigma^2_{C_1}$ (and similar estimators can be defined for the remaining variances and covariances). The latter also requires $E(C_i(z)^2) < \infty$, $E(B_i(z)^2) < \infty$. The INB and its variance can thus be estimated with

$$\widehat{INB} = K(\bar{B}_1 - \bar{B}_0) - (\bar{C}_1 - \bar{C}_0) > 0 \tag{8}$$

$$\hat{\Sigma} = K^2(\hat{\sigma}^2_{B_1} + \hat{\sigma}^2_{B_0}) + (\hat{\sigma}^2_{C_1} + \hat{\sigma}^2_{C_0}) - 2K(\widehat{cov}(C_1, B_1) + \widehat{cov}(C_0, B_0)). \tag{9}$$

We want to construct a confidence interval for INB using

$$Prob(a \leq \widehat{INB} - INB \leq b|F) = 1 - 2\alpha. \tag{10}$$

for a pre-specified significance level α and where F is the distribution of the data. F is not known and this precludes us from computing the constants a and b. However, $\widehat{INB}$ is a linear combination of sample means. Therefore, the central limit theorem can be applied to obtain a large sample ($N \to \infty$) approximation for its distribution. In particular, under the assumption of finite first two moments of $C(z), B(z)$, we have

$$\sqrt{N} \cdot (\widehat{INB} - INB) \xrightarrow{d} N(0, \Sigma) \tag{11}$$

and the continuous mapping theorem then yields the *studentized* $\widehat{INB}$,

$$\hat{\tau} = \sqrt{N} \cdot (\widehat{INB} - INB)/\sqrt{\hat{\Sigma}} \xrightarrow{d} N(0, 1) \tag{12}$$

from which we can now derive the quantities a, b. Critically, $\hat{\tau}$ does not depend on unknown quantities and is, therefore, an asymptotic pivot[5]. Tests of the null hypothesis $H_0 : INB = 0$

[5] It is convenient to remark here that the success of the bootstrap methods that we consider below depends on how well it approximates the distribution of $\widehat{INB} - INB$. This depends on four factors: simulation error, statistical error, the ability of the bootstrap algorithm to reproduce the true data generating process and the *pivotal* nature of the underlying statistic. Of these factors, the last is germane to the question of when the bootstrap works. Specifically, the distribution of $\widehat{INB} - INB$ depends on unknown quantities, which need to be estimated, thus introducing a systematic discrepancy between the estimated and true distributions (in the

can be based on $\hat{\tau}$. The null hypothesis would be rejected whenever the p-value $p(\hat{\tau}) = 1 - F(\hat{\tau})$ falls below a pre-specified significance level, α. The unknown cumulative distribution function of $\hat{\tau}$ under the null hypothesis, $F(.)$, can be replaced with the asymptotic distribution, and this is the standard approach in the cost-effectiveness literature. In that case, p-values follow from the quantiles of the standard normal distribution. A more successful approach, however, is to use Monte Carlo methods to approximate $F(.)$, as we now discuss.

3 Bootstrap and Randomization Inference

Bootstrap methods rely on the principle that a sample contains all (relevant) available information about the distribution F of a statistic. Under this principle, the standard bootstrap in Efron (1979), relies on B samples of size n, each obtained by sampling with replacement from the original data. The sequence $\{\tau_b^*\}_{b=1}^B$ of standard boostrap INB estimates of τ then has the ability to mimic the finite sample variation of the statistic $\hat{\tau}$and, as a result, the empirical distribution of the τ_b^*, say $\hat{F}^*$, serves as an approximation to F. Because B can be made arbitrarily large, any error due to simulation can be virtually eliminated. It follows that

$$\hat{p}^* = 1 - \hat{F}^*(\hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B} I(\tau_b^* > \hat{\tau}) \tag{13}$$

is a valid p-value for a one-sided test of the null hypothesis $H_0 : INB = 0$ against $H_a : INB > 0$ whereas a two-sided test can be evaluated with

$$\hat{p}^* = 2\min\left(\frac{1}{B}\sum_{b=1}^{B} I(\tau_b^* \leq \hat{\tau}), \frac{1}{B}\sum_{b=1}^{B} I(\tau_b^* > \hat{\tau})\right) \tag{14}$$

We can compute confidence intervals similarly (e.g. the lower bound for a confidence interval with significance level α corresponds to the $\alpha \cdot (B+1)/2$ quantile of the standard bootstrap distribution $\hat{F}^*$). When INB has finite moments, inference based on the standard bootstrap distribution of $\hat{F}^*$ will yield asymptotic refinements because $\hat{\tau}$ is a pivot.[6]. On the contrary, working with $\hat{INB}$ will fail to improve on asymptotic methods.

Overall, the performance of any bootstrap scheme is determined by how well it can replicate the key features of the process that generated the data and which most affect $\hat{\tau}$. The standard bootstrap relies on the existence of finite first two moments, but data exhibiting heavy tails violate this premise: we already noted in the introduction that the bootstrap distribution of a normalised sum of infinite variance random variables tends to a random distribution, and this leads to a failure of the stadndard bootstrap. The INB is a linear combination of sums

Normal case this means that $N(\hat{\mu}, \hat{\sigma}^2) \neq N(\mu, \sigma^2)$ even though both distributions might be close to each other if $\hat{\mu}, \hat{\sigma}^2$ are consistent estimators). When working with a pivot this problem is avoided.

[6] Specifically, if a pivot statistic has an exact distribution G and a standard normal asymptotic distribution, Φ, then the error in the bootstrap approximation to G vanishes at rate n^{-1} (in probability), whereas the error in the asymptotic approximation does so slower, at rate $n^{-1/2}$, where n is the sample size. If the statistic, however, is not a pivot, then the error in the bootstrap approximation is also of order $n^{-1/2}$. See Beran (1988) Hall (1992) and Davidson and Hinkley (1997) for theoretical demonstrations of this result. For empirical demonstrations, Orme (1990); Davidson and Flachaire (2007); Cameron et al. (2008b); Bugni (2010), constitute a few early representative examples in a variety of fields.

	k								
	0.1	0.5	1	1.5	1.6	1.7	1.8	1.9	2
Stability Index of INB	0.097209	0.48111	1.0151	1.6535	1.8367	2.0941	2.3388	2.4963	2.7310

Table 1: Hill estimates of the stability index of INB when all $C(z)$, $B(z)$ follow a Pareto distribution with minimum 1 and stability index k. Based on 100,000 draws of size 2000.

of random variables, so it will be vulnerable to the heavy tail problem. For example, if data come from a Pareto distribution, the INB will also be Pareto distributed, also with heavy tails. The latter can be shown numerically. Table 1 presents the Hill (1975) estimate[7] of the stability index of INB when $C(z), B(z)$ are Pareto distributed with minimum value 1 and under a range of values of k. For $k < 1.7$, the INB inherits the heavy tail property from the data and is Pareto distributed with infinite variance. This implies that the standard bootstrap will also fail in these situations.

Cavaliere et al. (2013) have proposed a wild bootstrap method that provides excellent inference for the sample mean in the infinite variance case. Given a sample of variables X_i, instead of resampling with replacement from the data, bootstrap samples are based on

$$X_i^* = \bar{X} + (X_i - q(X)) \cdot w_i \text{ for } i = 1, ..., n \tag{16}$$

where $q(X)$ is a centrality measure: either the sample mean of X or if data are heaviliy skweded, the median of X. The term w_i is a sequence of i.i.d. random variables with $E(w_i) = 0$ and $E(w_i^2) = 1$. There are various choices for the distribution of w_i in the literature, however the Rademacher distribution (which takes value 1 or -1 with probability 0.5 each) has been proved to deliver the best performance (Davidson and Flachaire, 2008). Intuitively, the above resampling scheme takes into account the variation of data around the median, thus being able to characterise the tails of the distribution of the mean. Interestingly, Cavaliere et al. (2013) prove that the above bootstrap scheme works also for situations when the data do not have a finite mean or a finite variance.

3.1 Randomization Inference.

In recent times, Fisher's Randomization Inference has gained popularity, due to its ability to yield exact tests which are robust to a variety of statistical problems (REFS). Randomization inference is particularly appealing in settings where treatment is unconfounded, such as the controlled experiments preceding cost-effectiveness analysis. Consider the sharp null hypothesis

[7] Hill's estimator is

$$\hat{k} = \left[q^{-1} \sum_{i=0}^{q-1} \log INB_{(n-i)} - \log INB_{(n-q+1)} \right]^{-1} \tag{15}$$

where $INB_{(j)}$ is the j^{th} ordered value of the simulated INB, and q determines the largest ordered statistic to use in the estimation.

that the new medical innovation is not cost-effective. Formally, this can be written as

$$H_0 : K \cdot B_i(1) - C_i(1) = K \cdot B_i(0) - B_i(0) \text{ for all } i = 1, ..., n \tag{17}$$

Under the above null, we can impute each treated (control) unit her/his missing INB under control (treatment respectively). That is, for a unit i with $Z_i = 1$, under the above sharp null, $K \cdot B(0) - C(0) = K \cdot B_i - C_i$ (and similarly for a unit with $Z_i = 0$). Further, in a controlled randomized experiment, the distribution of the treatment assignment is known. The latter then implies that we can study the variation of any statistics (such as INB or $\hat{\tau}$ under) 'all' possible $\binom{n}{n_1}$ permutations of the observed treatment assignment, $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$, where n_1 is the number of 'treated' units (which is determined a priori in randomized experiments). As a result, the null randomization distribution of these statistics is known, and they can be used construct a test of the above*sharp* null hypothesis.

In practice, even with moderate n, the number of possible permutations of $\mathbf{Z}$ is too large to consider, so researchers consider only a random selection, R, of all possible permutations. Strictly speaking, the ensuing tests will be approximately exact, but as with the bootstrap researchers can reduce the approximation error by selecting a large enough R. Thus, to construct a randomization inference test for INB, we can follow this algorithm,

1. Calculate the standardized INB, $\hat{\tau}$
2. Repeat the following $r = 1, ..., R$ times:
   (a) Draw a permutation, $\tilde{\mathbf{Z}}_r$ of the original assignment vector $\mathbf{Z}$.
   (b) Compute $\hat{\tau}$ given $\tilde{\mathbf{Z}}$, say $\hat{\tau}(\tilde{\mathbf{Z}}_r)$
3. Compare the observed $\hat{\tau}$ to its null distribution, to obtain the p-value for a one-sided tests:

$$p \approxeq R^{-1} \cdot \sum_{r=1}^{R} \mathbb{I}(\hat{\tau}(\tilde{\mathbf{Z}}_r) > \hat{\tau}) \tag{18}$$

4 Monte Carlo.

We illustrate the above discussion through two experiments. In the first instance, we drew data from a set of four standard distributions and computed asymptotic and bootstrap confidence intervals for the population mean. In the second instance, we estimated confidence intervals for the INB statistic. This second experiment is based on the design by Nixon et al. (2010). We calculated asymptotic confidence intervals and bootstrap confidence intervals based ($B = 199$ resamples) on the raw statistics and their studentized versions. We consider samples sizes of $N \in \{10, 20, 40, 80\}$ observations. In total, we generated data from 6 generating processes, and for each model we undertook $R = 100,000$ estimations of the confidence interval. The quantity of interest is the coverage of each confidence interval which was computed as the proportion of times that the true value of the parameters (population mean or INB) were contained within the computed confidence intervals. Nominal coverage was set at 95%.

Both the sample mean and INB have asymptotic normal distributions and, therefore, the corresponding asymptotic confidence intervals were computed in the usual manner. Let $\hat{T}$ denote the sample mean or INB and let θ be their true values. To compute the bootstrap confidence intervals we used the following two algorithms.

Percentile method.

1. Repeat the following $b = 1, \ldots, B = 199$ times:
   (a) Draw a sample of size N randomly with replacement from the original data set.
   (b) Compute and save the statistic of interest from each simulated sample, $\hat{T}_b$
2. Select the α and $1-\alpha$ percentiles of the series $(\hat{T}_b)_{b=1}^{B}$. Denote these $\hat{T}_{B,\alpha}$ and $\hat{T}_{B,1-\alpha}$ respectively.
3. A percentile method bootstrap confidence interval for θ is

$$(\hat{T}_{B,\alpha}, \hat{T}_{B,1-\alpha}) \tag{19}$$

Percentile-t method.

1. Compute $\hat{T}$ and its standard error $\hat{\sigma}_T$. Retain these values.
2. Repeat the following $b = 1, \ldots, B = 199$ times:
   (a) Draw a sample of size N randomly with replacement from the original data set.
   (b) Compute the statistic of interest from each simulated sample, $\hat{T}_b$, along with its standard error $\hat{se}(\hat{T}_b)$.
   (c) Compute and save $t_b = (\hat{T}_b - \hat{T})/\hat{se}(\hat{T}_b)$
3. Select the α and $1-\alpha$ percentiles of the series $(t_b)_{b=1}^{B}$. Denote these $t_{b,\alpha}$ and $t_{b,1-\alpha}$ respectively.
4. A percentile-t method bootstrap confidence interval for θ is

$$(\hat{T} - \hat{\sigma}_T t_{b,1-\alpha}, \hat{T} - \hat{\sigma}_T t_{b,\alpha}) \tag{20}$$

4.1 Experiment 1.

Data were drawn from four distributions, namely (i) $N(0,1)$, (ii) χ_2^2, (iii) $\Gamma(2,2)$, (iv) $N(0,1) + LN$ where LN is a standard log normal distribution. The parameter of interest was the population mean, $E(Y)$, which was estimated by the sample mean, $\bar{Y}$. Note that case (i) is the most favourable to asymptotic theory, as the central limit theorem approximates the distribution of the statistic exactly. Case (iv) is a simple version of the data generating process that we consider in the following section. The results of the simulation are given in table 2.

The top panel in the table contains the results obtained when data came from the $N(0,1)$ distribution. The confidence intervals produced by asymptotic theory exhibit excellent coverage,

Table 2: Experiment 1. Empirical coverage of asymptotic and bootstrap confidence intervals. $R = 100,000$. $B = 199$. Nominal coverage 95%.

	$Y \sim N(0,1)$		
N	Asymptotic	Percentile	Percentile-t
10	0.94048	0.90478	0.95098
20	0.94510	0.92852	0.94999
40	0.94804	0.94057	0.95064
80	0.94824	0.94482	0.95009
	$Y \sim \chi^2_2$		
N	Asymptotic	Percentile	Percentile-t
10	0.89076	0.84258	0.94167
20	0.91467	0.88848	0.94689
40	0.92941	0.91421	0.94778
80	0.94010	0.93239	0.95001
	$Y \sim \Gamma(2,2)$		
N	Asymptotic	Percentile	Percentile-t
10	0.91496	0.87209	0.94615
20	0.92777	0.90744	0.94802
40	0.93839	0.92701	0.92643
80	0.94333	0.93764	0.94888
	$Y \sim N(0,1) + LN$		
N	Asymptotic	Percentile	Percentile-t
10	0.89652	0.85612	0.91623
20	0.90857	0.88855	0.91958
40	0.91875	0.90651	0.92643
80	0.93144	0.92480	0.93541

close to the nominal 95%. However, the performance of the percentile method is poor. For $N = 10$, the coverage of the simulated confidence intervals is just 0.90 (compare with the value of 0.940 obtained with asymptotic methods). As the underlying statistics in the bootstrap and asymptotic methods are identical the discrepancy can only be due to simulation error, which suggests that $B = 199$ is too small. Yet, for identical B, the bootstrap algorithm based on the studentized mean,the percentile-t method, provides outstanding coverage, considerably superior to the two previous methods. Thus, for $N = 10$ we obtain a coverage of 0.95098, implying an error of 0.1031%. This is in contrast to the error of 1.002% obtained with asymptotic approximations. As N increases, the performance of the asymptotic and percentile methods improves, as expected, but so does the performance of the percentile-t method. Thus for $N = 80$ the discrepancies between empirical and nominal coverage levels are 0.18%, 0.54% and 0.009% for the asymptotic, percentile method and percentile-t method respectively. The conclusion remains invariant for each of the remaining three data generating processes which are included to illustrate the ability of appropriate bootstrap methods to capture non-normal features of distributions, such as skewness and kurtosis. Even in the least favourable scenario, case (iv), the percentile-t method clearly improves over asymptotic and percentile methods.

4.2 Experiment 2.

Here we revisit the data generating process in Nixon et al. (2010) and work with the INB described in section 1. We consider the coverage of three sets of confidence intervals for INB

		$\sigma_c = 0.05$			$\sigma_c = 0.5$	
N	Asyp.	Percentile	Percentile-t	Asyp.	Percentile	Percentile-t
10	0.92019	0.92024	0.92021	0.91985	0.91679	0.92205
20	0.93692	0.93691	0.93723	0.93371	0.93133	0.93599
40	0.94332	0.94332	0.94381	0.93445	0.93275	0.93634
80	0.94639	0.94698	0.94765	0.92582	0.92486	0.92822

Table 3: Monte Carlo Simulation. $B = 199$, $R = 100,000$. The left panel refers to a scenario with low skewness in the distribution of costs, while the right panel refers to a situation of high skewness in the distribution of costs.

when data were generated from the following design. Firstly,

$$\log C_0 = \sigma_c * N(0,1) + \log(50000) \quad (21)$$

$$\log C_1 = \sigma_c * N(0,1) + \log(50000 + 30000) \quad (22)$$

so that C_0, C_1 have log-normal distributions. Here σ also determines the skewness of the distributions. For instance, $\sigma = 0.05$ implies a skewness of 0.69, while $\sigma = 0.5$ implies a skewness of 2.93. The effectiveness variables are:

$$B_0 = \sigma_b N(5,1) + 0.3(\log C_0 - \log(50000)) \quad (23)$$

$$B_1 = \sigma_b N(6,1) + 0.3(\log C_1 - \log(50000 + 30000)) \quad (24)$$

where $\sigma_b = 2$. In the simulations the standard error were fixed at $\sigma_e = 2$ and $\sigma_c \in \{0.05, 0.5\}$. Willingness to pay was set at $K = 20,000$ which implies a true INB of -10,000.

The results of this simulation are given in table 3. We observe that all three methods fall short of producing outstanding results for the sample sizes considered. The percentile-t method provides the best coverage, followed by the asymptotic approximation and, lastly, the percentile method algorithm.

5 Conclusion.

In this note we show that 'studentizing' the Incremental Net Benefit statistic results in an asymptotic pivot quantity that provides improved inference in cost-effectiveness analysis. We provide some supporting Monte Carlo experiment. However, improvements are modest and both asymptotic and bootstrap methods tend to have weak coverage, particularly with skewed distributions. This is reminiscent of the results in Davidson and Flachaire (2007), who found that both asymptotic and standard bootstrap methods exhibit poor coverage in some standard inequality and poverty measures. Those authors suggested a parametric bootstrap that partly ameliorated coverage in at least moderate samples. Thus, future research could explore the validity of parametric bootstrap methods in cost-effectiveness analysis.

References

Athreya, K.B. (1987). Bootstrap of the Mean in the Infinite Variance Case. *The Annals of Statistics*, 15(2): 724 – 731

Ben, Â.J., **van Dongen, J.M.**, **Alili, M.E.**, **Heymans, M.W.**, **Twisk, J.W.R.**, **MacNeil-Vroomen, J.L.**, **de Wit, M.**, **van Dijk, S.E.M.**, **Oosterhuis, T.**, and **Bosmans, J.E.** (2022). The handling of missing data in trial-based economic evaluations: should data be multiply imputed prior to longitudinal linear mixed-model analyses? *The European Journal of Health Economics.* doi:10.1007/s10198-022-01525-y

Beran, R. (1988). Discussion: Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics*, 16(3): 956 – 959

Bugni, F.A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78(2): 735–753

Cameron, A.C., **Gelbach, J.B.**, and **Miller, D.L.** (2008a). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3): 414–427. ISSN 0034-6535. doi:10.1162/rest.90.3.414

Cameron, C., **Gelbach, J.**, and **Miller, D.** (2008b). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics.*, 90: 414–427

Cattaneo, M.D., **Frandsen, B.R.**, and **Titiunik, R.** (2015). *Journal of Causal Inference*, 3(1): 1–24. doi:doi:10.1515/jci-2013-0010

Cavaliere, G., **Georgiev, I.**, and **Taylor, A.M.R.** (2013). Wild bootstrap of the sample mean in the infinite variance case. *Econometric Reviews*, 32(2): 204–219

Cornea-Madeira, A. and **Davidson, R.** (2015). A parametric bootstrap for heavy-tailed distributions. *Econometric Theory*, 31(3): 449–470. ISSN 02664666, 14694360

Davidson, A. and **Hinkley, D.V.** (1997). *Bootstrap Methods and their Applications.* Cambridge University Press

Davidson, R. and **Flachaire, E.** (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics*, 141(1): 141–166. Semiparametric methods in econometrics

Davidson, R. and **Flachaire, E.** (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1): 162–169. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2008.08.003

Djogbenou, A.A., **MacKinnon, J.G.**, and **Nielsen, M.Ø.** (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics*, 212(2): 393–412. ISSN 0304-4076. doi:https://doi.org/10.1016/j.jeconom.2019.04.035

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, 7: 1–26

Fatoye, F., **Gebrye, T.**, **Mbada, C.E.**, **Fatoye, C.T.**, **Makinde, M.O.**, **Ayomide, S.**, and **Ige, B.** (2022). Cost effectiveness of virtual reality game compared to clinic based mckenzie extension therapy for chronic non-specific low back pain. *British Journal of Pain*, 16(6): 601–609. doi:10.1177/20494637221109108

Hall, P. (1992). *The bootstrap and Edegeworth expansion.* Springer

Hall, P. (1990). Asymptotic properties of the bootstrap for heavy-tailed distributions. *The Annals of Probability*, 18(3): 1342–1360. ISSN 00911798

Hall, P. and **Yao, Q.** (2003). Inference in arch and garch models with heavy-tailed errors. *Econometrica*, 71(1): 285–317. ISSN 00129682, 14680262

Hardle, W. and **Mammen, E.** (1993). Comparing Nonparametric Versus Parametric Regression Fits. *The Annals of Statistics*, 21(4): 1926 – 1947. doi:10.1214/aos/1176349403

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5): 1163–1174. ISSN 00905364

Ho, D.E. and **Imai, K.** (2006). Randomization inference with natural experiments. *Journal of the American Statistical Association*, 101(475): 888–900. doi:10.1198/016214505000001258

Imbens, G.W. and **Rosenbaum, P.R.** (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 168(1): 109–126. ISSN 09641998, 1467985X

Jenkins, S.P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica*, 84(334): 261–289

Jones, A.M., **Lomas, J.**, and **Rice, N.** (2014). Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, 29(4): 649–670

Kumar, D. (2017). The Singh–Maddala distribution: properties and estimation. *International Journal of System Assurance Engineering and Management*, 8(2): 1297–1311

Liu, R.Y. (1988). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, 16(4): 1696–1708. ISSN 00905364

MacKinnon, J.G. and **Webb, M.D.** (2020). Randomization inference for difference-in-differences with few treated clusters. *Journal of Econometrics*, 218(2): 435–450. ISSN 0304-4076. doi:https://doi.org/10.1016/j.jeconom.2020.04.024

Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics*, 21(1): 255 – 285. doi:10.1214/aos/1176349025

Mandelbrot, B. (1963). New methods in statistical economics. *Journal of Political Economy*, 71(5): 421–440

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

Orme, C. (1990). The small sample performance of the information matrix test. *Journal of Econometrics*, 46: 309–331

Politis, D.N. and **Romano, J.P.** (1994). Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *The Annals of Statistics*, 22(4): 2031 – 2050. doi: 10.1214/aos/1176325770

Romano, J.P. and **Wolf, M.** (1999). Subsampling inference for the mean in the heavy-tailed case. *Metrika*, 50(1): 55–69. doi:10.1007/s001840050035

Schluter, C. and **Trede, M.** (2002). Tails of lorenz curves. *Journal of Econometrics*, 109(1): 151–166

Singh, S.K. and **Maddala, G.S.** (1976). A function for size distribution of incomes. *Econometrica*, 44(5): 963–970

Tsiplova, K., **Jegathisawaran, J.**, **Mirenda, P.**, **Kalynchuk, K.**, **Colozzo, P.**, **Smith, V.**, and **Ungar, W.J.** (2022). Parent coaching intervention for children with suspected autism spectrum disorder: Cost analysis. *Research in Autism Spectrum Disorders*, 93: 101949. ISSN 1750-9467

Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4): 1261–1295. ISSN 00905364

Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press