A numerical investigation of the accuracy of parametric bootstrap for discrete data

Chris Lloyd, Melbourne Business School
A numerical investigation of the accuracy of parametric bootstrap for discrete data

BY CHRIS J. LLOYD

University of Melbourne, Carlton, 3053, AUSTRALIA
c.lloyd@mbs.edu

ABSTRACT: Standard first order tests have size error that decreases as \( m^{-1/2} \) where \( m \) is a measure of sample size. Parametric bootstrap tests use an exact calculation of the P-value, assuming nuisance parameters equal their null maximum likelihood estimates. It is commonly believed that their performance is driven by asymptotics, notwithstanding some confusion in the literature on asymptotic error rates.

For simple discrete models, parametric bootstrap tests can be calculated explicitly rather than simulated. Moreover, their accuracy can also be calculated exactly as a function of the nuisance parameter. This article reports the results of an intensive numerical investigation of the accuracy of first order and parametric bootstrap based tests, firstly of treatment effect in clinical trials, and secondly of association in simple logistic regression. We conclude that bootstrap tests have asymptotic error rate \( O(m^{-1}) \) but that their excellent small sample performance has little to do with asymptotics.

Key Words: bootstrap; exact test; nuisance parameters; r-star; tests of independence;

1 Introduction

Suppose we have a model \( \pi(y; \psi, \lambda) \) for discrete data \( Y \in \mathcal{Y} \) and want to test the alternative hypothesis \( \psi > \psi_0 \). We have a test statistic \( T \), larger values of which are considered more hostile to the null. The distribution of \( T \) is approximated by a known distribution \( H \) that does not depend on \( \lambda \), often normal or chi-squared. Then the achieved size of a nominal size \( \alpha \) test is

\[
a_1(\lambda) = \Pr(1 - H(T) \leq \alpha; \psi_0, \lambda).
\] (1)
The subscript "1" here refers to first-order methods. A more modern approach is to calculate the exact significance of $t = T(y)$ which is

$$S(t, \lambda) := \Pr(T(Y) \geq t; \psi_0, \lambda) = \sum_{y: T(y) \geq t} \pi(y; \psi_0, \lambda).$$

Instead of approximating this by $1 - H(t)$ we replace $\lambda$ by its estimate $\hat{\lambda}_0$ under the null and calculate the tail probability exactly. This gives rise to the parametric bootstrap P-value $\hat{P}(y) = S(T(y), \hat{\lambda}_0(y))$. The achieved size of this test is

$$a_2(\lambda) = \Pr(\hat{P}(Y) \leq \alpha; \psi_0, \lambda).$$

(2)

This paper concerns the size accuracy of tests based on the first order statistic $T$, tests based on second order statistics $\hat{P}(Y)$, and how this depends on sample size.

Example. In a clinical trial $n_1(n_0)$ patients are given treatment (placebo) and the number of positive outcomes recorded. We are interested in the probabilities $p_1(p_0)$ of positive outcome. A standard measure of treatment effect is the so-called risk ratio $p_1/p_0$. There are two common test statistics for testing the risk ratio, details of which will be given in section 3. Both have asymptotic normal distribution as $\min(n_0, n_1)$ increases, provided that $(p_0, p_1)$ are not on the boundary.

Figure 1 shows the actual size of two particular tests with nominal size 2.5%. More details of these tests will be given in section 3. The left panel is for a score-type statistic of Chan (1998), with sample sizes $(n_0, n_1) = (50, 250)$ when we are testing $p_1 > p_0$. The solid line is the actual size of the first order test, and the bold line is for the test based on $\hat{P}(Y)$. The right panel is for the likelihood ratio (LR) statistic with sample sizes $(n_0, n_1) = (30, 70)$ when we are testing $p_1/p_0 > 0.9$. It is no surprise to see the bootstrap test being more accurate. However, it is perhaps surprising how poor the first order methods are and how uniformly good the bootstrap tests are.

The purpose of this paper is to numerically investigate the extent to which bootstrap
Figure 1: Exact size $a_1(\lambda)$ of first order test (solid) and $a_2(\lambda)$ of bootstrap P-value based test (bold). *Left.* Statistic $T$ is the Score for testing $p_1 > p_0$ with binomial data from $(n_0, n_1) = (50, 250)$ trials. *Right.* Statistic $T$ is a LR statistic for testing $p_1/p_0 > 0.9$ with binomial data from $(n_0, n_1) = (30, 70)$ trials.

improves on first order tests, specifically how performance depends on the value of the nuisance parameter and on sample size, and to compare this with asymptotic theory.

## 2 Background on existing asymptotic theory

Henceforth all probabilities are understood to be at the null $\psi = \psi_0$ and we drop $\psi_0$ from the notation. Nuisance parameters are denoted $\lambda$ and we use $m$ as a measure of sample size. Standard first order test statistics such as Score, Wald or signed LR statistic have asymptotic normal distribution with error term $O(m^{-1/2})$, under conditions referred to as “local asymptotic normality”. In this case, size errors $a_1(\lambda) - \alpha$ are $O(m^{-1/2})$. For discrete models, local asymptotic normality breaks down at boundary parameter values so size accuracy may be poor when $\lambda$ is near the boundary.
The bootstrap P-value is an estimate $S(t, \hat{\lambda})$ of the exact significance $S(t, \lambda)$, where $\hat{\lambda}$ differs from $\lambda$ by $O_p(m^{-1/2})$. Under local asymptotic normality, $S(t, \lambda)$ is expanded in powers of $m^{1/2}$ and it is simple to show that $S(t, \hat{\lambda}) - S(t, \lambda)$ is $O_p(m^{-1})$. This argument is due to Beran (1988). Methods with error $O(m^{-1})$ are generically referred to as ”second order.”

The P-value under study in this paper is $\hat{P}(Y) = S(T(Y), \hat{\lambda}_0(Y))$ which uses the maximum likelihood (ML) estimator $\hat{\lambda}_0$ under the null. Building on results of DiCiccio and Stern (1994) and DiCiccio, Martin and Stern (2001) for continuous models, Lee and Young (2005) showed that for any $\alpha \in [0, 1]$

$$\Pr(\hat{P}(Y) \leq \alpha; \lambda) = \alpha + O \left(m^{-3/2}\right)$$

(3) for any $\lambda$ for which an Edgeworth expansion of the distribution of $T(Y)$ is valid. This third order accuracy result relies on the null covariance of $m^{1/2}(\hat{\lambda}_0 - \lambda)$ and $T$ being $O(m^{-1/2})$ rather than $O(1)$. This is not true of the unrestricted ML estimator.

Fraser and Rousseau (2008) suggest inferential errors will be $O(m^{-1})$ for discrete models for any estimator $\hat{\lambda}$ of $\lambda$ with $O(m^{-1/2})$ errors. Numerical illustrations by DiCiccio and Young (2008) also suggest this rate. Yet numerical results in Lloyd (2008) showed that using $\hat{\lambda}_0$ makes a huge practical difference.

None of these results address how inferential errors depend on $\lambda$. They do not apply at boundary values of $\lambda$ and will describe reality poorly when $\lambda$ is near the boundary. Moreover, results in Figure 1 suggest asymptotic effects may have little to do with the excellent performance of $\hat{P}(Y)$.

3 Numerical study of size accuracy

There has been no systematic numerical assessment of the accuracy of bootstrap methods for discrete data in the literature - only simulations at isolated parameter values.
Here I describe the key features of a numerical study to investigate the size accuracy of first order and second order tests for discrete models.

**Exact calculation.** For simple models, both the bootstrap P-value and the size of any test can be calculated exactly. This involves computing the test statistic $T(y)$ for all possible data sets $y$ and computing $(1,2)$ at a fine grid of values of $\lambda$. Figure 1 displayed just such an exact calculation. It is worth noting that for models for $d$-dimensional data, the cardinality of the sample space $N = O(m^d)$ and the number of data sets for which standard first order test statistics $T(y)$ are within a normal deviation range is $O(Nm^{-1/2})$. So for models of dimension $d \geq 2$ the discreteness of the distribution of $T$ will be of equal or smaller magnitude than the $O(m^{-3/2})$ effects we might like to detect. Computing the exact size $a_2(\lambda)$ of the bootstrap test involves $O(m^2)$ evaluations of $T(y)$. Our first numerical study will be $d = 2$ dimensional and our second is for logistic regression of dimension $d = 3$.

**Global measures of accuracy.** Size accuracy $a_j(\lambda)$ depends on $\lambda$ and can be summarised in various ways. Let $e_j(\alpha, \lambda) = a_j(\lambda)/\alpha - 1$ be the proportional error for $j = 1, 2$. The most obvious global measure is

$$E_j(\alpha) = (1 - 2\epsilon)^{-1} \int_{\epsilon}^{1-\epsilon} |e_j(\alpha, \lambda)| d\lambda$$

where $\epsilon$ is some small number chosen to avoid the measure being dominated by boundary effects. A second sensible global measure is

$$E_j^*(\alpha) = |\sup_{\lambda} e_j(\alpha, \lambda)|$$

which is sensitive to the worst violation of nominal size, and will be dominated by boundary effects. This measure is consistent with theory of Bickel and Doksum (1997) and Rohmel and Mansmann (1999) on the necessity of accounting for worst case parameter values in frequentist inference. Note in particular that using a test whose exact
size is zero at the boundary will not be harshly judged. A third measure of a similar nature is

\[ \tilde{E}_j(\alpha) = \int_0^1 I_{\{\epsilon_j(\alpha,\lambda)>0\}} e_j(\alpha, \lambda) d\lambda \]

which integrates all violations of the nominal size.

**Detecting asymptotic rates.** We will calculate these error measures \( E_{jm} \) as the sample sizes are scaled up by a factor \( m \) and want to investigate asymptotic rates. Let us suppose that

\[ \bar{E}_{mj}(\alpha) = b_{1j} m^{-1/2} + b_{2j} m^{-1} + b_{3j} m^{-3/2} \]  

(4)

A small value of \( b_{kj}/b_{k+1,j} \) indicates that the \( m^{-k/2} \) term is vanishing for the \( j \)th order test. Thus we will expect to see \( b_{1j}/b_{2j} \) become negligible for the \( j = 2 \)nd order bootstrap tests but not for the \( j = 1 \)st order test. We will be able to test the \( O(m^{-3/2}) \) assertion in (3) by looking at \( b_{22}/b_{32} \). A second cruder method of assessing the asymptotic rate is to regress \( \log(\tilde{E}_{mj}) \) on \( \log m \).

## 4 Non-inferiority test.

Consider testing the alternative hypothesis that \( \psi = \log(p_1/p_0) > 1 - \delta \) from two binomial samples, for some pre-chosen non-inferiority margin \( \delta \), often 0.1. Two basic statistics to be investigated are the score statistic of Chan (1998) and the signed LR statistic, exact formula for which are available in the working paper Lloyd (2010). The nuisance parameter is taken as \( \lambda = p_0 \) and the restricted ML estimator is obtained by solving a quadratic (Miettinen & Nurminen, 1985). For \( \delta = 0 \), the tests statistics are symmetric and first order methods enjoy second order accuracy. Therefore, for this study I chose \( \delta = 0.1 \).

Starting sample sizes \((n_0, n_1) = (7, 3)\) and \((4, 6)\), I calculated all \( N = (n_0+1)(n_1+1) \) possible values of the test statistic and the exact size \( a_1(\lambda) \) at an even grid of 101 values of \( \lambda \). I then scaled up the basic samples sizes by a factor \( m \) up to \( m = 225 \), resulting
in cardinality $N = 1,217,251$. I also calculated all possible values of the bootstrap P-value $\tilde{P}(Y)$ up to $m = 50$. For this case, $N = 60,501$ but computing all bootstrap P-values involves $O(N^2)$ computations.

![Figure 2](image)

Figure 2: Left. For the LR test, log-log plot of $\bar{E}_{1m}(0.025)$ (circle) and $E_{1m}^*(0.025)$ (cross) versus $m$. Line has slope -1/2. Right. For the bootstrap LR test, log-log plot of $\bar{E}_{2m}(0.025)$ (circle) and $E_{2m}^*(0.025)$ (cross) versus $m$. Line has slope -1.

Figure 2 shows some sample results for the LR statistic, starting sample sizes $(n_0, n_1) = (7, 3)$ and nominal size $\alpha = 0.025$. The left panel plots mean error $\bar{E}_{1m}$ and worst error $E_{1m}^*$ against $m$ on the log-log scale. The worst errors (cross symbol) are extremely large, decreasing from around 100% to around 10%, even for the largest sample sizes ($m = 225$). It is not surprising to see these errors persist since they are largely determined by boundary behaviour. Mean error is considerably smaller and decreases at an estimated rate $O(m^{-0.47})$.

The right panel is for the bootstrap based test. Mean errors $\bar{E}_{2m}$ decrease at an estimated rate $O(m^{-1.01})$. Fitting the model (4) we estimate $b_1/b_2 = 0.01$ which suggest the first order term has vanished but $b_2/b_3 = 1.72$ suggesting that the second term has not. The worst errors $E_{2m}^*$ are an order of magnitude smaller than the mean error.
Similar results (not presented) hold for the integrated violation measure $\tilde{E}_{2m}$.

Some of the points at the right of these plots took hours to compute. The "noise" is due to points entering and leaving the rejection set as sample size changes, which depends on the vagaries of integer arithmetic and is hard to predict. The structure in some of the errors is due to the same process.

Table 1: Error rates of standard and bootstrap P-values. For each model, three figures described in the text summarise the asymptotics. The fourth figure $\tilde{E}_{20}$ is the extrapolated error rate when $m = 20$. Upper section is for mean absolute error, $\bar{E}_m$. Lower section is for worst error $E^*_m$.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>First order P-values</th>
<th>bootstrap P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(n_1, n_2)$</td>
<td>$T$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>(7,3) score</td>
<td>.025</td>
<td>6.09</td>
</tr>
<tr>
<td>(7,3) score</td>
<td>.05</td>
<td>0.69</td>
</tr>
<tr>
<td>(7,3) LR</td>
<td>.025</td>
<td>0.64</td>
</tr>
<tr>
<td>(7,3) LR</td>
<td>.05</td>
<td>1.14</td>
</tr>
<tr>
<td>(4,6) score</td>
<td>.025</td>
<td>0.42</td>
</tr>
<tr>
<td>(4,6) score</td>
<td>.05</td>
<td>0.66</td>
</tr>
<tr>
<td>(4,6) LR</td>
<td>.025</td>
<td>0.26</td>
</tr>
<tr>
<td>(4,6) LR</td>
<td>.05</td>
<td>0.77</td>
</tr>
<tr>
<td>(7,3) score</td>
<td>.025</td>
<td>1.05</td>
</tr>
<tr>
<td>(7,3) score</td>
<td>.05</td>
<td>0.36</td>
</tr>
<tr>
<td>(7,3) LR</td>
<td>.025</td>
<td>0.05</td>
</tr>
<tr>
<td>(7,3) LR</td>
<td>.05</td>
<td>0.11</td>
</tr>
<tr>
<td>(4,6) score</td>
<td>.025</td>
<td>0.08</td>
</tr>
<tr>
<td>(4,6) score</td>
<td>.05</td>
<td>0.04</td>
</tr>
<tr>
<td>(4,6) LR</td>
<td>.025</td>
<td>0.12</td>
</tr>
<tr>
<td>(4,6) LR</td>
<td>.05</td>
<td>2.42</td>
</tr>
</tbody>
</table>

Summary results are presented in Table 1. For instance, row 3 summarise the plots in Figure 2. Column 8 suggests that the $O(m^{-1/2})$ does indeed vanish for the bootstrap P-values but that the $O(m^{-1})$ term does not. Asymptotic descriptors of $E^*_m(\alpha)$ given in the lower section of Table 1 are less clear. Dependence on $m$ is very erratic and no clear asymptotic pattern emerges; indeed there is no reason to suppose that worst error would follow the standard asymptotic rate. The columns labeled $\hat{E}_{20}$ list trend error rates when $m = 20$ from the log-log regression model. Looking at row 3 in the bottom
section, worst error is reduced from 87.5% to 3.4%. Essentially, a spike in the size $a(\lambda)$ is removed while the mean error is also slightly reduced (from 7.7% to 6.9%). However, much of the mean error of the bootstrap is the conservative error at the boundary.

5 Logistic regression.

A model of dimension 3 is studied with sample sizes $n = (3, 4, 5)$ scaled up by $m = 1, ..., 10$, with covariate $x = (-1, 0, 1)$ and model

$$p(x_i) = \frac{\lambda e^{\psi x_i}}{1 - \lambda + \lambda e^{\psi x_i}}.$$ 

The nuisance parameter $\lambda$ represents the value of $p(0)$ when $x_i = 0$ and ranges over $[0, 1]$. We test the alternative $\psi > 0$. Again, all computations of the test size $a_j(\lambda)$ are exact; no simulations are involved.

Table 2: **Mean absolute error rates of standard and bootstrap P-values.** The model is a logistic regression with sample sizes $(3, 4, 5)$ scaled up by a factor $m=1, ..., 10$. For each experiment, there are three summary figures as described in the text.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>First order P-values</th>
<th>Bootstrap P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b_1/b_2$ $b_2/b_3$</td>
<td>rate $\gamma$</td>
</tr>
<tr>
<td></td>
<td>$b_1/b_2$ $b_2/b_3$</td>
<td>rate $\gamma$</td>
</tr>
<tr>
<td>lr 05</td>
<td>3.69 0.15 -0.51 0.13</td>
<td>4.18 -0.82</td>
</tr>
<tr>
<td>lr 025</td>
<td>0.79 0.90 -0.55 0.08</td>
<td>1.50 -0.81</td>
</tr>
<tr>
<td>wls 05</td>
<td>3.02 0.54 -0.49 0.05</td>
<td>2.05 -0.87</td>
</tr>
<tr>
<td>wls 025</td>
<td>0.35 1.23 -0.46 0.05</td>
<td>2.03 -0.85</td>
</tr>
</tbody>
</table>

The computational burden of evaluating the bootstrap P-values increases as $m^6$. With $m = 10$ there are $N = 127,551$ points in the sample space. The first test statistic is the LR whose computational complexity becomes problematic when it must be computed for all possible data sets. A quick alternative is a weighted least squares version of the LR where null maximum likelihood estimates of mean values (i.e. the expected values) are replaced by weighted least squares on the empirical logits, with one iteration on the weights. This statistic is asymptotically equivalent to the LR.
statistic but can be conveniently computed on the whole sample space simultaneously since they are just linear transformations.

Table 1 presents summary statistics that address the asymptotics. Raw data plots of $\hat{E}_m$ and $\hat{E}_m^*$ are available in Lloyd(2010). There is again evidence that the $m^{-1/2}$ term is vanishing but no evidence that the $m^{-1}$ term is vanishing. In each case, the best estimate of the slope for the bootstrap P-values is somewhat less than -1, but this appears to be a limitation of the computations only proceeding to $m = 10$.

![Log-log plot of $E_{1m}^*(0.05)$ for LR P-value (cross) and $E_{2m}^*(0.05)$ for bootstrap version (circle)
Centre. Log-log plot of $E_{1m}^*(0.05)$ for WLS-LR P-value (cross) and $E_{2m}^*(0.05)$ for bootstrap version (circle)
Right. Pr($P \leq 0.05; p_0$) versus $p_0$ for the LR and bootstrap(bold) LR P-values with sample sizes (15, 20, 25).](image)

Figure 3: Left. Log-log plot of $E_{1m}^*(0.05)$ for LR P-value (cross) and $E_{2m}^*(0.05)$ for bootstrap version (circle) Centre. Log-log plot of $E_{1m}^*(0.05)$ for WLS-LR P-value (cross) and $E_{2m}^*(0.05)$ for bootstrap version (circle) Right. Pr($P \leq 0.05; p_0$) versus $p_0$ for the LR and bootstrap(bold) LR P-values with sample sizes (15, 20, 25).

As with the non-inferiority test, worst error $E^*$ tells a different story to mean error. In Figure 3 there is no clear tendency for worst error to decrease with sample size up to $m = 10$. What is clear is that bootstrap P-values have worst error an order of magnitude smaller than standard P-values. The right plot shows the probability profile Pr($P \leq 0.05; p_0$) versus $p_0$ for the LR and bootstrap LR based P-values for $m = 5$. This plot is typical of other cases.
6 Conclusion

I have conducted an intensive numerical study of the accuracy of standard first order tests and bootstrap tests across a fine range of parameter values and sample sizes. For the first time, we can see what nature says about the bootstrap rather than asymptotic theory.

We find that bootstrap P-values have extremely good size accuracy for the discrete models investigated, particularly when size accuracy is measured by worst error. It also emerges that size accuracy improves at rate $O(m^{-1})$ rather than $O(m^{-3/2})$. However, the most stark conclusion from the numerical data is that the good performance of bootstrap is non-asymptotic in nature; bootstrap performs much better than standard methods even at very small sample sizes. Existing extensive theory on asymptotics simply does not explain the observed behaviour. Three distinct non-asymptotic arguments that do explain the performance have been developed by the author but will be reported elsewhere.

Computing bootstrap P-values for larger models can be implemented using a version of importance sampling ideas. The algorithm also allows the full significance profile $S(t, \lambda)$ to be conveniently approximated, even for large sample sizes. Details are in Lloyd (2012).
References.


