How close are alternative bootstrap P-values?

Chris Lloyd, Melbourne Business School
How close are alternative bootstrap P-values?

BY CHRIS J. LLOYD
Melbourne Business School, Carlton, 3053, AUSTRALIA
c.lloyd@mbs.edu

SUMMARY

The bootstrap P-value based on a test statistic $T$ is the exact tail probability of the observed value $t$, with the null maximum likelihood estimate of the nuisance parameters substituted. This P-value is known to lead to tests whose size is closer to nominal asymptotically than the first order test. One issue that has not been addressed is whether bootstrap might reduce the impact of the choice of basic test statistic. It is shown that bootstrap P-values based on different first order statistics differ to second order i.e to $O(m^{-1})$ where $m$ is a measure of sample size. Just as importantly, it is shown numerically that this asymptotic rate can underestimate how close alternative bootstrap P-values are for small sample sizes.

Some Key Words: nuisance parameters; second order asymptotics; bootstrap

1 Introduction

Suppose we have a model $\pi(y; \psi, \lambda)$ and are interested in testing a null value $\psi_0$ using a test statistic $T$, larger values of which are considered more hostile to the null. The exact significance profile of the observed value $t = T(y)$ is

$$S(t, \lambda) := \Pr(T(Y) \geq t; \psi_0, \lambda).$$

Since this depends on $\lambda$ it is not available for inference. One approach is to approximate the distribution of $T$ by a distribution that does not depend on $\lambda$, often normal or chi-squared. This will obviously only be appropriate if the dependence of $S(t, \lambda)$ on $\lambda$ is slight. An obvious alternative approach is to replace $\lambda$ by its estimate $\hat{\lambda}_0$ under the null and quote the parametric bootstrap P-value $S(T(y), \hat{\lambda}_0(y))$. 
Example. The main issues are introduced with an example of Berger & Boos (1994). In a clinical trial, \( x_1 = 14 \) out of \( n_1 = 47 \) or 29.8\% of patients assigned the treatment survived while \( x_0 = 48 \) out of \( n_0 = 283 \) or 17\% of patients assigned the placebo survived. Under the null hypothesis of no treatment effect, the common survival rate is estimated by \( \hat{\lambda} = 62/330 = 18.8\% \) and the observed difference \( \hat{\psi} = 12.8\% \) in survival rates has null standard error of 6.15\%. The score statistic for testing \( \psi = 0 \) is the ratio \( 12.8\%/6.15\% = 2.085 \), which we label \( T_1 \). An alternative statistic is the likelihood root (LR) which will be labeled \( T_2 \) and here equals 1.983. Using the standard normal distributions to approximate the two P-values gives 0.0185 for \( t_1 \) and 0.0237 for \( t_2 \).

![Figure 1: Exact significance of approximate P-value (solid) and bootstrap P-value (dashed). Left. Statistic is score statistic. Right. Statistic is likelihood root. Vertical line is null estimate of \( \lambda \).](image)

The exact significance of \( t_1 = 2.085 \) is displayed in the left panel of Figure 1 with the approximate P-value of 0.0185 as a horizontal line. The approximation is poor except near \( \lambda = 0.5 \) and the strong dependence on \( \lambda \) means that no \( \lambda \)-free approximation will suffice. The vertical line is the null estimate \( \hat{\lambda}_0 = 18.8\% \) of \( \lambda \) from which we can read off the bootstrap P-value 0.0233. The right panel is for the statistic \( T_2 \) for which the bootstrap P-value is also 0.0233. Even though the two significance profiles are quite different, the bootstrap P-value are virtually identical.
Is it not at all uncommon for first-order P-values based on alternative statistics to give inferentially different results. Bootstrap P-value have very good frequentist properties, see Lee and Young (2005) and Lloyd (2009). Just as importantly though, it turns out that bootstrap reduces the difference between P-values based on alternative statistics such as LR, score and Wald statistics.

The plan of the paper is as follows. Section 2 gives a brief review of the main known asymptotic results concerning bootstrap P-values. In Section 3, the main result of the paper is proven. Section 4 presents a numerical study suggesting that actual performance may be much better than the asymptotic rate suggests.

2 Background on existing theory

This section gives a brief non-technical review of the main existing results on parametric bootstrap P-values. These all concern the size accuracy of the tests generated by a chosen statistic \( T \). There is another larger literature on non-parametric bootstrap which is not the focus of this paper.

Henceforth all probabilities are understood to be at the null \( \psi = \psi_0 \) and we drop \( \psi_0 \) from the notation. Throughout we use \( m \) as a measure of sample size. The bootstrap P-value estimates \( S(t, \lambda) \) at the true value by substituting an estimate \( \hat{\lambda} \) which differs from \( \lambda \) by \( O_p(m^{-1/2}) \). One might expect this to incur error of \( O_p(m^{-1/2}) \) but it is smaller if \( S(t, \lambda) \) can be asymptotically approximated by a \( \lambda \)-free distribution such as normal. Under standard conditions typically referred to as local asymptotic normality, \( S(t, \lambda) \) is expanded in powers of \( m^{1/2} \) as

\[
S(t, \lambda) = 1 - \Phi(t) + e_1(t, \lambda)m^{-1/2} + O(m^{-1})
\]

where \( e_1(t, \lambda) \) is a smooth \( O(1) \) function of \( \lambda \). It follows that

\[
S(t, \hat{\lambda}) - S(t, \lambda) = \left\{ e_1(t, \hat{\lambda}) - e_1(t, \lambda) \right\} m^{-1/2} + O_p(m^{-1})
\]

and since \( e_1(t, \hat{\lambda}) - e_1(t, \lambda) = O_p(m^{-1/2}) \) it follows that \( S(t, \hat{\lambda}) - S(t, \lambda) \) is \( O_p(m^{-1}) \).
This argument is by Beran (1988) and works for any $\sqrt{m}$-consistent estimator of $\lambda$.

More recent asymptotic results recommend using the restricted maximum likelihood (ML) estimator $\hat{\lambda}_0$ under the null. For continuous models, DiCiccio et al. (2001) then showed that $S(t, \hat{\lambda}) - S(t, \lambda)$ is $O_p(m^{-3/2})$ when $T$ is the LR statistic, extending results of DiCiccio and Stern (1994).

While it is desirable that $S(t, \hat{\lambda})$ is close to $S(t, \lambda)$ the more pertinent issue is the statistical properties of the bootstrap P-value in, and of, itself i.e. how close it is to uniformly distributed. Lee and Young (2005) provided an answer, not only for the LR statistic but for any statistic $T$ which differs from it by $O_p(m^{-1/2})$. They show that for any $\alpha \in [0,1]$

$$\Pr(S(T, \hat{\lambda}_0) \leq \alpha; \psi_0, \lambda) = \alpha + O \left(m^{-3/2}\right)$$

(2) for any $\lambda$ for which an Edgeworth expansion (1) of the distribution of $T(Y)$ is valid. Their result and those previously cited rely on the 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.

Nor is it necessarily true in discrete models. DiCiccio and Young (2008) and Fraser and Rousseau (2008) suggest inferential errors will be $O(m^{-1})$ for discrete models, without clarifying exactly why the previously cited results fail. Detailed numerical work in Lloyd (2009) however establishes without doubt that the error is $O(m^{-1})$ for the models considered there. While this means there is no apparent asymptotic justification for using the restricted ML estimator of $\lambda$, exhaustive numerical investigation of several discrete examples in Lloyd (2008,2009) showed that $S(t, \hat{\lambda}_0)$ has spectacularly good properties.

Most recently, results by DiCiccio and Young (2008) and also Fraser and Rousseau (2008) show that, for canonical parameters of continuous exponential families, the bootstrap P-value approximates the exact conditional P-value also to $O(m^{-3/2})$. This is the same order of error as achieved by competing analytic methods such as $r^*$, see Davison, Fraser and Reid (2006), Brazzale, Davison and Reid (2007).
All these quoted results concern the accuracy of the bootstrap P-value $S(t, \hat{\lambda}_0)$ based on a given test statistic $T$. Different test statistics could still lead to quite different inferences, each of which may be highly accurate. In the next section we study the extent to which bootstrap resolves the disagreement between inferences based on alternative test statistics.

3 Main result

For many statistical problems, the LR, score and Wald statistics can lead to practically different results, at least for some data sets. Under the standard asymptotic framework, the approximate P-values based on these alternative pivots differ by $O_p(m^{-1/2})$. Can bootstrap resolve this? To state the question formally, if $T_1$ and $T_2$ differ by $O_p(m^{-1/2})$ then how close are the bootstrap P-values asymptotically?

An informal argument suggests that bootstrap P-values will not be any closer asymptotically than first order P-values. The argument is that if two test statistics differ by $O(m^{-1/2})$ then their true significance functions will differ by the same order, and the bootstrap will simply result in highly accurate estimates of these different significance functions. This informal argument turns out to be wrong, see equation (7) below.

**Main result.** Suppose that $T_1$ and $T_2$ are asymptotic equivalents of the likelihood ratio statistic and both admit an expansion of their distribution function around a non-degenerate limit distribution as given in (3). Denote the survivor function of $T_j$ by $S_j(t; \lambda)$ and $\hat{\lambda}$ any $\sqrt{m}$-consistent estimator of $\lambda$. Then

$$S_1(T_1; \hat{\lambda}) - S_2(T_2; \hat{\lambda}) = O_p(m^{-1}).$$

**Proof.** Suppose that $T$ has null survivor function $S(t; \lambda)$ that admits the asymptotic expansion

$$S(t; \lambda) = \Phi(-t) + m^{-1/2}e_1(\lambda, t) + m^{-1}e_2(\lambda, t) + O(m^{-3/2})$$

(3)
where $e_j(\lambda, t) = O(1)$. Substituting $\hat{\lambda}$ for $\lambda$ we have

\[
S(t; \hat{\lambda}) = \Phi(-t) + m^{-1/2}e_1(\hat{\lambda}, t) + m^{-1}e_2(\hat{\lambda}, t) + O_p(m^{-3/2})
\]

\[
= \Phi(-t) + m^{-1/2}e_1(\lambda, t) + m^{-1/2}(\hat{\lambda} - \lambda)e_1'(t, \lambda) + m^{-1}e_2(\hat{\lambda}, t) + O_p(m^{-3/2})
\]

\[
= S(t; \lambda) + m^{-1}k(t, \lambda) + O_p(m^{-3/2}) \quad (4)
\]

where $k(t, \lambda) = m^{1/2}e_1'(t, \lambda) + e_2(\lambda, t)$ is $O_p(1)$.

Suppose that $T_j$ are asymptotic equivalents of the likelihood ratio statistic. One of these is the Wald statistic which we will denote by $T_0 = m^{-1/2}(\hat{\theta} - \theta)/\sigma(\theta)$ where $\sigma(\theta)$ is the appropriate element of the inverse Fisher information matrix. There is nothing special about $T_0$ however it is customary to expand other statistics in powers of $(\hat{\theta} - \theta)$ which a easily expressed in terms of $T_0$. Then all $T_j$ can be written as

\[
T_j = T_0 + m^{-1/2}\gamma_j(\lambda)T_0^2 + O(m^{-1}) \quad (5)
\]

for some $\gamma_j(\lambda) = O_p(1)$. We henceforth use subscript $j$ for quantities associated with $T_j$ and its expansions. Assuming each $T_j$ admits an expansion of form (3), we apply (4) to statistic $T_j$ evaluated at $t = T_j$ which gives

\[
S_j(T_j; \hat{\lambda}) = S_j(T_j; \lambda) + m^{-1}k_j(T_j, \lambda) + O_p(m^{-3/2}).
\]

The left hand side is nothing but the parametric bootstrap P-value based on $T_j$. Comparing two bootstrap P-values based on $T_i$ and $T_j$ we have

\[
S_j(T_j; \hat{\lambda}) - S_i(T_i; \hat{\lambda}) = S_j(T_j; \lambda) - S_i(T_i; \lambda) + O_p(m^{-1}) \quad (6)
\]

Cox and Reid (1987) have shown that, if $T_j = T_i + m^{-1/2}\Delta$ for some $O_p(1)$ random variable $\Delta$ then the survivor functions satisfy

\[
S_j(t; \lambda) = S_i(t; \lambda) - m^{-1/2}f_i(t, \lambda)E(\Delta|T_i = t) + O(m^{-1}).
\]

They require that $T_i$ and $T_j$ be continuous and several standard regularity conditions on the joint distribution of $T_i$ and $\Delta$, mainly finiteness of conditional moments, third
moments and some second order derivatives. The proof makes it clear that the result also holds for expansions of the form \( T_j = T_i + m^{-1/2} \Delta + O_p(m^{-1}) \) as well as for discrete data so long as the support of the statistic \( T_j \) has jumps of order no larger than \( O(m^{-1}) \). This is true when \( T_j \) is a standardised test statistic based on data of dimensions at least 2.

Applying their result to (5) where \( i = 0 \) and \( \Delta = \gamma_j(\lambda)T_0^2 \) we have

\[
S_j(t; \lambda) = S_0(t; \lambda) - m^{-1/2} f_0(t, \lambda) \gamma_j(\lambda) E(T_0^2 | T_0 = t) + O(m^{-1})
\]

Evaluating this at the random point \( t = T_j = T_0 + m^{-1/2} \Delta \) we have

\[
S_j(T_j; \lambda) = S_0(T_0 + m^{-1/2} \Delta; \lambda) - m^{-1/2} f_0(T_j; \lambda) \gamma_j(\lambda)(T_0 + m^{-1/2} \Delta)^2 + O_p(m^{-1})
\]

Finally, substituting this back into (6) it follows that

\[
S_j(T_j; \hat{\lambda}) - S_0(T_0; \hat{\lambda}) = S_j(T_j; \lambda) - S_0(T_0; \lambda) + O_p(m^{-1}) = O_p(m^{-1}).
\]

Consequently, the bootstrap P-values based on asymptotically equivalent test statistics \( T_j \) differ by \( O_p(m^{-1}) \). This result holds for any consistent estimator \( \hat{\lambda} \). I have not been able to establish a lower error rate for the restricted ML estimator. However, the method of proof here suggests that, even if the error term in (6) could be reduced to \( O(m^{-3/2}) \), the error rate in (7) cannot.

**Remark.**

It is important to realise that the asymptotic results just proven may under-estimate the closeness of two P-values, full analysis of which depends on the joint distribution of \((T_1, T_2)\), not just on how close \( T_1 \) is to \( T_2 \). For instance, suppose that \( T_{1m} = T_{2m} + \epsilon_m \) for
a non-random sequence $\epsilon_m$, even one that does not converge to zero. Then the upper tail sets $\{T_{1m}(Y) \geq T_{1m}(y)\}$ and $\{T_{2m}(Y) \geq T_{2m}(y)\}$ are identical for every $m$ and so the bootstrap P-values based on $T_{1m}$ and $T_{2m}$ will automatically agree. Potentially then, alternative bootstrap P-values could be much closer than the asymptotic error rate. The non-asymptotic effects of gross error removal may be more important than the asymptotic adjustment of the random differences.

4 Numerical investigation

We now present numerical evidence that the resolution of alternative bootstrap inferences may be dominated by non-asymptotic rather than asymptotic effects. The model is discrete which allows exact calculation of all quantities involved. Moreover, the non-asymptotic effects that emerge from this study seem more clear for discrete models than continuous models.

We consider a very simple model, namely testing $p_1 > 0.9p_0$ from binomial data. The results presented show how the differences between alternative P-values behave as the sample sizes are scaled up by a multiplier $m = 2, 3, \ldots, 50$, starting with denominators $(7, 3)$ for the two binomials. Similar results are obtained for alternative denominators, as well as for testing a difference of probabilities rather than the ratio. The alternative test statistics $T_1$ and $T_2$ are the score statistic of Chan (1998) and the likelihood root.

We consider two measures of the distance between two P-values. The first is the mean absolute difference under the null. This requires specification of a nuisance parameter which we will take to be $\lambda = \sqrt{p_1 p_0}$ and set to equal 0.5. Results are very similar for other values of $\lambda$. The left panel of Figure 2 shows this mean absolute deviation plotted against the multiplier $m$ on a log-log scale. Open circles are for the first order P-values, filled circles for the bootstrap P-values. Estimating the slopes of these plots to the nearest 0.1 gives $-0.5$ for the first order P-values and $-1.1$ for the
bootstrap P-values. What is more striking than the asymptotic rate is the fact that, even for small samples, bootstrap reduces the differences between alternative P-values by an order of magnitude.

![Graph showing differences between score and likelihood root P-values for testing \( p_1 > 0.9p_0 \).](image)

**Figure 2**: Differences between score and likelihood root P-values for testing \( p_1 > 0.9p_0 \).

*Left.* Mean absolute difference with \( \lambda = 0.5 \). *Right.* Measure of difference is one minus Kendall’s \( \tau \)-statistic. Open circles are for first order P-values, solid circles for bootstrap P-values.

The second method of comparing two P-values is to look at pairs of data sets \((s, t)\) for which \( P_1(s) < P_2(s) \) but \( P_1(t) > P_2(t) \). This means that there is a size \( \alpha \) test for which \( P_1 \) but not \( P_2 \) will reject the null from data \( s \) while \( P_2 \) but not \( P_1 \) will reject the null from data \( t \). If \( P_1 \) and \( P_2 \) are related by a monotone transform then there will be no such discordant pairs. A measure of distance between two P-values is the proportion of discordant pairs over the whole sample space, which is a simple transform of Kendall’s \( \tau \)-statistic. The right plot displays results for this alternative measure. The slope of the plot for the first order P-values is \(-0.5\) and for the bootstrap P-values is \(-1.4\).

The reason for the quicker convergence rate is that this measure is a simple average over the sample space, giving equal weight to probable and improbable points alike. The results then reflect that bootstrap is even more successful at resolving alternative
P-values for points that are improbable under the null i.e. outcomes for which the P-values are small.

References.


