Some non-asymptotic properties of parametric bootstrap p-values

Chris Lloyd, Melbourne Business School
Abstract. The bootstrap P-value is the exact tail probability of a test statistic, calculated assuming the nuisance parameter equals the null maximum likelihood (ML) estimate. For discrete data, bootstrap P-values perform amazingly well even for small samples, even as standard first order methods perform surprisingly poorly. Why is this? Detailed numerical calculations in Lloyd (2012a) strongly suggest that the good performance of bootstrap is not explained by asymptotics. In this paper, I establish several desirable non-asymptotic properties of bootstrap P-values. The most important of these is that bootstrap will correct ‘bad’ ordering of the sample space which leads to a more pivotal distribution. Most of the arguments only hold for discrete models and when the null ML estimate is used.

Some Key Words: bootstrap; exact test; nuisance parameters;

1 Introduction

Consider a parametric model \( \pi(y; \psi, \lambda) \) for discrete data \( Y \in \mathcal{Y} \) that depends on \((\psi, \lambda)\). We want to test the alternative \( \psi > \psi_0 \) using a test statistic \( T \), larger values of which lead to rejection of the null. The significance profile of the observed value \( t = T(y) \) is

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

Since \( S(t, \lambda) \) depends on \( \lambda \) it is not available for inference. One approach is to approximate the distribution of \( T \) by a known continuous distribution \( H \) that does not depend on \( \lambda \), often normal or \( \chi^2 \). This generates the approximate P-value \( Q(y) = 1 - H(T(y)) \).

This can only be appropriate if the dependence of \( S(t, \lambda) \) on \( \lambda \) is slight. A more modern
approach is to replace \( \lambda \) by the maximum likelihood (ML) estimate \( \hat{\lambda}_0 \) under the null and quote the parametric bootstrap P-value \( \hat{Q}(y) = S(T(y), \hat{\lambda}_0(y)) \).

This paper concerns frequentist properties of the statistic \( \hat{Q}(Y) \), which turn out to be extremely good.

**Example.** The key issues are introduced with an example of Berger & Boos (1994). In a clinical trial, \( y_1 = 14 \) out of \( n_1 = 47 \) or \( 29.8\% \) of patients assigned the treatment survived while \( y_0 = 48 \) out of \( n_0 = 283 \) or \( 17\% \) of patients assigned the placebo survived. We want to test the null hypothesis of no treatment effect. Two standard test statistics are the score statistic \( T_1 \), which here equals 2.085, and the signed likelihood ratio statistic \( T_2 \), which here equals 1.983.

![Figure 1: Exact significance of approximate P-value (solid) and bootstrap P-value (bold). Left. Score statistics. Right. Signed likelihood ratio statistic. Vertical line is null estimate of \( \lambda \).](image)

These are not small sample sizes so we would expect standard approximate methods to work well. The significance profile of \( t_1 = 2.085 \) is in the left panel of Figure 1 as a non-bolded curve and the approximate P-value \( Q_1(y) = 0.0185 \) based on the normal approximation as a horizontal line. The approximation is poor except near \( \lambda = 0.5 \). The vertical line is the null estimate \( \hat{\lambda}_0 = 62/330=18.8\% \) of \( \lambda \) from which we read
off the bootstrap P-value \( \hat{Q}_1(y) = 0.0233 \). The bold curve is the significance profile of the bootstrap P-value, which is to say it is the lower tail probability of the statistic \( \hat{Q}_1(Y) \) as a function of \( \lambda \). The curve is almost uniformly close to the quoted value of 0.0233, represented as a bold horizontal line. Note also that, while there is inevitably some breakdown at the boundary, the curve converges to zero at these endpoints. So the quoted value of 0.0233 is close to the true significance for most values of \( \lambda \) and is appropriately conservative when the method breaks down. The right panel is for \( t_2 = 2.085 \) for which the bootstrap P-value is \( \hat{Q}_2(y) = 0.0234 \). The exact significance of this bootstrap P-value is represented again as a bold curve and, as was the case for \( T_1 \), is uniformly close to nominal. Critically important from a strict frequentist point of view, the supremum of the significance profile is extremely close to nominal.

A large numerical study reported in Lloyd (2012a) shows that this behavior is typical in discrete models; that bootstrap P-values are consistently an order of magnitude more accurate than first order P-values, for even small sample size. It is surprising that this outstanding level of performance is not better known. The purpose of this paper is to explain this good performance. We give non-asymptotic arguments for three key conclusions, namely that bootstrap (a) makes the statistic more pivotal i.e. the significance profile flatter, (b) makes the statistic more uniform i.e. the significance profile is flat at the correct level, (c) behaves appropriately near the boundary.

2 Parametric bootstrap P-values

We start with a test statistic \( T \) for testing the alternative \( \psi > \psi_0 \) in the presence of nuisance parameters \( \lambda \). This could in principle be any statistic, but later we will restrict attention to likelihood based test statistics. After observing \( t = T(y) \), the exact significance is \( S(t, \lambda) \) as defined in (1). To quote a P-value we need to replace \( \lambda \) with some surrogate. What surrogate for \( \lambda \) should be used?

In general, consider replacing \( \lambda \) by a data-based surrogate \( g(Y, \psi_0) \) which gives the
general P-value $P_g(Y) := S(T(Y), g(Y, \psi_0))$. If $\lambda$ were known then we would use the ideal P-value $P_\lambda(Y) = S(T, \lambda)$. We will measure the size violation of a P-value by

$$d_\alpha(P(Y)) = \sup_\lambda \Pr\{P(Y) \leq \alpha\} - \alpha.$$  

We also let $\mu(X) = \inf_x \{x + \Pr(X \geq x)\}$ measure the size of a random variable $X$ defined on $[0, 1]$. When $X$ is stochastically larger, $\mu(X)$ is numerically larger. Several other natural properties also hold (results available from author).

**Result 1.** Let $\Delta_g(Y, \lambda) = |P_g(Y) - P_\lambda(Y)|$ measure the stochastic difference between the ideal P-value $P_\lambda(Y)$ and the general surrogate P-value $P_g(Y)$. Then

$$d_\alpha(P_g(Y)) \leq \sup_\lambda \mu(\Delta_g|\lambda).$$

Note that the right side does not depend on $\alpha$. This result means that to control the worst possible size violation, the best surrogate $g(Y, \psi_0)$ to use is one for which the magnitude of difference between $P_g(Y)$ and $P_\lambda(Y)$ is stochastically smallest, for all values of $\lambda$. This strongly suggests the restricted ML estimator $\hat{\lambda}_0$ since it is not only asymptotically optimal but also respects restrictions on the range of $\lambda$ for fixed $\psi_0$, unlike the unrestricted ML estimator. We henceforth call $\hat{Q}(Y) = S(T(Y), \hat{\lambda}_0(Y))$ the bootstrap P-value.

Beran (1988) shows that using any estimator $\hat{\lambda}$ that differs from $\lambda$ by $O_p(m^{-1/2})$, the difference between $S(T, \hat{\lambda})$ and $S(T, \lambda)$ is $O_p(m^{-1})$, where $m$ is a measure of sample size. On the other hand, building on results of DiCiccio and Stern (1994) and DiCiccio, Martin and Stern (2001)) for continuous models, Lee and Young (2005) showed that the distribution function of $\hat{Q}(Y) = S(T(Y), \hat{\lambda}_0(Y))$ differs from uniform by terms of order $O(m^{-3/2})$ for any $\lambda$ for which an Edgeworth expansion of the distribution of $T(Y)$ is valid. Their result is not true of the unrestricted ML estimator, and is not uniform in $\lambda$. Nor is it proven for discrete models. Fraser and Rousseau (2008) suggest that, for discrete models, inferential errors will be $O(m^{-1})$ so that from the asymptotic point of view any estimator of $\lambda$ may be used in the discrete case. Yet numerical results in
Lloyd (2008) showed that using $\hat{\lambda}_0$ makes a big difference, even for small samples.

In summary, there does not seem to be an asymptotic explanation for the very good small sample performance of bootstrap P-values for discrete models. Though there are good heuristic arguments for using the restricted ML estimator $\hat{\lambda}_0$, asymptotic accuracy rates seems to be $O(m^{-1})$ for any sensible estimator.

3 Likelihood based statistics and rankings

The results to be reported are non asymptotic, but instead focus on how standard and bootstrap test statistics rank points in the sample space. The arguments rely on the sample space $\{y_1, ..., y_N\}$ being discrete so that it can be indexed. We denote $T(y_j)$, $\hat{\lambda}_0(y_j)$ and $\hat{\psi}(y_j)$ by $t_j$, $\hat{\lambda}_j$ and $\hat{\psi}_j$ respectively. We also denote

$$\hat{\pi}_{j,k} = \pi(y_j; \hat{\psi}_0, \hat{\lambda}_k)$$

noting that $\hat{\pi}_{j,j}$ is the null likelihood of $y_j$. Therefore, $\hat{\pi}_{j,k}$ will be very small when $\hat{\lambda}_k$ is far from $\hat{\lambda}_j$ and takes its maximum value of $\hat{\pi}_{j,j}$ when $k = j$.

Initially, let us consider Wald type statistics of the form $t_j = (\hat{\psi}_j - \psi_0)/\sigma(\psi_0, \hat{\lambda}_j)$ where $\sigma^2(\psi, \lambda)$ is the so-called asymptotic variance of $\hat{\psi}$. Firstly, this statistic depends on the data only through $(\hat{\psi}_j, \hat{\lambda}_j)$. For exponential family models, this is also true of the likelihood ratio and score statistics. Secondly, for fixed $\hat{\lambda}_j$, $t_j$ is monotone increasing in $\hat{\psi}_j$. Again, this is also true of the likelihood ratio and score statistics (results available from author). This would seem to be an essential logical property of any statistic; surely if $\hat{\lambda}_j = \hat{\lambda}_k$ and $\hat{\psi}_j > \hat{\psi}_k$ then $y_j$ provides more evidence than $y_k$ that $\psi > \psi_0$.

Thirdly, let us write $\hat{\pi}_{j,j}$ in the alternative form $\pi((\hat{\psi}_j, \hat{\lambda}_j); \psi_0, \hat{\lambda}_j)$ where the data $y_j$ has been represented in its sufficient form $(\hat{\psi}_j, \hat{\lambda}_j)$. For fixed $\hat{\lambda}_j$ this will be a decreasing function of $\hat{\psi}_j$ when $\hat{\psi}_j > \psi_0$. Put simple, the null probability of data gets smaller as $\hat{\psi}$ exceeds $\psi_0$ by a larger and larger amount. Indeed, this null maximised probability was suggested as an ordering criterion itself by Freeman & Halton (1951). Combining this
with the second property, it follows that for fixed $\hat{\lambda}_j$, $t_j$ will be a decreasing function of $\hat{\pi}_{jj}$.

In summary, we can now described the main pattern of how any "sensible" statistic will rank the sample space. This pattern applies in particular to likelihood based statistics. Looking at the right panel of Figure 2 as a generic sample space, all sensible statistics agree on the ranking of data sets within a vertical column (for which $\hat{\lambda}_j$ is constant). The ranking will be monotone increasing in $\hat{\psi}_j$ or equivalently monotone decreasing in $\hat{\pi}_{jj}$. On the other hand, for two points $y_j$ and $y_k$ with $\hat{\lambda}_j$ far from $\hat{\lambda}_k$ (i.e. well separated horizontally), different statistics can disagree in their ranking. I will call this the natural ordering principle. These assertions are not true if $\hat{\lambda}_j$ is the unrestricted ML estimator. They are numerically confirmed in the following example.

**Example: Logistic regression.** For a fixed value of $\hat{\lambda}_0(y_j)$, it is claimed that a higher value of $\hat{\psi}(y_j)$ implies a lower null probability $\hat{\pi}_{jj}$. For canonical discrete models, this seems to be exactly true. The point is illustrated on the logistic regression model with binomial denominators $n = (5, 6, 7, 8)$ and covariates $x = (-3, -1, 1, 3)$ with slope $\psi$ and intercept $\lambda$. Sufficiency implies that inference should only be based on $V = (\sum x_i Y_i, \sum Y_i)$. For all of the 441 distinct values of $V$, I computed the signed likelihood ratio statistic $T(v_j)$ for testing $\psi > 0$, the null estimate of the slope $\hat{\lambda}_0(v_j)$ and the null probability $\hat{\pi}_{jj}$. The left panel in Figure 2 shows that larger values for the test statistic are associated with points whose null maximised probability is smaller. Points with a common value of $\hat{\lambda}_0 = \sum Y_i/26$ are joined by lines which reveals that, for fixed values of $\hat{\lambda}_0$, the relationship is monotone decreasing.

The right hand plot displays the sufficiency reduced sample space. Within a column, $\hat{\lambda}_j$ is constant and both the LR statistic and $\hat{\psi}_j$ monotonically increase while $\hat{\pi}_{jj}$ decreases. The dark points comprise those for which the null probability $\hat{\pi}_{jj}$ is less than 0.004, this value being chosen to roughly match the tail set $T > 2$. Tails sets based on Wald or Score would have the same ordering within columns but could disagree.
about the ranking of points well separated horizontally.

![Figure 2](image)

Figure 2: Left. LR statistic versus null probability for a logistic regression. Points for which $\hat{\lambda}_0$ is constant are joined by a line. Right. Sufficiency reduced sample space, with points highlighted for which null probability is less than 0.004.

4 Bootstrap pivotality and mis-ranking

We will show how bootstrap makes a P-value more pivotal by re-ranking points in the sample space. We demonstrate this by imagining a P-value $Q(Y)$ that is almost pivotal, deliberately vandalising it so that it is non-pivotal and then showing that applying bootstrap returns it to being almost pivotal. This process only works if the original and vandalised P-values respect the natural ordering principles explained in the previous section. We first need a definition of “almost pivotal.”

**Definition.** A statistic $Q(Y)$ is $\delta$-pivotal at $y$ if for some non-decreasing function $h$

$$
\Pr\{Q(Y) \leq Q(y); \psi_0, \lambda\} = h(Q(y)) + e(Q(y), \lambda)
$$

where $|e(y, \lambda)| \leq \delta$ for all $\lambda \in \Lambda_\delta$. The statistic is called $\delta$-uniform if $h$ is the identity.

The subset $\Lambda_\delta$ is necessary so as to exclude the boundary of the parameter space
where the tail probability will typically diverge, else no statistic could be almost pivotal. Boundary behaviour is studied explicitly in a later section.

Let $Q(Y) = 1 - H(T(Y))$ for some test statistic $T(Y)$, where $T$ follows the natural ordering principles. Denote $q_j = Q(y_j)$. Since $q_j$ is a monotone decreasing function of $t_j$, for fixed $\hat{\lambda}_0$, $q_j$ will be a decreasing function of $\hat{\psi}_j$ and increasing function of $\hat{\pi}_{j,j}$. It is convenient to henceforth index the sample space $\{y_1, ..., y_N\}$ according to the size of $q_j = Q(y_j)$ so that if $j < k$ then $q_j \leq q_k$. For tied values, the index can be defined in any consistent manner.

Suppose now that $Q(Y)$ is very close to pivotal. Consider two points $y_L$ and $y_m$ such that $q_L < q_m$. This means that $Q$ ranks $y_L$ as more hostile to the null than $y_m$. We vandalise $Q(Y)$ to $Q^*(Y)$ by reversing this ranking. This is simply achieved by defining

$$Q^*(y_j) = \begin{cases} q_L - \epsilon & j = m \\ q_j & j \neq m \end{cases}$$

where $\epsilon < q_L - q_{L-1}$. Denoting $q^*_j = Q^*(y_j)$, it is obvious that $q^*_L > q^*_m$ contradicting the ranking of $Q$. The profile of $Q^*(Y)$ at $y_L$ is

$$\Pr(Q^*(Y) \leq q^*_L; \lambda) = \Pr(Q(Y) \leq q_L; \lambda) + \pi(y_m, \lambda)$$

(3)

which is not approximately pivotal, as it includes an additional highly non-pivotal term $\pi(y_m, \lambda)$ which adds a peak of size $\hat{\pi}_{m,m}$ at $\hat{\lambda}_m$. Such spikes are visible in either panel of Figure 1. The next result shows that bootstrap restores the ‘correct’ ordering and returns the statistic to $\delta$-pivotality.

**Result 2.** Suppose that $Q(Y)$ is $\delta$-pivotal at $y_L$, $q_L < q_m$ and define $Q^*(Y)$ as above. Suppose

- $\hat{\pi}_{m,m} > \hat{\pi}_{L,L} + 3\delta$
- $\hat{\lambda}_m$ is far enough from $\hat{\lambda}_L$ that $\hat{\pi}_{L,M} < \delta$.

Let $\hat{Q}^*$ be the bootstrap P-value based on $Q^*$. Then, $\hat{Q}^*(y_L) < \hat{Q}^*(y_m)$ and as a consequence $\hat{Q}^*(Y)$ is $\delta$-pivotal at $y_L$. 

8
Condition (a) means $Q(Y)$ satisfies the logical ordering property that $q_j$ be increasing in $\hat{\pi}_{jj}$. Condition (b) says that the two points are well separated in their corresponding estimate of $\lambda$ so that the vandalised $Q^*$, in disagreeing with $Q$, does not violate the natural ordering principle.

It is easy to check numerically that the bootstrap does not correct gross errors in ordering, for instance swapping the ranking of two points with the same value of $\hat{\lambda}_0$. A tedious algebraic argument not presented here confirms that bootstrap will not correct such mis-ranking. Instead the significance of both points is seriously reduced, power is lost and pivotality compromised. Bootstrap then will not convert a bad P-value into a good one. It refines approximate likelihood based P-values.

5 Bootstrap induced uniformity

In this section we want to show that an almost pivotal but non-uniform P-value will be made almost uniform by applying the bootstrap once. The uniformity issue has been addressed by other authors, for instance Beran (1988) but the treatment below is non-asymptotic and emphasises total dependence on $\lambda$.

Bootstrap can automatically correct gross errors in a P-value $Q$. Indeed, the bootstrap P-value based on $Q(y) + c$ is the same as that based on $Q(y)$ for any constant $c$. If $Q$ is exactly pivotal then the proof is easy. The null distribution function $h$ of $Q$ does not depend on $\lambda$ and so the profile $\Pr(Q(Y) \leq Q(y); \lambda) = h(P(y))$. Hence $\hat{Q}(Y) = h(Q(Y))$ which has a discretised uniform distribution. So bootstrap adjusts the distribution to exact uniformity in one step, regardless of its distribution. The problem with this argument is that for discrete models no statistic is exactly pivotal, since tail probabilities are sums of polynomials. A more refined argument follows that allows formally for approximate pivotality and uniformity.

Result 3. Suppose that $Q(Y)$ is $\delta$-pivotal. Let $\hat{\lambda}_\delta$ be the null estimator of $\lambda$ restricted to $\Lambda_\delta$. Then $\hat{Q}(Y) = \Pr(Q(Y) \leq Q(y); \hat{\lambda}_\delta)$ is $\delta$-uniform.
We see in Figure 1 (and in the numerical results in Lloyd (2012a)) that bootstrap produces a P-value that is very close to pivotal - even more so than the argument in the previous section might suggest. Result 3 says that to the extent that bootstrap is successful in producing pivotality, it must also produce uniformity. It is not possible for the profile to be flat but at the wrong level.

6 Boundary correction

Discrete models typically become degenerate when the parameters are on the boundary. The model can be degenerate at a single point or on a lower dimensional subset of the sample space. Some may argue that properties of a statistical procedure on the boundary are practically irrelevant if it is post-hoc unlikely that the parameter is near the boundary. To argue this is to ignore the fundamentals of frequentist inference, which requires good performance for all parameter values. One cannot argue post-hoc that some parameter values do not matter.

We show that bootstrap P-values will be appropriately conservative at the boundary i.e. understates the actual significance. Throughout this section, denote \( Q(y_{\text{obs}}) = q_{\text{obs}} \) and \( \hat{Q}(y_{\text{obs}}) = \hat{q}_{\text{obs}} \). Suppose that for some value \( \lambda^* \) the model is degenerate at \( y^* \) i.e. \( \Pr(Y = y^*; \lambda^*) = 1 \). This is the case for instance when testing equality of two probabilities when the common value \( \lambda \) equals 0 or 1, or for a logistic regression where the baseline probability of success is 0 or 1. This can be seen in both panels of Figure 1 where the true significance converges to zero.

**Result 4A.** Suppose that \( \hat{q}_{\text{obs}} < 1 \). Then the true significance of \( \hat{Q}(y_{\text{obs}}) \) equals zero at \( \lambda^* \) i.e.

\[
\Pr(\hat{Q}(Y) \leq \hat{q}_{\text{obs}}; \lambda^*) = 0
\]

The proof in the appendix requires that the estimator of \( \lambda \) be the restricted ML estimator. Suppose next that for some value \( \lambda^* \), all probability is concentrated on a
lower dimensional subset $Y^*$ of $Y$ - typically a boundary set. Let $y^*$ minimise $\hat{Q}(y)$ over $y \in Y^*$, regardless of whether it is unique. In the proof of Result 4A it emerged that necessarily $\hat{q}_{\text{obs}} < \hat{Q}(y^*)$. For the following result this condition must be imposed.

**Result 4B.** Suppose that $\hat{q}_{\text{obs}} < \hat{Q}(y^*)$. Then the true significance $\hat{Q}(y_{\text{obs}})$ equals zero at $\lambda^*$ i.e.

$$\Pr(\hat{Q}(Y) \leq \hat{q}_{\text{obs}}; \lambda^*) = 0$$

**Remark.** The reason the result is of interest is that it will often be the case that the significance profile of $Q(y_{\text{obs}})$ is not zero at the boundary, for instance if $q_{\text{obs}} \geq Q(y^*)$. In this case,

$$\Pr(Q(Y) \leq q_{\text{obs}}; \lambda) \geq \Pr(y^*; \lambda)$$

and when $\lambda = \lambda^*$ the probability of $y^*$ can easily be quite large since the model is concentrated on the subspace $Y^*$.

**Example.** Consider testing the null hypothesis $p_1 = 0.95 p_0$ from binomial data, with $Q(Y)$ obtained from a standard normal approximation to the likelihood root. The nuisance parameter is taken as $\lambda = p_0$. When $\lambda = \lambda^* = 1$ the model is concentrated on the boundary set $Y^* = \{(y_0, y_1) : y_0 = n_0\}$. The significance profiles in Figure 3 are for the data $\hat{p}_1 = 7/25 = 0.35$ based on the LR statistic and $\hat{p}_0 = 9/20 = 0.36$. There is modest evidence that $p_1 > 0.95 p_0$ as measured by $q_{\text{obs}} = 0.095$ or $\hat{q}_{\text{obs}} = 0.102$. Referring to result 4B, the most significant point of the boundary set is $y^* = (25, 20)$ for which $Q(y^*) = 0.076$ while $\hat{Q}(y^*) = 0.358$. Thus $q_{\text{obs}} \leq \hat{Q}(y^*)$ so result 4B applies. Since $q_{\text{obs}} \geq Q(y^*)$ the profile of $Q(y_{\text{obs}})$ at $\lambda^* = 1$ will be greater than $\Pr(y^*; \lambda^*) = 0.358$. This is displayed in the left panel. The profile of the bootstrap P-value is bolded and equals 0 at the boundary. The attentive reader will have noted that in this example $q_{\text{obs}} = 0.095 > 0.076 = Q(y^*)$ but $\hat{q}_{\text{obs}} = 0.102 < 0.358 = \hat{Q}(y^*)$. There is something wrong with the way $Q$ ranks the sample space, which is corrected by bootstrap. The fact that the bootstrap and ordinary P-values give contradictory orderings of $y_{\text{obs}}$ and
\( y^* \) is not especially rare. There are 57 of the 546 points in the sample space of this example with this property.

![Profile](image)

Figure 3: **Left.** Profile of LR P-value and bootstrap version (bold). Restricted ML estimate and actual P-value (dotted lines)** **Right.** Profiles after artificially adjusting statistic at \((0,0)\).

To illustrate result 4A, observe that when \( \lambda = \lambda^* = 0 \) the model is concentrated at \( y^* = (0,0) \). To better reveal the effect of the bootstrap, the LR based P-value for the data set \((0,0)\) has been adjusted to a value less than 0.095 so that \( Q(y^*) < q_{\text{obs}} \). So we are beginning with an approximate P-value which is a deliberately vandalised version of the likelihood root based P-value. Consequently, the profile diverges to 1.0 at the left hand endpoint as displayed in the right panel. The bootstrap removes this anomaly. In fact, the profile of this bootstrap P-value is identical to that of the bootstrap based on the unvandalised P-value.

## 7 Conclusion

This paper concerns the empirical fact that bootstrap P-values from discrete models have exact significance that is very close to the quoted value apart from the boundary
where they are appropriately conservative.

I have investigated these features without recourse to asymptotics, but rather by looking at how bootstrap modifies the way the sample space is ordered. The results require that the nuisance parameters $\lambda$ be replaced by the restricted ML estimator, not just any estimator.

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

References.


Appendices

Proof of result 1. Recall the definition of $\Delta_g(Y) = |P_g(Y) - P_\lambda(Y)|$ in the text. We express $\Pr(P_g(Y) \leq \alpha; \lambda)$ as

$$
\Pr(\{P_g(Y) \leq \alpha\} \cap \{\Delta_g < x\}; \lambda)
+ \Pr(\{P_g(Y) \leq \alpha\} \cap \{\Delta_g \geq x\}; \lambda)
\leq \Pr(P_\lambda(Y) \leq \alpha + x; \lambda) + \Pr(\Delta_g \geq x; \lambda)
\leq \alpha + x + \Pr(\Delta_g \geq x; \lambda)
$$

where the last inequality follows from a standard result for the distribution function transform of a random variable. This is true for arbitrary choice of $x$ so it follows that for any $\lambda$

$$
\Pr(P_g(Y) \leq \alpha|\lambda) \leq \alpha + \mu(\Delta_g|\lambda).
$$

Taking supremum of both sides gives the result.

Proof of result 2. For simplicity, we suppose that $q_{L-1} < q_L$. In the event that $q_{L-1} = q_L$, the proof below follows unchanged except that the index $L - 1$ is replaced with the highest index $J$ such that $Q(y_J) < Q(y_L)$. The tail sets of $Q^*(Y)$ at $y_m$ and $y_L$ are

$$
\{Q^*(Y) \leq q^*_m\} = \{Q(Y) \leq q_{L-1}\} \cup \{y_m\}
\{Q^*(Y) \leq q^*_L\} = \{Q(Y) \leq q_L\} \cup \{y_m\}
$$

Therefore the bootstrap transformed values of $Q^*(Y)$ are

$$
\hat{q}^*_m := \hat{Q}^*(y_L) = F(q_L, \hat{\lambda}_L) + \hat{\pi}_{m,L},
\hat{q}^*_m := \hat{Q}^*(y_m) = F(q_{L-1}, \hat{\lambda}_m) + \hat{\pi}_{m,m}
$$

and the difference between these is

$$
\hat{q}^*_m - \hat{q}^*_L = \hat{\pi}_{m,m} - \hat{\pi}_{m,L} - (F(q_L, \hat{\lambda}_L) - F(q_{L-1}, \hat{\lambda}_m))
$$

Now

$$
F(q_L, \hat{\lambda}_L) - F(q_{L-1}, \hat{\lambda}_m) = F(q_{L-1}, \hat{\lambda}_L) + \pi(y_L, \hat{\lambda}_L) - F(q_{L-1}, \hat{\lambda}_m)
= q_{L-1} + e(q_{L-1}, \hat{\lambda}_L) + \hat{\pi}_{L,L} - q_{L-1} - e(q_{L-1}, \hat{\lambda}_m)
\leq \hat{\pi}_{L,L} + 2\delta
$$
since \(|e(q_{L-1}, \lambda)| < \delta\) for any \(\lambda\). Hence
\[
\hat{q}^*_m - \hat{q}^*_L \geq \hat{\pi}_{m,m} - \hat{\pi}_{L,L} - \hat{\pi}_{m,L} - 2\delta
\]
\[
= (\hat{\pi}_{m,m} - \hat{\pi}_{L,L} - 3\delta) + (\delta - \hat{\pi}_{m,L})
\]
which is positive by conditions (a) and (b) of the result.

By supposition, \(\hat{\pi}_{m,m} > \hat{\pi}_{L,L} + 2\delta\) and so \(\hat{Q}^*\) ranks \(y_m\) as less significant than \(y_L\) i.e. \(\hat{q}^*_m > \hat{q}^*_L\). Consequently, \(\hat{Q}(Y)\) is \(\delta\)-pivotal at \(y_L\) as a consequence of the assumption that \(Q(Y)\) is. More formally, the profile of \(\hat{Q}^*(Y)\) at \(y_L\) is

\[
\Pr(\hat{Q}^*(Y) \leq \hat{Q}^*(y_L)) = \Pr(\hat{Q}^*(Y) \leq Q^*(y_L)) - \pi(y_m, \lambda)
\]
\[
= F(q_L, \lambda) + \pi(y_m, \lambda) - \pi(y_m, \lambda)
\]
from (3), which equals the pivotal quantity \(F(q_L, \lambda)\).

**Proof of result 3.** Denote the distribution function of the approximate P-value \(Q(Y)\) by \(F(q, \lambda)\) which, by hypothesis, can be expressed as \(h(q) + e(q, \lambda)\) with \(e()\) small. Denote \(u_j = h(q_j)\) and note that the \(u_j\) order the sample space identically to \(q_j\). Note also that \(U(Y) = h(Q(Y))\) is \(\delta\)-uniform. The bootstrap P-value \(\hat{Q}\) when \(y = y_j\) is

\[
F(q_j, \hat{\lambda}_\delta(y_j)) = u_j + e(q_j, \hat{\lambda}_\delta(y_j))
\]
where \(\hat{\lambda}_\delta\) is the restricted ML estimator restricted to \(\Lambda_\delta\). So \(\hat{q}_j\) and \(u_j\) differ by \(e_j = e(u_j, \hat{\lambda}_\delta(y_j))\) which is less than \(\delta\). This suggest the approximate uniformity of \(\hat{Q}(Y)\) at \(y_j\). We now prove this. Consider the tail sets

\[
S_{Qj} = \{y : \hat{Q}(y) \leq \hat{q}_j\}, \quad S_{Uj} = \{y : U(y) \leq u_j\}
\]
keeping in mind that sample points are indexed by \(U\) which may not be the same ordering as \(\hat{Q}\). Let \(L = \max\{l : y_l \in S_{Qj} \cap \bar{S}_{Uj}\}\) and note that \(y_l \notin S_{Qj} \cap \bar{S}_{Uj}\) for \(l \leq j\). Therefore

\[
\Pr(S_{Qj} \cap \bar{S}_{Uj}; \lambda) \leq \Pr(u_j < U(Y) \leq u_L)
\]
\[
= u_L + e(u_L, \lambda) - u_j - e(u_j, \lambda)
\]
\[
\leq u_L - u_j + 2\delta
\]
Now $y_L \in S_Q \cap \bar{S}_{U_j}$ if, and only if, $\hat{q}_L < \hat{q}_j$ and $u_L > u_j$. Hence where the second inequality follows from the assumption that $Q$ is $\delta$-pivotal.

$$u_L - u_j = (u_L - \hat{q}_L) - (u_j - \hat{q}_j) - (\hat{q}_j - \hat{q}_L)$$

$$= \delta + \delta - (\hat{q}_j - \hat{q}_L)$$

$$\leq 2\delta$$

Therefore

$$\Pr(S_Q \cap \bar{S}_{U_j}; \lambda) \leq u_L - u_j + 2\delta \leq 4\delta.$$
Proof of result 4A. For any P-value, bootstrapped or otherwise, the profile at \( \lambda^* \) must equal either 0 or 1, depending on whether or not \( y^* \in \{ y : P(y) \leq p_{\text{obs}} \} \). This is determined by whether or not \( p_{\text{obs}} \geq P(y^*) \). We will show that for an bootstrap P-value \( \hat{P}(y^*) = 1 \). Note first that under the assumption that the model becomes degenerate at \( y^* \), the null probability \( \Pr(Y = y^* ; \lambda) \) is maximised with respect to \( \lambda \) at \( \lambda^* \) and so the restricted ML estimator \( \hat{\lambda}_0(y^*) = \lambda^* \). Therefore

\[
\hat{P}(y^*) = \Pr(P(Y) \leq P(y^*); \hat{\lambda}(y^*)) \geq \Pr(Y = y^*; \hat{\lambda}(y^*)) = \Pr(Y = y^*; \lambda^*) = 1.
\]

With the assumption that \( \hat{p}_{\text{obs}} < 1 \) this implies that \( \hat{p}_{\text{obs}} < \hat{P}(y^*) \) and therefore that

\[
y^* \notin \{ y : \hat{P}(y) \leq \hat{p}_{\text{obs}} \}.
\]

Hence the probability of this set at \( \lambda = \lambda^* \) is zero.

Proof of result 4B. Since \( \hat{p}_{\text{obs}} < \hat{P}(y^*) \leq \hat{P}(y) \) for all \( y \in \mathcal{Y}^* \), it follows that

\[
\{ \hat{P}(Y) \leq \hat{p}_{\text{obs}} \} \subset \hat{\mathcal{Y}}^*
\]

and so

\[
\Pr(\hat{P}(Y) \leq \hat{p}_{\text{obs}}; \lambda) \leq 1 - \Pr(\mathcal{Y}^*; \lambda)
\]

and the right hand side is zero when \( \lambda = \lambda^* \).