Electronic Journal of Statistics
Vol. 6 (2012) 2449–2462
ISSN: 1935-7524
DOI: 10.1214/12-EJS751

Some non-asymptotic properties of parametric bootstrap P-values in discrete models

Chris J. Lloyd

University of Melbourne, Melbourne Business School, 200 Leicester Street, 3052, Australia
e-mail: c.lloyd@mbs.edu

Abstract: For discrete data especially, standard P-values can misreport the true significance, even for moderately large sample sizes. The bootstrap P-value is the exact tail probability of an appropriate test statistic, calculated assuming the nuisance parameter equals the null maximum likelihood (ML) estimate. For basic discrete models and standard test statistics, bootstrap P-values have been found to be extremely close to uniformly distributed under the null ([1]). Detailed numerical results reported there suggest that this phenomenon is not explained by asymptotics. In this paper, we identify several desirable non-asymptotic properties of bootstrap P-values and provide arguments for why bootstrap P-values are so close to exact. The most important of these is that bootstrap will correct 'incorrect' ordering of the sample space and that this leads to a more pivotal distribution. Most of these arguments only hold for discrete models and when the null ML estimate is used.

AMS 2000 subject classifications: Primary 62F03; secondary 62F05.
Keywords and phrases: Bootstrap, exact test, exact P-value, nuisance parameters, size accuracy.

Received October 2012.

Contents

1. Introduction

Consider a parametric model $\pi(y;\psi,\lambda)$ for a discrete data vector $Y\in\mathcal{Y}$. We want to test the composite null hypothesis $\psi=\psi_0$ against the alternative $\psi>\psi_0$

in the presence of a nuisance parameter (vector) λ. To this end, we choose a test statistic T that measures deviation of the data from what would be expected under the null.

Commonly T is the likelihood ratio statistic, the score statistic or the Wald statistic, which are all members of the power divergence family of Cressie and Read ([2]). Other choices include the Euclidean distance statistic, see [3]. Different test statistics will be more powerful at detecting different alternatives. The precise choice of T is not critical to main ideas to be presented, though we will impose some quite weak conditions on T in Section 3.

To execute the test we need the null distribution of T. We will suppose that T has been defined so that larger values lead to rejection of the null hypothesis. This paper is about assigning the correct significance to the observed value $t = T(y)$. The *significance profile* is

$$S(t, \lambda) := \Pr(T(Y) \geq t; \psi_0, \lambda) = \sum_{y:T(y)\geq t} \pi(y; \psi_0, \lambda). \tag{1.1}$$

If λ were known then this would measure exactly how improbable the observed value is under the null, but since $S(t, \lambda)$ depends on λ it is not available for inference.

The classical solution is to approximate the distribution of T by a known continuous distribution H that does not depend on λ, often normal or χ^2. This generates the approximate P-value $Q(y) = 1-H(T(y))$. This can be appropriate only if the dependence of $S(t, \lambda)$ on λ is slight.

A less crude approach is to acknowledge that the distribution of T depends on λ and to replace λ by its best estimate from the data. In this paper, we use the maximum likelihood (ML) estimate $\hat{\lambda}_0$ under the null. A new result, supporting this choice, is given in Section 2. This generates the P-value $S(T(y), \hat{\lambda}_0(y))$ which we denote by $\hat{Q}(y)$. Note that $\hat{Q}(y)$ is no longer a monotone function of $T(y)$. So even though its construction is based on a particular statistic T, it may no longer be the case that larger values of T lead to rejection of the null. In practice, $\hat{Q}(Y)$ and $Q(Y)$ are highly correlated, and so lower values of $Q(y)$ will almost always lead to lower values of $\hat{Q}(y)$.

This paper concerns the null distribution of $\hat{Q}(Y)$, the so-called bootstrap P-value statistic. The null distribution turns out to be extremely close to uniform, regardless of the value of the nuisance parameters.

Example. The key issues are introduced with an example from [13]. 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.

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 nuisance parameter

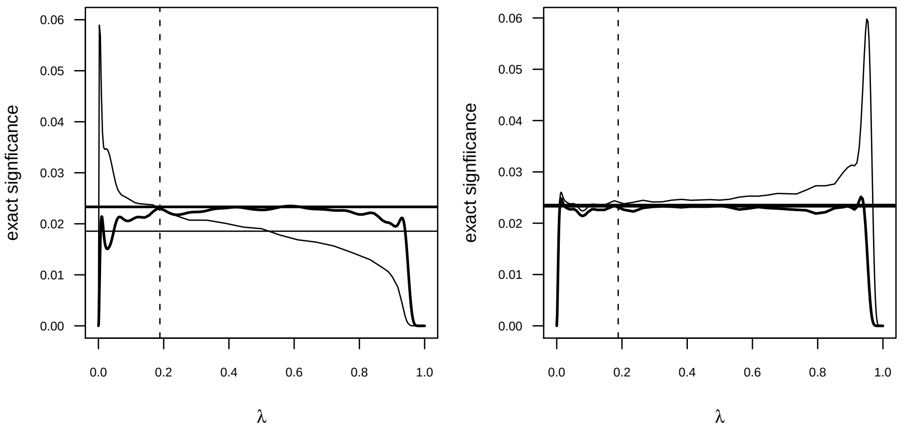

FIG 1. *Exact significance of approximate P-value (solid) and bootstrap P-value (bold).* Left. *Score statistics.* Right. *Signed likelihood ratio statistic. Vertical line is null estimate of* λ.

λ is the common probability of survival. The approximation is poor except near $\lambda = 0.5$. The vertical line is the null estimate $\hat{\lambda}_0 = 62/330$=18.8% of λ 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 λ. 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 λ 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 (see [5]), the supremum of the significance profile is extremely close to nominal.

A large numerical study reported in [1] 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 accuracy is not better known.

The purpose of this paper is to explain this behaviour. We give non-asymptotic arguments for three key features revealed in the above example, 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 null hypothesis $\psi = \psi_0$ against the alternative $\psi > \psi_0$ in the presence of nuisance parameters λ. This could in principle be any statistic, but we will impose some natural conditions in the next section. After observing $t = T(y)$, the exact significance is $S(t, \lambda)$ as defined in (1.1). To quote a P-value we need to replace λ with some surrogate. What surrogate for λ should be used?

In general, consider replacing λ 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 λ 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 α. 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 the difference between $P_g(Y)$ and $P_\lambda(Y)$ is stochastically smallest, for all values of λ. 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 λ for fixed ψ_0, unlike the unrestricted ML estimator.

A second reason for using $\hat{\lambda}_0$ is invariance to the choice of λ. If $\eta(\psi, \lambda)$ is such that the mapping from (ψ, λ) to (ψ, η) is one-to-one, then the restricted ML estimator $\hat{\eta}_0$ of η satisfies $\hat{\eta}_0 = \eta(\psi_0, \hat{\lambda}_0)$ and as a consequence $\hat{Q}(y)$ will be the same using either parametrisation.

It is shown in [4] that using any estimator $\hat{\lambda}$ that differs from λ 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 in [6] and [7] for continuous models, Lee and Young ([10]) 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 λ 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 λ. Nor is it proven for discrete models. It is suggested in [8] that, for discrete models, inferential errors will be $O(m^{-1})$ so that from the asymptotic point of view any estimator of λ may be used in the discrete case. Yet numerical results in [11] 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 seem 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 tests rank points in the sample space. The arguments rely on the sample space $\mathcal{Y} = \{y_1, y_2, \ldots\}$ being countable 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

$$\hat{\pi}_{j,k} = \pi(y_j; \psi_0, \hat{\lambda}_k)$$

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 $k = j$.

We now impose a very natural conditions on how the test statistic T depends on $\hat{\psi}$ and $\hat{\lambda}_0$. The key idea is that if two data sets lead to the same value of $\hat{\lambda}_0$, then the data set with the higher value of $\hat{\psi}$ provides more evidence that $\psi > \psi_0$. The most transparent example of a test statistic with this property is the Wald statistic. In our index notation this statistic is $t_j = (\hat{\psi}_j - \psi_0)/\sigma(\psi_0, \hat{\lambda}_j)$ where $\sigma^2(\psi, \lambda)$ is the so-called asymptotic variance of $\hat{\psi}$. For fixed $\hat{\lambda}_j$, t_j is monotone increasing in $\hat{\psi}_j$. For canonical exponential family models, this is also true of the likelihood ratio and the score statistics (results available from author). The author has investigated a range of test statistics, including the Euclidean statistic in [3], and hypotheses about non-canonical parameters, and found numerically that the condition above was always satisfied. An example is given below.

The second condition is actually a condition on the model, namely that $\hat{\pi}_{j,j}$ be a decreasing function of $\hat{\psi}_j$ for $\hat{\psi}_j > \psi_0$. Why is this natural? 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 smaller when $\hat{\psi}_j$ is further from ψ_0. Put simply, the further $\hat{\psi}_j$ is from ψ_0, the more improbable it is assuming $\psi = \psi_0$. Indeed, the null maximised probability $\hat{\pi}_{j,j}$ was itself suggested as an ordering criterion in [9].

Combining this second condition with the first condition on t_j, it follows that for fixed $\hat{\lambda}_j$, t_j is increasing in $\hat{\psi}_j$ and also decreasing in $\hat{\pi}_{j,j}$. We will say that T then satisfies the "natural ordering property".

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 $\hat{\pi}_{j,j}$. 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. These assertions are not true if $\hat{\lambda}_j$ is the unrestricted ML estimator. They are numerically confirmed in the following example.

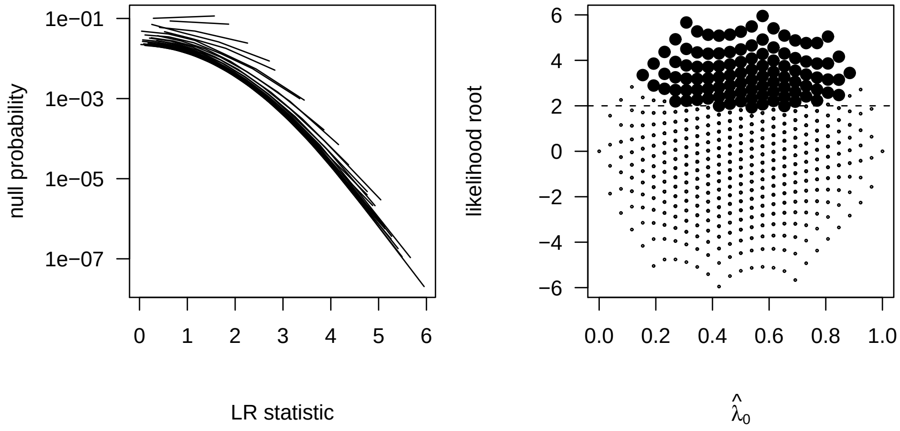

Fig 2. Left. *LR statistic versus null probability for a logistic regression. Points for which $\hat{\lambda}_0$ is constant are joined by a line.* Right. *Sufficiency reduced sample space, with points highlighted for which null probability is less than 0.004.*

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 $\hat{\pi}_{j,j}$. 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 ψ and intercept λ. 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 $t_j = T(v_j)$ for testing $\psi > 0$, the null estimate of the slope $\hat{\lambda}_j = \hat{\lambda}_0(v_j)$ and the null probability $\hat{\pi}_{j,j}$. 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 $\hat{\lambda}_j = \sum Y_i/26$ are joined by lines which reveals that, for fixed values of $\hat{\lambda}_j$, 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 t_j and $\hat{\psi}_j$ monotonically increase while $\hat{\pi}_{j,j}$ decreases. The dark points comprise those for which the null probability $\hat{\pi}_{j,j}$ 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.

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 δ-pivotal at y if for some non-decreasing function h

$$\Pr\{Q(Y) \leq Q(y); \psi_0, \lambda\} = h(Q(y)) + e(Q(y), \lambda) \tag{4.1}$$

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

The subset Λ_δ 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 principle. 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 $\{y_1, \ldots, y_N\}$ 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

$$\Pr(Q^*(Y) \leq q_L^*; \lambda) \quad = \quad \Pr(Q(Y) \leq q_L; \lambda) + \pi(y_m, \lambda) \tag{4.2}$$

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 $\lambda = \hat{\lambda}_m$. 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 δ-pivotality.

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

(a) $\hat{\pi}_{m,m} > \hat{\pi}_{L,L} + 3\delta$
(b) $\hat{\lambda}_m$ 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 δ-pivotal at y_L.

Condition (a) means $Q(Y)$ satisfies the natural ordering property that q_j be increasing in $\hat{\pi}_{j,j}$ (since t_j is decreasing in $\hat{\pi}_{j,j}$). Condition (b) says that the two points are well separated in their corresponding estimate of λ 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 is seriously reduced (i.e. q_j^* is much larger), power is lost and pivotality compromised. Bootstrap then will not convert a bad P-value into a good one. It refines approximate P-values that are based on test statistics that satisfy the natural ordering principle.

5. Bootstrap induced uniformity

In this section we 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 [4] but the treatment below is non-asymptotic and emphasises total dependence on λ.

Bootstrap can automatically correct gross errors in a P-value Q. If Q is exactly pivotal then the proof is easy. The null distribution function h of Q does not depend on λ 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 δ-pivotal. Let $\hat{\lambda}_\delta$ be the null estimator of λ restricted to Λ_δ. Then $\hat{Q}(Y) = \Pr(Q(Y) \leq Q(y); \hat{\lambda}_\delta)$ is δ-uniform.

We can see in Figure 1 (and more generally in the numerical results in [1]) 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 behaviour

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. understate 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 λ^* 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 λ 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 λ^* i.e.

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

The proof in the appendix requires that the estimator of λ be the restricted ML estimator. Suppose next that for some value λ^*, all probability is concentrated on a lower dimensional subset $\mathcal{Y}^*$ of $\mathcal{Y}$ - typically a boundary set. Let y^* minimise $\hat{Q}(y)$ over $y \in \mathcal{Y}^*$, 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 λ^* 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 $\mathcal{Y}^*$.

Example. Consider testing the null hypothesis $p_1 = 0.95p_0$ 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 $\mathcal{Y}^* = \{(y_0, y_1) : y_0 = n_0\}$. The significance profiles in Figure 3 are for the LR statistic, for the data $\hat{p}_1 = 7/25 = 0.35$ and $\hat{p}_0 = 9/20 = 0.36$. There is modest evidence that $p_1 > 0.95p_0$ 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 $\hat{q}_{\text{obs}} \leq \hat{Q}(y^*)$ 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_{obs} and y^* is not especially rare. There are 57 of the 546 points in the sample space of this example with this property.

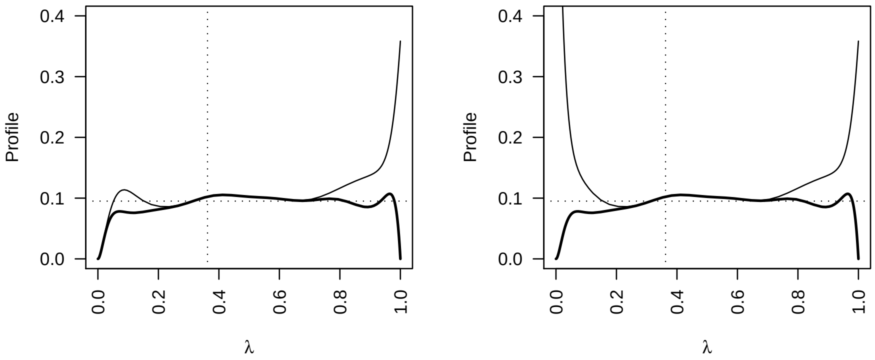

FIG 3. Left. *Profile of LR P-value and bootstrap version (bold). Restricted ML estimate and actual P-value (dotted lines).* Right. *Profiles after artificially adjusting statistic at* $(0, 0)$.

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 unconditional 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 λ be replaced by the restricted ML estimator, not just any estimator, and that the basic test statistics T satisfy some natural properties. The results do not imply that the bootstrap test will be more powerful. Rather, they implies that the bootstrap P-value gives a more honest assessment of the significance.

I have not addressed the controversial issues of conditionality that can arise in discrete data models. In principle, the model $\pi(y; \psi, \lambda)$ mentioned in the first sentence of this paper might be conditional. Provided that the model still involves a nuisance parameter, results 1–4B will apply conditionally. It is worth noting that for most discrete data models it is not possible to eliminate all nuisance parameters by conditioning.

Finally, for the examples in this paper I was able to compute the bootstrap P-values exactly. This become infeasible for models where the sample space $\mathcal{Y}$ is large. Computing bootstrap P-values 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 [12].

8. Appendices

Proof of result 1. Recall the definition of $\Delta_g(Y) = |P_g(Y) - P_\lambda(Y)|$ in the text. measure the difference between the general surrogate P-value and the "exact" P-value. We express $\Pr(P_g(Y) \leq \alpha; \lambda)$ as

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

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 λ

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

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

$$\begin{aligned}\{Q^*(Y) \leq q_m^*\} &= \{Q(Y) \leq q_{L-1}\} \cup \{y_m\}\\ \{Q^*(Y) \leq q_L^*\} &= \{Q(Y) \leq q_L\} \cup \{y_m\}\end{aligned}$$

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

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

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

$$\begin{aligned}F(q_L, \hat{\lambda}_L) - F(q_{L-1}, \hat{\lambda}_m) &= F(q_{L-1}, \hat{\lambda}_L) + \pi(y_L, \hat{\lambda}_L) - F(q_{L-1}, \hat{\lambda}_m)\\ &= q_{L-1} + e(q_{L-1}, \hat{\lambda}_L) + \hat{\pi}_{L,L} - q_{L-1} - e(q_{L-1}, \hat{\lambda}_m)\\ &\leq \hat{\pi}_{L,L} + 2\delta\end{aligned}$$

since $|e(q_{L-1}, \lambda)| < \delta$ for any λ. Hence

$$\begin{aligned}\hat{q}_m^* - \hat{q}_L^* &\geq \hat{\pi}_{m,m} - \hat{\pi}_{L,L} - \hat{\pi}_{m,L} - 2\delta\\ &= (\hat{\pi}_{m,m} - \hat{\pi}_{L,L} - 3\delta) + (\delta - \hat{\pi}_{m,L})\end{aligned}$$

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 that y_L i.e. $\hat{q}^*_m > \hat{q}^*_L$. Consequently, $\hat{Q}(Y)$ is δ-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

$$\begin{aligned}\Pr(\hat{Q}^*(Y) \leq \hat{Q}^*(y_L)) &= \Pr(Q^*(Y) \leq Q^*(y_L)) - \pi(y_m, \lambda) \\ &= F(q_L, \lambda) + \pi(y_m, \lambda) - \pi(y_m, \lambda)\end{aligned}$$

from (4.2), 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 δ-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 Λ_δ. So $\hat{q}_j$ and u_j differ by $e_j = e(u_j, \hat{\lambda}_\delta(y_j))$ which is less than δ. 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

$$\begin{aligned}\Pr(S_{Qj} \cap \bar{S}_{Uj}; \lambda) &\leq \Pr(u_j < U(Y) \leq u_L) \\ &= u_L + e(u_L, \lambda) - u_j - e(u_j, \lambda) \\ &\leq u_L - u_j + 2\delta\end{aligned}$$

where the second inequality follows from the assumption that $\hat{Q}$ is δ-pivotal. Now $y_L \in S_{Qj} \cap \bar{S}_{Uj}$ if, and only if, $\hat{q}_L \leq \hat{q}_j$ and $u_L > u_j$. Hence

$$\begin{aligned}u_L - u_j &= (u_L - \hat{q}_L) - (u_j - \hat{q}_j) - (\hat{q}_j - \hat{q}_L) \\ &= \delta + \delta - (\hat{q}_j - \hat{q}_L) \\ &\leq 2\delta\end{aligned}$$

Therefore

$$\Pr(S_{Qj} \cap \bar{S}_{Uj}; \lambda) \leq u_L - u_j + 2\delta \leq 4\delta.$$

In a very similar manner, let K be the smallest index such that $y_k \in \bar{S}_{Qj} \cap S_{Uj}$ and note that $y_k \notin \bar{S}_{Qj} \cap S_{Uj}$ for $k > j$. Therefore

$$\begin{aligned}\Pr(\bar{S}_{Qj} \cap S_{Uj}; \lambda) &\leq \Pr(u_K < U(Y) \leq u_j) \\ &= u_j + e(u_j, \lambda) - u_K - e(u_K, \lambda) \\ &\leq u_j - u_K + 2\delta.\end{aligned}$$

Now $y_K \in \bar{S}_{Qj} \cap S_{Uj}$ if, and only if, $\hat{q}_K > \hat{q}_j$ and $u_K \leq u_j$. Hence

$$\begin{aligned} u_j - u_K &= (u_j - \hat{q}_j) - (u_K - \hat{q}_K) - (\hat{q}_K - \hat{q}_j) \\ &= \delta + \delta - (\hat{q}_K - \hat{q}_j) \\ &\leq 2\delta \end{aligned}$$

Therefore

$$\Pr(\bar{S}_{Qj} \cap S_{Uj}; \lambda) \leq u_j - u_K + 2\delta \leq 4\delta.$$

Finally we look at the distribution function of $\hat{Q}(Y)$ at q_j and note that

$$\begin{aligned} |\Pr(\hat{Q}(Y) \leq \hat{q}_j); \lambda) - \Pr(U(Y) \leq u_j); \lambda)| &= |\Pr(S_{Qj} \cap \bar{S}_{Uj}; \lambda) - \Pr(\bar{S}_{Qj} \cap S_{Uj}; \lambda)| \\ &< 8\delta \end{aligned}$$

and so

$$\begin{aligned} \Pr(\hat{Q}(Y) \leq \hat{q}_j); \lambda) - \hat{q}_j &\leq |\Pr(\hat{Q}(Y) \leq \hat{q}_j); \lambda) - \Pr(U(Y) \leq u_j); \lambda)| \\ &+ |\Pr(U(Y) \leq u_j); \lambda) - u_j| + |u_j - \hat{q}_j| \\ &\leq 10\delta \end{aligned}$$

so $\hat{Q}(Y)$ is δ-uniform.

Proof of result 4A. For any P-value, bootstrapped or otherwise, the profile at λ^* 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 λ at λ^* 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

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

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 $\lambda = \lambda^*$.

The connection between bootstrap and likelihood can be seen by expressing

$$S(t, \hat{\lambda}_0) = \sum_{y:T(y) \geq t} \pi(y; \psi_0, \hat{\lambda}_0)$$

and noting that each summand $\pi(y;\psi_0,\hat{\lambda}_0)$ is the likelihood of y maximised under the null. Bootstrap P-values are thus sums of null maximised likelihoods. Entry into the tail set is determined by the initial statistic $T(Y)$. If this statistic is itself closely related to the likelihood then there is a match between the ordering and the calibration. This is the heuristic behind the result to be proven below. The argument relies on using the restricted ML estimator $\hat{\lambda}_0$, otherwise there is no connection with null likelihood. It also relies on the model being discrete.

References

[1] Lloyd, C.J. (2013). A numerical investigation of the accuracy of parametric bootstrap for discrete data. *Comp. Statist. Data Anal.* **57**. To appear.

[2] Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. B* **46** 157–214. MR0790631

[3] Perkins, W., Tygert, M. and Ward, R. (2011). Computing the confidence levels for a root-mean-square test of goodness-of-fit *Appl. Math. Comput.* **217** 9072–9084. MR2803971

[4] Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83** 687–697. MR0963796

[5] Rohmel, J. and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority or superiority. *Biometrical Journal* **41** 149–170. MR1693980

[6] DiCiccio, T.J., Martin, M.A. and Stern, S.E. (1984). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canad. J. Statist.* **29** 167–76. MR1834487

[7] DiCiccio, T.J. and Stern, S.E. (1994). Constructing approximate standard normal pivots from signed roots of adjusted likelihood ratio statistics. *Scand. J. Statist.* **21** 447–460. MR1310088

[8] Fraser, D.A.S and Rousseau, J. (2008.) Studentization and deriving accurate P-values. *Biometrika* **95** 1–16. MR2409711

[9] Freeman, G.H. and Halton, J.H. (1951). Note on exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* **38** 141–149. MR0042666

[10] Lee, S.M.S. and Young, G.A. (2005). Parametric bootstrapping with nuisance parameters. *Stat. Prob. Letters* **71** 143–153. MR2126770

[11] Lloyd, C.J. (2008). Exact P-values for discrete models obtained by estimation and maximisation. *Austral. and New Zealand J. Statist.* **50** 329–346. MR2474195

[12] Lloyd, C.J. (2012). Computing Highly Accurate or Exact P-values using Importance Sampling. *Comp. Statist. Data Anal.* **56** 1784–1794. MR2892377

[13] Berger, R.L. and Boos, D.D. (1994). P values maximised over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. MR1294746