To appear in: Maximum Entropy and Bayesian Methods, Eds. K. Hanson, R. Silver, 1995 (in press 1995)

THE BOOTSTRAP IS INCONSISTENT WITH PROBABILITY THEORY

David H. Wolpert
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501 USA (dhw@santafe.edu)

ABSTRACT.
This paper proves that for no prior probability distribution does the bootstrap (BS) distribution equal the predictive distribution, for all Bernoulli trials of some fixed size. It then proves that for no prior will the BS give the same first two moments as the predictive distribution for all size trials. It ends with an investigation of whether the BS can get the variance correct.

1. Introduction

Say we have a set of N data D that is created by IID sampling a distribution f, and a statistic $S(D)$ that assigns a number to that data. Our problem is to use D to estimate the probability of getting statistic value s in N more samples of f.

The vanilla version of the bootstrap procedure (BS) interprets this problem as asking for an estimate of the standard distribution of S, $P(S \mid f, N)$. To create this estimate, first it many times resamples D (with replacement) according to some D-dependent distribution $x(D)$ [1]. In this way it creates many new data sets D'. BS then estimates the standard distribution as the distribution of values of $S(D')$. (In this paper none of the variants of this vanilla version of the BS will be considered.) In particular, the BS can assign an error bar to the observed value of the statistic $S(D)$; its estimate of the standard distribution provides an estimate of the standard error. The power of BS is its wide applicability, the fact that its error bars behave quite "reasonably" (e.g., they often grow as the size of the data shrinks), and the empirical fact that it often gives good estimates of error bars.

This paper investigates whether and when the BS can be exactly correct in its answer to the problem. This question is vacuous if one views the problem of "estimating the probability of getting ... s in N more samples of f" as estimating $P(S \mid f, N)$—the BS will be correct if f is the empirical distribution given by D. However from a Bayesian perspective, in the real world we have knowledge of D and none concerning f, so it is more sensible to view the problem as estimating the predictive distribution $P(S$ of N new data points sampled from the unknown $f \mid D)$ than of estimating $P(S \mid f, N)$. In fact, *as its results are used in practice*, the BS is usually treated as though it produces the predictive distribution.

Accordingly, this paper analyzes whether the BS's distribution for S can equal the predictive distribution for some prior $P(f)$, and in this sense is "consistent with probability theory". The precise scenario investigated is Bernoulli trials, where S is the number of positive events in the sample. Intuitively, the question is of whether the "variance" (and higher moments) in the estimate of S given by the BS can agree with a Bayesian "variance" (and higher moments).

Recursive relationships are used to prove that there is no prior and no $x(D)$ such that the distribution of values of $S(D')$ always equals $P(S \mid D)$ for all D of fixed size N. Next it is shown that there is no prior and no $x(D)$ such the BS and the predictive distribution will always agree on the first two moments of $P(S \mid D)$ for all D of any size. Next the question is investigated of whether just the error bars generated by BS can equal those given by a Bayesian calculation (i.e., of whether the BS's variance can be correct). First a preliminary analysis of going from $x(D)$ to a prior giving the same error bar is presented. Then it is shown that for only certain Bernoulli scenarios are there $x(D)$'s that give the same error bar as the uniform prior's predictive distribution, and those $x(D)$'s are derived.

Of course, none of this means that one should not use the BS, even for Bernoulli trials. Rather it means that *if one is completely sure of one's prior*, then one should not use the BS for Bernoulli trials. How well the BS performs compared to a Bayesian calculation based on an incorrect prior is the subject of future work.

There has previously been some work on a "Bayesian variant of the BS" [2]. However one can argue that that technique is, in its essentials, equivalent to direct Monte Carlo sampling of the predictive distribution. As such, the question of whether it (or slight modifications of it) can give the predictive distribution is mute. Accordingly, in this paper only the conventional non-Bayesian variant of the BS—by far the more popular variant in the BS community—is addressed.

2. Preliminaries

To be more precise, say we have N IID Bernoulli trials (e.g., N flips of the same coin), giving n positives. We want the probability of k positives in the next N Bernoulli trials, or moments of that probability. (That number of positives k is the statistic S, and the n out of N positives is the data D.) Let p be the true probability of a positive in any single trial. Then $P(k \mid n) = \int dp \; P(k \mid p, n) \; P(p \mid n)$.

Now $P(k \mid p, n) = P(k \mid p) = C_k^N \; p^k \; (1-p)^{N-k}$, where $C_j^i \equiv \frac{i!}{j! \; (i-j)!}$. Furthermore, $P(p \mid n) = P(n \mid p) \; P(p) \; / \; P(n)$, which in turn is given by $p^n \; (1-p)^{N-n} \; P(p) \; / \; \int dp \; p^n \; (1-p)^{N-n} P(p)$. Therefore

$$P(k \mid n) = C_k^N \; \frac{\int dp \; p^{n+k} \; (1-p)^{2N-n-k} \; P(p)}{\int dp \; p^n \; (1-p)^{N-n} \; P(p)}. \tag{1}$$

As shorthand, define $P_{i,j} \equiv \int dp \; p^i \; (1-p)^j \; P(p)$, so that Eq. (1) becomes

$$P(k \mid n) = C_k^N \; \frac{P_{n+k, 2N-n-k}}{P_{n,N-n}}. \tag{2}$$

In the BS, one uses an estimator for p based on n and N, $x(n, N)$, and calculates the probability of k given that $p = x(n)$. (Rather than the probability of sampling a particular datum, from now on "$x(.)$" means the probability with which the set of all positives in the data is sampled.) So rather than directly calculate $P(k \mid n)$, in the BS one instead calculates the surrogate $P(k \mid n,\ p = x(n,N)) = C_k^N\ [x(n,N)]^k\ [1 - x(n,N)]^{N-k}$. (As the BS is usually practiced, this calculation is accomplished by Monte Carlo sampling, but that is not important for current purposes.) The question is whether this surrogate can equal $P(k \mid n)$.

The crucial difference between the BS's calculation and direct evaluation of the predictive distribution is that the BS is based on a single (estimate of) p, whereas the predictive distribution averages over all p. This is similar to the distinction between ML-II, where one fixes a hyperparameter to a single value, and the full hierarchical Bayesian approach, in which one averages over that hyperparameter. Since ML-II can be a poor approximation to the hierarchical calculation even when the posterior probability of the hyperparameter is sharply peaked [3], one might suspect that the BS has a difficult time agreeing with the predictive distribution.

3. Disagreement for the full distribution, for some k and n

Evidently the BS distribution will agree with the predictive distribution for all $k \in \{0, \ldots N\}$ and all $n \in \{0, \ldots N\}$ iff the following holds for all such k and n:

$$[x(n)]^k\ [1 - x(n)]^{N-k} = \frac{P_{n+k,2N-n-k}}{P_{n,N-n}}. \tag{3}$$

It turns out that the only estimator that (might) meet the constraint of Eq. (3) is the absurd estimator $\{$all $x(i)$ are the same constant, $x\}$. To see this, write

$$\frac{P_{n+k,2N-n-k}}{P_{n+k\pm 1,2N-n-k\pm(-1)}} = \frac{P_{n+k,2N-n-k}\ /\ P_{n,N-n}}{P_{n+(k\pm 1),2N-n-(k\pm 1)}\ /\ P_{n,N-n}}.$$

Now use Eq. (3):

$$\begin{aligned} \frac{P_{n+k,2N-n-k}}{P_{n+k\pm 1,2N-n-k\pm(-1)}} &= \frac{[x(n)]^k\ [1 - x(n)]^{N-k}}{[x(n)]^{k\pm 1}\ [1 - x(n)]^{N-k\pm(-1)}} \\ &= \{\frac{1 - x(n)}{x(n)}\}^{\pm 1}. \end{aligned} \tag{4}$$

For the positive exponent, this equality must hold for all $n \in \{0, \ldots N\}, k \in \{0, \ldots N - 1\}$ ($k \in \{1, \ldots N\}$ for the negative exponent). In particular, for the positive exponent it must hold for all $n \in \{1, \ldots N\}$ and $k = N - n$ (and similarly for the negative exponent). Therefore for the positive exponent, for all $n \in \{1, \ldots N\}$, ($\{0, \ldots N - 1\}$ for the negative exponent),

$$\frac{P_{N,N}}{P_{N\pm 1,N\pm(-1)}} = \{\frac{1 - x(n)}{x(n)}\}^{\pm 1}.$$

Since the left-hand side is independent of n, by the positive exponents we know that $[1 - x(i)]\ /\ x(i) = [1 - x(j)]\ /\ x(j)$ for all $i, j \in \{1, \ldots N\}$, which means that for fixed

$N, x(n)$ is independent of n for $n \in \{1 \ldots N\}$. Assuming $N > 2$, the negative exponent case then extends this to all $n \in \{0, \ldots N\}$. (As an aside, this extension of the exponents means that $P_{N-1,N+1}\ P_{N+1,N-1} = [P_{N,N}]^2$.)

The immediate conclusion is that the distribution calculated by the BS can not be $P(k \mid n)$ for any reasonable estimator of p, $x(n)$. Moreover, as is proven in the appendix, the only $P(p)$ which gives rise to a $P(k \mid n)$ of the form $C_k^N\ x^k\ [1-x]^{N-k}$ for all $\{k, N\}$ is $P(p) = \delta(p$ - constant), a clearly absurd prior.

Presumably BS probability distributions can, in many regimes of the Bernoulli problem, be good approximators of $P(k \mid n)$. However those distributions will never equal $P(k \mid n)$ exactly for all n and k for any reasonable prior.

4. Disagreement of the first two moments, for some n and N

This section addresses the issue of whether BS can even get the first two moments correct, for all n and N. First, write

$$\begin{aligned} E(k \mid n, N) &\equiv \sum_{k=0}^{N} k\ P(k \mid n) \\ &= \frac{\int dp\ P(p)\ p^n\ (1-p)^{N-n}\ \sum_{k=0}^{N}\ k\ C_k^N\ p^k\ (1-p)^{N-k}}{\int dp\ P(p)\ p^n\ (1-p)^{N-n}}. \end{aligned}$$

Using the fact that $\sum_{k=0}^{N}\ k\ C_k^N\ p^k\ (1-p)^{N-k} = Np$, we get

$$E(k \mid n, N) = N\ \frac{P_{n+1,N-n}}{P_{n,N-n}}. \tag{5}$$

Using similar reasoning,

$$E(k^2 \mid n, N) = \frac{N\ (N-1)\ P_{n+2,N-n}\ +\ N\ P_{n+1,N-n}}{P_{n,N-n}}. \tag{6}$$

Now by Eq. (5), for BS to get the first moment right, $x(n)$ must be equal to the ratio $P_{n+1,N-n}\ /\ P_{n,N-n}$. Then by Eq. (6), for BS to get the second moment right, $x^2(n)\ N\ (N-1) + Nx(n) = \frac{N\ (N-1)\ P_{n+2,N-n}\ +\ N\ P_{n+1,N-n}}{P_{n,N-n}}$. Therefore if the first moment is also correct, we have $x^2(n) = P_{n+2,N-n}\ /\ P_{n,N-n}$. Combining,

$$\frac{[P_{n+1,N-n}]^2}{P_{n,N-n}} = P_{n+2,N-n}. \tag{7}$$

We want this to hold for all pairs of values $\{N \geq 1, 0 \leq n \leq N\}$.

Define $D_i \equiv \int dp\ p_i\ P(p)$. Consider the $n = N$ case (so $n \geq 1$). By Eq. (7) $\frac{(D_{n+1})^2}{D_n} = D_{n+2}$, i.e., $\frac{D_{n+2}}{D_{n+1}} = \frac{D_{n+1}}{D_n}$. This must hold for all $n \geq 1$; for all such n, $\frac{D_{n+1}}{D_n} = \frac{D_2}{D_1} \equiv \alpha$. This in turn means that for all such n, $D_n = D_1\ \alpha^{n-1}$.

Now consider the case $N = n+1$ (so $n \geq 0$). We have $\frac{(D_{n+1} - D_{n+2})^2}{(D_n - D_{n+1})} = D_{n+2} - D_{n+3}$. Take $n = 0$, use $D_0 = 1$, and for the other D_n use $D_n = D_1 \alpha^{n-1}$. This gives $\frac{D_1^2 (1 - \alpha)^2}{1 - D_1} = D_1 [\alpha - \alpha^2]$. Cancelling terms and solving, we get $D_1 = \alpha$. So for all $n \geq 0, D_n = \alpha^n$.

This means that $D_2 - (D_1)^2$, the variance of $P(p)$, is 0. The only way this can be is if $P(p)$ is a Dirac delta function about some constant c. This in turn means that $E(k \mid n) = c$, which means that $x(n) = c$ for all n; a clear absurdity.

The preceding relied on looking at all n, N. In contrast, the proof concerning the full distribution over k (see the previous section) has N fixed, but varies over all (allowed by N) values of n and k. In addition, the arguments in both sections relied on allowing the $n = N$ and $n = 0$ cases. It is not immediately clear how things are changed if we simply decide to disallow those cases, as in [2].

5. Getting the variance right—going from $x(n)$ to $P(p)$

The results of the previous section notwithstanding, one might wish to use the BS with an $x(n)$ such that the standard deviation of the BS distribution over k, $C_k^N [x(n)]^k [1 - x(n)]^{N-k}$, is also the standard deviation of $P(k \mid n)$, for some $P(p)$, for all $N > 0, n \leq N$ (even though the BS distribution can not equal $P(k \mid n)$ or even get the first two moments exactly right). In general, the procedure for going from $x(n)$ to a $P(p)$ giving the same variance is the following.

The (squared) variance of k corresponding to $x(n)$ is $N\ x(n)\ (1 - x(n))$. To find the variance given by a $P(p)$, use Eq.'s (5) and (6). Next set the two variances equal and solve for $P_{n+2,N-n}$: for all $n \in \{0, \ldots N\}$,

$$P_{n+2,N-n} = \frac{x(n)\ (1 - x(n))\ P_{n,N-n} + (\frac{N\ P_{n+1,N-n}}{P_{n,N-n}} - 1)\ P_{n+1,N-n}}{N - 1}$$

To proceed further one must conduct an analysis similar to that in the appendix; expand the P's in terms of D_j, and solve. As in the appendix, having the requirement on the equalities hold for all possible $\{n, N\}$ might result in there being no $P(p)$ which solves the equations. Unfortunately, time constraints did not allow such an analysis for this paper.

However consider the special case where we use a frequency counts estimator, $x(n) = n/N$, and have $n = 0$ or N, or in any other way allow the estimate of the variance to equal 0. For such a scenario, for that n, $P(k \mid n)$ must be a delta function. So there are $N - 1$ values of k for which $P(k \mid n) = 0$. Looking at Eq. (1), we see that since $p^{n+k}\ (1 - p)^{2N-n-k}$ is greater than 0 for all $p \in$ the interval $(0, 1)$, the only way that $P(k \mid n)$ can equal 0 is if $P(p)$ is 0 for all $p \in (0, 1)$. So $P(p)$ must equal either $\delta(p)$ or $\delta(p - 1)$. This means that the variance always equals 0, regardless of n. This rules out the frequency counts estimator, and makes any estimator which can estimate the variance as 0 seem rather absurd.

6. Getting the variance right—going from $P(p)$ to $x(n)$

It is usually easier to go from a $P(p)$ to an $x(n)$ with the same variance than visa-versa. As an example, assume that $P(p)$ is uniform. Then using Eq. (5), $E(k \mid n, N) = N\ \frac{n+1}{N+2}$ which of course is just what one would expect from Laplace's law of succession. (In fact, regardless of the prior $P(p)$, $E(k \mid n, N) = N \int dp\, p\, P(p \mid n) = N \times$ the "Bayesian" estimate of the average p, given n.)

In a similar manner, we can derive

$$E(k^2 \mid n, N) = \frac{N\,(N-1)\,(n+2)\,(n+1)}{(N+3)\,(N+2)} + \frac{N\,(n+1)}{N+2}.$$

Collecting terms, after a bit of algebra we get

$$\begin{aligned} \chi(n, N) &\equiv \frac{E(k^2 \mid n, N) - [E(k \mid n, N)]^2}{N} \\ &= \frac{2(N+1)}{(N+3)\,(N+2)^2}\,\{(N+2)^2/4 - (n - (N/2))^2\} \end{aligned} \tag{8}$$

By inspection, $\chi(N - n, N) = \chi(n, N)$, as it should. We want $N\, x(n)\, [1 - x(n)] = N\, \chi(n, N)$ for all $n \in \{0, \ldots N\}$. The solution is

$$x(n) = \frac{1 \pm \sqrt{1 - 4\chi(n, N)}}{2}. \tag{9}$$

Now we would like to have $x(n) + x(N-n) = 1$, since intuitively $x(n)$ is the (estimate of) the probability of a positive event, and then by symmetry (redefine what is a "positive" versus a "negative" event), $x(N-n)$ is the probability of a negative event. To obey this equality we can use the negative root for the lower $N/2$ values of n in Eq. (9), and the positive root for the upper $N/2$ values.

So for BS to give the same variance one would have with a uniform $P(p)$, one should do the Monte Carlo sampling according to a distribution in which each of the positive events have probability $x(n)/n$ ($x(n)$ being the probability of the set of all positive events), and all the negative events have probability $\frac{x(N-n)}{N-n} = \frac{1-x(n)}{N-n}$. If no positive events occur ($n = 0$), then one must still assign probability $[1 - x(0)]/N$ to all of the negative events, but one must also assign probability $x(0)$ to a positive event, despite the fact that no such positive event is in the original sample. In other words, one must make up an event and add it to the original sample. (Similarly if no negative events occur.)

Note that $x(n)$ is only real if $\chi(n, N) \leq 1/4$. Therefore, since probabilities must be real, it is necessary that

$$(n - (N/2))^2 \geq (N + 2)^2\, [1/4 - \frac{N + 3}{8(N + 1)}]. \tag{10}$$

For $N = 1$, this condition is satisfied by all n. For $N = 2$, it reduces to $|n - 1| \geq \sqrt{2/3}$, which means that n can only equal 0 or 2. For large N, the requirement becomes $|n - (N/2)| \geq N/\sqrt{8}$, which is satisfied by $N(1 - 1/\sqrt{2})$ values of n.

In fact, for N large, we can write down immediately

$$x(n) = \frac{1 \pm \sqrt{-1 + 8(R - 1/2)^2}}{2}, \tag{11}$$

where $R \equiv n/N$. Taking the negative root for $R < 1/2$, and the positive root for $R > 1/2$, this $x(n)$ has the value 0 at $R = 0$, rises to $1/2$ for $R = (1/2)\,[1 - \sqrt{1/2}]$, is complex up to $(1/2)\,[1 + \sqrt{1/2}]$, where it again has the value $1/2$, and rises from there up to the value 1 at $R = 1$.

It is interesting to note that for large N, this $x(n)$, the estimator of p that gives the correct variance, agrees more and more with the frequency count estimator of p as one moves towards the limits $R = 0, 1$. In contrast, the Laplace's law of succession estimator of p *dis*agrees more and more with the frequency count estimator as one moves towards those limits. This despite the fact that the Laplace estimator, like the $x(n)$ estimator, is based on a uniform $P(p)$.

7. Future work

The kind of analysis done here can also be done when there are multiple possible events, and when the statistic is a more complicated function than counting the number of events of a given class. Other future work involves seeing how close BS can get to the predictive distribution. It may well be that although it can not given that distribution exactly, it can give a very close approximation to it.

8. APPENDIX

This appendix solves for the $P(p)$ which satisfies Eq. (3) for all N, n and k. First, since $x(n)$ is independent of n, using the positive exponent and defining $i \equiv n + k$, Eq. (4) tells us that for all $i \in \{0, \ldots 2N-1\}$, $\frac{P_{i+1,2N-(i+1)}}{P_{i,2N-i}}$ is the constant $x/(1-x)$ (x being the value shared by all N of the $x(n)$). Define $r \equiv x/(1-x)$:

$$P_{j,2N-j} = \alpha\; r^j, \tag{12}$$

where α, like r, is an as of yet undetermined constant.[1]

Now recall $D_j \equiv \int dp\; p^j\; P(p)$. Using this, the binomial expansion, Eq. (12), and the definition of $P_{n+k,2N-n-k}$, and defining $m \equiv n + k$, one derives

$$\sum_{i=0}^{2N-m} (-1)^i\; D_{m+i}\; C_i^{2N-m} = \alpha\; r^m \tag{13}$$

for all $m \in \{0, \ldots 2N\}$.

Start with $m = 2N$, and thereby get $D_{2N} = \alpha\; r^{2N}$. By iteratively decrementing m, we can solve for the other D_{2N-m}. This solution is unique. Therefore any formula for the D_j that solves Eq. (13) for all $2N$ of the m must be the unique solution to Eq. (13). This solution is given by the following:

$$D_j = \alpha\; r^{2N}\; (1 + 1/r)^{2N-j}. \tag{14}$$

[1] As an aside, note that by returning to Eq. (3) and setting k to 0, we see that $P_{n,N-n} = \frac{P_{n,2N-n}}{(1-x)^N} = \alpha\; r^n\; (1-x)^{-N}$; so with $\beta \equiv \alpha(1+r)^N$, we can write $P_{i,N-i} = \beta r^i$.

Proof: Plugging Eq. (14) into Eq. (13), we get

$$r^{2N} \sum_{i=0}^{2N-m} (-1)^i \; (1+1/r)^{2N-m-i} \; C_i^{2N-m} \;=\; r^m$$

as the equality that must be satisfied. This equality in turn implies

$$r^{2N-m} \; (\frac{r+1}{r})^{2N-m} \; \sum_{i=0}^{2N-m} (\frac{-r}{r+1})^i \; C_i^{2N-m} \;=\; 1. \qquad (15)$$

The sum $= [1 - \frac{r}{r+1}]^{2N-m} \;=\; (r+1)^{m-2N}$, so we do get 1 as required. QED.

Now D_0 must equal 1, since $P(p)$ is normalized. But by Eq. (14), $D_0 \;=\; \alpha \; r^{2N} \; (1 + 1/r)^{2N}$. Therefore $D_j \;=\; (1+1/r)^{-j}$. This in turn means that $D_2 \;=\; (D_1)^2$. An immediate consequence is that the variance of $P(p)$ is 0. The only way that can be is if $P(p)$ is a Dirac delta function. This completes the argument.

ACKNOWLEDGMENTS: I would like to thank Bill Macready for helpful discussions. This work was supported in part by TXN Inc.

References

[1] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap.* Chapman-Hall (1993).

[2] D. Rubin, *The Bayesian Bootstrap*, The Annals of Statistics, vol. 9, pp. 130-134 (1981).

[3] D. Wolpert and C. Strauss, *What Bayes says about the Evidence Procedure*, in *Proceedings of the 1993 Maximum Entropy and Bayesian Methods Conference*, G. Heidbreder (Ed), Kluwer, in press.