A BOOTSTRAP TEST FOR COMPARING PERFORMANCE OF PROGRAMS WHEN DATA ARE CENSORED, AND COMPARISONS TO ETZIONI'S TEST

Paul R. Cohen and John B. Kim

Computer Science Technical Report 93-52

Experimental Knowledge Systems Laboratory
Department of Computer Science, Box 34610
Lederle Graduate Research Center
University of Massachusetts
Amherst, MA 01003

Abstract

Experimental trials of programs are sometimes aborted when resource bounds are exceeded. The data from these trials are called censored data. This paper discusses the inferences that can be drawn from samples that include censored data. A key component of statistical inference, the sampling distribution, is generally not known for censored samples. However, the bootstrap procedure has been applied to estimate empirically the sampling distributions of many statistics. We show how to use the bootstrap to estimate the sampling distributions of the difference of means of two censored samples, enabling many comparisons that were previously ad hoc, such as the comparison of run times of algorithms when some run times exceed a limit. The reader will see how to extend the bootstrap to other tests with censored data. We also describe a test due to Etzioni and Etzioni for the difference of two censored samples. We show that the bootstrap test is more powerful, primarily because it does not make a strong guarantee that is a feature of the Etzioni's test.

Cohen and Kim. Bootstrap tests for comparing two samples with censored data.

The Problem of Censored Data

The subject of this paper is how to measure and make inferences about the performance of a program when trials of the program are occasionally aborted. This happens when resource bounds are exceeded; for example, when a program runs out of time or space before solving a problem. Imagine running ten trials of a search algorithm, recording the number of node expansions required to find a goal node if that number is less than 5000 and abandoning the trial otherwise. A hypothetical sample distribution of the number of node expansions, n, is:

Trial	1	2	3	4	5	6	7	8	9	10
Nodes	287	610	545	400	123	5000	5000	601	483	250

Table 1. A sample that includes two censored data.

Two of the trials were abandoned and the numbers we record in these cases (5000) are called *censored* data.

Censored data present no problems for descriptive statements about the *sample*, but they make it difficult to draw more general inferences. Provided we limit ourselves to the sample we can say, for example, that the mean number of nodes expanded in the previous ten trials is $\bar{n} = \left(\sum_{i=1}^{10} n_i / 10\right) = 1329.9$. If we are disinclined to include the censored data in the average[1], then we can leave them out and simply report the mean number of nodes expanded after the censored data are discarded: $\bar{n} = \left(\sum_{i \neq 6,7}^{10} n_i / 8\right) = 412.375$.

We run into problems, however, when we attempt to generalize sample results. For example, it is unclear how to infer the "population" mean number of nodes that would be expanded by the previous algorithm if we ran other experiments with ten trials. Statistical theory tells us how to make this generalization if no data are censored: the best estimate of the population mean is the sample mean. But our sample includes censored data, and we should not infer that the population mean is 1329.9, because we do not know how many nodes the censored trials might have expanded if we had let them run to completion. Nor should we infer that the population mean of uncensored trials is 412.375 because statistical theory does not explain the relationship between the mean of a sample that includes censored data and the mean of a population. We can draw no conclusions that depend on inferring the population mean; for example, we risk biased results if we try to infer that one algorithm expands significantly fewer nodes than another [7].

This paper describes a general method for drawing inferences from samples that include censored data. The method is an application of *bootstrap resampling*, a Monte Carlo technique for estimating *sampling distributions* of statistics [2,6]. We present two tests—one to tell us whether the mean of a sample is significantly different from a particular value, the other to determine whether two samples are significantly different; the reader will easily see how to construct other tests, including tests that depend on statistics

[1] In this example, the abandoned trials expanded more than ten times as many nodes as the others, which suggests that they are somehow different and not really comparable with the others, and should be left out of the sample.

other than the mean. We compare our two-sample test to one designed by Etzioni and Etzioni [5], and we show empirical power curves from which we conclude that our test is more powerful in many conditions. Bootstrap resampling is well-known and our one-sample test is similar in some respects to Efron's discussion of the sampling distribution of the trimmed mean [4]. The contributions of the paper are the two-sample test and comparisons with Etzioni and Etzioni's test, and bringing the constituent techniques to the attention of the AI community.

Background: Sampling Distributions

Statistical tests are commonly tests of whether sample results are *unusual.* Imagine we have two search algorithms, A and B, and two samples of ten trials for each algorithm. We want to know whether A expands significantly more nodes than B. A common way to answer the question is to subtract the sample mean number of nodes expanded by A, $\overline{n}_A$, from the same statistic for B's sample, $\overline{n}_B$, and ask whether $\overline{n}_A - \overline{n}_B$ is unusually large or small, given the *null hypothesis*, $H_0: \mu_A = \mu_B$ that the population means of the number of nodes expanded by A and B are equal. If the sample result, $\overline{n}_A - \overline{n}_B$, is unusual we reject the null hypothesis; we say μ_A is probably not equal to μ_B, or algorithm A expands a significantly different number of nodes, on average, than algorithm B. To say that a sample result is unusual, we must know the *sampling distribution* of the result: the probability distribution of all possible sample results, calculated from samples of a fixed size, given the null hypothesis, H_0. For example, the sampling distribution of $\overline{n}_A - \overline{n}_B$, given $H_0: \mu_A = \mu_B$, is the probability distribution of all possible values of $\overline{n}_A - \overline{n}_B$ that might be obtained by drawing samples of a fixed size from two populations with equal means. You can imagine what the sampling distribution looks like: small differences are likely (because H_0 says the population means are equal) and large positive and negative differences are unlikely. The sampling distribution of $\overline{n}_A - \overline{n}_B$ looks like a bell curve although it is not Gaussian. Rather, the sampling distribution of the difference of two means is a *t distribution*. To see whether a sample result is unusual, one simply converts it to a *t statistic* and sees where the statistic falls in the t distribution. If the t statistic falls in one of the tails of the distribution (as shown in Figure 1) then we know that the corresponding sample result has a relatively low probability, and we reject the null hypothesis.

Statisticians showed long ago that the t distribution is the sampling distribution of the difference of two means under the null hypothesis that the population means are equal, but no comparable results tell us the sampling distribution if the samples contain censored data. Moreover, for reasons discussed in [2,7], if we ignore the censored data, we get biased sampling distributions. Thus we cannot tell whether sample results are unusual, nor can we test hypotheses, at least, not by conventional means. The bootstrap resampling technique provides a way to estimate the sampling distribution of any statistic, given only the sample. In particular, bootstrapping permits us to estimate the sampling distribution of unusual statistics such as "the mean of all the sample values less than 5000," and, thus, the sampling distributions of statistics from samples with censored data.

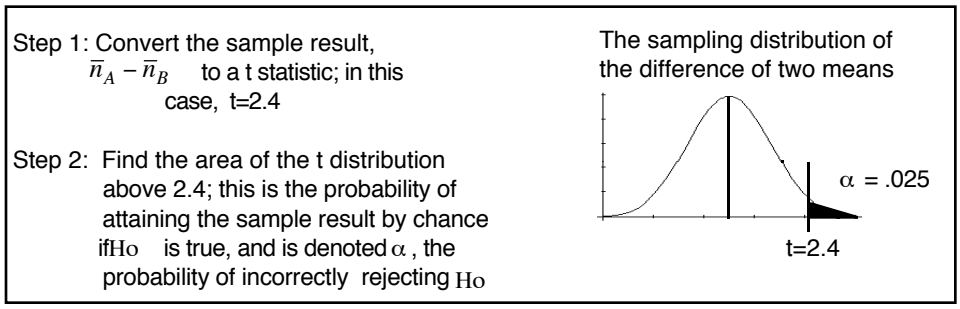


Figure 1. Sampling distributions and hypothesis testing.

The Bootstrap for a Censored One-sample Test of the Mean

We will illustrate the bootstrap method in the context of a one-sample test of the mean of a censored sample. More detailed discussions of the bootstrap (and related tests) can be found in [2,3,4,6]. In the following example we want to test whether the mean number of nodes expanded by an algorithm is less than 500. The null hypothesis is $H_0: \mu = 500$ and we will reject H_0 if the sample result is much lower than would be expected by chance, given H_0. Our sample is shown in Table 1, and our sample result is the mean of the uncensored data, $\bar{n}_{<5000} = 412.375$ (the subscript reminds us that we have the mean of the data with values less than 5000). Is this significantly lower than expected by chance under H_0? To find out, we need to estimate the sampling distribution of $\bar{n}_{<5000}$. Recall that the sampling distribution is the distribution of a statistic calculated from all possible samples of a fixed size drawn from a population. We want the distribution of $\bar{n}_{<5000}$ for all possible samples of ten items drawn from the population from which we obtained our sample. Unfortunately, we don't know anything about this population. But Efron proved that the best estimator of a population is a sample, leading to a remarkable procedure for constructing sampling distributions. Call the original sample S. We will draw K *bootstrap* samples $R_1 \ldots R_K$:

Procedure 1. Bootstrap Sampling for a One-sample Test

Repeat $i = 1 \ldots K$ times:

1. Draw a sample R_i of size N from S by sampling *with replacement* as follows:
   Repeat N times: select a member of S at random and add it to R_i
2. Calculate and record the value of $\bar{n}_{<5000}$ for R_i

Here are three bootstrap samples generated by this procedure:

	1	2	3	4	5	6	7	8	9	10	$\bar{n}_{<5000}$
R_1	610	601	610	483	483	610	287	5000	601	483	529.78
R_2	5000	601	250	250	5000	545	601	545	400	5000	456.0
R_3	250	287	400	400	123	545	601	250	545	545	394.6

After drawing a bootstrap sample R_i from S we calculate its $\bar{n}_{<5000}$ statistic. For example, R_1 contains nine values smaller than 5000 and their mean is 529.78, so $\bar{n}_{<5000}(R_1) = 529.78$.

You can see that sampling with replacement ensures that a datum in the original sample might be selected several times for inclusion in a bootstrap sample; for example, 610 shows up three times in R_1,

but just once in S. Similarly, items in S might not be selected for inclusion in a bootstrap sample; for example, 123 doesn't show up in R_1, and 5000 shows up just once instead of twice as in S. Resampling with replacement is justified in two rather different ways: First, if we resampled *without* replacement then every bootstrap sample R_i would be identical to S and every value of $\overline{n}_{<5000}$ for R_i would be identical to the original sample result, $\overline{n}_{<5000}$ for S. Clearly, this is no way to construct a sampling distribution of $\overline{n}_{<5000}$. Second, resampling with replacement is tantamount to assuming that the population (which we do not know) comprises the items in S, in the proportions that they appear in S, in essentially limitless quantities.

To construct a sampling distribution for $\overline{n}_{<5000}$ we simply repeat Procedure 1 many times. Figure 2 shows the sampling distribution of 1000 values of $\overline{n}_{<5000}$ calculated from bootstrap samples. The mean of this distribution is $\eta_{boot} = 412.75$ and its standard deviation is $\sigma_{\overline{n}} = 60.1$.

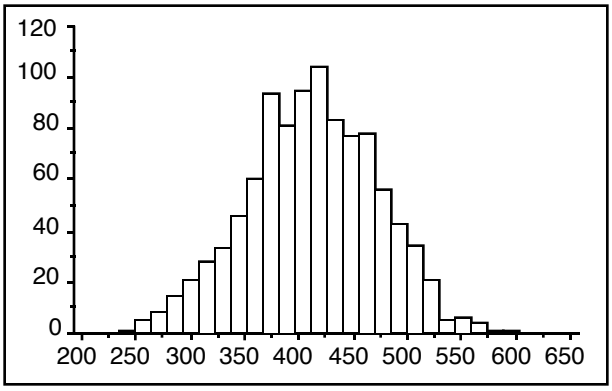


Figure 2. The bootstrapped sampling distribution for $\overline{n}_{<5000}$.

Now that we have a sampling distribution for $\overline{n}_{<5000}$, it would appear to be simple to test our hypotheses: $H_0: \eta_{<5000} = 500$, $H_1: \eta_{<5000} < 500$. Our immediate inclination might be to compare the sample result, $\overline{n}_{<5000} = 412.375$, to the sampling distribution in Figure 2. This is wrong. Figure 2 is the bootstrapped sampling distribution of $\overline{n}_{<5000}$, *not* the sampling distribution of $\overline{n}_{<5000}$ under the null hypothesis. If Figure 2 is not the sampling distribution of $\overline{n}_{<5000}$ under H_0, then what is? The question can be answered only by assuming some relationship between the bootstrapped sampling distribution and the null hypothesis sampling distribution. For brevity, we will refer to these distributions as $\mathbf{S}_{boot}$ and $\mathbf{S}_{Ho}$, respectively. One common assumption is that $\mathbf{S}_{Ho}$ has the same shape but a different mean than $\mathbf{S}_{boot}$. In this case, $\mathbf{S}_{Ho}$ is identical to the one in Figure 2 except it is shifted so its mean is 500 (because $H_0: \eta_{<5000} = 500$). The mean of $\mathbf{S}_{boot}$ is 412.75, so adding $500 - 412.75 = 87.25$ to every value in $\mathbf{S}_{boot}$ will shift it as desired. This is called the *shift method* of attaining $\mathbf{S}_{Ho}$. It transpires that 86 values in $\mathbf{S}_{Ho}$ are less than or equal to our sample result, $\overline{n}_{<5000} = 412.375$, so the probability of attaining this result by chance under H_0 is .086. Conventionally we adopt .05 as the probability required to reject the null hypothesis, so in this case we fail to do so. It is a simple matter to find a *critical value* for $\overline{n}_{<5000}$, a value sufficient to reject the null hypothesis. All we must do is sort the values in $\mathbf{S}_{Ho}$ and find the 50th one,

which happens to be 395.75. In other words, if our sample result had been $\bar{n}_{<5000} \leq 395.75$ then we could have rejected H_0 with $p \leq .05$.

The Bootstrap for a Censored Two-sample Test of the Mean

Two-sample tests are common in experiments that compare performance; for example, we might test whether one robot takes significantly longer to perform a task than another, and we might censor the data for the trials in which one or the other robot became trapped in a cul-de-sac. Sample data follow:

	1	2	3	4	5	6	7	8	9	10
Robot A	300	290	600	5000	200	600	30	800	55	190
Robot B	400	280	5000	5000	300	820	120	5000	120	400

Table 2. Hypothetical censored sample data for a comparison of two robots.

Trials 3 and 8 are *singly-censored*, which means that one or the other robot exceeded the time limit (5000, again), and trial 4 is *doubly-censored*, that is, both robots exceeded the limit.

Tests of differences of means are commonly run two different ways. We can find the mean times for robots A and B, $\bar{t}_A$ and $\bar{t}_B$, and ask whether $\bar{t}_A - \bar{t}_B$ is significantly different from zero, the value we expect under the null hypothesis that the robots perform equally. Or we can find the paired differences in performance between robots A and B on trials 1, 2, ..., 10, and ask whether the mean of these differences is significantly different from the value we expect under the null hypothesis. (The test statistic in this case is $\left(\sum_{i=1}^{10} t_{A_i} - t_{B_i}\right) / 10$.) The first test is called a *two-sample* test and the second is a *paired-sample* test, and they have different sampling distributions. However, the sampling distribution for both tests is unknown if the samples contain censored data. We now present two bootstrap procedures for estimating the sampling distribution of the difference of two means when the samples contain censored data, that is, a sampling distribution for a two-sample test. Later we will describe a paired-sample test (though not a test of means) due to Etzioni and Etzioni.

For each procedure, let S_A and S_B be the original samples of data from robots A and B, shown above. Let N_A and N_B be the sizes of the samples, which need not be equal for the following procedures. In Procedure 2, we will draw K bootstrap samples $A_1 \ldots A_K$:and $B_1 \ldots B_K$ as follows:

Procedure 2. Bootstrap Sampling for a Two-sample Test.

Repeat $i = 1 \ldots K$ times:

1. Draw a sample A_i of size N_A from S_A by sampling with replacement as described earlier
2. Draw a sample B_i of size N_B from S_B by sampling with replacement
3. Calculate and record the value of $\delta_i = \bar{t}_{A_i(<5000)} - \bar{t}_{B_i(<5000)}$, the difference of the means of the uncensored data in A_i and B_i.

In Procedure 3, we first combine S_A and S_B into a single sample $S_{A \cup B}$ and then draw K bootstrap samples $A_1 \ldots A_K$:and $B_1 \ldots B_K$ as follows:

Procedure 3. Bootstrap Sampling with Randomization for a Two-sample Test.

Repeat $i = 1 \ldots K$ times:

1. Draw a sample A_i of size N_A from $S_{A \cup B}$ by sampling with replacement as described earlier
2. Draw a sample B_i of size N_B from $S_{A \cup B}$ by sampling with replacement
3. Calculate and record the value of $\delta_i = \bar{t}_{A_i(<5000)} - \bar{t}_{B_i(<5000)}$, the difference of the means of the uncensored data in A_i and B_i.

The advantage of Procedure 3 is that the resulting distribution is the sampling distribution of the differences of means under H_0, whereas the distribution yielded by Procedure 2 must be shifted as described earlier to make it a sampling distribution under H_0. Details of this distinction are found in [2] but the basic intuition is that by sampling from $S_{A \cup B}$ in Procedure 3, we realize the implication of H_0 that a datum might as well have been produced by robot A as robot B. Whichever procedure we use, we end up with a sampling distribution to which we compare $\delta = \bar{t}_{A(<5000)} - \bar{t}_{B(<5000)}$, the difference of the means of the uncensored values in the original sample. If the probability of $\delta = \bar{t}_{A(<5000)} - \bar{t}_{B(<5000)}$ is less than .05 we reject H_0.

The Sign Test of Etzioni and Etzioni

Before evaluating the bootstrap tests, we will describe a different approach to the problem of censored data, due to Etzioni and Etzioni [5]. Theirs is a paired-sample test based on the sign test [1], so we will call it the Etzioni Sign Test, or EST. Briefly, if the samples for robots A and B contained no doubly-censored data, we could ask which robot "won" each trial. For example, on the first trial, robot A won because it had the shortest execution time, whereas robot B won on the second trial. Singly-censored data present no problem because the winning robot is obviously the one with the uncensored (i.e., smaller) execution time; for example, robot A wins trial three. Doubly-censored data, such as trial 4, is problematic. We cannot say which robot won. Etzioni and Etzioni propose a conservative interpretation of doubly-censored data: they count it as evidence for the null hypothesis, H_0. Imagine we are testing the hypothesis that robot A is faster than robot B. Then a "win" occurs when robot A completes a trial faster than robot B (e.g., trials 1,3). H_0 is that robot A and robot B are equally fast, which is equivalent to saying the expected number of wins is half the number of trials. H_0 will be rejected if the number of wins is unusually high. By counting each doubly-censored pair as a "loss," Etzioni and Etzioni provide the following strong guarantee: If we reject H_0 given the censored samples, then we would also have rejected H_0 if the censored trials had been allowed to run to completion.

The test statistic for EST is the number of wins, with doubly-censored data counting as losses. The sampling distribution of this test statistic is not known, but it is easy to show that comparing the test statistic to a binomial distribution provides the aforementioned guarantee. Imagine there are no doubly-censored data. N trials could therefore produce between zero and N wins, and if H_0 is true, the expected number of wins is $N / 2$ because H_0 says robots A and B are equally likely to win a trial. The binomial distribution gives the probability of m wins on N trials for $m = 0 \ldots N$, given the probability of a win, which under H_0 is .5. Thus the binomial is the sampling distribution for the number of wins—the probability distribution of all sample results. The probability of the sample result in Table 2—eight

wins in ten trials—given the null hypothesis that the robots perform equally, is .0547, marginally improbable enough to reject H_0.

EST does not take account of the magnitudes of differences between the robots. For example, it is surely important that when robot A was faster, it was a lot faster, whereas robot B was only a little faster in trial 2. Etzioni and Etzioni propose another test that uses magnitude information, but we will not study it here in part because it relies on some assumptions about population distributions, whereas the bootstrap tests and EST do not.

Evaluating the Performance of the Bootstrap Tests and EST

The tests were evaluated by constructing power curves for each. The power of a test is the probability that it will reject H_0 when H_0 is false, so, ideally, power should be 1.0. Practically, power depends on many factors, so the power of one or more tests is usually plotted against one of these factors. A test with a power curve that rises rapidly to 1.0 (or close to it) is preferred to a test with a slowly-rising curve, because the former test is more powerful over more of the range of a factor than the latter test. (See [2] for details on power curves.) The procedures for constructing power curves are somewhat involved, so the casual reader may wish to skip to the next section where the results are discussed. The discussion that follows assumes for the sake of brevity some knowledge of statistics.

Let Π_A and $\Pi_{B(k)}$ be two population distributions. Π_A is a uniform distribution[2] in the range $0...500$, so its mean and standard deviation are $\mu_A = 250$ and $\sigma_A = \sqrt{(500-0)^2/12} = 144.3$. $\Pi_{B(k)}$ is also a uniform distribution in the range $k\sigma_A...500 + k\sigma_A$. For example, Figure 3 shows $\Pi_{B(.5)}$ as a line shifted half a standard deviation with respect to Π_A. Let T be a censoring threshold; if a sample contains a datum $d_i > T$, that datum will be censored. Imagine for now that $T > \max[\Pi_B]$, so no data are censored. In this case, we could construct power curves for EST in a conventional manner:

To get the H_0 sampling distribution we let $k = 0$, so $\Pi_{B(0)} = \Pi_A$. The standard error of the H_0 sampling distribution is $\sigma_{H_0} = \sqrt{N/4}$. To get sampling distributions for alternative hypotheses H_k, which correspond to Π_A and $\Pi_{B(k)}$ being increasingly "pulled apart," we simply increase k. Under H_0, the probability of a "win" is .5; for example, the probability that robot A will complete a task before robot B is .5. Under an alternative hypothesis $H_{k>0}$, $\Pr(Win, H_k) > .5$, but it is easy to calculate.[3] Thus, the sampling distribution of the alternative hypothesis H_k will be binomial with parameters N and $\Pr(Win, H_k)$. For example, Figure 4 shows one null and two alternative sampling distributions for two cases: $\Pi_{B(.5)}, \Pr(Win, H_{.5}) = .716$ and $\Pi_{B(1.0)}, \Pr(Win, H_{1.0}) = .932$. We choose $\mu_A + 1.65\sigma_{H_0}$ as a critical value for H_0, giving $\alpha \approx .05$ (because the H_0 sampling distribution is approximately normal for large N). In the left pane of Figure 4, $\mu_A + 1.65\sigma_{H_0}$ is $N/2 + 1.65\sqrt{25/4} = 16.625$. The shaded area

[2] We chose uniform distributions because they made it easy to construct power curves. We do not believe our results depend on the choice but we will look at other population distibutions, such as the normal, in future.

[3] $\Pr(Win, H_{0.5}) = 2a + (b+c)^2/2$, where a,b and c are fractions of the uniform distribution in Figure 3.

of the alternative hypothesis distribution to the left of the critical value is β, the probability of not rejecting H_0 when we should, and the unshaded area of the $H_{.5}$ distribution to the right of the cutoff is $1-\beta$, the power of the test. Clearly, when $\Pi_{B(1.0)}$ is shifted one standard deviation to the right of Π_A (i.e., with a range $144.3...644.3$), the power of the test is 1.0 (shown in the second pane of Fig. 4).

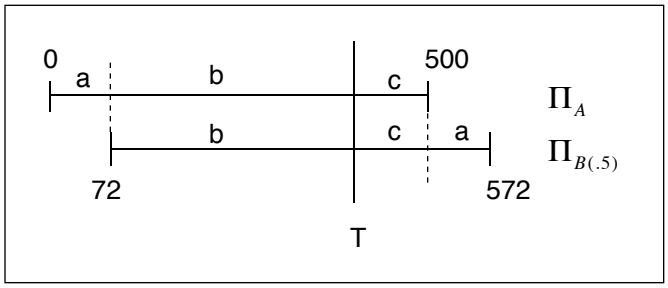


Figure 3. Uniform distribution populations and the censoring threshold.

The only effect of the censoring threshold T is to change the probability of a win for the null and alternative hypotheses. Under an alternative hypothesis $\Pi_{B(k)}$ is shifted above Π_A, so T divides the populations into three segments labelled a,b,c in Figure 3. The probability of a win, that is, drawing a pair of data $d_A < d_B$ from Π_A and $\Pi_{B(k)}$, respectively, decreases as the censoring threshold decreases T. Alternatively, the probability of a loss (which includes the probability of doubly-censored data) is $c\left(b^2/2\right)$, which increases as T decreases.

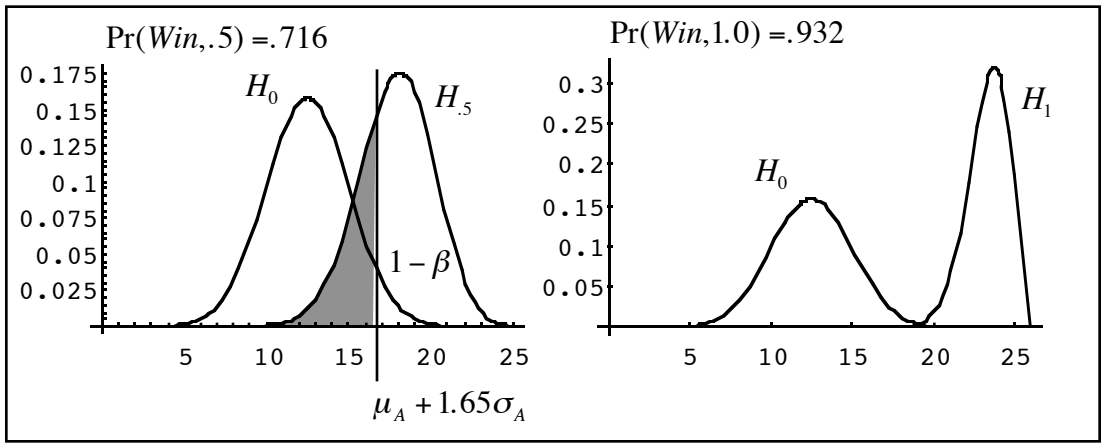


Figure 4. Computing the power of EST (see text for explanation).

We can now describe how the power tests were conducted. We assumed sample sizes of 25 throughout.

For EST:

- Loop over $k = \{.25, .5, .75, 1.0, 2.0, 3.0\}$
  ;;;(i.e., shift $\Pi_{B(k)}$ increasingly right of Π_A)
  - Loop over $p = \Pr(Win, H_k) = \{0, .1, \ldots, 1.0\}$
    ;;; (i.e., for each value of k, find a value of T to give the desired value of $\Pr(Win, H_k)$)
    - For a critical value c that ensures $\alpha \approx .05$, and a sample size N, the power of the test is the area

of the H_k distribution to the right of the critical value:

$$\sum_{i=c}^{N} \binom{N}{i} p^i (1-p)^{n-i}$$

For the two-sample bootstrap test:

- Loop over $k = \{.25, .5, .75, 1.0, 2.0, 3.0\}$
  ;;;(i.e., shift $\Pi_{B(k)}$ increasingly right of Π_A)
  - Loop over $p = \Pr(Win, H_k) = \{0, .1, \ldots, 1.0\}$
    ;;; (i.e., for each value of k, find a value of T to give the desired value of $\Pr(Win, H_k)$)
    - Loop over $x = \{1, 2, \ldots, 10\}$:
      ;;; (This is necessary to guard against the possibility that the results are biased by the sample from which we bootstrap)
      - Draw a sample A_x of size N from Π_A
      - Draw a sample B_x of size N from $\Pi_{B(k)}$
      - Derive a bootstrap sampling distribution for $\bar{t}_{A(<T)} - \bar{t}_{B(<T)}$, the difference of the means of the values in A_x and B_x smaller than T, following Procedure 2, above. This is the sampling distribution for $H_k: \mu_A - \mu_B = -k\sigma_A$.
      - Derive a bootstrap sampling distribution for $H_0: \mu_A - \mu_B = 0$ by Procedure 3 or by the shift method (both described above)
      - Determine c, the critical value that yields $\alpha = .05$. Determine P_x the area under the sampling distribution for $H_k: \mu_A - \mu_B = -k\sigma_A$ to the right of c. This is the power of the bootstrap test.
    - Average the values of P_x for each value of p and k.

Results

Pairs of sets of power curves are shown in Figure 5. Each pair corresponds to one setting of k, which is the number of standard deviations that $\Pi_{B(k)}$ is shifted right of Π_A (see Fig. 3). The first graph in the pair represents the effect on power of increasing $\Pr(Win, H_k)$, and the second represents the setting of T that was used to produce the corresponding setting of $\Pr(Win, H_k)$. Each graph contains three power curves, one for the Etzioni sign test (EST) and two for the two-sample bootstrap test. The bootstrap curves correspond to the two methods for deriving the bootstrap sampling distribution for $H_0: \mu_A - \mu_B = 0$, described above. For the bootstrap tests, the censoring threshold was constrained to be 10% more than the lower bound of $\Pi_{B(k)}$, to avoid getting bootstrap samples that contained no uncensored data. The missing points for the bootstrap tests are due to this constraint: no legal value of the censoring threshold would produce the desired $\Pr(Win, H_k)$.

Overall, the bootstrap test is more powerful than EST. This is probably because the bootstrap test is based on means (i.e., magnitude information) whereas EST is based on categorical data (i.e, whether a pair of data is a win or a loss.) Increasing k increases the power of all the tests, which is not surprising because increasing k corresponds to shifting $\Pi_{B(k)}$ increasingly above Π_A. Increasing the censoring threshold (T) increases the power of EST, but increasing T does not increase the power of the bootstrap tests monotonically. In fact, for most values of k, the power of the bootstrap tests drops initially and then rises as T increases. This is because T is the upper limit of the range of uncensored data, so as T decreases, so does the variance of the samples and of the bootstrap sampling distributions. (One can see in

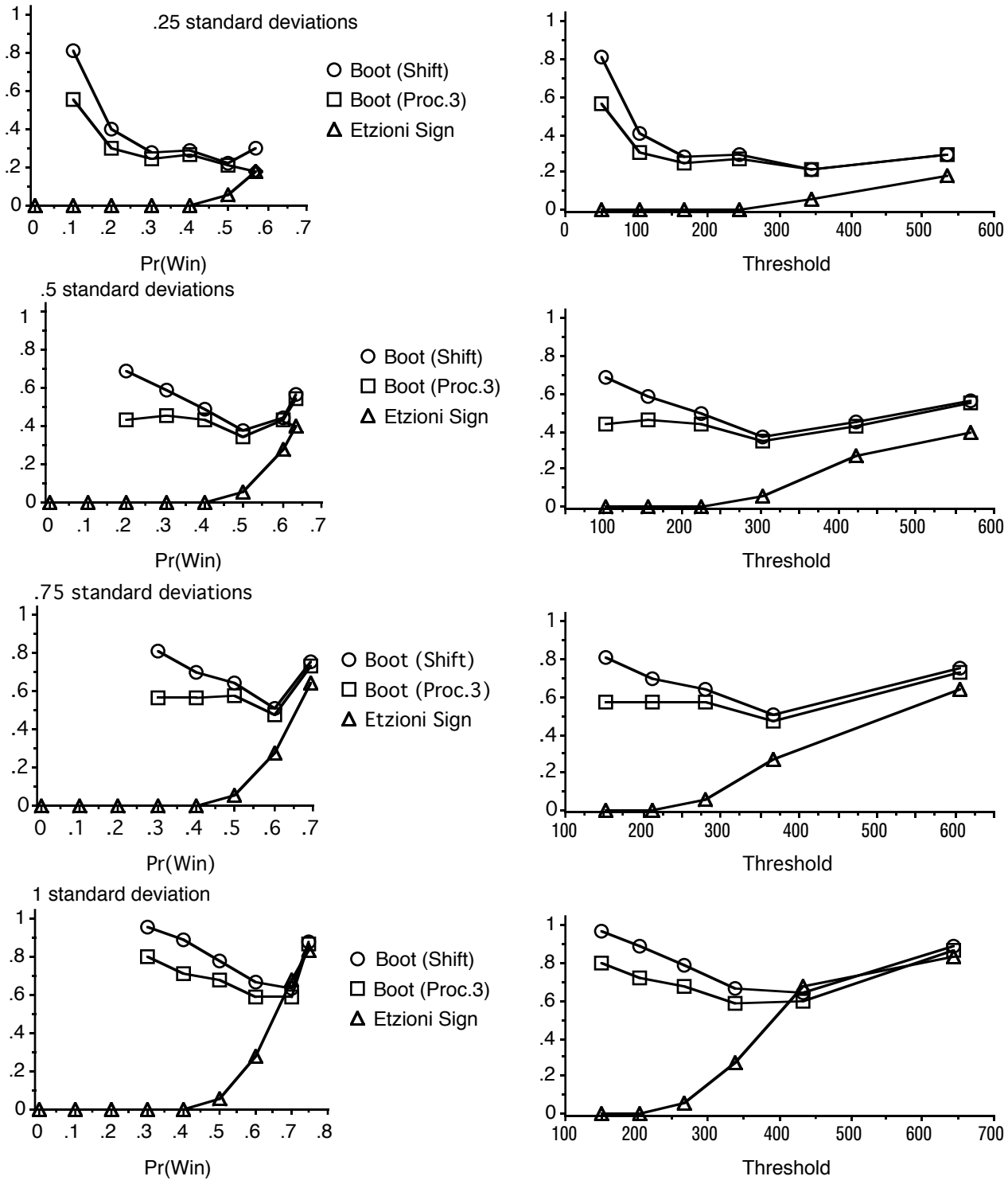
Figure 4 that, in general, power increases as the variance of the sampling distributions decreases.) Notably, the power of EST is often zero. This happens when the distance between $\Pi_{B(k)}$ and Π_A, combined with the censoring threshold, yields $\Pr(Win, H_k) < .5$, which means that the alternative hypothesis sampling distribution is actually to the left of the null hypothesis distribution and power is necessarily zero. In practical terms, if the censoring threshold is set low enough to ensure that half the pairs in a sample are doubly-censored, then the power of EST will be zero because EST treats doubly-censored pairs as losses. The bootstrap tests, in contrast, compare the means of the uncensored data, so maintain some power even when most of the data are censored. However, the bootstrap tests do not make EST's strong guarantee: if EST rejects H_0 given a sample that includes doubly-censored data, it would also have rejected H_0 if the doubly-censored trials had been allowed to run to completion. This guarantee accounts for the loss of power when the number of doubly-censored data is large.

Practically, EST is conservative but not as powerful as the bootstrap tests. If you want to be absolutely sure that censoring does not bias your results, then you should use EST but set the censoring threshold high enough to get a relatively small number of doubly-censored pairs. On the other hand, although the bootstrap tests do not offer EST's strong guarantee, they are more powerful over a wide range of censoring thresholds. Whereas EST ensures that you won't reject H_0 when you should not, the bootstrap tests are more likely than EST to reject H_0 when you should. A small additional advantage is that the bootstrap tests do not require your data to be paired. EST requires, for example, that in a trial robot A and robot B each solve the same problem, or that they solve two problems similar enough to make the question, "which one won this trial?" meaningful. Because the bootstrap tests take means over trials, the pairing of problems within trials is not required. Finally, both EST and the bootstrap tests are distribution-free—they make no assumptions about the distributions from which samples are drawn—unlike conventional methods such as the t test.[4]

In conclusion, experimental comparisons of AI systems and algorithms are becoming more common, so the problem of censored data—the bias it introduces into experimental results—is becoming more pressing. In many sciences, the problem is relatively minor because only small fractions of samples tend to be censored. But an AI researcher might inadvertently set a resource bound that censors half the data in a sample, which would certainly bias the results. Even if no sample data are censored, however, EST and the bootstrap tests might be preferred to conventional tests because they make no assumptions and they are just as easy to use.

[4]We assumed uniform population distributions only for the purpose of constructing power curves. The tests do not themselves rely on any assumptions about underlying populations.

Figure 5. Power curves for five pairs of uniform distribution populations separated by .25, .5, .75, 1.0, and 2.0 standard deviations.



References

1. Bradley, J.V. *Distribution-Free Statistical Tests.* Prentice-Hall, Inc., Englewood Cliffs, NJ (1968).

2. Cohen, P.R. *Empirical Methods for Artificial Intelligence*. In preparation.

3. Efron, B. and Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy . *Statistical Science,* Vol. 1, No. 1, 54-77 (1986).

4. Efron, B. and Tibshirani, R. Statistical data analysis in the computer age. *Science,* Vol. 253, 390-395 (1991).

5. Etzioni, O. and Etzioni, R. Statistical methods for analyzing speedup learning experiments. To appear in *Machine Learning*.

6. Noreen, E.W. *Computer Intensive Methods for Testing Hypotheses: An Introduction.* John Wiley and Sons, New York, NY (1989).

7. Segre, A., Elkan, C., and Russell, A. A critical look at experimental evaluations of EBL. *Machine Learning*, Vol. 6, No. 2 (1991).