Machine Learning, 14, 333-347 (1994)

Technical Note
Statistical Methods for Analyzing Speedup Learning Experiments

OREN ETZIONI (ETZIONI@CS.WASHINGTON.EDU)
Department of Computer Science and Engineering, FR-35, University of Washington, Seattle, WA 98195

RUTH ETZIONI
Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, Seattle, WA 98104 and Department of Biostatistics, University of Washington, Seattle, WA 98195

Editor: Steven Minton

Abstract. Speedup learning systems are typically evaluated by comparing their impact on a problem solver's performance. The impact is measured by running the problem solver, before and after learning, on a sample of problems randomly drawn from some distribution. Often, the experimenter imposes a bound on the CPU time the problem solver is allowed to spend on any individual problem. Segre et al. (1991) argue that the experimenter's choice of time bound can bias the results of the experiment. To address this problem, we present statistical hypothesis tests specifically designed to analyze speedup data and eliminate this bias. We apply the tests to the data reported by Etzioni (1990a) and show that most (but not all) of the speedups observed are statistically significant.

Keywords. speedup learning, statistics, explanation-based learning, experimental methodology

1. Motivation

Speedup learning systems are systems that automatically generate search-control knowledge (e.g., Etzioni, 1990b; Knoblock, 1990; Minton, 1988a; Mooney, 1989; O'Rorke, 1989; Shavlik, 1990). The effectiveness of a speedup learning system is typically evaluated by comparing the performance of a problem solver, guided by the learned knowledge, with the performance of the problem solver given no control knowledge, or given control knowledge acquired by a different learning system. The problem solver is run on a sample of problems randomly drawn from some distribution. In many experiments, the problem solver requires an inordinately long time to solve one or more of the problems due to the combinatorial nature of its search. To allow the experiments to complete in reasonable time, the experimenter imposes a bound on the CPU time that the problem solver is allowed to spend on any individual problem. When that bound is exceeded, the problem is marked "unsolved" and the problem solver moves on to the next problem. The same time bound

The statistical tests described in this article are encoded as COMMON LISP routines. The routines, and the data analyzed in the article, are available by sending mail to ETZIONI@CS.WASHINGTON.EDU. We hope that other researchers will use the routines to validate their own speedup learning experiments.

is imposed on each individual problem under all experimental settings. The information available regarding that problem's solution time is said to be truncated or *censored* due to the time bound.

In a recent paper, Segre et al. (1991) argue that the experimenter's choice of time bound can influence the results of the experiment. Segre et al. illustrate this point with a hypothetical example reproduced in tables 1 and 2. In table 1, using a time bound of 1000 CPU seconds, learning appears to increase total problem-solving time; in table 2, using a time bound of 3000 CPU seconds, learning is shown to actually reduce total problem-solving time.

1.1. Analysis

We agree with Segre et al. that this potential bias is undesirable. An obvious solution is to eliminate time bounds (or, more generally, resource bounds). In practice, this is not feasible, particularly as we scale our experiments to increasingly difficult problems. If we accept that some of our data may be censored, due to a resource bound, we need to analyze the impact a bound can have on the results of our experiment. Ideally, since the bound is under the experimenter's control, the bound should have *no* impact on the results in order to ward off claims that the experimenter could have manipulated the experiment to yield a particular outcome. In section 1.2, we propose statistical methods for analyzing censored data that have this property. Initially, however, we present several alternative approaches and identify their limitations. The fundamental question that all methods grapple with is this: how much weight should we assign to censored data?

An extreme approach is to discard all censored data, assigning it zero weight; the implicit assumption is that the relative performance of the two systems, as observed in the uncensored data, will extrapolate to the censored data. However, as tables 1 and 2 illustrate, this assumption can lead to erroneous conclusions.

Another alternative is to extend a standard test of average pairwise difference, such as the matched-pair *t*-test, which assumes that the observed differences between the pairs of solution times are drawn from a particular (e.g., normal) distribution. In such a test, even

Table 1. Segre et al.'s hypothetical speedup learning experiment, where the learned knowledge appears to slow down problem solving using a time bound of 1000 CPU seconds.

Problem	1	2	3	4	5	Total
Before learning	100	200	300	900	1000^+	2500
After learning	100	275	600	1000^+	1000^+	2975

Table 2. The learned knowledge turns out to speed up problem solving (as revealed when the time bound is increased to 3000).

Problem	1	2	3	4	5	Total
Before learning	100	200	300	900	3000^+	4500
After learning	100	275	600	1560	1078	3613

$^+$A problem whose solution time exceeds the bound.

though we do not know the true value of a censored difference (e.g., the difference in runtime due to learning on data points 4 and 5 in table 1), we impose a probability distribution on the possible difference, which enables us to compute the likelihood that a censored difference will turn out to be sufficiently large to change the outcome of the experiment; we then factor this likelihood into the test's result. Unfortunately, the distributional assumption severely restricts the generality of the test. In fact, the actual distributions observed in many speedup learning systems have no straightforward mathematical characterization—they are certainly not normal. Thus, using a test that presupposes a normal distribution, such as the matched-pair *t*-test (DeGroot, 1986), is inappropriate.

The ideal statistical test would test for "average speedup," which is the intuitive notion employed in the machine-learning literature, without making distributional assumptions. It is easy to see that, given censored data, no such test exists. Consider a sample containing at least one censored data point for each system. In the absence of distributional assumptions, the impact of the censored data points on the average performance of the two systems cannot be bounded. As a result, the systems could turn out to have the same performance, on average, or either system may turn out to be considerably faster than the other. There is no way to tell without increasing the resource bound.

To address this problem, we could posit an upper bound on the value of censored data points. We could then perform a worst-case analysis by replacing each censored data point, for the system purpored to be faster, with the upper bound and apply a standard statistical test for average difference to the transformed data set. This approach may be satisfying when a tight bound can be derived. However, when trying to investigate average speedup, a loose upper bound assigns too much weight to censored data points. Since the bound is under the experimenter's control, the experimenter would be open to claims that she manipulated the experiment. For illustration, refer back to table 1. If we replace each censored data point with *any* upper bound that exceeds 1000, learning would appear to be ineffective. Yet, as table 2 illustrates, this conclusion is misleading.

Worst-case analysis is appealing, in this context, because it could potentially eliminate *any* bias due to the experimenter's choices. However, we need to bound the impact of any single censored difference on the outcome of the experiment. Otherwise, assigning a worst-case value to a small number of censored data points will obscure a definitive trend in the rest of the data. We avoid this problem by reformulating the hypothesis being investigated. Instead of directly testing for average speedup, we use statistical methods that are not "swayed" by the value of a small number of data points. We return to this issue in section 4.4, after providing the appropriate background and describing our approach.

1.2. Maximally conservative tests

Our approach is based on a combination of standard statistical methods and worst-case analysis. The key idea is to eliminate bias from an experiment by interpreting censored data in a maximally conservative manner, without assigning too much weight to censored data points. We introduce statistical hypothesis tests *guaranteed* to draw reliable conclusions from censored data. The guarantee is the following:

> If our test provides evidence for a particular conclusion, given a resource bound, the evidence for that conclusion would be at least as strong if the experiment was run to completion without any resource bound.

We restate this guarantee in precise statistical terms at the end of section 2.

The remainder of this article is organized as follows. Section 2 reviews the statistical background necessary to understand the article, describes standard statistical methods for analyzing censored data, and considers their strengths and weaknesses. Section 3 describes the data set we use to illustrate our approach. The data set, taken from Etzioni (1990a), compares the performance of PRODIGY, EBL, STATIC, and human experts on PRODIGY's benchmark tasks. Section 4 introduces the statistical tests we propose. We apply each test to our speedup learning data, and discuss the results.

2. Statistical background

This section provides a precise but accessible exposition of statistical hypothesis tests, and considers standard statistical methods for analyzing censored data.

2.1. Hypothesis tests

Statistical hypothesis testing has become an important tool in any discipline in which observed data are subject to uncertainty. We provide a brief description of the basic concepts and procedure below. See Gibbons (1971), Wilks (1962), or any standard statistics textbook for a detailed exposition.

Broadly speaking, the goal is to estimate, using the data, a "state of nature," or an underlying data-generating mechanism from a finite space of possibilities. Consider the simple case of a coin toss, and let $P(h)$ be the probability of the coin turning up heads. If one is interested in whether the coin is biased in favor of heads or not, the two possible states of nature are $P(h) > 1/2$ or $P(h) \leq 1/2$. The aim of a hypothesis test is to use available data to decide which state prevails.

In formulating a hypothesis test, the researcher typically designates the hypothesis she wishes to establish as the *alternate hypothesis*; its negation is called the *null hypothesis*. The null hypothesis is denoted H_o, and the alternate hypothesis is denoted H_a. In our coin toss example, H_o might be $P(h) \leq 1/2$, and H_a would then be $P(h) > 1/2$. In a speedup learning experiment, H_o might be that the problem solver with no control knowledge performs at least as well as the problem solver with control knowledge, and the alternative hypothesis would be that the problem solver with control knowledge is superior. Note that the precise formulation of the null and alternative hypotheses is key to understanding exactly what conclusion is licensed by the test. See section 4.4 for further discussion of this issue.

The crux of the hypothesis test is the decision whether the data provide sufficient evidence against H_o to allow one to reject it. In essence, a hypothesis test is the statistical analog of a "proof by contradiction." The idea is to assume, tentatively, that H_o is true and to ask how *unlikely* is the experimental outcome observed, or one that favors H_a even more. If the likelihood of observing such extreme experimental outcomes is very low, then there is strong evidence for rejecting H_o. In the coin toss example, suppose that a coin turns up "heads" 5 out of 10 times. If the coin is fair (i.e., H_o holds), the probability that at least 5 out of 10 tosses will come out "heads" is 0.62. Thus, the evidence against H_o is weak. Suppose, on the other hand, that a coin turns up "heads" 98 out of 100. The probability of an observation at least as large as this, assuming that the coin is fair, is practically zero. Thus the evidence against H_o is very strong in this case.

The *p-value* is the probability, assuming H_o holds, of encountering data that favors H_a as much as or more than the data observed in the experiment. Thus, a small p-value leads one to reject H_o.

If H_o is rejected, the p-value is the probability that it has been rejeced in error; naturally, one would like this probability to be small. The threshold for the p-value is decided before the experiment and is called the *significance level*. If a test is performed at significance level α, then H_o is rejected if the p-value is less than α, and we say that the test is *statistically significant at level* α. All this means is that we are rejecting the null hypothesis with the caveat that we are making an error with probability at most α. Note that if H_o is not rejected, we do not accept it; all we can conclude is that we do not have evidence to reject it. In general, we can never conclude that the null hypothesis is true; we can only conclude that it is probably false.

In many cases, p-values are straightforward to compute. Consider a coin-tossing experiment where H_a is that $P(h) > 1/2$; h denotes that a coin turn up "heads." Suppose q out of n tosses turn up heads. The p-value is the probability of having q (or more) heads given that H_o is true, where H_o is that $P(h) \leq 1/2$. Although H_o covers an interval rather than exactly one value (i.e., $H_o : P(h) \leq 1/2$ as opposed to $H_o' : P(h) = 1/2$), the p-value corresponding to the H_o is bounded above by the p-value for H_o'.[1] Thus, we typically report this upper bound as the p-value in hypothesis tests with interval (or *composite*) null hypotheses. In the coin-tossing example, the p-value is simply the proportion of the n-long toss-sequences of a fair coin where q or more of the tosses turn up heads. This proportion can be computed using the binomial formula as follows:

$$\text{p-value} = \sum_{i=q}^{n} \frac{n!}{i!(n-i)!} \left(\frac{1}{2}\right)^n .$$

The normal approximation to the binomial distribution can be used if n is larger than 25 (DeGroot, 1986).

In many experiments, the significance level, α, is taken to be 0.05, but there is no reason that this value should be adopted in all situations. It is up to the experimenter to decide what significance level is appropriate. Factors to consider include the acceptable level of error and the number of tests being performed. If several tests are being performed, each at significance level 0.05, then the chance that at least one null hypothesis will be rejected in error is substantially larger than 0.05; this is called the *multiple comparisons problem* (Brown & Hollander, 1977). For k independent tests, each conducted at level α, the probability of at least one incorrect rejection of the null is $1 - (1 - \alpha)^k$. For example, if we conduct nine independent tests, each at level $\alpha = 0.05$, then the probability of incorrectly rejecting the null hypothesis at least once is 37%! To control the overall error probability in multiple testing, the experimenter has to reduce the significance levels in the individual tests. To achieve an overall significance level of 0.05, when conducting nine independent tests, the significance level of each test should be approximately one ninth of 0.05.

2.2. *Statistical tests for censored data*

The methodology described above applies to data that are not censored. Below, we consider statistical tests extended to analyze censored data. Censoring is not peculiar to speedup learning data. In fact, it abounds in reliability studies (where we encounter failure time data) and in medical studies (where we encounter survival or lifetime data). Consider, for instance, a clinical trial comparing two medical treatments. Patients are followed for five years, and their survival from the start of the trial is recorded. At the end of five years, some of the patients will have died, and their survival times will be known. However, the survival times of the patients who are still alive at the end of the trial will be unknown. These observed times are said to be censored at the end of the trial; all we know is that they exceed the trial duration. This is exactly the same situation as in a speedup learning experiment. The different treatments correspond to different problem solvers, or to the same problem solver with different control knowledge. The five-year length of follow-up corresponds to the resource bound. Problems that are solved within the resource bound are analogous to patients who die within the trial period; problems that remain unsolved at the resource bound are censored. Table 3 summarizes the analogy between survival analysis and speedup learning.

A large body of statistical theory has been developed for survival analysis; Kalbfleisch and Prentice (1980) is a classic reference. However, we have found that the theory relies on stronger assumptions than are warranted in the analysis of speedup learning data. Consider, for example, *doubly censored* data in which the problem-solving time (or the survival time) is truncated for both systems being studied.[2] It is standard statistical practice to discard such data and only analyze the singly censored and uncensored pairs in the sample (Holt & Prentice, 1974; Woolson & Lachenbruch, 1980). However, this practice amounts to assuming that the relative performance of the two systems as observed in the uncensored and singly censored data extrapolates to the doubly censored data. This assumption can introduce bias into the experiment.[3] The assumption is made in the medical and reliability

Table 3. The analogy between the speedup learning and survival analysis.

	Experiment	
	Speedup Learning Trial	Clinical Trial
Elements compared	Problem solvers	Treatments
Termination criterion	Problem solved	Death of patient
Data	Solution time	Survival time
Censoring due to	Resource bound	End of follow-up

contexts, where data may be extremely expensive to obtain, in order to enhance the ability of statistical tests to draw definitive conclusions from relatively small samples (samples containing only 20 to 30 data points are quite common).

To avoid making this assumption, we take a maximally conservative approach and interpret each doubly censored data point as supporting the null hypothesis. As a result, a larger sample may be needed to reject the null hypothesis in the presence of doubly censored data. Our maximally conservative choice decreases the sensitivity (also known as *power*) of our tests.[4] However, we feel that this tradeoff is appropriate because we can compensate for decreased sensitivity by increased sample size and, in speedup experiments, large samples are easy and inexpensive to generate.

Finally, with this statistical background in place, we are able to restate our guarantee from section 1 in statistical terms: whereas a hypothesis test computes a p-value, our tests compute an upper bound on the p-value that would be derived if the experiment were run without a resource bound. If this upper bound licenses the rejection of the null hypothesis, we can *guarantee* that the null hypothesis would have been rejected, with at least as much confidence, had the experiment been run without a resource bound. As with any hypothesis test, this guarantee is *one-sided*. If we fail to reject the null hypothesis, our tests are inconclusive; we can never conclude that the null hypothesis is true.

3. Speedup learning data

We demonstrate the value of our approach by applying it to speedup learning data taken from Etzioni (1990a). The data set compares the performance of the PRODIGY problem solver in the absence of control knowledge, to the performance of PRODIGY guided by the control rules generated by EBL, STATIC, and by human experts. Specifically, we analyze the pairwise comparisons PRODIGY versus STATIC, STATIC versus EBL, and EBL versus the human experts, on each of PRODIGY's benchmark tasks (the Blocksworld, Extended-Stripsworld, and Schedworld problem spaces). The problem sets, the human control rules, and the problem space definitions are taken from Minton (1988b). PRODIGY's total problem-solving time and the number of censored data points, in each experimental setting, are summarized in table 4. To the untrained eye, the table seems to indicate that STATIC and the human "significantly" outperform EBL, and that all three sources of control rules outperform PRODIGY, in each of the problem spaces. We will see how these intuitions fare under rigorous statistical scrutiny in section 4.

Table 4. Total problem-solving time in CPU seconds and number of censored data points in each experiment.

	Blocksworld		Stripsworld		Schedworld	
	Total	Censored	Total	Censored	Total	Censored
Human	46	0	193	0	948	4
STATIC	47	0	226	0	685	1
EBL	139	0	292	0	1262	6
PRODIGY	2182	12	4347	18	4391	23

Note: The number of problems in each problem set is roughly 100, and the resource bound is 150 CPU seconds on a SPARC Workstation.

From a statistical perspective, the central feature of speedup learning data is that they are censored. Another important feature of speedup learning data is that they are *paired*. Problems are generated at random, and *both systems* attempt to solve each problem. This is as opposed to unpaired, or independent data, where two sets of problems are generated and each system is allocated its own set of problems. The paired scenario is analogous to a trial comparing two opthalmic treatments, in which each patient receives both treatments, one in either eye. Applying both treatments to the same subject means that the two responses for a given subject are associated and cannot be treated as independent observations. Ignoring the paired nature of the data would amount to overlooking a key relationship in the data, and would result in formulating an overly conservative test. We rely exclusively on paired data techniques.

The third feature of our data is the presence of tied observations within a pair. If data are continuous, the likelihood of such an occurrence is practically zero, but in our data set such pairs make up a nontrivial fraction of the total. Although the actual running time of each system is some real number, the data were recorded on an integer scale. As a result, run times that are within one CPU second of each other can appear to be identical. We discuss how our statistical procedures are adapted to handle ties in the following section.

4. Statistical methods

Statistical methods for analyzing paired data are typically based on the differences between the paired observations. Below we describe two nonparametric statistical tests ordered by the amount of information they extract from their samples.[5] The first method, the *sign test*, relies only on the sign of the differences between pairs. The second method, the *signed rank test*, relies on both the sign and the rank, or order, of the differences. In this section, we apply both tests to our speedup learning data and discuss their limitations.

The sign test is a conceptually straightforward procedure that is readily extended to speedup learning data. However, as we shall see, it may lack the sensitivity to detect differnces between two systems. To address this problem, we turn to the more sensitive signed rank test. This test is a member of the class of linear rank methods, methods based on the rank of the observed pair differences (Hajek & Sidak, 1967).

4.1. The sign test

The sign test is based on the sign of the difference between the observations in a pair. Suppose that systems s and f are being compared. A *pair difference* is the difference in problem-solving time between system s and system f, on a given problem. The test's null hypothesis is that the probability of a positive pair difference is equal to the probability of a negative pair difference. That is, the probability that system f is faster than system s, on a given problem, equals 1/2. In what follows, we consider the one-sided alternative hypothesis H_a: the probability that system f is faster is greater than 1/2.

Given the number of pairs, and assuming pairs are independent, the observed number of positive differences is a binomial random variable. Suppose that q out of n differences are positive. Then the p-value is based on the binomial distribution with probability of success equal to 1/2, and is computed exactly as described in section 2. Note that the test considers the sign, but ignores the magnitude, of pair differences.

The above formulation does not consider the possibility of tied observations within a pair. As noted earlier, such pairs make up a nontrivial fraction of our data. One solution to the problem is to discard the tied pairs and to perform the sign test on the remaining data. A second solution is to count half the tied pairs as positive differences and half as negative differences. (See Hemelryk (1952) and Lehmann (1975) for analyses of the two solutions.) We choose the second solution here because it is more conservative. Our tied observations are almost certainly run times that are very close, and counting half of these pairs as positive and half as negative supports the null hypothesis, providing a conservative test statistic. The discreteness of our data is an artifact of the integer scale on which they have been recorded. In future experiments, we would recommend preserving the continuous nature of the data as much as possible. Then the number of tied pairs should be small, and both solutions to the problem of ties should yield the same inference.

4.1.1. Censored data

In speedup learning experiments, where censoring occurs because of a resource bound, all censored observations clearly exceed all non-censored observations. Thus, singly censored pairs represent complete data if all we are interested in is whether one member of a pair exceeds the other. On the other hand, doubly censored pairs represent no additional information whatsoever on the relative magnitudes of the observations within a pair. Instead of discarding doubly censored pairs from the sample, we take a maximally conservative approach and interpret each doubly censored pair as supporting the null hypothesis.

More precisely, we add the number of doubly censored pairs, d, to the number of negative differences, n^-. This extension has an elegant statistical interpretation: the test is now computing an upper bound on the p-value that would have been derived had the experiment been run without a time bound. This is precisely the figure we need in order to eliminate bias due to the experimenter's choice of time bound. If the p-value bound is sufficiently low, then we can *guarantee* that the null hypothesis would have been rejected even in the absence of the time bound. If the p-value bound is high, however, then, depending on the degree of double censoring, this high value may lead to a decision to increase the time bound and to repeat the experiment. Alternately, we may conclude that the experiment does not provide enough evidence to reject the null hypothesis. This conclusion may lead to a decision to run a substantially larger sample for greater sensitivity.

4.1.2. Application

We performed our censored-data extension of the sign test on the data described in section 3. Table 5 gives the resulting bounds on p-values. We find that, with the exception of the EBL-STATIC Stripsworld comparison, each of the pairwise comparisons in table 5 is significant in the Blocksworld and the Stripsworld, and no comparison is significant in the Schedworld.

Table 5. Upper bounds on p-values for our sign test to three decimal places.

	Blocksworld	Stripsworld	Schedworld
STATIC-EBL	0.000	0.04	> 0.5
Human-EBL	0.000	0.000	> 0.5
STATIC-PRODIGY	0.000	0.000	0.309

Note: The null hypothesis states that the source of control rules listed second is at least as effective as that listed first ("PRODIGY" refers to the PRODIGY problem solver run without learned control rules). For example, the "STATIC-EBL" row reports p-value bounds on the null hypothesis that EBL is as effective as STATIC. The sign test enables us to reject this hypothesis in the Blocksworld and Stripsworld, but not in the Schedworld.

The lack of a significant difference in the STATIC-PRODIGY Schedworld comparison is highly counter intuitive, considering that the STATIC's total problem-solving time is 685 CPU seconds compared with 4391 CPU seconds for PRODIGY (table 4). A closer look at the data indicates that in 48 problems STATIC is faster than PRODIGY, and that the reverse is true in 41 problems.[6] Thus, if we only look at the signs of the differences, the number of positive differences is roughly equal to the number of negative differences)—hence the lack of a significant result. However, it turns out that in this comparison, the negative differences are close to zero, while the positive differences are sizable, which explains the large difference in total problem-solving time.

These data illustrate an important limitation of the sign test. Since the sign test ignores all information about the magnitudes of the pair differences, it fails to reject the null hypothesis in such cases. This problem is particularly acute when measuring the impact of control knowledge on problem-solving time, as in the STATIC-PRODIGY comparison. Due to the overhead of utilizing control knowledge, we expect the unguided problem solver to run slightly faster on easy problems. If the control knowledge is effective, the unguided problem solver will be *much* slower on more difficult problems, but the sign test will not take this into account. The signed rank test, described below, is designed to remedy this deficiency of the sign test.

4.2. *The signed rank test*

The signed rank test weighs both the sign and magnitude of pair differences. The test procedure is as follows. The absolute values of the pair differences are ranked in increasing order. The smallest value is assigned the rank of one, the second smallest is assigned the rank of two, and so on. The signs of the differences are recorded along with the ranks. The null hypothesis is that the distribution of the pair differences is symmetric about zero. The alternate hypothesis is that the pair differences are slanted towards positive (or negative) values, in a sense made precise by Lehmann (1975, p. 157).

Under the null hypothesis, we expect the sum of the ranks corresponding to the positive differences to be at least as large as the sum of the ranks corresponding to the negative differences. The p-value is equal to the probability that sum of the positive ranks (denoted by T^+) is at least as large as that observed under the null hypothesis. Suppose there are n pairs. Then there are 2^n possible sign-rank configurations. In small samples, one can enumerate the sign-rank configurations yielding a value of T^+ at least as large as that observed; assuming the null hypothesis, each one has a probability of $1/2^n$. The p-value is the number of sign-rank configurations yielding a value of T^+ at least as large as that observed times this probability. If n is at least 25, a normal approximation to the distribution of T^+ may be employed.

As with the sign test, we handle zero pair differences by counting half the tied pairs as positive differences and half as negative differences. Since the value of the differences is zero, their rank is minimal. Thus, ties have less impact on the result of the signed rank test than on that of the sign test.

4.2.1. Censored data

The standard censored-data extension of the signed rank test is quite technical, so we omit its description here (see Woolson & Lachenbruch, 1980). We note, however, that the standard extension makes two important assumptions. First, although the procedure is rank based, p-values are in fact computed under a distributional assumption about the pair differences. Second, doubly censored pairs are effectively dropped, leading to a potential bias in the analysis due to the choice of resource bounds.

Instead of following the standard approach, we have developed a maximally conservative extension of the signed rank test for use with censored data. Specifically, if the alternate hypothesis is that system f is faster than system s, we assign a maximal negative rank to each difference in which system f is censored (including doubly censored pairs). This is the worst-case scenario that is still consistent with the data we have observed. In essence, we are checking whether the null hypothesis can still be rejected, if on each problem where system f was censored, f would have in fact taken much longer to solve the problem than system s. If so, then we can be confident that changing or eliminating the bound will not lead us to retract the rejection of the null hypothesis. After computing the ranks in this manner, we perform the standard signed rank test. As in our sign test, the p-value thus obtained is an upper bound on the p-value that would have been obtained had the experiment been run to completion.

4.2.2. Application

We performed our censored data extension of the signed rank test on our data set. The results appear in table 6. The comparisons that were statistically significant when using the sign test are significant here as well. In addition, the STATIC-PRODIGY comparison in the Schedworld shows a reduced p-value bound, demonstrating the increased sensitivity of the signed rank test over the sign test. The reduced p-value bound is very small (0.006), leading us to be fairly confident that STATIC outperforms PRODIGY in the Schedworld on the problem distribution used in the experiment.

Table 6. Upper bounds on p-values for our signed rank test to three decimal places.

	Blocksworld	Stripsworld	Schedworld
STATIC-EBL	0.000	0.009	> 0.5
Human-EBL	0.000	0.000	> 0.5
STATIC-PRODIGY	0.000	0.000	0.006

Note: The null hypothesis states that the source of control rules listed second is at least as effective as that listed first ("PRODIGY" refers to the PRODIGY problem solver run without learned control rules). For example, the "STATIC-PRODIGY" row reports p-value bounds on the null hypothesis that STATIC fails to speed up PRODIGY. The signed rank test enables us to reject this hypothesis in each of the benchmark problem spaces.

4.3. The effect of censoring on significance

As table 4 indicates, our data are very lightly censored. We can demonstrate that our tests are robust to heavier censoring by positing a lower time bound, resulting in additional censoring, and checking whether the differences observed remain significant. For example, in the Stripsworld STATIC-PRODIGY comparison, even if the time bound is reduced, resulting in 13 censored observations of STATIC's problem-solving time, the difference between STATIC and PRODIGY remains significant with p-value bound of 0.000 for both the sign and signed-rank tests. If the time bound is reduced further, and the number of censored observations of STATIC increases to 23, the p-value bound remains 0.000 for the sign test, but becomes 0.080 for the signed rank test.

We see that the outcome of the signed rank test is more sensitive to the number of censored data points than the sign test. This is not surprising, because the signed rank test assigns more weight to censored data. As in the sign test, censored data points count in favor of the system hypothesized to be slower but, in addition, these points are given a maximal rank that increases their weight.

In general, given the sample size and a significance level, it is easy to compute an upper bound on the number of censorings allowed before a test becomes inconclusive. The result of a test is *certain* to be inconclusive if the number of censored observations of system f (the faster system according to H_a) exceeds this bound. The upper bound is derived by calculating how many censored data points will cause the p-value bound to exceed the threshold α when the uncensored data maximally favor the alternate hypothesis. The calculations reveal that the number of censored data points should not exceed roughly 40% of the sample size for the sign test ($\alpha = 0.01$), and roughly 20% for the signed rank test ($\alpha = 0.01$). We omit the exact calculations here, but emphasize that these are only upper bounds. In general, the impact of censored data depends on the strength of the difference apparent in the uncensored data. If the uncensored data provide only "luke-warm" support for rejecting the null, then a small amount of censored data may well result in an inconclusive test. Note that our tests will not yield erroneous conclusions in the presence of heavy censoring; the tests will merely fail to report a significant result. If the proportion of censored data points is too high, a less restrictive resource bound may be necessary.

4.4. Limitations of statistical tests

While we advocate our tests as useful statistical tools, we caution the reader that the tests compare the total (or ranked total) of pair differences, not their actual magnitudes. This can potentially yield counterintuitive conclusions. For instance, a test can fail to find a significant difference in the rank metric when one appears to exist in the data metric. For instance, we were disappointed that our tests did not detect a statistically significant difference between STATIC and EBL in the Schedworld, despite the fact that STATIC produces control rules that appear to be almost twice as effective as EBL's (table 4). A close examination of the data revealed that the large difference in total problem-solving time is due to large differences on only 5 of the 100 problems in the sample (the difference between EBL and STATIC on these 5 problems is 649 CPU seconds). Since our tests place relatively little weight on any individual problem, it is not surprising that the tests were not "swayed" by large differences on 5% of the problems.

A more extreme consequence of outlying or atypical observations in the sample is the detection by the test of a significant difference in favor of one system when, according to the sample, the other system appears to be faster on average. A simple precaution, when interpreting a significant difference, is to make sure that the average problem-solving time of the system selected as faster by the test is in fact smaller than that of the competing system. Our software implementation of these tests enforces this restriction.

The term *significant difference* refers to *statistical significance* as defined in section 2, not to the magnitude of the difference. Given enough data, even a tiny difference may turn out to be statistically significant, although the *practical* significance of that difference may be questionable. Again, the solution is to examine the actual difference between the two systems. For instance, our tests show that there is a statistically significant difference between STATIC and EBL's search-control rules in the Blocksworld; table 4 confirms that the difference—almost a factor of three—is nontrivial.

Neither of our tests directly analyzes average problem-solving time. The sign test rejects the hypothesis that the number of positive differences is equal to the number of negative differences. The signed rank test rejects the hypothesis that the positive differences and the negative differences are of the same order. Only a parametric test, which assumes that differences are drawn from a particular (e.g., normal) distribution, can explicitly reject the hypothesis that the mean or average difference between the two systems is zero. However, both of our tests can be used as indirect *evidence* for an average speedup hypothesis.

While we hope that other researchers will use our tests to validate their own speedup experiments (see, for example, Kambhampati & Chen, 1993; Knoblock, 1993; Minton, 1993), we offer three final caveats. First, as with any statistical test, failure to reject the null hypothesis is inconclusive; it is *not* a basis for concluding that system s is at least as fast as system f. A more appropriate conclusion is that the experiment should be repeated with a higher resource bound or a larger sample size. If the sample size is already so large that the test is approaching maximal sensitivity (probability of detecting even small differences between the systems is greater than 90%), then failure to reject the null hypothesis can be regarded as suggestive that system s is at least as fast as system f. Second, as with any statistical tool, even when a significant result is obtained, the test does not substitute for a careful intuitive examination of the data, checking that the test is not "hiding" important characteristics of the data.

Finally, although statistical tests enable us to extrapolate from a small random sample of observations to the *particular* distribution from which the sample was drawn, the tests do *not* enable us to extrapolate from the behavior of the systems on small problems to their behavior on large problems without making further assumptions, namely, that the systems' relative behavior on small problems reflects their relative behavior on large ones. Thus, when carrying out statistical tests, it is important to generate a "representative sample," a sample that reflects the distribution of problems that the systems are expected to encounter.

5. Conclusion

We have described two statistical tests that determine whether observed differences in the performance of two systems are significant. The tests interpret truncated or censored data in a maximally conservative manner, eliminating bias due to the experimenter's choice resource bound. We applied both tests to the speedup learning data set taken from Etzioni (1990a) and have shown that most of the differences observed are statistically significant (see, in particular, the results of our extended signed rank test in table 6). We believe that this approach helps to allay the concerns regarding the use of resource bounds raised by Segre et al. (1991). Finally, although we have focused on speedup learning data, we note that our methodology can be used to analyze *any* quantitative comparison between two systems on a common set of problems.

Acknowledgments

We thank Mike Barley for an insightful observation that motivated us to develop the ideas in this article. We thank Steve Minton for distributing PRODIGY/EBL and his experimental data, making possible the experiments analyzed above. Mike Barley, Charles Elkan, Craig Knoblock, Steve Minton, Monte Zweben, and the anonymous reviewers provided very helpful comments on a previous draft. This research was funded, in part, by National Science Foundation grant No. IRI-9211045 and by Office of Naval Research Grant 92-J-1946.

Notes

[1] To see this, contrast the two null hypotheses $P(h) = 1/2$ and $P(h) = 1/10$. Under which null hypothesis are you more likely to see 60% heads? The answer is $P(h) = 1/2$. In general, we have that for all $q > 1/2$, and for all $k < 1/2$. prob($\geq q$ heads, given $p = 1/2$) is greater than prob($\geq q$ heads, given $p = k$). Thus, computing the p-value relative to $P(h) = 1/2$ is a conservative, and hence appropriate, choice.

[2] Data point 5 in table 1 is an example of a doubly censored data point.

[3] We thank Charles Elkan and Craig Knoblock for making this point.

[4] See Cohen and Kim (1993) for a more sensitive statistical test, which is contrasted with our own.

[5] Nonparametric tests are generally valid for a far wider range of distributions than their parametric counterparts.

[6] Ten of the remaining 11 problems are ties, and one problem is doubly censored.

References

Brown, B.W. Jr., & Hollander, M. (1977). *Statistics: A biomedical introduction*. New York: Wiley.

Cohen, Paul R., & Kim, John B. (1993). A bootstrap test for comparing performance of programs when data are censored, and comparisons to Etzioni's test. Unpublished manuscript, University of Massachusetts, Amherst.

DeGroot, Morris H. (1986). *Probability and statistics*, 2nd ed. Reading, MA: Addison Wesley.

Etzioni, Oren. (1990a). *A structural theory of explanation-based learning*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA. (Available as technical report CMU-CS-90-185.)

Etzioni, Oren. (1990b). Why Prodigy/EBL works. In *Proceedings of AAAI-90*.

Gibbons, Jean Dickinson. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

Hajek, J., & Sidak, Z. (1967). *Theory of rank tests*. New York: Academic Press.

Hemelryk, J. (1952). A theorem on the sign test when ties are present. *Indagationes Mathematica*, *14*, 322–326.

Holt, J.D. & Prentice, R.L. (1974). Survival analysis in twin studies and matched pair experiments. *Biometrika*, *61*, 17–30.

Kalbfleisch, J.D., & Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Kambhampati, Subbarao, & Chen, Jengchin. (1993). Relative utility of ebg based plan reuse in partial ordering vs. total ordering planning. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93)*. Cambridge, MA: MIT Press (AAAI).

Knoblock, Craig A. (1990). Learning abstraction hierarchies for problem solving. In *Proceedings of the Eighth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Knoblock, Craig A. (In press). Automatically generating abstractions for planning. *Artificial Intelligence*.

Lehmann, E.L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden Day.

Minton, Steven (1988a). Quantitative results concerning the utility of explanation-based learning. In *Proceedings of AAAI-88* (pp. 564–569).

Minton, Steven. (1988b). *Learning effective search control knowledge: An explanation-based approach*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA. (Available as technical report CMU-CS-88-133.)

Minton, Steven. (1993). Integrating heuristics for constraint satisfaction problems: A case study. In *AAAI-93 Proceedings*.

Mooney, Raymond J. (1989). The effect of rule use on the utility of explanation-based learning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 725–730).

O'Rorke, P. (1989). LT revisited: Explanation-based learning and the logic of Principia Mathematica. *Machine Learning*, *4(2)*, 117–160.

Segre, Alberto, Elkan, Charles, & Russell, Alexander. (1991). A critical look at experimental evaluations of EBL. *Machine Learning*, *6(2)*.

Shavlik, Jude W. (1990). Acquiring recursive concepts and iterative concepts with explanation-based learning. *Machine Learning*, *5(1)*.

Wilks, Samuel S. (1962). *Mathematical statistics*. New York: John Wiley & Sons.

Woolson, R.F., & Lachenbruch, P.A. (1980). Rank tests for censored matched pairs. *Biometrika*, *67*, 597–606.

Received July 9, 1992
Accepted October 7, 1992
Final Manuscript June 25, 1993