Performance Evaluation using Fast Permutation Tests

Tim Hesterberg
Research Department
Insightful, Inc.
1700 Westlake Ave. N, Suite 500
Seattle, WA 98109

Abstract Permutation tests provide exact performance evaluation in telecommunications applications, such as hypothesis tests for comparing equivalence of service time distributions, except for random error caused by inexact Monte Carlo sampling.

We describe software developed for Verizon for highly-accurate two-sample permutation tests. The software selects one of four algorithms, depending on samples sizes and accuracy required. Two are straightforward—exact calculation and Monte Carlo simulation. Two others, based on a combinationof analytical approximations and Monte Carlo calculations, provide fast calculations for medium to large sample sizes, and take into account outliers, lattice spacing and skewness for high accuracy.

The software performs hypothesis tests in 12 seconds or less on a 651 MHz Pentium over a wide range of sample sizes, for two test statistics (difference in means, and a modified t-statistic), with accuracy comparable to 500,000 replications or better. The software can be called from S-PLUS, SAS, or C.

1 Introduction

Verizon performs many tests of statistical significance, each comparing two groups of data. Let n_1 and n_2 be the numbers of observations in the two datasets. Typically one of these is much larger, say $n_1 >> n_2$. For example, there may be n_1 repair times for customers of the Incumbent Local Exchange Carrier (ILEC), and n_2 repair times for customers of a Competitive Local Exchange Carrier (CLEC). Let $\overline{x}_1$ and $\overline{x}_2$ be the means of the two sets of data, and $\hat{\sigma}_1$ the standard deviation of ILEC data. We may compute a statistic which compares the two sets of data, either the difference in means

$$\overline{x}_1 - \overline{x}_2, \tag{1}$$

or a "modified" t-statistic which uses the ILEC data to estimate the underlying variance

$$t_1 = \frac{\overline{x}_1 - \overline{x}_2}{(1/n_1 + 1/n_2)^{1/2}\hat{\sigma}_1} \tag{2}$$

The goal of the algorithms described here is to estimate "p-values" which measure the statistical significance of those test statistics under random permutations. The exact p-values are

$$p_1 = P(\overline{X}_1 - \overline{X}_2 \le \overline{x}_1 - \overline{x}_2) \tag{3}$$

and

$$p_2 = P\left(T_1 \le t_1\right) \tag{4}$$

where $\overline{X}_2$ is the mean of a sample of size n_2 chosen with replacement from all $n = n_1 + n_2$ observations, $\overline{X}_1$ and $\hat{\sigma}_{1,X}$ are the mean and standard deviation of the remaining n_1 observations, and

$$T_1 = \frac{\overline{X}_1 - \overline{X}_2}{(1/n_1 + 1/n_2)^{1/2}\hat{\sigma}_{1,X}} \tag{5}$$

is the corresponding modified t-statistic.

For small samples we may calculate the p-values exactly, by considering all possible random samples. Our "exact" algorithm, described in Section 2, does this. For medium-sized samples we may use pure random sampling, the "random" algorithm described in Section 3. For larger samples we may use analytical approximations, the "analytical" algorithm, or a combination of random and analytical techniques, the "mixed" algorithm; both are described in Section 4. Our "automatic" algorithm described in Section 5 selects the most appropriate of the other algorithms, depending on the sample sizes and numbers of outliers.

We note two points here before starting to describe particular algorithms. First, p-values based on the difference in means (1) are equal to

p-values based on using the usual two-sample t-statistic with a pooled estimate of variance; the proof is omitted here. Hence we omit further discussion of the usual t-statistic, and when we refer to a t-statistic we mean a modified t-statistic. Also equivalent to using the difference in means is to just use $\overline{x}_1$ as a test statistic, because $\overline{X}_1 \leq \overline{x}_1$ whenever $\overline{X}_1 - \overline{X}_2 \leq \overline{x}_1 - \overline{x}_2$, so for speed we use

$$p_1 = P(\overline{X}_1 \leq \overline{x}_1) \tag{6}$$

in place of (3).

Second, "sampling without replacement" and "permutation tests" are equivalent in these two-sample problems. In "sampling without replacement" note that there are $\binom{n}{n_1}$ ways to select a sample of size n_1 (or n_2) with replacement from n observations. Alternately, consider concatenating both samples into a single vector of data, randomly permuting, then treating the first n_1 elements as group 1 and the remaining as group 2. There are $n!$ different permutations, but if we ignore the order within each group this reduces to $\binom{n}{n_1}$ permutations, each of which corresponds to one of the samples without replacement. For example, if $n_1 = 3$ and $n_2 = 2$, there are $\binom{5}{3} = 10$ samples/permutations, whose indices are shown in this table:

1	2	3	4	5
1	2	4	3	5
1	2	5	3	4
1	3	4	2	5
1	3	5	2	4
1	4	5	2	3
2	3	4	1	2
2	3	5	1	4
2	4	5	1	3
3	4	5	1	2

We may view each row either as a single permutation, or the first three elements (or last two) within a row as a sample without replacement.

Our exact, analytical, and mixed methods correspond to sampling without replacement, while the random method corresponds to using permutations.

2 Exact Method

If $\binom{n}{n_1}$ is reasonably small, say less than 500,000 for routine use, then it is feasible to calculate p-values exactly by computing the test statistics for every random sample. The "exact" algorithm does this, by calculating the test statistics $\overline{X}_1$ and T_1 for each of the $\binom{n}{n_1}$ samples, and returning exact values:

$$\begin{aligned} p_1 &= \frac{\#\left\{\overline{X}_1 \leq \overline{x}_1\right\}}{\binom{n}{n_1}} \\ p_2 &= \frac{\#\left\{T_1 \leq t_1\right\}}{\binom{n}{n_1}} \end{aligned} \tag{7}$$

where $\#\left\{\overline{X}_1 \leq \overline{x}_1\right\}$ is the number of samples for which the mean of the ILEC sample is less than or equal to the mean of the original ILEC sample.

If $\binom{n}{n_1}$ is large, the running time for this algorithm could be very long. To avoid that, we set the `numPerms` argument (for either S-PLUS or SAS code) to the maximum number of random permutations to consider, by default 500,000. If $\binom{n}{n_1}$ is larger than the number the code will halt quickly and indicate an error.

3 Random Method

The "random" methods works by generating a large number (say 500,000) random partial permutations, computing the test statistics for each, and letting the estimate p-values be the fraction of times the random test statistics are less than or equal to the original values.

This algorithm is intended for use when $n_2 < 50$. It may be used for larger n_2, but will run slower; the running time is approximately proportional to $n_2 R$, where R is the number of random permutations.

For efficiency, we generate only partial random permutations, in which the first n_2 elements are selected randomly from the whole data, leaving the remaining n_1 elements in an arbitrary order; the fact that the remaining elements are not randomly ordered does not matter. To generate a random partial permutation we use Algorithm 2.

Algorithm 1 Random Method

numberUnder $\leftarrow 0$ (number of replications with random value $\leq$ s0)
Calculate sums and sums of squares for whole dataset
Calculate sum and sums of squares for smaller dataset
Calculate "Original Statistic" (mean or t-statistic)
for $r = 1$ to R **do**
  Generate a random partial permutation
  Calculate sum and sums of squares for smaller dataset
  Calculate "Random statistic"
  Increment numberUnder if "Random Statistic" $\leq$ "Original Statistic"
end for
$\hat{p} \leftarrow$ numberUnder / R (this is the estimated p-value)
standard error $\leftarrow \sqrt{\hat{p}(1-\hat{p})/R}$

Algorithm 2 Generate Partial Permutation

for $i = 1$ to n_2 **do**
  $j \leftarrow$ a random value between i and n
  Switch x_i and x_j
end for

For example, if $n_2 = 4$ and $n = 10$, and the random values of j chosen are $4, 7, 3, 9$, then the process of generating one partial permutation is shown in this table:

Original	1	2	3	4	5	6	7	8	9	10
i=1, j=4	4	2	3	1	5	6	7	8	9	10
i=2, j=7	4	7	3	1	5	6	2	8	9	10
i=3, j=3	4	7	3	1	5	6	2	8	9	10
i=4, j=9	4	7	3	9	5	6	2	8	1	10

In the first step, the fourth element is selected, and the original first element is placed in position 4; the original first element will later (in step 4) be moved to another position yet. In step 3 nothing happens, because $j = i$. Finally note that many of the final 6 elements are never moved.

4 Analytical and Mixed Methods

The next two methods, "analytical" and "mixed", share many of the same characteristics – assign outliers to one of two groups, then use analytical approximations for the conditional probabilities

$$P(\overline{X}_1 \leq \overline{x}_1 | \mathcal{A})$$
$$P(T_1 \leq t_1 | \mathcal{A}) \tag{8}$$

where $\mathcal{A}$ represents a particular assignment of outliers to the two groups; e.g. one assignment puts all outliers into group 1. They differ in how they assign outliers — "analytical" iterates over all possible outlier assignments, while "mixed" uses Monte Carlo for a random set of assignments.

The underlying probabilistic calculation involves conditioning,

$$P(\overline{X}_1 \leq \overline{x}_1) = \sum_{\mathcal{A}} P(\mathcal{A}) P(\overline{X}_1 \leq \overline{x}_1 | \mathcal{A}) \tag{9}$$

The sum is over all possible outlier allocations, and $P(\mathcal{A})$ is the probability of that particular assignment. The probability of assigning all

outliers to group 1 is $\frac{n_1}{n}\frac{n_1-1}{n-1}\ldots\frac{n_1-K+1}{n-K+1}$, or more generally, the probability of an assignment that puts K_1 outliers in group 1 is

$$P(\mathcal{A}) = \frac{\prod_{i=1}^{K_1}(n_1-i+1)\prod_{i=i}^{K_2}(n_2-i+1)}{\prod_{i=1}^{K}(n-i+1)} \tag{10}$$

The methods differ in how they assign outliers. In this section we first describe the common elements of the two methods, then describe how they assign outliers.

The basic algorithm in each case is

Algorithm 3 Analytical and Mixed Methods

```
Partially sort data, outliers first
Determine number of outliers, K
for r = 1 to R do
    Assign K outliers (each outlier to one of the two groups)
    analytical approximation for conditional probability given outliers
    accumulate results
end for
```

4.1 Determine Outliers

For accurate analytical approximations we need to exclude outliers from the data being handled by analytical approximations. For example, Figure 1 shows an example with a single outlier. Neither a normal nor skewness-adjusted (discussed below) analytical approximation can accurately represent the cumulative distribution of a test statistic. However, by conditioning on the outlier, both approximations are reasonably accurate. More generally, we handle multiple outliers.

In our code it is convenient to place all outliers at the beginning of the vector of n observations. In practice we assume that data are positively skewed, and that outliers occur only on the positive side of the data; this is customized for Verizon requirements, and could be easily generalized.

We class as outliers those observations whose assignment has a substantial impact on the final result, in particular that it exceeds a standard normal quantile by 0.3 times the standard deviation of the result.

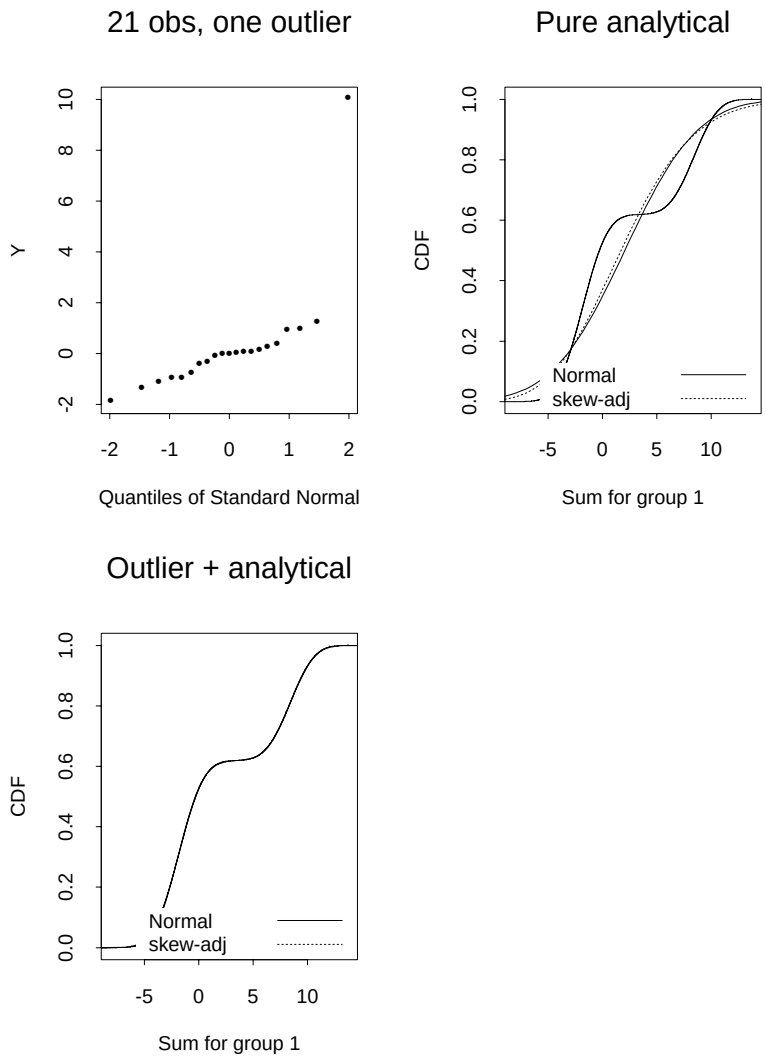

Figure 1: **Effect of an Outlier** Analytical approximations, with and without conditioning on the outlier. Note that handling the outlier separately is much more accurate.

We treat an observation as an outlier if

$$x > \overline{x}_3 + \Phi^{-1}(1-1/n)\sigma + 0.3\sigma\sqrt{n_1(1-n_1/n)} \tag{11}$$

where $\overline{x}_3$ is the mean of observations not already flagged as outliers, σ is the standard deviation of all observations, and Φ is the cumulative distribution function for a standard normal distribution. Note that if the underlying data really are normal, we expect the largest observation to exceed the mean by about $\Phi^{-1}(1-1/n)\sigma$.

The final term, $0.3\sigma\sqrt{n_1(1-n_1/n)}$, corresponds to how far beyond a normal value an observation should be in order to have a substantial impact on results. A factor of 0.3 results in small errors in p-values; less than .0005 based on comparing the cumulative distribution function (cdf) for a normal distribution for the cdf for a mixture distribution with the same overall mean and variance, but with two normal components, one of which has mean 0.3. And 0.0005 is an upper bound, occurring when $n_2/n \doteq 0.5$; the error is under 0.0002 when $n_2/n = 0.1$.

4.2 Conditional Probabilities for Means

Our goal is to estimate the conditional probability given in (8). There may or may not be outliers. Our description here assumes that there are; if there are not the description still holds, with the number of outliers set to zero. Let K be the number of outliers which have been assigned to either group, K_1 and K_2 the number assigned to the two groups, $K = K_1 + K_2$.

We begin with the conditional probability for sample means. It is convenient to work with sums rather than averages; let S_1 be the sum of the K_1 outliers assigned to group 1, and S be the sum of $n_1 - K_1$ observations chosen randomly without replacement from the $n-K$ non-outliers. Then

$$\begin{aligned} P(\overline{X}_1 \le \overline{x}_1|\mathcal{A}) &= P(n_1\overline{X}_1 \le n_1\overline{x}_1|\mathcal{A}) \\ &= P(S_1 + S \le n_1\overline{x}_1) \\ &= P(S \le n_1\overline{x}_1 - S_1) \end{aligned} \tag{12}$$

We consider three analytical approximations for estimating $P(S \le s)$ where S is the sum of a random sample of size J from N non-outlier observations ($N = n$ if no outliers, $N = n - K$ if there are K outliers). Let $\overline{x} = N^{-1}\sum_i x_i$, and let $y_i = x_i - \overline{x}$.

The approximations are based on moments of S. The moments for a single observation from $y_1, \ldots, y_N$ is

$$\begin{aligned} E(Y) &= 0 \\ \mathrm{Var}(Y) &= N^{-1}\sum y_i^2 \\ E(Y^3) &= N^{-1}\sum y_i^3. \end{aligned} \tag{13}$$

Now S is $J\overline{x}$ plus the sum of random values of Y, chosen with replacement. The moments for S are:

$$\begin{aligned} \mu &= E(S) = J\overline{x} \\ \sigma^2 &= \mathrm{Var}(S) = J\mathrm{Var}(Y)(1 - \frac{J-1}{N-1}) \\ m_3 &= E((S-E(S))^3) = JE(Y^3)(1 - 3\frac{J-1}{N-1} + 2\frac{(J-1)(J-2)}{(N-1)(N-2)}) \end{aligned} \tag{14}$$

Note that $\sigma^2 = 0$ when $J = N$, and $m_3 = 0$ when $J = N$ or $J = N/2$; there is no skewness when taking a sample of exactly half the size.

Let $z = (s-\mu)/\sigma$, $s_3 = m_3/\sigma^3$ be the skewness, and ϕ and Φ the standard normal density and distribution functions, respectively. The three approximations are:

$$\begin{aligned} \hat{P}_1(S \le s) &= \Phi(z) \\ \hat{P}_2(S \le s) &= \Phi(z) + \phi(z)(s_3/6)(-z^2+1) \\ \hat{P}_3(S \le s) &= F(z; \mu, \sigma^2, s_3) \end{aligned} \tag{15}$$

$\hat{P}_1$ is the normal approximation, $\hat{P}_2$ is the classical Edgeworth approximation, and $\hat{P}_3$ is based on matching moments of a translated gamma distribution, with F described below.

These are in order of increasing accuracy. Normal approximations are reasonable if skewness is negligible; see e.g. [1] page 42, as long z is not large.

The Edgeworth approximation is more accurate for small z if skewness is non-zero, but is inaccurate in the tails of the distribution; the approximation is not a monotone function of z, and can go outside the range $(0,1)$ — not a desirable property for a probability estimate.

The final approximation is the most accurate, in general. It avoids the problems of Edgeworth approximations. And translated gamma distributions have some nice properties for use in approximating distributions with only three moments known. First, they include normal distributions as a special case (when skewness is zero). Second, normal and translated gamma distributions are unique in having keeping the same basic shape (e.g. the familiar bell-shaped curve) under various operations. Third, their shapes are particularly suitable for accurately approximating statistics like those under consideration here. Finally, they can be accurately approximated using saddlepoint approximations; we do this for speed (see below).

For a gamma distribution with density

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{-1} e^{-\lambda x} \tag{16}$$

for $x > 0$, with shape parameter r and rate parameter λ (e.g. `dgamma` in S-PLUS), the moments are $\mu = r/\lambda$, $\sigma^2 = r/\lambda^2$, and $s_3 = 2/\sqrt{r}$.

Let X be a random gamma variable with shape $r = 4/s_3^2$ and rate $\lambda = \sqrt{r}/\sigma$, and Z a standard normal variable. The translated (reflected) gamma variable is

$$Y = \begin{cases} X + (\mu - r/\lambda) & \text{if } s_3 > 0 \\ -X + (\mu + r/\lambda) & \text{if } s_3 < 0 \\ \mu + \sigma Z & \text{if } s_3 = 0 \end{cases} \tag{17}$$

The corresponding distribution function is

$$F(x) = \begin{cases} F_G(x - (\mu - r/\lambda)) & \text{if } s_3 > 0 \\ 1 - F_G(-x + (\mu + r/\lambda)) & \text{if } s_3 < 0 \\ \Phi((x - \mu)/\sigma) & \text{if } s_3 = 0 \end{cases} \tag{18}$$

where F_G is the distribution function for a standard gamma variable with shape r and scale equal to 1.

Saddlepoint Approximations In an earlier version of the code, a substantial part of the computational time for the analytical and mixed algorithms was spent calculating the gamma distribution probabilities F_G. To reduce that time, we use a saddlepoint approximation to the gamma distribution, in particular the "r^*" variation of the Lugannani and Rice saddlepoint approximation (see [5]). For additional background see [2, 3].

Let

$$K(t) = \log(E(\exp(tX))) = -r\log(1-t)$$

be the cumulant generating function corresponding to F_G. Let $t = 1 - r/x$ be the solution to $K'(t) = x$, define $u = t\sqrt{K''(t)}$ and $w = \text{sign}(t)\sqrt{2(tx - K(t))}$, then the saddlepoint approximation is $\hat{F}_G(x) = \Phi(w + \log(u/w)/w)$ for $t \neq 0$, or $1/2 + s_3/(6\sqrt{2\pi})$ for $t = 0$.

The differences between $\hat{F}_G$ and F_G decrease as r increases, i.e. as the skewness decreases. The maximum absolute error is about 0.00009 for $r = 10$, with the maximum error occurring in the center of the distribution and the errors smaller in the tails. For $r = 25$ the maximum error is about 0.000025. We use this approximation for $r \geq 10$, and continue to calculate the gamma distribution probabilities for $r < 10$.

4.3 Conditional Probabilities for the t-statistic

In order to estimate conditional probabilities for the t-statistic, we begin with some algebraic manipulation:

$$\begin{aligned} P(T_1 \leq t_1) &= P\left(\frac{\overline{X}_1 - \overline{X}_2}{(1/n_1 + 1/n_2)^{1/2}\hat{\sigma}_1} \leq t_1\right) \\ &= P(\overline{X}_1 - \overline{X}_2 \leq t_1(1/n_1 + 1/n_2)^{1/2}\hat{\sigma}_1) \\ &= P(\overline{X}_1 - \overline{X}_2 - t_1(1/n_1 + 1/n_2)^{1/2}\hat{\sigma}_1 \leq 0) \end{aligned} \tag{19}$$

The previous relationships also hold if probabilities are calculated conditional on an outlier assignment. Let

$$h = \overline{X}_1 - \overline{X}_2 - t_1(1/n_1 + 1/n_2)^{1/2}\hat{\sigma}_1 \tag{20}$$

The top right panel of Figure 2 shows the relationship between T and h. Note that $T \leq t_1$ if and only if $h \leq 0$.

Next, we use standard Taylor-series techniques (see below) to obtain a linear approximation for h; that is, for every original observation x_i, we obtain an "influence" value L_i, and a single constant c_h, such that

$$h \doteq c_h + \sum_1 L_i \tag{21}$$

where the sum is over values assigned (in a random permutation) to group 1. The third panel in Figure 2 shows the relationship between x and L; the strong linear trend captures the effects that values have on the sample means for the two groups, and the slight nonlinearity captures the effect on $\hat{\sigma}_1$.

The final panel shows the excellent quality of the resulting linear approximation.

Now we may now estimate the p-value p_2 as the probability that the sum of a random sample of size n_1 of the L values is less than or equal to a constant,

$$p_2 = P(T_1 \leq t_1) \doteq P(S_L \leq -c_h) \tag{22}$$

where S_L is the sum of n_1 random values of L_i.

We then use the same type of analytical approximations described in the previous section, but for S_L (a sum of random values of L) rather than for S (a sum of original data values).

Furthermore, we may determine L for non-outlier observations conditional on where outliers were assigned, in order to more accurately estimate the conditional probability

$$P(T \leq t_1 | \text{outlier assignment}) \tag{23}$$

Calculating L The Taylor-series linear approximations used to obtain values L_i, conditional on outlier assignments, are also known as empirical influence function calculations in the statistical literature. The calculations are essentially obtaining the empirical influence function for h (20); but are complicated by the need for a conditional influence function, and because sampling (from the non-outliers) is without replacement.

Details of the calculations are omitted from this version of the document.

For other statistics the L values could be determined empirically rather than analytically, using regression methods based on bootstrap samples [4], using S-PLUS software currently described at

`www.insightful.com/Hesterberg/bootstrap`

This software was used for validation purposes here.

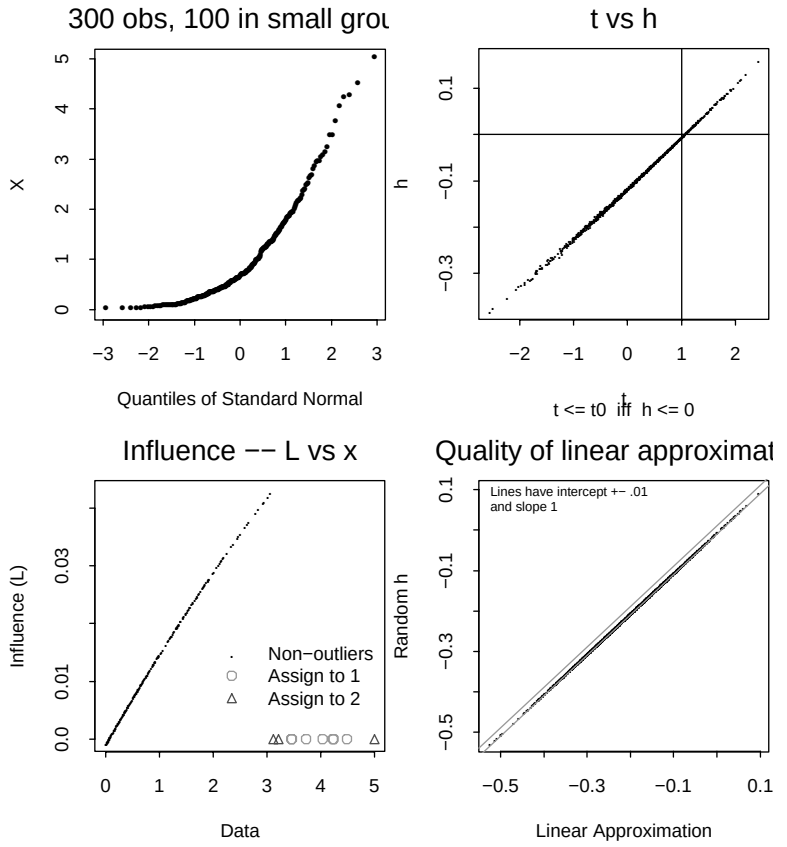

Figure 2: **Approximations for T-statistic** The top right panel shows the relationship between T (5) and h (20), for 1000 random permutations. Note that $T \leq t_0$ if and only if $h \leq 0$. The third panel shows the relationship between h and a linear approximation (21). The linear approximation is nearly exact. A line with slope 1 and intercept zero would have hidden the points; instead we show two lines with slope 1 and intercept ± 0.01. The final panel shows the relationship between the original data and the L values for the linear approximation. The final three plots are all conditional on a particular assignment of the largest 10 data values as outliers.

4.4 Lattice Spacing

For our analytical approximations for the p-value for the mean, it is important to determine whether the data are discrete, with differences between values all a multiple of a "lattice spacing", e.g. if all observations are integers, or a multiple of 10, or a multiple of 0.1; in these cases sample sums have the same spacing. If this is the case, we use a continuity correction based on the lattice spacing when doing analytical approximations.

Our code estimates the lattice spacing, using an algorithm optimized for Verizon requirements. It also allows the lattice spacing to be specified.

Once a lattice spacing is determined, we perform a continuity correction using half of this value; we use a Gamma (or normal or Edgeworth) approximation for

$$P(S \leq n_1 \overline{x}_1 - S_1 + \text{lattice}/2) \tag{24}$$

in place of the final line of (12).

We do not perform a continuity correction for p-values for the t-statistic. We found that it would be appropriate to perform a continuity correction if the original t-statistic t_1 is near zero, but not if t_1 is different from zero (which is when accuracy matters). The reason is that while the numerator of (5) has a lattice spacing, the denominator smears it out so that the ratio does not.

4.5 Analytical Method

The "analytical" method may be used when there are no outliers, or when the number of outliers K is small enough to loop over all 2^K possible allocations of outliers to the two groups (or, if $n_2 < K$, then there are fewer than 2^K possible allocations).

The probability of each outlier allocation is given by (10). No standard error is calculated because no random sampling is performed.

If 2^K is large, the running time for this algorithm could be very long. To avoid that, we set the `numPerms` argument (for either S-PLUS or SAS code) to the maximum number of random permutations to consider, by default 500,000. If 2^K is larger than the number the code will halt quickly and indicate an error.

Algorithm 4 Analytical Method (exhaustive treatment of outliers)

Partially sort data, outliers first
Determine number of outliers, K
$p \leftarrow 0$ (the p-value)
for all possible ways to assign outliers (typically 2^K) **do**
  Assign K outliers (each outlier to one of the two groups)
  $p \leftarrow p + P(\text{outlier assignment})P(T \leq t_1 | \text{outlier assignment})$
end for

4.6 Mixed Method

The "mixed" method is similar to the analytical method, except that outliers are assigned randomly to the two groups, and the number of permutations R is defined by the user.

Algorithm 5 Mixed Method (Monte Carlo treatment of outliers)

Partially sort data, outliers first
Determine number of outliers, K
for $r = 1$ to R **do**
  Randomly assign K outliers
  $p_r \leftarrow P(T \leq t_1 | \text{outlier assignment})$
end for
$\hat{p} \leftarrow \overline{p} = R^{-1} \sum p_i$
$\text{SE} \leftarrow \sqrt{(R(R-1))^{-1} \sum (p_i - \overline{p})^2}$

This yields a random sample observations whose expected value is the p-value; their average is the final p-value estimate, with standard error calculated using the usual formula for the standard error of an average of random values.

If $2^K < R$, the "analytical" method is faster and more accurate than "mixed".

5 Automatic Method

The "automatic" method chooses one of the other methods, depending on the values of

- sample sizes n, n_1, and n_2,
- maximum number of (random) permutations R to evaluate
- number of outliers K (supplied by the user, or calculated from the data)

according to the following rules:

exact if $\binom{n}{n_1} \leq R$, then exact results can be obtained by evaluating fewer than R permutations; this is the ideal situation, fast and exact. Run time is $O(n_2\binom{n}{n_1})$ (i.e. the running time grows at roughly this rate as sample sizes increase).

random if $n_2 < 50$, then that sample size is not sufficiently large to be sure of the accuracy of any method that includes analytical approximations. Run time is $O(n_2 R)$.

analytical if $2^K \leq R$, then combine exhaustive consideration of all 2^K possible allocations of outliers, followed by analytical approximations for the remaining observations. Run time is $O(K2^K)$.

mixed if $K \leq n_2$, perform R replications, in which outliers are allocated randomly and remaining observations treated analytically. Run time is $O(KR)$.

random if $K > n_2$, then it is faster to use this method than the "mixed" method. Run time is $O(n_2 R)$.

In addition, there are various initialization calculations whose run time is $O(n)$, or $O(n\log(n))$ if the data are sorted for determining outliers.

6 Conclusion

The methods described here are incorporated in software, which may be called from S-PLUS, SAS, or C.

We recommend the "automatic" method for routine use, which selects the most appropriate of the other methods. The other methods may be specified explicitly for validation purposes.

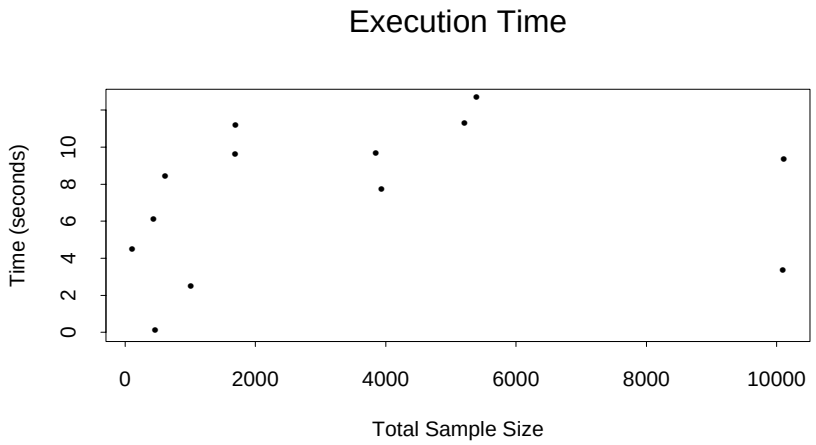

Figure 3: **Execution Time** Execution time on a 651 MHz Pentium, for sample data sets from Verizon. The times depend on the sample sizes in each group, and in some cases (where the analytical or mixed methods are used) on the number of outliers. The number of permutations for random methods is 500,000.

Two of the methods, "exact" and "random", are relatively straightforward, though some care was taken in programming these to yield fast results. The other two, which incorporate analytical approximations, incorporate a number of features to obtain high accuracy for medium and large samples.

The combination of methods yields results of high accuracy, quickly, as we see in Figure 3.

References

[1] W. G. Cochran. *Sampling Techniques*. John Wiley, New York, third edition edition, 1977.

[2] H. E. Daniels. Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650, 1954.

[3] Tim C. Hesterberg. Saddlepoint quantiles and distribution curves, with bootstrap applications. *Computational Statistics*, 9(3):207–212, 1994. Figures are in volume 10(2), page 193.

[4] Tim C. Hesterberg. Tail-specific linear approximations for efficient bootstrap simulations. In John Sall and Ann Lehman, editors, *Proceedings of the Conference on the Interface between Computing Science and Statistics*, volume 26, pages 472–481. Interface Foundation, 1994.

[5] John E. Kolassa. *Series Approximation Methods in Statistics*, volume 88 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1997.

Biographical Sketch Tim Hesterberg a Research Scientist at Insightful Corp. He holds a BA from St. Olaf College and an MS and Ph.D. in Statistics from Stanford University, and has received a DAAD Fellowship for study in Bonn, Germany, a National Science Foundation Graduate Fellowship, and a Performance Recognition Award from Pacific Gas & Electric Co. His research and consulting interests include bootstrap and other resampling methods, Monte Carlo methods, statistical computing, missing data, and electric demand forecasting. He is on the exective boards of the Interface Foundation of North America (Interface between Computing Science and Statistics, and the Statistical Computing Section and Nonparametric Statistics Sections of the American Statistical Association. A list of publications is available at `www.insightful.com/Hesterberg`.