Estimating extremely small Monte Carlo P-values cheaply

David L Duffy
Genetic Epidemiology Laboratory
Queensland Institute of Medical Research

Abstract

Davis and Resnick (Ann Statist 1984; 12:1467) describe a simple nonparametric procedure for the estimation of the tail of a distribution function based on a sample from that distribution, "the tail estimation problem". This relies on extreme value theory, and so is ideal for the types of very small P-values seen in modern large genome-wide association analyses, where permutation or simulation based tests are often preferred because they are perceived as being robust or correct in the face of familial or population clustering. Since only a small number of simulated statistics from the tail of the empirical distribution (usually the highest 10-20) need to be retained, this is computationally inexpensive. I present some applications, and show that the estimated P-values are conservative, but considerably better than the usual estimate 1/(1+B) (where B is the number of Monte-Carlo pseudo-samples) in the situation where the observed test statistic exceeded all simulated statistics:

$P_{extrapolated} = (m/B)(x/X_{(m+1)})^{-1/a*}$
$a* = m^{-1}\Sigma\ (\log(X_{(i)})-\log(X_{(m+1)}))$

where x is the observed test statistic, and m represents number of order statistics used.

Monte Carlo P-values

Geneticists are great consumers of Monte Carlo P-values (17% of the first 500 Google Scholar results for "Monte Carlo P-values", accessed 2011-03-29). The key advantage of Monte Carlo based significance testing [Barnard 1963] is that it is an easily implemented approach to very many statistical problems where analysis is intractable, or at least, too difficult for the present writer. The classical genetic situations it commonly finds use in involve testing hypotheses on correlated data, and multiple testing of correlated hypotheses.

$$p_{MC} = \frac{1}{B}\sum_{i=1}^{B}[T_i \geq t] \quad ,$$

where T_i is the value of the test statistic calculated for the ith of B datasets simulated under the null hypothesis, eg $E(T)=0$, and t is the test statistic value calculated for observed dataset.

The use of automatic digital computers makes this type of procedure reasonably quick, but quantities of data are expanding more quickly in this era of whole genome data. Sequential Monte-Carlo significance testing is one method to minimize the amount of calculation [Besag and Clifford 1991], which can be shown to always be more efficient than the usual fixed-size approach [Silva et al 2008]. This approach extends the standard sequential testing procedure (eg Wald 1945) to simulation under the null hypothesis.

However, in the multiple testing situation we are often interested in estimating very small P-values accurately. The minimum possible magnitude of a Monte Carlo P-value estimate is:

$$1/(1+B) \, ,$$

where B is the number of simulated samples, and the Monte Carlo error around this value is binomial. When the critical test threshold α is set to, say, 5×10^{-8}, the power to reject the null hypothesis is far less than the appropriate analytic test, unless B is large. The *resampling risk* is the probability that repeating the Monte Carlo test would reach a different conclusion as to whether a result was significant at the given α, and the above definition of a Monte Carlo P controls this resampling risk.

In the case of genome-wide data with many "significant" tests, this can become tediously slow.

Approximating the cumulative distribution function of the test statistic

In the above setup, we have carried out a lot of simulations, and one would think would give a lot of information about the shape of the null distribution of the test statistic. It would seem a more effective use of these data to carry out some type of curve fitting, and estimate the quantiles from this model: a higher-order approximation to the tail area. We can then assign P-values much smaller than 1/(B+1) to extreme observed values. Another advantage of this approach is in the multiple testing situation, where we can combine information about the *cdf* from all the tests.

In the case of the bootstrap, there has been much interest in using the saddlepoint approximation [Daniels 1954] for this purpose [Davison and Hinkley 1988] in that the relative error rates remain controlled in the tails. Unfortunately this property can only be guaranteed for functions such as the mean of the simulated samples (which arise naturally in bootstrap hypothesis testing).

Fitting curves from the Pearson system of distributions by maximum likelihood or matching the first four moments is the older approach. Most of the statistics arising in genetic linkage and association will come from the Gamma or Beta families.

Approximating the tail of the cdf *using extreme value theory*

Several authors have pointed out that we are not particularly interested in the shape of most of the distribution. The extreme tail of most distributions tend to resemble one another, and exceedances over a threshold fall into two families of the *extreme value distribution*, Exponential or Pareto, depending on the finiteness of the *index of regular variation*, *a*. Hill [1975] suggested a simple estimator of *a* based only on a set of the highest order statistics for the sample from that distribution,

$$a(n/m) = m^{-1} \sum_{i=1}^{m} [\log(X_{(i)}) - \log(X_{(m+1)})] \ ,$$

where $X_{(i)}$ is the *i*'th order statistic, *n* is the total sample size, and *m* is the number of order statistics.

With a finite estimate of *a* in hand, then Hill [1975], and Davis and Resnick [1984] suggested estimating the tail probability as (per the Pareto),

$$P = (m/n)(x/X_{(m+1)})^{-1/a}$$

Davis and Resnick [1984] show this estimate to be strongly consistent, and put bounds on this estimate. The variance of the estimate of $a(n/m) \sim 1/m$, but for optimal behaviour m/n should approach zero.

I am unaware of any previous application to a Monte Carlo type significance testing setup.

Fig 1. Performance of Davis-Resnick estimator versus *B* (constant *m*) for normally distributed test statistic. Black box-plots represents D-R estimator (*m*=10 highest statistics), upper blue boxes the naïve MC procedure, and solid line the asymptotic P-value for a true Z={4,5,6,7,8,9,10}. Panels represent results for 100, 1000, 10000, 100000 pseudosamples (1000 samples per condition).

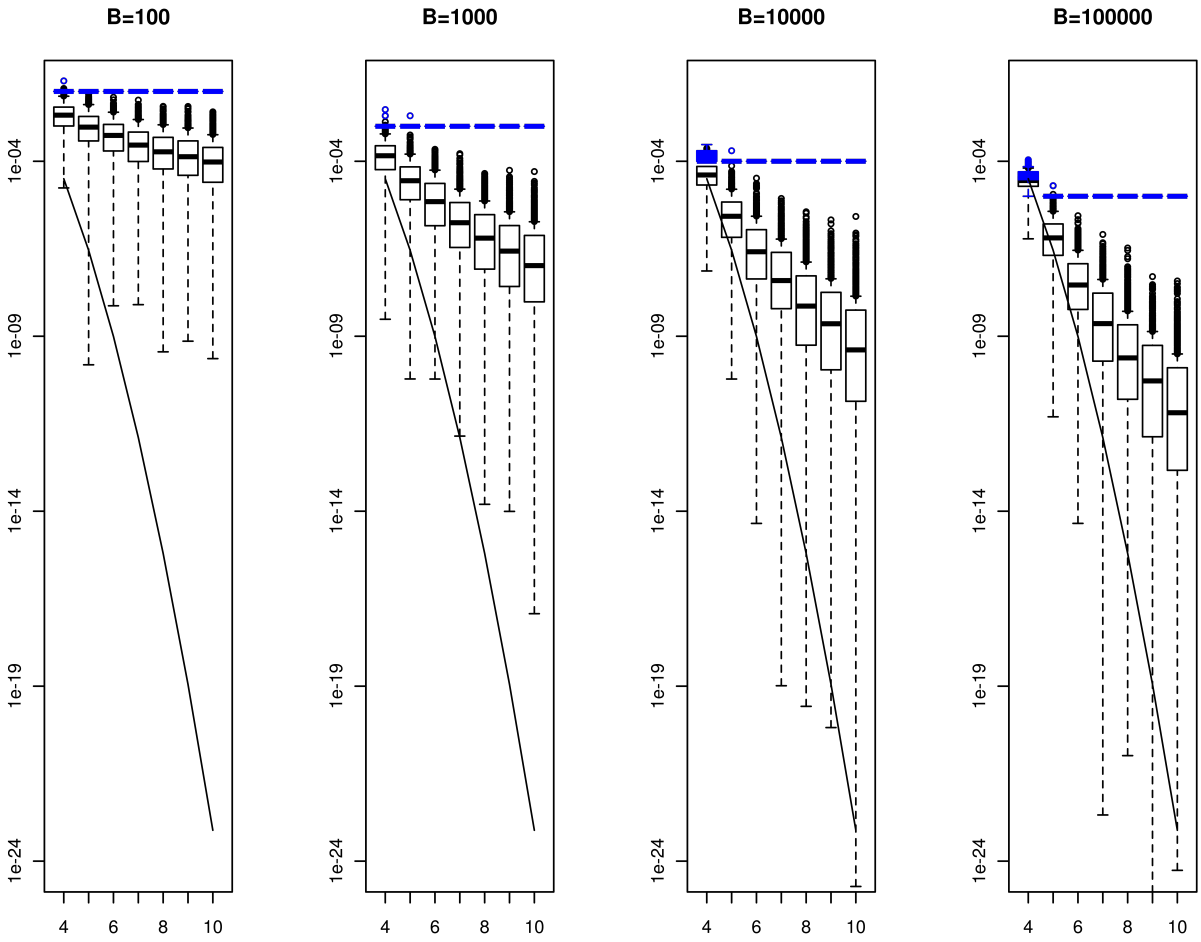


Fig 2. Performance of Davis-Resnick estimator versus *m* (constant *B*) for normally distributed test statistic versus *B*. Black box-plots represents D-R estimator, upper blue boxes the naïve MC procedure, and solid line the asymptotic P-value for a true Z={4,5,6,7,8,9,10}. Panels represent results for *m*= 5, 10, 20, 50 highest statistics in 10000 pseudosamples (1000 samples per condition).

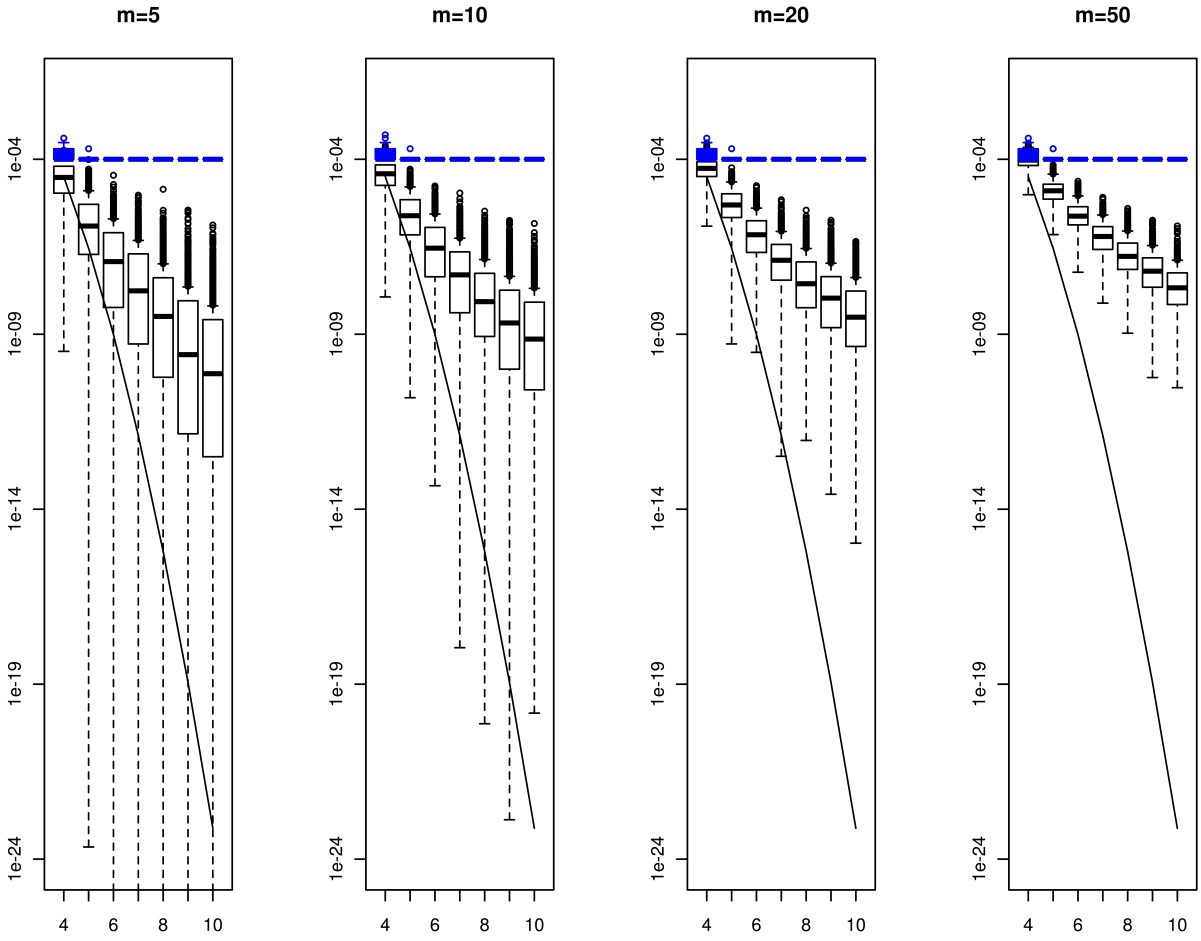
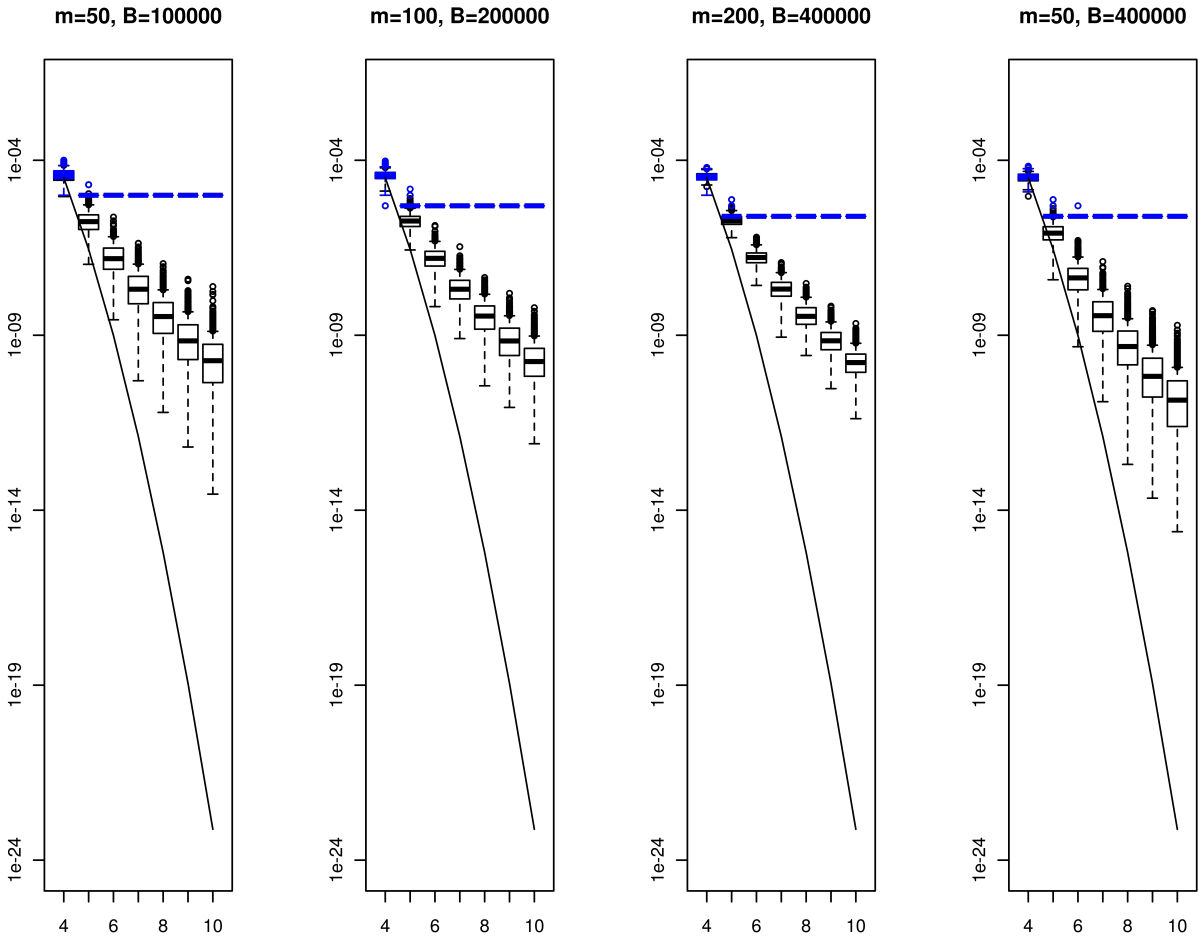


Fig 3. Performance of Davis-Resnick estimator with *larger* values of *B* for normally distributed test statistic. Black box-plots represents D-R estimator, upper blue boxes the naïve MC procedure, and solid line the asymptotic P-value for a true Z={4,5,6,7,8,9,10}. First three panels represent results for increasing *B*, holding constant *m/n*=0.0005; final panel shows reduction in bias by reducing *m/n* to 0.000125.



Resampling risk for this approach

This is easily defined, but hard to generalize upon since it depends on the size of the true effect, *n* and *m,* and the chosen significance threshold α.

A small example

Where the recombination distance is small, we can use a simple association X^2 with a gene-dropping test of significance to detect linkage within a single pedigree. Hall et al [1990] report on pedigree in which a BRCA1 pathogenic variant is segregating. The LOD under a fully penetrant dominant model is 3.01, equivalent to a two-tailed P-value of 2×10^{-4} (two-tailed). I use the two-tailed P so we can appropriately compare it to the gene-dropping P-value.

With *B*=100000 replicates, in 1000 runs the mean Monte Carlo P was 0.00021 (SD=4.6×10^{-5}). With a threshold equivalent to a LOD=3 (P=0.0002), the resampling risk is 50% and the alternative hypothesis is accepted 50% of the time, as one would expect straddling a hard threshold. If one used a threshold LOD of 2 to flag results for further iteration, then with only 200 iterations, the Davis-Resnick P-value would exceed this on 87% of occasions (obviously, no naïve MC P-values could exceed 0.005).

Estimator	*B*	*m*	Mean estimated P (SD)	Reject H0 (Resampling risk)
Asymptotic	-	-	0.0002	-
DR	200	10	0.0014 (0.0019)	14% (24%)
DR	500	10	0.0009 (0.0011)	23% (35%)
DR	1000	10	0.0006 (0.0007)	33% (44%)
DR	2000	10	0.00052 (0.00046)	35% (46%)
DR	5000	10	0.00038 (0.00023)	29% (41%)
DR	10000	10	0.00030 (0.00014)	39% (48%)
DR	10000	20	0.00030 (0.00014)	41% (48%)
Naive	100000	-	0.00021 (0.000046)	50.0% (50%)

Conclusions

This approach has the advantage of requiring relatively small amounts of computation. It tends to be biased towards the null, but is not exact, in the sense that it can sometimes exceed the correct P-value.

There is some room to develop adaptive methods to improve accuracy efficiently, but there is an interesting tradeoff between the sampling error due to m, the number of extreme values used to estimate the P-value, and the bias, which will increase as m/n increases.

The approach is used in my Sib-pair program [Duffy 2011] to augment the sequential testing algorithm used for Monte-Carlo P-values in a number of analyses.

References

Barnard GA (1963) Discussion of "The spectral analysis of point processes" by MS Bartlett. Journal of the Royal Statistical Society B, 25, 294.

Besag J, Clifford P (1991). Sequential Monte Carlo p-values. Biometrika, 78(2):301-304.

Daniels HE (1954). Saddlepoint approximations in statistics. Annals of Math. Stat. 25:631-650.

Davis R, Resnick S (1984): Tail Estimates Motivated by Extreme Value Theory. Annals of Statistics 12: 1467-87.

Davison AC, DV Hinkley (1988). Saddlepoint approximations in resampling methods, Biometrika 75: 417-431.

Davison AC, Hinkley DV (1997). Bootstrap methods and their application. Cambridge University Press.

Duffy DL (2011). Sib-pair Version 1.00Beta. [Computer Program]. QIMR: Brisbane.

Hill BM (1975). A simple general approach to inference about the tail of a distribution. Annals of Statistics 3:1163-1174.

Hope ACA (1968). A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society. Series B (Methodological), 30(3):582-598.

Silva R, Assuncao. R, Costa M (2009). Power of the Sequential Monte Carlo Test. Sequential Analysis: Design Methods and Applications 28: 163-174.

Wald A (1945). Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 16(2):117-186.