Journal of Statistical Computation and Simulation
Vol. 77, No. 1, January 2007, 55–61

Taylor & Francis
Taylor & Francis Group

Efficient calculation of p-values in linear-statistic permutation significance tests

PETER M. W. GILL*

Research School of Chemistry, Australian National University, ACT 0200, Australia

(Received 12 October 2004; in final form 1 December 2004)

It is shown that the exact p-values of permutation and bootstrap hypothesis tests of difference between groups can be written as an infinite series whose terms can be computed rapidly, even for large group sizes. Because of connections with the N-step random walk in the plane, the rate of convergence of the series improves as the size of the resampling distribution increases.

Keywords: Hypothesis tests; Permutation; Randomization; Bootstrap; Significance level

1. Introduction

Resampling significance tests possess a number of desirable properties: they are conceptually simple, unbiased, powerful and free of assumptions about the parent populations involved. There exists a huge literature in this area, including canonical works by Fisher [1], Pitman [2], Efron [3], Edgington [4], Davison and Hinkley [5] and Good [6]. Unfortunately, even if we restrict ourselves to the simplest case, in which a set of m observations and a set of n observations are pooled and then redistributed in all possible ways, the number of possible permutations can be vast (unless m and n are small), and the impossibility of exhaustively enumerating these in a reasonable time has restricted the extent to which exact permutation tests have been used in practice. For similar reasons, significance tests based on the exact bootstrap have rarely been implemented because of the formidable size of the resampling distribution. Most commonly, practitioners resort to random sampling from the permutation or the bootstrap distributions.

The practical problems associated with exhaustive enumeration and the statistical problems associated with random sampling are discussed in detail in Good's book. The author also devotes an entire chapter to the problem of increasing computational efficiency and gives an extensive bibliography of algorithms that have been proposed.

Our aim here is to show that, if the test statistic is linear, is possible to compute the p-value without exhaustive enumeration of the full permutation or bootstrap set and the computational

*Email: peter.gill@anu.edu.au

Journal of Statistical Computation and Simulation
ISSN 0094-9655 print/ISSN 1563-5163 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10629360500108053

time required is trivial, particularly when the resampling space is very large. Our approach is related to work by Pagano and Tritchler [7] but does not suffer from the numerical problems associated with the use of fast Fourier transforms [see ref. 8]. We hope that this will prove useful to statisticians who would like to use a resampling test but who would otherwise be deterred by the size of the total resampling distribution.

2. Theoretical formulation

Under the usual null hypothesis, all admissable resamples of the data are equally likely and we seek the fraction p of these whose corresponding test statistic is more extreme than that originally obtained. To quantify this, we introduce a statistic T, with observed value t, and then define the one-tailed significance level as $p = \mathrm{pr}(T > t) + \mathrm{pr}(T = t)/2$. The two-tailed extension of this is straightforward. If T is linear, scaling the data by a positive constant will not affect p and it will be convenient to prescale so that $\max |t_r^* - t| = 9\pi/10$, where t_r^* is the value of T in the rth resample.

If the total number of admissible resamples is N, we can write

$$p = N^{-1} \sum_{r=1}^{N} H(t_r^* - t)$$

$$H(x) = \begin{cases} 0 & x < 0 \\ 1/2 & x = 0 \\ 1 & x > 0 \end{cases}$$

and, if we substitute the Fourier expansion (whose $|x| < \pi$ validity is ensured by the prescaling earlier), we obtain

$$H(x) = \frac{1}{2} + \frac{2}{\pi} \Im \sum_{k'=1}^{\infty} \frac{\exp(ikx)}{k}$$

where $k = 2k' - 1$ and $\Im(z)$ is the imaginary part of z. Inverting the order of the finite and infinite sums then yields the convergent series

$$p = \frac{1}{2} + \frac{2}{\pi} \Im \sum_{k'=1}^{\infty} \frac{\Psi(k) \exp(-ikt)}{k}$$

where we have introduced the partition function

$$\Psi(k) = N^{-1} \sum_{r=1}^{N} \exp(ikt_r^*)$$

The potential usefulness of this reformulation depends on two questions.

Question 1: Does the infinite series for p converge satisfactorily? The answer to this depends to some extent on the characteristics of the distribution of the data. For small k, we can estimate $\Psi(k)$ by assuming that $t_r^* \sim N(\mu, \sigma^2)$ and replacing the sum over resamples by an

integral to obtain

$$\Psi(k) \approx \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] \exp(ikt)\,\mathrm{d}t$$
$$= \exp\left(ik\mu - \frac{\sigma^2 k^2}{2}\right)$$

which indicates that, if the resampling distribution of t is normal, the infinite series will initially converge rapidly. Numerical experiments suggest that the initial convergence is slower when the resampling distribution is non-normal but that it is still quite fast. For large k, we can estimate $\Psi(k)$ by assuming that the $kt_r^* \bmod 2\pi$ are uniformly distributed in $[0, 2\pi)$. This 'random-phase' assumption allows us to interpret $|\Psi(k)|$ as the net progress of an N-step random walk in the plane [see refs. 9–12] and to deduce that $E[|\Psi(k)|^2] = 1/N$. This implies that the later terms in the infinite series for p will fluctuate around zero with root-mean-square deviation $\mathrm{O}[1/(N^{1/2}k)]$. Thus, we expect the terms in the series to decay rapidly until they reach $\mathrm{O}[1/(N^{1/2}k)]$ and then to decay very much more slowly. Such convergence behaviour should be satisfactory in cases where N is large but will be less so if N is small. However, in the latter case, exhaustive enumeration is feasible.

The random-phase assumption is most reasonable in cases where the intrinsic resolution of the data is high. The exact distribution of $kt_r^* \bmod 2\pi$ is a sum of N delta functions and, if both k and N are large, this can closely approach a uniform distribution. However, when the data resolution is low, many of the delta functions coalesce and the uniform distribution assumption becomes much poorer. In such cases, the sequence of $\Psi(k)$ contains periodic 'spikes' that degrade the convergence rate of the series for p.

Question 2: Can $\Psi(k)$ be computed rapidly when N is large? The answer to this depends on the form of T and the definition of admissible resamples but, in the important case where T is linear, $\exp(ikt)$ factorises and the calculation of $\Psi(k)$ becomes almost trivial. Four particular cases will illustrate this.

3. Permutation test for matched pairs

A randomized matched-pair experiment to compare two treatments produces paired responses $\{x_q, y_q\}$ from which the paired differences $d_q = x_q - y_q$ are calculated for $q = 1, \ldots, n$. The null hypothesis H_0 of no treatment difference implies that the d_q are sampled from a distribution that is symmetric with zero mean. We define T to be the sum of the d_q and Fisher's permutation test considers the $N = 2^n$ resamples obtained by changing the signs of the d_q.

By definition, we have

$$\Psi(k) = 2^{-n} \sum_{s_1=\pm 1} \cdots \sum_{s_n=\pm 1} \exp\left[ik \sum_{j=1}^{n} s_j d_j\right] = \prod_{j=1}^{n} \cos k d_j$$

and the practical consequence of this factorisation is that, although each $\Psi(k)$ is a mean of $N = 2^n$ exponentials, it can be evaluated exactly in $\mathrm{O}(n)$ floating-point operations. The permutation-test significance level is thus

$$p = \frac{1}{2} - \frac{2}{\pi} \sum_{k'=1}^{\infty} \frac{\sin kt}{k} \prod_{j=1}^{n} \cos k d_j$$

Table 1. Matched-pair significance tests (Darwin data, $n = 15$).

	Permutation test ($N = 2^{15} = 32{,}768$)		Bootstrap test ($N = 30^{15} = 1.4 \times 10^{22}$)	
k	$\Psi(k)$	Partial sum	$\Psi(k)$	Partial sum
1	+0.8651202054	0.0264993197	+0.9500697835	0.1500657699
3	+0.2523199843	0.0245000457	+0.6292375852	0.0217413641
5	+0.0117796933	0.0258430766	+0.2725947929	0.0197723135
7	−0.0000090442	0.0258437504	+0.0751958197	0.0260880499
9	+0.0000055141	0.0258437068	+0.0126162029	0.0266835002
11	+0.0000391304	0.0258458043	+0.0011956147	0.0266501173
13	−0.0005027400	0.0258648633	+0.0000567313	0.0266473776
15	−0.0003475003	0.0258676014	+0.0000010952	0.0266473697
17	+0.0000005976	0.0258676227	+0.0000000058	0.0266473699
19	−0.0000005658	0.0258676364	$+3 \times 10^{-12}$	0.0266473699
21	+0.0000007006	0.0258676310	$+9 \times 10^{-18}$	0.0266473699
23	−0.0000264304	0.0258669199	-7×10^{-31}	0.0266473699
25	+0.0000294811	0.0258664159	-1×10^{-17}	0.0266473699
27	−0.0000893130	0.0258671105	-2×10^{-14}	0.0266473699
29	−0.0000015974	0.0258670759	-3×10^{-13}	0.0266473699
31	+0.0000095028	0.0258669560	-8×10^{-13}	0.0266473699
33	−0.0000308215	0.0258671935	-4×10^{-13}	0.0266473699
35	−0.0000030879	0.0258671376	-5×10^{-14}	0.0266473699
37	+0.0005511601	0.0258618885	-1×10^{-15}	0.0266473699
39	+0.0008794728	0.0258551889	-3×10^{-18}	0.0266473699
∞		0.0259094238		0.0266473631

and, for illustration, we have applied it to the self- and cross-fertilized plant data obtained by Darwin and discussed by Fisher [1], *viz.*

$$\{d_j\} = \{-67, -48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75\}$$

The exact value of p (for a one-tailed test) was computed by Fisher who considered all of the $N = 2^{15} = 32{,}768$ resamples and obtained $p = 849/32{,}768 = 0.0259094238$. Because $\max |t_r^* - t| = 858$, we scale the data by $3\pi/2860$. This yields $t = 471\pi/1430$ and the first 20 $\Psi(k)$ and partial sums of the series that result are listed in table 1. The $\Psi(k)$ initially decrease rapidly and the partial sum after three terms differs from the exact value by less than 0.0001. However, subsequent $\Psi(k)$ values are of variable sign and, although most are smaller than the random-phase estimate $N^{-1/2} \approx 0.0055$, they do not decay further and exhibit a spike whenever k is a multiple of 477. As a result, it is easy to obtain the first three or four decimal places of the exact p-value but many more terms in the series are needed if greater accuracy is required.

4. Bootstrap test for matched pairs

A bootstrap analogue of the Fisher permutation test considers the $N = (2n)^n$ resamples r obtained by sampling with replacement from the bootstrap space given by the union of set of differences d_q and their negatives $-d_q$ and seeks the fraction of these that yield $t_r^* > t$. By definition, we have

$$\Psi(k) = (2n)^{-n} \sum_{s_1=\pm1} \sum_{j_1=1}^{n} \cdots \sum_{s_n=\pm1} \sum_{j_n=1}^{n} \exp\left[ik \sum_{q=1}^{n} s_q d_{j_q} \right] = \left[\frac{1}{n} \sum_{j=1}^{n} \cos k d_j \right]^n$$

which shows that, as in the Fisher test, each $\Psi(k)$ can be evaluated exactly in $O(n)$ real floating-point operations. It is interesting to compare the bootstrap $\Psi(k)$ (the nth power of the mean of the cosines) with the analogous permutation $\Psi(k)$ (the product of the cosines) and it is easy to show that the two are identical to third order in the kd_j.

The bootstrap-test significance level is thus

$$p = \frac{1}{2} - \frac{2}{\pi} \sum_{k'=1}^{\infty} \frac{\sin kt}{k} \left[\frac{1}{n} \sum_{j=1}^{n} \cos kd_j \right]^n$$

We illustrate this by applying it to the Darwin data, after scaling them by $9\pi/14{,}390$ which leads to $t = 1413\pi/7195$. The first 20 $\Psi(k)$ and partial sums are shown in table 1. The $\Psi(k)$ decrease noticeably more rapidly than in the permutation test and, as a result, the partial sums converge more quickly. Although we cannot find the exact value by exhaustive enumeration (the bootstrap space is too large), we have found that the addition of a further 1000 terms of the series for p affects the partial sum only in the 9th decimal place.

5. Permutation test for two independent samples

Given two set of observations, $\{x_1, \ldots, x_m\}$ and $\{x_{m+1}, \ldots, x_n\}$, we define $t = x_1 + \cdots + x_m$. The Fisher permutation test considers the $N = {}_n\mathrm{C}_m$ resamples r formed by permuting the data between the sets and seeks the fraction of these that yield $t_r^* > t$.

By definition, we have

$$\Psi(k) = ({}_n\mathrm{C}_m)^{-1} \sum_{j_1=1}^{n} \sum_{j_2>j_1}^{n} \cdots \sum_{j_m>j_{m-1}}^{n} \exp\left[ik \sum_{q=1}^{m} x_{j_q} \right]$$

but this does not factorise because the resampling in this case is done without replacement. Nonetheless, if we define

$${}_a\Psi_b(k) = \sum_{j_1=1}^{a} \sum_{j_2>j_1}^{a} \cdots \sum_{j_b>j_{b-1}}^{a} \exp\left[ik \sum_{q=1}^{b} x_{j_q} \right]$$

then the pseudo-binomial recurrence relation

$${}_a\Psi_b(k) = {}_{a-1}\Psi_b(k) + {}_{a-1}\Psi_{b-1}(k) \exp(ikx_a)$$

can be used to generate $\Psi(k) \equiv {}_n\Psi_m(k)/N$ from the boundary values ${}_a\Psi_0(k) = 1$ and ${}_a\Psi_{a+1}(k) = 0$ in $O[m(n-m)]$ complex floating-point operations.

For illustration, we analyse the management scores given in table 2.1 of the textbook by Noreen [13], *viz.*

$$\{x_1, \ldots, x_{13}\} = \{10, 18, 22, 25, 25, 27, 28, 33, 34, 36, 37, 38, 38\}$$

$$\{x_{14}, \ldots, x_{47}\} = \{00, 07, 07, 10, 13, 17, 22, 22, 23, 25, 25, 25, 25, 25, 26, 26, 26,$$
$$26, 27, 27, 28, 28, 29, 30, 31, 31, 32, 34, 36, 36, 36, 39, 40, 40\}$$

Exhaustive examination of the $N = {}_{47}\mathrm{C}_{13} = 140{,}676{,}848{,}445$ possible resamples is a non-trivial task but is possible on a fast PC and one finds that 24,448,145,734 of these yield

Table 2. Two-sample significance tests (Noreen data, $n_A = 13$, $n_B = 34$).

	Permutation test ($N = {}_{47}C_{13} = 1.4 \times 10^{11}$)		Bootstrap test ($N = 47^{47} = 3.9 \times 10^{78}$)	
k	$\lvert\Psi(k)\rvert$	Partial sum	$\lvert\Psi(k)\rvert$	Partial sum
1	0.8997397012	0.2620946467	0.9779245468	0.3762969175
3	0.3790735618	0.1862707516	0.8180935447	0.2789269241
5	0.0601547051	0.1788258788	0.5729320155	0.2184755258
7	0.0030028372	0.1785535152	0.3364157254	0.1887321081
9	0.0000676002	0.1785545436	0.1660678581	0.1770293172
11	0.0000337348	0.1785528909	0.0692131305	0.1732871024
13	0.0000207736	0.1785519190	0.0245145575	0.1722867979
15	0.0000033650	0.1785517950	0.0074505327	0.1720528887
17	0.0000002173	0.1785517876	0.0019700177	0.1720020230
19	0.0000000012	0.1785517876	0.0004618448	0.1719911516
21	0.0000000094	0.1785517874	0.0000983990	0.1719888286
23	0.0000000469	0.1785517876	0.0000196342	0.1719883442
25	0.0000000406	0.1785517867	0.0000037933	0.1719882492
27	0.0000002382	0.1785517830	0.0000007326	0.1719882322
29	0.0000005052	0.1785517741	0.0000001449	0.1719882294
31	0.0000000329	0.1785517741	0.0000000297	0.1719882290
33	0.0000003338	0.1785517693	0.0000000063	0.1719882290
35	0.0000001374	0.1785517679	0.0000000013	0.1719882289
37	0.0000000133	0.1785517681	0.0000000003	0.1719882289
39	0.0000000176	0.1785517683	0.0000000001	0.1719882289
∞		0.1785761415		0.1719882289

$t_r^* > t$ and 1,346,766,114 yield $t_r^* = t$, so that the exact permutation significance level is $p = 0.1785761415$.

Alternatively, after scaling the data by $9\pi/1750$, we obtain $t = 477\pi/250$ and the $|\Psi(k)|$ and partial sums shown in table 2. As it was found in the Darwin example, the $\Psi(k)$ initially decay rapidly and then fluctuate. However, because the size of the resampling space is much larger for the Noreen data, the fluctuations are roughly three orders of magnitude smaller and the partial sums converge more rapidly.

6. Bootstrap test for two independent samples

Given two sets of observations, $\{x_1, \ldots, x_m\}$ and $\{x_{m+1}, \ldots, x_n\}$, we define

$$t = \frac{1}{m}\sum_{q=1}^{m} x_q - \frac{1}{n-m}\sum_{q=m+1}^{n} x_q$$

and the bootstrap test considers the $N = n^n$ alternative sets r obtained by resampling, with replacement, from the pooled data $\{x_1, \ldots, x_n\}$ and asks what fraction of these yield $t_r^* > t$.

By definition, we have

$$\begin{aligned}\Psi(k) &= n^{-n}\sum_{j_1=1}^{n}\cdots\sum_{j_n=1}^{n}\exp\left[ik\left(\frac{1}{m}\sum_{q=1}^{m}x_{j_q} - \frac{1}{n-m}\sum_{q=m+1}^{n}x_{j_q}\right)\right]\\ &= \psi\left(\frac{k}{m}\right)^m \psi\left(\frac{k}{m-n}\right)^{n-m}\end{aligned}$$

where

$$\psi(\omega) = \frac{1}{n}\sum_{j=1}^{n}\exp(i\omega x_j)$$

and thus, although $N = n^n$ can be very large, $\Psi(k)$ can be evaluated exactly in $O(n)$ complex floating-point operations.

We illustrate this approach by applying it to the Noreen scores. After scaling by $9\pi/400$, we obtain $t = 2817\pi/44{,}200$ and the $\Psi(k)$ and partial sums shown in table 2. Exhaustive examination of the full bootstrap space ($N > 10^{78}$) is clearly impossible but it is also clear that the series for p-value converges rapidly. As expected from the random-phase assumption, the $|\Psi(k)|$ values beyond $k = 39$ continue to decrease before eventually fluctuating around 10^{-39}. We can therefore be confident that the exact significance value is $0.1719882289\ldots$.

7. Conclusions

We have shown that it is possible to write the p-value for a variety of permutation and bootstrap significance tests as an infinite series whose terms can be computed rapidly, even when the group sizes are large. Moreover, if the resampling distribution is extremely large, such series often converge rapidly, allowing the p-value to be determined accurately with little computational effort. The series converges most rapidly when the intrinsic resolution of the data is high but convergence is usually satisfactory even when this is not the case.

Acknowledgements

I thank Dr Malcolm Gill for suggesting this problem to me, Dr Andrew Gilbert for useful discussions and Prof. Andy Wood and Dr Michael Martin for helpful suggestions concerning this manuscript.

References

[1] Fisher, R.A., 1935, *The Design of Experiments* (London: Oliver and Boyd).
[2] Pitman, E.J.G., 1937, Significance tests which may be applied to samples from any population. *Royal Statistical Society Supplement*, **4**, 119–130.
[3] Efron, B., 1979, Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.
[4] Edgington, E.S., 1995, *Randomization Tests* (New York: Marcel Dekker).
[5] Davison, A.C. and Hinkley, D.V., 1997, *Bootstrap Methods and Their Applications* (New York: Cambridge).
[6] Good, P.I., 2000, *Permutation Tests* (New York: Springer-Verlag).
[7] Pagano, M. and Tritchler, D., 1983, On obtaining permutation distributions in polynomial time. *Journal of American Statistical Association*, **78**, 435–440.
[8] Volsett, S.E., Hirji, K.F. and Elashoff, R.M., 1991, Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *Journal of American Statistical Society*, **86**, 404–409.
[9] Rayleigh, 1880, On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *Philosophical Magazine*, **10**, 73–78.
[10] Pearson, K., 1905, The problem of the random walk. *Nature*, **72**, 294, 342.
[11] Kluyver, J.C., 1905, A local probability problem. Proceedings Koninklijke Akademie van Wetenschappen te Amsterdam, **8**, 341–350.
[12] Hughes, B.D., 1995, *Random Walks and Random Environments* (Oxford: Clarendon).
[13] Noreen, E.W., 1989, *Computer Intensive Methods for Testing Hypotheses* (New York: John Wiley).