SIMPLE APPROXIMATIONS TO THE VARIATION NORM BETWEEN SAMPLING WITH AND WITHOUT REPLACEMENT

Chai Ho C. Wang, U.S. Department of Justice

The variation norm between sampling with and without replacement is given by Freedman [1]. Drawing a sample of k from a population of N ($0<k\leq N$), the variation norm is given by:

$$d = \|P-Q\| = 1-(N_k/N^k)$$

where P, Q are the induced probabilities for sampling with and without replacement respectively, and

$$N_k = N(N-1)\cdots(N-k+1)$$

Freedman's bounds for this norm are:

$$l_1 = 1-\exp[-k(k-1)/2N] < d < k(k-1)/2N = u_1$$

While l_1 is a fairly accurate approximation for d, u_1, as the first term of the Taylor series for l_1, is less adequate in representing d for certain combinations of N and k.

A simple approximation for d can be obtained via Stirling's formula for n!:

$$n! \simeq s_n = 2\pi n(n/e)^n$$

We have

$$\frac{N_k}{N^k} = \frac{N!}{(N-k)!N^k} \simeq e^{-k}\left(\frac{N}{N-k}\right)^{N-k+1/2}$$

and

$$l_3 = 1-e^{-k}\left(\frac{N}{N-k}\right)^{N-k+1/2} < d, \quad N\neq k$$

As will be justified later this formula provides accurate approximation for d. However, due to the involvement of large exponents, computational inconvenience and undesirable rounding error may arise. Taking an alternative approach, let us write, say, for k odd,

$$N(k) = \frac{N_k}{N^k} = \prod_{j=1}^{k-1}(1-j/N) = \prod_{j=1}^{(k-1)/2}\left(1-\frac{j}{N}\right)\left(1-\frac{k-j}{N}\right) = \prod_{j=1}^{(k-1)/2}\left(1-\frac{k}{N}+\frac{j(k-j)}{N^2}\right) \qquad (I)$$

Observing that

$$\min_{1\leq j\leq\frac{k-1}{2}} \frac{j(k-j)}{N^2} = \frac{k-1}{N^2},$$

$$\max_{1\leq j\leq\frac{k-1}{2}} \frac{j(k-j)}{N^2} = \frac{k^2-1}{4N^2}$$

we have

$$\left(1-\frac{k}{N}+\frac{k-1}{N^2}\right)^{(k-1)/2} < N(k) < \left(1-\frac{k}{N}+\frac{k^2-1}{4N^2}\right)^{(k-1)/2}$$

Further simplifications lead to the following two bounds for d:

$$l_2 = 1-\left(1-\frac{\dot{k}}{2N}\right)^{\ddot{k}} < d$$

and

$$u_2 = 1-\left(1-\frac{\dot{k}}{N}\right)^{\ddot{k}/2} > d$$

where $\dot{k}$ = the greatest odd integer less than or equal to k, and $\ddot{k}$ = the greatest even integer less than or equal to k.

Since

$$\sum_{j=1}^{(k-1)/2}\frac{j(k-j)}{N^2} = \frac{k(k^2-1)}{12N^2},$$

sharper approximations can be obtained by expanding (I). We have

$$u_3 = 1-(1-\dot{k}/N)^{(\ddot{k}/2)-1}\left[1-\frac{\dot{k}}{N}+\frac{\dot{k}(\dot{k}^2-1)}{12N^2}\right] > d$$

For practical computation purpose where N is large, u_3 is sufficiently accurate that no measurable precision will be gained by calculating the exact solution. To pursue this argument one step further, we note that the leading order term discarded when deriving u_3, again for k odd, is bounded above by

$$\hat{u}_3 = (1-k/N)^{(k-5)/2}\prod_{i<j}^{(k-1)/2}\frac{i(k-i)\cdot j(k-j)}{N^4} < 1/2(1-k/N)^{(k-5)/2}\left[\sum_{j=1}^{(k-1)/2}\frac{j(k-j)}{N^2}\right]^2 < (1-k/N)^{(k-5)/2}\frac{k^6}{288N^4}$$

This upper bound for $\hat{u}_3$ can be maximized at $k \simeq \sqrt{12N}$. It follows that

$$\hat{u}_3 < 6g(N)/N$$

where

$$g(N)=(1-\sqrt{12/N})(\sqrt{12N}-5)/2$$

g(N) is an increasing function of N. It increases very slowly for N up to, say, 10^{20}. Consequently,

$$\hat{u} << \frac{1}{N}$$

for all k, and for almost all N.

Let the relative error for approximating d by d' be defined by R=(d'-d)/d'. The difference between R and the usual relative error r=(d'-d)/d is small if R and r are small, because $|r| - |R| \leq |r-R| = |rR|$. The analysis of the trancation error for u_3 can be extended to the following proposition:

Proposition. The relative error for calculating N_k/N^k while using the (approximation) formulas associated with 1_1, u_1, 1_2, u_2, 1_3, and u_3 are $O[k^3/N(N-k)]$, $O[k^4/N(2N-k^2)]$, $O[k^3/N(N-k)]$, $O[k^3/N(N-k)]$, $O[k/N(N-k)]$, and $O[k^6/N^2(N-k)^2]$ respectively.

Proof: Since

$$\frac{\hat{u}_3}{u_3} < \frac{\hat{u}_3}{(1-k/N)^{(k-1)/2}} < \frac{k^6}{288N^2(N-k)^2}$$

the proposition for the case u_3 has been established. Similar leading-order-term argument is sufficient for proving all other cases except the case 1_3. For this case we note that

$$\ln(n!)-\ln s_n = \frac{1}{12n} - \frac{1}{360n^2} + \ldots < \frac{1}{12n}$$

then

$$0 < R = 1 - \frac{N!/s_N}{(N-k)!/s_{N-k}}$$

$$< 1 - \exp\left(\frac{1}{12N} - \frac{1}{12(N-k)}\right)$$

$$< \frac{k}{12N(N-k)}$$

as was to be shown.

From this proposition, it is apparent that both 1_3 and u_2 provide accurate approximation for d. Although as indicated in the proposition, 1_3 should be favored whenever $k^5 > 24N(N-K)$, only a rigorous analysis of rounding errors, which is outside the scope of this paper, can decide whether and when 1_3 is computationally more useful than u_3.

Table 1 displays numerical results computed with 16 (decimal) digit accuracy. The "exact" solution is computed by brute force using

$$\hat{d} = 1 - \prod_{j=1}^{k-1} (1 - \frac{j}{N})$$

For a fixed N, as k increases, all approximations except u_1 rapidly approach unity.

Relative computing error

$$\hat{R} = (\hat{d}-d')/d' = R + \text{rounding error}$$

for these approximations are found in Table 2. In certain cases, the rounding error becomes dominate over the relative error. For computation purpose u_3 appears to be a better approximation.

Reference

1. Freedman, D. (1977), "A Remark on the Difference Between Sampling With and Without Replacement" Journal of the American Statistical Association, Volume 72, 681.

Table 1

Approximations of $d = 1 - N_k/N^k$
[$a(J) = a \times 10^J$]

N	k	l_1	l_2	l_3	d	u_3	u_2	u_1
100	10	0.36237	0.36899	0.37179	0.37184	0.37185	0.37597	0.45000
1000	10	0.44003(-1)	0.44100(-1)	0.44139(-1)	0.44139(-1)	0.44139(-1)	0.44197(-1)	0.45000(-1)
	100	0.99292	0.99376	0.99404	0.99404	0.99406	0.99455	0.49500
10000	10	0.44899(-2)	0.44909(-2)	0.44913(-2)	0.44913(-2)	0.44913(-2)	0.44919(-2)	0.45000(-2)
	100	0.39043	0.39118	0.39143	0.39143	0.39143	0.39193	0.49500
	1000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.49950(+2)
100000	10	0.44990(-3)	0.44991(-3)	0.44991(-3)	0.44991(-3)	0.44991(-3)	0.44992(-3)	0.45000(-3)
	100	0.48295(-1)	0.48307(-1)	0.48310(-1)	0.48310(-1)	0.48310(-1)	0.48318(-1)	0.49500(-1)
	1000	0.99323	0.99331	0.99334	0.99334	0.99334	0.99340	0.49950
	10000	1.00000	1.00000	1.0000	1.00000	1.0000	1.00000	1.00000
1000000	10	0.44999(-4)	0.44999(-4)	0.44999(-4)	0.44999(-4)	0.44999(-4)	0.44999(-4)	0.45000(-4)
	100	0.49378(-2)	0.49379(-2)	0.49379(-2)	0.49379(-2)	0.49379(-2)	0.49380(-2)	0.49500(-2)
	1000	0.39317	0.39324	0.39327	0.39327	0.39327	0.39332	0.49950
	10000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
10000000	10	0.45000(-5)	0.45000(-5)	0.45022(-5)	0.45000(-5)	0.45000(-5)	0.45000(-5)	0.45000(-5)
	100	0.49488(-3)	0.49488(-3)	0.49488(-3)	0.49488(-3)	0.49488(-3)	0.49488(-3)	0.49500(-3)
	1000	0.48723(-1)	0.48724(-1)	0.48725(-1)	0.48725(-1)	0.48725(-1)	0.48725(-1)	0.49950(-1)
	10000	0.99326	0.99327	0.99327	0.99327	0.99327	0.99328	0.49995(+1)
	100000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.50000(+3)
100000000	10	0.45000(-6)	0.45000(-6)	0.46717(-6)	0.45000(-6)	0.45000(-6)	0.45000(-6)	0.45000(-6)
	100	0.49499(-4)	0.49499(-4)	0.48520(-4)	0.49499(-4)	0.49499(-4)	0.49499(-4)	0.49500(-4)
	1000	0.49825(-2)	0.49826(-2)	0.49826(-2)	0.49826(-2)	0.49826(-2)	0.49826(-2)	0.49950(-2)
	10000	0.39344	0.39345	0.39345	0.39345	0.39345	0.39345	0.49995
	100000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.50000(+2)
1000000000	10	0.45000(-7)	0.45000(-7)	0.20577(-6)	0.45000(-7)	0.45000(-7)	0.45000(-7)	0.45000(-7)
	100	0.49500(-5)	0.49500(-5)	0.51216(-5)	0.49500(-5)	0.49500(-5)	0.49500(-5)	0.49500(-5)
	1000	0.49938(-3)	0.49938(-3)	0.49959(-3)	0.49938(-3)	0.49938(-3)	0.49938(-3)	0.49950(-3)
	10000	0.48766(-1)	0.48766(-1)	0.48766(-1)	0.48766(-1)	0.48766(-1)	0.48766(-1)	0.49995(-1)
	100000	0.99326	0.99326	0.99326	0.99326	0.99326	0.99326	0.50000(+1)
	1000000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.50000(+3)

Table 2

Relative Computing Error

$R = (d'-d)/d'$

N	k	l_1	l_2	l_3	u_3	u_2	u_1
100	10	-0.26(-1)	-0.77(-2)	-0.16(-3)	0.27(-4)	0.11(-1)	0.17
1000	10	-0.31(-2)	-0.90(-3)	-0.18(-4)	0.29(-7)	0.13(-2)	0.19(-1)
	100	-0.11(-2)	-0.28(-3)	-0.56(-7)	0.22(-4)	0.51(-3)	0.80
10000	10	-0.32(-3)	-0.92(-4)	-0.18(-5)	0.29(-10)	0.13(-3)	0.19(-2)
	100	-0.26(-2)	-0.65(-3)	-0.13(-6)	0.51(-6)	0.13(-2)	0.21
	1000	0.00	0.00	0.00	0.00	0.00	0.98
100000	10	-0.32(-4)	-0.92(-5)	-0.17(-6)	-0.31(-13)	0.13(-4)	0.19(-3)
	100	-0.32(-3)	-0.82(-4)	-0.16(-7)	0.63(-9)	0.16(-3)	0.24(-1)
	1000	-0.11(-3)	-0.28(-4)	-0.56(-10)	0.23(-6)	0.56(-4)	0.80
	10000	0.00	0.00	0.00	0.00	0.00	1.00
1000000	10	-0.32(-5)	-0.92(-6)	0.83(-6)	0.31(-12)	0.13(-5)	0.19(-4)
	100	-0.33(-4)	-0.84(-5)	0.15(-7)	0.53(-12)	0.16(-4)	0.24(-2)
	1000	-0.26(-3)	-0.64(-4)	0.77(-10)	0.53(-8)	0.13(-3)	0.21
	10000	0.00	0.00	0.00	0.00	0.00	0.98
10000000	10	-0.32(-6)	-0.92(-7)	0.48(-3)	-0.31(-11)	0.13(-16)	0.19(-5)
	100	-0.33(-5)	-0.84(-6)	0.77(-6)	-0.81(-12)	0.16(-5)	0.24(-3)
	1000	-0.32(-4)	-0.81(-5)	0.16(-7)	0.67(-11)	0.16(-4)	0.25(-1)
	10000	-0.11(-4)	-0.28(-5)	0.90(-11)	0.24(-8)	0.56(-5)	0.80
	100000	0.00	0.00	0.00	0.00	0.00	1.00
100000000	10	-0.32(-7)	-0.94(-8)	0.37(-1)	-0.31(-10)	0.13(-7)	0.19(-6)
	100	-0.33(-6)	-0.84(-7)	0.44(-3)	-0.76(-11)	0.16(-6)	0.24(-4)
	1000	-0.33(-5)	-0.83(-6)	0.76(-6)	-0.13(-11)	0.17(-5)	0.25(-2)
	10000	-0.26(-4)	-0.64(-5)	0.13(-7)	0.53(-10)	0.13(-4)	0.21
	100000	0.00	0.00	0.00	0.00	0.00	0.98
1000000000	10	-0.40(-8)	-0.22(-8)	0.78	-0.12(-8)	0.00	0.19(-7)
	100	-0.33(-7)	-0.87(-8)	0.34(-1)	-0.11(-9)	0.16(-7)	0.24(-5)
	1000	-0.33(-6)	-0.83(-7)	0.43(-3)	-0.12(-10)	0.17(-8)	0.25(-3)
	10000	-0.33(-5)	-0.81(-6)	0.74(-6)	-0.29(-12)	0.16(-5)	0.25(-1)
	100000	-0.11(-5)	-0.28(-6)	0.56(-9)	-0.24(-10)	0.57(-6)	0.80
	1000000	0.00	0.00	0.00	0.00	0.00	1.00