Average correlations vs. correlated averages

WILLIAM P. DUNLAP

Tulane University, New Orleans, Louisiana 70118

MARSHALL B. JONES

Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033

and

ALVAH C. BITTNER

Naval Biodynamics Laboratory, New Orleans, Louisiana 70198

When repeated-measures data are collected with small or even moderate sample sizes, correlation matrices show considerable variability. If one's primary interest is in the correlations, some means of smoothing the coefficients may be desirable. Two methods of smoothing Pearson r s were investigated in the present study. The first method was to average repeated measures in blocks and then correlate block averages. In the second method, all repeated measures were correlated and then the correlation coefficients were averaged in blocks. The latter approach proved much superior, resulting in greatly reduced sampling variability and little distortion in the population correlation estimated.

In research involving repeated observations on the same group of subjects, the correlations between the various repeated measures are often of primary experimental interest. Frequently, however, one encounters data sets in which there are many repeated measures but rather few subjects. Given constraints on time and expense, the number of repeated measures and number of subjects in practical research are probably reciprocally related. In such cases, the correlation matrix among repeated measures may be quite variable, frustrating attempts to make general statements about the matrix structure or attempts to decompose that structure by techniques such as factor analysis. With small to moderate sample sizes, it may be of considerable benefit to smooth the correlation coefficients prior to examining the pattern of relationships. This poses the interesting question of how best to smooth these coefficients, improving their reliability while at the same time maintaining their representation of the underlying relationships.

Two possible approaches to this problem were examined by means of a computer simulation followed by some approximate mathematical predictions of the statistical properties of the two methods. The first approach examined was to average the data in blocks of

measures prior to calculating the statistic of interest. This procedure has often been used in psychological research to smooth consecutive observations, and we were interested in how blocking of repeated measures would affect the Pearson correlation coefficient. Blocking of data prior to computing correlations is a quite common procedure in many areas of psychology (as examples of this practice in the area of motor learning, see Bilodeau, 1955; Fleishman, 1960; Reynolds, 1952). The second, less obvious, approach examined was to find the simple correlations between all individual repeated measures and then to average these correlations in blocks. A problem with this approach is that the sampling distribution of r is markedly skewed when the population parameter differs substantially from 0, and thus the average r is biased toward more moderate values, underestimating the true population correlation (Kendall & Stuart, 1979). For this reason, the correlations were converted via Fisher's z transformation, which almost entirely corrects skew, prior to averaging.

METHOD

The first step in investigating the properties of correlations of averages as compared to average correlations involved a simulation. The computer was programmed to generate 12 simulated "observations" having a known common population correlation for each of N subjects. Pseudorandom normal deviates (mean = 0, SD = 1) were generated using Box and Muller's (1958) procedure. Seeds for the normal deviates were pseudorandom rectangular fractions generated by the DECsystem 20 internal function RAN (see Edgell, 1979, for statistical characteristics). First, 12 columns of independent normal data, X_{ij} , were produced: then a 13th column, W_i , was generated and was added to each of the original 12 columns to create new intercorrelated

This study was supported in part by Contract N00014-80-C-0715 from the Office of Naval Research. The findings and opinions contained in this report are those of the authors and should not be construed as an official position, policy, or decision of the Navy, unless so designated by other official documentation. Reprint requests should be addressed to William P. Dunlap, Department of Psychology, Tulane University, New Orleans, Louisiana 70118.

variables, Y_{ij} , by the following formula: $Y_{ij} = (1 - r)^{1/2} X_{ij} + r^{1/2} W_i$. It can be shown that the resulting columns, Y_{ij} , have population correlations equal to r .

The 12 measures were then averaged in blocks of 2, 3, 4, or 6, and the correlations between block means computed. Thus, for blocks of three, the mean of Observations 1, 2, and 3 was correlated with the means of Observations 4, 5, and 6, Observations 7, 8, and 9, and Observations 10, 11, and 12. The program also computed the correlations between all individual observations without blocking (blocks of one); then these correlations were averaged in blocks of two, three, four, or six. Thus, when the blocking factor was three, the average correlation for Blocks 1 and 2 was the mean of nine correlations, $r_{14}, r_{15}, r_{16}, r_{24}, r_{25}, r_{26}, r_{34}, r_{35},$ and r_{36} . Whenever correlations were averaged, they were first converted by Fisher's z transformation to avoid distortions in the means due to the skewed sampling distribution of r : $z = .5 \ln [(1 + r)/(1 - r)]$. After averaging, z values were converted back to r values by the reverse transformation, $r = [\exp(2z) - 1] / [\exp(2z) + 1]$.

Five hundred data sets were generated for each combination of sample size (10, 30, or 80) and population correlation (.2, .5, or .8). For each data set, all possible interblock correlations were computed, that is, 66 correlations for "blocks" of one observation, 15 for blocks of two observations, 6 for blocks of three observations, 3 for blocks of four observations, and 1 correlation for blocks of six observations. The means and standard deviations (standard errors) of all correlations resulting from a given blocking, over all data sets, were then computed, again using Fisher's z . The results are shown in Table 1.

RESULTS

As can be seen in Table 1, the effect of averaging the data in blocks prior to correlating produces an increase in the block-to-block correlation. This increase in correlation as a function of block size is entirely consistent with the Spearman-Brown formula, $r(B) = kr/[1 + (k - 1)r]$, where $r(B)$ is the correlation of block means, k is the number of observations per block, and r is the population correlation between observations. For

example, when the correlation equals .2, the columns headed $r(B)$ in Table 1 should approximately equal .20, .33, .43, .50, and .60, respectively, for blocks of Sizes 1, 2, 3, 4, and 6. As can be seen, when the sample size is 10, there is a tendency to overestimate the theoretical values, but as the sample size increases, the correlations between blocks closely approach their theoretical values. The standard errors of the $r(B)$ s, presented in the next column of Table 1, depend only on the sample size, and since Fisher's z transformation was used, the standard errors should approximately equal $1/(N - 3)^{1/2}$. Thus, the standard errors should equal .378, .192, and .114 for samples of 10, 30, and 80, respectively. Since r is always calculated from the same number of data points, there is no gain in precision when the data are averaged in blocks prior to correlating.

The next column of Table 1, headed $\bar{r}$, shows the means of correlations averaged in blocks. As one would expect, these values simply approximate the known population correlation, with the exception that with small N there is a tendency for overestimation. This small bias is not unexpected, since it is known that the mean of Fisher's z s is biased in this direction (Kendall & Stuart, 1979). Thus, the averaged r s are independent of the blocking factor, and the bias toward overestimation is not great even with small sample sizes. The next column of Table 1, which shows the standard errors of the $\bar{r}$ s, is both the most interesting and the most complicated. Clearly, these standard errors decrease as a function of the blocking factor, and thus there is a gain in precision by blocking. This gain in precision, however, is a function of both the blocking factor and the population correlation. When the population correlation is 0, the standard error should equal $1/[k(N - 3)^{1/2}]$,

Table 1
Correlations Between Block Means, $r(B)$, and Correlations Averaged in Blocks, $\bar{r}$, With Respective Standard Errors
for Population Correlations of .2, .5, and .8 and Sample Sizes (N) of 10, 30, and 80

Blocking Factor	N = 10				N = 30				N = 80			
	$r(B)$	SE	$\bar{r}$	SE	$r(B)$	SE	$\bar{r}$	SE	$r(B)$	SE	$\bar{r}$	SE
Population Correlation $r = .2$												
1	.209	.376	.209	.376	.201	.194	.201	.194	.199	.114	.199	.114
2	.352	.376	.210	.221	.335	.193	.201	.115	.331	.112	.199	.066
3	.446	.378	.210	.173	.430	.194	.201	.090	.428	.113	.199	.052
4	.524	.379	.211	.148	.503	.194	.201	.078	.498	.111	.199	.044
6	.626	.378	.210	.128	.606	.188	.201	.065	.596	.114	.198	.037
Population Correlation $r = .5$												
1	.515	.373	.515	.373	.508	.192	.508	.192	.500	.114	.500	.114
2	.682	.372	.515	.259	.673	.192	.508	.135	.667	.115	.500	.080
3	.765	.375	.514	.224	.758	.187	.508	.116	.750	.113	.500	.069
4	.812	.373	.514	.208	.806	.192	.508	.109	.800	.115	.500	.064
6	.867	.374	.514	.193	.862	.185	.508	.101	.857	.114	.500	.060
Population Correlation $r = .8$												
1	.821	.369	.821	.369	.808	.190	.808	.190	.803	.114	.803	.114
2	.902	.373	.821	.286	.893	.188	.808	.148	.891	.113	.803	.090
3	.933	.364	.821	.261	.927	.184	.808	.136	.925	.115	.803	.083
4	.948	.376	.820	.252	.944	.184	.808	.130	.943	.116	.803	.080
6	.965	.362	.821	.240	.961	.183	.808	.125	.961	.112	.803	.077

since k^2 independent estimates of r are averaged. When the population correlation is other than 0, the estimates of r are themselves intercorrelated, which affects the standard errors in such a way that less precision is gained the larger the population correlation.

Steiger (1980) presents asymptotic expressions for the variance-covariance matrix between the k^2 correlations to be averaged; however, he did not directly address the issue of average correlations and their standard errors in small samples. There are k^2 variances, $2k^2(k-1)$ covariances between r s sharing one common index, and $k^2(k-1)^2$ covariances between r s with no common indexes. Substituting the fact that all correlations between observations in the present problem are equal into Steiger's equations, one obtains a variance of $(1-r^2)^2$, a covariance of $r/2(2-r-4r^2+3r^3)$ for correlations sharing a common index, and a covariance of $2r^2(1-r)^2$ for correlations with no common index. To correct for the fact that Fisher z s were used, these quantities must be divided by $(N-3)(1-r^2)^2$. The variance of the average r , $V(r)$, may be found by pre- and postmultiplying the variance-covariance matrix by vectors containing $1/k^2$; thus, $V(r) = [(1-r^2)^2 +$

$2(k-1)r/2(2-r-4r^2+3r^3) + (k-1)^2 2r^2(1-r)^2] / [k^2(N-3)(1-r^2)^2]$. The standard error is, of course, the square root of the above expression.

Using the above formula, Table 2 was constructed; it gives theoretical values for the standard errors of correlations averaged in blocks under a greater variety of sample sizes and population correlations than were approximated in Table 1. Further, Table 2 presents a second column headed EN, which contains the sample size required to produce the same accuracy in estimating the population correlation had sample r s not been averaged; thus Column 2 provides an estimate of the "effective sample size" or efficiency gained in averaging. These values were calculated from the following formula, $EN = 1/SE^2 + 3$, which is simply the back solution from the standard error of Fisher's z , where SE is the standard error and EN is the effective sample size.

Comparing these theoretical values from Table 2 with the empirical values of Table 1, particularly when the sample size is smallest, one sees that the theoretical values for the standard error tend to be slightly larger and thus tend to be somewhat conservative in the case of small samples. It is clear that substantial efficiency is

Table 2
Standard Errors and Effective Sample Sizes for Correlations Averaged Over Blocks of Trials

Blocking Factor	$\bar{r} = 0$		$\bar{r} = .1$		$\bar{r} = .2$		$\bar{r} = .4$		$\bar{r} = .6$		$\bar{r} = .8$		$\bar{r} = .9$	
	SE	EN	SE	EN	SE	EN	SE	EN	SE	EN	SE	EN	SE	EN
N = 10														
1	.378	10	.378	10	.378	10	.378	10	.378	10	.378	10	.378	10
2	.189	31	.208	26	.225	23	.255	18	.279	16	.298	14	.306	14
3	.126	66	.152	47	.176	35	.217	24	.249	19	.275	16	.286	15
4	.094	115	.124	68	.152	46	.199	28	.235	21	.264	17	.276	16
6	.063	255	.097	110	.129	63	.182	33	.223	23	.254	18	.268	17
N = 20														
1	.243	20	.243	20	.243	20	.243	20	.243	20	.243	20	.243	20
2	.121	71	.133	59	.144	51	.163	40	.179	34	.191	30	.196	29
3	.081	156	.097	109	.113	82	.139	55	.160	42	.176	35	.183	33
4	.061	275	.079	161	.097	108	.128	64	.151	47	.170	38	.177	35
6	.040	615	.062	262	.083	149	.117	76	.143	52	.163	41	.172	37
N = 30														
1	.192	30	.192	30	.192	30	.192	30	.192	30	.192	30	.192	30
2	.096	111	.106	93	.115	79	.130	62	.142	53	.152	47	.156	44
3	.064	246	.077	171	.089	128	.110	85	.127	65	.140	54	.145	50
4	.048	435	.063	254	.077	170	.101	101	.120	73	.134	58	.141	54
6	.032	975	.049	414	.066	235	.093	119	.113	81	.130	63	.136	57
N = 50														
1	.146	50	.146	50	.146	50	.146	50	.146	50	.146	50	.146	50
2	.073	191	.080	159	.087	136	.098	107	.107	90	.115	79	.118	75
3	.049	426	.058	295	.068	221	.084	146	.096	111	.106	92	.110	85
4	.036	755	.048	440	.059	294	.077	173	.091	124	.102	99	.107	91
6	.024	1695	.037	719	.050	406	.070	206	.086	139	.098	107	.103	97
N = 100														
1	.102	100	.102	100	.102	100	.102	100	.102	100	.102	100	.102	100
2	.051	391	.056	325	.060	277	.068	217	.075	182	.080	159	.082	151
3	.034	876	.041	607	.047	452	.058	298	.067	226	.074	187	.077	173
4	.025	1555	.033	906	.041	604	.053	353	.063	253	.071	202	.074	184
6	.017	3495	.026	1480	.035	836	.049	421	.060	283	.068	217	.072	196

gained in averaging correlations, especially when the population correlation is small and the need for power is greatest.

DISCUSSION

First, a consideration of the limitations of the present study is in order. Our simulation covers a special case. Considerable work remains to be done by way of analyzing effective sample sizes for other specific situations, such as correlations between two sets of observations on different measures that may be differently correlated among themselves and across measures. The major point, however, is clear. Average correlations have much smaller standard errors (much larger effective N s) than correlations resulting from the commonly adopted practice of first averaging observations.

The present paper is concerned only with the correlations of one block of observations with another block of observations; thus, in calculating r , $k \times k$ correlations are averaged. These derivations are not applicable to the average intercorrelations within a single block of observations, in which there are $k(k-1)/2$ correlations to average and thus different counting of the various correlation types is necessary. This problem is addressed by Bittner (Note 1). Also, the present paper does not treat the issue of a block of observations of one task correlated with a block of observations of a second task. Here, some correlations involve different observations of the same task, whereas others involve different tasks; the latter will tend to be of a lower magnitude. Preliminary investigation suggests lower efficiencies of the averaging process under these conditions (Bittner, Dunlap, & Jones, Note 2).

Last, it may appear that we overlooked an obvious way to improve some aspects of $r(B)$, the coefficient obtained by correlating means of blocks of observations. The Spearman-Brown formula is easy to reverse; thus from a correlation of block means, one can estimate the unblocked correlation as $r = r(B)/[k - (k-1)r(B)]$. Preliminary simulations were done using this back transformation on $r(B)$, and two things were discovered. First, the mean of the back-transformed $r(B)$ s was badly biased, especially for smaller population correlations. This bias could perhaps be corrected statistically; however, ignoring the bias, the efficiencies were not nearly as great as the efficiencies of averaged correlations.

The conclusions then are rather clear. If one desires to smooth correlations, maintaining the representation of a single observation's correlation with another while improving the coefficient's accuracy or efficiency, the average correlation is superior to the correlation of averages. It can be argued that the size of $r(B)$, since it changes with the blocking factor, is determined by the experimenter's fancy, whereas r , since it remains

almost constant across blocking, is a fairer representation of the actual experimental unit used, a single measure's correlation with another single measure. Also, $r(B)$ fails to satisfy another primary goal of smoothing, that of improving accuracy or efficiency. One simply arrives at a bigger correlation with the same inherent instability using $r(B)$, whereas the efficiency of r increases dramatically, especially with lower population correlations in which greatest stability is really needed. Therefore, if one's interest is in the pattern of correlations across observations or a factor analysis of repeated-measures data, and smoothing prior to analysis appears warranted, averaging simple correlations in blocks seems most practical.

REFERENCE NOTES

1. Bittner, A. C., Jr. Averaged correlations between parallel measures: Reliability estimation. *Proceedings of the Eighth Psychology in DoD Symposium*, USAF Academy, Colorado Springs, Colorado, 1982.
2. Bittner, A. C., Jr., Dunlap, W. P., & Jones, M. B. Averaged cross-correlations with differentially stable variances: Fewer subjects required with repeated measures. *Proceedings of the 26th annual meeting of the Human Factors Society*, Seattle, Washington, 1982.

REFERENCES

- BILODEAU, E. A. Variations in knowledge of component performance and its effects on part-part and part-whole relations. *Journal of Experimental Psychology*, 1955, **50**, 215-224.
- BOX, G. E. P., & MULLER, N. E. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 1958, **29**, 610-611.
- EDGELL, S. E. A statistical check of the DECsystem-10 FORTRAN pseudorandom number generator. *Behavior Research Methods & Instrumentation*, 1979, **11**, 529-530.
- FLEISHMAN, E. A. Abilities at different stages of practice in rotary pursuit performance. *Journal of Experimental Psychology*, 1960, **60**, 162-171.
- KENDALL, M., & STUART, A. *The advanced theory of statistics* (Vol. 1). New York: Macmillan, 1979.
- REYNOLDS, B. The effect of learning on the predictability of psychomotor performance. *Journal of Experimental Psychology*, 1952, **44**, 189-198.
- STEIGER, J. H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 1980, **87**, 245-251.

(Received for publication March 2, 1983.)