Simulating Pearson's and Spearman's Correlations in Q-Sorts Using Excel: A Simulation Proof of A Widely Believed Result

K. Scott Alberts and Bruce Ankenmann

Social Science Computer Review 2001 19: 221
DOI: 10.1177/089443930101900208

Simulating Pearson's and Spearman's Correlations in Q-Sorts Using Excel

A Simulation Proof of A Widely Believed Result

K. SCOTT ALBERTS
Truman State University

BRUCE ANKENMANN
Northwestern University

A widely believed result about Q-sorts is shown by simulation; namely, that the hard-to-compute but statistically correct Spearman's ranked correlation, r_s, may be substituted by the more common Pearson's *r* correlation. Recent versions of Microsoft Excel are excellent for reasonably sized simulations, especially those used in classroom settings.

Keywords: simulation, Q-methodology, Q-sorts, Excel, correlation

Simulation is commonly used in engineering and other situations as a manner of finding quick answers to mathematically challenging (and perhaps unsolvable) problems. For instance, the arrival and service of a complicated queuing network, such as a hospital emergency room, is easily described through simulation but very difficult to describe using direct probabilistic queuing theory, except under trivial assumptions (Kelton, Sadowski, & Sadowski, 1998). The use of simulation is less frequent in the social sciences but can still be useful (Garson, 1994).

Q-methodology is an area of social research where we propose that simulation could be quite useful. In short, Q-methodology treats data from an individual as an entire vector, rather than doing an analysis of each item.

A Q-sort is a common tool in Q-methodology and can also be used outside of the methodology. It is a questionnaire variant in which participants are asked to complete a pseudo ranking of the items. To compare to the traditional questionnaire, consider the instrument modeled here. Sixty-four items are rated on a 7-point scale. However, instead of having a free distribution, exactly five items must be placed in each extreme category, eight in the next, and so forth, creating a 5-8-12-14-12-8-5 fixed distribution when completed. Brown (1980) and Block (1979) give more detailed information on the technical aspects of Q-sorts.

Two common ways of doing the vector analysis in Q-sorts are factor analysis and correlation. Correlation among subjects is the more simple and intuitive method of analysis. Pearson's *r* is the correlation taught in most introductory statistics courses, so most students are already aware of it; however, it is only appropriate for normal, continuous data. Spearman's rank correlation, r_s, which performs a standard Pearson correlation on ranks of data, is recommended for ordinal data such as that found in Q-sorts (Gibbons, 1985). Pearson's *r* uses about one third the resources and is less complex than Spearman's r_s, mak-

Social Science Computer Review, Vol. 19 No. 2, Summer 2001 221-226

ing it more useful in classroom settings. However, as Q-data has already been somewhat sorted, one could claim that the Pearson's *r* would be acceptable for most uses.

Brown (1971, 1980) claims that Pearson's *r* is an accurate and unbiased estimate of Spearman's r_s. He demonstrates that Pearson is unbiased but gives no proof to back his claim of accuracy, instead referring to several real-world studies of small sample size; however, most practitioners take this claim as a given. It would be useful to have a mathematical proof of some kind to reinforce this claim. Besides supporting a claim that most practitioners use and believe, this example would be excellent for classroom situations, especially methods courses that might otherwise cover Q-methodology, but do not want to take the time to teach Spearman's r_s.

In this exercise, 250 statistically simulated subjects are generated from random numbers that are then correlated by both metrics. Examining the distribution of differences between the sets of correlations will determine the accuracy of the estimation. In addition, the two sets of correlations may be correlated themselves, giving another measure of their similarity.

For this problem, a 64-item Q-sort with seven responses is simulated. It uses the forced distribution 5-8-12-14-12-8-5, explained above. Note that analyzing this phenomenon using pure probability techniques would be impossible due to both its complexity and sheer size. The model contains hundreds of 64-dimensional vectors, each of which is generated from a nonindependent pseudo-normal distribution.[1] Very little is known about the actual nature of nonindependent distributions, as most probability problems assume sets are independent and identically distributed. As a simulation, however, this problem is only slightly more than trivial.

TECHNICAL METHOD

Excel is used both to demonstrate its use as a teaching tool and as an actual vehicle for simulation.[2] It is simple to use, and most students and faculty already possess a copy on their personal computer.

First, a matrix of random numbers is needed. Here, the first worksheet (named "R#" in this example) is used to create (using the command `RND()`) a 64×250 matrix of uniform 0-to-1 random numbers.[3] The second worksheet, "ranks," is used to rank each row of "R#" from 1 to 64, using the `RANK(cell, array)` command.

A total of 10 worksheets is needed, one for each step in the simulation process. A description of each worksheet can be found in Figure 1, along with a sample command and sample data from an arbitrary entry, Z50.

The third worksheet, "composites," converts these columns to the fixed distribution required by the Q-sort, using the fixed distribution of 5-8-12-14-12-8-5 and arbitrary scores of –3, –2, –1, 0, 1, 2, and 3, respectively. This command line examines the ranks calculated above. It assigns the lowest five ranks to –3, the next eight to –2, and so on, using a nest of `IF` and `AND` commands. The fourth worksheet, "rankcomps," reranks the Q-data, using the rules for ties required by Spearman correlation.

"Composites" contains simulated Q-data. The data here look identical to what we would have after administering our instrument to random respondents. We now have the actual Q-data for the Spearman correlation and the ranked data required by the Spearman correlation. Now both correlation matrices can be generated and analyzed. The remainder of the worksheets do this analysis.

Performing row correlation on the worksheet "composite" yields the Pearson's *r* correlation matrix in a worksheet "Pearson," whereas using the worksheet "rankcomps" generates

Worksheet	Description	Sample Command (in entry Z50)	Sample Data (in Z50)
R#	Table of Random Numbers	=RAND()	0.514176043
Ranks	Rank within each row from "R#"	=RANK('R#'!Z50,'R#'!$A50:$BL50)	26
Comp	Transforms "Ranks" into Q-composite scores	=IF(Ranks!Z50<6,-3,IF(AND(Ranks!Z50>5,Ranks!Z50<14),-2, IF(AND(Ranks!Z50>13,Ranks!Z50<26),-1, IF(AND(Ranks!Z50>25,Ranks!Z50<40),0, IF(AND(Ranks!Z50>39,Ranks!Z50<52),1, IF(AND(Ranks!Z50>51,Ranks!Z50<60),2,3))))))	0
Ranked Comp	Re-ranks scores from "Comps" (with ties)	=RANK(Comp!Z50,Comp!$A50:$BL50)	12
Spearman	Correlation using "Ranked Comp"	=CORREL('Ranked Comp'!A26:BL26,'Ranked Comp'! A50:BL50)	0.004308382
Pearson	Correlation using "Comp"	=CORREL(Comp!A26:BL26,Comp!A50:BL50)	-0.06741573
Sp2	"Spearman" Matrix without the diagonal	=IF(OR(Spearman!Z50=1, Spearman!Z50=""),"",Spearman!Z50)	-0.058182528
P2	"Pearson Matrix" without the diagonal	=IF(OR(Pearson!Z50=1,Pearson!Z50 = ""),"",Spearman!Z50)	-0.06741573
Difference	Value in "Sp2"-Value "P2"	=IF(OR('P2'!Z50="",Sp2!Z50=""),"",('P2'!Z50-Sp2!Z50))	-0.009233203
AbsDiff	Absolute value of "Sp2" - "P2"	=IF(OR('P2'!Z50="",Sp2!Z50=""),"",ABS('P2'!Z50-Sp2!Z50))	0.009233203

Figure 1: Summary of Worksheets

the Spearman's r_s ranked correlation matrix in a worksheet "Spearman."[3] The diagonal terms were eliminated (any row perfectly correlates with itself under both correlation processes, so these diagonal terms are always one) and the remaining data were recorded in worksheets "P2" and "SP2," respectively.

RESULTS

The pairwise differences between the correlations in "P2" and "SP2"were calculated in worksheets "differences" and "absdiff," respectively. Over 30,000 (250*249) correlations were done, and Figure 2 shows a histogram of the natural differences, demonstrating that r_s is an unbiased estimator of r (the differences have mean and median 0).

The absolute differences, shown in Table 1 and as a histogram in Figure 3, reveal that 95% of the time, the error of the estimate of using Pearson in place of Spearman is less than 0.0475 (50% of the time the error is less than 0.0161). This is quite small, given the range of correlation from –0.5 to +0.5.[4] Future research will try to replicate this on more highly correlated data, using a genetic algorithm to generate vectors.

A metacorrelation, the correlation of the two correlation matrices (each a 250×250 lower diagonal matrix), gives a correlation above 0.98. This concurs with the claim that, in fact, the two methods of correlation give quite similar results.

CONCLUSIONS

As desired, the Pearson r can be used as a substitute for the Spearman's r_s when working with Q-data. This may allow the teaching of Q-methodology in methods courses and in other situations where students might know the simpler correlation, but not the nonparametric Spearman equivalent.

Q-methodology is an ideal candidate for simulation due to the complex probability patterns involved. Methods similar to these might be used in other complex situations arising in

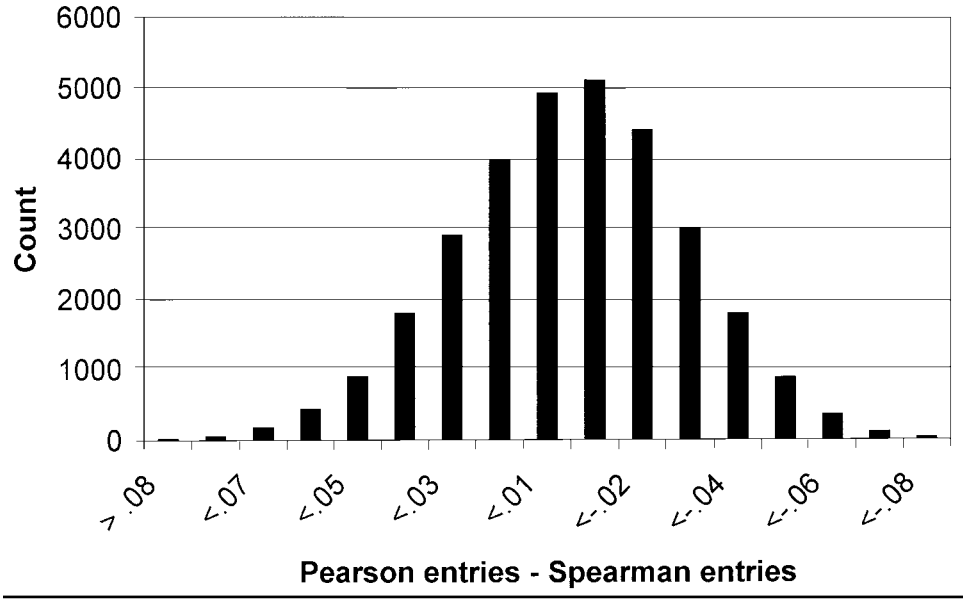

Figure 2: Raw Differences Between the Two Types of Correlations

TABLE 1
Summary Shape Statistics for Simulated Data From the Worksheets

	Diff	*Abs*	*P2*	*Sp2*
Maximum	0.1021	0.1021	0.4663	0.4827
99th percentile	0.0582	0.0621	0.2921	0.2938
95th percentile	0.0405	0.0475	0.2079	0.2082
90th percentile	0.0314	0.0396	0.1629	0.1610
75th percentile	0.0160	0.0276	0.0843	0.0847
50th percentile	–0.0004	0.0161	0.0000	–0.0014
25th percentile	–0.0162	0.0076	–0.0843	–0.0863
10th percentile	–0.0304	0.0030	–0.1629	–0.1624
5th percentile	–0.0388	0.0015	–0.2079	–0.2054
1st percentile	–0.0538	0.0003	–0.2921	–0.2900
Minimum	–0.0921	0.0000	–0.5000	–0.5207

NOTE: Correlation = 0.9817.

the social sciences, and Excel is an excellent piece of software for teaching demonstrations of simulation. Excel, however, takes between 3 and 10 times the amount of resources (both disk space and time) compared with simulations created in simulation software such as Arena or @Risk or by code written in C++ or other multipurpose languages.

Excel worksheets allow the convenient layering described below and allow the entire spreadsheet to be in one file. The total file described is about 27 MB, and the original computation took about 1 hour on a Pentium 166 machine running Excel 97 under Windows 95. A more modern P-6 450 running similar software allows this simulation to be run in less than

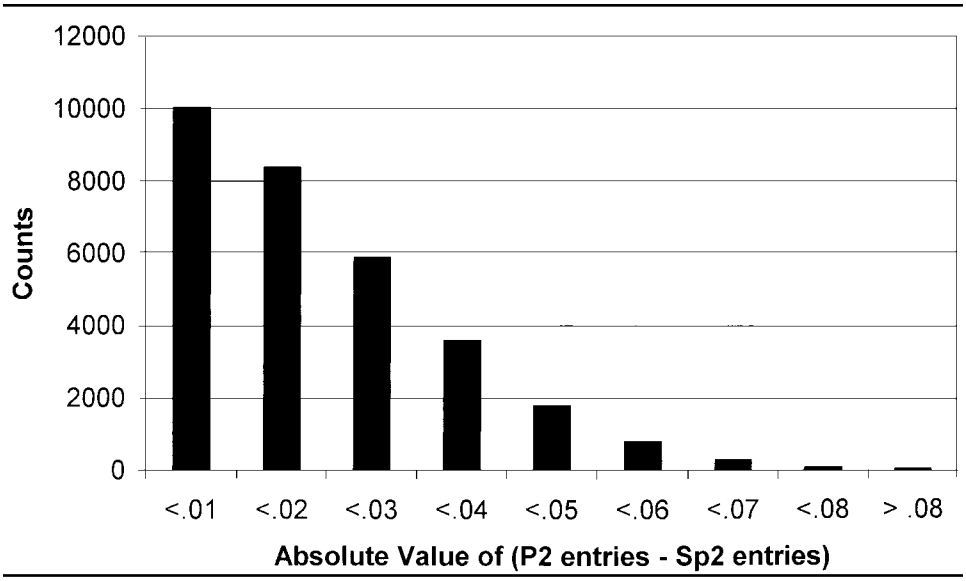

Figure 3: The Absolute Values of the Raw Differences

10 minutes. If, for a classroom demonstration, the simulation were done ahead of time, Excel can open the file in less than a minute.

Software Cited

Microsoft Excel 97 SR-1, part of Microsoft Office 97

The commands above should get you started on a simulation of this type. A copy of a truncated file (containing the same information, but only 50 replicates) is less than three MB and can be obtained from the Web site: `http://www.iems.nwu.edu/~bea/qsort.html`

NOTES

1. A quick power test shows that 50 replications should be sufficient for significance, so we are doing five times that minimum required number of replications to ensure a significant result.

2. However, the first thing that should be done in Excel is to turn off the Calculate automatically Feature, which can be done under Tools: Options. When this feature is enabled, Excel recalculates all cells, including those that generate random numbers with each carriage return. Given the size of the file, calculating parts of the spreadsheet separately is a must to avoid computer crashes and multiple-day procedures.

3. Correlation matrices can be easily generated using the Excel "data analysis add-in," resulting in a 250×250 lower triangular matrix, each point of which is the correlation of two of the simulated vectors.

4. A higher range of correlations would be desirable, but simulated independent data is very unlikely to have a high level of intercorrelation.

REFERENCES

Block, J. (1979). *The Q-sort method in personality assessment and psychiatric research*. Palo Alto, CA: Consulting Psychologists Press.

Brown, S. (1971). The forced-free distribution in Q-technique, *Journal of Educational Measurement*, *8*, 283-7.

Brown, S. (1980). *Political subjectivity.* New Haven, CT: Yale University Press.
Gibbons, J. (1985). *Nonparametric methods for quantitative analysis* (2nd ed.). Columbus, OH: American Sciences Press.
Garson, G. D. (1994). Social science computer simulation: Its history, design, and future. *Social Science Computer Review*, *12*(1), 55-82.
Kelton, W. D., Sadowski, R. P., & Sadowski, D. A. (1998). *Simulation with Arena*. Boston: McGraw-Hill.

K. Scott Alberts, Ph.D., is currently assistant professor of mathematics at Truman State University. This work was done as part of a dissertation completed in the Department of Industrial Engineering and Management Sciences of the McCormick School of Engineering and Applied Sciences at Northwestern University. His research tries to diagnose and describe decision-making groups in the real world and to create descriptive models of these groups. He can be reached at Truman State University, Kirksville, MO 63501; e-mail: alberts@truman.edu

Bruce Ankenmann, Ph.D., is an assistant professor of industrial engineering and applied sciences at Northwestern University. His research focuses on statistical design of industrial experiments, engineering design and development, quality improvement and quality control, and other applied statistical methods. He can be reached at the Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208; e-mail: ankenman@iems.nwu.edu