Judgment and Decision Making, Vol. 7, No. 1, January 2012, pp. 97–109

A statistical test of independence in choice data with small samples

Michael H. Birnbaum*

Abstract

This paper develops tests of independence and stationarity in choice data collected with small samples. The method builds on the approach of Smith and Batchelder (2008). The technique is intended to distinguish cases where a person is systematically changing "true" preferences (from one group of trials to another) from cases in which a person is following a random preference mixture model with independently and identically distributed sampling in each trial. Preference reversals are counted between all pairs of repetitions. The variance of these preference reversals between all pairs of repetitions is then calculated. The distribution of this statistic is simulated by a Monte Carlo procedure in which the data are randomly permuted and the statistic is recalculated in each simulated sample. A second test computes the correlation between the mean number of preference reversals and the difference between replicates, which is also simulated by Monte Carlo. Data of Regenwetter, Dana, and Davis-Stober (2011) are reanalyzed by this method. Eight of 18 subjects showed significant deviations from the independence assumptions by one or both of these tests, which is significantly more than expected by chance.

Keywords: preference reversals, transitivity, permutation test, independence, stationarity.

1 Introduction

Regenwetter, Dana, and Davis-Stober (2010, 2011) proposed a solution to the problem of testing whether choice data satisfy or violate transitivity of preference. Their probabilistic choice model assumes that on a given trial, a response can be represented as if it were a random sample from a mixture of transitive preferences. The model was used to analyze a replication of Tversky's (1969) study that had reported systematic violations of transitivity of preference (Regenwetter, et al., 2010, 2011). Reanalysis via this iid mixture model of new data concluded that transitivity can be retained.

Birnbaum (2011) agreed with much of their paper, including their conclusions that evidence against transitivity is weak, but criticized the method in part because it assumes that responses by the same person to repeated choices are independent and identically distributed (iid). If this assumption is violated, the method of Regenwetter et al. (2011) might lead to wrong conclusions regarding the tests of structural properties. Further, the violations of these assumptions can be analyzed by a more detailed analysis of individual responses to choice problems rather than by focusing on (averaged) binary choice proportions.

In the true and error model, a rival probabilistic representation that can also be used to test structural properties such as transitivity in mixture models, iid will be violated when a person has a mixture of true preferences and changes true preferences during the course of the study. Birnbaum (2011) showed how the Regenwetter et al. method might lead to wrong conclusions when iid is violated in hypothetical data, and described methods for testing between these two rival stochastic models of choice. These methods allow tests of assumptions of the Regenwetter, et al. (2011) approach against violations that would occur if a person were to change preferences during a study. He described conventional statistical tests that require conventional sized samples. Hypothetical examples illustrated cases in which the method of Regenwetter, et al. (2011) might lead to the conclusion that transitivity was satisfied, even when a more detailed analysis showed that the data contained systematic violations of both iid and transitivity.

The methods described by Birnbaum (2011) to test independence would require large numbers of trials, however, and might be difficult or impractical to implement. The experiment of Tversky (1969), which Regenwetter et al. (2011) replicated, does not have sufficient data to allow the full analyses proposed by Birnbaum (2011). Regenwetter, Dana, Davis-Stober, and Guo (2011) argued

I thank William Batchelder for suggesting the permutation method and helpful discussions. Michel Regenwetter kindly provided data for reanalysis and useful discussions of these issues. Regenwetter asked Chris Zwilling, Yun-shi Cha, and Michelle Choi to check my reanalysis of the Regenwetter, et al. (2011) data; thanks to their work it was determined that I had initially misunderstood the trial orders used for some of the participants. After this was corrected, our findings agreed. Thanks are also due Kathleen Preston for helpful suggestions. This work was supported in part by a grant from the National Science Foundation, SES DRMS-0721126.

*Department of Psychology, California State University, Fullerton, H-830M, Box 6846,Fullerton, CA 92834–6846, USA. Email: mbirnbaum@fullerton.edu.

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

that it would be difficult to collect enough data to provide a complete test of all iid assumptions, as proposed by Birnbaum (2011).

Nevertheless, this note shows that by building on the approach of Smith and Batchelder (2008), it is possible to test iid assumptions even in small studies such as that of Regenwetter, et al. (2011).

2 Testing IID assumptions in small studies of choice

Suppose a person is presented with *m* choice problems, and each problem is presented *n* times. For example, each of these *m* choice problems might be intermixed with filler trials and presented in a restricted random sequence such that each choice problem from the experimental design is separated by several fillers. These choices might also be blocked such that all *m* choices are presented in each of *n* trial blocks, but blocking is not necessary to this test. Let *x(i, j)* represent the response to choice *j* on the *i*th presentation of that choice.

Define matrix ***z*** with entries as follows:

$$z(i,k) = \sum_{j}[x(i,j) - x(k,j)]^2 \qquad (1)$$

where ***z*** is an *n* by *n* matrix showing the squared distance between each pair of rows of the original data matrix, and the summation is from *j* = 1 to *m*. If responses are coded with successive integers, 0 and 1, for example, representing the choice of first or second stimulus, then *z(i, k)* is a simply a count of the number of preference reversals between repetitions *i* and *k,* that is, between two rows of ***x***. In this case, the entries of *z(i, k)* would have a minimum of 0, when a person made exactly the same responses on all *m* choices in two repetitions, and a maximum of *m*, when a person made exactly opposite choices on all *m* trials.

Smith and Batchelder (2008) show that random permutations of the original data matrix allow one to simulate the distribution of data that might have been observed under the null hypothesis. According to iid, it should not matter how the data of ***x*** are permuted within each column. That is, it should not matter if we switch two values in the same column of ***x***; they are two responses to the same choice on different repetitions by the same person. For example, it should not matter whether we assign one response to the first repetition and the other to the last, or vice versa.

Assuming iid, the off-diagonal entries of matrix ***z*** should be homogeneous, apart from random variation. However, if a person has systematically changed "true" preferences during the study, then there can be some entries of ***z*** that are small and others that are much larger. That is, there can be a relatively larger variance of the entries in ***z*** when iid is violated.

Therefore, one can compute the variance of the entries in ***z*** of the original data matrix, ***x***, and then compare this observed variance with the distribution of simulated variances generated from random permutations of the data matrix. If iid holds, then random permutations of the columns will lead to variances that are comparable to that of the original data, but if the data violate iid, then the original data might have a larger variance than those of most random permutations. The proportion of random permutations leading to a simulated variance that is greater than or equal to that observed in the original data, taken from a large number of Monte Carlo simulations, is the p_v value for this test of iid. When $p_v < \alpha$, the deviations from iid are said to be "significant" at the α level, and the null hypothesis of iid can be rejected. When $p_v \geq \alpha$, one should retain both the null and alternative hypotheses.

A second statistic that can be calculated from the matrix of ***z*** is the correlation between the mean number of preference reversals and the absolute difference in replications. If a person changes gradually but systematically from one set of "true" preferences to another, behavior will be more similar between replicates that are closer together in time than between those that are farther apart (Birnbaum, 2011). This statistic can also be simulated via Monte Carlo methods, and the proportion of cases for which the absolute value of the simulated correlation is greater than or equal to the absolute value of the original correlation is the estimate of the p_r value for the correlation coefficient. (The use of absolute values makes this a two-tailed test).

A computer program in R (R Development Core Team, 2011) that implements these calculations via computer generated pseudo-random permutations is presented in Listing 1.

Appendix A defines independence and identical distribution (stationarity) for those who need a refresher, and it presents analyses of hypothetical data showing how the simulated variance method yields virtually the same conclusions as the two-tailed, Fisher exact test of independence in a variety of 2-variable cases with *n* = 20. It is noted that standard tests of "independence" usually assume stationarity, and it is shown that a violation of stationarity can appear as a violation of "independence" in these standard tests. For that reason, the variance test of this paper is best described as a joint test of iid.

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

Table 1: Analyses of actual data of Regenwetter (2011) replicating Tversky (1969). The mean of z is the mean number of disagreements (preference reversals) out of 10 choices between two rows, averaged over all possible pairs of rows. The variance of z is the variance of these entries for the original data. The p_v -values are the proportion of simulated permutations of the data for which the calculated variance of z in the permuted sample is greater than or equal that of the original data (each is based on 100,000 computer generated, pseudo-random permutations of the data). Correlations between mean rate of preference reversals and difference between repetitions are shown in the column labeled, r. Corresponding p_r -values are shown in the last column.

Subject	Mean of z	Variance of z	p_v	r	p_r
1	3.59	2.89	0.272	0.61	0.086
2	2.72	4.29	**0.000**	0.91	**0.000**
3	0.19	0.16	1.000	0.61	0.510
4	2.72	2.61	0.077	0.01	0.985
5	0.65	1.06	**0.011**	–0.82	0.104
6	2.67	2.18	0.114	0.67	**0.048**
7	1.70	1.75	0.238	0.88	**0.005**
8	0.35	0.26	1.000	0.12	0.905
9	4.33	3.12	0.782	–0.53	0.107
10	1.23	1.51	**0.047**	0.45	0.539
11	0.53	0.48	0.464	0.11	0.883
12	3.79	2.68	0.788	0.44	0.246
13	4.27	3.47	0.181	0.71	**0.011**
14	0.19	0.16	1.000	0.71	0.390
15	3.48	4.25	**0.000**	0.82	**0.003**
16	1.29	0.67	0.998	0.77	**0.024**
17	4.32	3.11	0.752	0.34	0.336
18	4.09	2.76	0.952	–0.38	0.316

3 Reanalysis of Regenwetter, et al. (2011)

Applying this approach to the data of Regenwetter, et al. (2011), the estimated p_v and p_r values based on 100,000 simulations are shown in Table 1. Four of the p_v values are "significant" at the $\alpha = 0.05$ level. Fifteen of the 18 correlation coefficients are positive, and 6 of the correlations are significantly different from 0 by this two-tailed test ($\alpha = 0.05$). Eight of the 18 subjects have significant deviations by one or both of these tests.

Since 18 tests were performed for each of two properties, we expect 5% to be significant with $\alpha = .05$; i.e., we expect about 1 case to be significant for each property. Can these data be represented as a random sample from a population of people who satisfy the iid assumptions? The binomial probability to find four or more people out of 18 with p_v significant at the .05 level by chance is 0.01. The binomial probability to find 6 or more cases with p_r significant at this level is 0.005. The binomial probability to observe 15 or more correlations positive out of 18, assuming half should be positive by chance, is .003. Therefore, using either criterion, variance or correlation, we can reject the hypothesis that iid is satisfied. Considering how small the sample size is for each person, compared to what would be ideal for a full test of iid such as proposed by Birnbaum (2011), it is surprising that these tests show so many significant effects.

Appendix B notes that the Regenwetter, et al. (2011) experiment has low power for testing iid, so these significant violations are probably an indication that the violations are substantial. Also discussed in Appendix B is the connection between finding significant violations of a property such as iid for some individuals and what inferences might be drawn for general conclusions concerning iid, based on summaries of individual significance tests. A philosophical dispute is reviewed there between those who "accept" the null hypothesis and those who "retain" both null and alternatives when significance is not achieved.

Table 2 shows data for Subject #2, whose data violated iid on both tests. These data show relatively more responses of "0" at the beginning of the study than at the end. Therefore, the first three or four repetitions resemble each other more than they do the next dozen repetitions, which in turn resemble each other more than they do the first repetitions. Random permutations of the data distribute the "0" values more evenly among rows, which resulted none (zero) of 100,000 random permutations of the data having larger variance than that in the original data. Figure 1 shows the estimated distribution of the variance statistic under the null hypothesis for this person.

4 Discussion

These tests show that the data of Regenwetter, et al. (2011) do not satisfy the iid assumptions required by their method of analysis. The assumption of iid in their paper is crucial for two reasons: first, iid is required for the statistical tests of transitivity; second, iid justifies analyzing the data at the level of choice proportions instead of at the level of individual responses. When iid is satisfied, the binary choice proportions contain all of the "real" information in the data. However, when iid is violated, it could be misleading to aggregate data across repetitions to com-

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

Table 2: Raw data for Subject #2 of Regenwetter, et al. (2011), who showed violations of iid on both indices. Each row shows results for one repetition of the study.

V:	12	13	14	15	23	24	25	34	35	45
1	1	0	0	0	1	1	1	0	1	1
2	0	0	0	0	1	0	1	0	0	1
3	0	1	0	0	0	0	1	1	0	1
4	0	0	1	1	1	1	1	0	1	0
5	1	1	1	0	1	0	1	1	1	1
6	1	1	1	1	1	1	1	1	0	1
7	1	1	1	1	1	1	1	1	1	1
8	0	1	1	1	1	1	1	1	1	1
9	0	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	0	1	1
11	1	1	0	1	1	1	1	0	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1
14	0	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	1	1
18	0	1	1	1	1	1	1	1	1	0
19	1	1	1	1	1	1	1	1	1	0
20	0	1	1	1	1	0	1	0	1	1

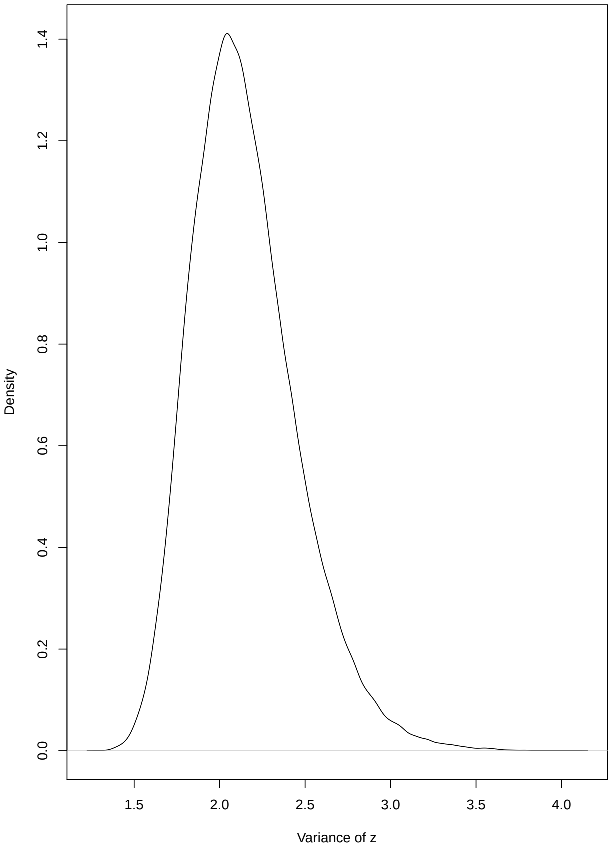
pute marginal choice probabilities (Smith & Batchelder, 2008; Birnbaum, 2011).

Appendix C describes three hypothetical cases that would be treated as identical in the Regenwetter, et al. (2011) approach but which are very different. These examples illustrate how cases with exactly the same choice proportions (column marginal means) could arise from very different processes, and how these different processes can be detected by examination of the individual response patterns. Appendix D presents further simulations of hypothetical data to compare the simulated variance method in three-variable cases with the results of standard Chi-Square and G^2 tests of independence.

These tests of iid are not guaranteed to find all cases where a person might change true preferences. For example, if a person had exactly two true preference patterns in the mixture that differed in only one choice, it would not produce violations of iid.

Each of these methods (variance or correlation) simplifies the z matrix into a single statistic that can be used to test a particular idea of non-independence. The variance method would detect cases in which a person randomly sampled from a mixture of true preference patterns in each block of trials, as in one type of true and error model.

Figure 1: Estimated distribution of variance of the entries in z matrices generated from random permutations of the original data matrix for case #2 (Table 2), based on 100,000 pseudo-random permutations of the data in Table 2. None of the simulations exceeded the value observed in the original data, 4.29. Case #2 was selected as showing the most systematic evidence against the iid assumptions.



The correlation method detects violations of iid in the z matrix that follow a sequential pattern; for example, a positive correlation would be expected if a person sticks with one true preference pattern until something causes a shift to another true pattern, which then persists for a number of trials. Violations of either type would be consistent with the hypothesis that there are systematic changes in "true" preference during the course of the study (Birnbaum, 2011; Birnbaum & Schmidt, 2008).

Furthermore, there may be more information in the data (and the z matrix) beyond what one or two indices could represent; for example, one might explore the z matrix via nonmetric multidimensional scaling (Carroll & Arabie, 1998) in order to gain additional insight into

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

the pattern of violation of iid. Note that each entry of z can be regarded as a squared Euclidean distance between two repetitions.

In summary, it is possible to test assumptions of iid in studies with small samples, and when these tests are applied, it appears that these assumptions are not consistent with data of Regenwetter et al. (2011). A larger study such as described in Birnbaum (2011) would have greater power and would certainly be a better way to identify and study violations of iid, but this note shows how these assumptions can also be tested with small samples. The fact that a number of cases are significant based on only 20 repetitions suggests that these violations are likely substantial in magnitude.

References

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: Comments on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*. 675–683.

Birnbaum, M. H., & Bahra, J. P. (2007). Transitivity of preference in individuals. *Society for Mathematical Psychology Meetings*, Costa Mesa, CA. July 28, 2007.

Birnbaum, M. H., & Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty, 37,* 77–91.

Carroll, J. D., & Arabie, P. (1998). Multidimensional scaling. In M. H. Birnbaum (Ed.), *Measurement, Judgment, and Decision Making* (pp. 179–250). San Diego: Academic Press.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Regenwetter, M., Dana, J. & Davis-Stober, C. P. (2010). Testing Transitivity of preferences on two-alternative forced choice data. *Frontiers in Psychology*, *1*, Article 148, 1–14.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.

Regenwetter, M., Dana, J., Davis-Stober, C. P., and Guo, Y. (2011). Parsimonious testing of transitive or intransitive preferences: Reply to Birnbaum (2011). *Psychological Review, 118*, 684–688.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. doi: 10.3758/PBR.15.4.713

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31–48.

Appendix A: Independence and stationarity in a repeated trials task

Consider an experiment that yields two dependent variables, X and Y. The experiment is repeated n times, and the data are labeled, X_i and Y_i for the observed responses on the ith repetition. For simplicity, assume that the values of the variables are binary, either 0 or 1. A hypothetical example of such a matrix is shown in Table A.1.

Table A.1. A hypothetical set of data with two choices and 20 repetitions.

Rep	X	Y
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	1	0
10	1	0
11	0	1
12	0	1
13	1	1
14	1	1
15	1	1
16	1	1
17	1	1
18	1	1
19	1	1
20	1	1

Let p_i and q_i represent probabilities that $X_i = 1$ and $Y_i = 1$, respectively. *Independence* is the assumption that the probability of the conjunction of X and Y is the product of the individual probabilities; i.e., $p(X_i = 1 \text{ and } Y_i = 1) = p_i\, q_i$. *Stationarity* is the assumption that $p_i = p$ and $q_i = q$, for all i. The term iid (independent and identically distributed) is the assumption that both of these properties are satisfied; i.e., i.e., $p(X_i = 1 \text{ and } Y_i = 1) = pq$ for all i.

The conditional probability of X given Y is the joint probability of X and Y divided by the probability of Y; i.e., $p(X_i = 1 \mid Y_i = 1) = p(X_i = 1 \text{ and } Y_i = 1)/p(Y_i = 1)$. If independence holds, $p(X_i = 1 \text{ and } Y_i = 1) = p(X_i = 1)p(Y_i = 1)$; that means that $p(X_i = 1 \mid Y_i = 1) = p(X_i = 1) = p_i$. Therefore, independence of X and Y can also be ex-

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

pressed in terms of conditional probabilities, as follows: $p(X_i = 1 \mid Y_i = 1) = p(X_i = 1 \mid Y_i = 0) = p(X_i = 1) = p_i$; similarly, independence also means that $p(Y_i = 1 \mid X_i = 1) = p(Y_i = 1 \mid X_i = 0) = p(Y_i = 1) = q_i$.

If the rows of Table A.1 represented different subjects who were tested separately and in random order, we could assume that rows are a "random effect" and we would test independence by studying the crosstabulation of X and Y, as shown in Table A.2. Counting the number of rows with $(X, Y) = (0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ we find that there are 8, 2, 2, and 8 cases, respectively.

Table A.2. Crosstabulation of the hypothetical data from Table A.1

	Y = 0	Y = 1	Row Sum/Total
X = 0	8	2	.5
X = 1	2	8	.5
Row Sum/Total	.5	.5	20

A Chi-Square test is often used to test if data in a crosstabulation satisfy independence. This test estimates the probabilities of X and Y from the marginal proportions (column marginal means in Table A.1 or the column and row marginal sums divided by n in Table A.2). The "expected" value (predicted value) in the crosstabulation table is then constructed using products of these estimates; for example, the predicted entry corresponding to the (1, 1) cell of Table A.2: $E(X_i = 1 \text{ and } Y_i = 1) = (.5)(.5)20 = 5$. That is, we multiply column marginal proportions of Table A.1 with each other and multiply this product by the total number of observations, in order to construct an "expected" value, based on independence. If X and Y were indeed independent, we would expect to observe frequencies of 5, 5, 5, and 5 in the crosstabulation. Thus, the frequencies in Table A.2 (8, 2, 2, and 8) indicate that the hypothetical data in Table A.1 are not perfectly independent. If these were sampled data, we might ask if the violations are "significant," which means we ask, "are such deviations unlikely to have arisen by random sampling from a population in which independence holds?"

The Chi-Square test compares expected (based on independence) frequencies with obtained frequencies. However, the Chi-Square test is a poor approximation when the sample size, n, is small, or when expected frequencies are small. For this reason, we need a more accurate way to compute the probability of observing a sample of data given the null hypothesis.

The Fisher test is called an "exact" test because it directly calculates the probabilities of any crosstabulation frequencies (as in Table A.2) given the assumption of independence, and it is therefore more accurate than the Chi-Square test as a test of independence, especially with small n. Let a, b, c, and d represent the frequencies of (0, 0), (0, 1), (1, 0), and (1, 1), respectively. Then the exact probability, p_F, assuming independence of X and Y, of any such array with the same marginal totals is given by the hypergeometric distribution:[1]

$$p_F = \frac{\left[\frac{(a+c)!}{a!c!}\right]\left[\frac{(b+d)!}{b!d!}\right]}{\left[\frac{n!}{(a+b)!(c+d)!}\right]}$$

As noted above, independence can be expressed in terms of conditional probabilities. We can estimate conditional probabilities in Table A.2 as follows: $p(Y_i = 1 \mid X_i = 0)$ and $p(Y_i = 1 \mid X_i = 1)$ from conditional proportions $b/(a + b)$ and $d/(c + d)$, respectively. If independence holds, these two conditional proportions should be equal to each other; when they are unequal, we can measure the degree of disproportion (violation of independence) from the following:

$$D = \left| \frac{b}{(a+b)} - \frac{d}{(c+d)} \right|$$

In Table A.2, $D = |.2 - .8| = 0.6$. The Fisher two-tailed test computes the sum of probabilities of all arrays with the same n and marginal sums, such that the degree of disproportion as measured by D is greater than or equal to that of the observed array.

In this case, the Fisher test yields the following two-tailed "p-value" for Table A.1: 0.023. From the Fisher test, if the hypothetical values in Table A.1 were considered a random sample of data, the hypothesis of independence of X and Y in Table A.1 would be rejected at the $\alpha = 0.05$ level of significance.

It is important to keep in mind that in order to interpret either the Fisher or Chi-Square test, that we had (implicitly) assumed another type of independence; namely, we assumed that each row in Table A.1 was obtained from a different person, and that the people were tested separately. We assumed that these people did not communicate with each other by talking, cheating, or via ESP. That is, we assumed that the rows in Table A.1 are a random factor; i.e., that "rows do not matter."

Statistics teachers often describe physical situations in which scientists believe, based on theoretical arguments supported by empirical evidence, in independence of replicated measures. For example, X in table 1 might represent whether or not a tossed "fair" coin comes up heads ($X = 1$) or tails ($X = 0$), and Y might represent whether a card drawn from a standard deck is red ($Y =$

[1] In this expression for the hypogeometric distribution, the denominator represents the number of ways of dividing n to achieve the row sums, $a + b$, and $c + d$; the numerator is the product of the number of ways of dividing the first and second column sums to arrive at entries of a and b with column sums of $a + c$ and $b + d$. The product in the numerator imposes the assumption of independence.

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

1) or black ($Y = 0$). We assume that coins and cards do not communicate with each other, and in recent years, we assume that no gods intervene in both physical situations to advise us of future events. Ample evidence has been obtained to check these hypotheses, and physical objects such as coins and cards do seem to conform to independence; at least they do when magicians do not handle the props.

The assumption that rows in Table A.1 do not matter implies "stationarity" with respect to Table A.1, therefore: $p_i = p$ and $q_i = q$, for all i. However, notice that for both X and Y, the likelihood of 1 in Table A.1 increases as the row number increases. We can test stationarity by asking if the X_i and Y_i are independent of i. In Table A.1, stationarity is violated (as well as independence).

In the physical example, stationarity is the assumption that the probability of heads does not change over time. Similarly, the probability of drawing a red card is assumed to be stationary, as long as we replace each card and reshuffle the deck. But if we don't return red cards to the deck, then stationarity would be violated.

Now consider the situation in which the "data" in Table A.1 arose instead from repeated trials performed by the same person over time. Suppose X and Y are scores on two learning tasks, for example, and rows represent repeated trials with feedback to the same person. That means that rows do matter. The first trial is not the same as the last trial of a learning experiment. If $X = 0$ is an error and $X = 1$ is correct, hypothetical data as in Table A.1 would indicate that there has been learning during the study, because there are more 1s at the end compared to the beginning.

When data arise from a single person, therefore, our standard tests of independence are confounded with the assumption of stationarity. We cannot justify the assumptions upon which the standard Chi-Square or Fisher test of independence are built. We can test the two assumptions of iid together by means of these tests, but it is less clear which of these was the culprit, when the (combined) test is rejected.

For the Fisher test of independence, we assumed stationarity of the probabilities when we created Table A.2, which collapsed across rows. We assumed stationarity when we simplified the statement of independence from testing $p(X_i = 1 \text{ and } Y_i = 1) = p_i\, q_i$ to $p(X_i = 1 \text{ and } Y_i = 1) = p\, q$. That means that the standard Fisher or Chi-Square test, when performed on repeated measures from the same person is a confounded test of independence and stationarity, and if the Fisher or Chi-Square test is significant, we cannot clearly blame one or the other without further information or assumption. So even though these tests are typically called "tests of independence", they also assume stationarity, so they would be better described as joint tests of iid.

Now, where does the Smith and Batchelder (2008) method using a computer to create pseudo-random permutations fit into this picture? Their method randomly permutes the entries in the columns in Table A.1. So the number of 0 and 1 in each column stay the same and therefore the column marginal means (column proportions) stay the same, but the pairings of X and Y will change because the columns are randomly and independently permuted. That means that the joint frequencies Table A.2 can (and usually will) change with each random permutation. It also means that the distribution of X and Y with respect to rows will change. Columnwise random permutations of the data therefore would remove two effects in Table A.1: First, it will break up the pattern of nonindependence between X and Y. Second, it will also break up the correlations between row number and X and between row number and Y.

Although Smith and Batchelder (2008) refer to testing stationarity while assuming independence, their procedure also tests independence, assuming stationarity. To illustrate this point, Monte Carlo simulations of 19 hypothetical cases in Table A.3 were conducted, using the R-program to implement pseudo-random permutations and the variance method to estimate p_v. Each hypothetical case was constructed (as in Table A.1) as a 20 by 2 array in which X and Y take on values of 0 and 1. The four entries in Table A.3, a, b, c, and d, represent the frequencies in the four cells (as in Table A.2): (0, 0), (0, 1), (1, 0), and (1, 1). respectively. The same data were also analyzed by the two-tailed Fisher exact test of independence. The last row of Table A.3 represents the example analyzed in Tables A.1 and A.2.

These simulations allow us to compare the variance method with the Fisher exact test. As shown in Table A.3, the simulated p_v values, based on 10,000 pseudo-random permutations, are very close to the p-values calculated by the two-tailed, Fisher exact test. The two procedures yield virtually the same conclusions.

The use of random permutations within columns with the variance method can therefore be considered a test of independence, since it yields the same conclusions as the Fisher "test of independence". But keep in mind that the Fisher test also implicitly assumes stationarity. When the results of the Fisher test are significant, it means that iid can be rejected, but we cannot say if one, the other, or both assumptions are violated. When we can safely assume "rows don't matter," as we would when each row of Table A.1 represented data from a different person tested separately, the Fisher test is indeed a "pure" test of independence because we assumed that people do not have ESP (based on considerable evidence); but when this test is applied to repeated data from a single person, it is a joint test of iid.

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

Table A.3. Results of Monte Carlo simulation of hypothetical data with 20 reps. Cells *a*, *b*, *c*, and *d* represent the frequencies of (0, 0), (0, 1), (1, 0), and (1, 1), respectively. The column labeled "Fisher" shows the calculated probability for the Fisher exact test of independence; The last column shows the simulated p_v, based on 10,000 random permutations.

a	*b*	*c*	*d*	Fisher	Simulated p_v
10	0	0	10	0.0000	0.0000
7	3	0	10	0.0031	0.0027
4	6	0	10	0.0867	0.0875
1	9	0	10	1.0000	1.0000
10	0	2	8	0.0007	0.0007
7	3	2	8	0.0698	0.0685
4	6	2	8	0.6285	0.6326
1	9	2	8	1.0000	1.0000
10	0	4	6	0.0108	0.0112
7	3	4	6	0.3698	0.3678
4	6	4	6	1.0000	1.0000
1	9	4	6	0.3034	0.3115
10	0	5	5	0.0325	0.0314
7	3	5	5	0.6499	0.6522
4	6	5	5	1.0000	1.0000
1	9	5	5	0.1409	0.1430
6	4	4	6	0.6563	0.6557
5	5	5	5	1.0000	1.0000
8	2	2	8	0.0230	0.0227

To understand how a violation of stationarity might appear as a violation of independence, consider the following case. A person samples randomly (with replacement) from two urns, *X* and *Y*, that each contain 20% Red and 80% white marbles. Code Red as "1" and White as "0." The two urns are independent and stationary (because we replace the sampled marbles and remix the urns after each draw), so in 100 trials, we expect the frequencies of (0, 0), (0, 1), (1, 0), and (1, 1) to be 64, 16, 16, and 4, respectively. Now consider a new pair of urns, each of which now contains 80% Red and 20% White, respectively. In 100 trials with these two urns, we expect 4, 16, 16, and 64, respectively. Each pair of urns satisfies independence. Now suppose we switched from the first pair of urns to the second pair of urns after 50 trials (which violates stationarity but not independence): we would expect to observe 34, 16, 16, and 34, when data are combined over the 100 total trials. These values violate "independence," which in this case implies 25, 25, 25, and 25. Even if one were to flip a coin on each trial to determine which pair of urns to use, we expect to observe a violation of independence despite the fact that the coin and urns are independent. Therefore, violations of stationarity could lead to an apparent violation of "independence," even though true (physical) independence was satisfied within each part of the experiment, and the culprit was actually a violation of stationarity.

The results of the variance method or of the Fisher test would not be affected by a random permutation method in which entire rows of data are permuted, because that would not change the connections between *X* and *Y*. From Expression 1, for the dissimilarity between two rows of data, note that the row number plays no role in the calculations except to index the cases. Further, the variance of dissimilarity is a computation that also is independent of the row numbering, except as an index. For example, randomly switching entire rows in Table A.1 would not alter the crosstabulation frequencies in Table A.2, nor would it affect the variance of dissimilarity of rows. However, the correlation method described here, which correlates the dissimilarity between rows with the gaps in trial order would indeed be affected by such a permutation method in which intact rows were exchanged. Therefore, the correlation test is clearly a test of stationarity, but it is best described as a test of stationarity that assumes independence. Furthermore, it tests only one type of violation that is related to the trial separation.

In conclusion, the random permutation method with the variance statistic is a joint test of iid and the permutation method with correlations between similarity and trial gap is one test of stationarity. However, it should be again noted that these two methods—variance and correlation—are not "pure" nor do they exhaust all of the information in the data that might reveal violations of iid. Aside from analyses described in Birnbaum (2011), these are the only two tests I have so far investigated with the Regenwetter, et al. (2011) data.

Appendix B: Comment on statistical power and multiple tests

Regenwetter, et al. (2010, 2011) studied designs in which there were ten choices comparing five stimuli. In the case of binary choices, there are two possible results for each choice. Therefore, there are $2^{10} = 1024$ possible response patterns, or cells, in the design. If we plan to look at a complete crosstabulation table to investigate independence, and if we plan to use a Chi-Square test, we should follow the rule of thumb that we should obtain an expected frequency of five responses per cell, so the experiment would require at least 5120 replicates, and even this large number might be considered just a bare minimum.

Regenwetter, et al. (2011) pointed out that such an experiment to properly test their iid assumptions in a ten

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

choice design would therefore be difficult, and in consequence, they concluded that one can assume iid for reasons of parsimony. They noted that their study had only 20 presentations per choice, so it would have very low power to test iid compared to a design with 5120 presentations per choice.

Despite the lack of power of the 20 repetition design, their data when analyzed by the simulation methods p_V and p_r reveal significant deviations from iid for 4 and 6 people, respectively. The fact that a study with so few repetitions detects significant deviations of iid suggests that the violations of iid are likely substantial.

How should one draw conclusions regarding theory from individual analyses? There are two philosophies of how to interpret individual subject analysis. To contrast them, I describe the views of Doctor 1 and Doctor 2, who wish to understand the safety of a new, hypothetical anesthetic called Propafool2.

The two doctors tested 18 subjects, using a triple blind, drug versus placebo study (the subjects, the doctor who administered the drugs, and the scientists who monitored the dependent variables were all blind with respect to the drug/placebo independent variable). The study measured both heart beat irregularities and breathing abnormalities within person comparing drug and placebo conditions. Both doctors agree that either irregularity causes an increase in the probability that a patient would die under the anesthetic, if not monitored. They performed a significance test on each subject on each dependent variable and found that 4 of 18 patients had significant heart abnormalities; 15 of 18 had reduced oxygen levels, of which 6 of 18 had significantly reduced oxygen under the drug compared to placebo; 8 of the 18 had "significant" abnormalities in at least one of the two measures.

The two doctors disagree on the implications of these results. Doctor 1 says that because we would expect approximately 1 person out of 20 to show significant irregularities in each test by chance (one of 20 at the 5% level would be 1), and because 4 and 6 of 18 are each significantly improbable, we would reject the null hypothesis (that drug is safe and the error rate accounts for the significant results) in favor of the alternative hypothesis (that the drug is dangerous to people in general); therefore, if it is to be used as an anesthetic, both heart rate and respiration should be monitored for every patient in every case.

Doctor 2 argues instead that those patients who did not show significant effects in the test have been "proven" to be immune to the drug. A doctor should be therefore be allowed to administer this drug as a sleeping agent at home to those 10 patients (for whom both effects were not statistically significant), without any requirement that heart rate or respiration be measured.

The first doctor, however, replies that nonsignificance does not prove there was no effect (retaining the null hypothesis is not the same as accepting it), so that the drug might be dangerous in a second administration to those people whose test scores were not significant. She argues that even if a person survived the drug on n previous tests, it is still possible that the next presentation might be lethal to that same person, and it would be a "recipe for disaster" to assume that the drug is safe, even for a person who showed no ill effects during previous administrations of the drug. If 18 men walk through a minefield and 8 are killed, the other 10 have not been shown to be invulnerable to mines.

The two doctors cannot agree on the applicability of individual subject analyses to predicting results for the same individuals, to predicting results with other individuals, nor do they agree whether nonsignificance proves a drug is safe. My own views are closer to those of Doctor 1, but I acknowledge that there is disagreement on these issues. I hold that failure to reject the null hypothesis does not prove the null hypothesis nor does it lead to refutation of the alternatives.

Appendix C: Three cases with choice proportions satisfying transitivity

Suppose we conducted a test of transitivity with 10 binary choices among five alternatives, and suppose we found that all binary choice proportions were 0.6. Such proportions are perfectly consistent with both weak stochastic transitivity and with the triangle inequality, and they are perfectly consistent with transitivity, according to the Regenwetter, et al. (2010, 2011) approach.

Table A.4 shows a hypothetical set of data that would yield such binary choice proportions. So do Tables 5 and 6. Note that in all three hypothetical arrays, the column marginal means are all 0.6. The person chose A over B 60% of the time, chose B over C 60% of the time, and chose A over C 60% of the time. According to the approach of Regenwetter, et al. (2010, 2011), which uses only the column marginal means, all three cases are perfectly consistent with transitivity, because all have the same column means.

However, Tables A.5 and A.6 are not consistent with the theory of Regenwetter, et al. (2010, 2011), because these data violate the assumptions of iid. In Table A.5, the data were constructed from the assumption that the subject started out with the transitive order, *ABCDE*, and then switched to the opposite transitive order, *EDCBA*. At least these data agree with the main conclusion in Regenwetter, et al., which is that behavior is consistent with a mixture of transitive orders. But their approach assumes that behavior can be modeled as iid samples from the mixture on each trial. That assumption is violated in Table A.5.

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

Table A.4. Hypothetical data consistent with transitivity and with the iid assumptions of Regenwetter, et al. (2010, 2011). These data are coded such that 1 = preference for the first stimulus in each choice and 0 = preference for the second stimulus in each choice.

Rep	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
1	1	0	1	1	1	1	0	1	1	1
2	0	0	0	0	1	0	1	0	0	1
3	0	1	1	1	1	1	1	0	1	1
4	1	1	0	0	1	1	1	0	1	0
5	0	0	1	1	0	1	0	1	0	0
6	1	1	1	1	0	1	0	0	1	0
7	1	1	1	0	0	1	1	1	1	1
8	1	1	0	1	1	0	1	1	0	1
9	1	0	0	0	1	0	0	1	1	0
10	0	1	1	1	0	0	1	1	0	1

Table A.5. Hypothetical data consistent with transitivity, but not with iid assumptions of Regenwetter, et al. (2010). In this case, the subject started with the transitive order, *ABCDE* for six blocks of trials, then switched to the opposite order for the last four blocks of trials.

Rep	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0

Table A.6. Hypothetical data that violate both transitivity and the iid assumptions of Regenwetter, et al. (2010, 2011). In this case, the person used an intransitive lexicographic semiorder for four blocks, followed by two transitive blocks of trials, followed by an opposite intransitive pattern for four blocks.

Rep	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
1	0	0	1	1	0	0	1	0	0	0
2	0	0	1	1	0	0	1	0	0	0
3	0	0	1	1	0	0	1	0	0	0
4	0	0	1	1	0	0	1	0	0	0
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	0	0	1	1	0	1	1	1
8	1	1	0	0	1	1	0	1	1	1
9	1	1	0	0	1	1	0	1	1	1
10	1	1	0	0	1	1	0	1	1	1

The hypothetical example in Table A.6 is more problematic for the Regenwetter, et al. approach. This case violates both transitivity and the assumptions of iid, but such data would be considered to be perfectly consistent with transitivity, according to the method of Regenwetter, et al. (2010, 2011).

Table A.6 was constructed from the assumption that the person used an intransitive lexicographic semiorder for 4 replicates, was then transitive for two replicates, and then used another intransitive lexicographic semiorder for the last four replicates. This hypothetical person was perfectly intransitive in 8 out of 10 blocks of trials. In one block of trials, this person chose A over B, B over C, C over D, D over E, and yet chose E over A. In other trials, the person had the opposite pattern of intransitive preferences. Thus, these data were constructed from assumptions that the Regenwetter, et al. model is false, and yet the procedures of Regenwetter, et al. would conclude that these data are perfectly consistent with their model.

In the approach of Regenwetter, et al. (2010), all three cases (Tables 4, 5, and 6) are treated as the same because the column marginal means are the same. If we assume that iid is satisfied, the column marginal choice proportions contain all of the information in the data, so we need not examine the actual data. But in this case, that would be wrong.

When we analyze these three cases using the variance method suggested here, however, we find that the simulated p_v -values for the test of iid are .0000 and .0000 for Tables 5 and 6, based on 10,000 pseudo-random permutations. The values of r and p_r are 0.943 and 0.0003 for Table A.5, and 0.943 and 0.0002 for Table A.6. Table A.4 is compatible with the assumptions of iid, according to the same methods (p_v = 0.8125, r = -0.105, p_r = 0.8159). Thus, this method correctly diagnoses these three hypothetical cases that are treated as if they are the same in the approach of Regenwetter, et al. (2010).

Do people actually show evidence of perfectly reversing their preferences between two blocks of trials? The answer is yes. Such cases of complete reversal have been observed in real data by Birnbaum and Bahra (2007). They separated blocks of trials by more than 50 intervening trials, and found that some people had 20 responses out of 20 choice problems exactly the opposite between two blocks of trials. Such extreme cases of perfect reversal mean that iid is not tenable because they are so improbable given the assumption of iid.

But how do we detect cases where a person switches between two or more different "true" patterns that are not perfect opposites? The methods in this paper are intended to do that.

The assumption of iid accomplishes two purposes: First, it justifies the decision not to examine the raw data as in Tables A.4, A.5, and A.6, but only to study

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

the marginal binary choice proportions (column marginal means of the tables). Second, it justifies the statistical tests in the approach of Regenwetter, et al. (2010, 2011). But if the iid assumptions are wrong, it means that not only are the statistical tests inappropriate, but that marginal choice proportions can be misleading as representative of the behavior to be explained.

Appendix D: Simulations in three variable case

Table A.7 shows results of 25 simulations in the three-variable case, in which the variance method is compared with two standard statistical tests of independence, χ^2 and G^2. The hypothetical data were constructed, as in Table A.1, except there were three variables, X, Y, and Z. Case 1 was constructed with 40 rows that perfectly satisfy independence in the crosstabulations.

In Table A.7, the response pattern, $(X, Y, Z) = (0, 0, 0)$ is denoted "000", (0, 0, 1) is denoted "001", and so on. Independence is the assumption that the probability of any combination of (X, Y, Z) is the product of the marginal probabilities. That is, the $p(000) = p(X = 0)p(Y = 0)p(Z = 0)$, $p(001) = p(X = 0)p(Y = 0)p(Z = 1), \ldots, p(111) = p(X = 1)p(Y = 1)p(Z = 1)$. Case 1 is perfectly consistent with independence because the "observed" frequencies of all response combinations are 5, which is exactly equal to the predicted values according to independence:

$$E(j,k,l) = n\,p(X=j)\,p(Y=k)\,p(Z=l)$$

where $E(j, k, l)$ are the expected frequencies in the crosstabulation (j = 0, 1; k = 0, 1; l = 0, 1), assuming iid; n is the total number of cases (number of rows in the hypothetical data matrix). In Case 1, there are exactly 40 rows; the column marginal means are estimates of $p(X = 1)$, $p(Y = 1)$, and $p(Z = 1)$, which are all 0.5, so each predicted frequency is 40(.5)(.5)(.5) = 5. The values in the first row of Table A.7 are the (hypothetical) "observed" values, which are counted from the crosstabulation of the hypothetical data.

The Chi-Squared test of independence in this case is defined as follows:

$$\chi^2 = \sum_j \sum_k \sum_l \frac{[F(j,k,l) - E(j,k,l)]^2}{E(j,k,l)}$$

where $F(j, k, l)$ are the observed frequencies, $E(j, k, l)$ are the expected frequencies assuming independence, and the summations are over j, k, and l. There are 8 "observed" frequencies in each row of Table A.7, which sum to n, so there are 8 — 1 = 7 degrees of freedom in the data. From these, 3 parameters are estimated from the marginal means (binary choice proportions), representing $p(X = j)$, $p(Y = k)$, and $p(Z = l)$, leaving 7 — 3 = 4 degrees of freedom. From the calculated value of χ^2, one can compute the probability of obtaining an equal or higher value of χ^2, according to the Chi-Square distribution with 4 df.

A statistic that is similar to χ^2 is G^2, which is defined as follows:

$$G^2 = 2\sum_j \sum_k \sum_l F(j,k,l)\,ln\left(\frac{F(j,k,l)}{E(j,k,l)}\right)$$

where $ln(x)$ is the natural logrithm, and the summations are over j, k, and l. This test has the same number of degrees of freedom, and is also assumed to be Chi-Squared distributed. This formula is a special case of a likelihood ratio test, and it is regarded as a better approximation to the Chi-Square distribution.

The approximation of either computed statistic, χ^2, or G^2 to the theoretical Chi-Square distribution is not good when n is small and when expected frequencies are small. There is a rule of thumb that expected cell frequencies should exceed 5 and n should exceed 35 for this approximation to be considered acceptable. Many of the cases in Table A.7 are near or even below this rule of thumb for considering either χ^2 or G^2 to be an accurate approximation. But these are the situations for which a simulation method is required.

Three methods are compared in Table A.7, where the p-value is calculated from χ^2 and G^2 assuming the Chi-Square distribution, or it is simulated using the R program and the variance method to estimate p_V with 10,000 pseudo-random permutations. Case 1 satisfies independence perfectly, and all three methods yield $p = 1$. Cases labeled "a" or "b" have larger values of n but the same relative frequencies as cases with the same number. So Case 1a is also perfectly consistent with independence, and it is also correctly diagnosed by all three methods.

Cases 2, 3, and 4 in Table A.7 are very close to satisfying independence (all frequencies are within rounding error of perfect independence). All three methods agree that independence is acceptable for these cases.

The other cases in Table A.7 have varying degrees of violation of independence, with more extreme departures in the lower rows of the table. Comparing the three methods, we find that the estimated p_V values by the simulation method show "regression" compared to the calculated values; that is, the simulated p_V -values are often lower for large p and higher for small p. The simulated p_V-values are more "conservative" in certain cases where one would consider rejecting the null hypothesis, as in Cases 6a and 10. An interesting exception is in Case 5a, where the χ^2 method would declare statistical significance but G^2 would not; in that case, p_V was smallest of the three methods. Summarizing the cases studied, the correlation between the p-levels calculated by χ^2 and G^2

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

Table A.7. Results of Monte Carlo simulations for Hypothetical Data with Three Variables. The hypothetical frequencies of response combinations, total n, and p-values given three methods. The last column shows Monte Carlo results based on 10,000 simulations.

Case	000	001	010	011	100	101	110	111	n	χ^2	G^2	p_V
1	5	5	5	5	5	5	5	5	40	1.000	1.000	1.000
1a	10	10	10	10	10	10	10	10	80	1.000	1.000	1.000
2	12	6	6	3	6	3	3	1	40	0.998	0.998	0.876
2a	24	12	12	6	12	6	6	2	80	0.992	0.990	0.830
3	3	1	3	1	11	5	11	5	40	0.998	0.998	0.880
3a	6	2	6	2	22	10	22	10	80	0.994	0.993	0.850
4	6	1	13	4	4	1	9	2	40	0.989	0.988	0.811
4a	12	2	26	8	8	2	18	4	80	0.960	0.957	0.803
5	8	4	4	4	4	4	4	8	40	0.308	0.363	0.197
5a	16	8	8	8	8	8	8	16	80	0.048	0.070	0.024
5b	24	12	12	12	12	12	12	24	120	0.006	0.011	0.002
6	8	2	2	8	5	5	5	5	40	0.126	0.103	0.383
6a	16	4	4	16	10	10	10	10	80	0.006	0.004	0.067
6b	24	6	6	24	15	15	15	15	120	0.000	0.000	0.016
7	8	2	2	8	6	4	4	6	40	0.092	0.074	0.111
7a	16	4	4	16	12	8	8	12	80	0.003	0.002	0.005
8	9	1	1	9	5	5	5	5	40	0.012	0.005	0.109
8a	18	2	2	18	10	10	10	10	80	0.000	0.000	0.006
9	8	2	2	8	8	2	2	8	40	0.006	0.004	0.003
9a	16	4	4	16	16	4	4	16	80	0.000	0.000	0.000
10	2	10	10	2	10	2	2	2	40	0.002	0.001	0.200
10a	4	20	20	4	20	4	4	4	80	0.000	0.000	0.033
10b	6	30	30	6	30	6	6	6	120	0.000	0.000	0.005
11	26	2	2	2	2	2	2	2	40	0.002	0.008	0.001
12	14	2	2	2	2	2	2	14	40	0.000	0.000	0.000

have a correlation exceeding 0.999; the correlation between the p-value calculated from χ^2 and p_V is 0.980. Apparently, all three methods give fairly similar results.

Listing 1. Listing of the program in R.

```
# This is R code to analyze independence in choice data for each subject
nchoices<-10 # nchoices is the number of choices (columns)
nreps<-20 # nreps is the number of repetitions of the study (rows)
nsubs<-18 # nsubs is the number of subjects.
nruns<-10000 # nruns is the number of random permutations (e.g., 10000)
outfile="results2.txt" # outfile is where the results will be printed
files1<-c("reg_01.txt","reg_02.txt","reg_03.txt",
  "reg_04.txt","reg_05.txt","reg_06.txt","Reg_700a.txt",
  "Reg_800a.txt","reg_09.txt","reg_10.txt","Reg_1100a.txt",
  "reg_12.txt","reg_13.txt","reg_14.txt","reg_15.txt",
  "reg_16.txt","reg_17.txt","reg_18.txt")
for (iii in 1:nsubs) {
file1<-files1[iii]
mm=read.table(file1) # read in the data for one subject
x <- mm # x (same as mm) is a matrix of the original data
z=array(0,c(nreps,nreps)) # Here arrays are initialized
zz=array(0,c(nreps*nreps))
xperm=array(0,c(nreps,nchoices))
zperm=array(0,c(nreps,nreps))
zzperm=array(0,c(nreps*nreps))
vardist=array(0,c(nruns))
cordist=array(0,c(nruns))
repdif=array(0,c(nreps,nreps))
rrdif = array(0,c(nreps*nreps))
zzap=array(0,c(nreps-1))
sum=array(0,c(nreps-1))
# These are calculations on the original data
# z is the matrix of disagreements between reps in original data
for (i in 1:nreps) { for (j in 1:nreps)
{ for (k in 1:nchoices) { z[i,j] = z[i,j]+ (x[i,k]-x[j,k])^2 }
repdif[i,j]<-abs(i-j)
}}
zz<-c(z)
a <- mean(zz)
b <- var(zz)
# here we calculate the correlation between rep difference and distance
nn<-nreps-1
for (id in 1:nn) {
sum[id]<-0
ni<-nreps – id
for (i in 1:ni) {
j<-(i+id)
sum[id]<- sum[id]+ z[i,j] }
```

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press

```
zzap[id]<-sum[id]/(nreps-id) }
repdif2<-c(1:nn)
c<-cor(zzap,repdif2)
# Here begin calculations on permuted data. Note that data are permuted across rows within columns. This leads to tests of iid independence.
# xperm is a permutation of the data
# zperm is the matrix of disagreements between reps in the permuted data
# totvar is the number of cases where the variance of permuted data exceeds the variance in the original data.
totvar=0.0
totcor=0.0
for (kk in 1:nruns) {
for (ii in 1:nreps){
for (jj in 1:nchoices) {xperm[,jj]<-x[sample(nreps,nreps),jj]} }
for (it in 1:nreps) {
for(jt in 1:nreps) {zperm[it,jt]=0} }
for (i in 1:nreps) {
for (j in 1:nreps) {
for (k in 1:nchoices) {zperm[i,j] = zperm[i,j]+ (xperm[i,k]-xperm[j,k])ˆ2 } }}
zzperm<-c(zperm)
a1<-mean(zzperm)
b1<-var(zzperm)
vardist[kk]=b1 # vardist a vector of variances of zperm
if (b1 >= b) {totvar=totvar+1}
# calculate correlation btn. rep difference and distance in permuted data
nn<-nreps-1
for (id in 1:nn) {
sum[id]<-0
ni<-nreps – id
for (i in 1:ni) {
j<-(i+id)
sum[id]<- sum[id]+ zperm[i,j] }
zzap[id]<-sum[id]/(nreps-id) }
repdif2<-c(1:nn)
c1<-cor(zzap,repdif2)
cordist[kk]<-c1
if (abs(c1) >= abs(c)) {totcor=totcor+1}
}
p=totvar/nruns # p is the p-value of the variance test of iid
p2=totcor/nruns # p2 is the p-value of the correlation test
o1=c(file1,a,b,p,c,p2,nruns) # This is the list for printout
sink(outfile,append=TRUE)
print(o1) # Here the results are printed to the output file
sink()
# hist(vardist) this would display histogram sampling distb. under H0
# plot(density(vardist)) this would display the density of above histogram
}
```

https://doi.org/10.1017/S193029750000187X Published online by Cambridge University Press