MEASURING SIMILARITY BETWEEN BINARY RESPONSES TO REPEATED QUESTIONS

Madhuri S. Mulekar and C. Scott Brown

University of South Alabama, 411 University Blvd, ILB 325, Mobile, AL 36688

Abstract
Similarity between objects is determined by comparing different characteristics, as a result the definition of similarity varies depending on the situation. Similarity measures are often used to study the association between two factors. In this study, we use similarity measures to determine the reliability of binary responses on repeated questionnaires. The technique is demonstrated using data collected from a nursing study.

Key Words: Similarity index, measure of association, repeated observations

1. Introduction

Pre- and post-test questions are used very commonly to measure the effectiveness of interventions. Suppose a question has binary responses such as yes/no, male/female, agree/disagree, present/absent, positive/negative, etc. Let us indicate two responses (or outcomes) on a question as positive/negative. Then the outcomes of pre- and post-test questions from *n* subjects can be summarized using a 2×2 table as follows (Table 1):

Table 1: A 2×2 contingency table

	Positive Post-test	Negative Post-test	Total
Positive Pre-test	*a*	*b*	*a* + *b*
Negative Pre-test	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*n* = *a* + *b* + *c* + *d*

Here, *a* indicates the number of subjects with a positive response on both pre- and post-tests, *b* indicates the number of subjects with a positive response on the pre-test but a negative response on the post-test, *c* indicates the number of subjects with a negative response on the pre-test but a positive response on the post-test, and *d* indicates the number of subjects with a negative response on both pre- and post-tests.

Such 2×2 tables have been used to measure the coexistence of two species at different locations (Sokal and Sneath, 1963), to measure observer agreement in classifying dichotomous objects (Fleiss, 1975), to compare two partitions of a set obtained using different clustering algorithms (Albatineh, Niewiadomska-Bugaj, and Mihalko, 2006), and to compare MRI image segmentation algorithms for evaluating the performance of skull stripping (Shattuck et al., 2009). Warrens (2008) studied the properties of a family of association coefficients that are linear transformations of the observed proportion of agreement given the marginal probabilities. Nekola and White (1999) studied changes in biography and ecology similarity with respect to the changes in distances in biological communities.

The literature is full of many different measures of similarity, association, and agreement. Here we consider a small subset of these measures and apply to outcomes from one study in nursing education. Seven measures considered here are: Dice's coefficient, Jaccard's coefficient, Yule's coefficient, Chi-sq measure, McNemar's statistic, simple matching coefficient, and similarity index.

2. Some Similarity Measures for Binary Responses

2.1 Dice's Coefficient (*D*):

Dice's coefficient of similarity for two strings of binary responses is given by,

$$D = \frac{2a}{2a + b + c}. \qquad (1)$$

Since Dice (1945) proposed this coefficient or index to measure the amount of ecologic association between species, scientists from other disciplines have very commonly used it to measure association or similarity between two sets of measurements. Because of its computationally simple nature, Dice's coefficient is used to determine the level of similarity between documents, images, and information retrieval (Maron and Kuhns, 1960, and Can and Ozkarahan, 1985), and to establish genetic similarity between species or diversity of strains (Yang, et al, 2010). It takes values in the range of (0, 1), where a value of 0 indicates no overlap and a value of 1 indicates perfect agreement with higher numbers indicating better agreement or more similarity.

2.2 Jaccard's Index (*J*):

Another popular similarity measure is the Jaccard index (Jaccard, 1901) given by,

$$J = \frac{a}{a+b+c}. \tag{2}$$

It also takes values between 0 and 1 with 1 indicating complete similarity and 0 indicating complete dissimilarity. Jaccard's index is similar to the Dice's index, but gives half the weight to agreement on the positive responses. It is a simple and easy to use index. It is popularly used in distance-decay studies as a measure of biodiversity. It works by comparing the species diversity between ecosystems. Sørenson (1948) modified Jaccard's index by taking into account the number of species and adjusting the numerator and denominator accordingly.

2.3 Yule's coefficient of association (*Q*):

Yule's coefficient (Yule, 1912) is used to measure proximity of algorithms. It assesses the predictability of the state of the characteristic – say positive or negative - for one item given the state of another item. It takes values between -1 and 1 where -1 indicates a negative association, 0 a no association, and +1 a positive association. It is a 2×2 version of Goodman and Kruskal's ordinal measure *gamma* (Goodman and Kruskal, 1954, 1959). For a 2×2 data it is given by,

$$Q = \frac{ad-bc}{ad+bc}. \tag{3}$$

It is associated with the odds ratio (*OR*) as $Q = (OR-1)/(OR+1)$.

2.4 Chi-square measure of association (χ^2):

Commonly used to compare observed data with the data expected under some assumptions, the χ^2 measure arises as a test statistic for a test of association. The properties of this measure were studied by Pearson (1900). This measure is available for $r\times c$ contingency tables where r = number of rows ≥ 2 and c = number of columns ≥ 2. For a 2×2 table, it is given by,

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+c)(b+d)(a+b)(c+d)}. \tag{4}$$

It takes values in the range of $(0,\infty)$ with the larger values indicating more dissimilarity.

2.5 McNemar's test statistic (*MN*):

McNemar's test was introduced by McNemar (1947) to assess the significance of the difference between two correlated proportions, unlike the χ^2 measure used for uncorrelated proportions. The test statistic for McNemar's test is given by,

$$MN = \begin{cases} \dfrac{(b-c)^2}{b+c} & \text{if} \quad b+c \geq 25 \\ \dfrac{(|b-c|-0.5)^2}{b+c} & \text{if} \quad b+c < 25. \end{cases} \tag{5}$$

The formula for a small sample size (*n* < 25) includes a Yates' correction for continuity (Yates, 1934). McNamar's test is useful to study the association between a genetic marker and a trait (Spielman, et al., 1993). It can be used to measures the over-transmission of an allele from heterozygous parents to affected offspring. The *n* affected offspring have total 2*n* parents who can be represented by the transmitted and the non-transmitted alleles resulting in a 2×2 table.

2.6 Simple matching coefficient (*p*):

Viewed as the simplest of all association measures, the simple matching coefficient is the intersection of two criteria. Also known as the observed proportion of agreement,

$$p = \frac{a+d}{n}, \tag{6}$$

i.e. it gives a ratio of positive and negative matches to the total sample. It is useful when both positive and negative responses carry equal weight or provide the same amount of information. Commonly used in information retrieval, it indicates the number of shared index terms.

2.7 Similarity Index (*S*):

Consider a population that is classified using two different criteria into a $2 \times C$ $(C \geq 2)$ contingency table, such as a question with $C \geq 2$ different options for an answer. It results in two subpopulations, each classified into *C* groups as shown in Table 2.

Here X_{1j} is the number of subjects providing the j^{th} response on the pre-test and X_{2j} is the number of subjects providing the j^{th} response on the post-test $(j = 1, 2, \ldots, C)$. Mulekar, Knutson, and Champanerkar (2008) studied behavior of a dissimilarity index computed from a $2 \times C$ contingency table under different configurations. They also compared different approximations for its expected value with the true value and different approximations for the variance of its estimate. Its complement, the index of similarity, is given as follows:

$$S = 1 - \frac{1}{2n}\sum_{j=1}^{C}\left|X_{1j} - X_{2j}\right|. \tag{7}$$

The value of *S* ranges from 0 to 1, where the value zero indicates complete dissimilarity and a value of 1 indicates complete similarity. The index of similarity is symmetric in nature and is invariant with respect to the sample size and scale. Refer to Mulekar, Knutson, and Champanerkar (2008) for more information about the behavior and uses of complement of *S*, the index of similarity. For $C = 2$, using relations $X_{11} = a + c$, $X_{12} = c + d$, $X_{21} = a + c$, and $X_{22} = b + d$, this index reduces to:

$$S = 1 - \frac{|b - c|}{n}. \tag{8}$$

Table 2: Classification of pre- and post-test outcomes with $C \geq 2$ possible responses to a question

	Possible response				Number of subjects taking the test
	1	2	$\cdots$	C	
Pre-test	X_{11}	X_{12}	$\cdots$	X_{1C}	n
Post-test	X_{21}	X_{22}	$\cdots$	X_{2C}	n

3. An Example using Nursing Education Data

The clinical experience prepares nursing students for practice. However, many students lack self-confidence and are stressed over meeting expectations; this leads to high level of stress. Such stress can lead to further reduction in self-confidence and the inability to perform tasks. Thus nursing educators use encouragement strategies to improve students' self-confidence. A study was designed at the University of South Alabama's College of Nursing to measure the effects of simulation-based learning on the novice students' first clinical experience (Dearmon, et al., 2012).

The goal was to determine if the simulations are successful in decreasing anxiety and improving confidence of nursing students preparing for their clinical experience. Pre- and post-tests were used to assess students' knowledge, anxiety level, and self-confidence before and after simulation-based orientation. Fifty students entering their first clinical course participated in the study. The summary of student responses for nine of the twelve questions from the knowledge assessment questionnaire is presented in Table 3. Responses to the remaining three questions were not used in the analysis presented here for reasons unrelated to computation of similarity measures. Note that terms *a*, *b*, *c*, *d*, and *n* are defined in Table 1.

Table 3: Outcome of nine knowledge-based questions on the pre- and post-tests

Question	*a*	*b*	*c*	*d*	*n*
Q1	3	4	6	37	50
Q2	12	5	12	21	50
Q3	2	2	3	43	50
Q4	8	3	13	26	50
Q5	29	4	7	10	50
Q6	13	6	8	23	50
Q7	2	1	1	46	50
Q8	0	0	3	47	50
Q9	3	4	8	35	50

Table 4: Values of different similarity measures for outcomes on nine knowledge-based questions before and after

Question	*S*	*Q*	*MN*-adj	Chi-sq	*p*	*D*	*J*
Q1	0.96	0.64	0.23	3.41	0.80	0.38	0.23
Q2	0.86	0.62	2.49	5.27	0.66	0.59	0.41
Q3	0.98	0.87	0.05	7.73	0.90	0.44	0.29
Q4	0.80	0.68	5.64	5.47	0.68	0.50	0.33
Q5	0.94	0.82	0.57	12.14	0.78	0.84	0.73
Q6	0.96	0.72	0.16	8.78	0.72	0.65	0.48
Q7	1.00	0.98	0.13	20.83	0.96	0.67	0.50
Q8	0.94	**	2.08	**	0.94	0.00	0.00
Q9	0.92	0.53	1.02	2.06	0.76	0.33	0.20

** Does not exist because of 0 in the denominator.

As noticed from the Figure 1, there is a considerable amount of variation in the computed values for different measures. It is not easy to decide which measure of similarity should be used. If we were to select similarity index *S*, then across all the questions, except question number 4, a very high degree of similarity was observed between the responses by students on the pre- and post-tests. Does it mean the intervention was not effective in improving the performance of students? On the other hand, if we were to select Jaccard's index *J*, then on all questions it resulted in the lowest numerical value. In fact, except question number 5, all the remaining questions resulted in index value below 0.5 indicating very little similarity. Does that mean the intervention was influential (do not know whether for better or worse) in changing student responses? It is a difficult choice for an experimenter. Which measure to use? There are ethical issues associated with the selection process if the experimenter selects measure of similarity depending on the outcome of experiment.

It is important to note that different measures take values on different scales. Not all of them take values between 0 and 1, which makes it more difficult to interpret values and compare values for different measures.

- Under the extreme configurations, some of these measures may not be useful in practice. For example, $a = b = 0$ resulted in Q and χ^2 to be undefined and $J = D = 0$.
- The measures p and Q resulted in very similar numerical values for all these data, except when $a = b = 0$.
- The measure *S* resulted in the consistently highest numerical values for all the questions considered here and *J* in the lowest.
- The measure *J* (as expected) is shifted upwards compared to *D* for all questions discussed here, except in the case of $a = b = 0$ when both resulted in the same value.

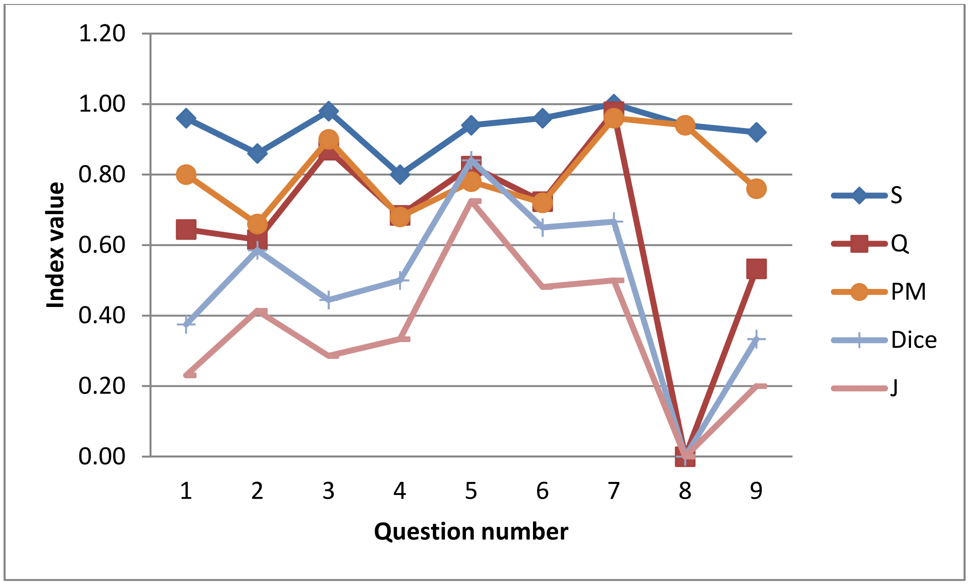

Figure 1: Comparison of Similarity indices

4. Conclusions

Different similarity measures computed from the same data may result in different numerical values. Under certain configurations, almost contradictory results may occur. Conclusions from a study may therefore depend on the choice of the similarity measure used. As a result, the questions that arise are: How is one supposed to choose one similarity measure from the possible selection to use in a given situation, and what is the ethical dilemma associated with such a selection?

References

Albatineh, A.N., Niewiadomska-Bugaj, M., and Mihalko, D. (2006). On similarity indices and correction for chance agreement, *Journal of Classification*, 23: 303-313.

Can, F. and Ozkarahan, E.A. (1985). Concepts of the cover coefficient-based clustering methodology, In: *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, 204-211.

Dearmon, V., Graves, R., Hayden, S., Mulekar, M.S., Lawrence, S., Jones, L., Smith, K.K., and Farmer, J.E. (2012). The effectiveness of a simulation-based orientation on knowledge acquisition, anxiety, and self-confidence in baccalaureate nursing students preparing for the first clinical experience, *Journal of Nursing Education*, Reviewed and accepted for publication.

Dice, L.R. (1945). Measures of the amount of ecologic association between species, *Ecology* 26 (3): 297–302.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait, *Biometrics*, 31: 651-659.

Goodman, L.A. and Kruskal, W.H. (1954). Measures of Association for Cross Classifications, *Journal of the American Statistical Association*, 49 (268): 732–764.

Goodman, L.A., and Kruskal, W.H. (1959). Measures of Association for Cross Classifications. II: Further Discussion and References, *Journal of the American Statistical Association*, 54 (285): 123–163.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37: 547–579.

Maron, M. E. and Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval, *Journal of the Association for Computing Machinery*, 7(3): 216–244.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, 12 (2): 153–157.

Mulekar, M.S., Knutson, J.C., and Champanerkar, J.A. (2008). How useful are approximations to mean and variance of the index of dissimilarity? *Computational Statistics & Data Analysis*, 52: 2098-2109.

Nekola, J.C. and White, P.S. (1999). The distance decay of similarity in biography and ecology, *Journal of Biogeography*, 26(4): 867-878.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine*, Series 5, 50(302): 157-175.

Shattuck, D.W., Prasad, G., Mirza, M., Narr, K.L., and Toga, A.W. (2009). Online resource for validation of brain segmentation methods, *NeuroImage*, 45(2): 431–439.

Sokal, R.R. and Sneath, P.H. (1963). *Principles of numerical taxonomy*, Freeman, San Francisco.

Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, Kongelige Danske Videnskabernes Selskab. *Biol. krifter*. Bd V. (4): 1-34.

Spielman R.S., McGinnis R.E., and Ewens W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics*, 52(3): 506–16.

Warrens, M.J. (2008). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions, *Psychometrika*, 73(4): 777-789.

Yang, X.L., Li, G.J., Li, S.F., and Wen, H.A. (2010). Genetic diversity of *Trametes versicolor* revealed by inter-simple sequence repeat markers, *Mycosystema*, 29(6): 886-892.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test., *Supplement to the Journal of the Royal Statistical Society*, 1(2): 217–235.

Yule, G.U. (1912). On the methods of measuring association between two attributes, *Journal of the Royal Statistical Society*, LXXV: 579-652.