The International Journal of Biostatistics

Volume 5, Issue 1

2009

Article 17

Measuring Agreement about Ranked Decision Choices for a Single Subject

Robert H. Riffenburgh, *Naval Medical Center San Diego
and San Diego State University*

Peter A. Johnstone, *Indiana University School of Medicine*

Recommended Citation:

Riffenburgh, Robert H. and Johnstone, Peter A. (2009) "Measuring Agreement about Ranked Decision Choices for a Single Subject," *The International Journal of Biostatistics*: Vol. 5: Iss. 1, Article 17.

DOI: 10.2202/1557-4679.1113

Measuring Agreement about Ranked Decision Choices for a Single Subject

Robert H. Riffenburgh and Peter A. Johnstone

Abstract

Introduction. When faced with a medical classification, clinicians often rank-order the likelihood of potential diagnoses, treatment choices, or prognoses as a way to focus on likely occurrences without dropping rarer ones from consideration. To know how well clinicians agree on such rankings might help extend the realm of clinical judgment farther into the purview of evidence-based medicine. If rankings by different clinicians agree better than chance, the order of assignments and their relative likelihoods may justifiably contribute to medical decisions. If the agreement is no better than chance, the ranking should not influence the medical decision.

Background. Available rank-order methods measure agreement over a set of decision choices by two rankers or by a set of rankers over two choices (rank correlation methods), or an overall agreement over a set of choices by a set of rankers (Kendall's W), but will not measure agreement about a single decision choice across a set of rankers. Rating methods (e.g. kappa) assign multiple subjects to nominal categories rather than ranking possible choices about a single subject and will not measure agreement about a single decision choice across a set of rankers.

Method. In this article, we pose an agreement coefficient A for measuring agreement among a set of clinicians about a single decision choice and compare several potential forms of A . A takes on the value 0 when agreement is random and 1 when agreement is perfect. It is shown that $A = 1 - \text{observed disagreement}/\text{maximum disagreement}$. A particular form of A is recommended and tables of 5% and 10% significant values of A are generated for common numbers of ranks and rankers.

Examples. In the selection of potential treatment assignments by a Tumor Board to a patient with a neck mass, there is no significant agreement about any treatment. Another example involves ranking decisions about a proposed medical research protocol by an Institutional Review Board (IRB). The decision to pass a protocol with minor revisions shows agreement at the 5% significance level, adequate for a consistent decision.

KEYWORDS: agreement, decision making, rank-order, tumor board, IRB, Institutional Review Board

Author Notes: We express heartfelt thanks to Dr. Bruce Levin, Professor and Chair, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, and to two

anonymous reviewers for insightful suggestions that have greatly improved this paper. Any errors that appear, however, are our own.

1. Introduction

Medical practitioners and investigators are often gathered to make a group decision, which is straightforward if they all agree. In many if not most cases, there is some disagreement. How are decisions made?

For example, consider an Institutional Review Board (IRB). Having sat on an IRB for nearly two decades, the first author has seen patterns of decision making remain consistent despite continual changes in membership. The primary patterns include consensus, decision rule by minority, and uneven influence by members. In the consensus pattern, a proposed protocol is discussed continually by the members until it appears that no new perceptions are forthcoming and at least several members hold a consensus of impressions about the protocol. Someone proposes a disposition and the board votes. However, a decision can be accepted when supported by only a minority. Suppose there are 12 members. A quorum is 7. A majority vote of the quorum is 4, only a third of the membership. Furthermore, at times one member exerts undue influence, due to an irrational passion over an issue, a persistence-type of personality in which he insists on his position until others vote his way to end the session, or a dominating personality in which he cajoles or bullies others into voting his way. While a majority of decisions are balanced and reasonable, some are not.

Many other forms of group decisions in medicine are subject to these or other shortcomings, including tumor boards, exemplified later in this paper, and choosing which to fund from among a number of proposals for a medical research project. What can be done to improve the balance?

One solution would be to rank-order the potential choices. Indeed, when faced with a medical classification, clinicians often rank-order potential diagnoses, treatment choices, assessments, or other categorizations as a way to focus on the likely ones without dropping rarer occurrences from consideration. A realistic recent example is the selection of salvage therapy for patients having rising PSA after prostate cancer therapy (Johnstone, Booth, Riffenburgh, et al, 2002). A “best” treatment on average was designated, but there was little agreement about it. To know how well clinicians agree on such rankings might help extend the purview of evidence-based medicine farther into the realm of clinical judgment. If rankings by different clinicians agree better than chance, the order of assignments and their relative likelihoods may be an improvement over the usual decision making process. For the members to see a significant ranking might mute the influence of the persistent or dominating member. However, if the ranking agreement is no better than chance, the ranking should not be allowed to influence the medical decision.

To pursue this issue of group decision by ranking if agreement is good, we propose a rank-based agreement statistic we have designated as A . We propose

that the probability of A being better than chance be found and that A not be used if it is not better than chance.

A is to be developed as a measure of concentration of rank agreement. Concentration of information has been seen in a number of guises, such as entropy (Hart, 1971), and has been developed for a number of applications, such as the Gini index and its generalizations in econometrics (Chakravarty, 1988), an indicator of concentration of business activity.

We are concerned with a number k of decision choices ranked by each of a number n of clinicians. We define the vector $[n_1, n_2, \dots, n_k]$ of frequencies of ranks. The vectors $[n, 0, \dots, 0]$ would represent perfect agreement (on the first decision choice) and $[n/k, n/k, \dots, n/k]$, random assignment. We define the measure of agreement A as a weighted sum of multinomial frequencies of ranks, a form having well known theory, but the nature of weighting must be defined. Our recommended form of A uses median rank as the basis of weight, where we define the median (m) for each possible decision choice as the rank position at which half the rankers chose a rank equal or below and half, equal or above. Later, other weighting schemes will be considered.

Sections numbered below will address the following topics. (2) An example will be posed. (3) Established methods of rank correlation and concordance will be examined to show why they will not provide the needed information. (4) Our form of A using median rank weight will be derived. (5) Various forms of A using other weighting schemes will be considered. (6) We will provide tables of 5% and 10% critical values at or above which A is significantly greater than chance for common numbers of categories and rankers. (7) We illustrate it with an example of potential treatment assignments by a Tumor Board to a patient with a neck mass and an example of decisions about a proposed protocol by an IRB. (8) Conclusions. At the end of the article is an Appendix. Readers of early versions of this paper have asked about the contrast of A with intra-class correlation coefficients. Although they are a fundamentally different type of measure than A , an attempt to answer these questions will be provided in the Appendix.

2. Example: Evaluation of a Proposed Research Protocol by an IRB

At an IRB meeting composed of $n = 7$ voting members, a proposed research protocol is considered. There are $k = 4$ possible decisions: (a) pass as is, (b) pass with minor revisions, (c) table for major revisions, or (d) reject. The Board must select one. We will designate these, respectively, *Pass*, *Revise*, *Table*, and *Reject*, for shorthand convenience. After a particular proposal was presented and discussed, members ranked their choices 1 through 4. The original data appear as Table 1.

Table 1. Rankings of four possible decisions about a medical research protocol proposal by 7 IRB members.

Decision	IRB Member Number						
	1	2	3	4	5	6	7
1 st Rank	<i>Pass</i>	<i>Revise</i>	<i>Revise</i>	<i>Revise</i>	<i>Revise</i>	<i>Revise</i>	<i>Revise</i>
2 nd Rank	<i>Revise</i>	<i>Pass</i>	<i>Pass</i>	<i>Pass</i>	<i>Table</i>	<i>Table</i>	<i>Pass</i>
3 rd Rank	<i>Table</i>	<i>Table</i>	<i>Table</i>	<i>Table</i>	<i>Pass</i>	<i>Pass</i>	<i>Table</i>
4 th Rank	<i>Reject</i>	<i>Reject</i>	<i>Reject</i>	<i>Reject</i>	<i>Reject</i>	<i>Reject</i>	<i>Reject</i>

We constructed another table with a row for each of the 4 choices with 4 cells in each row, one for each possible ranking. For each choice, we recorded the number of rankers selecting each rank for that choice. The data appeared as in the first four data columns of Table 2. For example, as the first choice, one member ranked *Pass*, 6 ranked *Revise*, and no one ranked *Table* or *Reject*, leading to the entries for the column under “1st Rank” in Table 2. The fifth column gives the median rank (m). (The agreement coefficient A will be derived in section 4.)

The challenge is to decide if the IRB members agree adequately on a choice, in this example, what to do with the protocol. In this case, the first choice (a median rank of 1) is to pass the protocol with minor revisions. However, if agreement should be no better than chance, the Board’s decision would be arbitrary and does not serve the system; random IRB decisions are worse than having no IRB.

Table 2. Results for seven IRB members’ ranking decisions for a medical research protocol proposal. Shown are number of rankers choosing each rank, medians of ranks, and an agreement coefficient A per decision choice for a median rank weighting scheme. Asterisks indicate coefficients significant at the 5% level.

Decision Choices	Rank				m	Agreement Coefficient A
	1 st	2 nd	3 rd	4 th		
<i>Pass</i>	1	4	2	0	2	0.57
<i>Revise</i>	6	1	0	0	1	0.86*
<i>Table</i>	0	2	5	0	3	0.71
<i>Reject</i>	0	0	0	7	4	1.00*

3. Other Ranking Methods

We could calculate a rank correlation coefficient [Spearman’s rho (1904) or Kendall’s tau (1938) (mathematically similar measures)] between members of each possible pair of rankers, resulting in 21 coefficients. This tells us little, tests of significance lead to an accumulating p value, and the rank correlation coefficients are not independent; such a scheme would not be helpful. We could

calculate Kendall's coefficient of concordance W (1948), indicating an agreement among Board members over all the choices simultaneously, but that admits the influence of possible disagreements over 2nd and 3rd, etc., choices not germane to a first choice agreement. For these data, $W = 0.66$. Friedman's test of significance yields $p = 0.003$. There is indeed significant agreement among the Board members overall, but it is not clear how much this is influenced by the perfect agreement that *Reject* should be the choice of last rank. Do they agree adequately on *Revise* for a first choice to become the Board's recommendation? To answer this question, the A statistic is needed.

4. An Agreement Coefficient

We should like to derive a measure of agreement among the rankers about a single decision choice. Let k denote the number of decision choices, $i = 1, 2, \dots, k$, which implies k possible ranks (cells per row in a data table) per choice; let n denote the number of rankers, $j = 1, 2, \dots, n$; and let n_i denote the number of rankers choosing the i^{th} rank. The median rank (m) was defined as the rank position at which half the rankers chose a rank equal or below and half, equal or above. Weights for these row entries are taken as the absolute value of difference between the median of ranks m and the rank order. In the IRB application, $k = 4$ choices, $n = 7$ rankers. The n_i 's for *Pass*, the first row of Table 2, are 1, 4, 2, and 0, the median rank $m = 2$, and the weights are, respectively, 1, 0, 1, and 2.

Let us think of agreement as the complement of disagreement ($1 - \text{disagreement}$) and of disagreement as the sum of choices in cells other than the median cell, weighted by the rank difference from the median. Denoting by d_{obs} the observed disagreement about a potential choice and as d_{exp} the expected disagreement under random assignment (the vector $[n/k, n/k, \dots, n/k]$), i.e. under the null hypothesis of random agreement, we may write

$$d_{\text{obs}} = \sum_{i=1}^k n_i \times |m - i| \quad , \quad (1)$$

where $|m - i|$ is the weight for the i^{th} rank, and

$$d_{\text{exp}} = \frac{n}{k} \sum_{i=1}^k \left| \frac{k+1}{2} - i \right| \quad . \quad (2)$$

In the IRB example, the observed disagreement d_{obs} for *Pass* is $1 \times 1 + 4 \times 0 + 2 \times 1 + 0 \times 2 = 3$ and the expected disagreement d_{exp} is $(7/4) \times (1\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1\frac{1}{2}) = 7$. Also, the maximum possible disagreement, d_{max} , in the sense of the largest value of d_{obs} that can occur in practice, may be calculated as

$$d_{\max} = (k-1) \times \frac{n}{2}, \quad n \text{ even},$$

$$d_{\max} = (k-1) \times \left(\frac{n-1}{2} \right), \quad n \text{ odd.} \quad (3)$$

Let us define level of agreement a as the complement of the disagreement ratio: observed disagreement to the maximum possible disagreement, so that

$$a_{\text{obs}} = 1 - \frac{d_{\text{obs}}}{d_{\max}} \quad (4)$$

and

$$a_{\text{exp}} = 1 - \frac{d_{\text{exp}}}{d_{\max}}. \quad (5)$$

In the example, a_{obs} for *Pass* is $1 - 3/9 = 0.67$ and $a_{\text{exp}} = 1 - 7/9 = 0.22$.

To render an agreement coefficient A easy to interpret, we should like A to take on the value 1 when agreement is perfect and 0 when the level of agreement is equal to that expected by chance. This occurs if

$$A = \frac{a_{\text{obs}} - a_{\text{exp}}}{1 - a_{\text{exp}}}. \quad (6)$$

By simple algebra, $d_{\max}$ drops out and

$$A = 1 - \frac{d_{\text{obs}}}{d_{\text{exp}}}. \quad (7)$$

Most of the time, this coefficient will lie between 0 and 1, indicating how much greater than chance the agreement is. If it is negative, it indicates that *disagreement* is greater than expected. In the example, A for *Pass* is $1 - 3/9 = 0.67$.

5. Other Forms of Weighting for A

Many forms of weighting are possible. We chose five that we found intuitively attractive: (a) unsigned deviation from median rank (already addressed) with weight $|\text{median} - i|$, (b) squared deviation from median rank with weight $(\text{median} - i)^2$, (c) unsigned deviation from mean rank with weight $|\text{mean} - i|$, (d) squared deviation from mean rank with weight $(\text{mean} - i)^2$, and (e) a form of chi-square. Schemes (c) and (d) require calculation assistance to find the mean, whereas the median can be seen by inspection. We find that $d_{\max}$ for squared weighting, i.e. (b) and (d), requires a different definition of weight, depending on n , as well as being computationally more difficult. This is not very practical, certainly less elegant than the simple weighting definition of (a). Therefore, we prefer (a) to the weighting schemes (b) through (d).

The chi-square weighting scheme, somewhat different, measures the “concentration” of agreement A_e , much as the Gini index measures the concentration of economic indicators. The chi-square form involves a commonly used calculation that has its long history and well established probability distribution to recommend it. However, the common chi-square, the sum of the squares of observed minus expected divided by expected, yields a form of A_e quite different from the other suggested A 's. The cause of this difference is straightforward. The dissenters from the concentration of primary agreement may be not very different from this concentration or quite different without affecting the measure A . In the IRB example, 6 members ranked *revise* as 1st choice and the 7th member chose *revise* as the 2nd choice. However, if the 7th member had chosen *revise* as his 4th choice, dissenting as much as possible from the other six, it would not affect A_e . Chi-square is not sensitive to the level of dissent (the “distance” of the dissenter along the vector of rank frequencies) and it is more difficult for the user to calculate.

For these several reasons, we recommend weighting scheme (a), the unsigned difference between the median and the rank position in the vector of rank frequencies.

6. Probabilities Associated with the Agreement Coefficient

To interpret the results, ideally the user would know how to relate a value of A to a clinical application, i.e., what is “high” and “low”. As with correlation and most other coefficients, the interpretation is somewhat arbitrary and lies in the judgment of the user. What can be approached unequivocally is if A is statistically significantly greater than chance, i.e. a test of the null hypothesis that $A = 0$. Because of lack of independence among choices, we recommend using only the choice with the highest median rank.

To conduct such a test, we need to know the probability distribution of randomly occurring agreement levels in order to decide if an observed A has less than probability α of occurring by chance alone. The probability distribution of possible outcomes of ranking assignments does not fall easily into a formula (or more likely, formulae, since the distributional form would be different for even and odd values of n). A computer was programmed to calculate, for each (n, k) , all values of d_{exp} and d_{obs} from all possible permutations, calculate the A occurring with each, and then tally the frequency with which each A occurs. The frequencies were ordered by possible values of A and the A identified at which 0.05 and 0.10 of the frequencies occurred as the critical value for A , i.e. the table entry.

(Interestingly, the number of permutations conveniently follows Pascal's triangle. While not proven in generality, the computer count agreed with the triangle entry for n, k from 3,3 to 23,12.)

Tables 3 and 4 provide critical values for 5% and 10% α , respectively, for commonly encountered numbers of clinical signs (k) and of rankers (n). Any A exceeding the tabulated value has α or less probability of occurring by chance.

The reader may note that table entries increase slightly in some regions as n increases, then steps down. This somewhat irregular, stair-step appearance of table entries is due to the integer nature of ranks, reducing with increasing k .

Table 3. $\alpha = 0.05$ critical values of chance occurrence for the agreement coefficient A for common values of k (3 to 9) and n (3 to 20). A value of A equal to or greater than the tabulated value will have less than 5% probability of occurring by chance.

$\alpha = 0.05$ critical values for A		k (number of clinical signs to be ranked)						
		3	4	5	6	7	8	9
n (number of rankers)	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	1.00	1.00	1.00	0.80	0.86	0.88	0.89
	5	1.00	1.00	0.83	0.86	0.78	0.80	0.73
	6	1.00	0.83	0.86	0.75	0.70	0.75	0.69
	7	1.00	0.86	0.75	0.67	0.75	0.71	0.69
	8	1.00	0.88	0.80	0.73	0.71	0.69	0.67
	9	1.00	0.78	0.73	0.67	0.67	0.67	0.60
	10	0.86	0.80	0.75	0.62	0.65	0.65	0.59
	11	0.86	0.82	0.69	0.67	0.63	0.64	0.58
	12	0.88	0.75	0.71	0.63	0.62	0.58	0.59
	13	0.89	0.77	0.69	0.59	0.59	0.58	0.55
	14	0.89	0.79	0.71	0.63	0.58	0.57	0.55
	15	0.90	0.73	0.67	0.60	0.62	0.57	0.55
	16	0.82	0.75	0.68	0.57	0.59	0.56	0.54
	17	0.82	0.76	0.65	0.57	0.59	0.56	0.54
	18	0.83	0.72	0.68	0.58	0.58	0.56	0.54
	19	0.85	0.74	0.65	0.56	0.58	0.55	0.53
	20	0.85	0.75	0.67	0.56	0.56	0.55	0.53

Table 4. $\alpha = 0.10$ critical values of chance occurrence for the agreement coefficient A for common values of k (3 to 9) and n (3 to 20). A value of A equal to or greater than the tabulated value will have less than 10% probability of occurring by chance.

$\alpha = 0.10$ critical values for A		k (number of clinical signs to be ranked)						
		3	4	5	6	7	8	9
n (number of rankers)	3	1.00	1.00	1.00	1.00	0.80	0.83	0.86
	4	1.00	1.00	0.80	0.80	0.71	0.75	0.78
	5	1.00	0.80	0.83	0.71	0.67	0.70	0.64
	6	1.00	0.83	0.71	0.63	0.60	0.67	0.62
	7	0.80	0.71	0.63	0.56	0.58	0.57	0.56
	8	0.80	0.75	0.70	0.64	0.57	0.56	0.56
	9	0.83	0.78	0.64	0.58	0.53	0.56	0.55
	10	0.86	0.70	0.67	0.54	0.53	0.55	0.50
	11	0.71	0.73	0.62	0.53	0.53	0.55	0.50
	12	0.75	0.67	0.57	0.50	0.52	0.50	0.48
	13	0.78	0.69	0.63	0.53	0.50	0.50	0.48
	14	0.78	0.64	0.59	0.53	0.50	0.50	0.48
	15	0.80	0.67	0.56	0.50	0.50	0.50	0.45
	16	0.73	0.69	0.58	0.48	0.48	0.47	0.45
	17	0.73	0.65	0.55	0.48	0.48	0.47	0.45
	18	0.75	0.67	0.59	0.46	0.48	0.47	0.43
	19	0.77	0.63	0.57	0.48	0.48	0.47	0.43
	20	0.69	0.65	0.54	0.48	0.47	0.45	0.42

7. Examples

IRB. For the IRB example with $k = 4$ decision choices and $n = 7$ rankers, it can be seen in Table 3 that the 5% critical value is 0.86 and the 10% critical value is 0.71. The choice with the highest median rank (1) is *Revise* with $A = 0.86$, significant at the 5% level. We may conclude that the Board's agreement on *Revise* is sufficient to serve as the recommendation.

Results of a Tumor Board's Deliberations. Tumor Boards are conferences used to provide multidisciplinary input into clinical decision making for cancer patients. A Tumor Board composed of 8 members considered the following case. A 68 year-old man had gradual onset of dysphagia associated with a right neck mass. Workup revealed a lesion of the right tongue base, staged T3N3 (AJCC Staging Manual, 2002), with pre-epiglottic space invasion. Further workup revealed multiple 8 mm bilateral lung nodules on a spiral computer tomography (CT) performed the prior week. Positron emission tomography revealed only the robust disease above the clavicles. Past medical history (PMH) indicated a significant pneumonia 5 years previously that cleared subsequently by CT, but that CT was not of spiral quality. The patient has been on dialysis for renal failure for 3 years. His performance status is 80, and he has lost 25 pounds from the baseline of 205 within the last 4 months. The Tumor Board may consider the

following 5 treatment choices: (a) *Surgery followed by radiation therapy (XRT)*, (b) *Hyperfractionated XRT alone*, (c) *Chemoradiation*, (d) *Induction chemotherapy followed by radiation or surgery depending on response*, or (e) *Palliative therapy*.

Each Board member ranks the possible treatments as judged most appropriate. A choice-by-rank table, Table 5, is generated, similar in format to Table 2. The last two columns provide the median ranks and the associated agreement coefficients. From Table 3, the 5% critical value for significance for 8 rankers and 5 choices is 0.80, the 10% value, 0.70. The choice with highest median rank has negligible agreement. Indeed, no choice demonstrates a significant agreement; the Board should not use ranking methods in making their decision.

This is unfortunately often the case with such a complex clinical picture. Had renal failure not been an issue, there would likely have been considerably more agreement on chemoradiation; and had the lung nodules been more virulent, palliation likely would have predominated. The crucial aspect of such a clinical

Table 5. Results of ranking of 5 potential therapies by 8 members of a Tumor Board. Shown are number of rankers choosing each rank, medians of ranks, and the agreement coefficient A for each therapy. No values of A are significant.

Potential Therapy	Rank					m	A
	1 st	2 nd	3 rd	4 th	5 th		
<i>Surgery, then XRT</i>	3	2	0	3	0	2	0.06
<i>Hyperfractionated XRT alone</i>	1	2	4	1	0	3	0.48
<i>Chemoradiation</i>	2	2	3	1	0	2.5	0.27
<i>Chemo, then XRT or surgery</i>	1	2	0	3	2	4	0.06
<i>Palliative therapy</i>	1	0	1	0	6	5	0.38

case is that each clinical option excludes the others—prudent clinicians must then rely on their own training and interpretation of the data for their ranking. Under such circumstances, the patient deserves the entire Board's evaluation of the entire universe of options as they interrelate. Forced ranking is useful in demonstrating the lack of agreement among Board members. Further discussion is required with perhaps consideration of aspects not yet included. To the patient, who clearly lacks clinical expertise, this lack of agreement should be disquieting but instructive.

8. Conclusions

Establishing a significant agreement about diagnosis, prognosis, or treatment improves the accuracy, reliability, and consistency of medical decisions, reducing arbitrariness and extraneous influence. Furthermore, research quality is improved,

ease of teaching is enhanced, and steps are taken toward evidence-based and automation-assisted medicine. On the other hand, to know the medical decision choice with highest median rank is of little use if the decision makers' agreement about this rank is no better than chance: ranking methods do not give sufficient information for decision and should not be used. This paper provides a method to measure agreement and its significance in the ranking of medical decisions.

Appendix. Contrast of A with Methods of Intra-class Correlations of Ratings

There has long been a variety of intra-class correlation coefficients using ratings (Ebel, 1951). Currently, the best known and most frequently used is kappa (κ), introduced by Cohen (1960). Fleiss (1975) showed that κ is superior in several respects to other forms of intra-class correlation coefficients, so we will use κ for our contrast. While different forms of κ exist (Kraemer 1980; Fleiss, Levin, & Paik, 2003, chapter 18), and there exists controversy on the utility of κ forms (Agresti, 2002, page 435), the distinctions to follow are valid for all forms.

κ is a fundamentally different type of measure from A . Categories used for κ are nominal; categories for A are rank-order. Data for κ are *ratings* per subject (or choice in the case of A). Each subject or choice is given a rating on its own merits independent of other subjects. However, data for A are *rankings* of subjects or choices. Each is ranked relative to others. Also, number of subjects is multiple for κ ; single for A . Number of ratings per subject is one for κ ; all possible rankings for A .

Consider the issue of selecting one of a set of possible diagnoses of a disease, given a particular constellation of symptoms. κ assesses the agreement of a group of physicians about the selection of a diagnosis in a sample of patients with the symptoms. A assesses the agreement as to what disease to treat in a single patient. For the simplest case of two raters/rankers (R_1 , R_2) and two categories (C_1 , C_2), κ data would be organized with R_1 's choices C_1 and C_2 vertically on the left, R_2 's choices C_1 and C_2 horizontally at the top, and proportions of assignment completing the table, while A data would have C_1 and C_2 vertically on the left, rankings 1st and 2nd along the top, and the number of times C_1 and C_2 were ranked 1st or 2nd completing the table. These two data sets can not be organized equivalently, illustrating the fundamental difference of κ and A . We cannot calculate a κ from the format of Table 2.

References

- Agresti A. *Categorical Data Analysis*, 2002, Wiley:NY, 2nd Ed.
- American Joint Committee on Cancer Staging Manual, 2002, 6th Ed.
- Chakravarty SR. Extended Gini indices of inequality. *International Economic Review*, 1988, **29**:147-156.
- Cohen J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 1960, **20**:37-46.
- Ebel RL. Estimation of the reliability of ratings. *Psychometrika*, 1951, **30**:407-424.
- Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 1975, **31**:651-659.
- Fleiss JL, Levin B, Paik CP. *Statistical Methods for Rates and Proportions*, 2003, NY:Wiley:NY, 3rd Ed.
- Hart PE. Entropy and other measures of concentration. *J Royal Statistical Society. Series A*, 1971, **134**, 1971:73-85.
- Johnstone PA, Booth R, Riffenburgh RH, Amling CL, Kane CJ, Moul JW. Initiation of salvage therapy for prostate cancer. *Prostate Cancer and Prostatic Diseases*, 2002, **5**:136-143.
- Kendall MG. A new measure of rank correlation. *Biometrika*, 1938, **30**:81.
- Kendall MG. *Rank Correlation Methods*, 1948, Griffin:London.
- Kraemer CK. Extension of the kappa coefficient. *Biometrics*, 1980, **36**:207-216.
- Spearman CE. The proof and measurement of association between two things. *American Journal of Psychology*, 1904, **15**:88.