Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 8 | Article 2

HOAJ
Herbert Open Access Journals

Technical Advances Open Access

Modification in inter-rater agreement statistics-a new approach

Sundus Iftikhar*

*Correspondence: sundusiftikhar@gmail.com

Head of Statistics and Training Unit, Indus Hospital Research Center, The Indus Hospital, Pakistan.

Abstract

Assessing agreement between the examiners, measurements and instruments are always of interest to health-care providers as the treatment of patients is highly dependent on the medical reports. Till now several agreement statistics have been developed and all of them have certain limitations. In 2002 Kilm Gwet introduced a more robust and unbiased agreement statistics named "Gwet's AC1 statistics". It has been shown by various researchers that AC1 statistics has the best statistical properties amongst all the other agreement statistics. Though it has been reported to be the better estimate still several inconsistencies existed in this agreement statistics. In this paper author aimed to develop a new formula that can overcome all the inconsistencies and dependencies of inter-rater agreement statistics.

Keywords: Inter-rater agreement, Agreement statistics, Kappa statistics, AC1 statistics

Introduction

In health sciences, health-care providers are always concerned about the accuracy of the measurements and the multiple outcomes. Precise results are essential in medical sciences as the life of patients depends on these results. Assessing the strength of agreement between the instruments, examiners, or a combination of the two is important to attain similar and unerring results [1]. Scoring techniques are widely used in medical science to determine certain deformities. Developing accurate scoring systems, therefore, requires substantial agreement between the raters. Several methods have been developed to assess the inter-rater agreement such as Yule's Y [2], Bennett, Alpert & Goldstein's S (1954) [3], Scott's π (1955) [4], Cohen's Kappa (1960) [5], Flesis Kappa (1971) [6], Krippendorff's Alpha (1980), Bandiwala's B statistics (1985) [7], Van Eerdewegh's V [8]. Amongst all these, Cohen's Kappa is the most popular and commonly used statistics even though; numerous inconsistencies in Kappa statistics have been pointed out. The interpretation of Kappa statistics becomes cumbersome as it is being affected by marginal probabilities, trait prevalence and skewness. Also, Kappa statistics produce awful results if the sample size is small.

Bryt T in 1993 highlighted two main dilemmas of Kappa statistics that isthe effect ofthe presence of bias between the raters and distribution of the data across the categories (prevalence). Brty T proposed new indices namely ***"Bias index"***, ***"prevalence index"*** and***"prevalence adjusted and bias-adjusted Kappa (PABAK)"***and suggested to use them instead of Kappa statistics. The author further suggested reporting bias and prevalence indices along with Kappa statistics as reporting Kappa statistics alone can be misleading [9]. More recently, in 2002 Kilm Gwet introduced a new formula of chance agreement and called the agreement statistics as "***AC1 statistics"*** and suggested to use it [10]. Many researchers have tested its validity and showed it to be robust; less affected by trait prevalence, not sensitive to marginal homogeneity and skewness and provides better results than Kappa statistics [11,12]. However, the interpretation and range of both PABAK and AC1statistics are same as of Kappa statistics.V Shankar (2014) have shown that B-statistics is a better agreement statistics amongst Kappa, delta, Aickin's alpha and AC1 statistics [13].

Kilm Gwet [10] proposed a new formula for the chance agreement that depends solely upon the values of category-1 for 2x2 cross-tables. It does not take into account the category-2 as well as the disagreement between the two raters.

In this paper, the author has shown that AC1 and other inter-rater agreement statistics are significantly affected by the change in individual cell probabilities, symmetry, marginal homogeneity and trait prevalence. Also, AC1 statistics provide no agreement in case of 50% observed agreement between the raters. To overcome these issues, the author has introduced a new, simple and easy to use a formula of chance agreement that takes into account both the categories as well adjusts for

 doi: 10.7243/2053-7662-8-2

discordant values. The chance agreement and agreement statistics are named as **Minimum expected chance agreement" and "SI statistics"** respectively.

Moreover, it has been shown that SI-statistics is not affected by trait prevalence, robust to change in cell values for the fixed observed agreement if a matrix is symmetric and marginal homogenous.

Types of inter-rater agreeement statistcs:

Statistics	Formula	Chance agreement
S -statistics	$\frac{(q*p_0)-1}{q-1}$	$\frac{1}{q-1}$
Where q: the total number of categories, p_0 is the observed agreement		
Scott's π-statistics	$\frac{p_0-e(\pi)}{1-e(\pi)}$	$e(\pi)=\sum\left(\frac{n_{i.}+n_{.i}}{2N}\right)^2;\ i=j$
Where, $n_{i.}$ is the row total of *ith* category and $n_{.i}$ is the column total of *ith* category		
Cohen's Kappa	$\frac{p_0-e_k}{1-e_k}$	$e_k=\frac{(a+b)(a+c)+(c+d)(b+d)}{N^2}$
B-Statistics	$\frac{\sum n_{ii}^2}{\sum n_{i.}*n_{.i}};\ i=j$	
Where, n_{ii} are the values of main diagonal, $n_{i.}$ is the row total of i^{th} category and $n_{.i}$ is the column total of i^{th} category		
PABAK	$2p_0-1$	where, $p_0$is the observed agreement
Gwet's AC1	$\frac{p_0-e_\gamma}{1-e_\gamma}$	$e_\gamma=2p_r(1-p_r)$
		Where, $p_r=\frac{(a+b)+(a+c)}{2N}$
Van Eerdewegh's V	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{(a+c)(b+d)}}$	
Yule's Y	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	
Cicchetti & Feinstein's p_{pos}	$\frac{2a}{(a+b)(a+c)}$	
Cicchetti & Feinstein's p_{neg}	$\frac{2d}{(c+d)(b+d)}$	

SI-Statistics

Let e_v minimum expected chance agreement calculated as

$$e_v=\frac{Average\left(Min(a+b,a+c),Min(c+d,b+d)\right)-Min(b,c)}{N}$$

$$SI-statistics=\frac{p_0-e_v}{1-e_v}$$

Where $p_0$is the observed agreement, and SI-statistics ranges from 0 to 1.

For nxn contingency table

$$\frac{Average\left(Min(n_{1.},n_{.1}),\ldots Min(n_{i.},n_{.i})\right)-Min(off-digonal\ values)}{N};i=1,\ldots.n$$

Table I. Interpretation of agreement statistics [14].

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect/Excellent
Scores	0.00	0.20	0.40	0.60	0.80	1.00
	Agreement					
0.00	No agreement					
0.01-0.20	Slight agreement					
0.21-0.40	Fair agreement					
0.41-0.60	Moderate agreement					
0.61-0.80	Substantial agreement					
0.81-0.99	Almost perfect/Excellent agreement					

A negative coefficient reflects weaker agreement than expected or discrepancy. In general, low negative values (0 to -0.10)can be interpreted as "no agreement". A large negative coefficient represents considerable discrepancy among raters [15].

Sensitivity to change in individual cell probabilities

To assess the sensitivity of SI-statistics and other inter-rater agreement statistics to change in cell values or cell probabilities, we simulated the data on sample size 24 where the observed agreement was fixed to 0.58 and the matrix was kept symmetric and marginally homogenous.

Table 1 shows the simulated data and **Table 2** shows the comparison of SI-statistics with Kappa, AC1 and other agreement statistics. Results showed that SI-statistics is unaffected by a change in individual cell probabilities for fixed observed agreement. However, all the agreement statistics are inconsistent and varies significantly except PABAK and S-statistics. Moreover, SI-statistics provides a reasonable estimate of actual (observed) agreement between the two raters.

Sensitivity to trait prevalence

Trait prevalence is defined as the likelihood that a randomly selected subject proved to be positive [10]. To assess the effect of trait prevalence, the author has fixed the sensitivity and specificity of each rater and allowed trait prevalence to vary from 0 to 1as suggested by Gwet K [10]. Data simulation was done using the following equations [10].

$$P_{A+}=p_r\alpha_A+(1-p_r)(1-\beta_A)$$
$$P_{B+}=p_r\alpha_B+(1-p_r)(1-\beta_B)$$

Where, P_r is trait prevalence, P_{A+} and P_{B+} are the probability of rater A and rater B to categorize subject as positive respectively, α_A and α_B are the sensitivity of rater A and rater B for the correctly classifying a subject as positive respectively and β_A and β_B are the specificities associated with rater A and rater B for the correctly classifying a subject as negative respectively.

Expected marginal totals for classifying a subject as positive for both rater A and rater B can be calculated as

$$A_+=nP_{A+}$$
$$B_+=nP_{B+}$$

Individual cell probabilities can be computed as follows

Rater B	Rater A		Marginal totals
	Positive (+)	Negative (-)	
Positive (+)	a (++)	b (-+)	B_+
Negative (-)	c (+-)	d (--)	B_-
Marginal totals	A_+	A_-	N

$$P_{++}=p_r\alpha_A\alpha_B+(1-p_r)(1-\beta_A)(1-\beta_B)\longrightarrow a$$
$$P_{-+}=P_{B+}-P_{++}\longrightarrow b$$
$$P_{+-}=P_{A+}-P_{++}\longrightarrow c$$
$$P_{--}=1-(P_{A+}+P_{B+}-P_{++})\longrightarrow d$$

In **Tables 3a** and **3b**, raters had common sensitivity and specificity of 0.9 and fixed observed agreement. Results showed that SI-statistics is unaffected by trait prevalence Whereas, other agreement statistics provided varying results except for PABAK and S-statistics.

doi: 10.7243/2053-7662-8-2

Table 1: Change in individual cell probabilities for symmetric and marginally homogenous data with fixed observed agreement.

		Individual cell probabilities				Row totals		Column totals		Observed agreement
Condition	**n**	**a**	**b**	**c**	**d**	**a+b**	**c+d**	**b+d**	**a+c**	**P_0**
1	24	0.083	0.208	0.208	0.500	7	17	17	7	0.583
2	24	0.125	0.208	0.208	0.458	8	16	16	8	0.583
3	24	0.417	0.208	0.208	0.167	15	9	9	15	0.583
4	24	0.583	0.208	0.208	0.000	19	5	5	19	0.583
5	24	0.000	0.208	0.208	0.583	5	19	19	5	0.583
6	24	0.042	0.208	0.208	0.542	6	18	18	6	0.583
7	24	0.542	0.208	0.208	0.042	18	6	6	18	0.583
8	24	0.292	0.208	0.208	0.292	12	12	12	12	0.583

Table 2: Comparison of inter-rater agreement statistics.

Condition	**SI statistics**	**Cohen's Kappa**	**Gwet's AC1**	**PABAK**	**S-statistics**
1	0.41	-0.01	0.29	0.17	0.17
2	0.41	0.06	0.25	0.17	0.17
3	0.41	0.11	0.22	0.17	0.17
4	0.41	-0.26	0.38	0.17	0.17
5	0.41	-0.26	0.38	0.17	0.17
6	0.41	-0.11	0.33	0.17	0.17
7	0.41	-0.11	0.33	0.17	0.17
8	0.41	0.17	0.17	0.17	0.17
Condition	**Scott's π**	**Van Eerdewegh's V**	**Yule's Y**	**Cicchetti & Feinstein's p_{pos}**	**Cicchetti & Feinstein's p_{neg}**
1	-0.01	-0.01	-0.01	0.29	0.71
2	0.06	0.07	0.07	0.38	0.69
3	0.11	0.11	0.12	0.67	0.44
4	-0.26	-0.51	-1.00	0.74	0.00
5	-0.26	-0.51	-1.00	0.00	0.74
6	-0.11	-0.13	-0.16	0.17	0.72
7	-0.11	-0.13	-0.16	0.72	0.17
8	0.17	0.17	0.17	0.58	0.58

In **Tables 4a** and **4b**, raters had common sensitivity of 0.8 and specificity of 0.9. Results showed that SI-statistics provided efficient results than other statistics. Additionally, it was observed that if an observed agreement had linear decreasing trend, so SI-statistics also showed liner decreasing trend. However, Kappa, AC1 and other statistics showed random variation except for PABAK and S-statistics.

In **Tables 5a** and **5b**, raters had different sFensitivity and specificity. Rater A had a sensitivity of 0.8 and specificity of 0.9, whereas Rater B had a sensitivity of 0.85 and specificity of 0.7. SI-Statistics was found to be the robust estimator amongst other statistics.

Sensitivity to equal cell distribution or 50% observed agreement

From **Tables 5a** and **5b** it follows that when cells distribution is equal and observed agreement is fixed to 0.5 then AC1, PABAK, S-statistics, Scott's π reports no agreement between the two raters. For equal observed agreement and disagreement (condition 1), all the agreement statistics estimated no agreement, whereas SI-statistics reported a fair agreement. Similarly, for condition 2 to 7 where the observed agreement was 0.5 but observed disagreement varies, SI statistics reported the moderate agreement (0.5), Kappa estimated slight agreement (0.2), Van Eerdewegh's V showed moderate agreement (0.58), however, Yule's Y estimated perfect agreement between the two raters.

For nxn contingency tables

Similar experiments were run for 3x3 (**Appendix A**) and 4x4 (**Appendix B**) contingency tables and results showed that SI-statistics provided better results.

Sensitivity to missing values

SI-statistics is less sensitive to missing values and provides

doi: 10.7243/2053-7662-8-2

Table 3a: Data simulation with common sensitivity of 0.9, common specificity of 0.9 and fixed observed agreement.

Condition	n	Probability of individual cell				Observed agreement
		a	b	c	d	P_0
1	24	0.01	0.09	0.09	0.81	0.82
2	24	0.018	0.09	0.09	0.802	0.82
3	24	0.05	0.09	0.09	0.77	0.82
4	24	0.09	0.09	0.09	0.73	0.82
5	24	0.17	0.09	0.09	0.65	0.82
6	24	0.25	0.09	0.09	0.57	0.82
7	24	0.33	0.09	0.09	0.49	0.82
8	24	0.41	0.09	0.09	0.41	0.82
9	24	0.49	0.09	0.09	0.33	0.82
8	24	0.57	0.09	0.09	0.25	0.82
9	24	0.65	0.09	0.09	0.17	0.82
8	24	0.73	0.09	0.09	0.09	0.82
9	24	0.81	0.09	0.09	0.01	0.82

Table 3b: Comparison of agreement statistics.

Condition	trait prevalence	SI statistics	Cohen's Kappa	Gwet's AC1	PABAK	S-statistics
1	0	0.69	0.00	0.78	0.64	0.64
2	0.01	0.69	0.07	0.78	0.64	0.64
3	0.05	0.69	0.25	0.76	0.64	0.64
4	0.1	0.69	0.39	0.74	0.64	0.64
5	0.2	0.69	0.53	0.71	0.64	0.64
6	0.3	0.69	0.60	0.67	0.64	0.64
7	0.4	0.69	0.63	0.65	0.64	0.64
8	0.5	0.69	0.64	0.64	0.64	0.64
9	0.6	0.69	0.63	0.65	0.64	0.64
8	0.7	0.69	0.60	0.67	0.64	0.64
9	0.8	0.69	0.53	0.71	0.64	0.64
8	0.9	0.69	0.39	0.74	0.64	0.64
9	1	0.69	0.00	0.78	0.64	0.64
Condition	**trait prevalence**	**Scott's π**	**Van Eerdewegh's V**	**Yule's Y**	**Cicchetti & Feinstein's p_{pos}**	**Cicchetti & Feinstein's p_{neg}**
1	0	0.00	0.00	0.00	0.10	0.90
2	0.01	0.07	0.10	0.14	0.17	0.90
3	0.05	0.25	0.31	0.37	0.36	0.90
4	0.1	0.39	0.43	0.48	0.50	0.89
5	0.2	0.53	0.55	0.57	0.65	0.88
6	0.3	0.60	0.61	0.61	0.74	0.86
7	0.4	0.63	0.63	0.63	0.79	0.84
8	0.5	0.64	0.64	0.64	0.82	0.82
9	0.6	0.63	0.63	0.63	0.84	0.79
8	0.7	0.60	0.61	0.61	0.86	0.74
9	0.8	0.53	0.55	0.57	0.88	0.65
8	0.9	0.39	0.43	0.48	0.89	0.50
9	1	0.00	0.00	0.00	0.90	0.10

doi: 10.7243/2053-7662-8-2

Table 4a: Data simulation with common sensitivity of 0.8 and common specificity of 0.9.

Condition	n	Probability of individual cell				Observed agreement
		a	b	c	d	P_0
1	24	0.01	0.09	0.09	0.81	0.82
2	24	0.02	0.09	0.09	0.80	0.82
3	24	0.04	0.09	0.09	0.77	0.81
4	24	0.07	0.10	0.10	0.73	0.81
5	24	0.14	0.10	0.10	0.66	0.79
6	24	0.20	0.11	0.11	0.58	0.78
7	24	0.26	0.12	0.12	0.50	0.76
8	24	0.33	0.13	0.13	0.43	0.75
9	24	0.39	0.13	0.13	0.35	0.74
8	24	0.45	0.14	0.14	0.27	0.72
9	24	0.51	0.15	0.15	0.19	0.71
8	24	0.58	0.15	0.15	0.12	0.69
9	24	0.64	0.16	0.16	0.04	0.68

Table 4b: Comparison of agreement statistics.

Condition	trait prevalence	SI statistics	Cohen's Kappa	Gwet's AC1	PABAK	S-statistics
1	0	0.69	0.00	0.78	0.64	0.64
2	0.01	0.69	0.05	0.78	0.64	0.64
3	0.05	0.68	0.20	0.76	0.63	0.63
4	0.1	0.68	0.31	0.73	0.61	0.61
5	0.2	0.66	0.43	0.67	0.58	0.58
6	0.3	0.64	0.48	0.61	0.56	0.56
7	0.4	0.62	0.50	0.55	0.53	0.53
8	0.5	0.60	0.49	0.50	0.50	0.50
9	0.6	0.58	0.47	0.47	0.47	0.47
8	0.7	0.56	0.43	0.46	0.44	0.44
9	0.8	0.55	0.35	0.47	0.42	0.42
8	0.9	0.53	0.22	0.49	0.39	0.39
9	1	0.52	0.00	0.53	0.36	0.36
Condition	**trait prevalence**	**Scott's π**	**Van Eerdewegh's V**	**Yule's Y**	**Cicchetti & Feinstein's p_{pos}**	**Cicchetti & Feinstein's p_{neg}**
1	0	0.00	0.00	0.00	0.10	0.90
2	0.01	0.05	0.08	0.12	0.15	0.90
3	0.05	0.20	0.25	0.31	0.31	0.89
4	0.1	0.31	0.36	0.41	0.43	0.88
5	0.2	0.43	0.46	0.48	0.57	0.86
6	0.3	0.48	0.49	0.51	0.64	0.84
7	0.4	0.50	0.50	0.51	0.69	0.81
8	0.5	0.49	0.50	0.50	0.72	0.77
9	0.6	0.47	0.47	0.47	0.75	0.73
8	0.7	0.43	0.43	0.43	0.76	0.66
9	0.8	0.35	0.36	0.37	0.78	0.57
8	0.9	0.22	0.24	0.26	0.79	0.43
9	1	0.00	0.00	0.00	0.80	0.20

Table 5a: Data simulation with different sensitivity and specificity.

Condition	n	Probability of individual cell				Observed agreement
		a	b	c	d	P_0
1	24	0.25	0.25	0.25	0.25	0.5
2	24	0.25	0.50	0.00	0.25	0.5
3	24	0.25	0.00	0.50	0.25	0.5
4	24	0.25	0.35	0.15	0.25	0.5
5	24	0.25	0.15	0.35	0.25	0.5
6	24	0.25	0.40	0.10	0.25	0.5
7	24	0.25	0.10	0.40	0.25	0.5

Table 5b: Comparison of agreement statistics.

Condition	SI statistics	Cohen's Kappa	Gwet's AC1	PABAK	S-statistics
1	0.33	0.00	0.00	0.00	0.00
2	0.50	0.20	0.00	0.00	0.00
3	0.50	0.20	0.00	0.00	0.00
4	0.50	0.20	0.00	0.00	0.00
5	0.50	0.20	0.00	0.00	0.00
6	0.50	0.20	0.00	0.00	0.00
7	0.50	0.20	0.00	0.00	0.00
Condition	**Scott's π**	**Van Eerdewegh's V**	**Yule's Y**	**Cicchetti & Feinstein's p_{pos}**	**Cicchetti & Feinstein's p_{neg}**
1	0.00	0.00	0.00	0.50	0.50
2	0.00	0.58	1.00	0.50	0.50
3	0.00	0.58	1.00	0.50	0.50
4	0.00	0.58	1.00	0.50	0.50
5	0.00	0.58	1.00	0.50	0.50
6	0.00	0.58	1.00	0.50	0.50
7	0.00	0.58	1.00	0.50	0.50

better estimates (**Appendix C**).

Concluding remarks

In literature, Gwet's AC1 statistic has been reported to be the most robust and less biased agreement statistics. In this article, it has been shown that AC1 statistic is sensitive to

1. Change in individual cell probabilities for fixed observed agreement
2. Trait prevalence
3. Equal cell distribution
4. Estimates zero agreement or no agreement between the raters if an observed agreement is 50%
5. SI statistic has been shown to be more stable and provide better results than other agreement statistics.
6. SI-Statistics can handle missing values and provides better results
7. Furthermore, SI-statistics ranges from 0 to 1 only. It does not have negative value inferring disagreement between the raters. So, SI-statistics only reports agreement.

Additional files

Appendix A
Appendix B
Appendix C

Competing interests

The author declares that she has no competing interests.

Publication history

EIC: Jimmy Efird, East Carolina University, USA.
Received: 10-Mar-2020 Final Revised: 19-May-2020
Accepted: 21-May-2020 Published: 04-Jun-2020

References

1. Miot H.A. **Agreement analysis in clinical and experimental trials**. *J Vasc Bras*. 2016; **15**:89-92. | Article
2. Yule G.U. **On the methods of measuring association between two attributes**. *Journal of the Royal Statistical Society*. 1912; **75**:579-652.
3. Bennett E, R. Alpert and A. Goldstein. **Communications through limited-response questioning**. *Public Opinion Quarterly*. 1954; **18**303-308. | Article
4. Scott W. **Reliability of content analysis: The case of nominal scale coding**. *Public opinion quarterly*. 1955.

5. J C. ***A coefficient of agreement for nominal scales***. *Educ Psychol Meas*. 1960; **20**:37-46. | Article
6. Fleiss J. **Measuring nominal scale agreement among many raters**. *Psychological bulletin*. 1971; **76**:378. | Article
7. Bangdiwala S. **A graphical test for observer agreement**. *in ISI Statistical Meeting*. 1985.
8. Spitznagel EL and Helzer JE. **A proposed solution to the base rate problem in the kappa statistic**. *Arch Gen Psychiatry*. 1985; **42**:725-8. | Article | PubMed
9. Byrt T, Bishop J and Carlin JB. **Bias, prevalence and kappa**. *J Clin Epidemiol*. 1993; **46**:423-9. | Article | PubMed
10. Gwet K. ***Inter-rater reliability: dependency on trait prevalence and marginal homogeneity***. *Statistical Methods for Inter-Rater Reliability Assessment Series*. 2002; **2**:1-9.
11. Wongpakaran N et al. **A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples**. *BMC medical research methodology*. 2013; **13**:61. | Website
12. Xie Q. **Agree or Disagree? A Demonstration of An Alternative Statistic to Cohen's Kappa for Measuring the Extent and Reliability of Agreement between Observers**.
13. Shankar V and Bangdiwala SI. **Observer agreement paradoxes in 2x2 tables: comparison of agreement measures**. *BMC Med Res Methodol*. 2014; **14**:100. | Article | PubMed Abstract | PubMed FullText
14. Viera AJ and Garrett JM. **Understanding interobserver agreement: the kappa statistic**. *Fam Med*. 2005; **37**:360-3. | PubMed
15. McHugh ML. **Interrater reliability: the kappa statistic**. *Biochem Med (Zagreb)*. 2012; **22**:276-82. | Article | PubMed Abstract | PubMed FullText

Citation:
Iftikhar S. **Modification in inter-rater agreement statistics-a new approach**. *J Med Stat Inform*. 2020; **8**:2. http://dx.doi.org/10.7243/2053-7662-8-2