African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

MEASURING ASSOCIATION BETWEEN SCREEN TEST RESULTS USING KAPPA K STATISTIC

Oyeka I.C.A[1], Nwankwo Chike H[1*], Onyiaoha C.A. Adaora[2] and Efobi Chilota C.N[2]

[1]Department of Statistics, Nnamdi Azikiwe University, Awka. Nigeria.

[2]Nnamdi Azikiwe University Teaching Hospital, Nnewi. Nigeria.

ABSTRACT: *We have in this paper developed and presented a more generalized and formatted method of estimating kappa k values often used for analyzing data obtained in comparative clinical trials and diagnostic screening tests of subjects for some condition such as disease in a population. The proposed method is generalized in the sense that it can be used to analyze data obtained in comparative, prospective studies and comparative retrospective studies with either two or more possible response classifications, categories or outcomes. A rule of thumb for determining the level or degree of concordance or agreement between test results as suggested or indicated by the estimated kappa k statistic is presented and used in the illustrative example. Test statistics are also presented for testing the significance of any desired hypothesized values of kappa k. Sensitivity and specificity dependent measure of association and odds ratios which are equivalently used as measures of association when there are only two possible response options or outcomes in comparative clinical trials and diagnostic screening tests are presented and discussed. These results are compared with the results obtained with kappa k statistics when there are only two response options or outcomes which are equally applicable in comparative studies with only two possible responses or classes. The results are shown to be quite similar leading to essentially the same conclusions.*

KEYWORDS: Association, Concordance, Diagnostic, Statistic, Generalised, Sensitivity

INTRODUCTION

Sometimes the interest of a research scientist or clinician may be in assessing the extent or degree to which his or her findings or test results are consistent with some existing knowledge or standard result in some research problem of interest. For example, in comparative clinical trials or diagnostic screening test of subjects for some condition or disease in a population, a clinician or diagnostician may need to compare his or her test results with other test results on the same population to know the level to which the two test results are in concordance or in agreement with each other. This would usually need the use of some statistical methods.

If each of the tests have many possible outcomes, categories or options, then the statistical tests often used include the extended sign test and the Friedman two-way analysis of variance test by ranks (Dawson and Trapp, 2004; Oyeka and Okeh, 2013). If each of the tests has only two possible responses, outcomes, categories or options then one can use either Odds ratios, Relative Risk, or Sensitivity and Specificity of the screening test or their modifications or the kappa k statistics (Fleiss 1973; Dawson and Trapp, 2004).

However, the kappa k statistic as presented by the authors cited is not clearly formatted and well-structured to make it easily useable by researchers. The purpose of this paper is therefore to formulate, present and discuss the kappa k statistic in a more generalized and formatted form for easier application in comparative clinical trials or diagnostic screening tests whether there are two or more than two possible response options or outcomes. A test statistic is also developed and presented for use in testing any desired hypothesis.

The Proposed Method

Now a statistic often used to measure agreement or concordance between two observers or clinicians on a condition with several possible outcomes, is kappa (k) statistic, defined as the agreement beyond chance divided by the amount of possible agreement beyond chance (Dawson and Trapp, 2004).

To assess the extent to which a given rater or clinician's characterization of a subject is reliable, one must have a number of subjects rated or assessed by more than one rater or clinician operating independent of each other and testing each subject or patient in a random sample or subject or patients for their response, to a condition or disease with several possible outcomes in a population of interest.

Now, to do this, suppose we have two raters or clinicians, clinician one and clinician two, operating independently. Clinician one may be a research scientist, while clinician two may also be a research scientist, or clinician two test results may be considered as the gold standard, state of nature or standard condition in a study population.

Furthermore suppose each of the two clinicians screens and tests a random sample of n = n.. subjects or patients or a certain condition or disease in a population, where the condition of interest has r possible outcomes or response options.

Let n_{ij} be the number of subjects in the random sample who are found or reported to be in condition i when screened and tested by clinician one, here referred to as clinician A, but are found or reported to be in condition j when screened and tested by clinician two, here referred to as clinician B; for $i, j = 1,2, \dots, r$ where $\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij} = n..$

These results are presented in a sample data format as in table 1 to help easily illustrate the estimation of the sample kappa k statistic of the population kappa value k.

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

Table 1: Sample Data Format for Estimating k Statistic
Clinician B (State of Nature, Standard Condition)

Test Result (*i*) Clinician A (Test Result (i))	**1**	**2**	**3**	**. . .**	**r**	**Total ($n_i.$)**	**Marginal Probability ($\frac{n_{i.}}{n_{..}}$)**
1	n_{11}	n_{12}	n_{13}	…	n_{1r}	$n_1.$	$p_1.$
2	n_{21}	n_{22}	n_{23}	…	n_{2r}	$n_2.$	$p_2.$
3	n_{31}	n_{32}	n_{33}	…	n_{3r}	$n_3.$	$p_3.$
.	.	.	.	…	.	.	.
.	.	.	.	…	.	.	.
R	n_{r1}	n_{r2}	n_{r3}	…	n_{rr}	$n_r.$	$p_r.$
Total (n.j)	$n._1$	$n._2$	$n._3$	…	$n._r$	$n..$ $(= n)$	_
Marginal prob $(n._j = \frac{n._i}{n_{..}})$	$p._1$	$p._2$	$p._3$	…	$p._r$	_	1.00

As noted above, table 1 is the data format showing frequency distribution of a random sample of $n.. = n$ subjects, individuals or patients by each subject's state nature or standard condition B and by the condition diagnosed for the subject or patient by clinician or diagnostician A

Thus in table 1, the cell frequency n_{ij} is the number of subjects or patients who are found by clinician or diagnostician A to be or to have condition i and found by clinician or diagnostician B to be, or whose actual condition or state of nature is in condition $j, for\ i\ j = 1,2,3,\ldots r;\ \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij} = n$ subjects. The notations of table 1 will be used to formulate or to obtain an expression for kappa k statistics. Reliability of diagnosis or association between clinicians' diagnosed condition or test results and state of nature or standard condition of subject or patient is often measured in terms of the magnitude of the resulting chi-square statistic (Fleiss 1973). But while statistical significance of a measure of association between observations or variables can be properly assessed using the chi-square statistic, this statistic cannot however be correctly used to assess the degree or strength of the same measure between the variables. What makes chi-square statistic test inadequate as a measure of the degree or level of association is that it measures only significance of association of any kind and not specifically agreement or concordance between any two raters, clinicians or between any test results and states of nature or actual condition in a population.

The first step in the calculation of chi-square test statistic is the determination of the frequencies expected in each of the cells under the null hypothesis of no association between the variables, such as the test results of clinicians A and B of table 1. Chi-square tells one nothing at all about the degree or strength of association that is between test results by clinicians or between a clinician's test result and actual state of nature or condition in a population.

In order to measure agreement or concordance between clinicians' or diagnostician's test results in a comparative clinical trial or diagnostic screening test one should focus attention

on the cells, that is, scores along the main diagonal of table 1 say, representing agreement or concordance between two raters or diagnosticians namely diagnosticians A and B or between clinician A's test result and actual state of nature or standard condition of subjects or patients in a population.

Each of the cell frequencies along the cells of the main diagonal of table1 say, is the number of subjects or patients in that cell in which there is an agreement or concordance between clinicians or diagnosticians A and B, that is between screening test results of clinicians or diagnosticians and actual states of nature or standard condition of subjects or patients in the population. For example, the cell frequency n_{ii} is the number of subjects or patients in the $(i,i)th$ cell in which clinicians or diagnosticians A and B are in agreement or in concordance that subjects or patients are in condition i in the population for some $i = 1,2,\ldots,r$.

In other words then n_{ii}'s, the cell frequencies along the main diagonal of table 1 say, are precisely equal to the number of subjects or patients on whom A and B are in agreement or in concordance that the subjects or patients are in condition i which agreement is to be expected to occur solely on the basis of chance or at random for all $i = 1,2,\ldots,r$. In these situations one would conclude that the degree, level, or strength of agreement or concordance between clinicians or diagnosticians A and B are no better than that expected or predicted by chance.

Hence, reliance on the significance of chi-square statistic would have led to the totally erroneous conclusion that the assessment or test results or diagnoses by clinician or diagnostician A are consistent with those of diagnostician B and are therefore reliable.

Just as chi-square statistic is inadequate for measuring reliability or agreement between screening test results so is the overall proportion of subjects, individuals or patients on whom there is agreement or concordance between clinicians or diagnosticians A and B in their screening test results (Cohen, 1960; Armitage, Blendis and Smyllie, 1966; Rogot and GoldBerg, 1966; Fleiss, 1973).

Now let the sum of the numbers n_{ii} of subjects or patients along the main diagonal of table1 say, that is the total number of subjects or patients in the (i,i)th cell of table 1 say, in which clinicians or diagnosticians A and B are in agreement by chance that subjects or patients are in condition i, that is in which diagnosed conditions or test results of subjects or patients are the same for clinicians A and B or diagnostician A and results of actual state of nature or standard condition in a population for all $i = 1,2,\ldots,r$, be represented by

$$n_c = \sum_{i=1}^{r} n_{ii} \qquad 1$$

Then the proportion of cases or test results in which there is agreement or concordance between clinicians or diagnosticians A and B test results or in which there is agreement or consistency between clinician A's test results and those of clinician B's or state of nature or standard condition, that is , in which diagnosis of subject's conditions in a population by diagnosticians A and B are the same is obtained as

$$p_c = \frac{n_c}{n_{..}} = \sum_{i=1}^{r} \frac{n_{ii}}{n_{..}} \qquad 2$$

The proportion p_c of agreement or concordance by chance by clinicians A and B of equation 2 is often not adequate for measuring agreement or concordance between clinicians or

diagnosticians A and B in their assessment of the consistence of these clinicians test results or of clinician A's test result and actual state of nature or standard condition of subjects or patients in a population.

Some degree of agreement or concordance between clinicians A and B in their assessment of subject's condition is to be expected purely on the basis of chance measured in terms of or as a function of the products of the marginal frequencies of table 1.

Now to find the number of cases or test results in which clinicians or diagnosticians A and B would agree by chance on their test results, use is made of a straight forward application of the rules of independence of events in statistical theory (Stirzaker, 1996). To do this there is a need to assume that clinicians A and B perform their screening tests independent of each other. In which case, one can then use the rule of independence of events and the multiplication rule of probability for two independent events to see how likely it is that clinicians A and B agree or concur in their findings merely by chance.

In fact, the total number of responses or test results expected by chance alone, which is a function of the off diagonal marginal frequencies of table 1, say, is obtained by straight forward application of the multiplication rule of independence of events as

$$n_o = \sum_{i=1}^{r} n_i . n_{.i} \qquad 3$$

Hence, the overall proportion of agreement or concordance between clinicians A and B in their test results expected by chance alone is

$$p_o = \frac{n_o}{n^2_{..}} = \sum_{i=1}^{r} \frac{n_i . \, n_{.i}}{n^2_{..}} = \sum_{i=1}^{r} p_i . p_{.i} \qquad 4$$

A more preferred, more reliable and better measure of agreement or concordance between clinicians or diagnosticians A and B in their test results, which is a better measure of agreement than $p_c$alone (Fleiss 1973) is

$$p_a = p_c - p_o \qquad 5$$

The index p_a measures how much agreement or concordance that exists between clinicians or diagnosticians A and B in their test results or diagnosis of a condition in a population beyond the amount expected by chance alone.

Note that as proportions, $0 \leq p_o, p_c \leq 1$. Also, strictly speaking, the index p_a as specified by equation 5 has no lower or upper bounds, it therefore needs to be normal or scaled to between -1 and +1 inclusively by dividing it with $1-p_o$, to ensure that as usual, 1 indicates perfect agreement and 0 and -1 indicates no agreement or concordance whatsoever between clinicians A and B in their diagnosis of a condition in a population.

The normal or scaled value of p_a of equation 5 is the sample estimate k of the well-known kappa k statistic (Dawson and Trapp 2004) defined as

$$k = \frac{p_a}{1-p_o} = \frac{p_c - p_o}{1-p_o} \qquad 6$$

If in equation 6, the index of similarity $p_c < p_o$, indicating even less than chance agreement, then k <0 indicate no agreement or concordance whatsoever between clinicians or

diagnosticians A and B in their test results. If $p_c = p_o$ so that $p_c - p_o = 0$, then k=0, indicating just chance agreement or no agreement whatsoever between clinicians or diagnosticians A and B in their test results. In each of these two cases knowing a subject's or patient's test result would be of no use whatsoever in predicting a subject's actual state of nature or condition in a population. If $p_c > p_o$, then k>0, indicating more than chance agreement. That is a high level of agreement or concordance between clinicians A and B in their test results. If the index of similarity p_c equal +1, then k=1, indicating perfect agreement or concordance between clinicians or diagnosticians A and B in their test results. In these two last cases, there is at least very good level of agreement, concordance or association between clinicians or diagnosticians A and B in their assessment of subjects or patient's condition in a population.

In other words, there is consistency between clinicians A and B test results or between clinician A's test result and actual state of nature or condition of subjects or patients in the population, in which case knowing a subject's test result would enable one make a reliable or accurate prediction of the subject's actual condition in the population. Byrt (1996) has proposed the following guidelines or rule of thumb for interpreting k and assessing the level or degree of similarity, agreement, association or concordance between two assessors A and B. In the similarity, agreement, concordance or association of their test results following the screening and testing of subjects or patients for some condition in a population:

S/N	Value of k	Conclusion
1	0.93≤k≤1.00	Excellent level or perfect agreement or concordance
2	0.81≤k≤0.92	Very good level of agreement or concordance
3	0.61≤k≤0.80	Good level of agreement or concordance
4	0.41≤k≤0.60	Fair level of agreement or concordance
5	0.21≤k≤0.40	Slight level of agreement or concordance
6	0.01≤k≤0.20	Poor level of agreement or concordance
7	k≤ 0.00	No agreement or concordance

Note from equation 7 that when k =0, agreement between assessors is only at the level expected by chance and one may conclude that the findings of clinicians or diagnosticians A and B are not in concordance or that there is no association or agreement between the test results by clinicians in their diagnosis of a condition in a population. When k is negative, that is when k < 0, the observed level of agreement is less than we would expect by chance alone and one may still conclude that there is no agreement or concordance between clinicians A and B in their test results.

This rule of thumb is however subjective. It is always better and more instructive to draw conclusions statistically using some test statistics to help one reject or accept any hypothesis regarding level of agreement or concordance between clinicians or diagnosticians on a condition of subjects or patients in a population.

Hence notwithstanding the rule of thumb for deciding the level or degree of a reliance or agreement suggested by equation 7 for k, a researcher may still, a priori, state and test the

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

assumed direction and magnitude expected or hypothesized for k as a measure of the degree of agreement or concordance or association between test results obtained in a study population of interest, stated as a hypothesis such as:

$$H_O : k \leq k_o \text{Versus } H_1 : k > k_o \ (-1 \leq k_o \leq 1) \text{ say} \qquad 8$$

If k_o= 0, the null hypothesis H_O of equation 8 would mean that there is no association or that there is no agreement or concordance between assessors, clinicians or diagnosticians in their diagnostic test results in a studied population against the complementary hypothesis that an association do in fact exist between test results in the population.

A sample estimate of the variance of kappa k statistic which is a function of the marginal frequencies and hence marginal proportion of table 1 can be shown to be (Everitt 1968; Fleiss, Cohen and Everitt 1969; Fleiss 1973)

$$\text{Var (k)} = \frac{1}{n_{..}(1-p_o)^2}\left(p_o + {p_o}^2 - \sum_{i=1}^{r} p_{i\cdot}\ p_{\cdot i}\ (p_{i\cdot}+p_{\cdot i})\right) \qquad 9$$

The null hypothesis H_O of equation 8 is tested using the chi-square test statistic

$$\chi^2 = \frac{(k-k_o)^2}{var\ (k)} = \frac{(((p_c-p_o) - k_o)/(1-p_o))^2}{(p_o + {p_o}^2 - \sum_{i=1}^{r} p_{i\cdot}\ p_{\cdot i}\ (p_{i\cdot}+p_{\cdot i})/n_{..}(1-p_o)^2)}$$

or

$$\chi^2 = \frac{n_{..}((p_c-p_o)-k_o)^2}{p_o(1-p_o) - \sum_{i=1}^{r} p_{i\cdot}\ p_{\cdot i}\ (p_{i\cdot}+p_{\cdot i})} \qquad 10$$

Which under H_O has approximately the chi-square distribution with

$(r-1)(r-1) = (r-1)^2$ degrees of freedom for sufficiently large sample size n.. = n

The null hypothesis H_O of equation 8 is rejected at the α level of significant if

$$\chi^2 \geq \chi^2_{1-\alpha;\,(r-1)^2} \qquad 11$$

Otherwise H_O is accepted.

Illustrative Example

Two clinicians A and B independently screened and tested a random sample of 150 male subjects aged 40 years and over to determine whether or not they have prostate cancer, one of the clinician B used histology test (which are considered as the gold standard), while the other clinician A used Magnetic Resonance Imaging (MRI) test.

The clinicians reported the prostate status of each subjects screened and tested as either positive, negative or non-definitive if the subjects test result is neither definitely positive nor definitely negative.

The screening test results are presented as Table 2

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

Table 2: Screening Test Results for Prostate Cancer Using Histology and MRI Tests
Histology Test Result (Gold Standard, B)

MRI Test Result(A)	Positive	Negative	Non - Definite	Total ($n_{i.}$)	Marginal Prob ($p_{i.}=\frac{n_{i.}}{n_{..}}$)
Positive	n_{11} (75)	n_{12} (3)	n_{13} (8)	$n_{1.}$ (84)	0.560
Negative	n_{21} (13)	n_{22} (29)	n_{23} (4)	$n_{2.}$ (46)	0.307
Non-Definite	n_{31} (7)	n_{32} (3)	n_{33} (8)	$n_{3.}$ (18)	0.120
Total ($n_{.j}$)	$n_{.1}$ (95)	$n_{.2}$ (35)	$n_{.3}$ (20)	150 ($n=n_{..}$)	___
Marginal Prob ($p_{.j}=\frac{n_{.j}}{n_{..}}$)	0.633($p_{.1}$)	0.23($p_{.2}$)	0.133($p_{.3}$)	___	1.00

The sample data of table 2 on prostate cancer test results are here used to illustrate the calculation of kappa k statistic.

Now from table 2 and equation 1, we have that the sum of the number of subjects n_{ii} along the main diagonal is

$$n_c = \sum_{i=1}^{3} n_{ii} = n_{11} + n_{22} + n_{33} = 75 + 29 + 8 = 112$$

Hence a sample estimate of the proportion of subjects in which there is agreement or concordance between clinicians A (MRI test) and B (Histology test) in their diagnostic test results is,a from equation 2:

$$p_c = \frac{n_c}{n..} = \frac{112}{150} = 0.747 \; or \; 74.7\%$$

Furthermore, the number of subject or test results expected by chance alone, which is a function of the off diagonal marginal frequencies of table 2, is from equation 3

$$n_o = \sum_{i=1}^{r} n_{i.} n_{.i} = n_{1.} n_{.1} + n_{2.} n_{.2} + n_{3.} n_{.3}$$

$$= (84)(95) + (46)(35) + (18)(20) = 7980 + 1610 + 360 = 9950$$

Hence the sample estimate of the proportion of agreement or concordance between clinicians A (MRI test) and B (Histology test) expected by chance alone is from equation 4

$$p_o = \frac{n_o}{n\ldots^2} = \frac{9950}{(150)^2} = 0.442 \; or \; 44.2\%$$

Therefore, from equation 6 we have that the sample kappa statistic k is:

$$k = \frac{p_c - p_o}{1 - p_o} = \frac{0.747 - 0.442}{1 - 0.442} = \frac{0.305}{0.558} = 0.547 \; or \; 54.7\%.$$

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

Now with an estimated kappa k value of k= 0.547, one may conclude, following the rule of thumb of equation 7 that the degree of association, agreement or concordance between the results of prostate cancer screening tests of subject obtained using histology test and MRI test independently is just fair or only moderate. Thus the level, degree or strength of association between MRI test results and histology test results, here considered as the gold standard, is only slightly, but nevertheless, highly statistically significant, as we would now proceed to test the null hypothesis H_o of equation 8 with $k_o = 0$.

To do this, we have from equation 9, that the variance of k is estimated as:

$$var(k) = \frac{1}{n_{..}(1-p_0)^2}(p_o + p_o^2 - (p_1.p_{.1}\,(p_1. + p_{.1}) + p_2.p_{.2}\,(p_2. + p_{.2}) + p_3.p_{.3}(p_3. + p_{.3}))$$

$$= \frac{1}{150(1-0.442)^2}\ (0.442 + 0.195 - ((0.560)(0.633)(0.560 + 0.633) + (0.307)(0.233)(0.307 + 0.233) + (0.120)(0.133)(0.120 + 0.133))$$

$$= \frac{0.172}{46.65} = 0.004$$

Using these results in equation 10, we have that the test statistic for testing the null hypothesis H_o of no association on $(H_o: k = 0)$ between the two diagnostic tests namely MRI (A) and Histology (B) in the results obtained using MRI and histology tests we have:

$$\chi^2 = \frac{k^2}{var(k)} = \frac{(0.547)^2}{0.004} = \frac{0.229}{0.004} = 74.80\ (p - value = 0.000)$$

which, with $(3-1)^2 = 4$, degrees of freedom, is statistically significant.

If the researcher's interest had been in only subjects whose responses in the screening test were specifically definitive, that is on only subjects who either test positive or test negative to prostate cancer test using MRI and Histology tests, we would obtain and use only the following sample data (table 3)

Table 3: Prostate Cancer Screening Test Results for Subjects Testing either Positive or Negative Using MRI Test and Histology (Gold Standard)

Histology test results (B)

MRI test result (A)	**Positive**	**Negative**	**Total($n_i.$)**	**Marginal Prob.($p_{i.} = \frac{n_{i.}}{n_{..}}$)**
Positive	75(n_{11})	3(n_{12})	78($n_1.$)	0.650($p_1.$)
Negative	13(n_{21})	29(n_{22})	42($n_2.$)	0.350($p_2.$)
Total ($n_{.j}$)	88($n_{.1}$)	32($n_{.2}$)	120($n.. = n$)	----------
Marg.Prob.($p_{.j} = \frac{n_{.j}}{n_{..}}$)	0.733($p_{.1}$)	0.267($p_{.2}$)	----------	1.00

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

Now from table 3 and equation 1, we have that

$$n_c = n_{11} + n_{22} = 75 + 29 = 104, hence\ \ p_c = \frac{n_c}{n_{..}} = \frac{104}{120} = 0.867\ 0r\ 86.7\%$$

Similarly, from equation 3, we have that

$$n_o = n_1.n_{.1} + n_2.n_{.2} = (78)(88) + (42)(32) = 6864 + 1344 = 8208$$

Therefore, the corresponding value of p_o from equation 4 is

$$p_o = \frac{n_o}{n_{...}^2} = \frac{8208}{(120)^2} = \frac{8208}{14400} = 0.570\ or\ 57.0\%$$

Hence the resulting sample kappa k statistic is from equation 6 is

$$k = \frac{p_c - p_o}{1 - p_o} = \frac{0.867 - 0.570}{1 - 0.570} = \frac{0.297}{0.430} = 0.691\ or\ 69.1\%$$

Which from equation 7 indicates that the association, agreement or concordance is a good one that is; there is a good level of agreement or concordance between prostate screening test results obtained using histology test (the good standard) and MRI test.

The variance of k from equation 9 is

$$\text{Var(k)} = \frac{1}{n_{..}\ (1-p_o)^2}(p_o + p_o^2 - ((p_1.p_{.1}(p_1. + p_{.1}) + p_2.\ p_{.2}\ (p_2. + p_{.2}))$$

$$= \frac{1}{120(1-0.570)^2}\ (0.570 + 0.325) - ((0.65)(0.733)(0.65+0.733)+(0.35)(0.267)(0.35 + 0.267))$$

$$= \frac{0.18}{22.2} = 0.008$$

The test statistic for the null hypothesis of no association $(H_o: k_o = 0)$ between MRI and Histology test results of prostate cancer screening in the population from equation 10 is

$$\chi^2 = \frac{k^2}{var(k)} = \frac{(0.547)^2}{0.008} = \frac{0.299}{0.008} = 37.375\ (p - value = 0.000)$$

which with 1 degree of freedom is statistically significant, leading to the conclusion that a significant association exist between the two methods of screening ad testing subjects for prostate cancer.

An equivalent measure of association or concordance when there are only two (r =2) possible response categories or options is either the odds ratio or sensitivity (s_e) and specificity (s_p) of the screening test.

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 3, 2020 (pp. 9-20)

www.abjournals.org

For example, the sensitivity (s_e) and specificity (s_p) of the screening test are respectively

$$S_e = \frac{n_{11}}{n_{\cdot 1}} ; S_p = \frac{n_{22}}{n_{\cdot 2}} \quad 12$$

For the prostate cancer data of table 3, we have that

$$S_e = \frac{75}{88} = 0.852 \qquad and \qquad S_p = \frac{29}{32} = 0.906$$

Sensitivity and specificity type measure of association in diagnostic screening test is

$$S = \frac{S_e \cdot S_p}{(1-S_e)(1-S_p)} \quad 13$$

For the sample data of table 3 on prostate cancer we have that

$$S = \frac{(0.852)(0.906)}{(1-0.852)(1-0.906)} = \frac{(0.852)(0.906)}{(0.148)(0.094)} = \frac{0.472}{0.014} = 55.143$$

Which is also the sample odds ratio, a value that is seen to be relatively high again indicating a high level of agreement or concordance between MRI and histology test results for prostate cancer as already shown above by kappa k statistic.

The test statistic for the statistical significance of sensitivity and specificity type measure of association and also for the odds ratio is

$$\chi^2 = \frac{n_{\cdot\cdot}(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{\cdot 1}n_{2\cdot}n_{\cdot 2}} \quad 14$$

Which, under the null hypothesis, H_o of no association or agreement, has approximately, the chi-square distribution with one degree of freedom for sufficiently large sample size$n_{\cdot\cdot} = n$. The null hypothesis H_o of no association is rejected at the α level of significance if equation 11 is satisfied with r = 2, otherwise, the null hypothesis H_o is accepted.

For the sample data of table 3 and from equation 14, we have that;

$$\chi^2 = \frac{120((75)(29)-(3)(13))^2}{(78)(88)(42)(32)} = 59.35 \text{ (p-value=0.00)}$$

This χ^2 value, for 1 degree of freedom, leads to the rejection of the null hypothesis (H_0: K_0 =0.00) of no concordance in the test results of the two clinicians, hence concluding significant concordance as was already found using Kappa K statistic in the population. One would therefore conclude that there is significant association between screening test results for prostate cancer using MRI test and the results obtained using Histology test, which is regarded as the gold standard. The results obtained from analysis of sample data in tables 2 and 3 on prostate cancer screening tests using Kappa K statistic and the Odds Ratios show that there is high level of agreement or concordance between results from these two tests , thereby suggesting that knowledge of a subjects screening test results using MRI test would enable a fairly accurate prediction on the basis of Histology test, the gold standard, on whether or not the subject is prostate cancer positive in the population of research interest.

SUMMARY AND CONCLUSION

We have in this paper developed and presented a more generalized and formatted method of estimating Kappa K values, often used for methods that could be equivalently used for the same purpose and were shown to be quite similar and led to similar conclusions.

REFERENCES

Armitage P, Blendis L. M. and Smyllie H. C. (1966). "The Measurement of Observer Disagreement in the Recording of Signs" Journal of Royal Statistical Society. Series A.129, 98.

Byrt Ted (1996). "How good is that Agreement" Epidemiology. Sept 1996 Vol 7, Issue 5. P 561.

Cohen, J. (1960): "A Coefficient of Agreement for Nominal Scales" Educational and Psychological Measurement, Volume XX, No 1, 1960.

Dawson, B. and Trapp R. G. (2004): "Basic and Clinical Biostatistics (Fourth Edition)" Lange Medical Books. McGraw-Hill Medical Publishing Division, New York.

Everitt B.S. (1968). "Moments of the Statistics Kappa and Weighted Kappa" British Journal of Mathematical and Statistical Psychology. Vol 2, Issue 1. Pp 97 – 103.

Fleiss J. L, Cohen J. and Evertt B. S. (1969)." Large Sample Standard Errors of Kappa and Weighted Kappa" Psychological Bulletin. 72(5). 332 – 327.

Fleiss, J.L (1973): "Statistical Methods for Rates and Proportions". New York. Wiley.

Oyeka, I.C.A and Okeh, U.M. (2013). "Coefficient of Concordance in Diagnostic Screening Test. Scientific Reports. 640. doi. 4172.

Rogot E. and GoldBerg I.D. (1966). " A Proposed Index for Measuring Agreement in Test Retest Studies". Journal of Chronic Diseases, Vol 19, pp 991 – 1006.

Stirzaker D (1996): Elementary Prbability. Cmbridge University Press (Low Price Edition.