# STAT.Dilman.Kareem@MSc.2024.Non Parametric

*by* Dilman Kareem

**Submission date:** 05-Jan-2024 06:48PM (UTC+0200)
**Submission ID:** 2267053043
**File name:** T.Dilman.Kareem_MSc.2024.Non_Parametric_-_dilman._statistics.pdf (827.64K)
**Word count:** 5376
**Character count:** 29870

**Salahaddin University -Erbil**
**College of Administration and Economics**
**High Education: MSc**
**Subject: Non-Parametric**
**Semester: First**

# A Review Article about Cohen's Kappa Test

"Prepared for the subject of Non-Parametric"

**By**

**Dilman Kareem Khudhur**
**M.Sc. Student- Statistics Dep.**
**dilman.statistics@gmail.com**

**Supervised by**

**Asst. Prof. Dr. Nazeera Sedeek Kareem**

**Academic year**
**2023-2024**

# Cohen's Kappa Test


## Abstract:

A common descriptive quantitative statistic to describe the cross-classification of two nominal variables with the same categories is the kappa coefficient. When two raters are assessing the same item, Cohen's kappa is used as a measure for the reliability that the raters would agree on the basis of chance. Rater reliability is significant because it indicates how closely the study's data reflects the real characteristics of the variables under investigation. Although there are many ways of evaluating reliability, historically it was expressed as percent agreement, which was determined by dividing the total number of points by the number of agreement points.

Jacob Cohen questioned the use of percent agreement in 1960 because it did not take chance agreement into consideration. To be attentive of the likelihood that raters may genuinely make educated guesses on at least some factors owing to uncertainty, he established the Cohen's kappa. The variable to be assessed by the two rates in the Cohen's Kappa instance is a nominal variable. The kappa has a range of -1 to +1, much like most of correlation statistics. Although one of the most used statistical methods to assess dependability is the kappa, it is not without limitations. There are disagreements over acceptable kappa levels for health studies. Cohen's interpretation means that a score less than 0.41 would be acceptable, which is inappropriate for studies concerning health. Levels for both kappa and percent agreement that need to be required in healthcare research are proposed after comparing kappa and percent agreement.



**Keywords:** Cohen's kappa, Interrater reliability, agreement

## 1. Introduction

In numerous instances, collecting research data in the medical field requires a team effort. The subject of agreement or consistency among the persons gathering data emerges right away since human observers differ from one another. Thus, procedures to evaluate agreement between different data collectors should be integrated into carefully well-designed projects for research. In most study designs, the data collectors are trained, and the degree to which they record identical ratings for identical phenomena is measured. Perfect agreement is rare, and the degree of disagreement or inaccuracy brought into a research due to inconsistent data collection practices among the participants influences the degree of trust that may be placed in the study's findings. "Reliability" refers to how closely data collectors agree with one another **(Hsu, and Field, R., 2003).**

For a few decades currently, the medical, biological, and social sciences have been the primary fields in which the Kappa statistic has been applied. However, Cohen's Kappa has not drawn much attention as an accuracy statistic in the expert systems, machine learning, and data mining sectors **(Knop. and Borkowski, 2011).**

In 1960, Jacob Cohen developed Cohen's kappa, a statistical measure designed to present a precise measure of the reliability among two raters in determining the appropriate classification for a given unit of study. In addition to calculating the percentage of agreement between two raters, kappa also determines the extent to which agreement is the result of chance **(Knop. and Borkowski, 2011)**.

### 1.1.Definition of Cohen's Kappa Statistic:

The degree of agreement between categorical variables X and Y is measured by Cohen's kappa. Kappa can be used, for instance, to examine how well various raters are able to arrange subjects into various categories **(Więckowsk et al., 2022).**

Only in the following situations may this statistic be computed:

- Two raters rate one trial on each sample, or.
- Two trials are rated by one rater for every sample.

### 1.2.Uses of Cohen's kappa

After accounting for random agreements between the categories, the Kappa statistic can be applied to assess the degree of agreement between two sets of categorizations of a dataset. This statistic, which measures the agreement between the prediction model and a set of field-surveyed sample points, can be particularly beneficial for determining the accuracy of predictive models in terms of landscape ecology and animal habitat analysis. To account for random agreement within categories, the Kappa statistic uses the model's overall accuracy as well as the accuracy within each category, as measured by the field-surveyed sample points and the predictive model **(Delgado &Tibau 2019).**

### 1.3. Importance of measurement Cohen's Kappa reliability & validity

Cohen's Kappa coefficient can only tell you how reliably both raters are measuring the same thing. It does not tell you whether what the two raters are measuring is the right thing **(Sun, 2011).**

Processes that measure agreement amongst several Collectors of data are essential components of well-designed research projects. In most study designs, the Collectors of data are trained, and the degree they record identical ratings for identical phenomena is measured. Efficient agreement is rare, and degree of disagreement or inaccuracy is needed to the study because of inconsistent collection practices among the participants influences the degree of trust which may be placed in the study's findings. "Reliability" refers to how closely data collectors agree with one another **(Delgado &Tibau 2019).**

In most large studies, interrater reliability is a concern to some extent because the phenomena of interest may be experienced and interpreted differently by different collectors of data. Variables prone to interrater errors can be found readily in literature of clinical research and diagnosis, for example, while data collectors can use measurement equipment for color, size, and swelling, these variables such as the degree of redness, edema, and erosion in the afflicted area—are very subjective when applied to pressure ulcer research. Researchers studying head trauma measure the size of the patient's pupils and the extent to which they constrict in the presence of light. It

has been discovered that there are differences in how laboratory readers of Papanicolaou (Pap) smears for cancer of the cervical cavity interpret the cells on the slides **(Sun, 2011).**

Researchers should instruct data collectors to minimize variability in data interpretation and viewing, as well as data recording on data collecting tools, as this might be a source of inaccuracy. Lastly, it is required of researchers to assess the success of the training as well provide information on the level of agreement (interrater reliability) amongst their data collectors **(Chicco et al., 2021).**

It measures the degree of agreement among raters on the relative ratings assigned to subjects and functions to indicate the accuracy as well as precision of rating procedure **(Chicco et al., 2021).**

### 1.4.Calculation of Cohen's kappa

**The following equation is used for calculating Cohen's kappa:**

$$k = (p_o - p_e) / (1 - p_e)$$

as:

- **Po**: The raters' relative reported agreement
- **Pe**: Assumed probability of a chance agreement

It aims to take into consideration the potential that the raters may agree on some items by chance in steed of merely measuring the proportion of items that the raters agree on ((**Chicco et al., 2021**).

### 1.5.Interpretation of Cohen's Kappa:

Cohen's Kappa consistently falls between 0 and 1, where 1 denotes complete agreement and 0 denotes no agreement at all between the two raters.

The following table summarizes how to interpret different values for Cohen's Kappa (8):

| Cohen's Kappa | Interpretation |
|---|---|
| 0 | No agreement |
| 0.10 - 0.20 | Slight agreement |
| 0.21 - 0.40 | Fair agreement |
| 0.41 - 0.60 | Moderator agreement |
| 0.61 - 0.80 | Substantial agreement |
| 0.81 - 0.99 | Near perfect agreement |
| 1 | perfect agreement |

## Calculation Example:

Majority of statistical tools can do k calculations. With basic data sets (two raters and two objects), figuring out k by hand is not too difficult. You should definitely utilize software like SPSS for greater data sets (8).

For two raters to agree, apply the following formula. A formula variation must be used if there are more than two raters (8).

For two raters, the formula to determine Cohen's kappa is:

$$k = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_0}$$

as:

Po is the raters' relative observable agreement.

Pe is hypothetical chance agreement probability.

Example:

The data is derived from a medical examination in which 2 radiologists scored fifty photographs as requiring more research. Either Yes (for more study) or No (no additional study needed) was the response from the researchers (A and B).

- Both gave 20 photographs a yes rating.
- Both gave 15 photos a "No" rating.
- Rater A rated 25 photos as yes and 25 as no overall.
- Rater B rated thirty photographs as yes overall and twenty as no.

Compute Cohen's kappa for this data.

**First, calculate the po, or the observed proportionate agreement.**

- Both gave 20 photographs a yes rating.
- Both gave 15 photos a "No" rating.

  Po = number in agreement / total = (20 + 15) / 50 = 0.70.

**Second stage: Calculate the probability that the two raters will respond "yes" at random.**

- Rater A gave yes for 25 out of 50 photos, or 50%(0.5).
- Rater B gave 30 out of 50 photos yes, or 60% (0.6),

• There is a 0.5× 0.6 = 0.30 probability that the raters are both going to randomly answer "yes."

**Third stage: Determine the probability that the raters are both going to answer "no" at random.**

- Rater A replied no for 25 out of 50 photos or 50%(0.5).
- Rater B replied no for 20 out of 50 photos, or 40% (0.6).
- The probability that the raters will both randomly respond "no" is 0.5 × 0.4 = 0.20.

**The fourth step: Calculate the Pe. You may calculate the total the probability that the raters would agree at random by adding your responses from the second and third steps.**

Pe = 0.30 + 0.20 = 0.50.

**The fifth step is to Solve the following equation by entering your calculations:**

k = (Po – pe) / (1 – pe = (0.70 – 0.50) / (1 – 0.50) = 0.40.

Fair agreement can be determined by k = 0.40.

## 1.6.Limitation of Cohen's Kappa

However, under unknown conditions, minimal sums might or might not predict the degree of chance rater agreement. Therefore, it is debatable if the kappa statistic's drop in the estimate of agreement really reflects the degree of chance rater agreement **(Vanbelle., 2016).**

The primary drawback is that it fails to include the likelihood that raters made educated guesses about scores. Hence, it can exaggerate the actual degree of agreement among raters. Although the kappa was created to account for the potential of guessing, it may overly decrease the estimate of agreement due to its poorly supported assumptions about rater independence and other issues. Moreover, because it is not interpretable directly, low kappa values are frequently accepted by researchers in their interrater reliability investigations **(Cicchetti, et al., 2017).**

The sensitivity of Cohen's Kappa to the degree of agreement in the data is one of its drawbacks. Cohen's Kappa has bias and may not fairly represent the actual agreement amongst raters when there is an imbalance in the categories being assessed or when one category is highly prevalent. Another drawback of Cohen's Kappa is its presumption of rater independence, or the absence of peer influence on rates. In certain instances, raters could be swayed by one another's evaluations, which could result in exaggerated estimations of agreement ((**Cicchetti, et al., 2017 &Wang & Xia, 2019).**

## 2. Review of literatures

**Steinijans, et al., 1997 reported in their article about Interobserver agreement: Cohen's kappa coefficient does not necessarily reflect the percentage of patients with congruent classifications that** One can obtain somewhat different values depending on whether the % agreement is calculated using Cohen's kappa coefficient, which corrects for chance, or not. This brief communication shows that the percentage of patients with congruent classifications does not correlate with Cohen's kappa coefficient of agreement between two raters or two diagnostic procedures based on binary (yes/no) replies. As such, its usefulness in evaluating advances in interrater reliability as a result of better diagnostic techniques may be restricted. Although it is simpler to understand clinically, the proportion of patients with congruent classifications does not take into consideration the percentage of agreement that would be predicted by chance. Therefore, we advise presenting Cohen's kappa coefficient with 95% confidence limits in addition to the percentage of patients with congruent classifications (**Steinijans, et al., 1997**). **Hsu, L.M. and Field, R., 2003 published article about Interrater agreement measures** and they showed that Kappa has a few disadvantages. It can be large when raters who arbitrarily classify things (such as patients) to categories (diagnoses) have sharply divergent opinions on base rates. It can also be significantly more significant when raters have significantly varied opinions about base rates than when they concur completely. We contend that Cohen's kappa, which is devoid of these significant drawbacks, is typically superior to kappa, in contrast to the opinions expressed by some of the more recent opponents of the method. Additionally, two other kappa-type statistics (Aickin's, 1990, α; Scott's, 1955, Π) are contrasted with Cohen's kappa. In contrast to Scott's Π, Cohen's kappa is more conservative than Aickin's αβ and is simpler to

compute. It can provide valuable insights into interrater agreement when marginal heterogeneity is present (**Hsu and Field 2003**).

**Nixon et al., 2005 conducted a study about the inter-rater reliability of the pressure Trial pressure ulcer diagnosis**. Qualified ward-based nurses and expert CRNs documented outcome skin examinations twice a week and daily, respectively. Seven different body sites had their skin evaluated in pairs of evaluations. The study established the percentage of agreement among nurses in diagnosing pressure ulcers, as well as the calculation of the Kappa statistic and confidence intervals. It was also discovered what percentage of nurses agreed when it came to categorizing skin for all grades. There were 378 pairs of people who completed assessments: 16 pairs of people who assessed patients' together (107 site comparisons) and 362 pairs of people who assessed patients together between CRNs and ward-based nurses. Diagnosis for a pressure ulcer was agreed upon by CRN team leader & CRNs 100% of the time, with 'very excellent' agreement shown by the Kappa statistics. Out of all the grades, there were just two (1.9%) differences in how these nurses classified the skin. The degree of agreement between CRNs and ward-based nurses on the diagnosis of pressure ulcers ranged from 93.6% to 100% depending on the skin site, with 'excellent' and'very good' agreement being indicated by the Kappa statistics (**Nixon et al ., 2005**).

**Simon, P., 2006. Published article titled that Including omission mistakes in the calculation of Cohen's Kappa and an analysis of the coefficient's paradox features**. They showed that Cohen's Kappa has a wider variety of applications for sequential observation data field, as observer omission errors can often occur. It is demonstrated how the omission errors may be included in the Kappa coefficient computation without affecting the underlying statistic. The Kappa Coefficient with Omission Calculation (Kwoc) is the name given to the improved coefficient. Furthermore, the observer bias, the base-rate problem, and the so-called paradox properties of the Kappa coefficient are explored. It is demonstrated that these characteristics are desirable Kappa coefficient qualities in the case of observation data (**Simon, 2006**).

**A systematic review of inter-rater reliability for PU classification systems by Kottner et al. (2009)**. Cohen's kappa was reported as the inter-rater reliability measure, or enough information was provided to calculate Cohen's κ; Cohen's κ estimates' standard errors were reported, or enough information to estimate the standard errors was measured; Cohen's κ is a suitable inter-rater reliability measure for the rating process. Authors, research sites, year of being published , categorization system, number of categories, rating technique, rater characteristics, number of raters, The overall number of skin sites, features of skin sites, and percentage of overall agreement

(p 0), and Cohen's κ estimate are the variables that were taken out of the six studies(**Kottner et al., 2009**).

**Vieira et al., (2010) published article about Cohen's kappa coefficient as a performance measure for feature selection.** They said that it is not a simple or easy process to measure a given classifier's performance. Depending on the application, if one or more of the classes perform poorly in the forecast, the overall classification rate might not be adequate. The feature selection procedure also has this issue, particularly when a wrapper technique is applied. A statistical indicator of inter-rater agreement for qualitative items is Cohen's kappa coefficient. Since it considers the possibility that the agreement may occur by coincidence, it is often seen as a more reliable metric than a simple % agreement estimate. Since kappa is a more restrictive measurement, using it for selecting wrapper features is appropriate when assessing the models' performance. In this research, a feature selection wrapper strategy utilizing the kappa measure as an evaluation metric is proposed. The suggested method uses fuzzy criteria to construct the feature selection issue and fuzzy models to test the feature subsets. The findings indicate that the kappa measure produces classifiers that are more accurate, which in turn produces feature subset solutions with more pertinent features (**Vieira et al., 2010**).

**Kvålseth, 2015. Published the importance of Measurement of interobserver disagreement: Correction of Cohen's kappa for negative values.** According to their research, Cohen's kappa coefficients are commonly utilized and have clear and significant implications. However, there are no set lower boundaries for the kappa coefficients for negative coefficient values, indicating that their interpretations are meaningless and may even be completely deceptive when disagreement is seen with a greater probability than it would be by chance. It is possible to get the fixed lower bound of −1 for the new coefficients regardless of the marginal distributions. For nominal classification categories, a coefficient is developed, and for ordinal categories, a weighted coefficient is suggested. In addition to the overall disagreement coefficients across categories, disagreement coefficients for specific categories are shown. Numerical examples and methodologies for statistical inference are developed (**Kvålseth, 2015.**).

**Wang and Xia (2019) Published an article about Relationships of Cohen's Kappa, Sensitivity, and Specificity for Unbiased Annotations.** They reported that contrary to its intended purpose of evaluating inter-annotator consistency, it is typically used for a quality

metric for data annotation. Nevertheless, it is not possible to use the developed connection functions of Cohen's kappa, sensitivity, and specificity in the literature to deduce classification performance from kappa values due to their complexity. In this work, we establish basic correlations between kappa, sensitivity, and specificity in the absence of bias in the annotations, using an annotation generation model as our foundation. Further, a link is discovered between kappa and Youden's J statistic, a binary classification performance parameter. Linear regression analysis is used to evaluate the obtained associations on a synthetic dataset (**Wang and Xia 2019**).

**Recently, Pérez et al., (2020) performed Systematic reviews in software engineering—enhancement of the study selection process using Cohen's Kappa statistic to reduce the bias and time spent in the study selection process,** according to the use of this method, they established an iterative procedure whereby the criteria are improved until almost perfect agreement ($k>0.8$) is achieved. By now, there is less prejudice because both researchers are using the same interpretation of the selection criteria. Dual review can be dropped beginning with this agreement, significantly cutting down on time. A tertiary research in software engineering on works published between 2005 and 2018 demonstrates the viability of this iterative procedure for study selection. According to their analysis, 28% of time was saved during the study selection process (involving 152 studies). It was shown that if the number of studies is high enough, the time saved will eventually approach 50% (**Pérez et al., 2020**).

**Rau, and Shih, (2021) Conducted a research on Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data**. In category evaluation, they inquire as to what term is acceptable, what statistical techniques are reliable for measuring it, and how much the choice of units influences the results. They discovered that while dependability and agreement could both be important, only agreement could be quantified using nominal data. Furthermore, because κ requires the units to be preset, fixed, and independent, kappa is problematic for move or component analysis, even though it may be appropriate for macrostructure or corpus analysis. Furthermore, κ makes the potentially incorrect assumption that every dispute in category assignment has the same probability. They also discussed various metrics, such as percent agreement, chi square, and correlation, and they showed that, in many cases, percent agreement is the sole useful metric despite its drawbacks. Lastly, they gave an

example of how the value calculated is greatly impacted by the unit selection. These results also hold for other applied linguistics research that makes use of nominal data. They came to the conclusion that, similar to any other statistical testing; the technique must be made transparent in order to verify that the standards have been satisfied (**Rau, and Shih, (2021**).

**Figuero et al., 2023, July performed an article about Generalized Cohen's kappa: a novel inter-rater reliability metric for non-mutually exclusive categories.** They present the Generalized Cohen's kappa, a unique technique for calculating inter-rater agreement. They first show it functions similarly to the commonly used Cohen's kappa under the latter's preconditions, and they then show that it may be used successfully in scenarios where there are non-mutually exclusive categories. They discovered that GCK remained robust under a range of qualitative coding scenarios, including the usage of non-mutually exclusive categories by sizable coding teams working with sizable datasets. A few of Cohen's kappa's acknowledged problems are also present in GCK. Having high rates of Type, I mistake, which means the data reports higher scores than real agreements, and reporting various agreements in data with similar beginning agreement but varied distribution of the data are two instances (**Figuero et al., 2023**).

## Discussion and comparison:

According to the studies' variation, an adequate comparison could have established. It is unclear how the variety of categories, the quality and instructional materials of raters, the characteristics of the evaluated individuals, and the application techniques affect the level of interrater reliability. In order to compare at least two distinct categorization methods, well-designed interrater reliability researches are required. These studies should be conducted using a representative sample of raters and applied to similar patient or inhabitant samples Such research results must to be calculated using appropriate and appropriate statistical techniques. In this article, 12 papers were addressed and illustrated Cohen's kappa and different techniques for assessment. **Simon, P., 2006** showed that Cohen's Kappa has a wider variety of applications for sequential observation data fields, as observer omission errors can often occur. It is demonstrated how the omission errors may be included in the Kappa coefficient computation without affecting the underlying statistic. **However, Rau, and Shih, (2021)** inquired about what statistical techniques are reliable for

measuring it, and how much the choice of units influences the results. They discovered that while dependability and agreement could both be important, only agreement could be quantified using nominal data. Furthermore, because κ requires the units to be preset, fixed, and independent, kappa is problematic for move or component analysis, even though it may be appropriate for macrostructure or corpus analysis.

**Kottner et al. (2009)** study included 24 studies out of 339 potentially relevant research were considered. According to the research's variability, a useful comparison was not feasible. Cohen's kappa was reported as the inter-rater reliability measure or enough information was provided to calculate Cohen's κ; Cohen's κ estimates' standard errors were reported, or enough information to estimate the standard errors was measured; Cohen's κ is a suitable inter-rater reliability measure for the rating process. There is insufficient data to suggest a particular pressure ulcer categorization scheme for use in routine clinical practice. It is necessary to conduct interrater reliability studies, in which similar raters classify comparable samples using various pressure ulcer categorization schemes.

Although the inclusion criteria were set extremely liberally to obtain as much interrater reliability data of PU categorization as feasible, all raters had to be handled symmetrically due to the usage of κ-coefficients as agreement metrics. This requirement was not met in the vast majority of the included research. When comparing one set of raters (such as ward nurses) to another, one or more of the examined sample of raters (such as researchers or PU experts) might be used as a standard. In these situations, the standard is probably more precise, hence the κ-statistic is no longer suitable **(Fleiss et al. 2003).** As a result, diagnostic accuracy was the focus of many investigations rather than interrater dependability.

**Kvålseth, 2015** indicated that their interpretations are meaningless and may even be completely deceptive when disagreement is seen with a greater probability than it would be by chance. It is possible to get the fixed lower bound of −1 for the new coefficients regardless of the marginal distributions. Accordingly, **Mudford et al., 1997** reported the overall level of agreement on occurrence was below average, averaging 63.5%, with five out of sixteen Subject X State Agreement Indices exceeding 80%. To measure the origins of disagreement, the percentage of disagreement on the occurrence of a metric that had not been published before was calculated. The

agreement data, despite their often insufficient dependability, were overlaid on behavior state profiles of participants to show how inferences might be made from the data. Researchers studying phenomena that are not immediately susceptible to observation are advised to use this method for interobserver agreement as well as information interpretation (**Mudford et al., 1997**).

Therefore, there are two implications when kappa is interpreted as the average of the category kappa. On the one hand, the total kappa cannot accurately represent the complexity of the agreement between the observers if the category kappa is significantly different, for instance, strong agreement on one category but poor agreement on another. It would be best practice to publish (different) category coefficients for each particular category if a researcher is interested in comprehending the patterns of agreement and disagreement, as this offers far more information than just giving a single value. As an alternative, one can represent agreement using log-linear or latent class models **(Kerr et al., 2015).**

Furthermore, Cohen's Kappa makes a presumption which the raters were selected with purpose. The kappa will be applied in place of the population of raters if the raters are selected at random. Interrater reliability was previously assessed using percent agreement, which is the number of agreement scores divided by the total number of scores. Just as a chance "correct" response on a multiple-choice exam is feasible, chance agreement resulting from rater guesswork is always a possibility. This element of chance is taken into consideration by the Kappa statistic **(Chang, 2014).**

## Conclusions:

Poor interrater reliability is unacceptable in clinical research and healthcare, particularly when study findings have the potential to alter clinical practice and negatively impact patient outcomes. Calculating both kappa and percent agreement is perhaps the best advice for researchers. In situations when rater guessing is expected to be high, the kappa statistic may be used; Nevertheless, the researcher may reasonably rely on percent agreement for evaluating reliability among respondents if raters have adequate training.

It is a different matter entirely whether rater-to-standard dependability is a viable notion and how to measure it; yet, it appears that Cohen's kappa was abused, hence the study cannot be included

in a meta-analysis. In fact, inter-rater dependability for a wide range of complex rating scenarios may now be estimated using Cohen's kappa. Unfortunately, empirical research seldom, if at all, uses those broader measurements. Clear and simple instructions on performing inter-rater reliability studies for selecting the right method of inter-rater reliability for various rating scenarios are necessary due to the abuse of Cohen's kappa and the abundance of increased inter-rater reliability indicators.

Finally, Cohen's kappa must be used appropriately in primary research. Since the accuracy of ratings is typically impossible to ascertain and all subjects are assessed by two raters who are equally competent, it follows that Cohen's kappa is a suitable indicator of inter-rater reliability. Cohen's kappa cannot be used to determine the consistency of ratings between two raters when certain presumptions are not met. For example, Cohen's kappa between hospital nurses and PU specialists was reported by Hart et al. (2006). Expert ratings were accepted as accurate classifications, and rater-to-standards dependability was defined as the consistency of ratings across nurses and experts.

## References:


1. Chang, C.H., 2014. Cohen's kappa for capturing discrimination. *International Health*, *6*(2), pp.125-129.

2. Chicco, D., Warrens, M.J. and Jurman, G., 2021. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*, *9*, pp.78368-78381.

3. Cicchetti, D.V., Klin, A. and Volkmar, F.R., 2017. Assessing binary diagnoses of bio-behavioral disorders: the clinical relevance of Cohen's Kappa. The Journal of Nervous and Mental Disease, 205(1), pp.58-65.

4. Delgado, R. and Tibau, X.A., 2019. Why Cohen's Kappa should be avoided as performance measure in classification. *PloS one*, *14*(9), p.e0222916.

5. Figueroa, A., Ghosh, S. and Aragon, C., 2023, July. Generalized Cohen's kappa: a novel inter-rater reliability metric for non-mutually exclusive categories. In International Conference on Human-Computer Interaction (pp. 19-34). Cham: Springer Nature Switzerland.

6. Hsu, L.M. and Field, R., 2003. Interrater agreement measures: Comments on Kappan, Cohen's Kappa, Scott's π, and Aickin's α. Understanding Statistics, 2(3), pp.205-219.

7. Kerr, G.H., Fischer, C. and Reulke, R., 2015, July. Reliability assessment for remote sensing data: beyond Cohen's kappa. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 4995-4998). IEEE.

8. Knop, K. and Borkowski, S., 2011. THE ESTIMATION OF ALTERNATIVE CONTROL EFFICIENCY WITH THE USE OF THE COHEN'S KAPPA COEFFICIENT. *Management & Production Engineering Review (MPER)*, *2*(3).

9. Kottner, J., Raeder, K., Halfens, R. and Dassen, T., 2009. A systematic review of interrater reliability of pressure ulcer classification systems. *Journal of clinical nursing*, *18*(3), pp.315-336.

**10.** Kvålseth, T.O., 2015. Measurement of interobserver disagreement: Correction of Cohen's kappa for negative values. Journal of Probability and Statistics, 2015.

**11.** Mudford, O.C., Hogg, J. and Roberts, J., 1997. Interobserver agreement and disagreement in continuous recording exemplified by measurement of behavior state. American Journal on Mental Retardation, 102(1), pp.54-66.

**12.** Nixon, J., Thorpe, H., Barrow, H., Phillips, A., Andrea Nelson, E., Mason, S.A. and Cullum, N., 2005. Reliability of pressure ulcer classification and diagnosis. *Journal of advanced nursing*, *50*(6), pp.613-623.

**13.** Pérez, J., Díaz, J., Garcia-Martin, J. and Tabuenca, B., 2020. Systematic literature reviews in software engineering—Enhancement of the study selection process using Cohen's kappa statistic. Journal of Systems and Software, 168, p.110657.

**14.** Rau, G. and Shih, Y.S., 2021. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. Journal of english for academic purposes, 53, p.101026.

**15.** Simon, P., 2006. Including omission mistakes in the calculation of Cohen's Kappa and an analysis of the coefficient's paradox features. Educational and Psychological Measurement, 66(5), pp.765-777.

**16.** Steinijans, V.W., Diletti, E., Bömches, B., Greis, C. and Solleder, P., 1997. Interobserver agreement: Cohen's kappa coefficient does not necessarily reflect the percentage of patients with congruent classifications. International journal of clinical pharmacology and therapeutics, 35(3), pp.93-95.

**17.** Sun, S., 2011. Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology*, *11*, pp.145-163.

**18.** Vanbelle, S., 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika*, *81*(2), pp.399-410.

**19.** Vieira, S.M., Kaymak, U. and Sousa, J.M., 2010, July. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems* (pp. 1-8). IEEE.

**20.** Wang, J. and Xia, B., 2019, August. Relationships of Cohen's Kappa, Sensitivity, and Specificity for Unbiased Annotations. In Proceedings of the 4th International Conference on Biomedical Signal and Image Processing (pp. 98-101).

**21.** Więckowska, B., Kubiak, K.B., Jóźwiak, P., Moryson, W. and Stawińska-Witoszyńska, B., 2022. Cohen's Kappa Coefficient as a Measure to Assess Classification Improvement following the Addition of a New Marker to a Regression Model. *International Journal of Environmental Research and Public Health*, *19*(16), p.10213.

# STAT.Dilman.Kareem@MSc.2024.Non Parametric

ORIGINALITY REPORT

| 22% | 12% | 15% | 7% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Shuyan Sun. "Meta-analysis of Cohen's kappa", Health Services and Outcomes Research Methodology, 2011<br>Publication | 2% |
| 2 | Submitted to University of Lincoln<br>Student Paper | 2% |
| 3 | www.statisticshowto.datasciencecentral.com<br>Internet Source | 2% |
| 4 | www.semanticscholar.org<br>Internet Source | 1% |
| 5 | link.springer.com<br>Internet Source | 1% |
| 6 | Matthijs J. Warrens. "New Interpretations of Cohen's Kappa", Journal of Mathematics, 2014<br>Publication | 1% |
| 7 | Gerald Rau, Yu-Shan Shih. "Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data", Journal of English for Academic Purposes, 2021 | 1% |

Publication

| | | |
|---|---|---|
| 8 | Jorge Pérez, Jessica Díaz, Javier Garcia-Martin, Bernardo Tabuenca. "Systematic literature reviews in software engineering - enhancement of the study selection process using Cohen's Kappa statistic", Journal of Systems and Software, 2020<br>Publication | 1 % |
| 9 | repository.library.georgetown.edu<br>Internet Source | 1 % |
| 10 | scales.arabpsychology.com<br>Internet Source | 1 % |
| 11 | Juan Wang, Bin Xia. "Relationships of Cohen's Kappa, Sensitivity, and Specificity for Unbiased Annotations", Proceedings of the 2019 4th International Conference on Biomedical Signal and Image Processing (ICBIP 2019) - ICBIP '19, 2019<br>Publication | 1 % |
| 12 | Submitted to Asian Institute of Management<br>Student Paper | 1 % |
| 13 | Patricia Simon. "Including Omission Mistakes in the Calculation of Cohen's Kappa and an Analysis of the Coefficient's Paradox Features", Educational and Psychological Measurement, 2016<br>Publication | 1 % |

| | | |
|---|---|---|
| 14 | openpublichealthjournal.com<br>Internet Source | 1 % |
| 15 | "Human Interface and the Management of Information", Springer Science and Business Media LLC, 2023<br>Publication | <1 % |
| 16 | Hsu, Louis M., and Ronald Field. "Interrater Agreement Measures: Comments on $Kappa_n$, Cohen's Kappa, Scott's Ï€, and Aickin's Î±", Understanding Statistics, 2003.<br>Publication | <1 % |
| 17 | www.researchgate.net<br>Internet Source | <1 % |
| 18 | Lisa Yang, Afreen Siddiqi, Olivier L. de Weck. "Urban Roads Network Detection from High Resolution Remote Sensing", IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019<br>Publication | <1 % |
| 19 | pure.rug.nl<br>Internet Source | <1 % |
| 20 | A. Geymen. "IMPACTS OF BOSPORUS BRIDGES oN THE ISTANBUL METROPOLITAN SETTLEMENT AREAS", Land Degradation & Development, 2013<br>Publication | <1 % |

| | | |
|---|---|---|
| 21 | Submitted to Anglia Ruskin University<br>Student Paper | <1 % |
| 22 | Juan Wang, Yongyi Yang, Bin Xia. "A Simplified Cohen’s Kappa for Use in Binary Classification Data Annotation Tasks", IEEE Access, 2019<br>Publication | <1 % |
| 23 | www.thieme-connect.com<br>Internet Source | <1 % |
| 24 | Submitted to Queensland University of Technology<br>Student Paper | <1 % |
| 25 | etheses.whiterose.ac.uk<br>Internet Source | <1 % |
| 26 | Submitted to Eiffel Corporation<br>Student Paper | <1 % |
| 27 | Louis M. Hsu, Ronald Field. " Interrater Agreement Measures: Comments on Kappa , Cohen's Kappa, Scott's π, and Aickin's α ", Understanding Statistics, 2003<br>Publication | <1 % |
| 28 | Submitted to University of Sheffield<br>Student Paper | <1 % |
| 29 | www.ijres.org<br>Internet Source | <1 % |
| 30 | repository.javeriana.edu.co<br>Internet Source | |

| | | |
|---|---|---|
| | | <1 % |
| 31 | Matthijs J. Warrens. "Cohen's weighted kappa with additive weights", Advances in Data Analysis and Classification, 2013<br>Publication | <1 % |
| 32 | doaj.org<br>Internet Source | <1 % |
| 33 | escholarship.org<br>Internet Source | <1 % |
| 34 | etheses.bham.ac.uk<br>Internet Source | <1 % |
| 35 | researchdirect.uws.edu.au<br>Internet Source | <1 % |
| 36 | Submitted to University of Melbourne<br>Student Paper | <1 % |
| 37 | www.deepdyve.com<br>Internet Source | <1 % |
| 38 | H. Lunt. "Cerebrovascular disease as a general medicine discharge diagnosis: assessment of diagnostic agreement on case note review", Internal Medicine Journal, 6/1999<br>Publication | <1 % |

| | | |
|---|---|---|
| 39 | Jane Nixon. "Reliability of pressure ulcer classification and diagnosis", Journal of Advanced Nursing, 6/2005<br>Publication | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | On | | |