k-Rater Reliability: The Correct Unit of Reliability for Aggregated Human Annotations

Anonymous ACL submission

Abstract

Since the inception of crowdsourcing, aggregation has been a common strategy for dealing with unreliable data. Aggregate ratings are more reliable than individual ones. However, many NLP datasets that rely on aggregate ratings only report the reliability of individual ones, which is the incorrect unit of analysis. In these instances, the data reliability is being under-reported. We present empirical, analytical, and bootstrap-based methods for measuring the reliability of aggregate ratings. We call this *k-rater reliability* (kRR), a multi-rater extension of inter-rater reliability (IRR). We apply these methods to the widely used word similarity benchmark dataset, WordSim. We conducted two replications of the WordSim dataset to obtain an empirical reference point. We hope this discussion will nudge researchers to report kRR, the correct unit of reliability for aggregate ratings, in addition to IRR.

1 Introduction

Crowdsourcing has become a mainstay for data collection in NLP (Geva et al., 2019; Sabou et al., 2014). It can produce data in a scalable and cost effective manner. However, these benefits come at a cost: quality. As researchers transitioned to replacing linguists with crowd workers for NLP labeling tasks, they understood data reliability was a concern. One common strategy to increase data reliability is to collect multiple, independent judgements and aggregate them. Indeed, early papers such as Snow et al. (2008) show that average ratings correlate more strongly with expert judgements. This makes sense, as average ratings are known to have a higher reliability than individual ones (Ebel, 1951).

A number of strategies have been proposed to address data quality issues, e.g. rater modeling, label correction, label pruning (Kumar and Lease, 2011), but aggregation remains very popular (Jung and Lease, 2011). Sheshadri and Lease (2013) present nine crowdsourced datasets across a wide range of NLP tasks to compare different aggregation methods. See Difallah and Checco (2021) for a recent review of aggregation techniques. Aggregation has become the default method for acquiring reliable data from the crowd.

After we adopted aggregation as a community, we forgot to update our reliability measures correspondingly. With aggregation, the data collection artifacts are no longer individual ratings, but average ratings or majority ratings. Focusing on IRR, we are unable to capture the increase in reliability due to aggregation.

By shifting our attention to the correct unit of analysis with a higher reliability, this may even have a side effect of lessening the stigma on low-IRR datasets. As a result, this may create a path forward towards reliable data on subjective tasks, where a high IRR is difficult to obtain, such as emotions (Wong et al., 2021) and toxicity (Wulczyn et al., 2017). With a reproducibility crisis looming in the background (Baker, 2016; Hutson, 2018), more frequent and accurate reporting of reliability is our primary safeguard (Paritosh, 2012).

We present *k-rater reliability* (kRR) as a multi-rater generalization of IRR to capture the reliability of aggregate ratings. We demonstrate a general empirical method for computing kRR, by conducting replications of a widely used word similarity dataset, WordSim-353 (Finkelstein et al., 2001). We discuss bootstrap as a simulation solution in situations with high rating redundancy. Then we present two techniques in the intra-class correlation (ICC) framework to compute kRR analytically. We conclude with recommendations for reporting reliability of crowdsourced annotations, and novel research questions to expand the usefulness of kRR.

2 Prior Work

Various authors have stressed the importance of measuring reliability for the correct unit of analy-

sis. Ebel (1951) asks "Is it better to estimate the reliability of individual ratings or the reliability of average ratings? If decisions are based upon average ratings, it of course follows that the reliability with which one should be concerned is the reliability of those averages." Similarly, Shrout and Fleiss (1979) ask "Is the unit of analysis an individual rating or the mean of several ratings?" The authors explain "the reliability of the mean rating is of interest" when the mean ratings is used. Hallgren (2012) reiterates, "the researcher must specify the unit of analysis" and decide whether to measure "the reliability of the ratings based on averages of ratings provided by several coders or based on ratings provided by a single coder."

The unit of analysis informs the reliability coefficient as well. Shrout and Fleiss (1979) list several types intra-class correlation coefficient, one of which is for average ratings. They call it ICC(k), where k is the number of ratings averaged over. ICC is designed for continuous scales. See Feldt (1965) for generalization to the the dichotomous case. McGraw and Wong (1996) use a slightly different notation ICC(1,k) to explicitly denote that it is for a one-way random effects model, where the raters are treated as interchangeable.

Another way to arrive at ICC(k) is via the Spearman-Brown (SB) prophecy formula (Spearman, 1910; Brown, 1910). de Vet et al. (2017) show that, originally designed to predict test reliability at various test lengths, SB can predict ICC(k) at any k based on ICC(1), reliability of individual ratings. Both ICC(k) and SB are set in the ICC framework. The authors are not aware of multi-rater generalization for other reliability coefficients, such as Cohen's (1960) *kappa* or Krippendorff's *alpha* (Krippendorff, 2011), used widely in linguistic annotations.

3 *k*-rater Reliability

Inter-rater reliability measures the reliability of individual raters. Based on this notion, we use k-rater reliability to denote the reliability of groups of k raters. The groups' reliability is defined as the chance-adjusted agreement between their aggregate judgements. kRR is analogous to IRR, where each rater is a committee and each rating is an group judgement.

Like IRR, kRR denotes a family of reliability indices for different rating scales, distance functions, and assumptions relevant to the annotation tasks. For continuous data, the aggregation function can be the mean, and the distance function the squared distance; for categorical data, the majority vote and equality; for ranks data, the mean reciprocal rank and Spearman's ρ. Much like IRR, kRR is a general notion and is agnostic to these choices. Any coefficients suitable for IRR are suitable for kRR. This allows one to build upon the rich IRR literature and the many different coefficients for different experimental conditions. For example, in a binary task, if all the items are rated by two fixed but distinct groups of raters (raters from different locales), Cohen's (1960) *kappa* is a suitable reliability index for kRR. Whereas if the raters groups are homogeneous, and the rating scale is ordinal (e.g. Likert), then Krippendorff's *alpha* (Krippendorff, 2011) can be used.

The most direct way to observe the chance-adjusted agreement between aggregate ratings is by replicating them, i.e., reproducing the entire annotation experiment and computing the reliability between the two vector of replicated means.[1] We call this the empirical approach and illustrate it with a word similarity dataset.

3.1 Replicating the WordSim Dataset

WordSim-353 (Finkelstein et al., 2001) is a widely used benchmark for measuring a system's ability to compute similarity between two words, and has been cited over 1500 times. The dataset contains 353 word pairs. Each word pair is rated by the same 13 workers for their similarity on a scale from 1 to 10. The 13 ratings on each word pair are then aggregated into a mean score. It is important to note that only the mean of the ratings are utilized by all the research using this dataset as a benchmark.[2] So the unit of analysis is the aggregate of the 13 ratings, not individual ratings.

Nearly twenty years have elapsed since the creation of the WordSim dataset. It is impossible to re-create the original experimental conditions due rater population changes. Therefore, we created two replications in order to approximate the kRR of the original dataset.[3] We used the original annotation guidelines on Amazon Mechanical Turk.[4] In each replication, we collected 13 judgements on each of the same 353 word pairs. These are our

[1] If the original experiment has a large number of annotation items, one can work with a random sub-sample instead.

[2] https://aclweb.org/aclwiki/WordSimilarity-353_Test

[3] We will open-source it with the publication of this paper.

[4] Raters were paid on average USD 9.5 per hour.

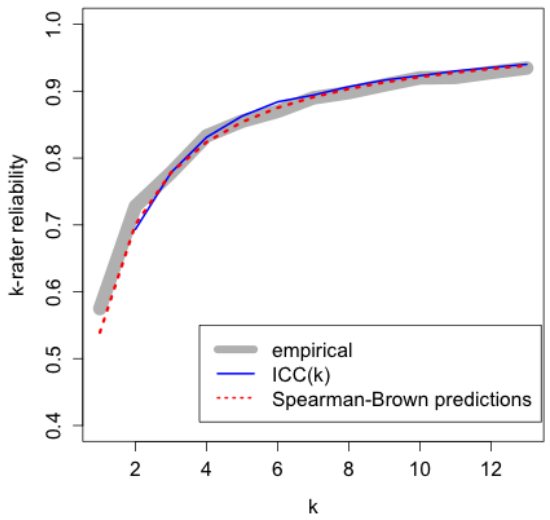

Figure 1: k-rater reliability for replications of WordSim benchmark, calculated using 3 different methods: 1) Empirical, based on replications, 2) ICC(k), analytical, and 3) SB predictions. Note ICC(1) is not available as we only have a single column of ratings available at $k = 1$. All SB predictions are based on only 2 ratings per item.

best attempts to replicate the original experiment.

3.2 Computing kRR Empirically

We take k column of ratings at random from each of the two replications, compute the k-rating mean scores for each replication, and measure the reliability between them using Krippendorf's *alpha*, the most widely used and general reliability index. We do this for $k = 1, 2, \ldots, 13$. The resulting kRR values are shown in Fig.1. At $k = 1$, the IRR is 0.574, slightly lower than the 0.6 originally reported in Finkelstein et al. (2001). At $k = 13$, the k-rater reliability is 0.94, quite a bit higher than the IRR. In addition, Fig.1 shows the marginal returns on increasing the number of ratings on the replicated datasets.

4 Other Approaches to Computing kRR

The empirical approach is general, as it is agnostic to the choice of rating scale, aggregation function, and reliability coefficient. However, it has a major drawback. As we see in Section3.1, it can be difficult to do a perfect replication post-fact. This backward incompatibility will present a challenge to computing kRR for existing datasets. Below we present two other approaches that can work on existing datasets under some conditions without requiring any additional data collection.

4.1 Bootstrap

Bootstrap (Efron and Tibshirani, 1994) is a re-sampling technique commonly used for quantifying uncertainty in statistical parameter estimation. One can bootstrap an NLP annotations dataset by re-sampling ratings within each annotation item with replacement at the same sample size. If one treats each bootstrap sample as a replication, then one can apply the technique discussed in Section 3 to obtain a *bootstrapped* kRR. Bootstrap is an approximate technique and works better with larger sample sizes, typically 20 observations and above for a single distribution. The 13-rating redundancy in the WordSim replications is arguably small for a typical bootstrap exercise, but it makes up for it with a large number of items.

Before we apply bootstrap to the original WordSim dataset, we first verify its soundness by comparing it against the empirical results obtained from Section 3.2. When applied to one of the two recent replications, the bootstrapped 13-rater reliability is 0.943. This is comparable to the 0.94 found empirically. When applied to the original dataset, the bootstrapped 13-rater reliability is 0.953. The exact method introduced below produces a very similar value of 0.95 (Table 1).

4.2 Intra-class Correlation

Intra-class correlation (ICC) is a popular reliability coefficient for continuous data in behavioral and medical sciences. ICC gives researchers granular control over assumptions about the raters. For example, each annotation item can be rated by the same set of raters, or different sets of raters (interchangeability). In the former, the raters can be treated as either fixed or randomly drawn from a population. Shrout and Fleiss (1979) and McGraw and Wong (1996) give very extensive treatment on different ICC types for different rater assumptions.

In this paper, we focus on the most basic definition, one that treats raters as interchangeable. The ICC for a k-rater average is denoted as ICC(k) using McGraw and Wong's notation. ICC(1) is hence just the reliability of individual ratings. ICC(k) can be computed by summing squares of differences on the data matrix (Shrout and Fleiss, 1979). Software implementations of ICC are also widely available, e.g. in R and Python.

We first verify ICC's accuracy by comparing it against the empirical results in Section 3.2. To do that, we calculate ICC(k) for one of the two

Unit of analysis	Method	reliability
single-rating	ICC	0.59
13-rating mean	ICC	0.95
13-rating mean	bootstrap	0.953

Table 1: Reliability of the original WordSim benchmark. First two rows are analytical estimates ICC(1) and ICC(13). Both computed using all 13 available ratings. Third row is a re-sampling based bootstrapped estimate based on 100 bootstrap samples.

recent WordSim replications for $k = 1, 2, \ldots, 13$ and overlay the results over the empirical curve in Fig.1. We can see ICC(k) matches the empirical results quite well.

After verifying the technique, we compute ICC(k) on the original WordSim dataset. We report in Table 1 both ICC(1) and ICC(13) to show the increase in reliability. They are respectively 0.59 and 0.95.[5]

4.3 Extrapolation of ICC(k)

ICC(k) quantifies the reliability of the k-rater average in the current experiment. If this reliability is too low, the researcher may want to increase the value of k. In this case, it would be helpful to know how additional ratings would impact reliability. This is analogous to calculating the required sample size for a given margin of error in a poll. For this purpose, the Spearman-Brown formula (SB) (Spearman, 1910; Brown, 1910) can be a useful tool. It predicts ICC(k) for any value of k based on single-rating ICC(1) in the current experiment:

$$\text{ICC}(k) = \frac{k \cdot \text{ICC}(1)}{1 + (k-1) \cdot \text{ICC}(1)}. \qquad (1)$$

Warrens (2017) and de Vet et al. (2017) show that SB and ICC(k) are indeed equivalent.[6] This finding merely confirms past observations that SB predicts empirical results accurately (Remmers et al., 1927). A limitation of SB is clearly that it only works with ICC. However, Fleiss and Cohen (1973) show ICC is actually equivalent to weighted-kappa with quadratic weights, so it likely has wider applicability.

To verify the formula, we apply SB to one of the two recent WordSim replications and overlay the results over the empirical curve obtained earlier. When computing SB, we only provide it with 2 ratings, in order to assess its predictive accuracy. That is, we first compute ICC(1) with 2 randomly drawn ratings from each word pair, then we plug this ICC(1) value into Eq.1 for $k = 1, 2, \ldots, 13$. The SB curve is overlaid over the empirical curve in Fig.1. We see that SB tracks the empirical results very well even at high k. This is remarkable as the empirical approach requires 26 ratings for $k = 13$, whereas SB merely requires 2 for any value of k.

5 Conclusions and Discussion

We pointed out where aggregated ratings are used, as is the case in many crowdsourced datasets, reliability of aggregate ratings is a more accurate accounting of data reliability. We introduced k-rater reliability (kRR) as a multi-rater extension of IRR. We demonstrated empirical, analytical, and bootstrap-based methods for computing the kRR on the original WordSim dataset and our recent replications. All three methods produce similar estimates for 13-rater reliability ranging from 0.94 to 0.953.

While aggregation makes it possible to have reliable benchmarks on subjective topics, some readers may feel uneasy about increasing reliability via replication, as opposed to other traditional means such as improving annotation guidelines. This concern can be mediated by reporting both IRR and kRR. In fact, kRR is not meant to replace IRR, but rather complement it. IRR speaks to the reliability of the experiment, whereas kRR the aggregate ratings we consume. We urge researchers to report both where possible. In fact, Hallgren (2012) states, "In cases where single measures ICCs are low but average-measures ICCs are high, the researcher may report both ICCs to demonstrate this discrepancy."

This research also raises interesting questions for future research:

1. How do we derive multi-rater generalizations for coefficients other than ICC?

2. Is the Landis and Koch (1977) kind of interpretation for IRR suitable for kRR?

We urge researchers to report both IRR and kRR of aggregated human annotations, and for further inquiry around the above fundamental questions about reliability.

[5]The former is computed using two-way random without interaction ICC(1), the latter two-way random without interaction ICC(13). The equivalent one-way models yield identical point estimates.

[6]The only exception is two-way mixed model with interaction (Warrens, 2017).

References

Monya Baker. 2016. Reproducibility crisis. *Nature*, 533(26):353–66.

William Brown. 1910. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, 3(3):296–322.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Henrica C.W. de Vet, Lidwine B. Mokkink, David G. Mosmuller, and Caroline B. Terwee. 2017. Spearmanâbrown prophecy formula and cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85:45–49.

Djellel Difallah and Alessandro Checco. 2021. Aggregation techniques in crowdsourcing: Multiple choice questions and beyond. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4842–4844.

Robert L Ebel. 1951. Estimation of the reliability of ratings. *Psychometrika*, 16(4):407–424.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Leonard S Feldt. 1965. The approximate sampling distribution of kuder-richardson reliability coefficient twenty. *Psychometrika*, 30(3):357–370.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34.

Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

Hyun Joon Jung and Matthew Lease. 2011. Improving consensus accuracy via z-score and weighted voting. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Abhimanu Kumar and Matthew Lease. 2011. Modeling annotator accuracies for supervised learning. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 19–22.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74.

Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30.

Praveen Paritosh. 2012. Human computation must be reproducible. In *WWW 2012, Lyon.*

HH Remmers, NW Shock, and EL Kelly. 1927. An empirical study of the validity of the spearman-brown formula as applied to the purdue rating scale. *Journal of Educational Psychology*, 18(3):187.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866. Citeseer.

Aashish Sheshadri and Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Charles Spearman. 1910. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295.

Matthijs J Warrens. 2017. Transforming intraclass correlation coefficients with the spearman–brown formula. *Journal of clinical epidemiology*, 85:14–16.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.