References

Your Name

Table of contents

Agresti, A. (2003). *Categorical data analysis* (Vol. 482). John Wiley & Sons.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to cohen's kappa. *Biometrics*, 293–302.

Ascari, R., & Migliorati, S. (2021). A new regression model for overdispersed binomial data accounting for outliers and an excess of zeros. *Statistics in Medicine*, *40*(17), 3895–3914.

Bassett, R., & Deride, J. (2019). Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, *174*, 129–144.

Bennett, E. M., Alpert, R., & Goldstein, A. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, *18*(3), 303–308.

Bonett, D. G. (2022). Statistical inference for g-indices of agreement. *Journal of Educational and Behavioral Statistics*, *47*(4), 438–458.

Brennan, R. L., Measurement in Education, N. C. on, et al. (2006). *Educational measurement*. Praeger Publishers,.

Button, C. M., Snook, B., & Grant, M. J. (2020). Inter-rater agreement, data reliability, and the crisis of confidence in psychological research. *Quant Methods Psychol*, *16*(5), 467–471.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429.

Carpenter, B. (2008). Multilevel bayesian models of categorical data annotation. *Unpublished Manuscript*, *17*(122), 45–50.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chaturvedi, S., & Shweta, R. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, *41*(3), 20–27.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, *10*, 92–110.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 20–28.

Delgado, R., & Tibau, X.-A. (2019). Why cohen's kappa should be avoided as performance measure in classification. *PloS One*, *14*(9), e0222916.

Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement*, *13*, 321–335.

Engelhard Jr, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56–70.

Eubanks, D. (2017). (Re)visualizing rater agreement:beyond single-parameter measures. *Journal of Writing Analytics*, *1*.

Eubanks, D. A. (2014). Causal interfaces. *Arxiv.org Preprint*. http://arxiv.org/abs/1404.4884v1

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. *arXiv Preprint arXiv:2011.01808*.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Grilli, L., Rampichini, C., & Varriale, R. (2015). Binomial mixture modeling of university credits. *Communications in Statistics - Theory and Methods*, *44*(22), 4866–4879. https://doi.org/10.1080/03610926.2013.804565

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Hodgson, R. T. (2008). An examination of judge reliability at a major US wine competition. *Journal of Wine Economics*, *3*(2), 105–113.

Holley, J. W., & Guilford, J. P. (1964). A note on the g index of agreement. *Educational and Psychological Measurement*, *24*(4), 749–753.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., & Hovy, E. (2013). Learning whom to trust with MACE. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1120–1130.

Krippendorff, K. (2013). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. *Annals of the International Communication Association*, *36*(1), 481–499.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Krippendorff, K., & Fleiss, J. L. (1978). Reliability of binary attribute data. *Biometrics*, *34*(1), 142–144.

Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., & Subrahmanian, V. (2018). Rev2: Fraudulent user prediction in rating platforms. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 333–341.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

McLachlan, G., & Peel, D. (2000). Wiley series in probability and statistics. *Finite Mixture Models*, 420–427.

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., & Poesio, M. (2018). Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, *6*, 571–585.

Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements in symposium on scientific objectivity, vedbaek, mau 14-16, 1976. *Danish Year-Book of Philosophy Kobenhavn*, *14*, 58–94.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* MESA Press.

Ross, V., & LeGrand, R. (2017). Assessing writing constructs: Toward an expanded view of inter-reader reliability. *Journal of Writing Analytics*, *1*.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 321–325.

Shabankhani, B., Charati, J. Y., Shabankhani, K., & Cherati, S. K. (2020). Survey of agreement between raters for nominal data using krippendorff's alpha. *Arch Pharma Pract*, *10*(S1), 160–164.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Team, S. D. (2022). *Stan user's guide 2.34*. Stan Development Team.

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for cohen's kappa–a comparison of basic properties. *MethodsX*, 102212.

Williams, D. (1975). 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 949–952.