Bibliography

[1] Aickin, M. (1990), "Maximum Likelihood Estimation of Agreement in the Constant Predictive Probability Model, and Its Relation to Cohen's Kappa." *Biometrics*, **46**, 293-302.

[2] Agresti, A. (1988), "A Model Agreement Between Ratings on an Ordinal Scale." *Biometrics*, **44**, 539-548.

[3] Agresti, A. (1992), "Modeling patterns of agreement and disagreement." *Statistical Methods in Medical Research*,1, 201-218.

[4] Altman, D. G. (1991). *Practical Statistics for Medical Research.* Chapman and Hall.

[5] Benini, R. (1901). *Principii di Demongraphia: Manuali Barbera Di Scienze Giuridiche Sociali e Politiche* (No. 29). Firenze, Italy : G. Barbera.

[6] Bennett, E. M., Alpert, R., and Goldstein, A. C. (1954), "Communications through limited response questioning." *Public Opinion Quarterly*, **18**, 303-308.

[7] Berry, K. J., and Mielke, Jr., P. W. (1988), "A Generalization of Cohen's Kappa Agreement Measure To Interval Measurement and Multiple Raters," *Educational and Psychological Measurement*, **48**, 921-933.

[8] Bland, M. J., Altman, D. G. (1986),"Statistical Methods for Assessing Agreement between two Methods of Clinical Measurement." *Lancet*, **1**, 307-310.

[9] Bland, M. J., Altman, D. G. (1996), "Statistics Notes: Measurement error." *British Medical Journal*, **312**, p. 1654.

[10] Brennan, R. L., and Prediger, D. J. (1981). "Coefficient Kappa : some uses, misuses, and alternatives." *Educational and Psychological Measurement*, 41, 687-699.

[11] Byrt, T., Bishop, J., and Carlin, J. B. (1993). "Bias, prevalence and Kappa." *Journal of Clinical Epidemiology*,46, 423-429.

[12] Cantor, A. B. (1996). "Sample-Size Calculations for Cohen's Kappa.", *Psychological Methods*, 1, 150-153.

[13] Carletta, J. (1996). "Assessing Agreement on Classification Tasks: the Kappa Statistic." *Computational Linguistics*,**22**, 1-6.

[14] Carmines, E. G., and Zeller, R. A. (1979), *Reliability and Validity Assessment*,Sage Publications.

[15] Cicchetti, D. V., and Feinstein, A. R. (1990). "High Agreement but low Kappa : II. Resolving the paradoxes." *Journal of Clinical Epidemiology*, 43, 551-558.

[16] Cochran, W. G. (1977). *Sampling Techniques*, John Wiley & Sons, Inc. : New York.

[17] Cohen, J. (1960). "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement*, 20, 37-46.

[18] Cohen, J.(1968). "Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit." *Psychological Bulletin*, 70, 213-220.

[19] Conger, A. J. (1980), "Integration and Generalization of Kappas for Multiple Raters," *Psychological Bulletin*, **88**, 322-328.

[20] Cronbach, L. J. (1951), "Coefficient Alpha, and the Internal Structure," *Psychometrika*, **16**(3), 297-334.

[21] Doros, G. and Lew, R. (2010), "Design Based on Intr-Class Correlation Coefficients." *American Journal of Biostatistics*, **1**(1), 1-8.

[22] Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement.* Peter Lang, Internationaler Verlag der Wissenschaften.

[23] Efron, B. (1979). "Bootstrap methods : another look at the jackknife." *Annals of statistics*, **7**,1-26.

[24] Everitt, B. S. (1992). *The Analysis of Contingency Tables (2nd Ed.)* Chapman and Hall, London.

[25] Feinstein, A. R., and Cicchetti, D. V. (1990), "High agreement but low kappa : I. The problems of two paradoxes," *Journal of Clinical Epidemiology*, **43**, 543-549.

[26] Fenning, S., Craig, T. J., Tanenberg-Karant, M., & Bromet, E. J. (1994). "Comparison of facility and research diagnoses in first-admission psychotic patients," *American Journal of Psychiatry*, **151**, 1423-1429.

[27] Finn, R. H. (1970), "A Note on Estimating the Reliability of Categorical Data," *Educational and Psychological Measurement*, **30**, 71-76.

[28] Flack, V. F. (1987), "Confidence intervals for the interrater agreement measure kappa," *Communications in Statistics - Theory and Methods*, **16**, 953-968.

[29] Flack, V. F., Afifi, A. A., Lachenbruch, P. A., and Schouten, H. J. A. (1988), "Sample Size Determinations for the Two Rater Kappa Statistic," *Psychometrika*, **53**, 321-325.

[30] Fleiss, J. L. (1971). "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, **76**, 378-382.

[31] Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions.* John Wiley & Sons.

[32] Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). "Large sample standard errors of kappa and weighted kappa," *Psychological Bulletin*, **72**, 323-327.

[33] Fleiss, J. L., and Davies, M. (1982). "Jackknifing Functions of Multinomial Frequencies, with an Application to a Measure of Concordance," *American Journal of Epidemiology*, **115**, 841-845.

[34] Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). Wiley Series in Probability and Statistics.

[35] Fleiss, J. L., Nee, J. C. M., and Landis, J. R. (1979). "The large sample variance of kappa in the case of different sets of raters." *Psychological Bulletin*, **86**, 974-977.

[36] Friedman, M. (1937). "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." *Journal of the American Statistical Association*, **32 (200)**, 675-701.

[37] Gartner, J. B. (1991). "The standard error of Cohen's kappa." *Statistics in Medicine*, **10**, 767-775.

[38] Giraudeau, B., and Mary, J. Y. (2001). "Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient." *Statistics in Medicine*, **20**, 3205-3214.

[39] Goodman, L. A., and Kruskal, W. H. (1954). "Measures of Association in Cross Classifications." *Journal of the American Statistical Association*, **49**, 1732-1769.

[40] Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., and Shapiro, R. W. (1981). "Reliability Studies of Psychiatric Diagnosis." *Archives of General Psychiatry*, **38**, 408-413.

[41] Guttman, L. (1945). "The test-retest reliability of qualitative data." *Psychometrika*, **11**, 81-95.

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

[43] Gwet, K. L. (2008*b*). "Variance estimation of nominal-scale inter-rater reliability with random selection of raters." *Psychometrika*, **73**, 407-430.

[44] Gwet, K. L. (2008*c*). Intrarater Reliability. In R. B. D'Agostino, L. Sullivan, and J. Massaro (Eds.), *Wiley Encyclopedia of Clinical Trials* (pp. 473-485). Wiley-Interscience

[45] Gwet, K. L. (2010*a*). *How to Compute Intraclass Correlation Using Excel: A Practical Guide to Inter-Rater Reliability Assessment for Quantitative Data*, Advanced Analytics, LLC.

[46] Gwet, K. L. (2010*b*). *The Practical Guide to Statistics: Basic Concepts, Methods, and Meaning*, Advanced Analytics, LLC.

[47] Gwet, K. L. (2010*c*). *Inter-Rater Reliability Using SAS: A Practical Guide for Nominal, Ordinal, and Interval Data*, Advanced Analytics, LLC.

[48] Hale, C. A., and Fleiss, J. L. (1993). "Interval estimation under two study designs for kappa with binary classifications," *Biometrics*, **49**, 523-533.

[49] Holley, J.W., and Guilford, J. P. (1964), "A note on the G index of agreement." *Educational and Psychological Measurement*, **24**, 749-753.

[50] Holsti, O.R. (1969). *Content Analysis for the Social Sciences and Humanities*, Reading, MA: Addison-Wesley.

[51] Hubert, L., "Kappa revisited." *Psychological Bulletin*, **84**, 289-297.

[52] Janson, S., and Vegelius, J. (1979). "On generalizations of the G index and the PHI coefficient to nominal scales." *Multivariate Behavioral Research*, **14**, 255-269.

[53] Janson, H., and Olsson, U. (2001). "A Measure of Agreement for Interval or Nominal Multivariate Observations," *Educational and Psychological Measurement*, **61**, 277-289.

[54] Janson, H., and Olsson, U. (2004). "A Measure of Agreement for Interval or Nominal Multivariate Observations by Different Sets of Judges," *Educational and Psychological Measurement*, **64**, 62-70.

[55] Jung, H. W. (2003). "Evaluating interrater agreement in SPICE-based assessments," *Computer Standards & Interfaces*, **25**, 477-499.

[56] Kendall, M. G. (1938). "A New Measure of Rank Correlation," *Biometrika*, **30**, 81-93.

[57] Kendall, M. G., and Babington-Smith, B. (1939). "The Problem of *m* Rankings," *Annals of Mathematical Statistics*, **10**, 275-287.

[58] Kolmogorov, A. N. (1999). The Theory of Probability. In A. D. Aleksandrov, A. N. Kolmogorov, and M. A. Lavrent'ev (Eds.), *Mathematics - Its Contents, Methods and Meaning* (Chapter XI, pp. 229-264). Dover Publications - Dover Books on Mathematics.

[59] Kraemer, H. C. (1979). "Ramifications of a population model for κ as a coefficient of reliability," *Psychometrika*, **44**, 461-472.

[60] Kraemer, H. C., Peryakoil, V. S., and Noda, A. (2002). "Kappa Coefficients in Medical Research," *Statistics in Medicine*, **21**, 2109-2129.

[61] Krippendorff, K. (1970). "Estimating the reliability, systematic error, and random error of interval data," *Educational and Psychological Measurement*, **30**, 61-70.

[62] Krippendorff, K. (1978). "Reliability of binary attribute data," *Biometrics*, **34**, 142-144.

[63] Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*, Thousand Oaks, Calif, USA.

[64] Krippendorff, K. (2011). "Computing Krippendorff's alpha reliability." http://www.asc.upenn.edu/usr/krippendorff/mwebreliability5.pdf

[65] Krippendorff, K. (2011). "Agreement and Information in the Reliability of Coding," *Communication Methods and Measures*, **5.2**, 1-20.

[66] Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology, 3rd. Edition*, Thousand Oaks, CA: SAGE Publications, Inc.

[67] Kruskal, W. H. (1952). "A nonparametric test for the several sample problem," *Annals of Mathematical Statistics*, **23**, 525-540.

[68] Kruskal, W. H., and Wallis, W. A. (1952). "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, **47**, 583-621.

[69] Kuder, G. F., and Richardson, M. W. (1937). "The Theory of the Estimation of Test Reliability," *Psychometrika*, **2**, 151-160.

[70] Landis, J. R, and Koch G. (1977). "The measurement of observer agreement for categorical data," *Biometrics*, **33**, 159-174.

[71] Lee, J. J., and Tu, Z. N. (1994). "A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes." *Journal of Computational and Graphical Statistics*, **3**, 301-321.

[72] Leone, M.A., Gaviani, P., and Ciccone, G. (2006). "Inter-coder agreement for ICD-9-CM coding of stroke," *Neurological Sciences*, **27**, 445-448.

[73] Lindeman, R. H., Meranda, P. F., and Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*, Glenview, IL: Scott, Foresman and Company.

[74] Likert, R. (1932). "A Technique for the Measurement of Attitudes." *Archives of Psychology*, **140**, 1-55.

[75] Lindsay,B. G., Markatou, M., Ray, S., Yang, K., Chen, S. (2008). "Quadratic Distances on Probabilities: A Unified Foundation." *The Annals of Statistics*, **36**, 983-1006.

[76] Lipsitz, S. R., Laird, N. M., and Brennan, T. A. (1994). "Simple moment estimates of the κ-coefficient and its variance," *Applied Statistics*, **43**, 309-323.

[77] Light, R. J. (1971). "Measures of response agreement for qualitative data : some generalizations and alternatives," *Psychological Bulletin*, **76**, 365-377.

[78] Likert, R. (1931). "A Technique for the Measurement of Attitudes," *Archives of Psychology*, New York: Columbia University Press.

[79] Mann, H., and Whitney, D. (1947). "On a test of whether one of two random variables is stochastically larger than the other." *Annals of Mathematical Statistics*, **18**, 50-60.

[80] Marascuilo, L. A., and McSweeney, L. (1977). *Nonparametric and Distribution-free Methods for the Social Sciences.* Monterey, CA: Brooks/Cole Publishing Company.

[81] Maxwell, A. E. (1977). "Coefficient of agreement between observers and their interpretation." *British Journal of Psychiatry*, **130**, 79-83.

[82] McCarthy, P. J. (1966). "Replication : An approach to the analysis of data from complex surveys." National Center for Health Statistics, Washington, D.C., Series, 2, 14.

[83] McGraw, K. O., and Wong, S. P. (1996). "Forming Inferences About Some Intraclass Correlation Coefficients." *Psychological Methods*, **1**, 30-46.

[84] McIver, J. P., and Carmines, E. G. (1981). *Unidimensional Scaling*, Thousand Oaks, CA: Sage.

[85] Metropolis, N., and Ulam, S. (1949). "The Monte-Carlo Method." *Journal of the American Statistical Association*, **44**, 335-341.

[86] Miller, R. G. (1974). "The jackknife - a review." *Biometrika*, **61**, 1-15.

[87] Neyman, J. (1934). "On the Two Different Aspects of the Representative Method : the Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society*, **97**, 558-606.

[88] Nunnally, J. C. (1978). *Psychometric Theory (2nd ed.)*. New York: McGraw-Hill.

[89] Osgood, C.E. (1959). The Representational Model and Relevant Research Methods. In I. de Sola Pool (Ed.), *Trends in Content Analysis* (pp. 33-88). Urbana: University of Illinois Press.

[90] Park, H. M., and Jung, H. W. (2003). "Evaluating Interrater Agreement with Intraclass Correlation Coefficient in SPICE-based Software Process Assessment," *Proceedings of the Third International Conference On Quality Software*, 308-314.

[91] Pearson, K. (1896). "Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia," *Philosophical Transactions of the Royal Society of London*, Series A 187, 253-318.

[92] Pearson, K. (1900). "On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can

Reasonably be Supposed to have arisen in a Random Sampling," *Philosophical Magazine*, 5, 157-175.

[93] Perreault, W.D., and Leigh, L.E. (1989). "Reliability of nominal data based on qualitative judgments," *Journal of Marketing Research*, **26**, 135-148.

[94] Quenouille, M. H. (1949). Approximate tests of correlation in time series. *Journal of The Royal Statistical Society*, Series B, 11, 68-84.

[95] Quenouille, M. H. (1956). "Notes on bias in estimation." *Biometrika*, **61**, 353-360.

[96] Rowland, W. J. (1984), "The relationships among nuptial coloration, aggression, and courtship in male Threespine Sticklebacks," *Canadian Journal of Zoology*, **51**, 453-466.

[97] Särndal, C. E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*, Springer-Verlag New York, Inc.

[98] Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics*, **2**, 110-114.

[99] Schouten, H. J. A. (1986), "Nominal scale agreement among observers," *Psychometrika*, **51**, 453-466.

[100] Schuster, C. and von Eye, A. (2001). "Models for ordinal agreement data," *Biometrical Journal*, **43**(7), 795-808.

[101] Scott, W. A. (1955). "Reliability of content analysis : the case of nominal scale coding." *Public Opinion Quarterly*, **XIX**, 321-325.

[102] Searle, S. R. (1997). *Linear Models (Wiley Classics Library).* Wiley-Interscience: John Wiley & Sons, Inc.

[103] Shoukri, M. M (2010). *Measures of Interobserver Agreement and Reliability, Second Edition (Chapman & Hall/CRC Biostatistics Series).* CRC Press.

[104] Shrout, P. E., and Fleiss, J. L. (1979), "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin*, **86**(2), 420-428.

[105] Siegel, S., and Castellan, N. J., Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.). New York: McGraw-Hill Book Company.

[106] Sim J., and Wright C. C. (2005), "The kappa statistic in reliability studies : use, interpretation, and sample size requirements." *Physical Therapy* **85**(3), 257-268.

[107] Spearman, C. (1904), "The Proof and Measurement of Association between two Things." *American Journal of Psychology*, **15**, 72-101.

[108] Stein, C.R., Devore, R.B., and Wojcik, B.E. (2005). Calculation of the Kappa Statistic for Inter-Rater Reliability: The Case Where Raters Can Select Multiple Responses from a Large Number of Categories. In *SAS Institute Inc.*

2005. Proceedings of the Thirtieth Annual SAS® Users Group International Conference. Cary, NC : SAS Institute Inc .

[109] Tanner, M.A. and Young, M.A. (1985). "Modeling agreement among raters," *Journal of American Statistical Association*, **80**, 175-180.

[110] Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications*, Sage Publications, Beverly Hills.

[111] Tukey, J. W. (1958). "Bias and confidence in not quite large samples (Abstract)." *Annals of Mathematical Statistics*, **29**, 614.

[112] von Eye, A., and Mun, E. Y. (2006). *Analyzing Rater Agreement: Manifest Variable Methods*, Lawrence Erlbaum Associates; Pap/Cdr edition.

[113] Wallis, W. A. (1939). "The Correlation Ratio for Ranked Data," *Journal of the American Statistical Association*, **34**, 533-538.

[114] Wilcoxon, F. (1949). *Some Rapid Approximate Statistical Procedures.* Stamford, CT: Stamford Research Laboratories, American Cyanamid Corporation.

[115] Zhao, X., Liu, J.S., and Deng, K. (2013). Assumptions behind Intercoder Reliability Indices. In C.T. Salmon (Ed.), *Communication Yearbook*, **36** (pp. 419-480). Routledge