Bibliography

Agresti, A. (1988), "A model agreement between ratings on an ordinal scale." *Biometrics*, 44, 539–548.

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

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.

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

Axelson, R.D. and C.D. Kreiter (2009), "Reliability." In *Assessment in Health Professions Education* (Steven M. Downing and Rachel Yudkowsky, eds.), 57–74, Taylor and Francis.

Benini, R. (1901), *Principii di Demongraphia*: *Manuali Barbera Di Scienze Giuridiche Sociali e Politiche*. 29, G. Barbera, Firenze, Italy.

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

Berry, K.J. and P.W. Mielke Jr. (1988), "A generalization of cohen's kappa agreement measure to interval measurement and multiple raters." *Educational and Psychological Measurement*, 48, 921–933.

Brennan, R.L. and D.J. Prediger (1981), "Coefficient kappa: some uses, misuses, and alternatives." *Educational and Psychological Measurement*, 41, 687–699.

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

Cantor, A.B. (1996), "Sample-size calculations for cohen's kappa." *Psychological Methods*, 1, 150–153.

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

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

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

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

Conger, A.J. (1980), "Integration and generalization of kappas for multiple raters." *Psychological Bulletin*, 88, 322–328.

Cook, R.Dennis and Sanford Weisberg (1982), *Residuals and Influence in Regression*, 1 edition. Monographs on Statistics and Applied Probability, Chapman and Hall, 733 Third Avenue, New York NY 10017.

Cronbach, L.J. (1951), "Coefficient alpha, and the internal structure." *Psychometrika,*, 16, 297–334.

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

Efron, B. (1979), "Bootstrap methods: another look at the jackknife." *Annals of Statistics,*, 7, 1–26.

Everitt, B.S. (1992), *The Analysis of Contingency Tables*, 2 edition. Chapman and Hall, London.

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

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

Finn, R.H. (1970), "A note on estimating the reliability of categorical data." *Educational and Psychological Measurement*, 30, 71–76.

Flack, V.F., A.A. Afifi, P.A. Lachenbruch, and H.J.A. Schouten (1988), "Sample size determinations for the two rater kappa statistic." *Psychometrika*, 53, 321–325.

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

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

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

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

Goodman, L.A. and W.H. Kruskal (1954), "Measures of association in cross classifications." *Journal of the American Statistical Association*, 49, 1732–1769.

Grove, W.M., N.C. Andreasen, P. McDonald-Scott, M.B. Keller, and R.W. Shapiro (1981), "Reliability studies of psychiatric diagnosis." *Archives of General Psychiatry*, 38, 408–413.

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

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

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

Gwet, Kilem L. (2016), "Testing the difference of correlated agreement coefficients for statistical significance." *Educational and Psychological Measurement*, 76, 609–637.

Gwet, Kilem L. (2020), "Large-sample variance of fleiss generalized kappa." *Educational and Psychological Measurement*, URL https://doi.org/10.1177%2F0013164420973080.

Gwet, Kilem L. (2021), *Handbook of Inter-Rater Reliability - Volume 2: Intraclass Correlation Coefficients for Quantitative Ratings*, 5 edition, volume 1. AgreeStat Analytics, Maryland, USA.

Hayes, A.F. and K. Krippendorff (2007), "Answering the call for a standard reliability measure for coding data." *Communication Methods and Measures*, 1, 77–79.

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

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

Hripcsak, G. and A.S. Rothschild (2005), "Agreement, the f-measure, and reliability in information retrieval." *Journal of the American Medical Informatics Association*, 12, 296–298.

Hubert, L. (1977), "Kappa revisited." *Psychological Bulletin*, 84, 289–297.

Janson, H. and U. Olsson (2001), "A measure of agreement for interval or nominal multivariate observations." *Educational and Psychological Measurement*, 61, 277–289.

Janson, H. and U. Olsson (2004), "A measure of agreement for interval or nominal multivariate observations by different sets of judges." *Educational and Psychological Measurement*, 64, 62–70.

Janson, S. and J. Vegelius (1979), "On generalizations of the g index and the phi coefficient to nominal scales." *Multivariate Behavioral Research*, 14, 255–269.

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

Kendall, M. and A. Stuart (1976), *The Advanced Theory of Statistics*, 3 edition, volume 3. Griffin, London.

Khan, L., G. Mitera, L. Probyn, M. Ford, M. Christakis, J. Finkelstein, A. Donovan, L. Zhang, L. Zeng, J. Rubenstein, A. Yee, L. Holden, and E. Chow (2011), "Interrater reliability between musculoskeletal radiologists and orthopedic surgeons on computed tomography imaging features of spinal metastases." *Current Oncology*, 18, 282–287.

Klein, D. (2018), "Implementing a general framework for assessing interrater agreement in stata." *The Stata Journal*, 18, 871–901.

Kolmogorov, A.N. (1999), "The theory of probability." In *Mathematics - Its Contents, Methods and Meaning* (A.D. Aleksandrov, A.N. Kolmogorov, and M.A. Lavrent'ev, eds.), chapter Chapter XI, 229–264, Dover Publications - Dover Books on Mathematics.

Kottner, Jan and David L. Streiner (2011), "The difference between reliability and agreement." *Jornal of Clinical Epidemiology*, 64, 701–702.

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

Kraemer, H.C., V.S. Peryakoil, and A. Noda (2002), "Kappa coefficients in medical research." *Statistics in Medicine*, 21, 2109–2129.

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

Krippendorff, K. (1978), "Reliability of binary attribute data." *Biometrics*, 34, 142–144.

Krippendorff, K. (2012), *Content Analysis: An Introduction to Its Methodology*, 3 edition. Thousand Oaks, CA: SAGE Publications, Inc., California, USA.

Krippendorff, Klaus (2004), "Measuring the reliability of qualitative text analysis data." *Quality and Quantity*, 38, 787–800.

Krippendorff, Klaus (2011), "Agreement and information in the reliability of coding." *Communication Methods and Measures*, 5, 93–112.

Krippendorff, Klaus (2018), *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications, Inc.

Kuder, G.F. and M.W. Richardson (1937), "The theory of the estimation of test reliability." *Psychometrika*, 2, 151–160.

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

Leone, M.A., P. Gaviani, and G. Ciccone (2006), "Inter-coder agreement for icd-9-cm coding of stroke." *Neurological Sciences*, 27, 445–448.

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

Likert, Rensis (1932), "A technique for the measurement of attitudes." *Archives of Psychology*, 22, 3–55.

Lindsay, B.G., M. Markatou, S. Ray, K. Yang, and S. Chen (2008), "Quadratic distances on probabilities: A unified foundation." *The Annals of Statistics*, 36, 983–1006.

Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015), "The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment." *Computational Linguistics*, 41, 437–479.

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

McCarthy, P.J. (1966), "Replication: An approach to the analysis of data from complex surveys." Technical Report 14, National Center for Health Statistics, Vital and health statistics, Washington, D.C. Series 2: Data evaluation and methods.

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

Metropolis, N. and S. Ulam (1949), "The monte-carlo method." *Journal of the American Statistical Association*, 44, 335–341.

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.

Nunnally, J.C. (1978), *Psychometric Theory*, 2 edition. McGraw-Hill, New York.

Osgood, C.E. (1959), "The representational model and relevant research methods." In *Trends in Content Analysis* (I. de Sola Pool, ed.), 33–88, University of Illinois Press, Urbana.

Park, H.M. and H.W. Jung (2003), "Evaluating interrater agreement with intraclass correlation coefficient in spice-based software process assessment." In *Proceedings of the Third International Conference On Quality Software*, 308–314, Dallas, TX, USA.

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

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

Quenouille, M.H. (1956), "Notes on bias in estimation." *Biometrika*, 61, 353–360.

Rowland, W.J. (1984), "The relationships among nuptial coloration, aggression, and courtship in male threespine sticklebacks." *Canadian Journal of Zoology*, 51, 453–466.

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

Savkov, A., J. Carroll, R. Koeling, and J. Cassell (2016), "Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus." *Lang Resources & Evaluation*, 50, 523–548.

Schuster, C. and A. von Eye (2001), "Models for ordinal agreement data." *Biometrical Journal*, 43, 795–808.

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

Shoukri, M.M. (2010), *Measures of Interobserver Agreement and Reliability*, 2 edition. CRC/Biostatistics Series, Chapman and Hall/CRC Press.

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

Stein, C.R., R.B. Devore, and B.E. Wojcik (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 *Proceedings of the Thirtieth Annual SAS® Users Group International Conference*, SAS Institute Inc., SAS Institute Inc, Cary, NC.

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

Tinsley, Howard E. A. and David J. Weiss (1975), "Interrater reliability and agreement of subjective judgments." *Journal of Counselling Psychology*, 22, 358–376.

Tinsley, Howard E. A. and David J. Weiss (2000), "Journal of counselling psychology." In *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (S. D. Tinsley, H. E. A. & Brown, ed.), 94–124, Academic Press, New York.

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

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

von Eye, A. and E.Y. Mun (2006), *Analyzing Rater Agreement: Manifest Variable Methods*, pap/cdr edition. Lawrence Erlbaum Associates.

Wongpakaran, Nahathai, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet (2013), "A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples." *BMC Medical Research Methodology*, 13, URL https://doi.org/10.1186/1471-2288-13-61.

Zhao, X., J.S. Liu, and K. Deng (2013), "Assumptions behind intercoder reliability indices." In *Communication Yearbook* (C.T. Salmon, ed.), volume 36, 419–480, Routledge.

List of Notations

AC_1, 65, 70, 118, 138, 141, 142, 150, 157
AC_1, 185
AC_2, 113, 118, 150, 161
AC_2, 185
ACM, 241
α, 138, 146, 148, 330–332
α_A, 149, 152
$\widehat{\alpha}_A$, 143, 149
$\widehat{\alpha}_K$, 70, 79, 118, 194
$\widehat{\alpha}_{K|k_0}$, 274
$\widehat{\alpha}^{(i)}_{K|k_0}$, 275
$\widehat{\alpha}_{K|i}$, 194
$\widehat{\kappa}_{K|i}$, 266
$\widehat{\alpha}^{\star}_{K|i}$, 194, 266
$\widehat{\alpha}'_K$, 112, 194
α_K, 65
$\alpha^{\star(i)}_{K|k_0}$, 275
$\alpha_{K|i}$, 194
α_S, 334

BP, 117
BRR, 202

$\overline{c}$, 332
C^n_N, 178
C^r_R, 178
CROSSTAB, 38

ε_{ik}, 262
ε_n, 65, 71, 79, 113, 118

f, 185, 193
FC_1, 179
FC_2, 179

$\widehat{\ }$, 61

`IAA`, 288
ICD-9-M, 50
I_r, 93

κ, 61, 173
κ_0, 173
κ_{1G}, 151, 152, 154, 157, 182
κ_{2G}, 182
κ_C, 61
κ_F, 180
$\kappa_{F|i}$, 193
$\kappa^{\star}_{F|i}$, 193
$\kappa_{G|i}$, 193
$\kappa^{\star}_{G|i}$, 193
$\widehat{\kappa}$, 173
$\widehat{\kappa}_{1G}$, 142, 158, 182, 185
$\widehat{\kappa}_2$, 66
$\widehat{\kappa}_{2G}$, 113, 161, 164, 182, 185, 259
$\widehat{\kappa}_{BP}$, 112, 188, 259
$\widehat{\kappa}_{BP|k_0}$, 258
$\widehat{\kappa}_{BP|i}$, 267
$\widehat{\kappa}^{\star}_{BP|i}$, 267
$\widehat{\kappa}_C$, 61, 69, 78, 83, 89, 105–107, 116, 187, 259, 267
$\widehat{\kappa}_{C|k_0}$, 255
$\widehat{\kappa}^{\star(i)}_{C|k_0}$, 276
$\widehat{\kappa}_{C|i}$, 195, 268
$\widehat{\kappa}^{\star}_{C|i}$, 195, 268

$\widehat{\kappa}'_{\mathrm{C}}$, 108
$\widehat{\kappa}_{\mathrm{F}}$, 77, 171, 181, 265
$\widehat{\kappa}_{\mathrm{F}|k_0}$, 272, 276
$\widehat{\kappa}^{(i)}_{\mathrm{F}|k_0}$, 273, 277
$\widehat{\kappa}^{\star(i)}_{\mathrm{F}|k_0}$, 273
$\widehat{\kappa}_{\mathrm{F}|i}$, 265
$\widehat{\kappa}^{\star}_{\mathrm{F}|i}$, 265
$\widehat{\kappa}'_{\mathrm{F}}$, 117
$\widehat{\kappa}^{(-g)}$, 201
$\widehat{\kappa}_{\mathrm{G}}$, 66, 70, 78, 119, 263
$\widehat{\kappa}_{\mathrm{G}|k_0}$, 251, 271
$\widehat{\kappa}^{(i)}_{\mathrm{G}|k_0}$, 271
$\widehat{\kappa}^{\star(i)}_{\mathrm{G}|k_0}$, 271
$\widehat{\kappa}_{\mathrm{G}|i}$, 264
$\widehat{\kappa}^{\star}_{\mathrm{G}|i}$, 264
$\widehat{\widehat{\kappa}}$, 173
$\widehat{\kappa}_q$, 70, 78, 117
$\widehat{\kappa}_{q|i}$, 194
$\widehat{\kappa}_{\mathrm{S}}$, 64, 69, 186, 259
$\widehat{\kappa}_{\mathrm{S}|k_0}$, 256
$\widehat{\kappa}_{\mathrm{S}}$, 111

λ, 93

M, 130
M_{kl}, 127

$\overline{n}$, 76
$\overline{n}_{\cdot k}$, 75, 76
N_{E}, 148
n_g, 77, 116
n_{gk}, 75–77, 116, 159
$N^{(\mathrm{E})}_k$, 147
N^{EH}_{kk}, 152
N^{EE}_{kk}, 152
N^{HE}_{kk}, 152
N^{HH}_{kk}, 152
n_{kl}, 68, 185
$N^{(\mathrm{H})}_{kl}$, 147
NLP, 48

n', 75, 79, 116, 118, 119, 158, 262

p_a, 58, 59, 69, 75, 89, 108, 116, 119, 142, 158, 164, 181, 195, 259
$p_{a|k_0}$, 251, 271
$p_{a|i}$, 192, 195, 264
p'_a, 65, 70, 79, 118, 189
$p^{\star}_a$, 113
$\overline{p}_{+k}$, 116
$\overline{p}_{\cdot k}$, 78
$\overline{p}_{l+}$, 187
$\overline{p}_{+k}$, 187
PC, 50
PC_2, 179
PCA, 31
P_a, 25, 157, 180
$P_{a|i}$, 25
P_e, 157, 180
P_{ik}, 180
P_{kl}, 148, 153
p_e, 60, 61, 69, 78, 89, 95, 108, 111, 113, 116, 117, 119, 142, 158, 164, 181, 187–189, 259
$p_{e|k_0}$, 251, 255, 256, 271
$p_{e|i}$, 193, 195, 264
$p^{\star}_{e|i}$, 265, 267, 268
$p^{\star\star}_{e|i}$, 265, 267, 268
p'_e, 159
p_{gk}, 77, 116, 159
$\overline{\pi}_k$, 187, 193, 195
$\overline{\pi}_{kl}$, 187, 188
$\overline{\pi}_{k+}$, 193, 195
$\overline{\pi}_{+l}$, 193, 195
$\widehat{\pi}_k$, 69, 70, 77, 79, 181
π_k, 69, 111, 113, 117, 119, 142, 158, 180, 185, 189
p_{ik}, 181
$\pi_{k|k_0}$, 249, 271
$\pi^{\star}_{k|k_0}$, 272
$\pi^{\star}_k$, 265
p_k, 245, 263
p_{k_0}, 271

p'_{k_0}, 271
$p^{(\mathrm{A})}_{k|\mathrm{H}}$, 143, 148
p_{kk}, 142
p_{kl}, 68, 185
$p_{kl|k_0}$, 247
$p^{(k_0)}_{kl}$, 247
p'_{kl}, 161
p_{k+}, 68, 142, 185, 189
p_{+k}, 68, 142, 185
p_{+l}, 189

r, 74
$\overline{r}$, 75, 118, 334
$\overline{r}_{\cdot k}$, 74, 75
$\overline{r}^{\star}$, 332
$\mathcal{R}$, 153
R_i, 180
R_{ik}, 25, 180
$R^{\star}_{ik}$, 180
RCM, 242
r_i, 74, 75, 116, 118, 158, 181, 262
r_{ik}, 74, 75, 116, 158, 181, 262
$r^{\star}_{ik}$, 116, 118, 119, 164, 181, 263

S, 93
s_i, 331
s_k, 159
s^2_k, 78
s_{kl}, 116
s_n, 176, 177
SRS, 201
$s^{\star}_r$, 176, 177, 184
s_{T}, 332

T_w, 161, 186
$T^{\star}_w$, 263

U, 130
$\mathcal{U}_{\mathcal{R}}$, 176
$\mathcal{U}_{\mathcal{S}}$, 176

$v(\widehat{\alpha}_{\mathrm{K}})$, 189, 194, 266
$v(\widehat{\alpha}_{\mathrm{K}|k_0})$, 275
$\overline{v}$, 332
$v(\widehat{\kappa}_{\mathrm{BP}})$, 188, 194, 267
$v(\widehat{\kappa}_{\mathrm{C}})$, 187, 195, 268
$v(\widehat{\kappa}_{\mathrm{C}|k_0})$, 276
$v(\widehat{\kappa}_{\mathrm{F}})$, 193, 265
$v(\widehat{\kappa}_{\mathrm{F}|k_0})$, 273
$v(\widehat{\kappa}_{\mathrm{G}})$, 185, 193, 264
$v(\widehat{\kappa}_{\mathrm{G}|k_0})$, 271
$v(\widehat{\kappa}_{\mathrm{S}})$, 186
$v(p_a)$, 190, 194

RDIST, 39
WDF, 40
w_{kl}, 90, 108, 127, 129, 130, 161, 262

x_k, 108, 128–130
x_l, 111

Author Index

Afifi, A.A., 206
Agresti, A., 11, 175
Aickin, M., 33, 138, 140, 141, 143, 146, 148, 168
Alpert, R., 58
Altman, D.G., 223
Andreasen, N.C., 140
Axelson, R.D., 17

Benini, R., 58
Bennett, E.M., 58
Berry, K.J., 100, 102, 172
Bishop, J., 97, 364
Brennan, R.L., 58, 70, 78, 95, 100, 111, 117, 194, 364
Bromet, E.J., 56
Byrt, T., 97, 364

Cantor, A.B., 206
Carletta, J., 8
Carlin, J.B., 97, 364
Carmines, E.G., 29, 331
Chen, S., 156
Chow, E., 338
Christakis, M., 338
Cicchetti, D.V., 82, 84, 86, 97, 142
Ciccone, G., 8
Cochran, W.G., 176
Cohen, J., 26, 54, 58, 60–62, 84, 89, 90, 100, 102, 108, 121, 150, 152, 172, 187, 317, 350, 363–365
Conger, A.J., 39, 73, 74, 76, 77, 87, 96, 115, 116, 159, 171, 195, 201, 350
Craig, T.J., 56
Cronbach, L.J., 26, 29, 330

Deng, K., 8, 98, 136
Devore, R.B., 10, 50, 51
Donovan, A., 338

Eckes, T., 7, 10, 11
Efron, B., 202
Everitt, B.S., 187, 222, 317

Feinstein, A.R., 82, 84, 86, 97, 142
Fenning, S., 56
Finkelstein, J., 338
Finn, R.H., 359
Flack, V.F., 206
Fleiss, J.L., 29, 40, 69, 74, 76, 77, 87, 95, 117, 158, 180, 185, 187, 222, 241, 317, 351, 356, 365
Ford, M., 338

Gaviani, P., 8
Goldstein, A.C., 58
Goodman, L.A., 93
Grove, W.M., 140
Guilford, J.P., 58, 66
Guttman, L., 58
Gwet, K.L., 33, 66, 74, 83, 85, 97, 112, 118, 138, 140, 141, 143, 156, 168, 184, 185, 192, 204, 286, 310, 351

Hayes, A.F., 64

Holden, L., 338
Holley, J.W., 58, 66
Holsti, O.R., 56
Hripcsak, G., 50
Hubert, L., 76

Janson, H., 100, 102, 115
Janson, S., 58, 93
Jung, H.W., 8, 199

Keller, M.B., 140
Kendall, M., 204
Khan, L., 338
Klein, D., 365
Koch, G., 62, 221, 222, 352
Kolmogorov, A.N., 11
Kottner, J., 5
Kraemer, H.C., 33, 85, 172, 201
Kreiter, C.D., 17
Krippendorff, K., 5, 8, 63, 64, 79, 112, 117, 129, 194, 290, 291, 293
Kruskal, W.H., 93
Kuder, G.F., 333

Lachenbruch, P.A., 206
Landis, J.R., 62, 221, 222, 351, 352
Leigh, L.E., 93–95
Leone, M.A., 8
Light, R.J., 28, 29, 95, 241, 278
Likert, R., 331
Lindsay, B.G., 156
Liu, J.S., 8, 98, 136

Métivier, J.P., 287, 288, 290, 291
Markatou, M., 156
Mathet, Y., 287, 288, 290, 291
Maxwell, A.E., 58
McCarthy, P.J., 202
McDonald-Scott, P., 140
McIver, J.P., 331
Metropolis, N., 225
Mielke, P.W., 100, 102, 172
Mitera, G., 338

Mun, E.Y., 175

Nee, J.C.M., 351
Neyman, J., 176
Noda, A., 33, 85, 172
Nunnally, J.C., 333

Olsson, U., 100, 102, 115
Osgood, C.E., 56

Park, H.M., 199
Perreault, W.D., 93–95
Peryakoil, V.S., 33, 85, 172
Prediger, D.J., 58, 70, 78, 95, 100, 111, 117, 194, 364
Probyn, L., 338

Quenouille, M.H., 202

Ray, S., 156
Richardson, M.W., 333
Rothschild, A. S., 50
Rowland, W.J., 74
Rubenstein, J., 338

Särndal, C.E., 176
Savkov, A., 49
Schouten, H.J.A., 206
Schuster C., 11
Scott, W.A., 8, 63, 69, 77, 97, 100, 172, 186, 317
Shapiro, R.W., 140
Shoukri, M.M., 175
Sim, J., 39, 67
Stein, C.R., 10, 50, 51
Streiner, D.L., 5
Stuart, A., 204
Swensson, B., 176

Tanenberg-Karant, M., 56
Tanner, M.A., 11
Tinsley, H.E.A., 4
Traub, R.E., 29
Tukey, J.W., 202

Ulam, S., 225

Vegelius, J., 58, 93
Verzani, J., 350
von Eye, A., 11, 175

Wedding, D., 66
Weiss, D.J., 4
Widlöcher, A., 287, 288, 290, 291
Wojcik, B.E., 50, 51
Wojcik, B.F., 10
Wongpakaran, N., 66
Wongpakaran, T., 66
Wretman, J., 176
Wright, C.C., 39, 67

Yang, K., 156
Yee, A., 338
Young, M.A., 11

Zeller, R.A., 29
Zeng, L., 338
Zhang, L., 338
Zhao, X., 8, 98, 136

Subject Index

Absolute Category Membership, 241
AC_1 Coefficient, 65, 70, 78, 113, 118, 142, 150, 151
  Multiple-Rater, 157
  Variance, 185, 192
AC_2 Coefficient, 113, 118, 160
  Conditional Reliability, 271
  Interval/Ordinal Ratings, 160
  Multiple-Rater, 164
  Variance, 185, 192
Affiliation, 180
  Weight, 108
Agreement
  for Cause, 142
  Full/Partial, 108
Aickin's Alpha, 143
Alpha
  Aickin, 143
  Cronbach, 331
  Krippendorff, 64, 79, 112
  Standardized Cronbach, 334
Alternate Hypothesis, 304
Alternative Hypothesis, 304

BAK Coefficient, 97
Balanced Repeated Replication, 202
Benchmark, 34, 62, 220
  Altman, 222
  Fleiss, 222
  Landis and Koch, 222
  Probabilities, 234
  Scale, 62
Benchmarking, 34, 220
Benchmarking Model, 234
Bipolar Weights, 130
Brennan-Prediger Coefficient
  Conditional, 258, 275
  Three Raters or More, 78, 117
  Two Raters, 70
  Validity Coefficient, 267
  Variance for Multiple Raters, 194
  Variance for Two Raters, 188
  Weighted, 112

Carroll, J., 290
Cassell, J., 290
Categorization, 289
Chance Agreement, 13, 26, 57, 60, 62
  Correction, 60
  Percent, 60
Chance-Corrected Coefficient, 57, 115
Circular Weights, 129
Cohen's Kappa
  Variance for 2 Raters, 187
Complex Sampling, 179
Conditional
  AC_2, 251, 271
  Brennan-Prediger, 258, 275
  Conger, 276
  Fleiss, 272
  Inference, 182
  Kappa, 255
  Krippendorff, 273
  Krippendorff's Alpha, 256

Percent Agreement, 242, 250
Pi, 256
Probability, 148
Reliability, 245
Standard Error, 190
Use probability, 252
Validity, 245
Conditioned Event, 243
Conditioning, 242
Conditioning Event, 242
Confidence Bound, 231
Lower, 231
Upper, 231
Confidence Interval, 207
Conger's Kappa, 77, 116
Conditional Reliability, 276
Validity Coefficient, 267
Variance, 195, 268, 276
Contingency Table, 38
Continuum, 291
Critical Value, 225, 306

E-subjects, 147
Error Margin, 206, 231
Estimand, 32, 173
Estimate, 173
Estimator, 32, 173
Expected Chance Agreement, 60

Fleiss' Kappa
Coefficient, 76, 117
Conditional Reliability, 272
Validity Coefficient, 265
Variance, 193, 265, 273
For-cause agreement, 146
Full Agreement, 89, 108, 110
Fully-Crossed, 179

G-Index, 66, 188
Gold Standard, 30, 241
Gwet's AC_1
Coefficient, 65, 113
Construct, 151
Estimand, 151
Multiple-Rater, 78, 157, 158
Probabilistic Model, 153
Variance, 185
Weighted, 113
Gwet's AC_2
Coefficient, 118, 160
Multiple-Rater, 118, 164
Variance, 193

H-subjects, 147
Hypothesis
Alternate/Alternative, 304
Null/Research/Statistical, 304

Identity Weights, 110, 164
Inference, 31
Statistical, 31
Influence Analysis, 314
Inter-Annotator
Agreement, 288
Reliability, 8
Inter-group Reliability, 337
Inter-Rater Reliability
Applications, 7
Definition, 4, 10
Parameter, 32
Sample Size Calculation, 211, 213, 214, 216
Type of, 27
Intercoder Reliability, 8
Internal Consistency, 29, 330
Interval
Data, 102
Estimation, 182
Intra-Rater Reliability, 6, 287
Item-Total Correlation, 336

Jackknife, 201

k-subject, 245
Kappa, 58, 105

Conger, 77
Fleiss, 76
Marginal Homogeneity Dependency, 86
Multiple raters, 72
Multiple-level scale, 72
Paradoxes, 82
Trait Prevalence Dependency, 83
Koeling, R., 290
KR-20, 333
Krippendorff's Alpha, 70, 79, 112, 117
Conditional Reliability, 273
Validity Coefficient, 266
Variance, 266, 274

Left-tailed Test, 305
Level of Confidence, 182
Linear Weights, 90, 128
Linearization Method, 310

Margin of Error, 206

Nominal Scale, 25
Nondeterministic, 141
Null Hypothesis, 304, 306

One-tailed Test, 305
Ordinal
Data, 102
Scale, 25
Weights, 127

P-value, 306
PABAK Coefficient, 97
Paradox, 82
Parameter, 173
Partial Agreement, 26, 89, 102, 108, 110, 180
Partially Crossed, 179
Percent Agreement, 56, 59, 75
Variance, 189
Weighted, 89
Perreault & Leigh Index, 93
Pi Coefficient, 69
Pivot Statistic, 305
Point Estimation, 182
Principal Component Analysis, 31
Psychometrics, 330

Quadratic Weights, 90, 108

Radical Weights, 129
Random Rating, 224, 225
Rater Sample, 176
Rater Sampling, 15
Ratings
Raw Representation, 103
Vector Representation, 103
Ratio Data, 102
Ratio Weights, 129
Relative Category Membership, 242
Reliability, 4
Inter-Annotator, 8
Inter-Rater, 4
Intercoder, 8
Internal Consistency, 29
Intra-Rater/Test-Retest, 6
versus Validity, 30
Reliability Measure, 259
Replicate Sample, 202
Replication Methods, 202
Research Hypothesis, 304
Right-tailed Test, 305

Sample Size Calculation
AC_1 Coefficient, 213
Brennan-Prediger Coefficient, 214
Fleiss Generalized Kappa, 216
Sampling
Distribution, 173
Error, 177
Fraction, 185, 201
Plan, 16
Rater/Subject Population, 15
Savkov, A., 290
Scoring Rubric, 18

Scott's Pi
Definition, 63, 69
Variance, 186
Weighted, 111
Significance Level, 306
Simple Ordinal Weights, 127
Simple Random Sampling, 178
Standard Error, 14
Standardized Cronbach's Alpha, 334
Statistical
Hypothesis, 304, 306
Independence, 63
Inference, 61, 182
Noise, 304
Significance, 304
Statistically Significant, 221
Subject Sample, 15, 176
Subject Sampling, 15

Test of Hypothesis, 182
Test Statistic, 305
Test-Retest Reliability, 6
Text Annotation, 287
Textbook Cases, 140
Total Disagreement, 102
Total Variance
Calculation, 204
Definition, 203
Two-Tailed Test, 305

Unconditional
Inference, 182
Reliability, 245
Unidimensionality, 335
Unitization Analysis Zone, 292
Unitizing, 289

Validity, 245
Analysis, 243
Measure, 259
Validity Coefficient
AC_2, 263
Brennan-Prediger, 267
Conger, 267
Fleiss, 265
Krippendorff, 266
Unconditional, 259
Variable, 19
Variance
AC_1 - 2 Raters, 185
AC_2 - 2 Raters, 185
AC_2 - 3 Raters+, 193
Brennan-Prediger - 2 Raters, 188
Brennan-Prediger - 3 Raters+, 194
Cohen's Kappa - 2 Raters, 187
Conditional Conger, 276
Conditional Fleiss Reliability, 273
Conditional Krippendorff, 274
Conger's Kappa, 195
Definition, 184
Estimators, 184
Fleiss' Kappa - 3 Raters+, 193
Krippendorff - 2 Raters, 188
Krippendorff - 3 Raters+, 194
Percent Agreement - 2 Raters, 189
Scott's Pi - 2 Raters, 186
Three Raters or More, 192
Unconditional, 203

Weighted
Kappa, 89, 108
Percent Agreement, 89
Weighting, 89
Weights
Bipolar, 130, 132
Circular, 129, 132
Custom, 122
Identity, 131
Linear, 90, 91, 128, 132
Ordinal, 127, 132
Quadratic, 90, 91, 108, 132
Radical, 129, 131
Ratio, 129, 132
Use for Defining Agreement, 121

Handbook of Inter-Rater Reliability, 5th Ed.

Vol 1: Analysis of Categorical Ratings

Inter-rater reliability assessment has become an essential component in the process of evaluating the quality of experimental data in almost all fields of research. The 4th edition of the *Handbook of Inter-Rater Reliability* covered the analysis of categorical and quantitative ratings in a single volume. In response to comments on previous editions, the current 5th edition is released in 2 volumes. Volume 2 is devoted to the analysis of quantitative ratings, whereas the current volume 1 focuses on the analysis of categorical ratings.

Here is the link to the webpage of volume 1, where you can find, a link to the errata page, some example workbooks, and other datasets:

www.agreestat.com/books/cac5/

Here are a few topics that are new to the 5th edition:

- Chapter 2 describes various methods for setting up your dataset of ratings before analysis.
- New sample size calculation procedures for chance-corrected agreement coefficients are presented in chapter 6.
- Several new techniques for analyzing categorical ratings are described in chapter 9. Among these are the inter-annotator agreement, useful in Natural Language Processing, the testing of the difference of 2 agreement coefficients for statistical significance, and many others.

About the Author

Kilem L. Gwet, Ph.D.
Statistical consultant, mathematical statistician, researcher, and instructor. Over 20 years of experience in various industries, and several publications in peer-reviewed journals.

$42.99
ISBN 978-1-7923-5463-2
54299>
9 781792 354632

AgreeStat Analytics
PO Box 2696
Gaithersburg, Maryland 20886-2696 – USA