

Contents

<u>chap: acknow</u> Acknowledgment	xvii
Part I: Preliminaries	1
1 Introduction July 17, 2019	2
1.1 <i>What is Inter-Rater Reliability?</i>	4
1.2 <i>Scope and Design of Inter-Rater Reliability Experiments</i>	10
1.2.1 <i>Scope of the Investigation</i>	11
1.2.2 <i>Experimental Design</i>	14
1.3 <i>Scoring of Subjects/Items</i>	17
1.4 <i>Formulation of Agreement Coefficients</i>	23
1.4.1 <i>Nominal Ratings</i>	24
1.4.2 <i>Ordinal Ratings</i>	25
1.4.3 <i>Interval and Ratio Ratings</i>	26
1.5 <i>Different Reliability Types</i>	26
1.5.1 <i>Undefined Raters and Subjects</i>	27
1.5.2 <i>Conditional Reliability</i>	27
1.5.3 <i>Reliability as Internal Consistency</i>	28
1.5.4 <i>Reliability versus validity</i>	28
1.5.5 <i>Multivariate Inter-Rater Reliability</i>	29
1.6 <i>Statistical Inference</i>	30
1.7 <i>Book's Structure</i>	31
1.8 <i>Choosing the Right Method</i>	33
2 Setting Up Databases of Ratings for Analysis	36
2.1 <i>Introduction</i>	37
2.2 <i>Dealing with the Notions of Subject and Characteristic</i>	41
2.3 <i>Dealing with the Notion of Agreement in a Multiple-Level Process</i>	44
2.4 <i>Multiple Ratings per Rater and per Subject</i>	46
2.5 <i>Dealing with the Notion of Rater</i>	47
Part II: Chance-Corrected Agreement Coefficients	50

3	Agreement Coefficients for Nominal Ratings: A Review	51
3.1	<i>The Problem</i>	53
3.2	<i>Agreement for two Raters and two Categories</i>	56
3.2.1	<i>Cohen's Kappa Definition</i>	57
3.2.2	<i>What is Chance Agreement?</i>	59
3.2.3	<i>Scott's Pi Coefficient</i>	60
3.2.4	<i>Krippendorff's Alpha Coefficient</i>	61
3.2.5	<i>Gwet's AC_1 Coefficient</i>	62
3.2.6	<i>G-Index</i>	63
3.3	<i>Agreement for 2 Raters and q Response Categories ($q \geq 3$)</i>	63
3.4	<i>Kappa for r Raters and q Categories ($r > 2$ and $q > 2$)</i>	69
3.4.1	<i>Defining Agreement Among 3 Raters or More</i>	70
3.4.2	<i>Computing Inter-Rater Reliability</i>	71
3.5	<i>Kappa Coefficient and its Paradoxes</i>	79
3.5.1	<i>Kappa's Dependency on Trait Prevalence</i>	80
3.5.2	<i>Kappa's Dependency on Marginal Homogeneity</i>	82
3.5.3	<i>Paradoxes in Multiple-Rater Studies and Other Agreement Coefficients</i>	83
3.6	<i>Weighted Kappa: A Review</i>	85
3.7	<i>More Alternative Agreement Coefficients</i>	89
3.8	<i>Concluding Remarks</i>	93
4	Agreement Coefficients for Ordinal, Interval and Ratio Data	97
4.1	<i>Overview</i>	99
4.2	<i>Generalized Kappa for Two Raters</i>	100
4.2.1	<i>Calculating the Kappa Coefficient</i>	102
4.2.2	<i>Kappa: a Function of Squared Euclidean Distances</i>	103
4.3	<i>Agreement Coefficients for Interval Data: $2 \times q$ Tables</i>	107
4.4	<i>Agreement Coefficients for Interval Data and Multiple Raters</i>	111
4.4.1	<i>Defining the Multiple-Rater Agreement Coefficient</i>	111
4.4.2	<i>Formulating the Multiple-Rater Agreement Coefficient</i>	112
4.5	<i>On the Use of Weights for Defining Agreement</i>	117
4.5.1	<i>Defining Agreement When Two Measurement Scales Are Used</i>	117
4.5.2	<i>Defining Agreement When Raters Assign Some Subjects to Multiple Categories</i>	121
4.6	<i>More Weighting Options for Agreement Coefficients</i>	123
4.7	<i>Concluding Remarks</i>	130
5	Constructing Agreement Coefficients: AC_1 and Aickin's α	134
5.1	<i>Overview</i>	135
5.2	<i>Gwet's AC_1 and Aickin's α for two Raters</i>	137
5.2.1	<i>The AC_1 Statistic</i>	137

5.2.2	<i>Aickin's α-Statistic</i>	138
5.2.3	<i>Example</i>	139
5.3	<i>Aickin's Theory</i>	141
5.3.1	<i>Aickin's Probability Model</i>	143
5.4	<i>Gwet's Theory</i>	144
5.4.1	<i>The Probabilistic Model</i>	147
5.4.2	<i>Quantifying the Probability $P(\mathcal{R})$ of Selecting an H-Subject</i>	149
5.5	AC₁ <i>for Multiple Raters</i>	152
5.6	AC₂ : <i>the AC₁ Coefficient for Ordinal and Interval Data</i>	155
5.6.1	<i>AC₂ for Interval Data and two Raters</i>	155
5.6.2	<i>AC₂ for Interval Data and for three Raters or More</i>	157
5.7	<i>Concluding Remarks</i>	161
6	Agreement Coefficients and Statistical Inference	164
6.1	<i>Introduction</i>	166
6.1.1	<i>The Problem</i>	166
6.1.2	<i>The Challenge and How to Go About It</i>	168
6.2	<i>Finite Population Inference</i>	171
6.2.1	<i>The Notion of Sample</i>	172
6.2.2	<i>Assigning Raters to Subjects</i>	174
6.2.3	<i>The Notion of Parameter in Finite Population Inference</i>	175
6.2.4	<i>The Nature of Statistical Inference</i>	177
6.2.5	<i>Independence of Subjects and Its Impact on Statistical Inference</i>	178
6.3	<i>Conditional Inference</i>	178
6.3.1	<i>Inference Conditionally Upon the Rater Sample</i>	179
6.3.2	<i>Inference Conditionally Upon the Subject Sample</i>	194
6.4	<i>Total Variance</i>	197
6.4.1	<i>Definitional Equation of Total Variance</i>	197
6.4.2	<i>Computational Equation of Total Variance</i>	199
6.5	<i>Sample Size Estimation</i>	201
6.5.1	<i>The Mechanics of Sample Size Calculation</i>	202
6.5.2	<i>Applications</i>	203
6.5.3	<i>Optimal Number of Subjects for the Percent Agreement</i>	205
6.5.4	<i>Optimal Number of Subjects for Gwet's AC₁ Coefficient</i>	207
6.5.5	<i>Optimal Number of Subjects for Brennan-Prediger Coefficient</i>	209
6.5.6	<i>Optimal Number of Subjects for Fleiss' Generalized Kappa</i>	210
6.6	<i>Concluding Remarks</i>	211
7	Benchmarking Inter-Rater Reliability Coefficients	213
7.1	<i>Overview</i>	214
7.2	<i>Benchmarking the Agreement Coefficient</i>	215
7.2.1	<i>Existing Benchmarks</i>	216

7.2.2	<i>Agreement Coefficient's Sources of Variation</i>	218
7.3	<i>The Proposed Benchmarking Method</i>	223
7.3.1	<i>The Method</i>	224
7.3.2	<i>The Benchmark Probabilities and the Interpretation of the New Method</i>	228
7.4	<i>Concluding Remarks</i>	231
Part III: Miscellaneous Topics on the Analysis of Inter-Rater Reliability Experiments		233
8	Inter-Rater Reliability: Conditional Analysis	235
8.1	<i>Overview</i>	236
8.2	<i>Two-Rater Conditional Agreement for ACM Studies</i>	240
8.2.1	<i>Basic Conditional Probabilities for ACM Studies</i>	240
8.2.2	<i>Conditional Reliability for 2 Raters in ACM Reliability Studies</i>	244
8.2.3	<i>Unconditional Validity Coefficient for 2 Raters in ACM Studies</i>	252
8.2.4	<i>Concluding Remarks on Section ^{sec:12acm2}8.2</i>	254
8.3	<i>Multiple-Rater Coefficients for ACM Studies</i>	255
8.3.1	<i>Validity Coefficients for three Raters or More</i>	255
8.3.2	<i>Conditional Agreement Coefficients for three Raters or More</i>	262
8.4	<i>Conditional Agreement in RCM Studies</i>	270
8.5	<i>Concluding Remarks</i>	274
9	Analysis of Nominal-Scale Inter-Rater Reliability Data	276
9.1	<i>Overview</i>	278
9.2	<i>Inter-Rater Reliability Coefficients under the PC₂ Design</i>	278
9.2.1	<i>The FC₁ and PC₂ Designs</i>	278
9.2.2	<i>Calculating Agreement Coefficients and their Variances under the PC₂ Design</i>	280
9.3	<i>Testing the Difference of Agreement Coefficients</i>	284
9.3.1	<i>Testing Uncorrelated Agreement Coefficients for Statistical Significance</i>	284
9.3.2	<i>Testing Correlated Agreement Coefficients for Statistical Significance</i>	286
9.4	<i>Influence Analysis</i>	288
9.5	<i>Multivariate Agreement Coefficients for Nominal Ratings</i>	289
9.6	<i>Intra-Rater Reliability</i>	289
9.7	<i>Inter-Annotator Reliability in Natural Language Processing</i>	292
9.8	<i>Inter-Rater Reliability between Groups of Raters</i>	292
9.9	<i>Cronbach's Alpha</i>	294
9.9.1	<i>Defining Cronbach's Alpha</i>	295
9.9.2	<i>How Does Cronbach's Alpha Evaluate Internal Consistency?</i>	296

9.9.3	<i>Use of Cronbach's Alpha</i>	299
Part IV:	Appendices	302
A	Data Tables	303
B	Software Solutions	315
B.1	<i>The R Software</i>	315
B.2	<i>AgreeStat for Excel</i>	327
B.3	<i>Online Calculators</i>	328
B.4	<i>SAS Software</i>	329
B.5	<i>SPSS & STATA</i>	329
B.6	<i>Concluding Remarks</i>	330
C	Sample Size Calculations	331
	Bibliography	340
	List of Notations	350
	Author Index	353
	Subject Index	355
