

Limitations of Kappa and its Extensions

N. Clay Mann, PhD, MS



Limitations to Kappa

- Affected by base rates of diagnoses.
 - Can't easily compare across studies that have different base rates, either in the population, or in the reliability study.
- Chance agreement is a problem?
 - When the null hypothesis of rater independence is not met (which is most of the time), the estimate of chance agreement is inaccurate and possibly inappropriate).

Standard Kappa

Rater 2

		Rater 2		
		Pres	Abse	
Rater 1	Pres	50	15	65
	Abes	15	20	35
		65	35	100

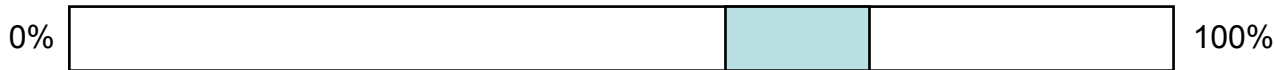
Observed Agreement = $50 + 20 / 100 = 70\%$



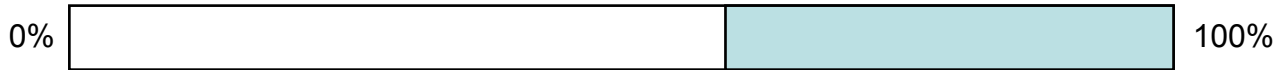
Agreement Expected by Chance = $(65\% \times 65) + (35\% \times 35) / 100 = 54.5\%$



Actual Agreement Beyond Chance = $70\% - 54.5\% = 15.5\%$



Potential Agreement Beyond Chance = $100\% - 54.5\% = 45.5\%$



$Kappa = \text{Agreement Beyond Chance} / \text{Potential Agreement Beyond Chance} = 15.5\% / 45.5\% = 0.34$

High Prevalence

Rater 2

		Rater 2		
		Pres	Abse	
Rater 1	Pres	65	15	75
	Abes	15	5	15
		80	20	100

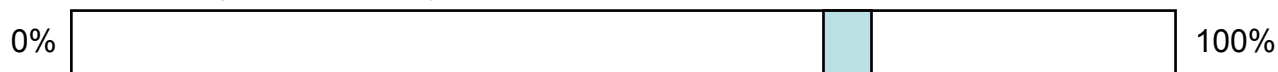
Observed Agreement = $65 + 5 / 100 = 70\%$



Agreement Expected by Chance = $(80\% \times 75) + (20\% \times 15) / 100 = 63\%$



Actual Agreement Beyond Chance = $70\% - 63\% = 7\%$



Potential Agreement Beyond Chance = $100\% - 63\% = 37\%$



*Kappa = Agreement Beyond Chance / Potential Agreement
Beyond Chance = $7\% / 37\% = 0.18$*

Rater Bias

Rater 2

		Rater 2		
		Pres	Abse	
Rater 1	Pres	50	25	75
	Abes	5	20	25
		55	45	100

Observed Agreement = $50 + 20 / 100 = 70\%$



Agreement Expected by Chance = $(55\% \times 75) + (45\% \times 25) / 100 = 52.5\%$



Actual Agreement Beyond Chance = $70\% - 52.5\% = 17.5\%$



Potential Agreement Beyond Chance = $100\% - 52.5\% = 47.5\%$



*Kappa = Agreement Beyond Chance / Potential Agreement
Beyond Chance = $17.5\% / 47.5\% = 0.37$*



What to do with Polytomous Categories?

Two raters classify n cases into k mutually exclusive categories.

		Rater 2						
Category		1	2	.	.	j	k	$\sum_j n_{ij}$
Rater 1	1	n_{11}	n_{12}					$n_{1.}$
	2	n_{21}	n_{22}					$n_{2.}$
	.							
	i				n_{ii}	n_{ij}		$n_{i.}$
	.							
	k							
	$\sum_i n_{ij}$	$n_{.1}$	$n_{.2}$			$n_{.j}$		$n_{..}$

n_{ij} = number of cases falling into cell = freq of joint event ij
 $n_{..}$ = total number of cases
 $p_{ij} = n_{ij} / n_{..}$ = proportion of cases falling into particular cell.

Reliability by Percentage Agreement = $\sum_i p_{ii} = 1/n \sum_i n_{ii}$

Chi-Square Test of Association as Proposed Solution

- Can perform a Chi-Square Test of Association to test null hypothesis that the two raters' judgments are independent.
- To reject independence, show that observed agreement departs from what would be expected by chance alone. $\text{Chi-Square} = \sum_{\text{cells}} (\text{Observed} - \text{Expected})^2 / \text{Expected}$
- Problem: In example below, we have a perfect association between the Raters with zero agreement. Chi-Square is a test of Association, not Agreement. It is sensitive to any departure from chance agreement, even when the dependency between the raters' judgments involves perfect non-agreement.
- So, we cannot use Chi-Square Test to assess agreement between raters.

Rater 2

		Pres	Abse	Other	
Rater 1	Pres	0	5	0	5
	Abse	0	0	5	5
	Other	5	0	0	5
		5	5	5	n=15

Kappa Coefficient (Cohen, 1960)

- High reliability requires that the frequencies along the diagonal should be > chance and off diagonal frequencies should be < chance.
- Use marginal frequencies/probabilities to estimate chance agreement.

Proportion agreement observed, $p_o = \sum_i p_{ii} = 1/n \sum_i n_{ii}$

Proportion agreement expected by chance, $p_c = \sum_i p_{i\cdot} \times p_{\cdot i}$

		Rater 2			$n_{i\cdot}$	$p_{i\cdot}$
		Pres	Abse	Other		
R a t e r 1	Pres	106 .53 (78) .39	10	4	120	.6
	Abse	22	28 .14 (15) .075	10	60	.3
	Other	2	12	6 .03 (2) .01	20	.1
1		$n_{\cdot j}$	130	50	20	200
		$p_{\cdot j}$.65	.25	.1	1
		$p_{i\cdot} \times p_{\cdot i}$.39	.075	.01	

$$\text{Kappa, } K = \frac{p_o - p_c}{1 - p_c}$$

$$p_o = .53 + .14 + .03 = .7$$

$$p_c = .39 + .075 + .01 = .475$$

$$K = \frac{.7 - .475}{1 - .475} = .429$$

K = 1, perfect agreement
 K = 0, chance agreement
 K < 0, agreement worse than chance.

Weighted Kappa Coefficient

Can assign weights, w_{ij} , to classification errors according to their seriousness using some ratio scale of weights.

$$K_w = \frac{p_{o(w)} - p_{c(w)}}{1 - p_{c(w)}}$$

		Rater 2			$n_{i\cdot}$	$p_{i\cdot}$
		Pres	Abse	Other		
R a t e r 1	Pres	106 .53 .39 0	10 .05 .15 1	4 .02 .06 4	120	.6
	Abse	22 .11 .195 1	28 .14 .075 0	10 .05 .03 1	60	.3
	Other	2 .01 .065 4	12 .06 .025 1	6 .03 .01 0	20	.1
$n_{\cdot j}$		130	50	20	200	
$p_{\cdot j}$.65	.25	.1		1.0



Weighted Kappa

So ... for polytomous categories

- Weighted Kappa is equivalent to un-weighted Kappa when assume equal weights

Problem: How to assign weights? Standard is to use the square of the deviation from perfect

Set your weights using something like:

kap obs1 obs2, tab wgt(w) (for linear weights)

kap obs1 obs2, tab wgt(w²) (for quadratic weights)

Reference:

Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991. pp403-409.



Weighted Kappa

So ... for ordinal categories

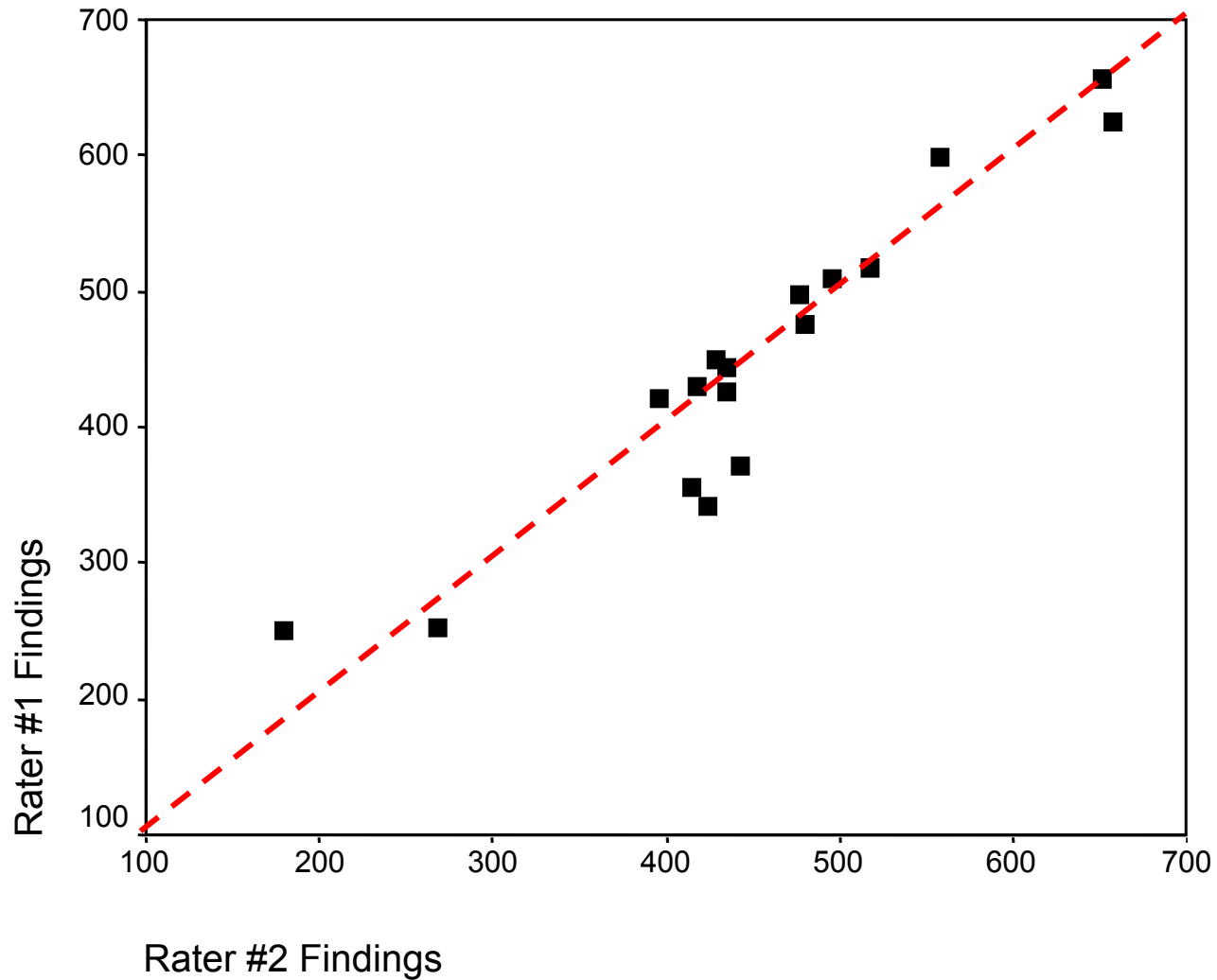
- Weighted Kappa with quadratic weights and uniform marginal distributions is exactly equivalent to a Intraclass Correlation Coefficient.
- Primarily associated with a repeated-measures ANOVA...used to measure agreement for continuous data...where the Rater is the within-subject factor.



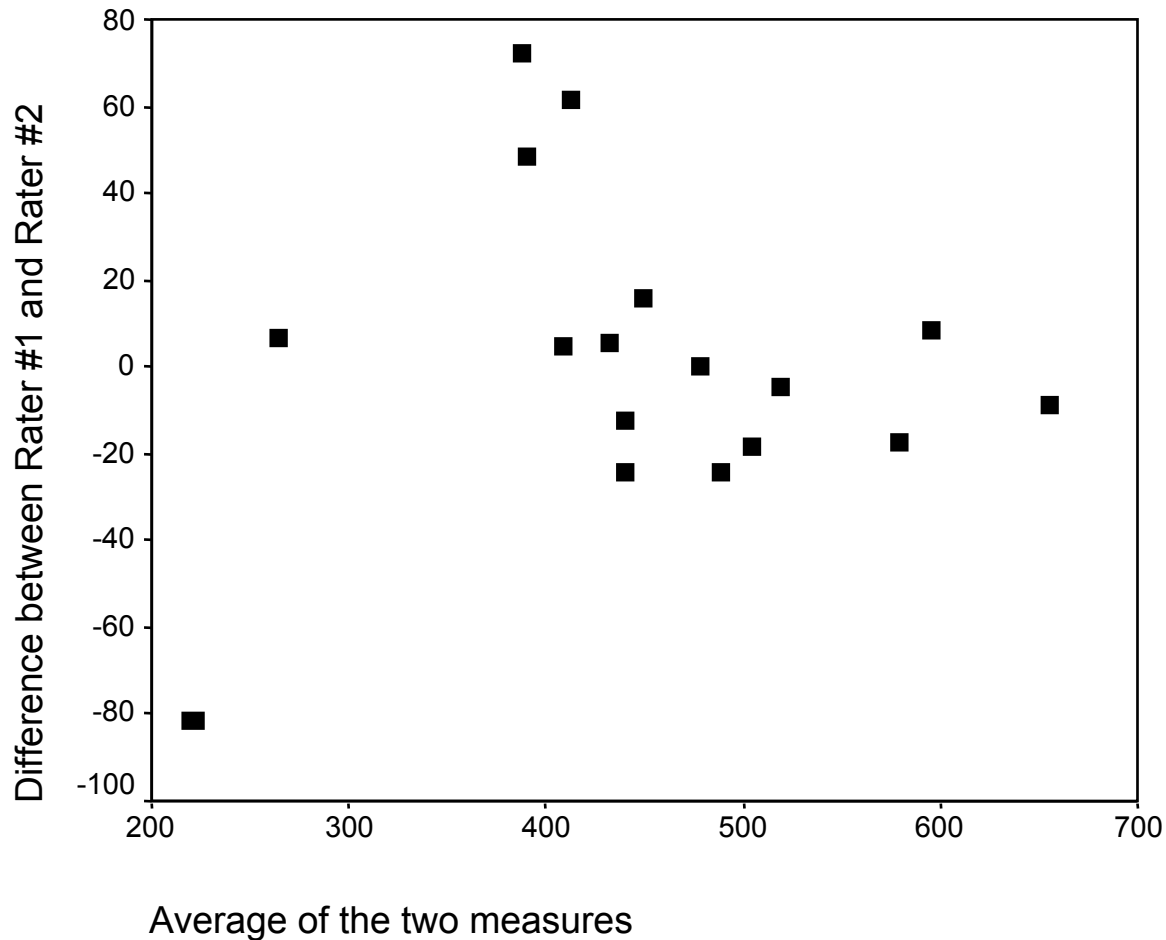
There is a Valuable Method for
Demonstrating Agreement among
Continuous Variables

Bland, Altman Charts

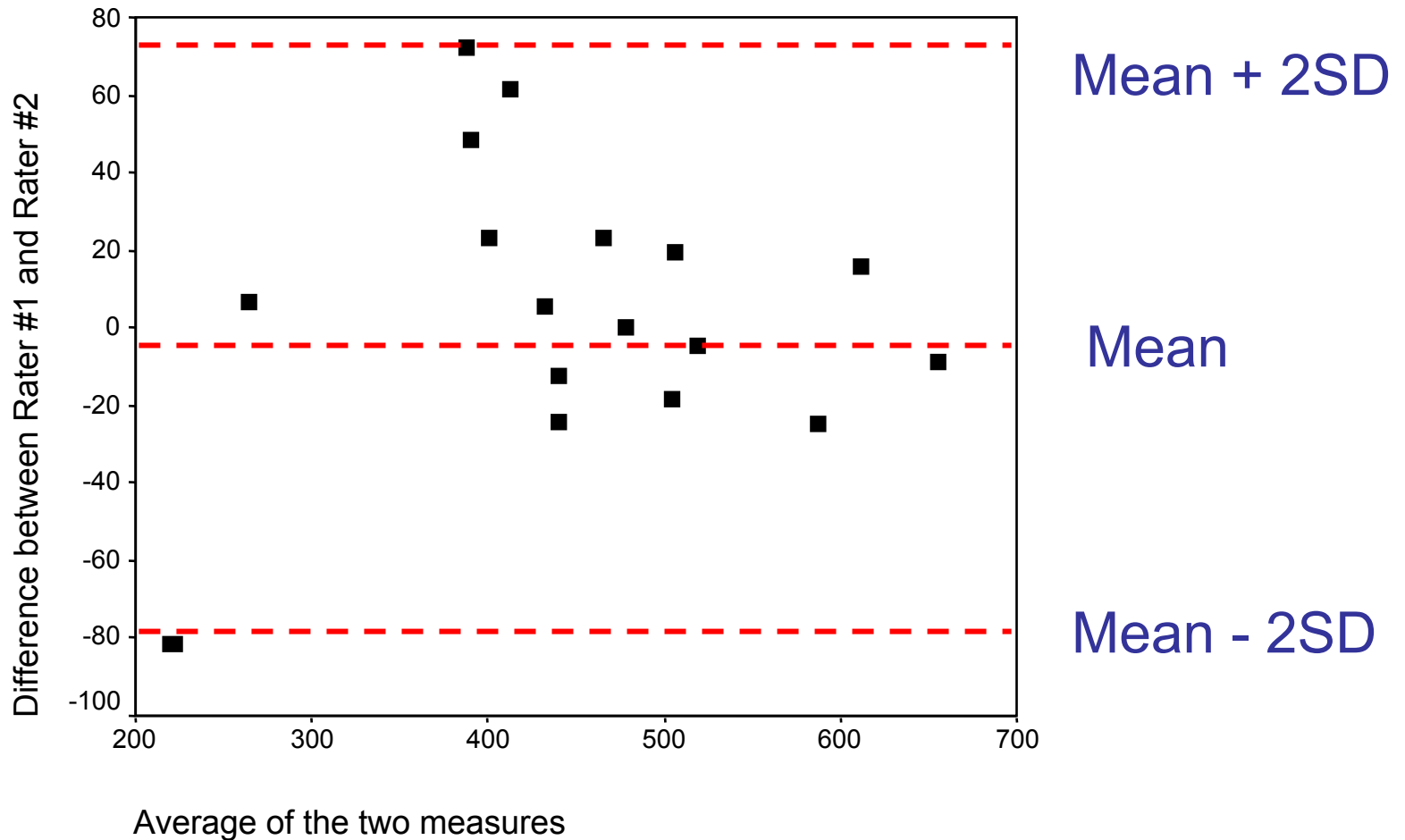
Scatterplot of two measures with line of equality



More informative Scatterplot looks at the difference between the two raters against the average of the two measures:



Provides the mean difference (bias) and limits of agreement (variation):





Assumes normal dist. & constant variance.

References:

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; i: 307-310.

Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991. pp398-403.




What does Consensus Theory offer?

- Decision-making Model
 - Pools information provided by raters to:
 - Select the most “correct answer”
 - By weighting rater’s responses
 - Provide a degree of confidence in the selection
 - Provides an estimate of rater competency
 - Probability a rater knows the correct answer

Can Agreement Equate to Precision?





How does Consensus Theory Work?

- Raters independently answer a question
- Each rater's responses are compared to the aggregate of responses
 - Competency is calculated based upon how often the rater agrees with the majority
- Each person's responses are then weighted by their own competency
- The correct answer is estimated
- Bayes Theorem is used to determine the probability of a true consensus

Medical Example





Assessment of Radiographs

- Study by Kisson, et. al
 - 15 radiographs of injured elbows
 - Assessed by one EM resident, one Ped resident, two Ped EM physicians, one Ped radiologist, two Ortho residents and one Ortho surgeon.
 - Gold standard based upon surgical findings

Actual Proportion of Cases Correctly Classified by Each Rater and Estimated Proportion Correct Based on Consensus Theory.

Rater	Actual % Correct	Estimated % Correct	Estimated D_i	Actual P_{1i}^*	Actual $1 - P_{0i}^\dagger$
Ped EM Physician #1	.720	.752	.504	.583	.846
Ped EM Physician #2	.760	.830	.661	1.00	.538
Ortho Surgeon	.880	.857	.715	.916	.692
Ped Radiologist	.920	.965	.930	1.00	.846
Ped Resident	.880	.898	.797	.917	.846
EM Resident	.720	.768	.536	.917	.538
Ortho Resident #1	.800	.860	.721	.916	.629
Ortho Resident #2	.880	.942	.885	1.00	.769

* Sensitivity.

† Specificity.




Repeated Experiments

- 11 Cardiologists assess 13 cinarteriograms
- 6 Dermatologists asses 15 skin tone panels
- 20 Otolarnngologists assess 10 CT scans
- 6 Radiologists assess 150 mammograms
- 19 Pathologists assess 180 cervcal smears
- 6 Radiologists assess 42 urograms
- 27 Dentists assess dental radiographs
- 72 raters of colposcopic photos for sexual abuse
- 6 Neuroradiologists assess CT scans



Monte Carlo Simulations





True and Mean Estimated Competency Scores for Data Sets Satisfying Goodness-of-Fit Criteria.

D_i	Low Bias	High Bias
	(.67)	(.83)
<hr/>		
.5	.52 ±.14 (800)*	.51 ±.13 (870)
.6	.61 ±.13 (960)	.60 ±.12 (1000)
.7	.71 ±.11 (980)	.71 ±.10 (970)
.8	.81 ±.09 (980)	.81 ±.09 (990)
.9	.91 ±.06 (1000)	.91 ±.06 (1000)
<hr/>		
	* Total number of estimated competencies	



Sensitivity and Specificity of Consensus Diagnostic Predictions

D_i	SENSITIVITY*		SPECIFICITY*	
	Low Bias (.67)	High Bias (.83)	Low Bias (.67)	High bias (.83)
.5	.987	1.000	.875	.734
.6	.994	.999	.937	.858
.7	.998	1.000	.983	.951
.8	.999	1.000	.996	.990
.9	1.000	1.000	1.000	.999

* Table headings must be interchanged to provide results based on negatively biased data.



Questions

