University of York Department of Health Sciences

Measurement in Health and Disease

Suggested answers: Kappa statistics

Question 1

a) *For fusion status, kappa = 0.51. What does this mean and what conclusions could we draw?* Kappa measures the amount by which the agreement exceeds that expected by chance. It is the proportion of subjects for which there is agreement minus the proportion expected to agree, divided by the maximum value this difference could have. Conventionally, 0.51 is thought to represent 'moderate' agreement.

b) *Why is kappa less than the proportional (percentage) agreement for fusion status?* Because kappa is the proportional agreement greater than that which would be expected by chance. It should never be greater than the proportional agreement.

c) *What hypothesis is the 'P<0.0001' testing? What does it tell us?* It is testing the null hypothesis that there is no agreement. It tells us that there is good evidence that the agreement in the population which these data represent is greater than might be expected by chance, given the proportions in the different categories. Evidence of some agreement is not the same as evidence of good agreement, so the P value does not tell us much.

d) *Why was weighted kappa used for bone mineralization ratings?* The bone mineralization was classified in four ordered categories: absent, mild, moderate, or extensive. If one observation is mild and the other moderate, the disagreement is not so bad as between mild and extensive. Weighting allows for this.

Question 2

e) *What is meant by '"good" or "very good" Kappa agreement'?* Good agreement is usually taken to be kappa between 0.6 and 0.8, very good agreement to mean kappa above 0.8.

f) *What do the kappas tell us about initial intra-observer and inter-observer agreement?* Intra-observer agreement was much better than inter-observer agreement, i.e. the student observers were much more consistent with their own assessments than with those of the experts. The student observers are not using the same criteria as the experts for their judgments.

f) *Why did inter-observer agreement improve after training?* The observers have been trained to use the same criteria for assessment as the experts. This led to closer agreement between the students and the experts.

g) *How would you change the design to estimate inter-observer agreement between the students?* We would need to have more than one student assessing the same tooth. One way would be to have all students assess all the teeth. Another would be have some of the students, three or more, students assess each tooth. In either case, we could then apply kappa for multiple observers. We could pair the students and have two assess each tooth, then use the ordinary kappa. In all these designs the students would be treated as interchangeable.