5.2 Abdominal Tenderness to Palpation

Yen et al[1] compared abdominal exam findings suggestive of appendicitis, such as tenderness to palpation, between pediatric emergency physicians and pediatric surgical residents.

Assume that the emergency physician and the surgeon each examine the same 10 patients for right lower quadrant tenderness with the following results:

Emergency Physician	Surgeon		
	Tender	**Not Tender**	**Total**
Tender	3	2	**5**
Not Tender	2	3	**5**
Total	**5**	**5**	**10**

a) Note that the observed agreement is 3 + 3 = 6/10 = 60%. Calculate kappa.

Observed Agreement = (3 + 3)/ 10 = 6 / 10 = 60%
Expected Agreement = (0.5×5 + 0.5×5)/10 = (2.5 + 2.5)/10 = 5/10 = 50%
Kappa = (60% − 50%) / (100% − 50%) = 10% / 50% = 0.20

Now, assume that the emergency physician and the surgeon both find a higher prevalence of right lower quadrant tenderness, but still have 60% observed agreement:

Emergency Physician	Surgeon		
	Tender	**Not Tender**	**Total**
Tender	5	2	**7**
Not Tender	2	1	**3**
Total	**7**	**3**	**10**

b) Calculate kappa.

Observed Agreement = (5 + 1)/10 = 60%
Expected Agreement = (0.7×7 + 0.3×3)/10 = (4.9 + 0.9)/10 = 5.8/10 = 58%
Kappa = (60% - 58%) / (100% - 58%) = 2% / 42% = 0.048

c) Compare the values of kappa for the tables in part (a) and part (b). The observed agreement was 60% in both cases, why is kappa different?

You can think of the second kappa calculation as assuming that the two physicians knew ahead of time that the right lower quadrant would be tender in 7 out of the 10 patients. The kappa of 0.048 says that they really didn't do much better than if they each had just skipped the exam and randomly selected the 7 patients to classify as tender. If the two observers agree that the prevalence of the finding is high or low, it is hard for them to have a high kappa.

Now, assume that the surgeon has a higher threshold than the emergency physician for calling tenderness. This is a source of systematic disagreement.[1] Results follow:

Emergency Physician	Surgeon		
	Tender	**Not Tender**	**Total**
Tender	3	4	**7**
Not Tender	0	3	**3**
Total	**3**	**7**	**10**

d) Note that the observed agreement is still 6/10 or 60% and calculate kappa.

Observed Agreement = (3 + 3)/10 = 60%
Expected Agreement = (0.7×3 + 0.3×7)/10 = (2.1 + 2.1)/10 = 4.2/10 = 42%
Kappa = (60% - 42%) / (100% - 42%) = 18% / 58% = 0.31

e) If you answered (a), (b) and (d) correctly, you found that the highest value of kappa occurred in (d) when disagreements were unbalanced. Why?

Unbalanced disagreement, as occurred here because the surgeon often said "not tender" when the emergency physician said "tender," but never vice versa, leads to lower levels of *expected* agreement. Since the observed agreement was constant in parts a to d, the value for Kappa increased as expected agreement decreased. The lower one's expectations, the more easily they are exceeded! Note, however, that with the level of unbalanced disagreement observed in part d, the kappa is as high as it can be; there is no way to keep these marginals and place numbers inside the table that will give a higher kappa.)

1. Yen K, Karpas A, Pinkerton HJ, Gorelick MH. Interexaminer reliability in physical examination of pediatric patients with abdominal pain. Arch Pediatr Adolesc Med. 2005;159(4):373-6.

[1] In fact, in the Yen et al. study, abdominal tenderness was reported much more frequently by the emergency department residents (73.5%) and attendings (72.1%) than by the surgical residents (43.5%).