

Lesson 22: Rater Agreement and the ICC

MA376 Applied Statistics, AY27-1

Cadet: _____

Section: _____

Objectives

1. **Frame.** A rater is a classifier with no truth column, so agreement replaces accuracy.
2. **Apply.** Percent agreement, kappa, PABAK, and the intraclass correlation coefficient (ICC) from a two-way ANOVA with raters as blocks.
3. **Mean.** When almost every rep is good, or almost every rep is bad, graders agree a lot by chance alone, so kappa drops even though percent agreement stays high. The ICC is the share of the variation in a call that comes from the push-up.
4. **Communicate.** Tell the first sergeant whether the graders agree and which grader is the problem.

1. No answer key

Lesson 21's clinic model scored accuracy, sensitivity, specificity, precision and F_1 . Each one needs the _____ column.

1. The push-up graders below have no such column. What can you still measure?

2. Two graders

40 hand-release push-ups on video. Five graders (A to E) call each one a rep or a no rep. Grader B against Grader E:

	B: no rep	B: rep	Total
E: no rep			
E: rep			
Total			40

Percent agreement $p_o = \frac{\text{both no rep} + \text{both rep}}{n} = \underline{\hspace{2cm}}$

Chance agreement. Take away the video and keep each grader's habit. $P(B) = P(\text{B calls no rep}) = \underline{\hspace{2cm}}$
 $P(E) = P(\text{E calls no rep}) = \underline{\hspace{2cm}}$

By the multiplication rule from MA206, $P(B \cap E) = P(B) P(E)$.

	B: no rep	B: rep
E: no rep	$P(B \cap E) = \underline{\hspace{2cm}}$	$P(B^c \cap E) = \underline{\hspace{2cm}}$
E: rep	$P(B \cap E^c) = \underline{\hspace{2cm}}$	$P(B^c \cap E^c) = \underline{\hspace{2cm}}$

$$p_e = P(B \cap E) + P(B^c \cap E^c) = \underline{\hspace{2cm}}$$

Kappa $\kappa = \frac{p_o - p_e}{1 - p_e} = \underline{\hspace{2cm}}$ Landis and Koch rating: $\underline{\hspace{2cm}}$

2. B and E agree on 87.5% of the push-ups. Why is kappa so much lower?

3. A drawback of kappa

Hold percent agreement at $p_o = 0.875$. Both graders call no rep on the same share of push-ups.

Called no rep (prevalence)	10%	30%	50%	70%	90%
p_e					
κ					

3. The graders did not change between columns. What did? Can you compare kappas from two groups with different prevalence?

4. PABAK

If both graders call no rep half the time, $p_e =$ _____. Put that into kappa:

$$\kappa = \frac{p_o - \text{_____}}{1 - \text{_____}} = \text{_____}$$

B and E: $\kappa =$ _____ PABAK = _____

Of the _____ push-ups either B or E called a no rep, they both called _____.

4. What does PABAK say about B and E? What does that last line say?

5. Grader D

Grader A against Grader D: $p_o =$ _____ $p_e =$ _____ $\kappa =$ _____ PABAK = _____

5. How many no reps did D call? Which number catches D, and which one misses D?

6. All ten pairs

Kappa for every pair.

	A	B	C	D	E
A					
B					
C					
D					
E					

6. Which graders are consistent with each other? Does that make them right?

7. The ICC

`lm(no_rep ~ pushup + grader)`, graders as blocks. From the **Mean Sq** column:

MS_R (pushup) = _____ MS_C (grader) = _____ MS_E (Residuals) = _____

$$\text{Push-up} \quad \hat{\sigma}_{\text{pushup}}^2 = \frac{MS_R - MS_E}{k} = \underline{\hspace{2cm}}$$

$$\text{Grader} \quad \hat{\sigma}_{\text{grader}}^2 = \frac{MS_C - MS_E}{n} = \underline{\hspace{2cm}}$$

$$\text{Error} \quad \hat{\sigma}_{\text{error}}^2 = MS_E = \underline{\hspace{2cm}}$$

$$\text{ICC} = \frac{\hat{\sigma}_{\text{pushup}}^2}{\hat{\sigma}_{\text{pushup}}^2 + \hat{\sigma}_{\text{grader}}^2 + \hat{\sigma}_{\text{error}}^2} = \underline{\hspace{2cm}} \quad \text{Koo and Li rating: } \underline{\hspace{2cm}}$$

	All five	Without D
Grader effect, p-value		
ICC		

7. Tell the first sergeant, in two or three sentences, whether the graders agree and which grader is the problem.

8. Your turn: is it a sandwich?

Your group: _____ raters $p_o =$ _____ $\kappa =$ _____ PABAK = _____ ICC = _____

8. Who in your group is the Grader D?

Key Takeaway. With no answer key, you cannot ask who is right. You can ask whether the raters agree more than chance would, and which rater does not. Kappa removes chance but moves with prevalence. PABAK fixes prevalence but hides a biased rater. The ICC splits the variation into push-up, grader and error.