

## การวัดความสอดคล้อง

### Measuring agreement: Kappa & Weight Kappa

ผู้ช่วยศาสตราจารย์ นิคม ถนอมเสียง  
สาขาวิชาวิทยาการระบาดและชีวสถิติ  
คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น  
Email: nikom@kku.ac.th  
WebPage: http://home.kku.ac.th/nikom

## การวัดความสอดคล้อง

- การวัดความสอดคล้อง เป็นการวัดค่าความเที่ยงของ rater หรือ เครื่องมือหลายชนิดวัดในสิ่งเดียวกัน
- เพื่อพิจารณาว่าสิ่งที่วัดสิ่งเดียวกันโดยผู้วัดหลายคน หรือเครื่องมือวัดหลายชนิด ให้ผลการวัดได้สอดคล้องกัน

## พิจารณาใช้เมื่อไร

- แพทย์ 2 คน ฟังเสียง Murmur ได้สอดคล้องกันหรือไม่
  - การศึกษาเปรียบเทียบระดับ Cholesterol ด้วยวิธีทดสอบ 3 วิธี เพื่อพิจารณาว่าให้ผลสอดคล้องกันหรือไม่
- etc

## สถิติที่ใช้วัดความสอดคล้อง

### Continuous data

- Lin concordance correlation coefficient
  - Limit of agreement
  - Intraclass correlation
- etc

### Categorical data

- ✓ Kappa Statistic (2x2)
  - ✓ weight Kappa (rxc)
- etc

สถิติ Kappa: ใช้วัดความสอดคล้อง ข้อมูลที่เป็นสิ่งเดียวกันหรือ คนเดียวกัน โดยผู้วัด 2 คน/ เครื่องมือวัด 2 ชนิด ข้อมูลที่ใช้ในการวิเคราะห์ เป็นข้อมูลจำแนกนับ

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT  
Vol. XX, No. 1, 1960

### A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES\*

JACOB COHEN  
New York University

CONSIDER Table 1. It represents in its formal characteristics a situation which arises in the clinical-social-personality areas of psychology, where it frequently occurs that the only useful level of measurement obtainable is nominal scaling (Stevens, 1951, pp. 25-26), i.e. placement in a set of  $k$  unordered categories. Because the categorizing of the units is a consequence of some complex judgment process performed by a "two-legged meter" (Stevens, 1958), it becomes important to determine the extent to which these judgments are reproducible, i.e., reliable. The procedure which suggests itself is that of having two (or more) judges independently categorize a sample of units and determine the degree, significance, and

## ลักษณะของข้อมูลที่ใช้ในการวิเคราะห์ kappa & weight Kappa

Rater A	Rater B		รวม
	+	-	
+	a	b	$N_{1+}$
-	c	d	$N_{2+}$
รวม	$n_{+1}$	$n_{+2}$	N

Rater A	Rater B		รวม
	+	-	
+	$p_{11}$	$p_{12}$	$p_{1+}$
-	$p_{21}$	$p_{22}$	$p_{2+}$
รวม	$p_{+1}$	$p_{+2}$	p

kappa

a = ให้ผลเชิงบวกเหมือนกัน  
d = ให้ผลเชิงลบเหมือนกัน

ผู้วัด/ เครื่องมือวัด 1	ผู้วัด/เครื่องมือวัด 2				รวม
	1	2	...	k	
1	✓				$n_{1+}$
2		✓			$n_{2+}$
...			✓		...
k				✓	$n_{k+}$
รวม	$n_{+1}$	$n_{+2}$	...	$n_{+k}$	N

weight Kappa

✓ ให้ผลสอดคล้องกัน

$$\text{สถิติ Kappa } \hat{k} = \frac{p_o - p_e}{1 - p_e} \text{ or } \frac{f_o - f_e}{N - f_e}$$

$p_o$  = ความน่าจะเป็นความสอดคล้องของค่าสังเกต (observe probability of agreement)

$p_e$  = ความน่าจะเป็นความสอดคล้องของค่าคาดหวัง (hypothetical expected probability of agreement)

$f_o, f_e$  = ความถี่ของค่าสังเกตและความถี่คาดหวังของค่าสังเกต

The coefficient  $\kappa$  is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

Expressed in frequencies to facilitate computation,

$$\kappa = \frac{f_o - f_e}{N - f_e} \quad (2)$$

$$\hat{k} = \frac{p_o - p_e}{1 - p_e} \text{ or } \frac{f_0 - f_e}{N - f_e}$$

$$\hat{k} = \frac{p_o - p_e}{1 - p_e};$$

$$p_o = \frac{(a + d)}{N},$$

$$p_e = \frac{\left[ \frac{(n_{1+})(n_{+1})}{N} \right] + \left[ \frac{(n_{2+})(n_{+2})}{N} \right]}{N}$$

$$\hat{k} = \frac{f_0 - f_e}{N - f_e}; f_0 = a + d, f_e = \sum \frac{n_{i+} n_{+i}}{N}$$

Rater	Rater B		
A	+	-	รวม
+	a	b	n <sub>1+</sub>
-	c	d	n <sub>2+</sub>
รวม	n <sub>+1</sub>	n <sub>+2</sub>	N

  

Rater	Rater B		
A	+	-	รวม
+	p <sub>11</sub>	p <sub>12</sub>	p <sub>1+</sub>
-	p <sub>21</sub>	p <sub>22</sub>	p <sub>2+</sub>
รวม	p <sub>+1</sub>	p <sub>+2</sub>	p

**การทดสอบสมมุติฐาน**  
 การทดสอบสมมุติฐาน ใช้สถิติ z  
 -โดยคำนวณค่า se<sub>0</sub> จากการศึกษาของ  
 Fleiss, Cohen & Everitt (1969) ดังนี้

$$z = \frac{\hat{k}}{se_0(\hat{k})}$$

$$se_0(\hat{k}) = \frac{1}{(1 - p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^k p_{i+} p_{+i} (p_{i+} + p_{+i})}$$

การแปลความหมายของค่าสถิติ kappa พิจารณาดังนี้  
 (Landis & Koch, 1977)

ค่าสถิติ Kappa	ขนาดความสอดคล้อง (Strength of Agreement)
<0.00	แย่มาก (Poor)
0.00-0.20	น้อย (Slight)
0.21-0.40	พอใช้ (Fair)
0.41-0.60	ปานกลาง (Moderate)
0.61-0.80	ดี (Substantial)
0.81-1.00	ดีมาก/ค่อนข้างสมบูรณ์ (Almost Perfect)

**ตัวอย่าง**  
 ต้องการทราบแพทย์ 2 คน ตรวจสอบ Cardiac Murmur ได้  
 สอดคล้องกันหรือไม่

รังสีแพทย์คนที่ 1	รังสีแพทย์คนที่ 2		รวม
	+	-	
+	7	3	10
-	2	6	8
รวม	9	9	18

**Kappa=.44, Z=1.897 (p-value=0.0289)**

Rater A	Rater B		รวม
	+	-	
+	7(5)	3	10
-	2	6(4)	8
รวม	9	9	18

Rater A	Rater B		รวม
	+	-	
+	.39(.28)	.17	.56
-	.11	.33(.22)	.44
รวม	.50	.50	1.00

$f_0 = 7 + 6 = 13$        $p_o = 0.39 + 0.33 = 0.72$   
 $f_e = \left[ \frac{(10)(9)}{18} \right] + \left[ \frac{(8)(9)}{18} \right] = 5 + 4 = 9$        $p_e = 0.28 + 0.22 = 0.50$   
 $\hat{k} = \frac{13 - 9}{18 - 9} = .44$        $\hat{k} = \frac{.72 - .50}{1 - .50} = .44$

$$\hat{k} = \frac{p_o - p_e}{1 - p_e} \quad p_o = \frac{7+6}{18} = .72222; \quad p_e = \frac{(10)(9)}{18} + \frac{(8)(9)}{18} = .5$$

$$\hat{k} = \frac{.72222 - .5}{1 - .5} = .44444$$

ทดสอบสมมุติฐาน ใช้สถิติ z

$$z = \frac{\hat{k}}{se_0(\hat{k})} \quad se_0(\hat{k}) = \frac{1}{(1 - p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^k p_{i+} p_{+i} (p_{i+} + p_{+i})}$$

$$\sum_{i=1}^k p_{i+} p_{+i} (p_{i+} + p_{+i}) = \left[ \left( \frac{10}{18} \right) \left( \frac{9}{18} \right) \left( \frac{10}{18} + \frac{9}{18} \right) \right] + \left[ \left( \frac{8}{18} \right) \left( \frac{9}{18} \right) \left( \frac{8}{18} + \frac{9}{18} \right) \right] = 0.50309$$

$$se_0(\hat{k}) = \frac{1}{(1 - .5)\sqrt{18}} \sqrt{.5 + .5^2 - .50309} = .23424$$

$$z = \frac{\hat{k}}{se_0(\hat{k})} = \frac{.4444}{.23424} = 1.897367$$

```

.do "H:\cat2011\kappa.do"
. qui tabi 7 3\ 2 6, replace
. rename row rater_A
. rename col rater_B
. kap rater_A rater_B [freq=pop], tab

```

rater_A	rater_B	Total
1	2	10
2	2	8
Total	9	18

```

-----
Expected
Agreement Agreement Kappa Std. Err. Z Prob>Z
-----
72.22% 50.00% 0.4444 0.2342 1.90 0.0289

```

ปฏิเสธสมมุติฐาน  $H_0 : \kappa = 0$   
 รังสีแพทย์ 2 คน ตรวจ Cardiac murmur ผู้ป่วยได้สอดคล้องกัน  
 อย่างมีนัยสำคัญ ( $Z=1.90, p=0.029$ ) โดยมีขนาดความสอดคล้อง  
 ในระดับปานกลาง ( $\kappa=0.44$ )

ประเด็นศึกษาเพิ่มเติม

กรณีพบว่า: ค่าความสอดคล้องกัน (concordance)  
 สูงแต่พบว่าค่า kappa ต่ำ

Byrt, T., Bishop, J., Carlin, J. B. (1993). Bias, prevalence and kappa. J. Clin. Epidemiol. 46, 423-429.

Cicchetti, D. V., Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43, 551-558

Weight Kappa

- ❖ การคำนวณค่า Kappa พิจารณาจากข้อมูลที่มีความเห็นสอดคล้องกันเท่านั้น (diagonal)
- ❖ Weight Kappa พิจารณาในเรื่องที่ไม่สอดคล้องกัน ร่วมด้วย เช่น การวินิจฉัยโรคมาเร็งโดยการอ่านผลจากการทำ xeromammograms โดยรังสีแพทย์ 2 คน (normal, Benign, Suspected cancer, cancer)

Assessments of 85 xeromammograms by two radiologists (Boyd et al., 1982)

Radiologist 1	Radiologist 2				Total
	Normal	Benign	Suspected Cancer	Cancer	
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected Cancer	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

ความไม่สอดคล้องเกิดขึ้นได้มากกว่ากรณีที่มีการวัดใกล้เคียงกัน มากกว่าผลแตกต่างกันมาก เช่น Normal-Benign, Normal-Cancer

Weight Kappa

$$\hat{\kappa}_w = \frac{p_{0(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

$$p_{0(w)} = \sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{ij} \text{ or } p_{0(w)} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^k w_{ij} f_{ij}$$

$$p_{e(w)} = \sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{+j} p_{+i} \text{ or } p_{e(w)} = \frac{1}{n^2} \sum_{i=1}^k \sum_{j=1}^k w_{ij} r_i c_j$$

การคำนวณ Weight (I)

การคำนวณค่าน้ำหนัก ( $w_{ij}$ ) คำนวณโดย Cohen (1968) และ Cicchetti & Allison (1971) ดังนี้

$$w_{ij} = 1 - \frac{|i-j|}{k-1}$$

I=ลำดับที่แนวคอลัมน์, j=ลำดับที่แนวแถว, k=จำนวนกลุ่ม

การคำนวณ Weight (II)

การคำนวณ ค่าน้ำหนัก ( $w_{ij}$ ) คำนวณโดย Fliess & Cohen (1973) ~ Intraclass Correlation Coefficient ICC

$$w_{ij} = 1 - \frac{|i-j|^2}{(k-1)^2}$$

I=ลำดับที่แนวคอลัมน์ j=ลำดับที่แนวแถว, k=จำนวนกลุ่ม

### การทดสอบสมมติฐาน

ทดสอบสมมติฐานเพื่อทดสอบ  $H_0: k_w = 0$

$$z = \frac{\hat{k}_w}{se_0(\hat{k}_w)}$$

$$se_0(\hat{k}_w) = \frac{1}{(1 - p_{e(w)})\sqrt{n}} \sqrt{\sum \sum p_i p_j [w_{ij} - (\bar{w}_{i+} + \bar{w}_{+j})]^2 - p_{e(w)}^2}$$

$$\bar{w}_{i+} = \sum_{j=1}^k p_{ij} w_{ij} \quad \text{and} \quad \bar{w}_{+j} = \sum_{i=1}^k p_{i+} w_{ij}$$

### การทดสอบสมมติฐาน

ทดสอบสมมติฐานเพื่อทดสอบ  $H_0: k_w = k$

$$z = \frac{|\hat{k}_w - k_w|}{se_0(\hat{k}_w)}$$

$$se_0(\hat{k}_w) = \frac{1}{(1 - p_{e(w)})\sqrt{n}} \sqrt{\sum_{i=1}^k \sum_{j=1}^k p_{ij} [w_{ij} - (\bar{w}_i + \bar{w}_j)(1 - \hat{k}_w)]^2 - [\hat{k}_w - p_{e(w)}(1 - \hat{k}_w)]^2}$$

$$\bar{w}_{i+} = \sum_{j=1}^k p_{ij} w_{ij} \quad \text{and} \quad \bar{w}_{+j} = \sum_{i=1}^k p_{i+} w_{ij}$$

การวินิจฉัยโรคมะเร็งโดยการอ่านผลจากการทำ xeromammograms โดยรังสีแพทย์ 2 คน (normal, Benign, Suspected cancer, cancer)

รังสีแพทย์ คนที่ 1	รังสีแพทย์คนที่ 2				รวม
	normal	Benign	Suspected cancer	cancer	
normal	21	12	0	0	33
benign	4	17	1	0	22
Suspected cancer	3	9	15	2	29
cancer	0	0	0	1	1
รวม	28	38	16	3	85

Weight Kappa

$K_w = .57, Z=7.22$  (p-value < .0001)

### ใช้การวิเคราะห์ Weight Kappa

$$\hat{k}_w = \frac{p_{0(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

$$p_{0(w)} = \sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{ij} \quad \text{or} \quad p_{0(w)} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^k w_{ij} f_{ij}$$

$$p_{e(w)} = \sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{+j} p_{+i} \quad \text{or} \quad p_{e(w)} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^k w_{ij} r_i c_j$$

คำนวณ weight  $w_{ij} = 1 - \frac{|i-j|}{k-1}$

รังสีแพทย์ คนที่ 1	รังสีแพทย์คนที่ 2			
	normal	Benign	Suspected cancer	cancer
normal	1	.6667	.3333	0
benign	.6667	1	.6667	.3333
Suspected cancer	.3333	.6667	1	.6667
cancer	0	.3333	.6667	1

. di 1-abs(1-1)/(4-1)  
1  
. di 1-abs(1-2)/(4-1)  
.66666667  
. di 1-abs(1-3)/(4-1)  
.33333333  
. di 1-abs(1-4)/(4-1)  
0

$$p_{0(w)} = (1/85) \times [(21)(1) + (12)(.6667) + (0)(.3333) + (0)(0)] + [(4)(.6667) + (17)(1) + (1)(.6667) + (0)(.3333)] + [(3)(.3333) + (9)(.6667) + (15)(1) + (2)(.6667)] + [(0)(0) + (0)(.3333) + (0)(.6667) + (1)(1)]$$

$$= .866676$$

$$p_{e(w)} = (1/85^2) \times [(1)(33)(28) + (.6667)(33)(38) + (.3333)(33)(16) + (0)(33)(3)] + [(.6667)(22)(28) + (1)(22)(38) + (.6667)(22)(16) + (.3333)(22)(3)] + [(.3333)(29)(28) + (.6667)(29)(38) + (1)(29)(16) + (.6667)(29)(3)] + [(0)(1)(28) + (.3333)(1)(38) + (.6667)(1)(16) + (1)(1)(3)]$$

$$= .691082$$

$$\hat{k}_w = \frac{p_{0(w)} - p_{e(w)}}{1 - p_{e(w)}} = \frac{.866676 - .691082}{1 - .691082} = .568418$$

```
. tabi 21 12 0 0 \ 4 17 1 0 \ 3 9 15 2 \ 0 0 0 1, replace
```

row	1	2	3	4	Total
1	21	12	0	0	33
2	4	17	1	0	22
3	3	9	15	2	29
4	0	0	0	1	1
Total	28	38	16	3	85

```

Pearson chi2(9) = 77.8111 Pr = 0.000
. rename row rater_A
. rename col rater_B

```

ถ้าไม่ rename ใช้คำสั่งใน kap row col

```
. kap rater_A rater_B [freq=pop], wgt(w)
```

Ratings weighted by:

1.0000	0.6667	0.3333	0.0000
0.6667	1.0000	0.6667	0.3333
0.3333	0.6667	1.0000	0.6667
0.0000	0.3333	0.6667	1.0000

Agreement	Expected Agreement	Kappa	Z	Pr>Z
86.67%	69.11%	0.5684	7.22	0.0000

```
. kap row col [freq=pop], wgt(w2)
```

Ratings weighted by:

1.0000	0.8889	0.5556	0.0000
0.8889	1.0000	0.8889	0.5556
0.5556	0.8889	1.0000	0.8889
0.0000	0.5556	0.8889	1.0000

Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
94.77%	84.09%	0.6714	0.1079	6.22	0.0000

### คำสั่ง kappa

```
. tabi a b \ c d, replace
. rename row raterA
. rename col raterB
. kap raterA raterB(kap row col
freq=pop], tab
```

หรือ kap var1 var2 (มีข้อมูล)

```
. kap row col
. kap raterA raterB
```

### คำสั่ง weight kappa

```
. tabi # #...n\##...n\..., replace
. rename row raterA
. rename col raterB
. kap raterA raterB [freq=pop], wgt(w)
. kap row col [freq=pop], wgt(w)
. Kap var var2 ,wgt(w)
```

### เช่น

```
. kap row col, wgt(w2)
. kap raterA raterB, wgt(w)
```

### Reference

- Boyd NF, Wolfson C, Moskowitz M, Carlile T, Petitclerc M, Ferri HA, Fishell E, Gregoire A, Kiernan M, Longley JD, Simor IS, Miller AB. Observer variation in the interpretation of xeromammograms. *J Natl Cancer Inst.* 1982 Mar; 68(3):357-63.
- Cicchetti, D.V. & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 11,101-109
- Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* , 20: 37-46.
- Cohen, J. (1968). "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin*, 70: 213-220.
- Fleiss, J. L. (1981) *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley. 38-46
- Fleiss, J.L.; Cohen, J., & Everitt, B.S. (1969). "Large sample standard errors of kappa and weighted kappa". *Psychological Bulletin* 72: 323-327.
- Fleiss, J.L.; Cohen, J. (1973). "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability". *Educational and Psychological Measurement* 33: 613-619.
- Landis, J.R.; & Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics* 33 (1): 159-174.