評定者間信頼性

教育学研究科修士1回生
西村優美子

評価者間信頼性

- 被験者のパフォーマンスや病状などを評価する場合、評価者の判断に主観性が入ってしまう恐れがある。

- たとえば…
  教師1はAさんの成績を5と判定
  教師2はAさんの成績を3と判定

教師たちの評価はどれくらい信用できるのか？

評価者信頼性を求める必要がある

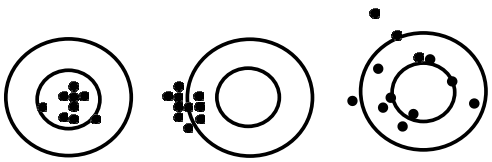
流れ

- 信頼性
- 級内相関係数
- 級内相関係数ー演習
- カッパ係数
- カッパ係数ー演習
- カッパ係数ー発展

妥当性と信頼性 (対馬, 2007)

- 測定を行うにあたって…
  妥当性 (validity) のある尺度を使って、**信頼性** (reliability) の高いデータを得ることが大切である。
  ＊妥当性：その尺度が測定すべきものを測定しているか
  ＊信頼性：測定が安定して正確であるか

- ダーツのアナロジー
  妥当性の高い場合は信頼性も高くなる必要がある
  →信頼性は妥当性も満たす必要条件となる。



妥当性：高い	低い	低い
信頼性：高い	高い	低い

信頼性とは (対馬, 2007)

- 尺度が測定している構成概念をどの程度高い精度で測定しているか
  →**安定性・一貫性**

  安定性：被験者にもう一度同じ項目を回答してもらった場合に同じ効果が得られるか
  一貫性：異なった項目であっても同じ特性をたずねる問に対しては、同じような回答が得られるか
- 信頼性係数
  →基礎となる理論として
  古典的テスト理論が設定されている。

古典的テスト理論 (対馬, 2007)

- 観測値は目に見えない真値と誤差から成り立っている。

  観測値(測定値) = 真値 + 誤差
  ※真値と誤差の相関は0

  ⇩

  観測値の分散 = 真値の分散 + 誤差分散

  信頼性係数 ρ = 真値の分散/観測値の分散
  = 真値の分散/真値の分散+誤差の分散
- 信頼性係数ρ (ロー)
  0から1までの値をとる。
  .80以上の場合、観測データの80%以上が構成概念の説明をしていることになり、信頼性の高いデータといえる。
- しかし…
  真値がわからないので**信頼性係数を求めることができない**。

信頼性の推定方法

そこで、信頼性係数を推定する必要がある

①再テスト法：同じテストを同一の受験者に二回実施

②平行テスト法：ほぼ同一の難易度のテストを同一の被験者に実施

③内的一貫性：同じ構成概念を測定する尺度内で、受験者の個々の項目の得点がどの程度一貫しているのかをみる。(折半法、スピアマン･ブラウン公式、アルファ係数…など)

④評価者信頼性

＊評価者内信頼性
＊評価者間信頼性

今回はこの評定者間信頼性について説明します！

評価者信頼性

- 被験者のパフォーマンスや病状などを評価する場合、評価者の判断に主観性が入ってしまう恐れがある。
  →評価者信頼性を求める必要がある。
- 測定値の一致度を検討するには…
  ･級内相関係数(ICC)：間隔尺度、比率尺度
  ･κ（カッパ）係数：名義尺度、順序尺度
  ･ケンドールの一致係数：順序尺度
  など…

一般化可能性理論(対馬, 2007)

級内相関係数の前に…

- 古典的テスト理論の延長として、**一般化可能性理論**というものがある。

→分散分析を利用して、分散成分を分割し、その一般化によって測定の信頼性を推定する方法

→級内相関係数は、この一般化可能性理論における信頼性係数の一部である。

一般化可能性理論において…
＊G研究（一般化可能性研究）：各分散を推定
＊D研究（決定研究）：分散分析の影響を推定し、適切なテスト使用計画をたてる過程

たとえば…ある測定法Xの信頼性をしるために
数人の評価者によってXを測定し、そのばらつき（分散）を求める（G研究）。次に、複数評価者におけるXの信頼性を求める（D研究）。

級内相関係数

- 全部で3つのタイプが存在する
  さらにそれぞれのCaseにおいて、1回のみの測定と繰り返し測定の2タイプが存在するため、計6種類係数がある。

Case1:評価者内信頼性：
測定値=測定誤差+被験者の真の値

Case2:**評価者間信頼性**：
測定値=被験者の真の値+測定誤差+交互作用
+評価者の効果

Case3:評価者が特定されているときの評価者間信頼性：
測定値=被験者の真の値+測定誤差+交互作用
(+評価者の効果（AorB）)
↑()内は分析対象でない

評価者間信頼性

評価者間信頼性の中にも･･･
①k人の評定者がn人の被験者の測定をした際の信頼性
②k人の評定者がn人の被験者をm回測定したときの平均値の信頼性
の2種類がある。

※このmの妥当性については、別途考慮が必要である。
→ある測定をしたときの評価者内信頼性が ρ 1=0.7であるとき、ρ 2=0.9以上にすることが目標とすると、

$$\frac{\rho_1(1-\rho_2)}{\rho_2(1-\rho_1)}$$

↑の式で一人当たり何回繰り返し測定した平均をデータで用いるとρ 2=0.9となるかを検討する。(対馬, 2007より)

評価者間信頼性 (対馬, 2007)

測定値　=期待値
+被験者の効果+評価者の効果
+評価者と被験者の交互作用
+被験者の測定誤差
と表すことができる。

①k人の評定者がn人の被験者の測定をした際の信頼性

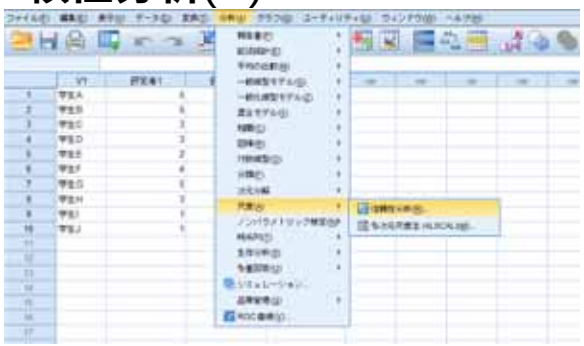
$$\frac{\text{被験者の効果の分散}}{\text{被験者の効果の分散}+\text{評価者の効果の分散}+\text{交互作用の分散}+\text{測定誤差の分散}}$$

②k人の評定者がn人の被験者をm回測定したときの平均値の信頼性

$$\frac{\text{被験者の効果の分散}}{\text{被験者の効果の分散}+\frac{(\text{評価者の効果の分散}+\text{交互作用の分散}+\text{測定誤差の分散})}{k}}$$

SPSSで級内相関係数を算出してみよう！

①分析(A)→尺度(A)→信頼性分析(R)

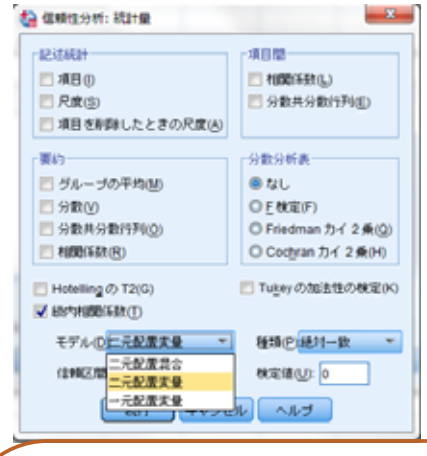


②評定者1~3を項目へ移動

SPSSで級内相関係数を算出してみよう！

③級内相関係数にチェック
モデル：二元配置変量
種類：絶対一致　を選択

④続行をクリック→OKをクリック



モデル(井上俊哉, 2011)
評定者要因が変量効果を持つ（評定者が評定者母集団から無作為に選ばれた）とみなされる。対象要因も変量効果を持つとみなされるので、「2元配置変量」である。

ちなみに・・・
評価者が特定されているときの評価者間信頼性
評定者要因は固定効果、対象要因は変量効果として扱われる（固定効果と変量効果が混ざっている）ので「2元配置混合」
評価者内信頼性
評定者要因のみが変量効果をもつので、「1元配置変量」

種類(井上俊哉, 2011)
「一致性」では対象相互の相対的位置の一貫性が問題とされるのに対して、「絶対一致」では評定値間の完全な一致が要求される。
そのため、評価者間信頼性では「絶対一致」を選択する。

ちなみに・・・
評価者が特定されているときの評価者間信頼性では、評価者の効果は考慮されず、相対的に平行な関係にあれば信頼性は高いと判断するので、一致性を選択する。

級内相関係数(ICC)

- **単一測定値**
  1回の測定値の級内相関係数
- **平均測定値**
  平均測定値の級内相関係数
- ICCの判定基準は右図
  しかし、後で紹介するカッパ係数の判定を応用したものであり、理論的根拠はない。
  他の基準と照らし合わせて、おおよそ0.7以上あれば信頼性が高いといわれている。

信頼性分析

[データセット4]

スケール: すべての変数

ケース処理の要約

		N	%
ケース	有効数	10	100.0
	除外数[a]	0	.0
	合計	10	100.0

a. 手続きのすべての変数に基づいたリストごとの削除。

信頼性統計量

Cronbach のアルファ	項目の数
.895	3

級内相関係数

	級内相関[b]	95% 信頼区間		真の値 0 を使用した F 検定			
		下限	上限	値	df1	df2	有意確率
単一測定値	.748	.448	.922	9.511	9	18	.000
平均測定値	.899	.709	.973	9.511	9	18	.000

人的効果と測定効果の両方が変量であるときの二元変量効果モデル。

a. 交互作用効果の有無にかかわらず、推定量は同じです。

b. 完全一致定義を使用したタイプ A 級内相関係数。

ICC の値	判定
0.0 – 0.20	*slight*
0.21 – 0.40	*fair*
0.41 – 0.60	*moderate*
0.61 – 0.80	*substantial*
0.81 – 1.00	*almost perfect*

http://www.hs.hirosaki-u.ac.jp/~pteiki/research/stat/icc.pdf

評価者信頼性

- 被験者のパフォーマンスや病状などを評価する場合、評価者の判断に主観性が入ってしまう恐れがある。
  →評価者信頼性を求める必要がある。
- 測定値の一致度を検討するには…
  ・相関係数、級内相関係数(ICC):間隔尺度、比率尺度
  ・**κ（カッパ）係数**:名義尺度、順序尺度
  ・ケンドールの一致係数:順序尺度

カッパ係数

- κ係数
  評価者間信頼性･評価者内信頼性のどちらの場合も使用可能。
- 今回は評価者間信頼性を例に説明します！

例えば…
二人の医師A、Bが患者のある病気の陽性･陰性を診断したとする。

患者	医師A	医師B
1	陽性	陽性
2	陰性	陰性
3	陽性	陰性
4	陽性	陽性
5	陽性	陽性
6	陰性	陽性
7	陰性	陰性
…	…	…

→クロス表に直すと→

		医師B		
		陽性	陰性	合計
医師A	陽性	20	7	27
	陰性	15	28	43
	合計	35	35	70

カッパ係数

		医師B		
		陽性	陰性	合計
医師A	陽性	20	7	27
	陰性	15	28	43
	合計	35	35	70

(20+28)/70=0.69
一致率は、69%！

しかし
この一致率には、医師ABが偶然に同じ判断を下した可能性を考慮していない！

→見かけ上の一致率
偶然による一致を考慮する必要がある！
医師ABが偶然に「陽性」と診断する確率
(35/70)×(27/70) = 0.19
医師ABが偶然に「陰性」と診断する確率
(35/70)×(43/70) = 0.31
→ABが偶然に一致した確率
0.19 + 0.31 = 0.5

カッパ係数

		医師B		
		陽性	陰性	合計
医師A	陽性	20	7	27
	陰性	15	28	43
	合計	35	35	70

κ係数（偶然によらない一致率）

$$\frac{\text{見かけ上の一致率のうち、偶然によらない一致率}}{\text{全体の一致率のうち、偶然によらない一致率}}$$

$$\frac{= 0.69 - 0.5}{1.00 - 0.5} = 0.38$$

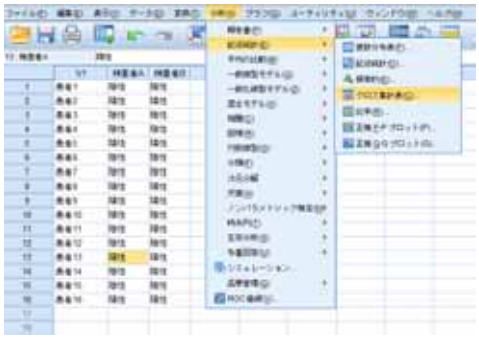
これがカッパ係数となる！

κ係数の判定基準	
値	一致度
.00- .20	低い
.21- .40	やや低い
.41- .60	中程度
.61- .80	かなり高い
.81-1.00	ほぼ一致

カッパ係数をSPSSで算出してみよう！

- とても簡単な計算式なので、手でも算出できます･･･が、一応やってみましょう。

①分析(A)→記述統計(E)→クロス集計表(C)



②行に検査者Aを、列に検査者Bを入れる

カッパ係数をSPSSで算出してみよう！

③統計量(S)をクリックし、カッパ (K)にチェック
④続行→OKをクリック

カッパ係数

- カッパ係数は
  κ = .746、p = .003
  で有意となる。
- 今回の判定は
  “かなり高い”となる。

➔ クロス集計表

［データセット5］

処理したケースの要約

	ケース					
	有効数		欠損		合計	
	N	パーセント	N	パーセント	N	パーセント
検査者A * 検査者B	16	100.0%	0	0.0%	16	100.0%

検査者A と 検査者B のクロス表

度数

		検査者B		
		陰性	陽性	合計
検査者A	陰性	8	1	9
	陽性	1	6	7
合計		9	7	16

対称性による類似度

		値	漸近標準誤差[a]	近似 T 値[b]	近似有意確率
一致の測定方法	カッパ	.746	.168	2.984	.003
有効なケースの数		16			

a. 帰無仮説を仮定しません。
b. 帰無仮説を仮定して漸近標準誤差を使用します。

κ係数の判定基準

値	一致度
.00- .20	低い
.21- .40	やや低い
.41- .60	中程度
.61- .80	かなり高い
.81-1.00	ほぼ一致

カッパ係数～発展

		Observer A		
		Yes	No	Total
Observer B	Yes	a	b	g_1
	No	c	d	g_2
	Total	f_1	f_2	N

カッパ係数に関する2つのパラドックス
BIAS, PREVALENCE AND
KAPPA(Byrt, Bishop & Carlin, 1993)**より**

①Bias
不均一なf1,f2,g1,g2はより高いカッパ係数を生み出す
たとえば…ObserverAはNoの割合が高い一方、ObserverBはYesの割合が高いとバランスのとれたクロス表よりもκ係数は高くなる

②Prevalence
偶然による一致率 (f1*g1+f2*g2) /N*N**がおおきいと、見かけ上の一致率がたとえ同じでも、κ係数は低くなる**

Table 3

		Observer A		
		Yes	No	Total
Observer B	Yes	45	15	60
	No	25	15	40
	Total	70	30	100

Table 4

		Observer A		
		Yes	No	Total
Observer B	Yes	25	35	60
	No	5	35	40
	Total	30	70	100

Table 1

		Observer A		
		Yes	No	Total
Observer B	Yes	40	9	49
	No	6	45	51
	Total	46	54	100

Table 2

		Observer A		
		Yes	No	Total
Observer B	Yes	80	10	90
	No	5	5	10
	Total	85	15	100

BI

		Observer A		
		Yes	No	Total
Observer B	Yes	a	b	g_1
	No	c	d	g_2
	Total	f_1	f_2	N

- Bias Index
  2人の評定者間の"Yes"**の割合の違いを表す。**
  BI = (a+b)/N – (a+c)/N = **(b-c)/N**
  bとcが等しいとき、BI=0
  b=N、c=0**のとき** (**あるいはその逆**) 、BI=1

BI=0
κ=0.17

Table 5

		Observer A		
		Yes	No	Total
Observer B	Yes	40	20	60
	No	20	20	40
	Total	60	40	100

- **右の例→**
  見かけ上の一致率は等しいにもかかわらず、不一致の数に偏りがある (Biasがかかっている) **ために、カッパ係数が高くなってしまっている。**

Table 6

		Observer A		
		Yes	No	Total
Observer B	Yes	40	35	75
	No	5	20	25
	Total	45	55	100

- <u>Biasを修正する方法</u>
  bとcに、(b+c)/2をおく。
  Biasが大きくなるとカッパ係数も大きくなってしまう。
  →平均値を使うことでBiasを下げ、κ係数の価値を高める。
  (κ係数は小さくなる)

BI=0.3
κ=0.24

PI

		Observer A		
		Yes	No	Total
Observer B	Yes	a	b	g_1
	No	c	d	g_2
	Total	f_1	f_2	N

- Prevalence Index
  “Yes”と”No”の割合の違いを表す
  →PI = (a-d)/N
  a=0、d=Nのとき、PI=-1
  a=N、d=0のとき、PI=+1
  “yes”と“No”の割合が同じとき、PI=0

PI=0
K=0.6

- 右の例→
  見かけ上の一致率は等しいにもかかわらず、一致している数のうち、Yesに偏っている(PIの絶対値が大きい)ために、カッパ係数が小さくなってしまっている。

- Prevalenceを修正する方法
  aとdに、(a+d)/2をおく。

Table 7

		Observer A		
		Yes	No	Total
Observer B	Yes	40	10	50
	No	10	40	50
	Total	50	50	100

Table 8

		Observer A		
		Yes	No	Total
Observer B	Yes	70	10	80
	No	10	10	20
	Total	80	20	100

PI=0.6
K=0.375

PABAK (PREVALENCE AND BIAS ADJUSTED KAPPA)

- PABAKはBiasとPrevalenceを調整したカッパ係数である。
  Biasを修正 → bとcに、(b+c)/2 = nをおく。
  Prevalenceを修正 → aとdに、(a+d)/2 = mをおく。

		Observer A	
		Yes	No
Observer B	Yes	n	m
	No	m	n

- 偶然の一致の確率は、計算すると0.5になる。
- 見かけ上の一致率は2n/N

- よって、

$$\text{PABAK} = \frac{(2n/N) - 0.5}{1 - 0.5} = 2p_o - 1.$$

となる。

- PABAKは-1から1までの値をとり、0のとき、観察者の一致は50%となる。

カッパ係数とPABAKの関係

- BIが大きくなると、κは大きくなる。
- PIの絶対値が大きくなると、κは小さくなる。
  →BIやPIの大きさによって、
  　κ係数は大きくなったり小さくなったりする。

$$K = \frac{PABAK - PI^2 + BI^2}{1 - PI^2 + BI^2}.$$

カッパ係数とPABAK

- Biasの影響
  PABAKが0.8のとき、Biasの影響はほとんどないが、PABAKがゼロに近いとき、BIは最大0.5まで値をとる
  →PABAKとカッパ係数の差は最大0.2生まれる。(つまり、PABAKよりも高いカッパ係数の数値が出る可能性がある)

- Prevalenceの影響
  PABAKは0.8でも、PIが高いとカッパ係数は低くなる可能性がある。

- カッパ係数のみの報告ではミスリーディングの可能性も出てきてしまう
  →調査者はBiasやPrevalenceの影響も議論するべきであるといえる。

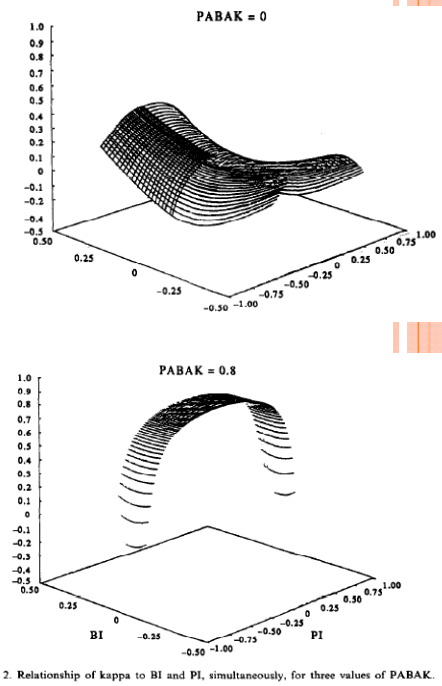


Fig. 2. Relationship of kappa to BI and PI, simultaneously, for three values of PABAK.

引用文献

- **対馬栄輝**(2007). SPSSで学ぶ医療系データ解析 東京図書株式会社
- **対馬栄輝 統計学資料② 信頼性指標としての級内相関係数** http://www.hs.hirosaki-u.ac.jp/~pteiki/research/stat/icc.pdf
- **平井明代 (編)** (2012). **教育・心理系研究のためのデータ分析入門ー理論と実践から学ぶ**SPSS**活用法 東京図書株式会社**
- Byrt, T., Bishop, J., and Carlin, J.B. (1993). Bias, prevalence and kappa. *J. Clin. Epidemiol., 46,* 423-429
- **井上俊哉**(2011). **シリーズ臨床心理学研究と統計学** ***東京家政大学附属臨床相談センター紀要***, *6,* 73-77
- **下井研究室** http://shimoi.iuhw.ac.jp/hout_lec_reliability_210422.pdf