

# 信頼性評価指標としての $\kappa$ 係数の性質について

武田薬品工業株式会社 医薬開発本部 統計解析部 統計グループ

関根 直樹

## 1.はじめに

測定・評価を通じて得られたデータには、通常、誤差が含まれており、その誤差が無視できないほど大きい場合、データから得られた結果の「信頼性」が疑われることになる。そのため、データから適切な結果を導こうとする場合、妥当性ととも信頼性が十分吟味された測定・評価方法を用いる必要があり、実際、ICH ガイドライン「臨床試験のための統計的原則」でも、信頼性の確立された変数を主要変数として用いるよう求めている。

信頼性(関連性)を評価する方法としては、評価指標が幾つか提案されているが、その中で特に頻繁に用いられているものの1つに「 $\kappa$  係数」なる指標がある。 $\kappa$  係数は、提案されている他の指標よりは、信頼性の評価指標として望ましい性質を持つものの、評価指標として好ましくないと思われる特徴も内包している。一方、「信頼性」という概念そのものが測定・評価方法を評価する上で必ずしも適切かつ十分と言える指標ではなく、その意味でも  $\kappa$  係数の解釈には十分注意する必要がある。

そこで、ここではまず「信頼性」の概念について簡単に触れた後、信頼性評価指標としての適切性について  $\kappa$  係数と他の指標を比較した場合、 $\kappa$  係数の方がより望ましい性質を持つことを示す。その後、 $\kappa$  係数の特徴・問題点について、2 値データの場合に限定して検討した後、信頼性評価指標が持つべき望ましい性質について議論する。また、測定・評価方法について「信頼性」のみ評価することの問題についても触れることとする。

## 2.信頼性 (Reliability) の定義およびその欠如による影響

信頼性 (Reliability) とは、同じ状態の対象に関する測定値は状況に依らず同じになるという特性のことであり、評価者間信頼性 (複数の測定者が測定した場合の結果の一致性)、評価者内信頼性 (同一測定者が異なる時点で測定した場合の結果の一致性) に分類されることが多い。

信頼性の低い測定・評価から得られたデータは、「相関の希薄化」、「症例数の増加」を始めとして治験に様々な悪影響を及ぼす。

## 3.関連性 (Association) と信頼性 (Reliability)

関連性 (Association) とは変数間の従属性 (Dependence) とほぼ同義語であり、ある変数の変動に伴い、他の変数も変動する性質のことである。一方、信頼性 (Reliability) とはある変数同士が同じ値になるという性質のことであり、関連性の

特別な場合であると言える。そのため関連性が高い場合に必ずしも信頼性が高いとは限らない。例えば、測定者 A、B の測定値が高い相関を示していたとしても、A の測定値が B の測定値よりも全体的に高い場合、測定値間の関連性は高いが、信頼性は低いということになる。

このように、関連性と信頼性という概念は混同されることが多いうえ、同時にこれらの評価指標も混同されることが少なくないので注意が必要である。

#### 4. 代表的な関連性、信頼性の評価指標

K カテゴリーの分類尺度を用いて、評価者 A、B が同一の対象を評価した場合、そのデータは K×K クロス表形式で表現することができる（図表 1）。分類尺度データにおける評価の関連性（信頼性）を評価する代表的な指標には、次のようなものがある。

##### χ<sup>2</sup>系の評価指標（χ<sup>2</sup>検定）

$$\hat{\phi} = \sum_i \sum_j^K (np_{ij} - np_{i.}p_{.j})^2 / n^2 p_{i.}p_{.j}$$

を基本的な構造としてもつ関連性評価指標。Cramer's V、Pearson's C 等がある。

##### オッズ比（2×2 分割表の場合）

$$\hat{\phi} = p_{11}p_{22} / p_{12}p_{21}$$

##### 一致率

$$\hat{P}_o = \sum_i^K p_{ii}$$

##### κ 係数

$$\hat{\kappa} = \frac{\hat{P}_o - \hat{P}_e}{1 - \hat{P}_e} = \frac{\sum_i^K p_{ii} - \sum_i^K p_{i.}p_{.i}}{1 - \sum_i^K p_{i.}p_{.i}}$$

図表 1. K×K クロス表における各セルの発現割合

|      |                | 評価者B            |    |    |                 |    |                 |                 |
|------|----------------|-----------------|----|----|-----------------|----|-----------------|-----------------|
|      |                | C <sub>1</sub>  | .. | .. | C <sub>j</sub>  | .. | C <sub>K</sub>  |                 |
| 評価者A | C <sub>1</sub> | p <sub>11</sub> | .. | .. | p <sub>1j</sub> | .. | p <sub>1K</sub> | p <sub>1.</sub> |
|      | :              | :               |    |    | :               |    | :               | :               |
|      | C <sub>i</sub> | p <sub>i1</sub> | .. | .. | p <sub>ij</sub> | .. | p <sub>iK</sub> | p <sub>i.</sub> |
|      | :              | :               |    |    | :               |    | :               | :               |
|      | C <sub>K</sub> | p <sub>K1</sub> | .. | .. | p <sub>Kj</sub> | .. | p <sub>KK</sub> | p <sub>K.</sub> |
|      |                | p <sub>.1</sub> | .. | .. | p <sub>.j</sub> | .. | p <sub>.K</sub> | 1 (n)           |

χ<sup>2</sup>系の評価指標は期待値からのズレを評価する指標であり、必ずしも一致性を評価している指標ではない。オッズ比、一致率はともに一致度が高くなれば、指標値も高くなるため、信頼性評価指標となりうる。κ 係数も一致度が高くなれば指標値も高くなる傾向を示すが、更に偶然による一致度も考慮しているという点で、オッズ比、一致率よりも望ましい指標であると言える。

#### 5. κ 係数の性質

信頼性評価指標として頻繁に用いられる κ 係数には幾つか特記すべき特徴がある。ここでは、その特徴について、特に 2 カテゴリーのデータに限定して検討する。

### 5.1.反応率 $\theta$ への依存性

ある対象について、評価者 A が反応「有」と判定した場合  $i=1$ 、反応「無」と判定した場合  $i=2$  とする。評価者 B の場合も同様に  $j=1, 2$  とする。この時、同一対象について評価者 A が  $i$ 、かつ B が  $j$  と判定する確率を  $\pi_{ij}$  とすると、 $\pi_{ij}$  は図表 2 に示すように  $2 \times 2$  クロス表形式で表現することができる。更に、反応が「有」である真の割合を  $\theta$ 、反応「有」である症例を評価者 A、B がともに「有」と判定する確率（感度）を  $1-\beta$ 、反応「無」である症例をとともに「無」と判定する確率（特異度）を  $1-\alpha$  とする。

図表 2.  $2 \times 2$  クロス表における各セルの発現確率

|      |    | 評価者B            |                 | 合計             |
|------|----|-----------------|-----------------|----------------|
|      |    | 有               | 無               |                |
| 評価者A | 有  | $\pi_{11}$      | $\pi_{12}$      | $\pi_{1\cdot}$ |
|      | 無  | $\pi_{21}$      | $\pi_{22}$      | $\pi_{2\cdot}$ |
|      | 合計 | $\pi_{\cdot 1}$ | $\pi_{\cdot 2}$ | $\pi_{..}=1$   |

この時、各セルの発現確率に対し、(a)のモデルが成り立つと仮定し、 $\kappa$  係数が(b)のように定義されるとする。

$$\begin{aligned}
 \pi_{11} &= \theta(1-\beta)^2 + (1-\theta)\alpha^2 \\
 \pi_{12} &= \theta\beta(1-\beta) + (1-\theta)\alpha(1-\alpha) \\
 \pi_{21} &= \theta\beta(1-\beta) + (1-\theta)\alpha(1-\alpha) \\
 \pi_{22} &= \theta\beta^2 + (1-\theta)(1-\alpha)^2
 \end{aligned} \tag{a}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{(\pi_{11} + \pi_{22}) - (\pi_{1\cdot}\pi_{\cdot 1} + \pi_{2\cdot}\pi_{\cdot 2})}{1 - (\pi_{1\cdot}\pi_{\cdot 1} + \pi_{2\cdot}\pi_{\cdot 2})} = 1 - \frac{\alpha(1-\beta)}{\alpha + (1-\alpha-\beta)\theta} - \frac{\beta(1-\alpha)}{1-\alpha - (1-\alpha-\beta)\theta} \tag{b}$$

(b)より、 $\kappa$  係数は評価・測定方法の特徴を表すパラメータ  $\alpha$ 、 $\beta$  だけでなく、その特徴とは無関係であるパラメータ  $\theta$  に依存する指標であることが分かる。

### 5.2.一致・不一致セルにおける偏りの影響

同一対象について評価者 A が  $i$ 、かつ B が  $j$  と判定した実際の割合を  $p_{ij}$  とすると、そのデータは図表 3 に示すように  $2 \times 2$  クロス表型式で表現され、 $\kappa$  係数は(c)により算出される。

図表 3.  $2 \times 2$  クロス表における各セルの発現割合

|      |    | 評価者B          |               | 合計           |
|------|----|---------------|---------------|--------------|
|      |    | 有             | 無             |              |
| 評価者A | 有  | $p_{11}$      | $p_{12}$      | $p_{1\cdot}$ |
|      | 無  | $p_{21}$      | $p_{22}$      | $p_{2\cdot}$ |
|      | 合計 | $p_{\cdot 1}$ | $p_{\cdot 2}$ | $p_{..}=1$   |

$$\hat{\kappa} = \frac{\hat{P}_o - \hat{P}_e}{1 - \hat{P}_e} = \frac{(p_{11} + p_{22}) - (p_{1.}p_{.1} + p_{2.}p_{.2})}{1 - (p_{1.}p_{.1} + p_{2.}p_{.2})} \quad (c)$$

ここで、図表 4～図表 6に示す 3つのデータが得られたとする。これらの一致率は全て 0.6 と等しいが、 $\kappa$  係数は全て異なっている。図表 4に比べ図表 5では、評価者 A,B の評価が一致したセル  $\{(1,1), (2,2)\}$ 、すなわち対角セルの割合が一方(1,1)に偏っており、 $\kappa$  係数は低い値を示している。一方、図表 6では評価が一致しなかったセル  $\{(1,2), (2,1)\}$ 、すなわち非対角セルの割合が一方(1,2)に偏っており、 $\kappa$  係数は高い値を示している。つまり、一致率が一致している場合でも、このような偏りが存在すると、 $\kappa$  係数は異なる値を示す。Byrt,et all.(1993)は、対角セルおよび非対角セルにおける割合の偏りについて、それぞれ Prevalence Index(PI)、Bias Index(BI)なる概念を導入し、 $\kappa$  係数が  $P_o$ 、PI、BI を用いて次式のように定式化できることを示した。

$$\hat{\kappa} = \frac{(2\hat{P}_o - 1) - PI^2 + BI^2}{1 - PI^2 + BI^2}, \quad PI = p_{11} - p_{22}, \quad BI = p_{12} - p_{21} \quad (d)$$

同様に Lantz,et all.(1996)は、観測された一致率の下での  $\kappa$  係数の最小値  $\kappa_{\min}$ 、最大値  $\kappa_{\max}$ 、および偏りが存在しない場合の  $\kappa$  係数  $\kappa_{\text{nor}}$  が次式により与えられることを示した。

$$\hat{\kappa}_{\min} = \frac{\hat{P}_o - 1}{\hat{P}_o + 1}, \quad \hat{\kappa}_{\max} = \frac{\hat{P}_o^2}{(1 - \hat{P}_o)^2 + 1}, \quad \hat{\kappa}_{\text{nor}} = 2\hat{P}_o - 1 \quad (e)$$

Byrt,et all.(1993)、Lantz,et all.(1996)はいずれも、一致セルおよび不一致セルの偏りを評価するために、 $\kappa$  係数と併せて PI、BI、または  $\kappa_{\min}$ 、 $\kappa_{\max}$ 、 $\kappa_{\text{nor}}$  を算出することを薦めている。

図表 4. 2×2 クロス表データ 1

|      |    | 評価者B |     | 合計  |
|------|----|------|-----|-----|
|      |    | 有    | 無   |     |
| 評価者A | 有  | 0.3  | 0.2 | 0.5 |
|      | 無  | 0.2  | 0.3 | 0.5 |
|      | 合計 | 0.5  | 0.5 | 1   |

$P_o = 0.600$   
 $\kappa = 0.200$

図表 5. 2×2 クロス表データ 2

|      |    | 評価者B |      | 合計   |
|------|----|------|------|------|
|      |    | 有    | 無    |      |
| 評価者A | 有  | 0.55 | 0.2  | 0.75 |
|      | 無  | 0.2  | 0.05 | 0.25 |
|      | 合計 | 0.75 | 0.25 | 1    |

$P_o = 0.600$   
 $\kappa = -0.067$

図表6. 2×2クロス表データ3

|          |    | 評価者B |      | 合計   |
|----------|----|------|------|------|
|          |    | 有    | 無    |      |
| 評価者<br>A | 有  | 0.3  | 0.35 | 0.65 |
|          | 無  | 0.05 | 0.3  | 0.35 |
|          | 合計 | 0.35 | 0.65 | 1    |

$P_0 = 0.600$   
 $\kappa = 0.266$

## 6. $\kappa$ 係数における問題点

「 $\kappa$  係数の性質」に示したように「反応率  $\theta$  に依存していること」、「推定値の算出において一致・不一致セルにおける偏りが影響を与えること」が、 $\kappa$  係数の大きな特徴であるといえるが、 $\kappa$  係数を信頼性評価指標として捉えた場合、の性質は好ましいものとはいえない。なぜなら評価・測定方法を評価する指標はその評価・測定方法の特徴を表すパラメータのみに依存すべきであり、その特徴とは無関係であるパラメータ  $\theta$  に依存すべきではないと考えられるためである。しかし、図表3に示す 2×2 クロス表データしか利用できない状況において、 $\theta$  に依存しない信頼性評価指標やその推定量を定義することは非常に困難であると思われる（信頼性の指標ではないが、例えば、(a)の  $\alpha$ 、 $\beta$  を  $\theta$  が未知の下で推定することは難しい）。

一方、 $\kappa$  係数は(d)式にあるように  $P_0$  のみの関数ではないため、 $P_0$  と  $\kappa$  係数を比較した場合、解釈上、大きな差が生じる場合がある。よって  $\kappa$  係数の安定性を評価する意味では の性質を考慮して、PI、BI、または  $\kappa_{\min}$ 、 $\kappa_{\max}$ 、 $\kappa_{\text{nor}}$  等を算出することは重要であると考えられる。

## 7. 信頼性評価指標のみで測定・評価方法を評価することの問題

「信頼性」とは評価の「一致性」を表す概念であるが、信頼性が高いということだけでその測定・評価方法が良いと判断することは適切ではない。例えば式(a)において、評価者 A、B の評価が一致したセル{(1,1)、(2,2)}におけるモデル式の項  $(1-\theta)\alpha^2$ 、 $\theta\beta^2$  は、実際には反応「無」であるが A、B とともに反応「有」と評価してしまう確率、実際には反応「有」であるが A、B とともに反応「無」と評価してしまう確率をそれぞれ示している。つまり  $\pi_{11}$ 、 $\pi_{22}$  には、真の状態と両評価者の評価がすべて一致する確率だけでなく、真の状態とは異なるが両評価者の評価は一致する確率も含まれることになる。 $\kappa$  係数を始めとする信頼性評価指標の多くは  $\pi_{11}$ 、 $\pi_{22}$  の観測値である  $p_{11}$ 、 $p_{22}$  に基づき評価するため、真の状態とは異なる場合であっても評価が一致している割合が高ければ、信頼性は高いと判断されることになる。

このように、測定・評価方法を評価するに際して、信頼性は1つの指標になりうるが、信頼性を評価するだけでは十分とは言えない。通常は信頼性だけでなく、目的とする事象を正しく測っているかどうかという性質、すなわち「妥当性」も十

分に評価する必要がある。しかし、信頼性と比較して妥当性の概念は非常に多岐に渡っており、「妥当性」を評価するための方法論も確立しているとは言い難い。そのため、信頼性のみを評価すること（ $\kappa$ 係数を算出すること）に終始しがちであるが、少なくとも信頼性は一致性のみを評価している概念であり、信頼性を評価するだけでは不十分であることに注意すべきであると言える。

## 参考文献

- Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* **46**, 293-302.
- Byrt, T., Bishop, J., and Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology* **46**(5), 423-429.
- 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.
- Fliess, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378-382.
- Fliess, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edition. New York: Wiley.
- Fliess, J. L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.
- Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics* **36**, 207-216.
- 楠 正 (1996). 客観的指標と主観的評価の考え方. 臨床薬理 **27**(2), 499-500.
- Lantz, C. A. and Nebenzahl, E. (1996). Behavior and interpretation of the  $\kappa$  statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology* **49**(4), 431-434.
- 小田 英世 (1997). 評価尺度の信頼性と妥当性, 8Bios テキスト, 日本科学技術連盟.
- Thompson, W. D. and Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* **41**(10), 949-958.
- Walter, S. D., Eliasziw, M. and Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine* **17**, 101-110.

## 連絡先

〒540-8645 大阪府大阪市中央区道修町4丁目1番1号

武田薬品工業株式会社 医薬開発本部 統計解析部 統計グループ 関根直樹

Tel. 06-6204-2271 Fax. 06-6204-2368 E-mail. Sekine\_Naoki@takeda.co.jp