

統計学資料②

信頼性指標としての級内相関係数

弘前大学 医学部

保健学科 理学療法学専攻

対馬 栄輝

目次

第 1 章 検者間・検者内信頼性指標としての級内相関係数	1
1.1 はじめに	1
1.2 基礎的理論	1
1.2.1 対象データの特性	1
1.2.2 古典的テスト理論	2
1.2.3 一般化可能性理論	4
1.2.4 母数モデル，変量モデル，混合モデル	4
1.2.5 分散値の加法性	5
1.2.6 交互作用	5
1.2.7 平均平方和の期待値	6
1.3 級内相関係数 (ICC)	8
1.3.1 Case1	9
1.3.2 Case2	10
1.3.3 Case3	12
1.4 ICC 適用の留意点	15
1.4.1 ICC (1, 1), ICC (2, 1), ICC (3, 1) の大きさの関係	15
1.4.2 Case1 と Case2 の違い	15
1.4.3 ICC の判定基準	15
1.4.4 測定回数・標本の大きさの決定	16
1.4.5 統計的検定の方法	16
1.4.6 標準誤差の算出	17
1.5 シミュレーションによる特性の検討	18
1.5.1 推定値の性質	18
1.5.2 信頼区間の性質	20
1.5.3 データの特性によってどのような値をとるか	20
1.6 データ解析の実際	23
1.6.1 解析手順－検者間信頼性を求める－	23
1.6.2 データの観察	25
1.6.3 測定回数の決定	26

第1章 検者間・検者内信頼性指標としての級内相関係数

1.1 はじめに

ある検査の検者間または検者内信頼性の指標として、最近では級内相関係数 (*Intraclass correlation coefficients*; 以下, ICC^a) が用いられる。もちろん ICC 以外にも信頼性を表す指標は様々存在するし, ICC のみから信頼性を検討できるわけでもない。統計学分野の著書では信頼性と ICC を別記しており [1] [2], 恐らく“ 級内相関係数 ”という用語を最初に用いた Fisher の著書 [3] の級内相関の章では“ 信頼性 ”という用語は記載されていない。信頼性の指標の一つとして ICC が利用できるという解釈が妥当であろう。

Shrout [4] によると, ICC には3つの形式があり合計6つの公式が提唱されている。現在のところ, Shrout の分類に基づいて使用している研究・解説論文 [5] [6] [7] [8], 著書 [9] は多い。このことから, 本章では Shrout の分類に従って解説していくことにする。ICC の3つの形式は Case1, Case2, Case3 と呼び, それぞれの下位モデル (全6公式) は $ICC(n, 1)$, $ICC(n, k)$ と記することができる ($n = 1, 2, 3$)。また, 単に ICC と記すときは6つの公式すべてを指すものとする。なお, Bartko [10] は単に“ ICC ”と呼称し, One-way Classification, Two-way Random Model, Two-way Mixed Model の3つの場合に分類している。これらは前述の Case1, Case2, Case3 に相当する。また, 後に Bartko [11] は $ICC(1)$, $ICC(2)$, Two-way, Winer's anchor point と名称を変えている。このように ICC は一貫した記載方法が示されていないから, どの記載方法が正しいとはいえない現状であることを留意しておく。

(本文中, ※補足, ※注, 部分は若干専門的理解が必要なので, 無視して構わない。)

1.2 基礎的理論

1.2.1 対象データの特性

ICC は, 平均値や分散値の特性を利用する分散分析 [→他章を参照] の結果を活用している。従って, 分散分析の適用条件,

1. データは無作為抽出による。
2. データは正規分布に従う。
3. 各要因の水準間で等分散性が保証されている。

が満たされているのが理想である。これらのうち, 少なくとも 1. と 2. の条件は満たされている必要がある。2. に関しては離散変数であっても正規分布に従えば, 名義尺度や順序尺度でない限り適用可能である。3. の条件は, ICC を算出する上で必ずしも保証されていなければならないものではない。

^a 統計学分野では項目特性曲線 (item characteristic curve) の略称を ICC として用いるため, 級内相関係数を ICC と略すのは一般的ではないゆえに混乱を招く。論文などで利用する際は, まず級内相関係数と断ってその後に略す必要がある。

1.2.2 古典的テスト理論

古典的テスト理論 (classical test theory) は 1950 年代までに体系づけられ、教育学分野、心理学分野のテストで基礎となっている理論である。対して、項目応答理論 (item response theory; IRT) は現代テスト理論と呼ばれ、古典的テスト理論の問題を解決すべく 1950 年代以後に発展してきたものである。

信頼性モデルの構造、理論は後に述べるとして、ここでは古典的テスト理論で用いる代表的な信頼性係数の推定方法を述べておく。なおこれらの数理的な詳細については解説しない。

1.2.2.1 再検査法 (test-retest method)

同じ被検者を対象に数回の検査を繰り返して、その一貫性を見ようとするのが再テスト信頼性である。この推定には再テスト信頼性 (test-retest reliability) が用いられる。繰り返しテスト間の真の値と誤差の分散は同じであると仮定して検査間の相関係数を求める。しかし、計算の前提として“繰り返し測定間の誤差の共分散を 0 とする”といった条件が付く^b。同一の対象を繰り返し測定するといった場合は、この条件を受け入れることが難しくなるため、問題が生じてしまう。この使用に関しては否定的意見が多い。

1.2.2.2 平行テスト法 (parallel test method)

内容的に同等と認められる複数の検査を繰り返し適用させてそれらの信頼性を検討する時は、平行測定法による信頼性 (parallel form reliability) が用いられる。これには強平行 (strongly parallel measurement; 厳密平行とも呼ばれる) と弱平行 (weakly parallel measurement) があり、前者はテスト間の被検者の真の値が同一であり、誤差分散は等しいという仮定が存在し、後者は真の値は異なってもよいとする点で異なる。なお、弱平行によって求められた値は Cronbach の α 係数 [→ 1.2.2.4 節参照] や ICC (3, 1) と一致する。

1.2.2.3 折半法 (split-half method)

上述の問題を解決すべく考えられた折半法は、検査項目の内容を 2 つに分割して、その 2 つの評価項目ごうしの信頼性を求める手順である。平行テスト法と比較して、2 つのテストを同時に施行できるという利点がある。特に、偶数番号の項目と奇数番号の項目を 2 つに分ける奇偶法 (odd-even method) はよく用いられる。計算には Spearman-Brown の公式を用いる。これをもとに一般化した公式として Kuder-Richardson の公式 20 (KR-20) と Kuder-Richardson の公式 21 (KR-21) が導かれた。下位項目が 2 項分布に従うデータ (0-1 型のダミーデータ) である場合は KR-20 を、下位項目間で正答率が等しい (難易度が均等な) 場合は KR-21 を利用することができる。

1.2.2.4 α 係数 (Cronbach's coefficient alpha)

上記は Cronbach の α 係数へ拡張された。 α 係数は内部一貫性または内的整合性を求めるものである^c。計算も簡単であり利用される機会が多い。評価法 A が下位項目数 n で構成されているとき、 n 個項目の内

^b 現在は、この仮定を必要としない方法も存在する。

^c 信頼性そのものではないことに注意。高い内的整合性が高い信頼性を意味するのではない。

的整合性の下限値を求めることができる。なお、ICC (3, k) は α 係数と一致する。

※補足

①－古典的テスト理論での仮定条件

測定値 X_i が真の値 T_i と誤差 E_i で構成されているとすると、古典的テスト理論では以下の2つの仮定が必要とされる。

1) 誤差の平均は0である。

$$E(E_i) = 0 \quad (2.1)$$

従って,

$$\mu(X) = E(X_i) = E(T_i + E_i) = E(T_i) + E(E_i) = \mu(T) \quad (2.2)$$

2) 誤差と真の値は独立である。

$$Cov(T, E) = \rho(T, E) = 0 \quad (2.3)$$

従って,

$$\begin{aligned} \sigma^2(X) &= Var(T) + Var(E) + 2Cov(T, E) \\ &= Var(T) + Var(E) \end{aligned} \quad (2.4)$$

以上より信頼性係数 ρ は,

$$\rho(X) = \frac{Var(T)}{Var(X)} \quad (2.5)$$

と表せる。古典的テスト理論には、現実には上記の仮説に適用できないということと、誤差分散を単一項でまとめてしまっているという問題がある。

②－Spearman-Brown の公式

2つ（または2つ以上）の下位テストを X_1, X_2 とし、 T_1, T_2 を真の値、 E_1, E_2 を誤差とする。なお、 X_1, X_2 は強平行測定であると仮定する。従って、 $X = (X_1 + X_2)/2, T = (T_1 + T_2)/2$ である。 T_1, T_2 は強平行測定だから、

$$Var(X_1) = Var(X_2) \quad (2.6)$$

$$Var(T_1) = Var(T_2) \quad (2.7)$$

である。ところで $X_1(= (X_{11}, X_{12}, \dots, X_{1i}))$, $X_2(= (X_{21}, X_{22}, \dots, X_{2i}))$ の共分散の性質は、

$$\begin{aligned} Cov(X_1, X_2) &= E \left[\sum_i (X_{1i} - \mu(X_1)) \left(\sum_i X_{2i} - \mu(X_2) \right) \right] \\ &= E \left[\sum_i (T_i - \mu(T) + E_{1i}) \left(\sum_i T_i - \mu(T) + \sum_i E_{2i} \right) \right] \\ &= E \left[\sum_i (T_i - \mu(T))^2 + \sum_i (T_i - \mu(T)) \sum_i E_{2i} + \sum_i E_{1i} \sum_i (T_i - \mu(T)) + \sum_i E_{1i} E_{2i} \right] \\ &= Var(T) \end{aligned} \quad (2.8)$$

と変形できる。これと、

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{Var(X_1)Var(X_2)} = \frac{Cov(X_1, X_2)}{2Var(X_1)} \quad (2.9)$$

を利用して, (2.5) 式は,

$$\begin{aligned}
 \rho(X) &= \frac{Var(T)}{Var(X)} = \frac{Var(T_1 + T_2)}{Var(X_1 + X_2)} = \frac{Var(T_1) + Var(T_2) + 2Cov(T_1, T_2)}{Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)} \\
 &= \frac{2Var(T_1) + 2Cov(T_1, T_2)}{2Var(X_1) + 2Cov(X_1, X_2)} = \frac{4Var(T_1)}{2Var(X_1)[1 + \rho(X_1, X_2)]} \\
 &= \frac{2 \cdot 2Cov(X_1, X_2)}{2Var(X_1)[1 + \rho(X_1, X_2)]} \\
 &= \frac{2\rho(X_1, X_2)}{1 + \rho(X_1, X_2)} \\
 &= \frac{2\rho(X_1)}{1 + \rho(X_1)} \tag{2.10}
 \end{aligned}$$

に変形される。(2.10) 式が Spearman-Brown の公式である。しかし、上述した仮定が必要とされることに注意しなければならない。

1.2.3 一般化可能性理論

古典的テスト理論の延長として一般化可能性理論 (generalizability theory) がある。一般化可能性理論とは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である。被検者個人のまたは検者の差異、課題差などの誤差が入り組む状況の信頼性研究では、どこまで一般化が可能であるかといった理論的基礎を与える一般化可能性理論が要される。テストの信頼性とは、“テストが目的とする対象をどの程度精密に測定できるかを示す指標 [2]”であり、算出される信頼性の指標は“一般化可能性係数”と呼称される。ICC はその一般化可能性係数の一部に属する。

一般化可能性理論では、分散を推定する実験計画の過程を“G 研究 (generalizability study; 一般化可能性研究)”, 分散分析の影響を推定し、適切なテスト使用計画を立てる過程を“D 研究 (decision study; 決定研究)”と呼ぶ。例えば、A という評価法の信頼性を知りたいとしよう。無作為に選ばれた検者によって A を測定し、測定の“ばらつき (分散)”を求めることが G 研究である。ICC (1, 1), ICC (2, 1), ICC (3, 1) 式は G 研究として用いられる。A のばらつきが推定できたら、次に複数検者の評点の平均値を利用した A の信頼性を求める (D 研究)。D 研究では特定の検者または被検者を対象として“信頼性”を検討する。ICC (1, k), ICC (2, k), ICC (3, k) 式は D 研究のために用いる。まず G 研究として A 評価法の測定を行い、ばらつきがどの程度か推定してから、D 研究として検者の平均値を利用して信頼性を検討する。

このように信頼性の検討をする際には、解析の前に“データはどのように構成されていて、その中での何をみたいのか”という目的を明確に決めておかないと、手法選択の決定が困難となり、誤った結果を招くことになる。

1.2.4 母数モデル, 変量モデル, 混合モデル

分散分析には母数モデル (fixed model; 固定モデルとも呼ばれる), 変量モデル (random effect model), 混合モデル (mixed model) があり, それぞれで平均平方和の期待推定値が異なる。

母数モデルは, 要因を母数要因としている場合である。母数要因とは, 要因の水準が固定されて変化しな

いときである^d。例えば、季節要因の水準“春、夏、秋、冬”は母数要因である。別の例を挙げると、地域の要因として“日本、米国、英国”といった水準を規定する場合である。

変量モデルは、要因を確率変数として扱えるとき^eである。例えば、被検者要因と検者要因の2元配置分散分析を行うとき、被検者要因を“脳卒中患者”とする。脳卒中患者であれば年齢、性別、入院・外来に関わりなく対象にできるならば変量要因である。検者要因も“理学療法士”であれば経験年数、所属施設に関わりなく誰でも選ばれるというように、要因内の水準を広く一般的に適用できるときは変量要因と考える。しかし、被検者要因であっても、入院施設・時期を限定したり、年齢を限定すると母数要因となる^f。

混合モデルとは、2元配置以上の分散分析を用いるときに一方の要因が母数要因で他方が変量要因であるような場合をいう。信頼性のデザインでは、検者要因を“施設別の理学療法士”とし、その水準を“A施設、B施設、C施設”と定め、被検者要因を“脳卒中患者”とした場合が混合モデルに相当する。もちろん、それぞれ検者自身が所属する施設の被検者のみを対象とすれば、被検者要因も母数要因になってしまう。しかし、あらかじめ他の施設の患者も含めた母集団を想定しているのであれば混合モデルと考えても間違いではない。以上のモデルの扱いについては研究者の考え方、解析の目的によって変化する。

1.2.5 分散値の加法性

分散分析に基づく ICC では、この性質を基本として各分散値を加減する。

要因 i 水準の j 番目のデータ x_{ij} ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$) が与えられたとき、 $x_{11}, x_{21}, \dots, x_{nm}$ において水準ごとの平均値を $\bar{x}_i$ 、総平均値を $\bar{x}$ とする。このとき、

$$(x_{ij} - \bar{x}) = (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i) \quad (2.11)$$

の関係がある。両辺を2乗したその和は、

$$\begin{aligned} (x_{ij} - \bar{x})^2 &= (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)^2 \\ (x_{ij} - \bar{x})^2 &= (\bar{x}_i - \bar{x})^2 + (x_{ij} - \bar{x}_i)^2 + 2(\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_i) \\ \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x})^2 &= \sum_{i=1}^n (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2 + 2 \sum_{i=1}^n (\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_i) \quad (2.12) \end{aligned}$$

となる。ここで、(2.11) 式の右辺最右項は水準 i に属するデータから水準 i の平均値を引いた偏差であり、その和は常に0となる^g。従って、(2.12) 式の右辺第3項は0となり、結局、

$$\sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x})^2 = \sum_{i=1}^n (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2 \quad (2.13)$$

となる [→ (1.2.2) 節の「※補足」も参照]。

1.2.6 交互作用

交互作用の存在は二元配置分散分析を基本として求められる Case2 と Case3 において問題となってくる。図 1.1 に交互作用の参考例を挙げる。

^d 水準が固有名詞のときと考えてもよい。

^e 水準が固有名詞以外のときと考えてもよい。

^f 母数要因とするか変量要因とするかは結局、実験者の考え方に委ねることになる。

^g 数理的には $(\bar{x}_i - \bar{x})$ と $(x_{ij} - \bar{x}_i)$ は独立であると仮定し、共分散は0と考える。

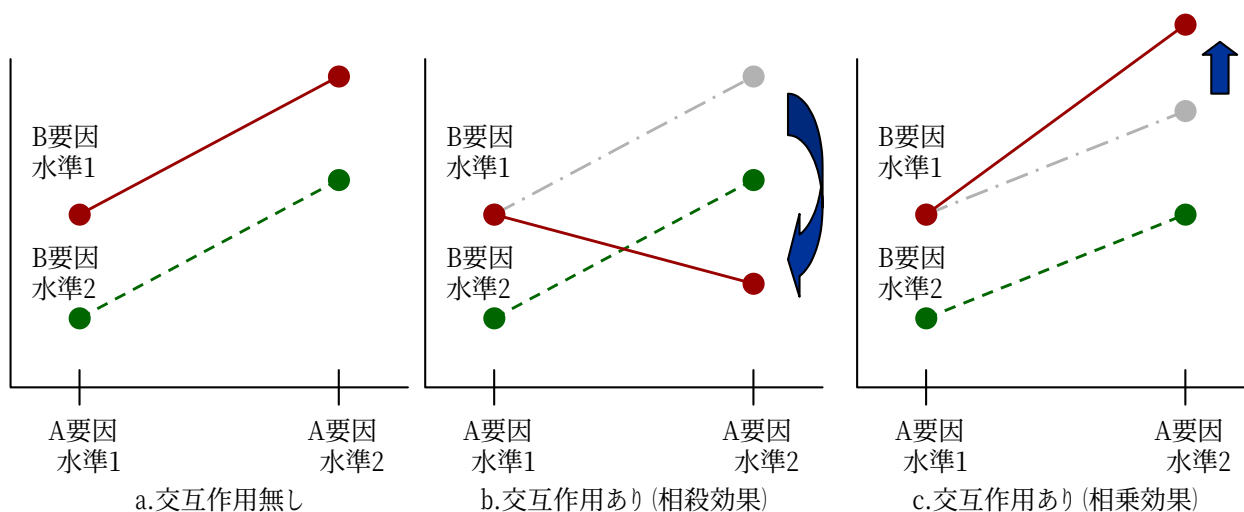


図 1.1 交互作用の例

図 1.1 中 a. は交互作用が存在せず、A 要因と B 要因が独立となっている図である。つまり、A 要因の影響を受けず、B 要因の水準は一定して平行に変化している。

交互作用が存在するとは、b. と c. のように A 要因の水準によって B 要因の少なくとも 1 つの水準は変化のパターンが異なる場合である。これが誤差によるものか A 要因の第 2 水準の影響によるものかについては断言できない。交互作用が存在すると B 要因の単調増加（減少）が確定できなくなり、B (A) 要因の変化の影響で A (B) 要因も変化すると考えられるため、解釈は複雑となる。一般に交互作用は二元配置以上の分散分析で取り上げられ、この存在は主効果を解釈する上でも大きな弊害となる。

1.2.7 平均平方和の期待値

ICC を計算する上で平均平方和の期待値の構成を理解しておくことは重要である。ICC は分散分析で算出される平均平方和の期待値を利用して推定するからである。その平均平方和と期待値は表 1.1 のようにまとめられる。この表の値について特に理論的な面を理解する必要はないが、表に記した値の構造を念頭に置いておくと ICC の理解が平易となる。

表 1.1 各平均平方和の期待値 [4]

変量	df	MS	期待値		
			一要因の変量モデル	二要因の変量モデル	二要因の混合モデル
被検者間	$n - 1$	BMS	$k\sigma_T^2 + \sigma_W^2$	$k\sigma_T^2 + \sigma_I^2 + \sigma_E^2$	$k\sigma_T^2 + \sigma_E^2$
被検者内	$n(k - 1)$	WMS	σ_W^2	$\sigma_J^2 + \sigma_I^2 + \sigma_E^2$	$\theta_J^2 + f\sigma_I^2 + \sigma_E^2$
検者間	$(k - 1)$	JMS	—	$n\sigma_J^2 + \sigma_I^2 + \sigma_E^2$	$n\theta_J^2 + f\sigma_I^2 + \sigma_E^2$
残差	$(n - 1)(k - 1)$	EMS	—	$\sigma_I^2 + \sigma_E^2$	$f\sigma_I^2 + \sigma_E^2$

$f = k/(k-1)$; $\theta_J^2 = a_1^2/(k-1)$ と不偏推定値となっていることに注意

n は被検者数、 k は検者数に該当する。分散分析では表 1.1 中 2 列目の MS — BMS, WMS, JMS, EMS — が求められる。実際にはこれらを利用して ICC を求める。平均平方和の期待値は 3 列目以降に記載されて

いる。一要因の変量モデルとは一元配置分散分析、二要因の変量モデルまたは混合モデルは二元配置分散分析に該当する。

二要因の変量モデルと比べて、二要因の混合モデルでは期待値の中に“ f ”と“ θ ”の係数が付記されている点で異なる。混合モデルでは検者間要因と交互作用（表 1.1 中では“残差”）要因における期待値が母数として与えられるため、不偏推定値 [→他章を参照] としてある^h。なお、Bartko [10] では二要因の混合モデルにおける期待値は表 1.1 と一部異なっている。

^h 変量モデルと母数モデルの違いの特徴である。文献によっては混合モデルにおける EMS を θ_j^2 とせず σ_j^2 とし、 σ_1^2 の係数 f を略記しているものも多い。

1.3 級内相関係数 (ICC)

ICC は Case1, 2, 3 に大別され、さらにそれぞれ 2 つの公式に分けられる。従って、以降で説明する公式は全部で 6 つになる。

6 つの公式を説明する前に、まず、なぜ Case を分けておく必要があるかを簡単に述べる。

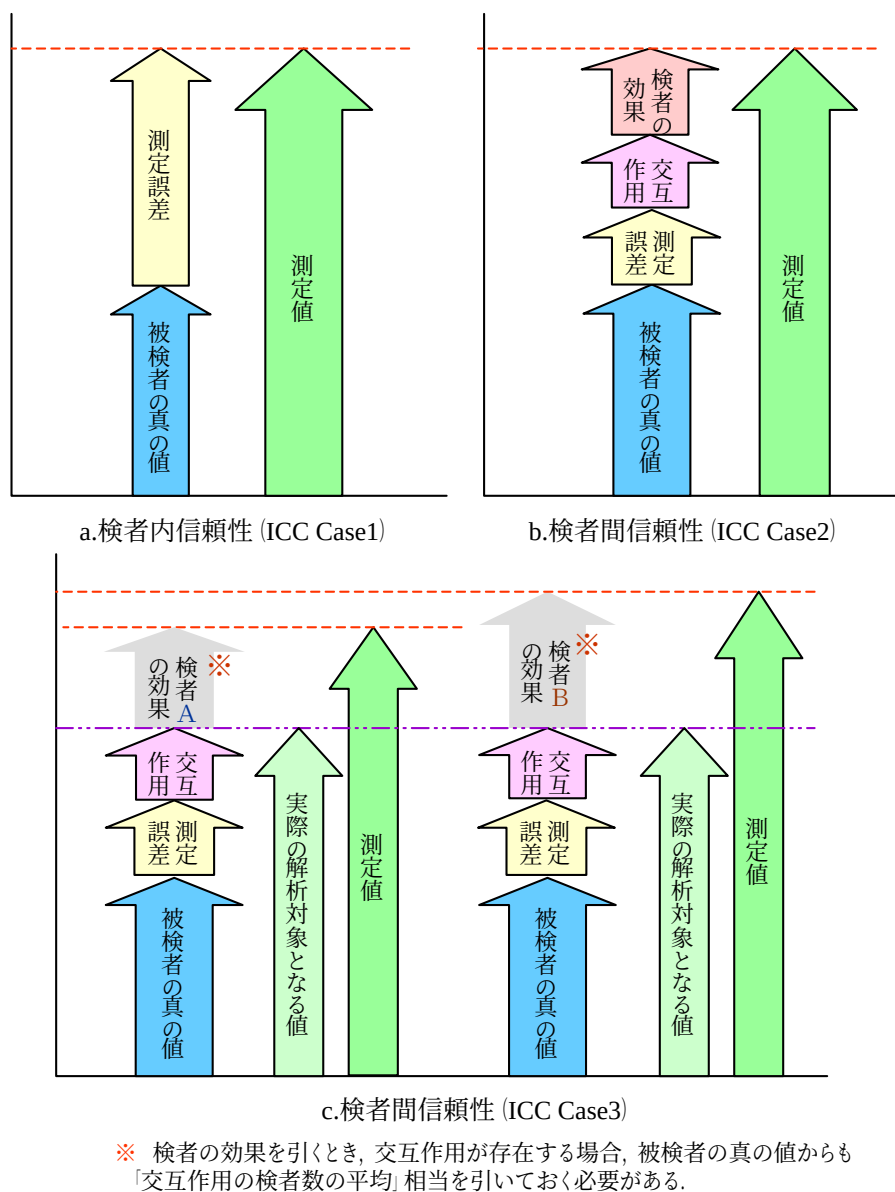


図 1.2 信頼性の構成

図 1.2 の a. は Case1 に相当する。測定値は真の値とその他の誤差により構成されていると考える。その中で真の値の比率を求めることになる。ここでは検者の影響（誤差）は考慮しないため、同一検者による検者内信頼性を求めることになる。

図 1.2 の b. は Case2 に相当し、今度は検者の効果も考慮して患者の真の値の比率を求めている。平均平方和は二元配置分散分析により求められ、検者要因と被検者要因の 2 要因を考慮すると同時に、交互作用

も考慮する必要がある。

図 1.2 の c. は Case3 である。検者間信頼性を求めるものとして Case2 と同義であるが、検者要因が母数要因となっている点で異なるⁱ。Case2 は変量モデルであるから、あらゆる検者が母集団となり、どのように選ばれてもよいとする。Case3 では固定した特定の母集団、例えば A 病院勤務の医師と B 病院勤務の医師間の信頼性比較や、C 病院勤務の理学療法士のみと設定したときの信頼性を比較する。ただし、Case3 では常に検者は同一者と限らない。例えば“ 検者は理学療法士養成校卒業 3 年目 ”という母集団を想定すれば、検者は卒後 3 年目の理学療法士であれば誰でもよい。

Case3 で求める値は、検者間の効果を除いた被検者の真の値の比率である。A 評価表を用いて被検者 a, b を測定するとき、真の値よりも一定量過大評価する傾向のある検者 I と一定量過小評価する傾向のある検者 II、評価の訓練を受けた検者 III が用意されているとしよう。真の値は a が 85 点、b が 75 点であるとする。これらは検者 III が最も正しく評価できるのだろうが、検者 I も検者 II も“ 常に a は b よりも 10 点低い ”という“ 相対的に平行な関係 ”にあれば信頼性は高いと判定するのが Case3 である。対して、“ 検者間の完全一致 ”を要するのが Case2 である。

1.3.1 Case1

1 人の検者が複数の被検者を評価したときの検者内信頼性 (*Intra-rater reliability*) を知りたいときに用いる式が Case1 (表 1.1 中、一要因の変量モデル) の ICC (1, 1) と ICC (1, k) である。いま、1 人の検者が r 人の被検者に対して k 回繰り返し測定を行ったとすると測定値 x_{ij} は、

$$x_{ij} = \mu + T_i + W_{ij} \quad (3.14)$$

(μ : 期待値, T_i : 被検者 i の効果, W_{ij} : 被検者 i の j 回目の測定誤差, $i = 1, 2, \dots, r; j = 1, 2, \dots, k$)

のような構造モデルで表せる。

それぞれの項の分散値は、

$$\sigma_x^2 = \sigma_T^2 + \sigma_W^2 \quad (3.15)$$

となる。結局求めるものは全体的な分散 $\sigma_T^2 + \sigma_W^2$ に対する真の値 σ_T^2 なので、

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_W^2} \quad (3.16)$$

である。この (3.16) 式は検者内信頼性を求める ICC (1, 1) である。これを求めるために表 1.1 中の BMS, WMS を利用すると、Case1 における BMS, WMS の期待値は、

$$BMS = k\sigma_T^2 + \sigma_W^2 \quad (3.17)$$

$$WMS = \sigma_W^2 \quad (3.18)$$

となっている。(3.17) 式と (3.18) 式から $\sigma_T^2 = (BMS - WMS)/k$ が導かれ、これらを (3.16) 式に代入すると、

$$ICC(1, 1) = \frac{(BMS - WMS)/k}{(BMS - WMS)/k + WMS}$$

ⁱ つまり混合モデルである。

$$\begin{aligned} &= \frac{BMS - WMS}{BMS - WMS + k \cdot WMS} \\ \therefore ICC(1, 1) &= \frac{BMS - WMS}{BMS + (k - 1) WMS} \end{aligned} \quad (3.19)$$

が求められる。

以上では、A という評価法の検者内信頼性 ICC (1, 1) を求めてばらつきの信頼性を推定した。今度は A 評価法を用いて k 人の検者が、被検者 n 人を m 回ずつ測定して、それぞれの平均値（例えば、 $x_{11} = (x_{111} + x_{112} + \dots + x_{11m})/m$ ）をデータとした場合の信頼性を知りたいとする。k 回測定の信頼性を知るためには、(3.16) 式を利用して、

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_W^2/k} \quad (3.20)$$

を求めるとよい。この (3.20) 式と (3.16) 式は右辺分母の第 2 項が σ_W^2/k となっている点で異なる。つまり、 k 回繰り返し測定の平均的な誤差分散を利用するのである。そうしなければ、 k を増やすほど、誤差分散が大きくなる。(3.20) 式は ICC (1, k) になる。

(3.17) と (3.18) 式を (3.20) 式に代入すると、

$$ICC(1, k) = \frac{BMS - WMS}{BMS} \quad (3.21)$$

が求まる。任意の信頼性を得るためには 1.4.4 節を参考にして測定回数を決めるとよい。

次に信頼区間の算出方法 [4] を述べる。ICC (1, 1) の $100(1 - \alpha) \%$ 信頼区間は、以下の式で求める^j。

$$\begin{aligned} F_0 &= BMS/WMS \\ F_U &= F_0 \cdot F_{(1-\alpha/2, n(k-1), n-1)} \\ F_L &= F_0 / F_{(1-\alpha/2, n-1, n(k-1))} \\ \frac{F_L - 1}{F_L + (k - 1)} &< \rho_{ICC(1,1)} < \frac{F_U - 1}{F_U + (k - 1)} \end{aligned} \quad (3.22)$$

例えば $F_{(1-\alpha/2, n(k-1), n-1)}$ とは、 $\alpha = 0.05$ とすると、自由度 $df1 = n(k - 1)$ 、 $df2 = n - 1$ の $1 - 0.05/2 = 0.975$ 点の F 値を代入する。

ICC (1, k) の $100(1 - \alpha) \%$ 信頼区間は、以下の式で求める。

$$1 - \frac{1}{F_L} < \rho_{ICC(1,k)} < 1 - \frac{1}{F_U} \quad (3.23)$$

1.3.2 Case2

複数の検者によって複数の被検者を測定する場合の検者間信頼性 (*Inter-rater reliability*) を求めるのが Case2 と Case3 である。ここでは特に変量モデル [→ 1.2.4 節参照] である Case2 の ICC (2, 1), ICC (2, k) を説明する。

k 人の検者が n 人の被検者を測定した場合を想定する。Case1 と同様に構造モデルを考えると、

$$x_{ij} = \mu + T_i + J_j + I_{ij} + E_{ij} \quad (3.24)$$

^j 信頼区間の上限値は 1 を越えないと考えて、常に上限値 = 1 とする方法もある [13]。

(μ : 期待値, T_i : 被検者 i の効果, J_j : 検者 j の効果, I_{ij} : 検者と被検者の交互作用,
 E_{ij} : 被検者 i の j 回目の測定誤差, $i = 1, 2, \dots, n; j = 1, 2, \dots, k$)

と表せる. これは (3.14) 式に J_j と I_{ij} が加えられた形となっている. このモデルでは検者間信頼性も問うため, J_j が存在するのである. I_{ij} は二元配置分散分析での交互作用に該当し, 検者と被検者の相互関係を考慮している. 例えば, 検者 A による測定値が被検者 $a > b > c$ で, 検者 B の測定値も $a > b > c$ であったとする. 逆に検者 C は被検者 $a < b < c$ となれば, このパターンの異なりは検者 C 固有の影響なのか, 被検者の影響なのか判別できない. このようなときに交互作用として考慮しておく必要がある.

Case1 の時と同様に, 様々なばらつきの中の真の値の分散 σ_T^2 を求めるために,

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2} \quad (3.25)$$

と考える. これは検者間信頼性 (ばらつきの程度) を求める ICC (2, 1) と呼ばれる.

表 1.1 の推定値を用いると,

$$BMS = k\sigma_T^2 + \sigma_I^2 + \sigma_E^2 \quad (3.26)$$

$$JMS = n\sigma_J^2 + \sigma_I^2 + \sigma_E^2 \quad (3.27)$$

$$EMS = \sigma_I^2 + \sigma_E^2 \quad (3.28)$$

となっていることがわかる. これらを変形して,

$$\sigma_T^2 = (BMS - EMS)/k \quad (3.29)$$

$$\sigma_J^2 = (JMS - EMS)/n \quad (3.30)$$

を (3.25) 式に代入すると,

$$\begin{aligned} ICC(2, 1) &= \frac{(BMS - EMS)/k}{(BMS - EMS)/k + (JMS - EMS)/n + EMS} \\ &= \frac{BMS - EMS}{BMS - EMS + k \cdot (JMS - EMS)/n + k \cdot EMS} \\ \therefore ICC(2, 1) &= \frac{BMS - EMS}{BMS + (k - 1)EMS + k \cdot (JMS - EMS)/n} \end{aligned} \quad (3.31)$$

となる.

Case1 と同様に A という評価を, 検者 k 人で被検者 n を m 回繰り返し測定した平均値に対して検者間信頼性を知りたいときには, (3.25) 式分母の誤差を平均して,

$$\begin{aligned} ICC(2, k) &= \frac{\sigma_T^2}{\sigma_T^2 + (\sigma_J^2 + \sigma_I^2 + \sigma_E^2)/k} \\ &= \frac{(BMS - EMS)/k}{(BMS - EMS)/k + ((JMS - EMS)/n + EMS)/k} \\ \therefore ICC(2, k) &= \frac{BMS - EMS}{BMS + (JMS - EMS)/n} \end{aligned} \quad (3.32)$$

を利用するとよい.

ICC (2, 1) の $100(1 - \alpha) \%$ 信頼区間は, 以下の式で求める.

$$\hat{\rho} = ICC(2, 1)$$

$$\begin{aligned}
F_J &= JMS/EMS \\
\nu &= \frac{(k-1)(n-1) \{k\hat{\rho}F_J + n[1 + (k-1)\hat{\rho}] - k\hat{\rho}\}^2}{(n-1)k^2\hat{\rho}^2F_J^2 + \{n[1 + (k-1)\hat{\rho}] - k\hat{\rho}\}^2} \\
F^* &= F_{(1-\alpha/2, (n-1), \nu)} \\
F_* &= F_{(1-\alpha/2, \nu, (n-1))}
\end{aligned}$$

$$\frac{n(BMS - F^*EMS)}{F^*[kJMS + (kn - k - n)EMS] + nBMS} < \rho_{ICC(2,1)} < \frac{n(F_*BMS - EMS)}{kJMS + (kn - k - n)EMS + nF_*BMS} \quad (3.33)$$

ICC (2, k) の $100(1 - \alpha)$ % 信頼区間は、以下の式で求める。

$$\rho_L = \frac{k\rho_L^{**}}{1 + (k-1)\rho_L^{**}} \quad (3.34)$$

$$\rho_U = \frac{k\rho_U^{**}}{1 + (k-1)\rho_U^{**}} \quad (3.35)$$

$$\rho_L < \rho_{ICC(2,k)} < \rho_U \quad (3.36)$$

ここで、(3.34) と (3.35) 式の ρ_L^{**}, ρ_U^{**} はそれぞれ (3.33) 式から求める ICC (2, 1) の信頼区間のそれぞれ下側値と上側値を表す。

1.3.3 Case3

変量要因の検者が変量要因の被検者を測定する場合の検者間信頼性— ICC (2, 1) と ICC (2, k) —を述べたが、ここでは検者が混合モデル [→ 1.2.4 節参照] となる ICC (3, 1) と ICC (3, k) を説明する。

構造モデルは (3.24) 式と同様であるが、検者間の効果を省いた混合モデルとしている点が異なる。このことから、

$$\rho = \frac{\sigma_T^2 - \sigma_I^2 / (k-1)}{\sigma_T^2 + \sigma_I^2 + \sigma_E^2} \quad (3.37)$$

として求める。(3.37) 式の分子で σ_T^2 から $\sigma_I^2 / (k-1)$ を引いている理由は、分母において検者間の効果 (σ_J^2) を除いたと同時に、分子からも検者間に影響する交互作用の平均値を引いておく必要があるからである^k (図 1.3 または下記注参照)。

Case1, Case2 と同じく、表 1.1 の BMS, EMS を変形して、

$$\sigma_T^2 = (BMS - \sigma_E^2) / k \quad (3.38)$$

$$\sigma_I^2 = (k-1)(EMS - \sigma_E^2) / k \quad (3.39)$$

とし、(3.37) 式に代入すると、ICC (3, 1) が求まる。

$$\begin{aligned}
ICC(3, 1) &= \frac{(BMS - \sigma_E^2) / k - (k-1)(EMS - \sigma_E^2) / k \cdot 1 / (k-1)}{(BMS - \sigma_E^2) / k + (k-1)(EMS - \sigma_E^2) / k + \sigma_E^2} \\
&= \frac{BMS - \sigma_E^2 - EMS + \sigma_E^2}{BMS - \sigma_E^2 + (k-1)(EMS - \sigma_E^2) + k \cdot \sigma_E^2} \\
\therefore ICC(3, 1) &= \frac{BMS - EMS}{BMS + (k-1) EMS} \quad (3.40)
\end{aligned}$$

^k 交互作用が存在しないとは限らないためである。もし交互作用が存在しなければ、分子分母の σ_I^2 は 0 となる。

Case1, Case2 と同様に m 回測定した平均値を対象とした信頼性を知りたいときには, 交互作用がないと仮定して ICC (3, k) を求める. (3.37) 式を変形すると,

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2/k} \quad (3.41)$$

となる. 実際には表 1.1 最終列の期待値から交互作用 $\sigma_I^2/(k-1)$ を引くと,

$$BMS = k\sigma_T^2 + \sigma_E^2 \quad (3.42)$$

$$EMS = \sigma_E^2 \quad (3.43)$$

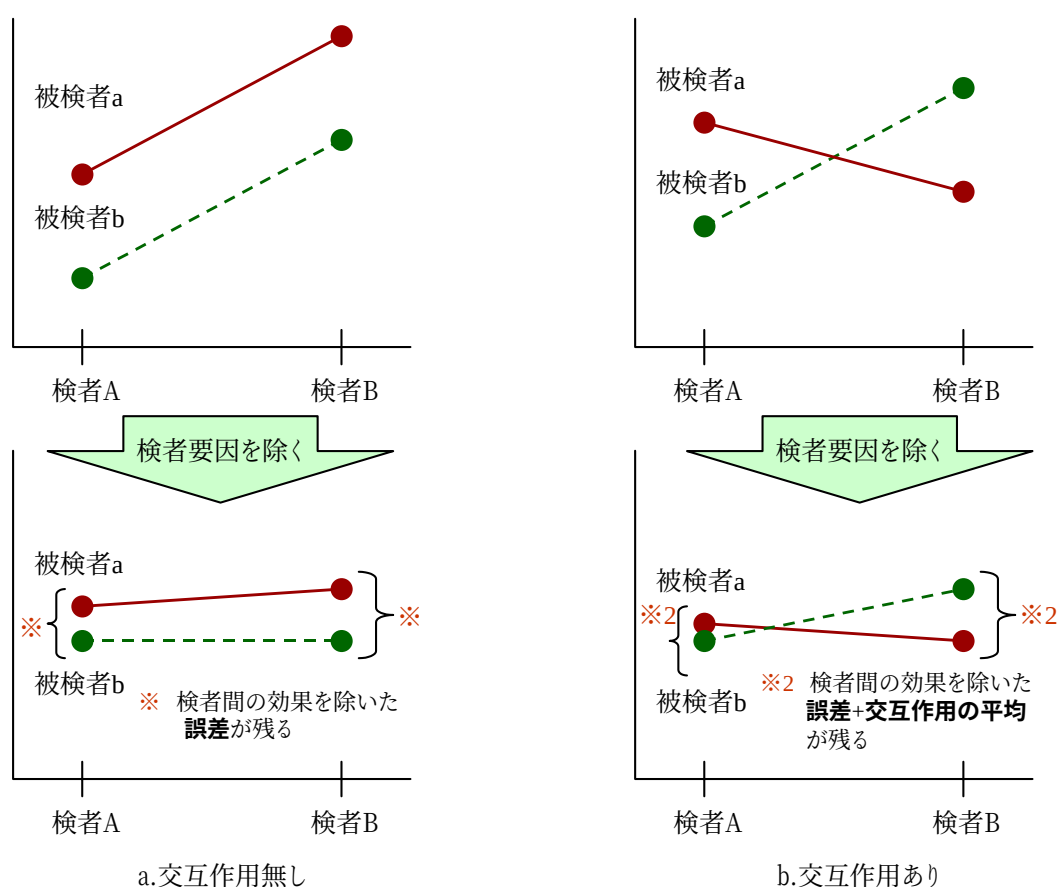


図 1.3 検者要因を引いた後の交互作用の影響

となるから, これらを (3.41) 式に代入して,

$$ICC(3, k) = \frac{BMS - EMS}{BMS} \quad (3.44)$$

が求める式である.

信頼区間は, ICC (3, 1) の場合,

$$F_0 = BMS/EMS \quad (3.45)$$

$$F_L = F_0 / F_{(1-\alpha/2, (n-1), (n-1)(k-1))} \quad (3.46)$$

$$F_U = F_0 \cdot F_{(1-\alpha/2, (n-1)(k-1), (n-1))} \quad (3.47)$$

として、以下で求まる．

$$\frac{F_L - 1}{F_L + (k - 1)} < \rho_{ICC(3,1)} < \frac{F_U - 1}{F_U + (k - 1)} \quad (3.48)$$

ICC (3, k) の信頼区間は、以下のようになる．

$$1 - \frac{1}{F_L} < \rho_{ICC(3,k)} < 1 - \frac{1}{F_U} \quad (3.49)$$

※注－交互作用の考え方－

k 人の同一被検者群を対象として n 人の検者によって測られたデータは、互いに従属となる．従って分散値の性質 [→他章を参照] から、交互作用 I_{ij} の分散の和は、

$$V\left(\sum_{j=1}^k I_{ij}\right) = k \cdot V(I_{ij}) + k(k-1) \text{COV}(I_{ij}, I_{i^0j}) \quad (3.50)$$

と表せる．固定モデルでは、検者間のばらつきは存在せず、かつ検者間の分散は等しいとなるため、(3.50) 式 = 0 となり、かつ $V(I_{ij}) = V(I_{i^0j})$ である ($i \neq i'$)．従って、 $\text{COV}(I_{ij}, I_{i^0j}) = V(I_{ij})$ であり、これらから $\text{COV}(I_{ij}, I_{i^0j}) = -\sigma_I^2/(k-1)$ が導かれる．これは交互作用に影響する検者間の共分散であるから、(3.37) 式では分子からその分、引いている．

交互作用項の処理については上記とは別の考え方もある．

Bartko [10] では (3.37) 式を、

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_I^2 + \sigma_E^2} \quad (3.51)$$

と定義している．また、表 1.1 の EMS の期待値を $EMS = \sigma_I^2 + \sigma_E^2$ と記載している．

この EMS と BMS を (3.51) 式に代入して、

$$ICC(3, 1) = \frac{BMS - EMS + \sigma_I^2}{BMS + (k-1)EMS + \sigma_I^2} \quad (3.52)$$

が得られる．まず、交互作用項が存在するかないか不明な状態と仮定すると、

$$ICC(3, 1) \cong \frac{BMS - EMS}{BMS + (k-1)EMS} \quad (3.53)$$

となり、これを下限値とする ((3.40) 式と同じになる)．

次に交互作用は必ず存在すると仮定すると、 $EMS \leq \sigma_I^2$ であるから、最大値は $EMS = \sigma_I^2$ となり、

$$ICC(3, 1) = \frac{BMS}{BMS + kEMS} \quad (3.54)$$

が求まる．つまり、ICC (3, 1) 式は (3.53) 式 $\leq$ ICC (3, 1) $\leq$ (3.54) 式の範囲に存在するとして表す考え方である．

1.4 ICC 適用の留意点

1.4.1 ICC (1, 1) , ICC (2, 1) , ICC (3, 1) の大きさの関係

ある特定のデータを対象に ICC (1, 1) , ICC (2, 1) , ICC (3, 1) , それぞれの値を求めたとき, ほとんどは $ICC(1, 1) \leq ICC(2, 1) \leq ICC(3, 1)$ の関係にあるといわれる [8]. しかし, ICC 値が 0 に近似するときや被検者のばらつきが大きいときなどは $ICC(1, 1) > ICC(2, 1)$ の場合もある. 同一のデータであっても ICC の Case によって値は大きく異なることがある. Case 適用に誤りがないように注意を払う必要がある.

1.4.2 Case1 と Case2 の違い

前述した Case1 の式 [→ 1.3.1 節] は, あまり使用しない方がよいといわれる [4] [8] が, その根拠は示されていない. また, 信頼性の考え方が正しく理解されているならば, このようなことはいえない.

ICC(1, 1) と ICC (2, 1) を比較してみよう. まず, ICC (1, 1) の BMS は,

$$BMS = k\sigma_T^2 + \sigma_W^2 \quad (3.17) \text{ 式再掲}$$

であり, ICC (2, 1) の BMS は,

$$BMS = k\sigma_T^2 + \sigma_I^2 + \sigma_E^2 \quad (3.26) \text{ 式再掲}$$

となっており, それぞれ異なっていることがわかる. σ_W^2 は表 1.1 より,

$$\sigma_W^2 = \sigma_J^2 + \sigma_I^2 + \sigma_E^2 \quad (4.55)$$

の様に表せる. これと (3.30) 式, (3.28) 式, ICC (2, 1) の BMS を利用して ICC (1, 1) の σ_T^2 を表すと,

$$\begin{aligned} \sigma_T^2 &= BMS - ([JMS - EMS]/n + EMS) \\ &= BMS - EMS - (JMS - EMS)/n \end{aligned} \quad (4.56)$$

である. (4.55) 式と (4.56) 式を ICC (1, 1) に代入して変形すると,

$$ICC(1, 1) = \frac{\sigma_T^2 - \sigma_J^2}{\sigma_T^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2 - \sigma_J^2} \quad (4.57)$$

になり, ICC (2, 1) 式の分子と分母から, $\sigma_J^2 (= (JMS - EMS)/n)$ の分がそれぞれ引かれていることになる.

繰り返し述べるが, ICC (1, 1) では検者間の分散を真の値から引いてしまっている¹. 従って, 検者要因を考慮しない被検者のばらつきに対して, 誤差のばらつきがどれくらいかを問題にしたいときは ICC (1, 1) を適用する. これはモデルの考え方によるものである.

1.4.3 ICC の判定基準

求めた ICC 値は, 例えば以下の表 1.2 のような基準をもとにして判定する. 但し, この表は Landis ら [12] による Kappa 係数の指標を ICC の判定に応用したものであり, 理論的根拠は全くない.

¹ 分散分析の平方和の構成を参照.

これ以外の判定基準 [6] [9] も存在するが、全体をまとめると ICC が 0.7 以上であれば信頼性は良好であると考えてよいようである。ここで注意すべきは、求める ICC はあくまで点推定値であることである。ICC の点推定値以外に区間推定値、すなわち信頼区間を提示することも適切である。

表 1.2 判定基準 [12]

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

1.4.4 測定回数・標本の大きさの決定

D 研究として ICC(n, k) ($n = 1, 2, 3$) を利用する手順を述べよう。

例えば、ある測定を行い、ICC (1, 1) = 0.7 という結果を得たとする。研究で要求される値は 0.9 であるとして、この値を得るためには何回測定の平均値を用いればよいか。そのとき、以下の式^m を利用する。

$$k = \frac{\rho_1(1 - \rho_2)}{\rho_2(1 - \rho_1)} \quad (4.58)$$

なお、この式の ρ_1 は目標とする信頼性係数値、 ρ_2 は得られた信頼性係数値 である。この例では、

$$k = \frac{0.9 \times (1 - 0.7)}{0.7 \times (1 - 0.9)} = 3.8571 \quad (4.59)$$

であり、少数桁を四捨五入すると“ 4 回 ”となるから一人あたり 4 回繰り返し測定した平均値をデータとして用いればよいことになる。このようにして信頼性の向上を計画することができる。

標本の大きさの決定については確定したものはないが、Eliasziv らの報告 [13] に例が挙げられている。これは一元配置モデルによる信頼性係数で検討しているので、一般化するには理論的保証がない。

1.4.5 統計的検定の方法

ICC の検定の方法については Eliasziv ら [13] に記載されているが、繰り返し測定の二元配置モデルのデザインⁿ であり、さらに理論的に正しい方法か否かについてここでは断言できない。従って、信頼区間を利用した仮説検定が無難であろう。

なお、 $H_0: \rho = 0$ の帰無仮説に対する検定は単に母信頼性係数が“ 0 か否か ”を検定しているに過ぎないため、例えば表 1.2 の様な基準値を基に $H_0: \rho \leq 0.6$ とし、*substantial* 以上になるかといった実質的に意味のある検定を推奨する。

^m 平行テスト法で用いられる Spearman-Brown の公式を利用している [→ 1.2.2 節参照]。

ⁿ 上述してきた例では一元配置モデル、または繰り返しのない二元配置モデルである。

1.4.6 標準誤差の算出

測定の標準誤差（以下、SEM）の算出は、ICC 各公式の分母から分子の値を引いて平方を取るとよい。つまり、各 Case の構造モデルで真の値の項以外の項を平方したものが SEM となる。SEM は、

$$SEM = \sqrt{\frac{\sigma_{total}^2(1 - \rho)}{n}} \quad (4.60)$$

である。 σ_{total}^2 は全変動のことである。

ICC (1, 1) の標準誤差 SEM_1 は、

$$SEM_1^2 = \sigma_W^2 = WMS \quad (4.61)$$

で求められる。ICC (1, 2) と ICC (1, 3) の標準誤差（それぞれ SEM_2 , SEM_3 ）は、

$$SEM_2^2 = \sigma_J^2 + \sigma_I^2 + \sigma_E^2 = (JMS - EMS)/n + EMS \quad (4.62)$$

$$SEM_3^2 = \sigma_I^2 + \sigma_e^2 = EMS \quad (4.63)$$

で求めることができる。SEM の簡単な例題については、Stratford ら [14] で取り上げられている。

SEM の信頼区間は分散の性質を利用して以下の式で求めることができる。

$$\frac{WMS}{\chi_{(\alpha/2, df)}^2}, \frac{WMS}{\chi_{(1-\alpha/2, df)}^2} \quad (4.64)$$

なお、上式中 $\chi_{(\alpha, df)}^2$ は自由度 df における $\alpha/2$ %点の χ^2 値である。これを利用すると検定が可能となる。

2つの信頼性研究から得られた SEM が有意に異なるかどうかの検定を行うことも可能である。 SEM_A と SEM_B がわかっているとして、これらの自乗比が“ 帰無仮説 H_0 : 2つの分散の比は等しい ”の下で自由度 (df_A , df_B) の F 分布に従うことを利用する。つまり $F_{(df_A, df_B)} = SEM_A^2/SEM_B^2$ を求めるとよい。

1.5 シミュレーションによる特性の検討

ICC は理論的に確立した手法とは言い難いため、上述してきた様々な計算式が妥当であるかどうかは不明である。また、データの特性によっては単に ICC どうしを比較できない可能性もある。そこで、この節では様々な状況を想定してシミュレーションを行い、ICC の特性を検討する。

1.5.1 推定値の性質

例えば、Case2 における線形モデルの各項は $T_i \sim N(0, \sigma_T^2)$, $J_j \sim N(0, \sigma_J^2)$, $I_{ij} \sim N(0, \sigma_I^2)$, $E_{ij} \sim N(0, \sigma_E^2)$ であり、互いに独立である。シミュレーションデータは構造モデル (3.14) 式または (3.24) 式に基づくデータとしての統計モデルを考えて構築するのが正当である。ここでは実際に扱うデータの形式として乱数を発生させ、シミュレーションを行う。

いま、A という評価基準に基づいて k 人の検者が n 人の被検者を評価したとする。興味があるのは k 人の検者間信頼性である。求めたい ICC は ICC (2, 1) とする。検者要因を J_j 、被検者要因を T_i とすると、測定したデータ行列 $X = (x_{11}, x_{12}, \dots, x_{ij})(i = 1, 2, \dots, n; j = 1, 2, \dots, k)$ は、

$$x_{ij} = \mu + T_i + J_j + (TJ)_{ij} + e_{ij} \quad (5.65)$$

と表せる。 $(TJ)_{ij}$ は検者と被検者の交互作用である。ここで、 $\mu = 0$, $X \sim N(\mu_X, \sigma_X^2)$ と仮定しよう。

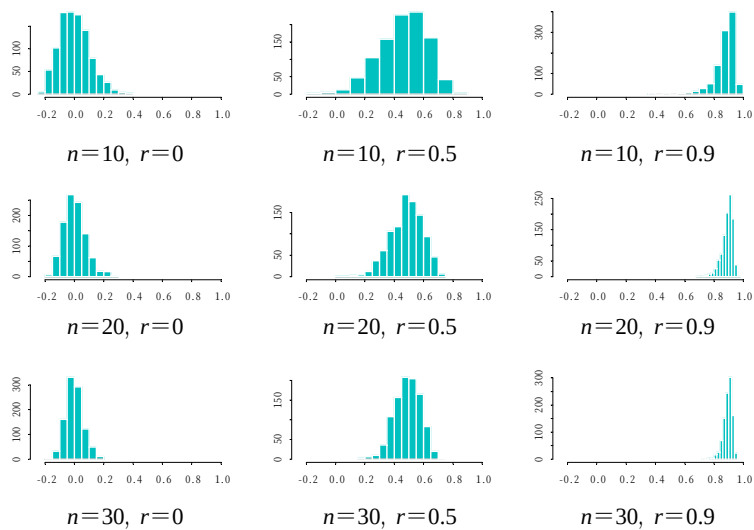
このデータは反復測定モデルであるから、 X の列ベクトル $x_{n1}, x_{n2}, \dots, x_{nk}$ 間は互いに従属な関係がある。従って、これらの条件を満たした乱数データを発生させる必要がある。また、分散分析は検者要因を変量効果、被検者要因をブロック要因とする乱塊法 (反復測定による分散分析または二元配置分散分析) に基づいて計算を行う。

検者 k 人 ($k = 5, 10, 20$) によって、被検者 i 人 ($i = 10, 20, 30$) を評価した状況を想定する。このとき検者間の相関性、つまり一致性を $r = 0.0, 0.5, 0.9$ と変化させる。これは同時に被検者間の従属性を持たせることになる。要素 r の $k \times k$ 母相関行列 R に対して固有値問題を解き、 $R = AA^t$ なる A 行列を求める。正規乱数の $n \times k$ 行列 $D \sim N(0, 1)$ を作成し $A^t D$ の行列積を求め、これをデータ行列 X としてそれぞれの条件で ICC (2, 1) を求めた。この作業を 1,000 回繰り返し、ヒストグラムにより性質を検討した。なお、計算の詳細は省略する。

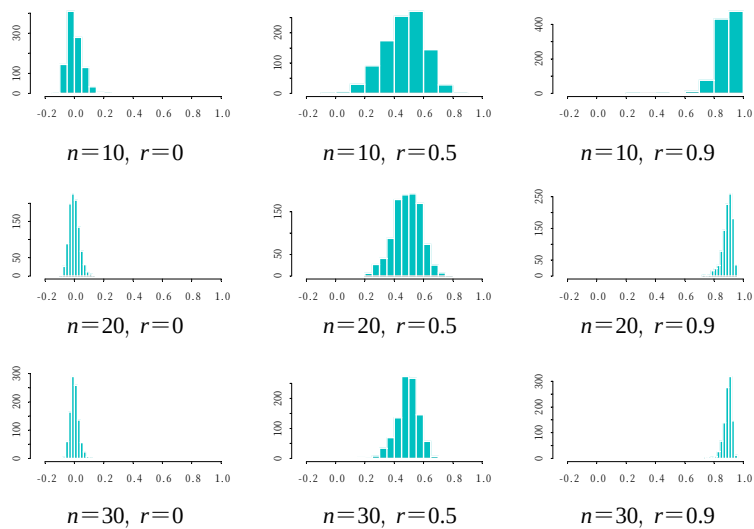
結果は、図 1.4 の通りであった。何れの条件においても、被検者数 n が増加するとばらつきは小さくなった。検者数 k の増加によっては、ややばらつきが小さくなった。また、母相関係数 r (ICC) = 0 のときは半数近くが負の値をとった。分布は、正規分布に従っているとは言い切れず、不偏性が保たれていない。この点については、Fisher [3], Shrout ら [4] も指摘しており^o、推定値の性質のうち一致性のみは保たれると述べている。実際に各ヒストグラムの平均値を求めると期待値よりも常に小さい値をとっていた。ICC は推定値として、あまりよい性質とはいえない。また、ICC の期待値が 0 に近いときには、算出される ICC が負の値を取ることが想定されるが、そのときには信頼区間も述べておくべきである。なお、今回のシミュレーションでは値が 1 を越えたケースは存在しなかった。

^o Fisher によると標本からの級内相関係数は常に負の方向に偏り、修正が必要であると述べている。その値は $+\frac{1}{2} \log \frac{n^0}{n^0-1}$ である。これは近似的に $+\frac{1}{2n^0-1}$ で修正可能である。ここで $n' = n + 2$ 。

検者数
 $k=5$ のとき



検者数
 $k=10$ のとき



検者数
 $k=20$ のとき

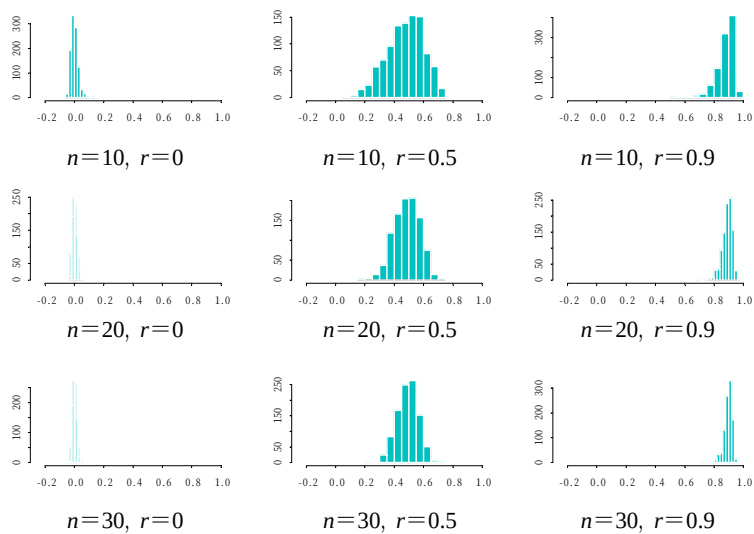


図 1.4 検者数, 被検者数, 相関係数を変化させたときの ICC (2, 1) のヒストグラム

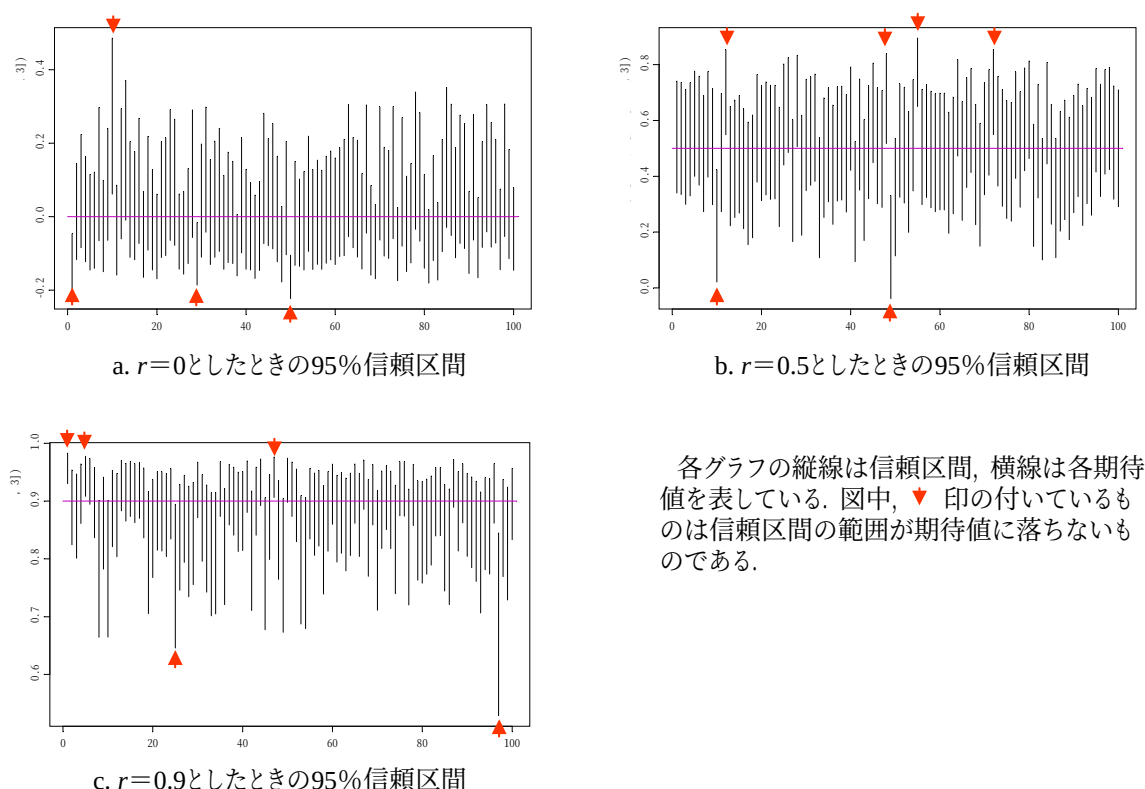


図 1.5 ICC (2, 1) の 95 %信頼区間

1.5.2 信頼区間の性質

信頼区間の求め方については1.3節で述べた。1.5.1節で推定値の性質を検討した結果、よい推定値とはいえなかったので、信頼区間の性質も知りたい。

シミュレーションは1.5.1節と同様の手順でデータ行列 $X = A^t X$ を作成し、ICC (2, 1) の 95 % 信頼区間を求めた。ただし、 $k = 5$ 、 $n = 20$ に固定し、 r のみ 0, 0.5, 0.9 と変化させた。各々 r の条件で信頼区間を 100 回ずつ求め、その結果を図 1.5 に示した。

信頼区間はほぼ理論に従っているようである。Eliaszew [13] らは信頼区間の上限を“1”としている。

1.5.3 データの特性によってどのような値をとるか

ICC を利用する者は最低限データの特性によって ICC がどのような値をとるか、その特徴を知っておく必要がある。図 1.6 では様々な架空のデータを作成し、ICC (1, 1)、ICC (2, 1)、ICC (3, 1) がどのように変化するか検討したものである。

図 1.6-a. を基本値とする。A～D の検者は a～d の被検者に対して全く一致した判定をしているからすべての ICC は“1”を示す。b. のように A の検者が d の被検者を+2 だけ多く見積もったとき ICC 値は低くなる。仮にデータ全体の値が大きくなって、A の検者が d の被検者を+2 だけ多く見積もっても（図 1.6-c.）,

$k=4, n=4$ として以下のようなデータが得られたとする。

	検者				平均		
	A	B	C	D			
a	1	1	1	1	1	ICC(1,1)=	1.0000
b	2	2	2	2	2	ICC(2,1)=	1.0000
c	3	3	3	3	3	ICC(3,1)=	1.0000
d	4	4	4	4	4	SEM1=	-
平均	2.5	2.5	2.5	2.5	2.5	SEM2=	-
						SEM3=	-

a. 基本値

	検者				平均		
	A	B	C	D			
a	11	11	11	11	11	ICC(1,1)=	0.9684
b	12	12	12	12	12	ICC(2,1)=	0.9684
c	13	13	13	13	13	ICC(3,1)=	0.9684
d	15	14	14	14	14.25	SEM1=	0.2500
平均	12.75	12.5	12.5	12.5	12.563	SEM2=	0.2500
						SEM3=	0.2500

c. 図a+10の後, A-dセルの値に“2”加える

	検者				平均		
	A	B	C	D			
a	11	10	9	8	9.5	ICC(1,1)=	0.4286
b	12	11	10	9	10.5	ICC(2,1)=	0.5000
c	13	12	11	10	11.5	ICC(3,1)=	1.0000
d	14	13	12	11	12.5	SEM1=	1.2910
平均	12.5	11.5	10.5	9.5	11	SEM2=	0.6455
						SEM3=	-

e. 検者 (列) 要因に定数差分の順序的な偏りを持たせる ($A > B > C > D$)

	検者				平均		
	A	B	C	D			
a	110	90	70	50	80	ICC(1,1)=	-
b	120	100	80	60	90	ICC(2,1)=	0.2000
c	130	110	90	70	100	ICC(3,1)=	1.0000
d	140	120	100	80	110	SEM1=	25.8199
平均	125	105	85	65	95	SEM2=	12.9099
						SEM3=	-

g. 検者 (列) 要因に定数差分の順序的な偏りを持たせる (図fよりも列差は大きい)

	検者				平均		
	A	B	C	D			
a	110	90	70	50	80	ICC(1,1)=	-0.3169
b	110	90	70	50	80	ICC(2,1)=	0.0093
c	110	90	70	50	80	ICC(3,1)=	1.0000
d	115	95	75	55	85	SEM1=	25.8199
平均	111.25	91.25	71.25	51.25	81.25	SEM2=	12.9099
						SEM3=	-

i. hと同一のデータで被検者d (d行) が定数値 (+5) 分大きい値を取る

	検者				平均		
	A	B	C	D			
a	1	1	1	1	1	ICC(1,1)=	0.9684
b	2	2	2	2	2	ICC(2,1)=	0.9684
c	3	3	3	3	3	ICC(3,1)=	0.9684
d	5	4	4	4	4.25	SEM1=	0.2500
平均	2.75	2.5	2.5	2.5	2.5625	SEM2=	0.2500
						SEM3=	0.2500

b. A-dセルの値に“2”加える

	検者				平均		
	A	B	C	D			
a	110	110	110	110	110	ICC(1,1)=	0.9684
b	120	120	120	120	120	ICC(2,1)=	0.9684
c	130	130	130	130	130	ICC(3,1)=	0.9684
d	150	140	140	140	142.5	SEM1=	2.5000
平均	127.5	125	125	125	125.63	SEM2=	2.5000
						SEM3=	2.5000

d. 図c×10の後, A-dセルの値に“20”加える

	検者				平均		
	A	B	C	D			
a	110	100	90	80	95	ICC(1,1)=	0.4286
b	120	110	100	90	105	ICC(2,1)=	0.5000
c	130	120	110	100	115	ICC(3,1)=	1.0000
d	140	130	120	110	125	SEM1=	12.9099
平均	125	115	105	95	110	SEM2=	6.4550
						SEM3=	-

f. 図e×10とする ($A > B > C > D$)

	検者				平均		
	A	B	C	D			
a	110	90	70	50	80	ICC(1,1)=	-0.2698
b	110	90	70	50	80	ICC(2,1)=	0.0361
c	110	90	70	50	80	ICC(3,1)=	1.0000
d	120	100	80	60	90	SEM1=	25.8199
平均	112.5	92.5	72.5	52.5	82.5	SEM2=	12.9099
						SEM3=	-

h. 被検者d (d行) が他の行よりも, 定数値 (+10) 分大きい値を取る

	検者				平均		
	A	B	C	D			
a	110	111	112	113	111.5	ICC(1,1)=	-0.1111
b	110	111	112	113	111.5	ICC(2,1)=	0.1304
c	110	111	112	113	111.5	ICC(3,1)=	1.0000
d	111	112	113	114	112.5	SEM1=	1.2910
平均	110.25	111.25	112.25	113.25	111.75	SEM2=	0.6455
						SEM3=	-

j. 検者 (列) 要因に順序性を持たせ, かつ被検者dが定数値 (+1) 分大きい値を取る

図 1.6 さまざまなデータに対する ICC の変化

値は b. と変わらない. このことから, ICC は平均値の大きさに対する相対的な誤差の大きさではなく, データのばらつき分に対する誤差の大きさに影響を受けることがわかる. ところで c. の値を 10 倍して図 1.6-d. のようにしても ICC 値は変わらない. しかし, a. や b. のように“ A の検者が d の被検者を +2 だけ多く ”どころか, “ A の検者が d の被検者を +20 も多く ”見積もっているにも関わらず ICC 値は同じというのは不合理である. つまり, ICC は被検者間の値のばらつきが大きければ検者の誤差の大きさは見逃されやすい. 従って, 事前の分散分析で被検者要因が有意となるような場合は, その解釈に注意が必要である. なお, SEM を見ると図 1.6 中 b. と c. に比較して d. の値は大きくなっているので, SEM は判断の一助となる.

図 1.6-e. は検者に一定した順序性 (偏り) を持たせた場合である. 被検者間のばらつきが生じているばかりでなく, 検者間の偏りも生じている. もちろん, ICC 値は小さくなるが ICC (3, 1) は検者間の定数的な偏りは考慮しないために“ 1 ”を示したままである. 従って e. の値を 10 倍した f. でも, これら ICC の

値は変わらない。検者間の偏りを更に大きくとった g. では ICC (1, 1) と ICC (2, 1) の値は小さくなるが、ICC (3, 1) の値は不変である。結局、被検者間の値と比較して検者間の定数的な偏りが大きくなれば ICC (1, 1) ICC (2, 1) 値は小さくなるが、ICC (3, 1) は不変の性質がある。

図 1.6-h. では被検者 d の値を全体的に“ +10 ”過大に評価した場合を想定している。このときには ICC (1, 1) と ICC (2, 1) はかなり小さくなるが、検者 A~D の相対的な差は一定しているので、ICC (3, 1) は不変を保っている。図 1.6-i. では被検者 d の値を他の被検者よりも“ +5 ”過大に評価した場合である。ここでは、被検者 d の値が図 1.6-h. の被検者 d よりも小さいにも関わらず、ICC (1, 1), ICC (2, 1) は更に小さくなっている。この理由は、被検者間のばらつきが $h. > i.$ となっても検者間のばらつきは $h. = i.$ を保っており、相対的に大きいからである。j. では被検者間、検者間ともにばらつきを小さくしているから、当然 ICC (1, 1) と ICC (2, 1) 値は大きくなる。やはりここでも被検者間のばらつきに注意する必要がある。

以上のような傾向は既に確認されているが、基本的に押さえておきたい事項である。ICC の提示のみでは不十分となり、SEM やローデータの提示も必要である。分散分析の結果を利用して ICC を求める場合は、効果の大きさの観察も欠かせない。

まとめ

- ICC は検者数よりも被検者数の増加で推定制度が向上する。
- 推定量の性質のうち、不偏性については保証されない。
- ICC の母係数が 0 に近いとき、推定値（データから求める係数）は負の値を示すことがある。
- ICC は被検者間の値のばらつきが大きければ検者の誤差の大きさを見逃す。

1.6 データ解析の実際

下表のデータ例は、患者（被検者 1～10；計 10 名）を対象として理学療法士（検者 A～D；計 4 名）により、膝関節屈曲可動域と足関節背屈可動域のそれぞれを計測した結果である。

膝屈曲	A	B	C	D	足背屈	A	B	C	D
被検者 1	126	122	131	125	被検者 1	6	5	4	7
2	137	143	141	141	2	6	8	6	8
3	113	119	115	105	3	15	14	12	15
4	153	143	135	144	4	4	4	1	0
5	146	157	150	149	5	11	10	11	11
6	161	157	160	160	6	15	14	15	18
7	110	109	105	113	7	9	12	9	12
8	145	151	152	156	8	5	2	4	5
9	126	141	132	122	9	14	12	14	16
10	114	126	130	125	10	9	8	7	8
平均値	133.1	136.8	135.1	134.0	平均値	9.4	8.9	8.3	10.0
標準偏差	17.95	16.79	16.75	18.63	標準偏差	4.20	4.23	4.67	5.50

測定の順序は無作為とし、検者間の臨床経験年数には差がないとする。ここでは膝関節屈曲と足関節背屈のそれぞれで検者間信頼性を知ることが目的である。

1.6.1 解析手順－検者間信頼性を求める－

検者間の信頼性係数を求めるためには、二元配置分散分析を用いて以下のような表を作成しておくとう便利である。

表 1.3 二元配置分散分析の結果

変動要因	偏差平方和	df	MS		分散比	有意性
膝屈曲						
被検者要因	10319.5	9	1146.6	← BMS	40.4	**
検者要因	76.1	3	25.4	← JMS	0.9	
総変動	765.9	27	28.4	← EMS		
足背屈						
被検者要因	740.6	9	82.3	← BMS	47.5	**
検者要因	15.7	3	5.2	← JMS	3.02	*
総変動	46.8	27	1.7	← EMS		

** $p \leq 0.01$, * $p \leq 0.05$

この表の BMS, JMS, EMS を利用して信頼性係数を求める。あらゆる理学療法士を母集団と仮定して無作為抽出された検者と想定するから、係数は ICC (2, 1) が適当である。計算手順を以下に示す。

膝関節屈曲の場合

$$\begin{aligned}
 ICC(2, 1) &= \frac{BMS - EMS}{BMS + (k-1)EMS + k \cdot (JMS - EMS)/n} \\
 &= \frac{1146.6 - 28.4}{1146.6 + (4-1) \times 28.4 + 4 \times (25.4 - 28.4)/10} = \frac{1118.2}{1230.6} \\
 ICC(2, 1) &= 0.909
 \end{aligned}$$

比較的高い値が得られた。さて、この95%信頼区間を求めよう[→1.3.2節参照]。

$$\begin{aligned}
 \hat{\rho} &= ICC(2, 1) = 0.909 \\
 F_J &= JMS/EMS = 0.89 \\
 \nu &= \frac{(k-1)(n-1) \{k\hat{\rho}F_J + n[1 + (k-1)\hat{\rho}] - k\hat{\rho}\}^2}{(n-1)k^2\hat{\rho}^2F_J^2 + \{n[1 + (k-1)\hat{\rho}] - k\hat{\rho}\}^2} \\
 &= \frac{(4-1) \times (10-1) \times 4 \times 0.909 \times 0.89 + 10 \times [1 + (4-1) \times 0.909] - 4 \times 0.909^2}{(10-1) \times 4^2 \times 0.909^2 \times 0.89^2 + 10 \times [1 + (4-1) \times 0.909] - 4 \times 0.909^2} \\
 &= 29.95 \\
 F^* &= F_{(1-\alpha/2, (n-1), \nu)} = F_{(0.975, 9, 29.95)} = 3.561 \\
 F_* &= F_{(1-\alpha/2, \nu, (n-1))} = F_{(0.975, 29.95, 9)} = 2.576
 \end{aligned}$$

$$\begin{aligned}
 &\frac{n(BMS - F^*EMS)}{F^*[kJMS + (kn - k - n)EMS] + nBMS} < \rho_{ICC(2,1)} < \frac{n(F_*BMS - EMS)}{kJMS + (kn - k - n)EMS + nF_*BMS} \\
 &\frac{10 \times (1146.6 - 3.561 \times 28.4)}{3.561 \times [4 \times 25.4 + (4 \times 10 - 4 - 10) \times 28.4] + 10 \times 1146.6} < \rho_{ICC(2,1)} \\
 &< \frac{10 \times (2.576 \times 1146.6 - 28.4)}{4 \times 25.4 + (4 \times 10 - 4 - 10) \times 28.4 + 10 \times 2.576 \times 1146.6} \\
 &0.7232 < \rho_{ICC(2,1)} < 0.963 \\
 SEM &= \frac{(JMS - EMS)/n + EMS}{\frac{1}{q} \frac{(JMS - EMS)/n + EMS}{(25.4 - 28.4)/10 + 28.4}} = \frac{5.30}{\frac{1}{q} \frac{(JMS - EMS)/n + EMS}{(25.4 - 28.4)/10 + 28.4}}
 \end{aligned}$$

以上から、95%信頼区間の下限値は0.7232であるため、 $\rho = 0$ はもちろん、 $\rho \leq 0.7$ の帰無仮説も5%水準で棄却され、信頼性の高いことがわかる。

足関節背屈の場合

足関節背屈も同様に計算すると、それぞれの結果は、

$$\begin{aligned}
 ICC(2, 1) &= 0.906 \\
 0.776 &< \rho_{ICC(2,1)} < 0.973 \\
 SEM &= 1.43
 \end{aligned}$$

となる（計算過程の有効桁数は異なる）。これも膝関節屈曲と同様に帰無仮説 $H_0: \rho \leq 0.7$ は5%水準で棄却される。各推定値の値のみをみると両者とも同程度の値であるから、“膝関節屈曲の可動域測定と足関節背屈の可動域測定は同程度に検者間信頼性が高い”と結論づけてしまう。しかし、SEMは膝関節屈曲の方が大きい。

1.6.2 データの観察

データを観察すると上記の結論を補助する知見を得ることができる。まず、各検者の平均値と標準偏差値を比べ、次に分散分析の結果を観察してみる。

前述したデータの表から平均値、標準偏差値を観察すると、膝関節屈曲の平均値は大きく、足関節背屈の平均値は小さい。同時に、標準偏差値も足関節背屈の値の方が小さい。平均値の差の原因は対象としている関節が異なるわけだから、当然の結果である。標準偏差が異なる原因も関節の違いのためであろう。被検者の中に、関節疾患を有していた者または既往症は存在しないかを確認しておくべきである。また、被検者自体の値の再現性も問題となってくる。

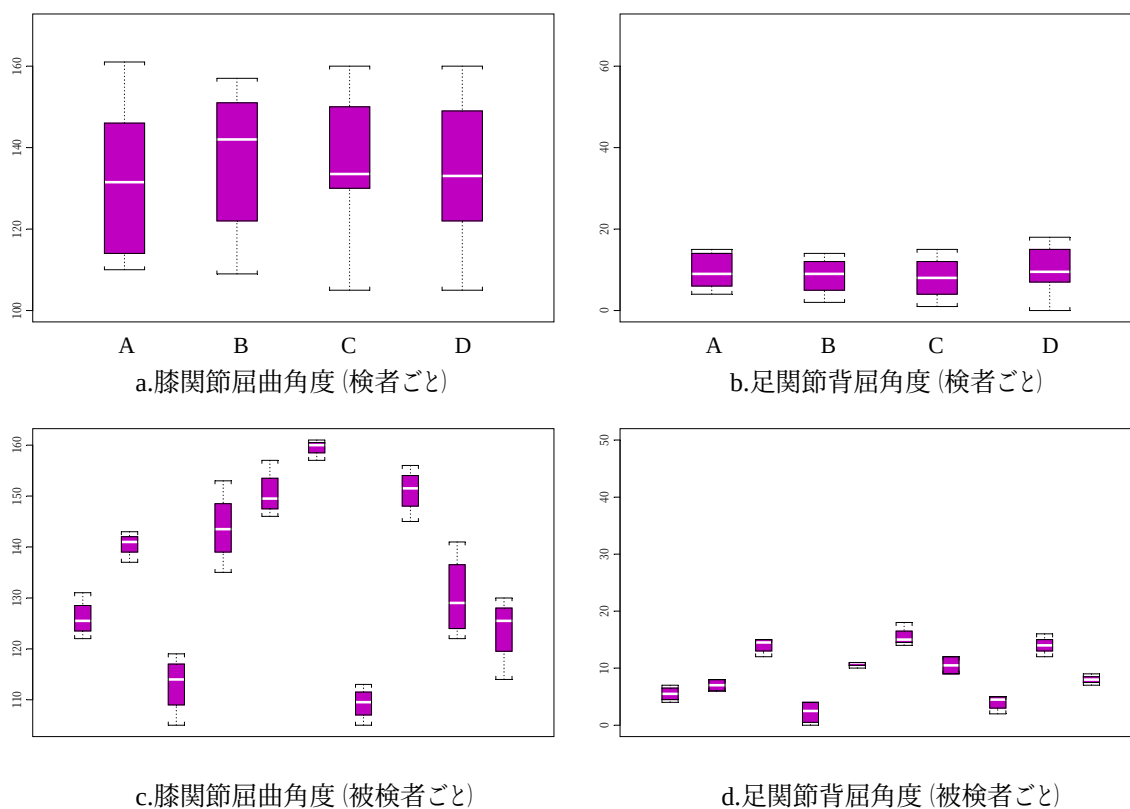
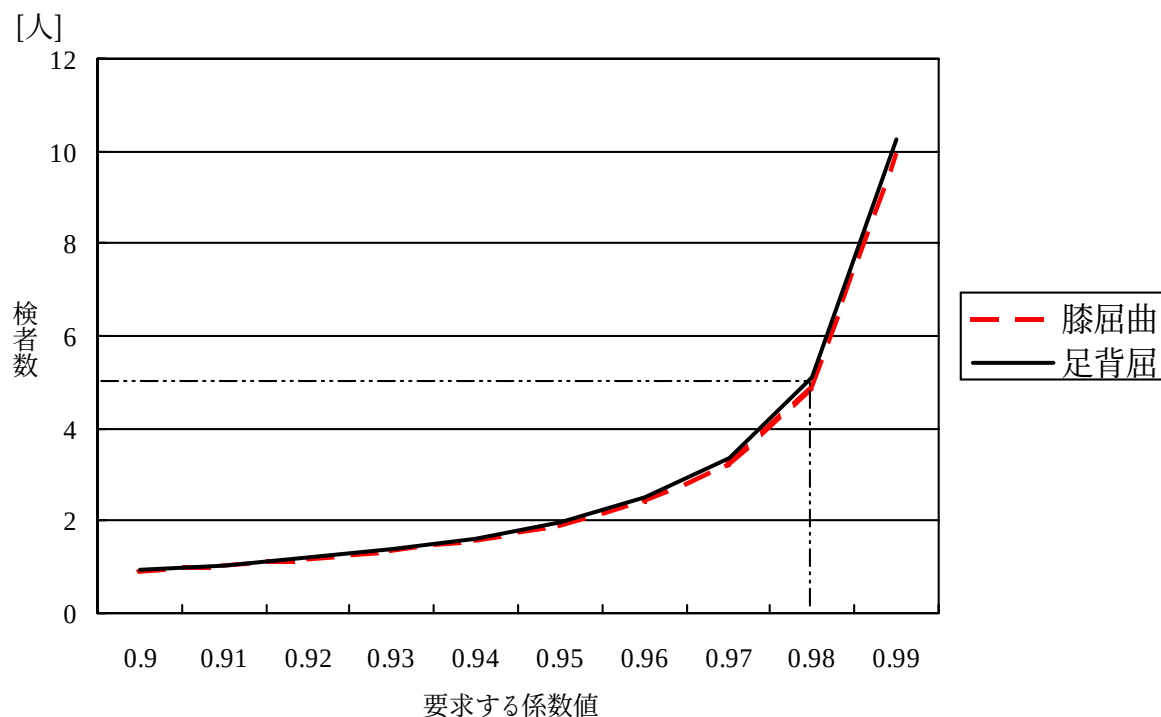


図 1.7 各データの箱ひげ図

図 1.7 は、2 種類のデータを検者ごと、被検者ごとに box plot したものである。比較のために a. と b., c. と d., それぞれのスケールを同じにし、縦軸最大値または最小値は同じにしていない。膝関節屈曲角度の検者のばらつき、または平均値の変動は大きい (a. と b. の比較)。また、被検者のばらつきも膝関節屈曲角度の方が大きい。これに対して SEM は有効な情報を与えてくれる。膝関節屈曲における被験者間のばらつきは、足関節背屈のそれよりも大きくなって当然と考えるか、絶対的な値は同一でなければならないとするかによって結果の見方も変化するはずである。とにかく SEM の提示、できるならばグラフ表示は必要であろう。

1.6.3 測定回数の決定

D 研究として、測定回数を決定する。膝関節屈曲の ICC (2, 1) = 0.909, 足関節背屈の ICC (2, 1) = 0.906 であった。これから得たい係数値を 0.90~0.99 まで 0.01 刻みに変化させ, (1.4.4) 節の式に代入して求めていくと, 図 1.8 のようなグラフを得ることができる。



※ $\rho \geq 0.98$ の係数値を得たいならば, 検者数5人の平均値を用いる必要があるということがわかる。

図 1.8 検者数に対する係数値変化の推定グラフ

膝関節屈曲と足関節背屈の ICC 値はほとんど同じ値であったので, 変化量もほぼ同値である。例えば, $\rho = 0.98$ の値を得たいならば, 5 人の検者の平均値を用いればよいと推定される。 $\rho = 0.99$ の値を得たいならば 10 人以上の検者数が必要である。

※なお、この資料に関する質問・意見がある場合は、下記まで連絡下さい。

弘前大学医学部 保健学科 理学療法学専攻 対馬栄輝（つしま えいき）

〒036-8564 青森県弘前市本町 66-1

電話 & FAX 0172 - 39 - 5981

E - mail : pteiki@cc.hirosaki-u.ac.jp

URL : <http://www.hs.hirosaki-u.ac.jp/pt/eiki/>

関連図書

- [1] 芝 祐順ほか（編）：統計用語辞典，新曜社，1995.
- [2] 竹内 啓：統計学事典．東洋経済新報社，1992.
- [3] Fisher RA：研究者のための統計的方法．（遠藤健児ほか訳），森北出版，1975，p167-197.
- [4] Shrout PE,Fleiss JL：Intraclass Correlations:Uses in Assessing Rater Reliability. *Psychol Bull* **86**：420-428, 1979.
- [5] James WY,Connie LB,et.al.:Reliability of goniometric measurements and visual estimates of ankle joint active range of motion obtained in a clinical setting.*Arch Phys Med Rehabil* **74**:1113-1118,1993.
- [6] 栗原洋一，斉藤俊弘，他：検者内および検者間の Reliability（再現性，信頼性）の検討. 呼と循 **41**:945-952,1993.
- [7] 谷 浩明：評価の信頼性. 理学療法科学 **12**:113-120,1997.
- [8] Gabrielle R,Maria S:Reliability of assessment tools in rehabilitation:an illustration of appropriate statistical analyses.*Clinical Rehabilitation* **12**:187-199,1998.
- [9] Portney LG,Watkins MP:Foundations of clinical research-Applications to practice-,Appleton & Lange,USA,1993,p505-516.
- [10] Bartko JJ：The intraclass correlation coefficient as a measure of reliability.*Psychological Reports* **19**：3-11, 1966.
- [11] Bartko JJ:On various intraclass correlation reliability coefficients.*Psychological Bulletin* **83**:762-765, 1976.
- [12] Landis JR,Koch GG:The measurement of observer agreement for categorical data.*Biometrics*. **33**,159-174,1977.
- [13] Eliasziw M,Young SL,et al.:Statistical methodology for the concurrent assessment of interrater and intrarater reliability:Using goniometric measurements as an example.*Physical Therapy* **74**:777-788,1994.
- [14] Stratford PW,Goldsmith CH:Use of the error as a reliability index of interest:an applied eample using elbow flexor strength data.*Physical Therapy* **77**:745-750,1997.