測れないものを測る方法－尺度化について－

高木　廣文

平成12年8月
超音波検査技術 Vol. 25 No. 4

統計講座（第8回）

測れないものを測る方法ー尺度化についてー

高木　廣文

はじめに

身長，体重，血液学的な多くの検査項目のように，はじめから数値としてデータが得られるもの以外に，概念上では当然存在するはずに思えても，いざ測定しようとすると極めて難しいことがある．例えば，難病患者のQOL（Quality Of Life，生活の質）などは，概念的には理解できても，実際にどのように測定したらよいのか困ることだろう．このように，現実には測りたいがそれほど容易に測定するわけにはいかない事象は，どのように測ればよいのだろうか．

通常の測定方法では測れないような特性を測定するための手段として，簡単でよく用いられる方法がある．すなわち，その概念に関係すると考えられる項目を多数作成し，それぞれの項目に対する回答に適当な得点を与え，その総合得点でQOLなどを測定するという方法である．

この方法は，尺度化，尺度構成などと呼ばれ，分野によっては得点（スコア）を与えるので，スコア化などとも呼ばれることがある．尺度（Scale，すなわち「ものさし」）を複数の項目から構成する場合，それらの項目の内容がその尺度の概念に適切なものでなければならないことはいうまでもない．

尺度構成された測定用具の適切さは，一般に「妥当性 validity」と「信頼性 reliability」の2つの側面から評価する必要がある．[1] 以下に，これらの考え方について簡単に説明しよう．

Hirofumi TAKAGI：新潟大学医学部保健学科
〒951-8518　新潟県新潟市旭町通2番町746番
e-mail：takagi@clg.niigata-u.ac.jp

尺度構成での妥当性について

妥当性とは，「測定しようとしている内容を，どの程度適切に測定しているか」を示す用語であり，作成した尺度の妥当性の評価は，以下に述べるようにいくつかのタイプに細分して考えられる．

1．基準関連妥当性

大学入学試験の妥当性は，大学入学後の成績とその相関が高いことで評価することができるし，質問紙で塩分の摂取量を推定した結果は，実際の食事調査での塩分摂取量と相関が高ければ，妥当性をもつと考えてよいだろう．このように，適当な評価のための基準（外的基準）との相関で評価される妥当性は「基準関連妥当性 criterion-related validity」と呼ばれている．

さらに，基準関連妥当性は問題にしている尺度の測定と，外的基準の測定が，同時に実施されているか，時間的に将来かによって，「同時的妥当性 concurrent validity」（質問紙調査と栄養調査による食塩摂取量など），「予測的妥当性 predictive validity」（大学入学試験と入学後の成績など）とに区別される．ただし，どちらの場合でも，評価の方法は，問題となっている尺度と外的基準との相関を調べることにより行われるのが普通である．

2．内容的妥当性

算数の計算能力を測定する問題を作成する場合，すべての問題が足し算ばかりで，引き算，かけ算，わり算などを含んでいなければ，その問題は「内容的に妥当ではない」ことになる．このような妥当性のタイプは「内容的妥当性 content validity」と呼ばれている．

一般に内容的妥当性は，算数の問題の例のよう

に簡単な訳でない．QOLの内容は何か，不安を構成するものは何か，など実際には明確に定義できないことも多いだろう．そのような場合には，内容的妥当性は，抽象的な概念の測定に対する努力目標的な意味をもつものと考えられよう．

他のタイプの妥当性として，「構成概念妥当性construct validity」がある．その尺度の測定結果が，他の諸概念・構成概念の尺度の測定結果と，理論的に導かれる仮説に関して，どの程度符合しているかにより評価される．例えば，QOL得点の高い患者は，生活満足度や社会的適応が高いはずであり，情緒不安定性などは低いものと考えられるだろう．

3．妥当性が疑われた場合

理論的な仮説と，実際の測定間に破綻がある場合，いくつかの理由が考えられる．すなわち，

①基礎にある理論の誤り．

②理論の検証のための方法などの誤り．

③測定の信頼性の低さなど．

が考えられよう．しかし，実際には，どれが原因かを特定することは，困難な場合が少なくない．そのような場合，尺度構成に使用した項目を用いて因子分析（多変量解析の方法の一つで，多数の変数間に共通する因子が何かを検討するための方法）などを行うことで，項目間の相関構造を検討したり，複数の尺度間の解析を行うのも，妥当性を評価するうえでは有用なものとなるかもしれない．

尺度構成での信頼性について

尺度構成における信頼性とは，同一項目に関して繰り返し測定をした場合の一致性を示すものである．したがって，2つの反復する測定の結果が一致的であれば，その尺度の信頼性は高いことになる．以下に，信頼性を測るための統計理論を簡単に解説しよう．

ある事象の測定において，真の得点をt，測定得点をx，測定誤差をeとすると，$x=t+e$，と書くことができる．このとき，信頼性を，真の得点tと測定得点xの相関によって定義する．

実際には真の得点を直接測定することは不可能である．そのため，2つの同等な尺度の同時的な測定（平行測定）や，調査の反復測定により，2つの繰り返し測定された尺度の相関を求めることで，信頼性の推定を行うことができる．

1回の測定で信頼性を推定するのは，以下のような方法を用いる．なお，ここでは，複数の項目への回答からの合成得点（「スコア」と呼ぶこともある）について，信頼性係数を求める方法について解説しよう．

いま，p個の変数$x_1, x_2, \ldots, x_p$の合計点（尺度得点，スコア）をY，ケース数をNとする．さらに，$x_i=(x_{i1}, x_{i2}, \ldots, x_{iN})'$，$(i=1, 2, \ldots, p)$，および，$Y=(y_1, y_2, \ldots, y_N)'$とする．したがって，合計得点$y_j=x_{1j}+x_{2j}+\cdots+x_{pj}$，がその事象を表すためのケースjの尺度得点となる．

変数X_i（$i=1, 2, \ldots, p$）と合計点Yの平均値をm_i，m_Yとし，分散をV_i，V_Yとする．すなわち，

$$m_i=\Sigma_j x_{ij}/N, \text{ および } m_Y=\Sigma_j y_j/N$$

である．ここで，Σ_jはすべてのケースについての和を計算することを示す．また，

$$V_i=\Sigma_j(x_{ij}-m_i)^2/N, \text{ および } V_Y=\Sigma_j(y_j-m_Y)^2/N$$

である（$i=1, 2, \ldots, p$；$j=1, 2, \ldots, N$）．

一般によく用いられている「クロンバッハ（「クロンバック」とも呼ばれる）のα信頼性係数（Cronbach's α）」は，

$$\alpha=[p/(p-1)]\cdot[1-\Sigma_i V_i/V_Y]$$

によって求められる．なお，Σ_iはすべての項目についての和を求めることを意味している．

なお，クロンバッハのα信頼性係数を求める場合，尺度得点の計算に用いる項目間の相関係数は，すべて正である必要がある．他の項目と負の相関である項目は，尺度得点の計算に用いないか，得点のつけ方を逆にしなければならない．

尺度得点の計算に用いる各項目が，その尺度と関係があることが尺度構成の前提となっているが，それを検討するための方法がいくつかある．

もっとも簡単な方法は，すべての項目を用いた場合のα信頼性係数と，各項目についてその項目を除いた場合についてを計算し，その違いを比較すればよい．すなわち，α信頼性係数の減少幅が大きな項目は，その尺度に重要な役割をもつと考えてよい．逆に，その項目を除いた方がαの値が

大きくなるようならば，その項目は用いない方がよいことになる．

信頼性係数を求めるのに，主成分分析（多変量解析の方法の一つで，多くの変数のもつ情報を表すような，より少数の合成得点を求める方法）を利用することもできる．項目間に相互に高い相関があり，その和を計算できるのであれば，すべての項目の第1主成分（各項目に最適の重み付けをした合成得点）の負荷量（主成分との相関係数）が正に大きくなるはずである．したがって，求めた第1主成分が負の項目については，その得点を逆順にして尺度得点の計算に用いる必要があり，負荷量が0に近いような項目はその尺度の計算に用いる意味がないことを示すことになる．

左室壁運動異常の重症度の尺度化の例

ここでは，尺度化の例として，左室壁運動異常の重症度を示すための方法を用いることにする．その重症度を判定する方法として，American Society of Evhocardiographyが推奨している左

表1　ドブタミン負荷エコー前後の左室壁運動異常の重症度のスコア

No	負荷	分画																Score	Index
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
1	前	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	19	1.1875
1	後	1	1	1	1	2	2	3	2	2	1	1	1	1	1	1	1	22	1.375
2	前	1	1	1	1	1	1	1	1	2	2	3	2	1	2	1	1	22	1.375
2	後	1	1	1	1	1	1	1	1	1	2	3	1	1	2	1	1	20	1.25
3	前	2	2	2	3	4	4	4	2	2	1	1	1	1	1	1	1	32	2
3	後	2	3	3	3	4	4	4	3	3	2	1	1	1	1	1	1	37	2.3125
4	前	1	1	1	1	1	2	2	3	2	2	2	1	1	1	1	1	23	1.4375
4	後	1	1	1	1	1	2	3	4	3	2	1	1	1	1	1	1	25	1.5625
5	前	1	1	1	2	2	2	1	1	1	1	2	2	3	1	2	2	25	1.5625
5	後	1	1	1	3	3	3	2	2	1	1	2	3	2	1	2	2	30	1.875
6	前	1	1	1	1	1	1	2	2	2	2	1	1	1	1	1	1	20	1.25
6	後	1	1	1	1	1	1	1	2	2	1	2	1	1	1	1	1	19	1.1875
7	前	1	1	1	1	1	1	1	1	1	2	2	2	2	1	1	1	20	1.25
7	後	1	1	2	1	1	1	2	2	2	2	2	2	2	1	1	1	24	1.5
8	前	1	1	1	1	1	2	2	2	3	1	1	1	1	1	1	1	21	1.3125
8	後	1	1	1	1	1	1	2	2	2	1	1	2	2	1	1	1	21	1.3125
9	前	1	1	1	1	1	1	1	3	4	4	4	3	1	1	1	1	29	1.8125
9	後	1	1	1	1	1	1	1	4	4	4	4	4	2	2	2	2	35	2.1875
10	前	2	2	2	3	4	1	1	1	1	1	2	2	2	1	1	1	27	1.6875
10	後	2	2	1	2	2	1	1	1	1	1	2	2	2	1	1	2	24	1.5
11	前	1	1	1	3	3	3	2	2	1	1	1	1	1	1	1	1	24	1.5
11	後	1	1	1	1	2	2	2	2	1	1	1	1	1	1	1	1	20	1.25
12	前	1	1	1	1	1	1	2	2	2	3	3	1	1	1	1	1	23	1.4375
12	後	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	17	1.0625
13	前	1	1	1	3	3	3	1	1	1	1	1	1	1	1	1	1	22	1.375
13	後	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	19	1.1875
14	前	1	1	1	1	1	2	2	2	3	2	2	1	1	2	2	2	26	1.625
14	後	1	1	1	1	1	1	1	2	1	1	1	1	2	2	2	2	21	1.3125
15	前	2	2	3	1	1	1	2	2	2	1	1	1	1	1	1	1	23	1.4375
15	後	1	1	2	1	1	1	1	1	1	1	1	2	2	1	1	2	20	1.25

1 : normal, 2 : hypokinesis, 3 : akinesis, 4 : dyskinesis.

室壁16分画モデルを用いた点数化データを**表1**に示した.

表1のデータでは，各分画において壁運動normalを1点，hypokinesisを2点，akinesisを3点，dyskinesisを4点として，その合計得点をScoreとしており，この得点が高いほど，より重度が高いとする方法である．また，Scoreを分画数で割った値，すなわち平均得点をIndexとして表示している．なお，**表1**はドブタミン負荷エコー前後の各点数である．まず，この方法による尺度構成の信頼性の検討を行い，次いでドブタミン負荷前後でのスコアに差はあるのかを検討しよう．

1．第1主成分負荷量とα信頼性係数[2)]

表2は，16分画ごとの平均値，不偏分散，その分画を除いた場合のα信頼性係数，第1主成分負荷量を示したものである．さらに，第1主成分負荷量が第1分画～7分画まで値が負なので，点数の付け方を逆順にした場合について，その分画を除いた場合のα信頼性係数を最右欄に示した.

なお，尺度構成を行う場合，すべてのケースで同一の測定結果になるような項目を使用することはできない．例えば，子宮癌の重症度を示すようなスコアを作成したい場合，「性」という項目は意味を持たないだろう．また，全員ではないが大部分の者が同一の回答をするような項目も，あまり尺度構成では使用されない．この辺りのチェックは，(不偏) 分散の値を調べることで検討できる．すなわち，分散が0ならば回答はすべて同一であり，そのような項目は尺度構成には用いられない.尺度構成に使用する項目の検討を行う場合，調査対象者数が多ければあまり問題にならないが，少なすぎると項目によっては回答のばらつきが小さく，すなわち分散が小さくなってしまい，相関係数や主成分分析などの計算が適切に行えない場合が起こり得る．ここでの例では，15ケースという極めて小標本なので，ドブタミン負荷前後のデータを個々のケースとして分析した（実際には，前後を別々に解析してもほとんど結果に変わ

表2　各分画の基礎統計と信頼性分析結果

分画	平均値	不偏分散	α[1)]	負荷量[2)]	α[3)]
1	1.167	0.144	0.655	−0.693	0.818
2	1.200	0.234	0.643	−0.733	0.815
3	1.267	0.340	0.656	−0.612	0.819
4	1.500	0.672	0.647	−0.691	0.806
5	1.667	1.057	0.647	−0.783	0.800
6	1.700	0.838	0.652	−0.699	0.806
7	1.733	0.754	0.657	−0.637	0.819
8	1.900	0.714	0.640	0.112	0.834*
9	1.800	0.855	0.640	0.237	0.826*
10	1.600	0.731	0.667	0.552	0.804
11	1.700	0.838	0.679	0.636	0.801
12	1.500	0.603	0.651	0.477	0.812
13	1.367	0.309	0.682	0.243	0.830*
14	1.167	0.144	0.678	0.531	0.819
15	1.167	0.144	0.662	0.400	0.823
16	1.233	0.185	0.669	0.347	0.824
合計得点	23.667	23.195	0.674	——	0.826

1) 各分画については，その分画を除いた場合のα信頼性係数を示す.
2) 主成分分析により第1主成分負荷量.
3) 第1主成分負荷量が負の分画の得点の付け方を逆順にした場合のα信頼性係数を示す.
*) その分画を除いた方がα信頼性係数が大きい場合を示す.

りはなかった).

さて，**表2**の合計得点のα信頼性係数は0.673であった．一般に，αの値は0.8以上あることが望ましいので，その点からはやや値が小さいといえる．この主な理由は，第1分画から第7分画までと第8分画から第16分画までの重症度の傾向が，このデータでは逆転していることによる．これは，**表2**の第1主成分負荷量の値が正負に分かれていることで，すぐに見つけることができるだろう．

このような場合，単純に各分画の点数を合計することはできない．第1主成分負荷量の値が負の分画については，点数の付け方を他のものの逆順にして計算をする必要がある．**表2**の最右欄はそのようにしてα信頼性係数の計算を行った場合の結果を示している.このような操作を行うことで，合計得点のα信頼性係数の値が0.826と大幅に増加していることがわかるだろう．もっとも，点数化の方法は各分画の重症度の程度にしたがっているので，ここで行っているように，部分的に重症度を逆転させて和をとるというのは，内容的には妥当ではないかもしれない．第1分画～7分画と第8分画～16分画の2つのグループごとの重症度の計算を個々に行った方がよいのかもしれないが，ここでは一般的な手順を紹介するために，このような方法を採ることにする．

さて，個々の分画の重要度については，それを除いた場合のαの値を調べればよい．**表2**から，第8，9，13分画を除いた場合のα値が増加することがわかる．すなわち，これらの分画のデータは,左室壁運動異常の重症度をスコア化する場合，必要のないものであるといえる．ただし，この例はわずか15例についてのものであり，実際的な尺度構成の検討を行う場合には，より多数の標本をもとに解析を行う必要があるので注意が必要である．

2．ドブタミン負荷前後でのスコアの差

表3にドブタミン負荷前後で左室壁運動異常重症度判定のための尺度得点の平均値の比較のための統計量を示した．**表1**のデータは，同一人の繰り返し測定によるものなので，検定は「対応のある場合の母平均値の差の検定（t検定)」を用いた[3]．また，第1分画から第7分画までの点数化をそのまま用いた場合と，逆順にした場合についてそれぞれの統計量を示した．

第1分画から第7分画までの点数化をそのまま使用する場合，ドブタミン負荷前の平均得点は23.7，負荷後の平均得点は23.6と，両者には0.1とほとんど差がなかったが，負荷前後の尺度得点には0.755と高い相関が認められた．一方，第1分画から第7分画の点数を他と逆順につけると，ドブタミン負荷前後の平均得点差は1.067とやや大きくなるが，標本数が極端に小さいため統計学的にはp値が約0.2と有意とはならなかった．しかし，負荷前後の単相関は0.880とさらに大きくなり，尺度構成の安定性を示していた．

ところで，上記の結果は尺度得点を，身長や体重の測定値のような量的データとして解析したものであった．実際のところ，ここで説明したように，尺度化とは順序尺度で構成された多数の項目から量的データとなるような「ものさし」を作る方法といえるので,用いた方法に別段問題はない．しかし，検定の基礎として，尺度得点の母集団での分布が正規分布に従うか不明であることから，ノンパラメトリック検定を用いることも多い．表には示していないが，このような研究デザインに

表3　対応のある場合の2グループの尺度得点の比較

ドブタミン負荷	左室壁運動異常の重症度	
	平均値	標準偏差
(1)　1～7分画の点数化の反転前		
前	23.733	3.575
後	23.600	5.938
前後の差	0.133	3.998
母平均値の差の検定	t＝0.129	p＝0.899
単相関係数	r＝0.755	p＝0.001
(2)　1～7分画の点数化の反転後		
前	37.667	6.264
後	38.733	6.227
前後の差	－1.067	3.058
母平均値の差の検定	t＝1.351	p＝0.198
単相関係数	r＝0.880	p＜0.001

対して用いられる「Wilcoxonの符号付き順位和検定」を行うと[3]，第1分画から第7分画の点数化をそのまま使用する場合，負荷前後のz値0.032，p値0.975となる．また，第1分画から第7分画の点数化を逆順にした場合，z値1.263，p値0.206となり，検定結果は**表3**と大差ないものが得られる．

上記の結果は，いずれにしてもドブタミン負荷により左心室壁運動異常の重症度に相違は認められなかったことを示している．ただし，統計学的仮説検定法は仮説が棄却されない場合（有意差が認められない場合），態度保留が原則であり[4]，かつ用いた標本数が少ないこともあり，重症度の尺度化の方法の妥当性の検討とともに，標本数を増やして検討する必要性を示す結果といえるだろう．

おわりに

今回は，通常の測定方法では測れないような事象の測定の方法として，尺度構成法（いわゆる「スコア化」の方法）を紹介した．尺度構成法では，作成している尺度の信頼性の検討を行う場合，多変量解析の手法を用いるので，理解するのがそれほど容易ではないかもしれない．また計算の方法もそれほど簡単ではない．多変量解析に関しては，今後機会があれば解説していきたいと考えている．なお，今回の計算はすべて，私が作成した統計学ソフトHALWIN（はるうぃん）を用いて行った[2]．正式なバージョンはやや値段が高いので，ちょっと使ってみたいと思う読者は，私のホームページ（http://www.clg.niigata-u.ac.jp/~takagi/）の「HALBAU/HALWINの世界」に「HALWINアップグレード」があり，そこから試作版をダウンロードすることができるので，試してみるとよいだろう．

参考文献

1）水野欣司，野嶋栄一郎 訳：テストの信頼性と妥当性，朝倉書店，1983（E.G. Carmines and R.A. Zeller: Reliability and Validity Assessment, SAGE Pub., Beverly Hills, 1979）．
2）高木廣文：HALWINによるデータ解析，1998，現代数学社（京都）．
3）高木廣文：対応のあるデータの比較，超音波検査技術，24，pp424-429，1999．
4）高木廣文：統計学の基本的な考え方，超音波検査技術，23，pp329-334，1998．