項目反応理論について

1　項目反応理論の概略

1.1　項目反応理論でできること

豊田 (2002)[*1]は、項目反応理論で以下のことができると述べている。

- 複数のテスト間の結果の比較を容易にする
- 測定精度をきめ細かく確認できる
- 平均点をテスト実施前に制御できる
- テスト得点の対応表が作成できる
- 受験者毎に最適な問題を瞬時に選び、その場で出題できる

最後のものなんかは、コンピュータに項目のストックがあって初めてできる、かなり進歩的な使い方だが、コレができたらすごいなぁと誰しも普通に感心してもらえるのではないだろうか。

さらに、項目反応理論にはもうひとつメリットがある。それは、

- 無作為抽出をしなくてよい

ということだ。古典的な尺度理論では、サンプルの母集団をもとに尺度値を決めたり、尺度の基準を求めていたので、しっかりとしたサンプリングによって母集団を代表するような被験者を集めてこなければならない。しかし現実問題として、それではコストがかかりすぎるので、実際は大教室の授業を受けに来た学生にデータを取らせてもらって、はいオワリ、ということが少なくない。これでは学生心理学とか、大学生尺度と言われても言い返せないだろう。

項目反応理論は、サンプルに基づいて尺度を標準化するわけではない。一言で言えば、被験者の特性値の母数を使わずに項目を標準化しているので、無作為抽出の縛りから解放されるのである。

こんなことができるなんて、なんて素敵なんでしょう。と強烈なあこがれを抱いて、実際的なステップに入っていこう。

[*1] 豊田秀樹 (2002)『項目反応理論　入門編』朝倉書店

1.2　項目反応理論でやっていること

さて、能書きはこれぐらいにして、ではそのスゴイ項目反応理論とは一体何をやっていることなのだろうか？

数学者には怒られるかもしれないが、例によって感覚的に説明してみよう。

尺度であれテストであれ、一度実施すれば、平均点や分散がわかるので、その情報をもとに洗練し、より一般的で、頑健で、使い勝手の良い、標準化された尺度をつくることができる。その作り方は前章までに述べたとおりである。しかし、尺度はそれを適用される被験者に依存していると言える。そのとき、たまたま質の悪い (失礼) 被験者が相手であれば、それをもとに標準化して、一方的に出来の悪い項目だな、というのはどうも項目に分が悪い。被験者はその時々で、調子が良かったり悪かったりするだろうから、被験者の方ももっと標準化された、いいサンプルをもってこいよ、といいたくなる。

項目反応理論は、まさにこの要望に応えるものだ。尺度と被験者の組み合わせによりデータが得られたら、それをもとに尺度も標準化するし、被験者も標準化するのである。どちらか一方が良いとか悪いとか言わずに済むように、どちらの情報も十分活かせるような標準化をする。尺度は因子分析によって、潜在変数に影響されるものとして考えられるが、この考えを応用して、被験者の能力にも潜んでいるであろう潜在変数を見いだすのだ。いわば、顕在化した被験者の反応ではなく、それの誤差を取り除いた、潜在変数としての被験者の反応 (あるいはテストで測る能力) を引き出す。これが項目反応理論の狙いなのである。

余談であるが、項目反応理論は、尺度が一因子であることを前提に話を進める。なんらかの能力を測定するとき、多元的なものは考えずに、一次元的な尺度の上での潜在的能力の差異が反映される、と考えるのである。じゃあ下位概念の違いが取り出せないじゃないの、と憤慨する向きもあるかもしれない。しかし、既に述べたように (**??**節)、尺度というのは一つのものに対して多角的にアプローチするものであり、そもそも複数の概念を引き出すためのものではないから、その反論は当たらないので、悪しからず。

1.3　項目反応理論の考え方のヒント

さて、先に項目反応理論は尺度と被験者の両方を標準化しちゃう、と書いた。尺度の標準化はこれまで述べてきたように、下記のような方法で行われる。すなわち、

- 項目の平均点や分散をもとに、各アイテムカテゴリの相対度数を求め、標準正規分布から尺度値を算出
- 項目間の相関関係から[*2]、尺度の内部構造を明らかにする
- 妥当性の基準に従って、因子構造を標準化 (一般化) する

[*2] 相関係数は標準得点同士を掛け合わせたもの、つまりこのプロセスは既に標準化を含んでいる

これらに基づいて、一般化された因子構造から因子得点を算出する、といった方法で被験者が査定される。これが古典的尺度理論なのである。これだと、上で述べたような被験者の潜在的特性は査定されない。きちんとした尺度で測ったので、あなたの得点 (〜度) はこれこれです、と顕在的に示されるだけである。被験者の潜在的な能力[*3]は、テスト実施時にいかなる誤差が紛れ込もうとも、項目側の誤差変動に吸い取られて、測定されないことになる。

じゃあ何とかして、被験者の潜在的な能力を算出できないだろうか。ちょっと考えると、n 個の項目、N 人の被験者では $n < N$ である。しかも、N の方が圧倒的に多い (普通 n は二桁、N は三〜四桁のオーダーになる)。被験者の能力を算出するのだから、結果は一人一人に当てはまらなければならず、N 個の点数を算出することになる。そんなに多くの未知数を推定できるのだろうか？と不安になってくる人もいるかもしれない。

もちろん、このままでは無理だ。しかし数学的発想によくあるように、いくつかの制限 (条件) をもうけてやれば、これは可能である。その制限のヒントになるのが、**正規分布**と **ICC(Item Characteristic Curve, 項目特性曲線)** である。

2　項目反応理論の数理

2.1　正規分布の応用

被験者を多く取ると、能力の分布は正規分布に近似していると考えられる。また、能力測定の特徴から、能力が高いものはより下位の問題をパスする、ということだ。偏差値 70 の人は、偏差値 30 の人が解けた問題は当然解ける。逆はそうならない。当たり前ですね。となると、ある問題が解ける人数というのは、正規分布に含まれる度数に従って累積的に増加して行くに違いない。

図 1 に、正規分布の確率密度 (その確率に該当する人がいる度数) を示した。横軸にあるのは能力である。横軸はこのままにして、累積度数を表現すると図 2 のようになる[*4]。

数学的にはこれを関数として表現する。正規分布の確率密度関数は、変数 x が標準化されていたとすると (標準正規分布の密度関数)、

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$$

である。この累積分布関数は、

$$\Phi(f(x)) = \int_{-\infty}^{f(x)} f(x)dx$$

[*3] 尺度によっては、必ずしも「能力」という言葉は正しくない。ただ、項目反応理論はテスト理論から来ているので、それを考えるとこの表現の方がわかりやすい。本書では以下も「能力」で通すが、これは測定しようとするもの (因子) と被験者の相関を意味しているものとする。

[*4] この図は Excel で簡単に描写できる。まず、能力の散らばる範囲を決めて、一列に等間隔に区切って入力していく。範囲を −3 から 3 にする場合、たとえば $A1 = -3.0$、$A2 = -2.95$、$A3 = -2.90 \cdots A121 = +3.0$ とする。次に、関数 NORMSDIST を使って、$B1 = NormDist(A1)$ のようにするだけである。ちなみに、1 は、累積度数を算出した k 行目を参照しながら、$C_k - C_{k-1}$ のようにすればよい。

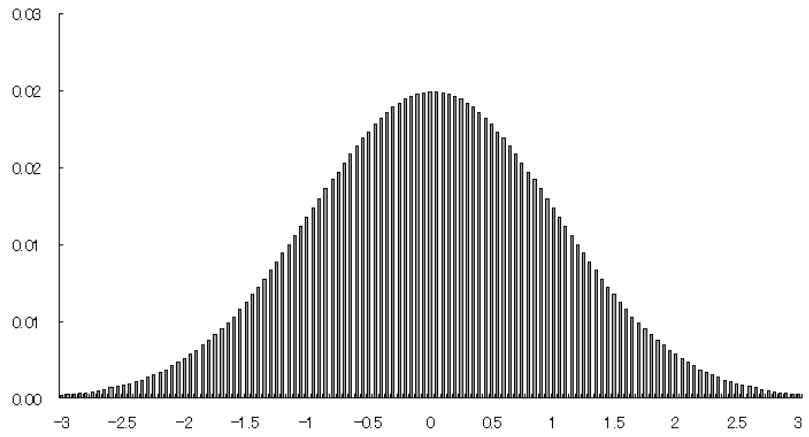

図 1　Excel で描いた正規分布の確率密度

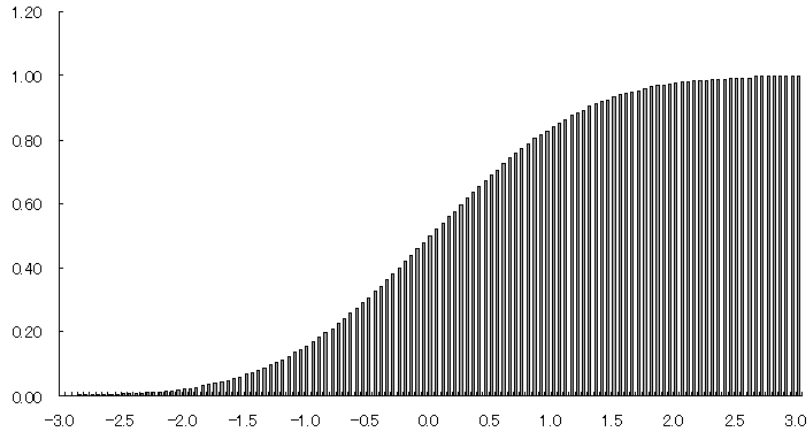

図 2　同じく累積確率密度

で表される。ただ、この式は積分を含んだ形なので、計算に不便だということで、以下の近似式を用いることが慣例になっている。

$$\int_{-\infty}^{f(x)} f(x)dx \simeq \frac{1}{1+exp(-1.7*f(x))} \tag{1}$$

この式は積分が入ってないので、変数を増やして曲線を変形するのに向いている。関数にエクスポネンシャル (exp) が入ったこの関数は、一般にロジスティック関数と呼ばれる。このロジスティック関数を変形するために、三つの変数、a, b, c を加えた一般的なモデルは以下のようなも

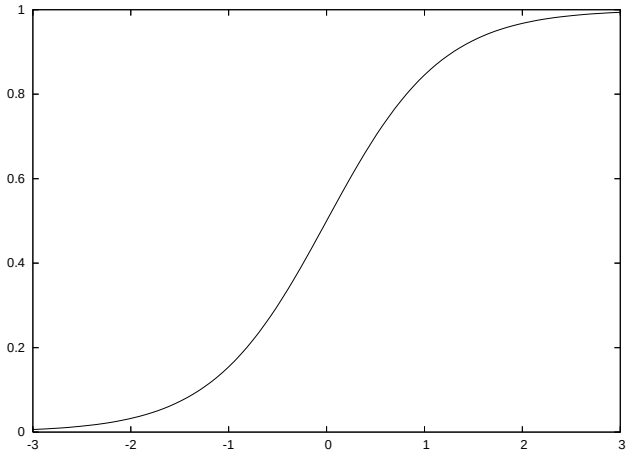

図 3　式 1 で描かれるグラフ

のである。

$$p_j(x) = c_j + \frac{1 - c_j}{1 + exp(-1.7a_j(x - b_j))} \tag{2}$$

ここで j は項目である。

さて、a, b, c という三つも一度に変数が出てくると、なにがどんな意味を持っているのかわかりにくい。そこで順に変数を操作して、関数カーブがどのように変わるか見ていこう。これらのデフォルトは、$a = 1.0, b = 0.0, c = 0.0$ であることを忘れずに。

まず a から。これは元が 1.0 だったので、ここから増やしたり減らしたりして考えてみよう。$a = 1.0$ と $a = 2.0$ のグラフを図 4 に示す。

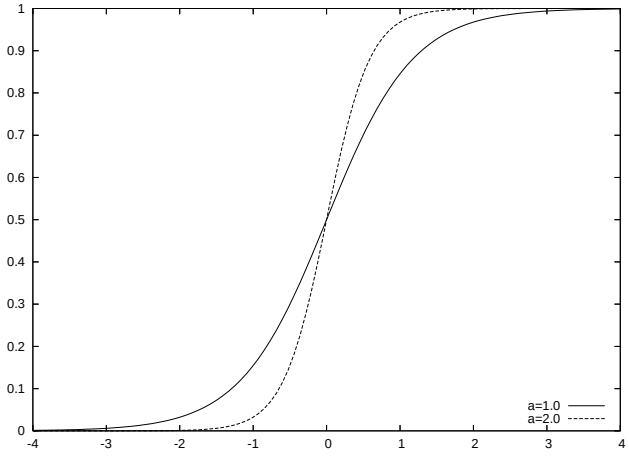

図 4　$a = 1.0$ と $a = 2.0$ のグラフ

なにやら角度が急になっていることにお気づきだろうか。これはもともと、正規分布の累積確率密度関数だから、累積する前の形に戻して正規分布と比較すると、図 5 のようになっているこ

とがわかる。このように、正規分布をぐっと狭めた形になっており、分散が小さくなったことが

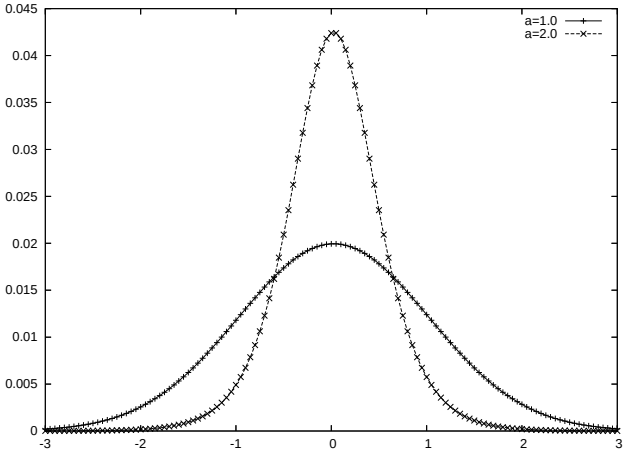

図 5　$a = 1.0$ と $a = 2.0$ のもとになる正規分布の形

わかる。正確には分散が 1/4 になっている。このことから、a の数値を変えるとロジスティック関数の傾きが変わり、それはつまり正規分布の分散が変化することを意味していることがわかる。

この a の値が小さくて、グラフの傾きが緩やかであれば、被験者母数 (=被験者の能力) が上がるに連れて徐々に正答率が高くなる問題であると言える。逆に傾きが急であれば、あるレベルに達すると、グッと正答率が上がるような問題である。そこでこの a のことを、**識別力**といい、普通 0.3 から 2.0 ぐらいの数値を取る。マイナスの数値は、能力が上がるほど正答率が低くなる問題なので、テスト項目として不適切であるか、逆転項目として考えられるため、ここでは考慮しない。

次に b の数値に目を向けてみよう。b はデフォルトが 0.0 であるが、+1.0 や +2.0 のときはどうなるか、グラフにしてみよう (図 6)。グラフが順に右にずれて行くことがわかるだろうか。右にずれる、ということは、能力が高くないと正解率が上がらないことを意味しているから、この b は項目の難しさについての係数であることがわかる。b は別名、**困難度**である。もちろん簡単な問題もあるので、この b は正の数も負の数も取りうる。

最後に c について見てみよう。c はデフォルトが 0.0 であるが、0.3 や 0.5 のときは図 7 のようになる。これみれば、今度はグラフ全体が上に上がっていくことがわかる。グラフ全体が上がる、ということは全体の正答率が上がっていくことである。つまり、能力があろうと無かろうと、これぐらい (の点数) は取れちゃうんですよ、というテストの意味を無視した偶然性を表していることになる。この c は**当て推量**との別名があり、0 から 1 までの値で推定される。

このようにしてみると、ロジスティック関数を使う利点－関数の変形が容易であること、が実感できたことと思う。ところで、数学的にはこの a, b, c の三つの変数を全て使うモデルがもっとも一般的であるが、実際の運用は a, b の二つまでで済ますことも多い。二変数しか使わないものを 2 母数ロジスティックモデルと呼び、同様にひとつ (a だけ推定)、三つの場合も 1 母数、3 母

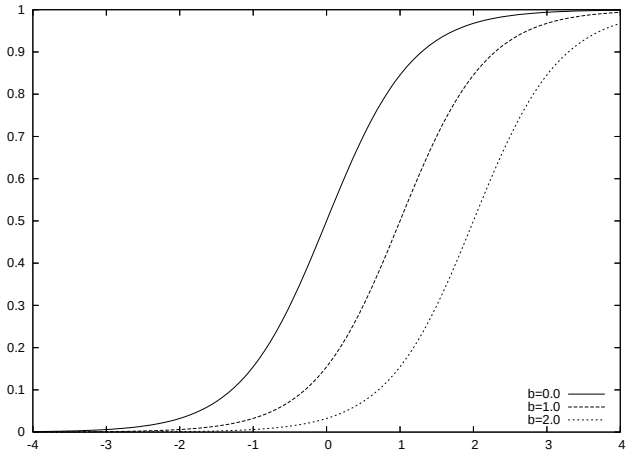

図 6　$b = 0.0, 1.0, 2.0$ のグラフ

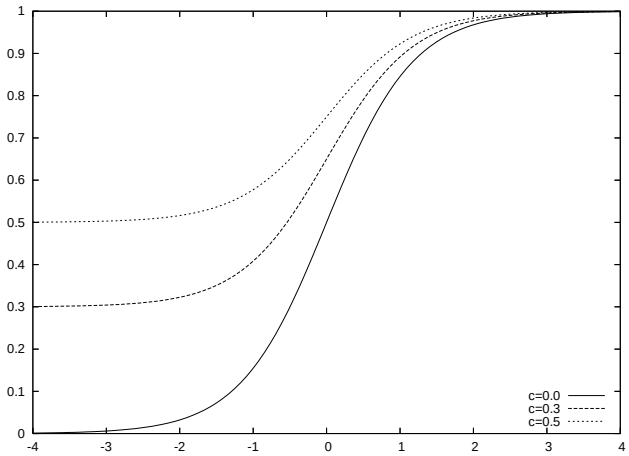

図 7　$c = 0.0, 0.3, 0.5$ のグラフ

数ロジスティックモデルとよぶ。推定すべき母数の数が多い場合は、データ数もたくさん要るようになってくる。

2.2　ICC とは？

では次に、ICC について考えてみよう。

図 3 で示されているのは、被験者の能力が正規分布に従うと仮定したとき、ある問いについてどれぐらいの回答者が出るか、を示しているグラフだと言い直しても良い。つまり、偏差値が 70 とか 80 ある人 (標準得点で +2.0, +3.0) は、96% 以上の確率でその問題に正解する、とも読める。つまり、このグラフはある項目の測定精度についての理論値になっていたのだ。

理想的にはこの形として、実測値はどうなのだろうか。項目特性曲線 ICC とは、この実測値の

ことを指している。理論値の横軸が成績だったように、実測値の横軸にはテスト全体での成績がくる。次に、被験者を成績順にいくつかのグループに分ける。5〜7 等分して、頭のいい人グループからそうでないグループにまで、何段階かに区分する。

次に、グループごとに**通過率**を計算する。通過率とは、そのグループでどれぐらいの割合の人間がその問いにパス (pass, 正答) したか、を表す割合である。その問題に正解していれば 1、間違っていれば 0 とコード化した変数 P_{ij} を作り、$P_j = \sum P_{ij}/N$ とすれば求まる。当然のことながら、全体での得点が高い人はある項目に正解している確率も高かろうし、逆もまた真であから、高成績者グループになるほど通過率は上がっていくだろう。

図 8 に示すのは、とあるテストにおける回答者のデータを、成績が高いものから順に五等分し、通過率を描いたものである。

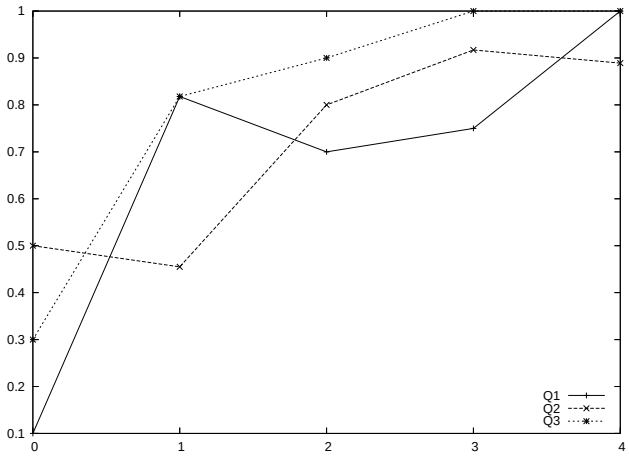

図 8　通過率の折れ線グラフ

この折れ線、つまり成績順にソートされたある項目の通過率を結んだものが、実測値としての ICC である。当然のことながら、この実測値は理論値のように美しい形で得られるわけはない。実測値は折れ線だし、理論値はなめらかな曲線なのだ。ただ、散布図に回帰直線を当てはめるように、実測値のデータがあればそれに理論モデルの曲線を当てはめることができる。項目反応理論とは、要するにこの理論カーブを実測値に当てはめる理論なのである。

2.3　母数の推定

なぁんだ、式の当てはめなら最小二乗法を使えばいいんじゃないか、簡単簡単。と思った人は、半分正解で半分間違いである。ICC で描かれている横軸は、五つの段階に分けた離散変数であり、最終的に求めたいのは、被験者の (目に見えない) 特性 θ なのである。しかも、通過率は実際に何割の人がその問題に正答を出したか、ということから得られた数値だが、理論的カーブであるロジスティック関数の縦軸は「特性値が θ_i であれば、x% の確率で正答する」という確率変数についての話になっているのである (この辺でロジックが飛躍する感じがするかと思うので、しっか

りとついてくること)。

我々が手にしているデータは、i さんが項目 j に正答したかどうか、というデータである。被験者 i の回答パターンを全部まとめたベクトルを、

$$\boldsymbol{u}_i = [100101\cdots 1]$$

と表してみよう。これと未知の母数、つまり項目母数 a_j, b_j, c_j のもと、被験者母数 θ_i で観測された、と考えると

$$f(\boldsymbol{u}_i \mid \theta_i, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{j=1}^{n} f(u_{ij} \mid \theta_i, a_j, b_j, c_j) \tag{3}$$

と表現される[*5]。

データ全体は、というと被験者それぞれが独立だと考えられるので、

$$f(\boldsymbol{u} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{i=1}^{N} \prod_{j=1}^{n} f(u_{ij} \mid \theta_i, a_j, b_j, c_j) \tag{4}$$

である。

この関数で、$\boldsymbol{U}$ はデータとして得られており、θ_i, a_j, b_j, c_j が確率関数である。データがある確率変数の元で得られたとき、確率変数の値がどのようなモノであったかを推定する方法を最尤推定法と呼び、この方法を用いれば $\boldsymbol{U}$ が最も得られやすい (尤もらしい) とされる母数の推定値が得られる。最尤推定については**??**節を参照して欲しい。数学的には大変面倒で、特殊な計算ソフトを使わないと算出できない。IRT の有名なソフトとして、BILOG-MG というのがあるが、250 ドルもするので、熊谷さんのフリーソフトウェア、Easy Estimation(http://itranalysis.main.jp) などを使って計算してみるといいだろう。

2.4 事例：多変量解析法のテスト

ある大学で、「多変量解析法」という講義があり、そこでのテスト結果を IRT で分析してみた例を示そう。テストは 20 問で、受験生は 52 名である。テストの内容は、表 1 のようなものだった。

まず、それぞれの問いについて ICC を描いた (図 9〜15)。

なるほど、これを見ると F2 は難しすぎたことがわかるし、C1 は簡単すぎ、B1 は途中で凹むので変な ICC だな、ということがわかる。これらのデータをもとに、項目母数を推定するべくソフトを動かしてみたが、D3 と E1 は推定に不向きだったようで[*6]、この二つを除くと表 2 のような結果が得られた[*7]。識別力のもっとも高い項目は G2 で、もっとも低いのは C3 である。ちなみに、G2 は「重回帰分析の結果の表を見て、どの変数が最も説明力があるといえるか、理由と共に答えなさい」というものである。これはあるレベルに達すると正解できるという意味で、被験

[*5] 条件付き確率については**??**節を参照

[*6] 推定しにくい値は、極端に易しい項目や難しい項目である。通過率が高すぎるか、低すぎるものは除外する。次に、合計得点との相関係数が低すぎると、一次元性が疑わしくなるので除外する。

[*7] EasyEstimation を使った結果で、このソフトは二母数モデルの推定しかできないのが残念である。

表 1　テストの内容

問題	内容	詳細
A	概念・用語	多変量解析の用語を解説する
B	記述統計	計算式を書くもの
C	記述統計 2	計算式が何を表しているか読み取るもの
D	回帰分析	回帰式の基本性質
E	行列の計算	行列の計算問題
F	因子分析	因子分析の基本定理について
G	応用問題	統計パッケージの出力から結果を読み取る

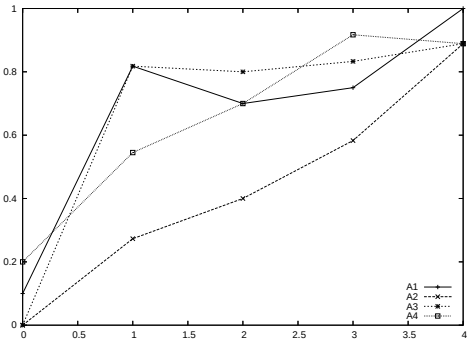

図 9　A の問いについての ICC

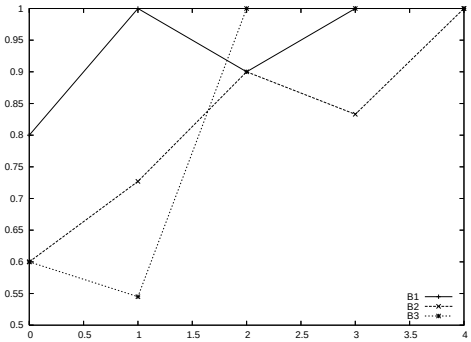

図 10　B の問いについての ICC

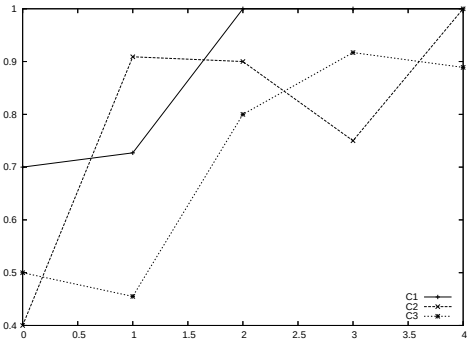

図 11　C の問いについての ICC

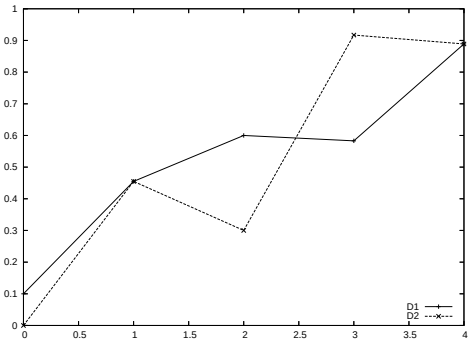

図 12　D の問いについての ICC

者の能力の有無を最もはっきりと区別する質問だったことがわかる。対する C3 は、「$\frac{1}{N}\sum z_{wi}z_{yi}$ は何を表しているか」、という問題で[*8]、これはあんまりよい問題だったとはいえないようだ。図 16 に G2 と C3 の ICC を描いた。$x = 0$ 近くでの傾きの違いが明らかである。

[*8] 念のため、正解は「変数 w と y の相関係数」。

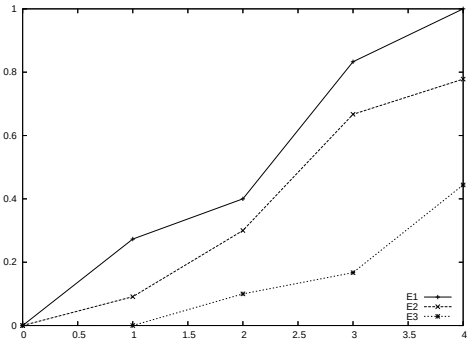

図 13　E の問いについての ICC

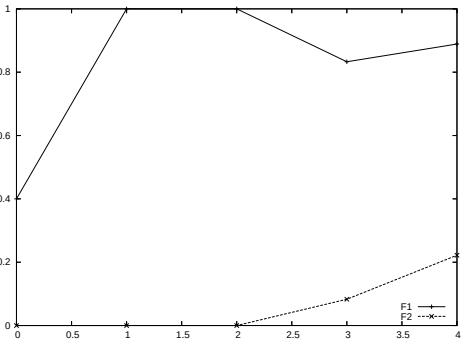

図 14　F の問いについての ICC

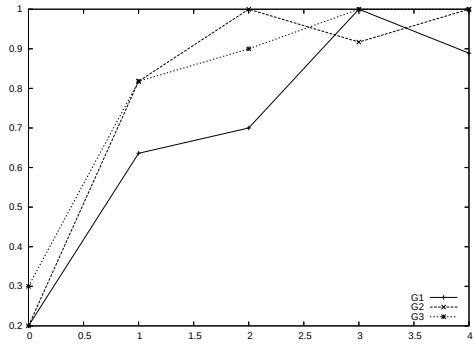

図 15　G の問いについての ICC

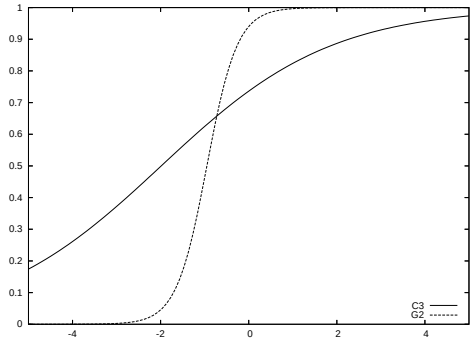

図 16　最も識別力の高い G2 ともっとも低い C3 の ICC

次に困難度を見てみよう。困難度がもっとも低いのが B1 の問いで、「w の平均 $\bar{w}$ を求める式を書け」である。確かに簡単な問題だ。一方、最も困難度が高い F2 は「共通性を式で表せ」というものである。この二つの ICC は図 17 に示した。こういったデータから、テスト項目として良かったもの、悪かったものを選別することができる。また、受験者ひとり一人の θ_i を推定することも可能で、テストの善し悪しに依存しない評価ができるのである。

表 2　IRT で推定した結果

itemID	通過率	合計との相関	識別力 a_j	困難度 b_j
A1	0.673	0.546	0.67656	-0.82464
A2	0.423	0.516	1.05612	0.28116
A3	0.673	0.602	0.74076	-0.77277
A4	0.654	0.544	0.7805	-0.65497
B1	0.942	0.442	0.49457	-4.22575
B2	0.808	0.450	0.54271	-1.93461
B3	0.827	0.554	0.82433	-1.58988
C1	0.885	0.537	0.67481	-2.37387
C2	0.788	0.494	0.46359	-2.0053
C3	0.712	0.449	0.3038	-1.98659
D1	0.519	0.467	0.49567	-0.13879
D2	0.519	0.606	err	err
E1	0.500	0.654	err	err
E2	0.365	0.573	0.99804	0.49137
E3	0.135	0.353	1.53279	1.27066
F1	0.827	0.485	0.59517	-1.97173
F2	0.058	0.227	0.51262	3.5166
G1	0.692	0.625	1.06214	-0.69208
G2	0.788	0.695	1.70839	-0.94984
G3	0.808	0.637	1.23395	-1.17999

2.5　テスト情報関数

ここでは、構成されたテスト (尺度) の精度の良さ (=信頼性) を表現する、IRT 独自の方法をみておこう。テストをした結果、ある被験者 i の尺度値が θ_i と推定された、としよう。しかし、これはただの推定値なので、$\hat{\theta}_i$ と表現すべきものである。これは確率変数で、うまく推定できている場合もあれば、そうでない場合もある。一応、最尤法というやり方で推定しているんだから、尤もらしい値にはなっているんだろうけれども。

推定値じゃない θ_i があったとして、$\hat{\theta}_i$ が必ずしも一致しないのであれば、その誤差

$$e = \theta_i - \hat{\theta}_i$$

がどれぐらいなのかを見積もる必要がある。これが IRT における信頼性に関わってくるのである。また、IRT 独特の利点は、尺度値ごとにこの誤差の大きさ=測定精度を算出できる点にある。

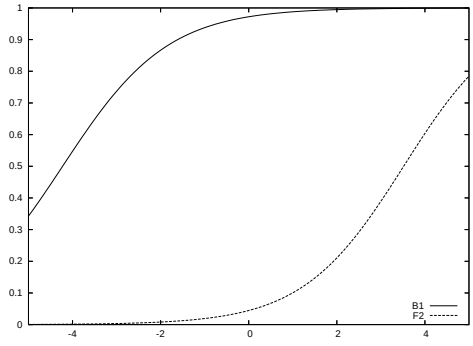

図 17　最も困難度の高い F2 ともっとも低い B1 の ICC

尺度値が高いときはうまく推定できているけど、低いときはうまく推定できてないよ、という情報があれば、テスト運用のときに大変有益であるし、古典的テスト理論ではできなかったことでもある。

ここでやりたいことは θ_i を与えたときに、ロジスティックモデルによって推定される値がどれほどの分散をもっているか、を見ることにある。最尤推定で解を求めたことの特徴は、

- n が大きくなれば、推定値の分散は正規分布に近づく
- n が大きくなれば、推定値の平均は真の値に近づく
- n が大きくなれば、推定値の分散は $1/I(\theta)$ に近づく

というものである。ここで、$I(\theta)$ はフィッシャー情報量と呼ばれる数値で、3 母数モデルの場合

$$I(\theta_i) = 1.7^2 \sum_{j=1}^{n} \frac{a_j^2 (p_j(\theta_i) - c_j)^2 q_j(\theta_i)}{p_j(\theta_i)(1 - c_j)^2}$$

となる。2 母数、1 母数モデルの場合はそれぞれ

$$I(\theta_i) = 1.7^2 \sum_{j=1}^{n} a_j^2 p_j(\theta_i) q_j(\theta_i)$$

$$I(\theta_i) = 1.7^2 a^2 \sum_{j=1}^{n} p_j(\theta_i) q_j(\theta_i)$$

となる。ここでの $p_j(\theta)$ は、尺度値 θ の人が項目 j に正答する確率であり、$q_j(\theta)$ は同じく誤答する確率。つまり $q_j(\theta) = 1 - p_j(\theta)$ である。このフィッシャー情報量は、IRT の文脈では特にテスト情報量、あるいは θ を使ったテスト情報関数 $I(\theta)$ と呼ばれる。

では 2.4 節のテスト事例で、2 母数モデルによるテスト情報関数を描いてみよう。

図 18 に描いたのがテスト情報関数である。縦軸には情報量の平方根 $\sqrt{I(\theta)}$、つまり標準誤差をとった。これを見ると、$\theta = 0$ より小さい値の時に関数が Max になる。つまり、平均が $\theta = 0$ より少し低い被験者にこのテストを実施した方が、テストの精度が高いことがわかる。

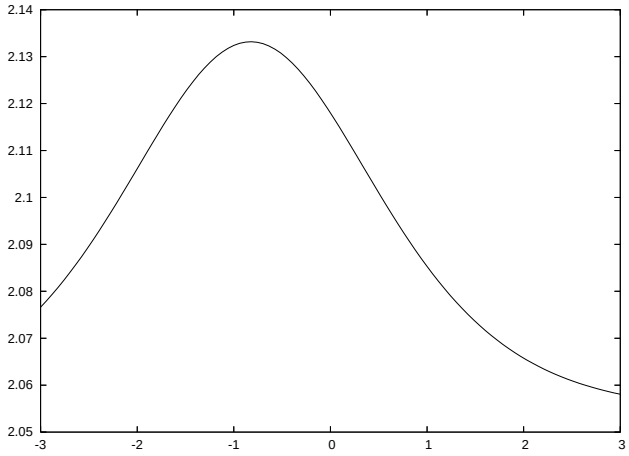

図 18　テスト情報関数

さて次に、テストの平均点予測などに用いられる、テスト特性曲線 (Test Characteristic Curve) を紹介しておこう。ある推定値 $\hat{\theta}$ の人が、何度もテストを受けると、平均 θ、分散 $1/I(\theta)$ である程度散らばることがわかった。では、推定値ではなくて θ_i とはっきりわかっている場合は、テスト得点の平均値 (=期待値) はいくつになるか、というと

$$E[y_i \mid \theta_i] = \sum_{j=1}^{n} w_j p_j(\theta_i)$$

である。ここで w_j は項目 j における重み (配点) である。この式で θ_i を固定せずに変数とすると、この式はテストの平均点の推移が現れる関数となる。この関数に描かれる曲線を、テスト特性曲線と呼ぶ。

2.4 節のテスト事例の、テスト特性曲線は図 19 のようになる。

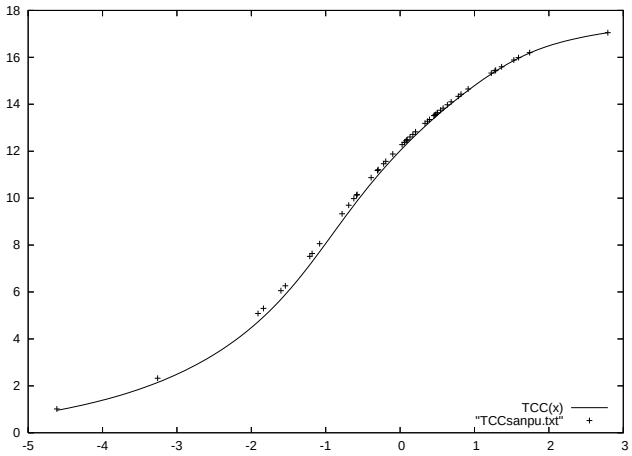

図 19　テスト特性曲線

図 19 には、あわせて推定された $\hat{\theta}_i$ の散布図も描いてある。この事例では N の数が少ないのでほとんど TCC 曲線からの解離はないが、推定値の分散が大きくなるところは尺度の精度が悪くなるところである。

さて、ここまでは、項目反応理論によってテストを開発するものとして話を進めてきたが、心理学などの分野では尺度開発法としてこれを用いたいと思う方も多いだろう。次のセクションでは、尺度構成法としての IRT の使い方を紹介する。

3　項目反応理論を使った尺度作り

心理学などの分野で、尺度として IRT の技術を応用する場合、まず問題になるのが尺度水準である。今までの例のように、正答・誤答で 0/1 のデータをするのではなく、できたら「非常にそう思う」「そう思う」「どちらともいえない」「そう思わない」「全くそう思わない」などの数段階の反応を得たい、と考えるだろう。これらの尺度が間隔尺度水準で得られた、といえるのは実は五件法でギリギリ、七件法でやっと、というぐらいである。理論的には、順序尺度として扱うのが望ましい。そこで IRT では、順序尺度データの分析に段階反応モデルを導入する。

3.1　段階反応モデル

IRT の段階反応モデルは、一つひとつのカテゴリ (「全くそう思わない」から「非常にそう思う」までの、各点) に 2 母数ロジスティックモデルを当てはめる。このとき、識別力母数 a_j は等しいと仮定する。これは段階的反応を記述するためにする制限で、段階ごとの違いは困難度母数 b_j のほうで区別する。項目 j に c と答えるときの関数用困難度を b_{jc} と表す。このような 2 母数ロジスティックモデル、

$$p^*_{jc}(\theta) = \frac{1}{1 + exp(-1.7(\theta - b^*_{jc}))}$$

で描かれる例のカーブは、項目 j について、母数 θ の人が c 以上の反応、すなわち $u_j \geq c$ をする確率、であるとする。これは境界特性曲線 (Boundary Charactaristic Curve) と呼ばれる。ここから、尺度値 θ の人が、$u_j = c$ とする確率 $p(u_j = c \mid \theta)$ を以下のように表現する。

$$p(u_j = c \mid \theta) = p^*_{jc}(\theta) - p^*_{jc+1}(\theta) \tag{5}$$

何でこんなコトになるのか、このままではイメージしにくいので、例を使おう。先ほどの四件法、「全くそう思わない」「そう思わない」「そう思う」「非常にそう思う」を考えよう。順に項目値は $c = 0, 1, 2, 3$ である。また、仮に b^*_{jc} をそれぞれ、$b^*_{j1} = -1, b^*_{j2} = 0.5, b^*_{j3} = 1.2$ とする。

まず、$p(u_j = 0 \mid \theta) = p^*_{j0}(\theta) - p^*_{j1}(\theta)$ を考える。これは日本語に書き下すと、項目 j について、「全くそう思わない」と回答する確率は、「全くそう思わない」以上の回答をする確率から「そう思う」以上の回答をする確率を引いたもの、となる。ところで、“「全くそう思わない」以上の回答” とはつまり 0 以上だから、$u_j = 0, 1, 2, 3$ の確率である。つまりあらゆる確率、という意味な

ので $p^*_{j0} = 1$ である[*9]。次に、「そう思わない」以上の回答、つまり $1, 2, 3$ を選択する確率なので、これはロジスティックモデルを使って算出すればよろしい。つまり、

$$p(u_j = 0 \mid \theta) = 1 - \frac{1}{1 + exp(-1.7(\theta - b^*_{j1}))}$$

である。この関数を図にしたのが図 20 である。

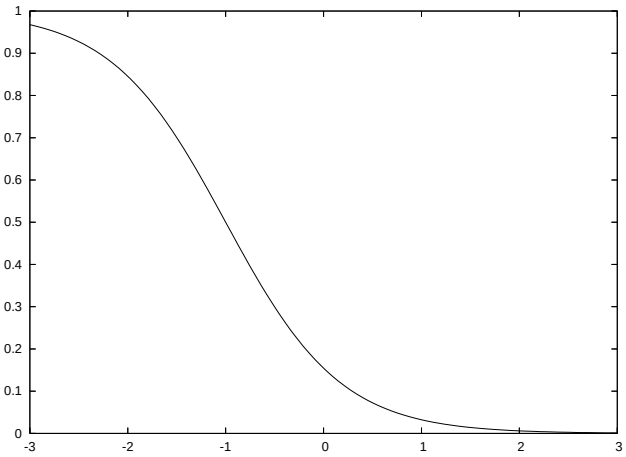

図 20　「全くそう思わない」と回答する確率の推移

これを見ると、θ が低ければ低いほど「全くそう思わない」と回答する確率は高く、θ が上がるに連れてロジスティックの逆カーブでその確率が減っていくことがわかる。なるほど、これだとイメージは難しくない。

次に、逆の極端「非常にそう思う」と回答する確率を考えよう。つまり $p(u_j = 3 \mid \theta) = p^*_{j3}(\theta) - p^*_{j4}(\theta)$ を考えるのである。しかし、p^*_{j4} の値とは一体なんだろう。「非常にそう思う」が 3 なので、4 というのは用意した段階カテゴリ以上のものである。こんな確率はあり得ない。ので、$p^*_{j4} = 0$ とおいて良い。そうすると、「非常にそう思う」と回答する確率は、

$$p(u_j = 3 \mid \theta) = \frac{1}{1 + exp(-1.7(\theta - b^*_{j3}))}$$

であり、図 21 のようになる。これは、θ が上がるに連れてその回答が出現する確率が増えて行く関数だから、なるほどと理解できる。

それでは中間の値、「そう思わない」や「そう思う」はどうなるだろうか。これは引き算する二つの項目が 0 や 1 にならず、関数のままなので、

$$p(u_j = 1 \mid \theta) = \frac{1}{1 + exp(-1.7(\theta - b^*_{j1}))} - \frac{1}{1 + exp(-1.7(\theta - b^*_{j2}))}$$

[*9] 当然これは、何件法になっても変わらない。

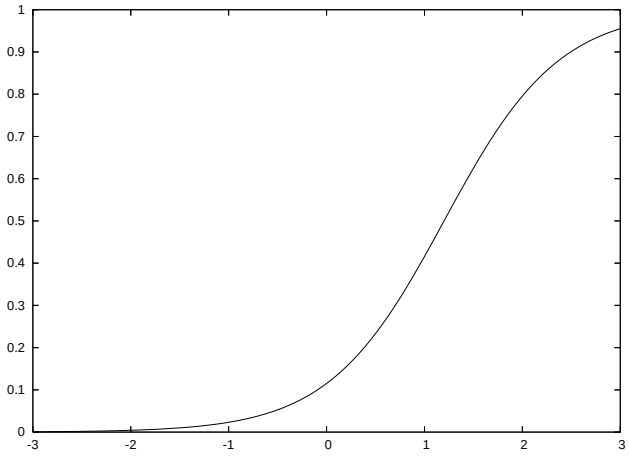

図 21　「非常にそう思う」と回答する確率の推移

$$p(u_j = 2 \mid \theta) = \frac{1}{1 + exp(-1.7(\theta - b^*_{j2}))} - \frac{1}{1 + exp(-1.7(\theta - b^*_{j3}))}$$

と表すしかない。グラフにすると図 22 のようである。

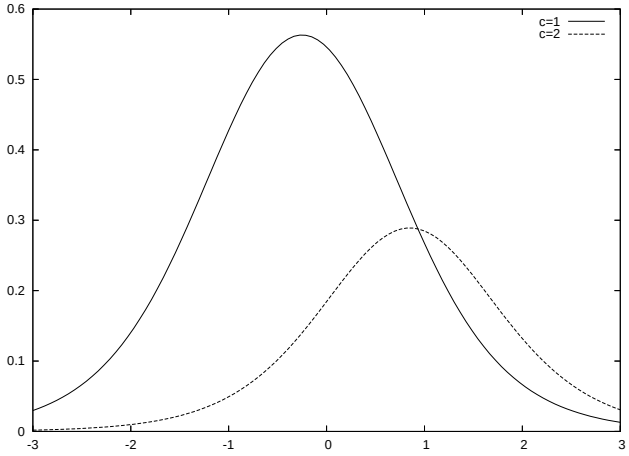

図 22　「そう思わない ($c = 1$)」と「そう思う ($c = 2$)」と回答する確率の推移

見慣れた ICC と形が変わるので、ちょっと意外かもしれないが、この釣り鐘型のカーブが途中の段階に反応する確率として得られる。

さて、それでは次に、どのようにして b^*_{jc} を導出するか考えよう。

仮に被験者のカテゴリに対する反応が正規分布に従っているとしよう (リッカート法と同じ前提)。ここで、被験者の尺度値 θ は連続変量だが、顕在化する被験者の反応としては (例えば四段階の)「そう思わない」になる。さて、ここで四つのカテゴリに分別する閾値 τ を考えよう。「全くそう思わない」と「そう思わない」を分けるのは、$\tau_1 = -1.1$ ぐらい、同じく「そう思わない」

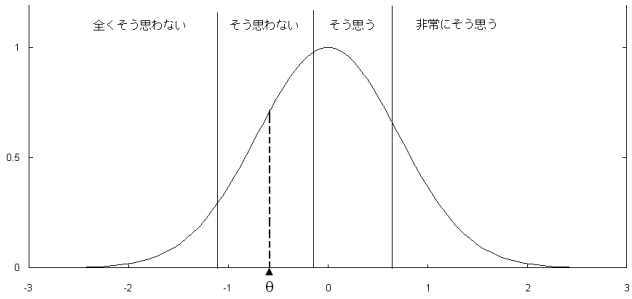

図 23　被験者の尺度値 θ とカテゴリ反応との対応

と「そう思う」を分ける $\tau_2 = -0.2$ ぐらい、といったように設定できるだろう。このような項目 j についての標準正規分布 z_j が、被験者の尺度値 θ と項目との関係で、

$$z_j = \alpha_j \theta + e_k$$

と表されたとき、これを 1 因子カテゴリカル因子分析モデルという。この 1 因子カテゴリカル因子分析モデルは、実は IRT の段階反応モデルと非常に相性がよい。

ここから因子負荷量 α_j と閾値 τ_{jc}(項目 j で c と反応する閾値) を推定し、その値から

$$a_j = \alpha_j / \sqrt{1 - \alpha_j^2}$$

$$b_{jc}^{*} = \tau_{jc} / \alpha_j$$

が算出できる。因子得点 f は被験者の尺度値 θ と等しく、この変換を用いると尺度値 $f = \theta$ の被験者がテスト項目 u_j に c と反応する確率は式 5 と一致する。

つまり、段階反応モデルを実際にやってみよう、という場合は SEM でカテゴリカル因子分析モデルを構成する必要がある。カテゴリカル因子分析モデルを扱えるソフトは少なく、LISREL でもがんばったら作れるみたいだし、Mplus というソフトも結構いいらしいが、ユーザーとしては IRT 専門のソフトが手軽に手にはいるようになって欲しいものである。