項目特性図作成方法の精緻化による項目分析の新たな展開

秋山隆

目 次

第1章　問題と目的

第 1 章では，項目分析 (item analysis) について概説した後に，本稿を通じて扱っていく項目特性図 (item characteristic chart) を用いた項目分析の方法について述べる。

1.1　項目分析

項目分析はテストを構成し，実施，運用するために欠かすことのできない作業工程である。項目分析の重要性は池田 (1992)，植野・荘島 (2010)，豊田 (2012) においても指摘されている。また，項目の正答比率や，選択肢の振る舞いの分析の重要性は，Schmeiser, C. B., & Welch, C. J (2006, p.329, p.340) において指摘されている。テストの準備，作成段階において，項目分析を行うことで，テストを実施する運営者は，当該項目が測るべき事柄を適切に，どれだけの精度で測定できているのかを確認することが可能となる。この確認作業を通じて，項目をふるいにかけ，実際に運用するテストに含めるに足る項目を見定めていくのである。

項目の検討を経ずに構成したテストをそのまま運用することは，そのテストが信頼できるものであるのか，テストが測定している対象が妥当であるのかといった事柄に懸念を残すこととなる。

1.1.1　項目の形式

テストは原則的には複数の項目（質問紙調査における質問や，学力試験における問題）から構成される。

論述形式

論述形式項目は少数大課題設定方式（池田, 1992）において採用される項目形式である。論述形式項目では，受験者は問われた課題内容について,（制限時間以内に）文章を構成し，解答することが求められる。本形式の項目を採用したテストでは，出題項目が一つのみとなる場合もある。

受験者に，課題に関連した論述を求めることは，受験者の知識の程度や正確さ，更に当該知識を論理的に説明する能力を有しているかを測ることが可能であるとされている。しかしながら，測定対象となる受験者の能力を適切に測ることができるのかについては意見の一致は見られていない。例えば池田 (1992, p.22) は，時間制限があるために受験者の最大限の解答が得られるとは限らないこと，課題の選定が結果に及ぼす影響の大きさ，評価基準の設定や評価者によって評点が変化しうること，の三つの問題点を指摘している。

多肢選択式項目

多肢選択式項目 (multiple choice item) は，主として細目積み上げ方式（池田, 1992）のテストにおいて採用される項目形式である。多肢選択式項目は，出題内容に関する正誤判断や，あらかじめ用意された複数の選択肢の内から正答となる選択肢を回答することを受験者に対して求める。多肢選択式項目は例えば

例：**5** 肢選択形式の項目例

問

ドストエフスキーが 1880 年に刊行した小説作品を示す選択肢として適切なものを以下から選べ。

A アンナ・カレーニナ

B カラマーゾフの兄弟

C 外套

D スペードの女王

E かもめ

のように受験者に呈示される。なお正答は「B：カラマーゾフの兄弟」である。「A」の作者はトルストイ,「C」はゴーゴリ,「D」はプーシキン,「E」は戯曲であり，チェーホフの作品である。

選択肢数は任意に定めることが可能である。最小選択肢数は 2 肢であり，出題内容の正誤のみを回答者に対して問う場合が該当する。選択肢数の上限は存在しないが，5 肢までとすることなどが一般的である。本形式の解答方法として，選択肢に割り振られた記号を答えさせる場合や，マークシート方式が採用される。

多肢選択式項目の場合は，論述形式と比較すると，各項目で受験者に要求される解答時間は短くなる。また，各出題項目が被覆する領域は非常に限定されたものとなる。そのため，一般に細目積み上げ方式を採用したテストは測定対象となる領域を被覆できるように十分とされる数の項目群によって構成されることが望ましい。

この他，短答形式として，文章中の任意の語句が空欄となっており，適切な語句を簡潔に記述，もしくは複数選択肢中から順に選択する形式も存在する。テストの目的に応じて，これらの項目形式を組み合わせて一つのテストを構成する場合もある。これ以降，本稿においては多肢選択式の項目を扱うこととする。

1.1.2 項目反応の整理

構成したテストを実施した結果，得られた受験者の項目に対する反応（解答）を項目反応 (item response) という。得られた項目反応を元に採点，分析を行うため，まずは項目反応の整理を行うこととなる。ここでは数理的に整理する方法について概説する。

任意のテストにおける受験者は添え字記号 i を用いて表すこととする。i は 1 から N までの範囲をとる。また，テストを構成する項目数は添え字記号 j を用いて表現される。j は，1 から J までの範囲をとる。項目ごとに受験者数が異なりうる場合には，N_j と表現する。

任意の j 番目の多肢選択式項目の選択肢をカテゴリと呼び，添え字記号 c_j を用いて表す。c_j は 1 から C_j までの範囲で変化する。c_j は項目ごとに選択肢数（カテゴリ数）が異なりうる場合にも用いることができるが，全項目にわたり，カテゴリ数が同じである場合には，項目を表す添え字記号を省略し，c として表す。

例えば，第 1 問目が 4 肢選択式の項目であるならば，$C_1 = 4$ と表される。また，第 2 問目が 6 肢選択式の場合には $C_2 = 6$ である。ただし，出題形式が 3 肢以上の多肢選択式であったとしても，受験者の反応の正誤にのみ関心がある場合には，受験者が誤答選択肢のいずれを選んだかを区別せずに，すべての項目においてカテゴリ数を $C_j = 2$ の 2 値として整理する。

表 1.1 は架空のテスト受験結果について，選択肢コードを整理した状態を表している。行に受験者を配し，列に項目を配している。第 2 行目は正答選択肢のコードを示している。ここでは例 1 のような 5 肢選択式項目を想定しているため，選択肢コードは A,B,C,D,E の 5 つとなっている。何れの選択肢が選ばれたのか以外に，無効解答をもコード化する場合には，F 以降の適当なコードを割り当てればよい。例えばここでは 1 番目の受験者の 10 項目目の解答が無効解答となる。

表 1.2 は表 1.1 の選択肢のコードに対して，A=1 から順に，E=5 までの数値

表 1.1 架空テスト生データ

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10
正答	A	B	B	C	D	E	D	E	A	B
1	A	B	B	C	D	E	D	E	A	F
2	B	B	B	C	A	E	C	D	A	B
3	B	B	B	E	A	E	C	D	B	B
4	B	B	B	C	A	E	C	E	A	A
5	A	B	C	D	C	D	B	A	A	A
6	B	B	C	C	D	E	D	E	A	A

表 1.2 数値コード化済架空テスト生データ

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10
正答	1	2	2	3	4	5	4	5	1	2
1	1	2	2	3	4	5	4	5	1	6
2	2	2	2	3	1	5	3	4	1	2
3	2	2	2	5	1	5	3	4	2	2
4	2	2	2	3	1	5	3	5	1	1
5	1	2	3	4	3	4	2	1	1	1
6	2	2	3	3	4	5	4	5	1	1

コードを割り当て，整理し直した状態を示している。この場合も無効解答は 1 から 5 以外の適当な数値コードを割り当てればよい。

表 1.3 は表 1.1 もしくは表 1.2 を元に，正答か誤答かを区別する 2 値のみでコード化した場合の受験結果整理例である。本例の場合には，誤答=0，正答=1 としてコード化している。2 値データとして扱う場合には，一般に無効解答も誤答として整理されることとなる。表 1.3 から表 1.1 や表 1.2 を復元することはできないため，2 値データ化後も，生データは保持しておくことが望ましい。

テスト得点

ここで，表 1.3 の受験者の反応パタンをベクトル表記を用いて $\boldsymbol{x}_i$ と表現する。例えば

$$\boldsymbol{x}_1 = (1, 1, 1, 1, 1, 1, 1, 1, 1, 0)'$$

表 1.3 2 値化された架空テストデータ

$i \diagdown j$	1	2	3	4	5	6	7	8	9	10
正答	1	2	2	3	4	5	4	5	1	2
1	1	1	1	1	1	1	1	1	1	0
2	0	1	1	1	0	1	0	0	1	1
3	0	1	1	0	0	1	0	0	0	1
4	0	1	1	1	0	1	0	1	1	0
5	1	1	0	0	0	0	0	0	1	0
6	0	1	0	1	1	1	1	1	1	0

である。テスト全体の反応パタンは行列表記を用いて

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]' \tag{1.1}$$

と表される。

表 1.3 の行に関して，

$$x_i = \sum_{j=1}^{J} w_j x_{ij} \tag{1.2}$$

を計算することで，受験者個人のテスト得点（和得点）を得ることができる。w_j は項目ごとに定められた配点の重みであり，x_{ij} は i 番目の受験者の j 番目の項目における反応を表している。w_j はテストの採点者が任意に定めることが可能であるものの，一般的には $w_j = 1$ とされ，その場合には w_j を省略して表記することもある。例えば $w_j = 1$ とした場合の 1 行目の受験者のテスト得点は $x_1 = \sum_{j=1}^{10} x_{1j} = (1+1+1+1+1+1+1+1+1+0) = 9$ となる。

1.1.3 選択肢の選択率

多肢選択式項目の受験結果に関しては，その選択肢ごとの選択数によって，どの程度多くの受験者に選ばれたのか，人数の多寡を確かめることができる。項目選択肢の選択率は

$$p_{jc} = \frac{n_{jc}}{N_j} \tag{1.3}$$

として表される。ここで N_j は第 j 番目の項目における受験人数，n_{jc} は項目 j の c 番目のカテゴリにおける，選択人数を表している。カテゴリごとの選択率

p_{jc} を参照することで，正答選択肢以外に選ばれやすい誤答選択肢を確認することができる。

選択率を参照する際には，受験者を任意の数の群に分けることが，更なる詳細な分析を行うために便利である。群に分ける際には，受験者をテスト得点順に並べ，低得点群から高得点群となるように分割する。この際，低得点群，中得点群，高得点群となるように受験者を 3 群に分割する場合や，更に分割数を増やし，5 群とする等，群の数は任意に定めることが可能である。一般的には 5 群に分割されることが多い。群ごとの選択率は

$$p_{jgc} = \frac{n_{jgc}}{N_{jg}} \tag{1.4}$$

と表される。添え字 $g(= 1, 2, \ldots, g, \ldots, G)$ は，g 番目の群を表している。つまり，n_{jgc} は j 番目の項目の，g 番目の群における，カテゴリ c の選択人数を示している。選択率 p_{jgc} を用いた項目分析を行うことで，受験者の得点（能力）群ごとの選択傾向を検討することが可能となる。

1.1.4 選択率を用いた項目分析

表 1.4 は架空の文学史のテスト内に含まれる多肢選択式の項目における，各選択肢の選択数の整理例を示した。行に選択肢を，列に受験者について，得点で分けた群を配している。ここでは受験者を低得点群としての第 1 群から，高得点群を表す第 5 群までの 5 群に分割している。受験者は 250 人おり，各群の人数が等しくなるように分割している。表の各セルは選択人数が格納されている。

表 1.4 例題における各選択肢の被選択数

選択肢＼群	1	2	3	4	5
A	20	15	12	11	5
B	7	16	20	27	41
C	10	8	9	6	3
D	7	8	7	4	1
E	6	3	2	2	0
群人数	50	50	50	50	50

表 1.5 は表 1.4 を群における選択率として再表現したものである。項目反応を整理し，表 1.4 や表 1.5 を作成することで，それぞれの群で，どの選択肢が選択されやすかったのかを値の大小から分析することが容易となる。

ここでは選択肢 “B” が正答選択肢となっている。表 1.5 を観察すると，第 1 群から群が大きくなるにつれて，選択率も上昇している様子を見て取ることがで

表 1.5 例題における各選択肢の被選択率

選択肢＼群	1	2	3	4	5
A	0.40	0.30	0.24	0.22	0.10
B	0.14	0.32	0.40	0.54	0.82
C	0.20	0.16	0.18	0.12	0.06
D	0.14	0.16	0.14	0.08	0.02
E	0.12	0.06	0.04	0.04	0.00
確率和	1.00	1.00	1.00	1.00	1.00

きる。このことはテスト得点が高くなるほど，正答率も高くなるという関係を表しており，当該項目が測定したい対象を測定できていることを示唆する傍証となる。また，いずれの誤答選択肢も，群が大きくなるにつれて，選択率が減少しており，テスト得点の高い受験者ほど，誤答選択肢に惑わされる率が少ないことを意味している。このこともまた，項目の適切性を示唆するものである。誤答選択肢の中でも，特に魅力のある選択肢として “A” が挙げられる。この特徴に注目し，出題文と特に魅力的であった誤答選択肢について誤答分析を行うことで，誤答者への教科教育を再検討することが可能となる。

項目識別力

識別力（項目テスト相関）

項目の測定性能を確認するために利用される数値的な道具として，項目識別力（item discrimination; 豊田, 2012）がある。項目識別力はテスト得点と，当該項目における項目得点とのピアソンの積率相関係数によって定義される。項目識別力は項目テスト相関 (item-total correlation) とも呼ばれる（植野・荘島, 2010）。

識別力は -1.0 から 1.0 の範囲をとり，識別力が高い項目，つまり 1.0 に近い項目ほど，当該項目が含まれるテストが測定対象としている特性を，項目が識別することができていると判断できる。一方で，項目識別力が低く，0 に近い場合は，対象とする特性を，項目得点によって識別できていないと判断を下すこととなる。

なお，識別力が負となる項目は，項目得点（あるいは正答率）が下がるほど，テスト得点が高くなる傾向にある項目であることを表している。そのため，学力試験等の項目分析においては，特別な理由がない限り，識別力が負となる項目が観察されることは少ない。また，例え負の識別力が観察されたとしても，

一般には識別力が負となる項目を，そのままテスト内に残しておくことはなく，項目内容の改訂，あるいは削除の検討対象となる。

項目残余相関

識別力はテスト得点と項目得点の相関係数であるが，テスト得点それ自体に，項目分析の対象としている当該項目得点が含まれているという性質がある。この場合，項目数が少ない場合や，配点が大きい場合には識別力が高くなる傾向にある（豊田, 2012, p.11）。こうした特徴を回避するために，テスト得点から，現在分析対象としている項目の項目得点を差し引いた残余テスト得点を計算し，その残余テスト得点と，当該項目得点との相関係数を計算した値を識別力として利用する場合もある。

当該項目も含む，テスト得点を用いる場合の識別力と区別するために，この場合の識別力を項目残余相関，あるいは項目リメインダ相関（item-remainder correlation; 植野・荘島, 2010）と呼ぶこともある。

項目残余相関もまた，−1.0 から 1.0 の範囲をとり，解釈も識別力と同様に行うことが可能である。項目残余相関が高い（1.0 に近い）項目ほど，当該項目に正答した受験者のテスト得点が高くなる傾向にあることを示している。ただし，項目識別力とは異なる点として，“分析対象項目を除いたときの”テスト得点との相関が高いことを示している点に注意が必要である。項目残余相関が負となる項目もまた，識別力の場合と同様に，一般的にはテストにふさわしい項目とは見なされない。

1.2 項目特性図を用いた項目分析

テストを構成している設問項目の性質（測定性能）を評価する際に，当該設問項目における受験者の正答率（通過率）を用いて項目の難しさを調べる方法がある（Millman & Greene, 1989）。そのための有用な道具として項目特性図（item characteristic chart, 豊田, 2012）を挙げることができる。項目特性図は設問解答率分析図（菊地, 1999）もしくは設問回答率分析図（吉村, 2009），層別解答率図（大津, 2006）といった呼称の下で，広く利用されている。

1.2.1 項目特性図の作成方法

本項で項目特性図の作図方法を述べる。まずテストの受験結果を整理し，受験者のテスト得点を算出する。算出したテスト得点を昇順に並べ，それらを任意の G 個の群 (赤根・伊藤・林・椎名・大澤・柳井・田栗（2006）では $G = 5$ としている）に分割する。次に各群の，各項目の正答率を計算する。最後に群を

横軸に，確率を縦軸にとったグラフに正答率をプロットし，各プロットを直線で結ぶ。これを項目特性曲線[1]と呼ぶ。

誤答選択肢についても興味がある場合は，更に誤答選択肢の選択率についても，正答率と同様にプロットし，プロット間を線で結ぶ。なお，項目特性図作図における受験者の群への分割方法では，各群の所属人数を等人数とする場合と，各群間の得点間隔が等しくなるように受験者を振り分ける場合の 2 通りが主要な方法として存在する（豊田, 2012, p.4）。

1.2.2 項目特性図の解釈例

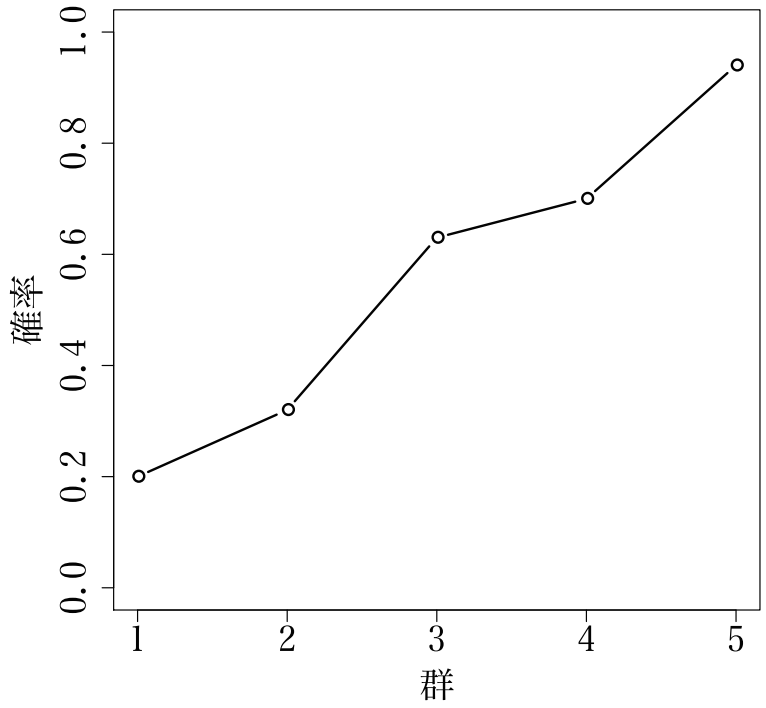


図 1.1 2 値で整理した場合の項目特性図の例

図 1.1 に項目反応を 2 値データとして整理した場合の項目特性図を例示した。群内における全選択肢の選択率の和が 1 となる制約から，2 値反応の場合の誤答選択肢に関する反応確率は，正答選択肢の項目特性曲線と対称の関係となるため，描画が省略されることが一般的である。図 1.1 では，テスト得点最下位群においては 2 割程度が正答できており，より上位の群になるにつれ正答率が上昇し，最上位群においてはほとんどの受験者が正答していたことが，また第 3 群から第 4 群にかけては他の群間に比べて正答率のグラフの上昇が鈍化傾向にあることがわかる。

[1]項目特性曲線はトレースライン (trace line) とも呼ばれる。

項目特性図の分類

前述のように，項目特性図は項目分析を行う際の道具として用いられる。その際，表現される項目特性の違いに応じて，項目特性図を分類することが，項目の取捨選択，改善を行っていく上で有効である。ここでは，図 1.2 に豊田（2012）に基づいた項目特性図の分類型を示した。

項目特性図は識別力の高い項目として G (General) 型, L (Low rank) 型，H (High rank) 型，識別力の低い項目として E (Easy) 型， M (Middle) 型， D (Difficult) 型，B (Bad) 型にそれぞれ分類可能である。

ここで識別力の高い項目とは，テスト得点下位群よりも上位群の正答確率が明らかに高く，折れ線グラフが右上がりの形状を示す項目を指している。反対に，識別力が低い項目とは，群によらず正答率が比較的一定であり，折れ線が横に這っているような形状を示す項目を指している（豊田, 2012）。

G 型項目はテスト得点低位群から上位群の全群にわたって測りたい特性が比較的精度よく識別されている項目であることを示している。当該項目が良質な項目であることを意味しており，通常はこのような項目がテストに多く含まれていることが望ましい。

L 型項目は下位群において特性がよく識別されている項目である。上位群の特性を識別するには力不足ではあるが，下位群の識別を目的としたテストの場合には L 型項目が必要となる。

一方で H 型項目は上位群の特性がよく識別される項目である。選抜試験など，より高い特性を有する受験者を識別したい場合には H 型項目がテストに含まれているとよい。

E 型項目は下位群も上位群も正答率が同様に高く，特性の識別自体にはあまり寄与していない。ただし基礎事項の確認として E 型項目が必要となる場合もある。

M 型項目は下位群から上位群まで正答率が中間を漂っており，受験者が低特性か高特性かに関係なく偶然正答するような項目であり，無駄な設問となってしまっている。

D 型項目は全体的に正答率が低い位置に張り付いており，受験者の多くが誤答する項目となっている。特性の識別には不適当かつ高特性者が時間を空費するような項目であり，テスト全体に悪影響を与える場合がある。

B 型項目は低位から高位群にかけて正答率が下がってしまうような項目であり，測定対象特性とは無関係であったり，設問の誤りが疑われる。

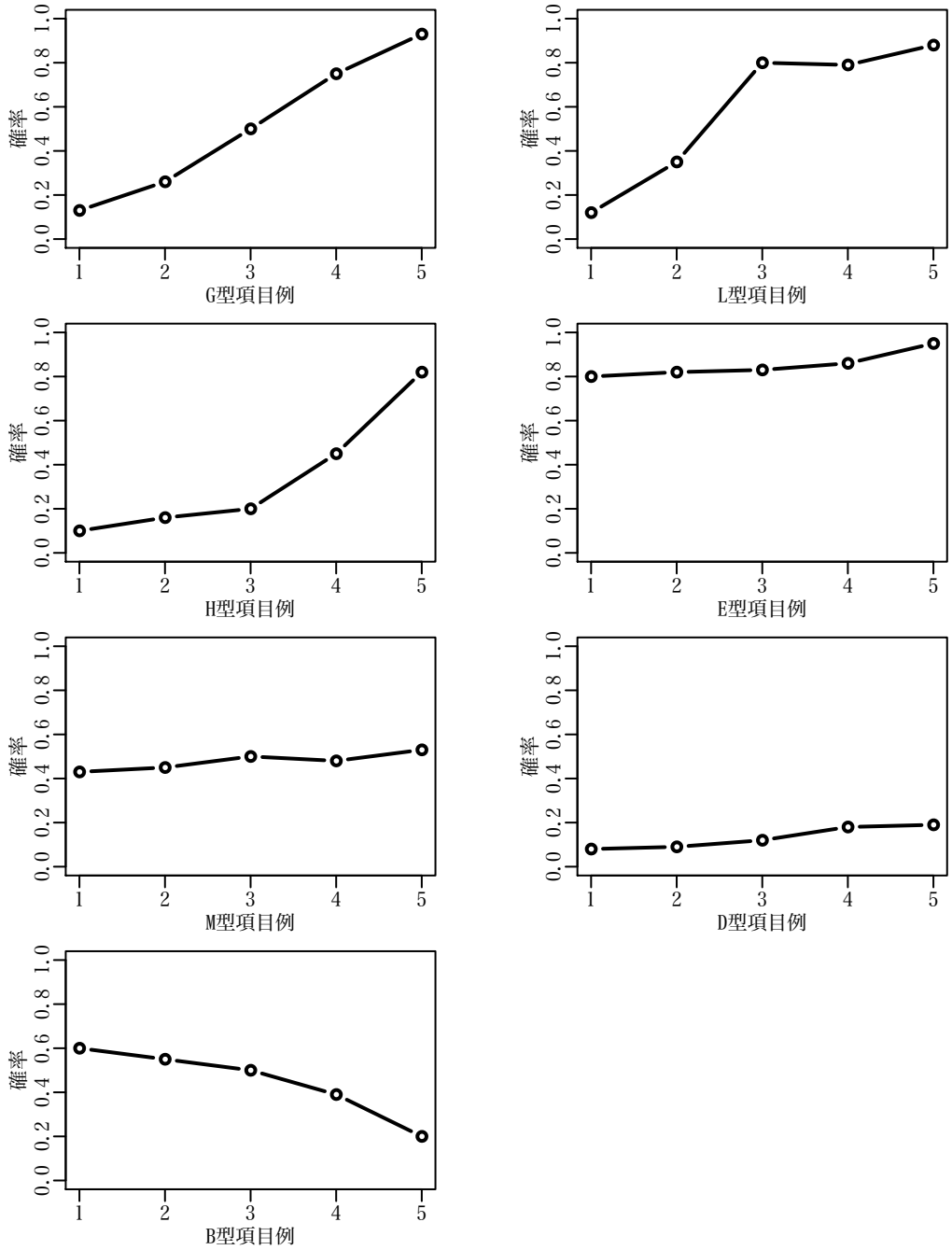


図 1.2 項目特性の分類例

1.2.3　正答分析

項目特性図の基本的な利用方法として，正答分析に用いる方法が挙げられる。項目特性図を観察することで，群ごとに正答選択肢の選択傾向を確認することが可能となる。テスト得点が高い受験者群になるほど，正答選択肢の選択率が高くならなければ，当該テスト項目は何らかの問題を有しているものと解釈することが可能である。

正答分析における項目特性図の解釈は，前掲の項目特性図の正答傾向の分類に従って行えばよい。

1.2.4　誤答分析

項目特性図は正答分析のみならず，誤答分析に用いることも可能である。この場合には，正答選択肢以外の選択肢，すなわち誤答選択肢の選択率もプロットし，プロット間を直線でつなげばよい。正答分析を行った後に，誤答分析へと分析を進めることで，どの得点群の受験者がいかなる誤答選択肢に魅力を感じ，誤答してしまったのかを確認することが可能となる。詳しくは第 3 章で述べる。

1.3　項目特性図作成時の問題点

項目特性図は非常に有用な項目分析のための道具ではあるが，その作成方法については，注意すべき点が大別して三つ存在する。

1.3.1　群数決定方法の恣意性

項目特性図の作成時に受験者を群に分ける際，群数 G の値を決定するための明確な基準は知られていない。現状では分析者によって経験的に決定されており，一般的には $G = 5$ とされるが，その根拠は明確には示されていない。

1.3.2　項目特性図の視覚的特徴と解釈の検証可能性

項目特性図における誤答選択肢の表現は，その性質上，選択率が似たような傾向を示しやすい。そのために，プロット間の直線が重なり，項目特性図の長所の一つである，選択率の視認性が低下してしまうこととなる。

このような項目特性図の難視性を回避するためには，例えば任意の基準を設定し，その基準を下回った誤答選択肢に関しては誤答分析の対象から外すなど，項目特性図における描画する誤答選択肢の数を減らすという対応が考えられる。

しかしながら，項目特性図の視認性を確保する際の基準は広く知られていない。

誤答選択肢の惑わしの機能が類似しているために，項目特性図における誤答曲線[2]の類似性が推測される場合には，こうした仮説に基づいて，誤答曲線をまとめることが考えられる。あるいは，学習の特定の段階において誤用される方略，知識によって導かれる誤答機能に焦点を絞るために，当該選択肢以外の部分をまとめたい場合も想定可能である。

ただし，作成した（オリジナルの）項目特性図を観察し，観察結果から構成した仮説（誤答選択肢の機能の類似性や特異性）に基づいて複数の誤答選択肢をまとめることは可能であっても，仮説を反映して作成した項目特性図を検証する方法は未だ確立されていない。

1.3.3 同一項目に関する重複した項目特性図の解釈の恣意性

テストの実施方法の一つとして，ブックレット（小冊子）方式による項目の出題が挙げられる。ブックレット方式とは，テストが大規模なために同一日時，場所でテストを実施することが困難であり，複数の地域で，複数の受験可能日を設定している場合に有用なテスト実施方法である。

同じテストを数回に渡り実施することは，問題項目の漏洩や，同じ受験者がまったく同じテストを受験することに繋がり，結果的に望ましくない影響が生じる恐れがある。この場合には，項目は（場合によっては項目数も）異なるが，同じ特性を測定できるようなテストの版（ブックレット）をいくつか作成し，テストの実施ごとにブックレットを変更することで対処する。つまり，各受験者集団に対して異なるブックレットが割り当てられる。

このとき，異なるブックレット間であっても，後の分析においてテストや受験者の特性が比較可能となるように，ブックレット間でいくつかの共通する項目（共通項目）を含ませておく。

基礎的な項目分析はブックレットごとに行われるため，共通項目については含まれるブックレットの数だけ項目特性図を作成する[3]ことが可能となる。

共通項目についての項目特性図は，ブックレット間の特性曲線（トレースライン）の表現が一貫して全く同じ傾向を示していることが望ましい。しかしながら，特性曲線は必ずしも同一表現とはならず，ブックレット間で異なり得る。これは，ブックレットごとに共通項目以外の項目や，項目数，そして受験者が異なるという要素の影響を受けるためである。

このとき，当該共通項目の項目特性を表現する項目特性図として，何れを採用して項目分析を行うべきか，基準は広く知られておらず，分析者が任意の一つの図に注目して分析が行われている。

---

[2]正答選択肢の特性曲線を正答曲線，誤答選択肢の特性曲線を誤答曲線と呼称する。

[3]和得点としてブックレット得点を用いる。

1.4　項目特性図の作成方法における基準の導入

本論文では，前節における項目特性図の恣意性を取り除くための統計的基準の利用方法を提案する。そこで，統計モデルを選択するために用いられる，情報量規準 (information criterion) と呼ばれるモデル評価規準を用いることとする。情報量規準を用いた選択を行うことで，群数選択基準や，項目特性図の統合基準，同一項目における複数の項目特性図の選択といった判断をモデル選択の観点から行うことできるようになる。まず本節では情報量規準について概説する。

1.4.1　赤池情報量規準

我々はある観測データが得られた場合に，そのデータがどのような構造の下で発生したのか，ということを考える。しかしながら，多くの分野において，真の構造は未知である。データが得られたモデルを完全に再現しようとすると，多くのパラメタを想定しなければならなくなる。しかしそうして構成したモデルは，手元のデータに対してのみ，有効なモデルとなってしまう。赤池情報量規準を用いることで，手元のデータにおいて正しいモデルよりも，より予測精度のよいモデルを選ぶ（例えば赤池・甘利・北川・樺島・下平, 2007, p.62）ことが可能となる。

1.4.2　カルバック–ライブラー情報量

本項で，カルバック–ライブラー情報量 (Kullback–Leibler divergence; Kullback & Leibler, 1951) について，小西・北川 (2004) の記述に従って概説する。いま，n 個のデータ $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ が確率分布関数 $G(x)$ から発生したものとする。ここでは $G(x)$ を真のモデル（分布，あるいは構造）と呼ぶこととする。また，$G(x)$ 自体は未知であるものとする。一方，研究者が想定する（提案する）モデルを $F(x)$ とする。確率分布関数 $G(x)$ と $F(x)$ が，それぞれ密度関数 $g(x)$ と $f(x)$ をもつ場合は，連続モデルという。一方 $g(x)$ と $f(x)$ が有限もしくは可算無限個の離散点 $\{x_1, x_2, \cdots, x_k, \cdots\}$ に対して，以下のように事象 $\{\omega; X(\omega) = x_i\}$ の確率

$$
\begin{aligned}
g_i =& g(x_i) \equiv \Pr(\{\omega; X(\omega) = x_i\}), \\
f_i =& f(x_i) \equiv \Pr(\{\omega; X(\omega) = x_i\}),\ \ i = 1, 2, \cdots
\end{aligned}
$$

で表される場合は，離散モデルという。

Akaike (1973) は，真のモデル $g(x)$ と提案モデル $f(x)$ の確率分布としての近さをカルバック–ライブラー情報量

$$
I(G; F) = E_G \left[\log \frac{G(X)}{F(X)} \right] \tag{1.5}
$$

を用いて算出し、モデルの相対的なよさとして評価することを提案した。カルバック–ライブラー情報量は連続型モデルの場合，

$$I(g;f) = \int_{-\infty}^{\infty} \log\left(\frac{g(x)}{f(x)}\right) g(x)d(x) \tag{1.6}$$

と表され、離散型モデルの場合は，

$$I(g;f) = \sum_{i=1}^{\infty} g(x_i) \log \frac{g(x_i)}{f(x_i)} \tag{1.7}$$

と表現される（小西・北川, 2004, p.28）。

カルバック–ライブラー情報量には二つの性質が存在する。一つ目は $I(g;f) \geq 0$ であり，二つ目は $I(g;f) = 0$ であるとき，かつそのときに限り，$g(x) = f(x)$ である。これらの性質から，カルバック–ライブラー情報量の値が 0 に近いほど，提案モデル $f(x)$ は真のモデル $g(x)$ に近いということができる。この性質の証明は小西・北川 (2004, pp.28-29) に与えられている。

カルバック–ライブラー情報量を利用することで，提案モデルのよさを評価することが可能となる。しかしながら，実際には，カルバック–ライブラー情報量を利用することは難しい場合が多い。カルバック–ライブラー情報量 (1.5) 式自体がその内に真の分布を含んでしまっているために，これが未知である場合には直接的にカルバック–ライブラー情報量を得ることができない。

1.4.3 平均対数尤度

カルバック–ライブラー情報量は

$$\begin{aligned} I(g;f) =& \int \log\left(\frac{g(x)}{f(x)}\right) g(x)dx \\ =& E_G[\log g(X)] - E_G[\log f(X)] \end{aligned} \tag{1.8}$$

と分解可能である（小西・北川, 2004, p.33; 赤池・甘利・北川・樺島・下平, 2007）。ここで E_G は確率変数 X が真の分布 g に従うものとしての期待値を示す。(1.8) 式の最右辺第 1 項 $E_G[\log g(X)]$ は，未知なる真のモデルに関する値であり，直接は知り得ないにしろ，ただ一つの値に定まる。ところで，カルバック–ライブラー情報量の性質から，$I(g;f) \geq 0$ であり，なおかつカルバック–ライブラー情報量の値が 0 に近いほど，提案モデル $f(x)$ は真のモデル $g(x)$ に近いということが可能であった。この性質から，$I(g;f)$ を 0 へと近づけるためには，(1.8) 式の最右辺第 2 項 $E_G[\log f(X)]$ が大きいほどよいということになる (Burnham & Anderson, 2002, p.58; 小西・北川, 2004, p.33)。平均対数尤度 (expected log-likelihood) とい

うときは，この第 2 項 $E_G[\log f(X)]$ のことを指す（小西・北川, 2004, p.33）。平均対数尤度は

$$E_G[\log f(X)] = \int \log f(x) dG(x)$$
$$= \int_{-\infty}^{\infty} g(x) \log f(x) dx \tag{1.9}$$

によって算出されるが，これは未知なる真のモデル g に依存しているため，そのままでは求めることができない。

1.4.4 平均対数尤度の推定量としての対数尤度

前項までで，構成したモデルと真のモデル間の距離としてのカルバック–ライブラー情報量を導入した。また，直接算出することが困難なカルバック–ライブラー情報量そのものに代えて，その構成要素である平均対数尤度を用いてモデル比較を行うことを概説した。しかしながら，この平均対数尤度もまた，真の分布に依存するために直接利用することに難があることが分かった。そこで，平均対数尤度を算出することに代えて，何らかの方法で推定し，この推定値を用いて，モデル比較を行うことが考えられる。その方法が，対数尤度による平均対数尤度の推定である。

真の分布 $g(x)$ から生成されたデータ $x_1, x_2, \cdots, x_n$ が与えられると，

$$l = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i) \tag{1.10}$$

によって平均対数尤度を推定することが可能となる。データ数 n が $n \to \infty$ となるにしたがい，大数の法則から l は $E_X[\log f(X)]$ に収束する。このことから，(1.10) 式の l が大きい構成モデルほど，比較しているモデル間において，よいモデルであると判断することができる。データが手元に得られれば対数尤度を計算することができ，計算した値は，平均対数尤度の不偏推定値となる。この推定値を用いて，モデルのよし悪しを評価することが可能となる。

それでは，いま，モデル評価のために対数尤度 l を求めたいものとする。ただし，実際のモデル内には，未知のパラメタ $\boldsymbol{\theta}$ が含まれる。このとき対数尤度はパラメタ $\boldsymbol{\theta}$ の値によって変わるため，l を $\boldsymbol{\theta}$ の関数 $l(\boldsymbol{\theta})$ とみなす。つまり，対数尤度関数 $l(\boldsymbol{\theta})$ を最大とするパラメタ $\boldsymbol{\theta}$ を求めることで，それは近似的に最もよい構成モデルを与えるパラメタを求めることが可能となる。

カルバック–ライブラー情報量の場合，それは最尤推定量 $\hat{\boldsymbol{\theta}}(= \hat{\boldsymbol{\theta}}(\boldsymbol{x}))$ を用いればよいことになる（北川, 2007）。

1.4.5　対数尤度のバイアス

平均対数尤度を直接計算することはできなくとも，手元のデータから計算され得る対数尤度を，平均対数尤度の推定値として用いればよいことを前節で述べた。つまり，最尤推定による推定量 $\hat{\boldsymbol{\theta}}$ によって構成されるモデル $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))$ を，対数尤度 $l(\hat{\boldsymbol{\theta}}(\boldsymbol{x}))$ を用いて評価すればよいこととなる。

しかし，最尤推定量 $\hat{\boldsymbol{\theta}}$ を代入した対数尤度 $l(\hat{\boldsymbol{\theta}}) = \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ を用いて，モデルのよさを評価することにはまだ問題が存在する。

対数尤度はが平均対数尤度の不偏推定量となるのはパラメタ $\boldsymbol{\theta}$ が固定されている場合である。その一方で，パラメタに最尤推定量を代入した場合には，対数尤度は平均対数尤度の不偏推定量ではなく，偏り（バイアス; bias）のある推定量となる。このバイアスは，パラメタ数の次元の変化にともない変化する。最尤推定量を代入した場合の対数尤度の，平均対数尤度に対するバイアスは，同一データをパラメタ $\boldsymbol{\theta}$ の推定および，推定されたモデルの平均対数尤度の推定の 2 度に渡り使用したことによって生じたものである（北川, 2007）。

バイアスの補正

ここで，対数尤度と平均対数尤度の差を

$$\begin{aligned}\delta =& \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{x})) - E_{\boldsymbol{y}}[\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] \\ =& \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{x})) - \log f(\boldsymbol{x}|\boldsymbol{\theta}_0) \\ &+ \log f(\boldsymbol{x}|\boldsymbol{\theta}_0) - E_{\boldsymbol{y}}[\log f(\boldsymbol{y}|\boldsymbol{\theta}_0)] \\ &+ E_{\boldsymbol{y}}[\log f(\boldsymbol{y}|\boldsymbol{\theta}_0)] - E_{\boldsymbol{y}}[\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}(\boldsymbol{x}))] \end{aligned} \tag{1.11}$$

と分解し、和の記号ごとに $\delta_1 + \delta_2 + \delta_3$ と置く（北川, 2007, p.83）。$\boldsymbol{\theta}_0$ は

$$E_G\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}|\boldsymbol{\theta})\right] = \int g(\boldsymbol{y})\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y} = \mathbf{0} \tag{1.12}$$

の解である（小西・北川, 2004, p.50)。ここで $\boldsymbol{y}$ は未来に得られるデータを表している。上式，δ_2 は

$$E_X[\delta_2] = 0 \tag{1.13}$$

となり，δ_1, δ_3 はそれぞれ漸近的に

$$E_X[\delta_1] = E_X[\delta_3] = \frac{1}{2}\mathrm{tr}(\boldsymbol{I}\boldsymbol{J}^{-1}) \tag{1.14}$$

と近似される。$\delta_1, \delta_2, \delta_3$ の期待値のより詳細な導出は小西・北川 (2004, pp.51-52) に与えられている。δ の期待値，すなわち漸近的な偏りは，上記の分解から，

$$\begin{aligned}\delta =& \delta_1 + \delta_2 + \delta_3 \\ =& \frac{1}{2}\mathrm{tr}(\boldsymbol{I}\boldsymbol{J}^{-1}) + 0 + \frac{1}{2}\mathrm{tr}(\boldsymbol{I}\boldsymbol{J}^{-1}) \\ =& \mathrm{tr}(\boldsymbol{I}\boldsymbol{J}^{-1})\end{aligned} \tag{1.15}$$

となり，$\mathrm{tr}(\boldsymbol{I}\boldsymbol{J}^{-1})$ であることが示される。なお $\boldsymbol{I}$ はサイズ $p \times p$ のフィッシャー情報量行列 (Fisher information matrix) を表し，$\boldsymbol{J}$ はサイズ $p \times p$ のヘッセ行列 (Hessian matrix) の期待値の符号を反転させたものである（北川, 2007, pp.82-83）。つまり，

$$\boldsymbol{I} = E_X\left[\frac{\partial \log f(X|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \log f(X|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right] \tag{1.16}$$

であり，また，

$$\boldsymbol{J} = -E_X\left[\frac{\partial^2 \log f(X|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] \tag{1.17}$$

と定義される。Akaike (1973) は，平均対数尤度の推定量として最尤推定量 (maximum likelihood estimator) における対数尤度関数 (log-likelihood function) を最大化した最大対数尤度 (maximum log likelihood) を用い，さらに，δ の期待値（漸近バイアス）がパラメタ（次元）数 p で近似可能であることを示し，対数尤度にパラメタ数によるバイアス修正を行った

$$\mathrm{AIC} = -2(\text{最大対数尤度}) + 2(\text{パラメタ数}) \tag{1.18}$$

$$= -2l(\hat{\theta}) + 2p \tag{1.19}$$

を提案した。(1.19) 式は，赤池情報量規準 (Akaike's Information Criterion, AIC; Akaike, 1973) と呼ばれる。赤池情報量規準は近似的に平均対数尤度の不偏推定量となる。

AIC を用いることで，我々はデータが得られた真なる構造が未知のままであったとしても，複数の提案モデルを，真のモデルとの近さ，という観点から評価することが可能となる（赤池・甘利・北川・樺島・下平, 2007, p.19）。

AIC の値が最小となるモデルを最良のモデルとして評価することとなる。なお，小西・北川 (2004) は，AIC が最小となることで選択された推定値やモデルを，最小 AIC 推定値と呼んでいる。ただし，AIC は比較対象となるモデル間の相対的なよさの指標であることや，相対的に AIC が最小となったモデルが必ずしも真のモデルであるとは限らないことに注意が必要である。

1.4.6 ベイジアン情報量規準

Schwarz (1978) はベイズ統計学[4]の立場からモデル選択規準を提案した。ここでは小西・北川 (2004, pp.151-156)，甘利 (2007) の記述に基づき，Schwarz (1978) によって提案されたモデル選択規準について概説する。

いま，手元に n 個のデータ $\boldsymbol{x}_n$ が得られた状況を考える。このデータに関して，比較候補モデル $f(\boldsymbol{x}|\boldsymbol{\theta})$ を I 個作成するものとする。ここで，それぞれの候補モデルはパラメタ $\boldsymbol{\theta}_i$ を用いて特定される。すると候補モデルは，それぞれ $f_1(\boldsymbol{x}_n|\boldsymbol{\theta}_1), f_2(\boldsymbol{x}_n|\boldsymbol{\theta}_2), \cdots, f_I(\boldsymbol{x}_n|\boldsymbol{\theta}_I)$ と表される。ここで，パラメタ $\boldsymbol{\theta}_i$ は事前分布 $\pi_i(\boldsymbol{\theta}_i)$ によって，その事前情報が表現されるものとする。

このとき，候補モデル $f_i(\cdot|\cdot)$ から観測データ $\boldsymbol{x}_n$ が得られる確率は，パラメタ $\boldsymbol{\theta}_i$ に関して積分を行うことで得られる，候補モデル分布の周辺尤度

$$\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \tag{1.20}$$

によって与えられる（小西・北川, p.151）。上式の周辺尤度と，ベイズの定理を用いることで，候補モデル $f_i(\cdot|\cdot)$ に関して確率

$$\Pr\{f_i(\boldsymbol{\theta}_i|\boldsymbol{x}_n)\} = \frac{\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \Pr\{f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\}}{\sum_{i=1}^{I}\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \Pr\{f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\}}, \quad i = 1, 2, \cdots, I \tag{1.21}$$

が得られる。ここで $\Pr\{f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\}$ は候補モデル $f_i(\cdot|\cdot)$ の事前確率と呼ばれる情報を表す。また，$\Pr\{f_i(\boldsymbol{\theta}_i|\boldsymbol{x}_n)\}$ は事後確率と呼ばれ，データ $\boldsymbol{x}_n$ が手元に得られたときに，当該データが候補モデル f_i から生成される確率を表している。つまり，I 個の候補モデルの内からどれか一つのモデルを選択するとした場合，事後確率最大 (maximum a posteriori, MAP) となるようなパラメタ推定値を有するモデルを採用すればよい。なお，この基準は，(1.21) 式の分母がすべてのモデルにおいて共通であることから，分子を最大にするモデルの選択と同義である（小西・北川, p.152）。

また，事前確率 $\Pr\{f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\}$ が候補モデルに依らず等しい場合には，分子を最大にするモデルの選択とは，周辺尤度を最大にするモデルの選択を意味することとなる。

ただし，(1.21) 式の周辺尤度の近似には，積分計算が含まれており，計算する上で困難が生じる場合がある。ベイジアン情報量規準 (Bayesian Information Criterion, BIC) は，この積分をラプラス近似によって

$$-2\log\left\{\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i\right\} \approx -2\log f_i(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_i) + p_i\log n \tag{1.22}$$

[4]ベイズ統計学については第 4 章において記述する。

と近似される。なおこの近似は標本数が十分に大きいことが想定されている。ここで，$\hat{\boldsymbol{\theta}}_i$ は候補モデル $f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)$ の p_i 次元パラメタベクトル $\boldsymbol{\theta}_i$ の最尤推定量を表している。BIC においては，各モデルのパラメタ数を表す項が AIC とは異なり，$\log n$ となっている。

AIC には，標本数を増やしても，AIC によって選択した構成モデルのパラメタ次数は，真のモデルの次数に一致しない，つまり一致性を持たないという主張が存在する（北川, 2007, p.84; 甘利, 2007, pp.65-66）。一方で，BIC におけるモデル選択で選ばれるモデルの次数は一致性をもつこととなる。ただし，このことをもって BIC が AIC よりも優れているということにはならない（例えば甘利, 2007, p.67; 北川, 2007, p.85-86）。

I 個の比較候補モデルを最尤法によって推定し，その内から，BIC の値が最小となるモデルを採用する。

1.5 研究の目的

次章以降では，本章で挙げた項目特性図作成の際に直面する問題へと対処する方法の提案を行い，それぞれの提案手法に関して，具体的な適用例を通じて，手法の有用性の検討を行う。

第 2 章では研究 I として，赤池情報量規準，ベイジアン情報量規準を利用した項目特性図の群数選択法を提案する。提案手法 I を用いることで，統計的な基準に基づき，群数を選択することが可能となる。第 2 章では，シミュレーションと実際のテストデータへの適用例を通じて，提案手法が有用であることを示す。

第 3 章において，研究 II として，ある特定の項目において，カテゴリ特性曲線の併合状況（モデル）が複数想定され得る場合に，分析者の仮定（ある特定のカテゴリ特性曲線傾向が同等であると見做すことができるか）を確かめるための情報量規準による統計的指標を算出する方法を提案する。本提案手法を用いることで，効果的な誤答分析を行うための項目特性図を作成することが可能となる。また，研究 II における提案手法 II を用いることで，分析者が項目特性図を考察した結果得られた仮説を検証し，なおかつ図の視認性を確保することが可能となる。

第 4 章では，後の第 5 章で用いられるハイブリッドモンテカルロ (hybrid Monte Carlo, HMC) 法[5] について概説する。

第 5 章の研究 III においては，同一項目において複数の項目特性図が作成可能な場合に，何れの項目特性図の表現を採用すればよいのか，もしくはそれらの項目特性図の統合は可能かについて，その選択基準を提案する。提案手法 III について，シミュレーションを行い，実データへの適用例を示す。

最後に第 6 章において，本稿で示した研究全体の考察を行う。

[5] ハミルトニアンモンテカルロ (Hamiltonian Monte Carlo) 法とも呼ばれる。

第2章　項目特性図の群分け基準（研究 I ）

2.1　問題 I

第 1 章で述べたとおり，項目特性図は項目分析の際に，項目の測定性能を目視で確認することを可能とする有用な道具である。ただし，項目特性図の作成時に恣意的な作業が必要となる。作成時には受験者を，得点に基づいていくつかの群へと分けることが求められる。この際に，分割する数，すなわち群数は任意に設定可能である。受験者数が比較的少数である場合には，群数の変更によって，項目特性図の表現が大きく変化することへとつながる。

項目特性図の作成時に受験者を群に分ける際，群数 G の値を決定するための明確な基準は知られておらず，現状では分析者によって経験的に決定されており，一般的には $G = 5$ とされる。しかし，その根拠は明確には示されていない。

受験者をいくつかの群に分けることはテスト理論，および教育的な見地から関心を向けられ，電子計算機が発達する以前より研究されている。池田（1973）では，受験者を得点群別に分け，各群の選択肢の分布を用いて項目分析を行う際に，D 指数法において，受験者上位群と下位群の所属割合を 27%とすることが示唆されている。27%という基準は Kelly（1939）において，変数が正規分布する場合に標本平均の差の臨界比（critical ratio）を最大化する比率として 0.2702678 が導かれたことに由来する（赤根ら, 2006）。また，Croon（2002）では，潜在クラス分析を行い，適合度指標を用いてクラス数を検討している。近年では Shojima（2008）が，潜在ランク理論（荘島, 2010）における項目参照プロファイル（item reference profile, IRP; 荘島, 2010, p.86）を作成する際に，受験者を分割する潜在ランク数を適合度指標および情報量規準を用いて決定することを検討している。

項目特性図の作図の際にも，群の分割数の選択について，統計的な基準や根拠を与えることができれば，項目特性図を用いて項目の性質を調べる上で便利である。

2.2 研究 I 目的

本研究では情報量規準（赤池情報量規準，ベイジアン情報量規準）を利用した項目特性図の群数選択法を提案する。

本研究における群数選択法には二つの利点が挙げられる。第一点目は，項目特性図を描く際に，受験者の多寡に応じて，安定した特性の表現を行うために利用可能であるということである。

項目の性質を考察するという目的のためには，分割する群の数が少なすぎると正答確率の変化に関する直線の表現が大味になり過ぎ，大域的考察のみが可能となる。例えば 2 群に分割した場合，項目特性図は 1 本の直線で表されることになり，当該直線の傾きのみが考察の対象となる。受験者が多い場合には，群数を多くしてもグラフ上の表現は安定するため，項目の性質を細やかに考察することが可能となる。一方で，受験者の全体人数が少数である場合は，群数を増やそうとすると各群に属する受験者数が少なくなるために，グラフ上の表現が安定しなくなってしまう。この場合，無理に群数を増やそうとはせず，むしろ減らすことによって，安定した項目の性質を考察することが期待できる。

第二点目は，近年我が国において広く用いられている項目反応理論（item response theory, IRT; Lord, 1980）が適用できないテスト項目に関しても項目特性図は作図可能であり，その特性表現が有する情報は，決して IRT における ICC (item characteristic curve) と比較しても見劣りがするものではない，ということである。

項目特性を考察する道具の一つとして IRT における ICC がある。ICC を利用するためには，まずテストデータに対して IRT が適用可能であることが前提条件となる。IRT が適用可能なテストデータを研究 I における提案手法で分析することは可能だが，この逆は必ずしも成り立たない。また，後述する提案手法は IRT よりも簡単に分析を行うことが可能である。更に IRT はパラメトリックな分析手法であるため，項目母数が定まると自動的に $-\infty$ から ∞ の範囲までにロジスティック曲線も定まることとなる。一方で項目特性図においては，低得点群から中得点群にかけて一旦正答率が下がりながらも，高得点群においては再び上昇傾向に転じる，低得点群から中得点群においては特性表現が単調増加傾向にありながら，高得点群においては正答率が下がる，あるいはこの両方の傾向が観察され，群間で正答率が上下する，といったロジスティック曲線では表せないような状況をも表現することが可能である。この点を鑑みれば，項目特性図を用いた折れ線グラフによる考察は項目特性について豊かな情報を提供するといえよう。上記の理由より，以下に提案する手法は IRT の普及によってもその重要性が損なわれることはないものといえる。

2.3 情報量規準を用いた群数選択法の提案（方法 I）

いま，J 個の項目によって構成されるテストに関して，第 $j(=1,\cdots,J)$ 番目の項目に対する受験結果が正誤の 2 カテゴリに整理されるものとする。

項目 j に対する，第 $g(=1,\cdots,G)$ 群に属する i 番目の受験者の反応 x_{jgi} が正答反応として観測される確率を p_{jg} とする。第 j 項目の第 g 群における受験者が N_{jg} 人おり，その内で正答した受験者が n_{jg} 人いるとき，正誤反応パタンベクトル

$$\boldsymbol{x}_{jg} = (x_{jg1},\cdots,x_{jgi},\cdots,x_{jgN_{jg}})'$$

を得る確率は

$$p(\boldsymbol{x}_{jg}|p_{jg}) = p_{jg}^{n_{jg}}(1-p_{jg})^{N_{jg}-n_{jg}}$$

となる。なお，$n_{jg} = \sum_{i=1}^{N_{jg}} x_{jgi}$ は正答者数を意味しており，これは十分統計量である。項目 j の群ごとの反応パタンベクトルの集まり

$$\boldsymbol{x}_j = (\boldsymbol{x}'_{j1},\cdots,\boldsymbol{x}'_{jg},\cdots,\boldsymbol{x}'_{jG})'$$

を観測する確率は，p_{jg} を

$$\boldsymbol{p}_j = (p_{j1},\ldots,p_{jg},\ldots,p_{jG})'$$

のように配した母数ベクトル $\boldsymbol{p}_j$ を用いて，

$$p(\boldsymbol{x}_j|\boldsymbol{p}_j) = \prod_{g=1}^{G} p_{jg}^{n_{jg}}(1-p_{jg})^{N_{jg}-n_{jg}}$$

となる。p_{jg} の最尤推定量は

$$\hat{p}_{jg} = \frac{n_{jg}}{N_{jg}}$$

であり，これを $\boldsymbol{p}_j$ の各要素として配することで，最大対数尤度を

$$\begin{aligned}\log l_{\max} &= \log l(\hat{\boldsymbol{p}}_j) \\ &= \sum_{g=1}^{G}\{n_{jg}\times\log\hat{p}_{jg} \\ &\quad + (N_{jg}-n_{jg})\times\log(1-\hat{p}_{jg})\}\end{aligned} \tag{2.1}$$

と導くことができる。

(2.1) 式を用いて，2 カテゴリで整理される項目 j に関して，群数 G で描く項目特性図の AIC および BIC，

$$AIC_{j(G)} = -2 \times \log l_{\max} + 2 \times G \tag{2.2}$$

$$BIC_{j(G)} = -2 \times \log l_{\max} + G \times \log(N_j) \tag{2.3}$$

を算出可能である。ここで $N_j = \sum_{g=1}^{G} N_{jg}$ である。

(2.2) 式および (2.3) 式によって項目ごとに情報量規準を用いて群数を選択することが可能であるが，実用場面においてはテスト項目全体で分割する群数を統一した方が扱い易い場合もある。テスト項目全体で同じ群数を用いて項目特性図を描く場合，AIC と BIC はそれぞれ

$$AIC_{(G)} = \sum_{j=1}^{J} AIC_{j(G)} \tag{2.4}$$

$$BIC_{(G)} = \sum_{j=1}^{J} BIC_{j(G)} \tag{2.5}$$

で算出する。ただし，項目間の独立性が本質的に否定される場合には，(2.4) 式，(2.5) 式は成立しないことに注意する必要がある。

2.4 シミュレーションによる提案方法の検討

前節「方法 I」において提案された手法が適切な群数を推奨可能であるかを調べるために，様々な条件でシミュレーションデータを 1000 回推定し，提案手法によって真の群数を何回当てられるかを調べた。複数のモデルの良さを比較する場合に，情報量規準を算出し，その値が最も小さいモデルが最適なモデルであると判断することが可能である（坂元・石黒・北川, 1983）。ここでは，ある項目について，群数を変化させた場合の情報量規準の値を比較し，値が最小となる群数を当該項目の項目特性図作成時における群数として採用する。本章ではシミュレーション手順および結果について示す。本研究では，シミュレーションにおける群への分割方法は，等人数，等間隔の両方で行うこととした。

シミュレーションデータのパラメタ設定　　シミュレーションに用いるテストデータを人工的に発生させるためのパラメタを設定する。対象となるパラメタは

1. 受験者数：100 人，200 人，300 人，400 人，500 人
2. 真の群数：3 群，4 群，5 群，6 群，7 群

の二つである。なお，ここではテスト項目数を 100 問，各項目の配点を 1 点に固定した。

等人数データ発生手順　まずシミュレーションを行う受験者数と，当てるべき真の群数を設定する。次に架空の受験者数を真の群数で割り，各群の所属人数を決定する。その際に余りが出た場合は，その分の人数は低い方の群から一人ずつ含めることとした。最後に真の群数（3・4・5・6・7）における各群の各項目への正答確率を一様分布から乱数として発生させる。なお，各群の一様乱数発生範囲は表 2.1 のように設定した。例えば真の群数が 3 群の場合の，第 1 群の各項目への正答確率は，0.00 から 0.34 の範囲において発生させた一様乱数によって設定されることとなる。

表 2.1 真の群数における群ごとの一様乱数の設定範囲

3 群	始点	終点
第 1 群	0.00	0.34
第 2 群	0.31	0.67
第 3 群	0.64	1.00
4 群	始点	終点
第 1 群	0.00	0.25
第 2 群	0.20	0.50
第 3 群	0.45	0.75
第 4 群	0.70	1.00
5 群	始点	終点
第 1 群	0.00	0.20
第 2 群	0.15	0.40
第 3 群	0.35	0.60
第 4 群	0.55	0.80
第 5 群	0.75	1.00
6 群	始点	終点
第 1 群	0.00	0.17
第 2 群	0.12	0.33
第 3 群	0.29	0.50
第 4 群	0.46	0.67
第 5 群	0.63	0.84
第 6 群	0.80	1.00
7 群	始点	終点
第 1 群	0.00	0.15
第 2 群	0.10	0.30
第 3 群	0.25	0.45
第 4 群	0.40	0.60
第 5 群	0.55	0.75
第 6 群	0.70	0.90
第 7 群	0.85	1.00

等間隔データ発生手順　等間隔法では，受験者のテスト得点の分布に正規分布を仮定し，データを発生させる。まずシミュレーションを行う受験者数と，当てるべき真の群数を設定する。

次に偏差値の 30 から 70 の範囲を，設定した真の群数で割る。真の群数に対応した累積点を求め，これらの各点の間を，群間の区間として設定する。最後に区間内の面積を正規分布表より求め，各群の所属人数をこれらの区間の面積に従って決定する。

なお，0 から 30，および 70 から 100 部分の面積はそれぞれ 30 からの最低位群と 70 までの最高位群の面積に含めることとした。また，面積を所属人数に換算する際に，四捨五入等により余りが出た場合は，各群のうちで最も多い人数から一名ずつ引くこととした。

受験者の正誤パタン発生方法は等人数データの発生手順と同様である。

受験者数，真の群数，それぞれのパラメタ設定について発生させた等人数，等間隔のシミュレーションデータに関して，(2.4) 式，(2.5) 式に基づき，2 群から 10 群までテスト全体の AIC，BIC を算出，比較し，それぞれで最小となる群数を確認した。なお，等人数データの場合は等人数分割とし，等間隔データの場合は等間隔に群を分割することとした。

2.5　シミュレーション結果

表 2.2 に 1000 回分のシミュレーションにおける AIC による推奨群数の頻度を示した。表 2.2 では，縦に群（2 群から 10 群）を，横に受験者数を配し，各セルはそれぞれの条件における情報量規準による推奨頻度（情報量規準が相対的に最小となった度数）を示している。

表 2.2 等人数データ AIC 推奨群数の頻度

群数	100	200	300	400	500
真の群数 (3)					
2	0	0	0	0	0
3	1000	1000	999	978	652
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	1	22	348
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (4)					
2	0	0	0	0	0
3	0	0	0	0	0
4	1000	1000	1000	1000	998
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	2
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (5)					
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	1000	1000	1000	1000	1000
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (6)					
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	11	0	0	0	0
6	989	1000	1000	1000	1000
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (7)					
2	0	0	0	0	0
3	0	0	0	0	0
4	19	0	0	0	0
5	149	0	0	0	0
6	254	1	0	0	0
7	578	999	1000	1000	1000
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

表 2.2 より，真の群数が比較的少なく，3 群の場合，受験者数が 100 人から 400 人では，高い割合で真の群数として設定した値が推奨された。人数が 500 人の場合には，試行数のうち 35%が 6 群を推奨する結果となった。このことから受験者数が多く，真の群数が比較的少ない場合には，無駄に群数を増やしてしまう傾向がやや認められる。

真の群数が 4，5，6 群の場合には，いずれの受験者数であってもほぼ，真の群数と同じ値を推奨する結果となった。

真の群数が 7 群と比較的多く，受験者数が 100 人の場合，AIC では真の群数と同じ群数を一貫して推奨することはできず，7 群よりも小さい群数が推奨される結果となった。これは 7 群にまで分割してしまうと，各群に所属する受験者の人数が減少し，表現が不安定となってしまうためと推察される。

表 2.3 にシミュレーションにおける BIC による各推奨群数の頻度を示した。AIC と比較すると，BIC では真の群数が 3・4 群という比較的少ない群数である場合は AIC による推奨結果よりも多く，1000 回とも真の群数として設定した値を推奨する結果となった。

表 2.3 等人数データ BIC 推奨群数の頻度

群数	100	200	300	400	500
真の群数 (3)					
2	0	0	0	0	0
3	1000	1000	1000	1000	1000
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (4)					
2	0	0	0	0	0
3	0	0	0	0	0
4	1000	1000	1000	1000	1000
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (5)					
2	0	0	0	0	0
3	946	0	0	0	0
4	16	0	0	0	0
5	38	1000	1000	1000	1000
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (6)					
2	0	0	0	0	0
3	998	730	48	0	0
4	2	269	249	0	0
5	0	1	35	2	0
6	0	0	668	998	1000
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (7)					
2	0	0	0	0	0
3	993	112	0	0	0
4	7	888	973	533	73
5	0	0	27	431	339
6	0	0	0	22	63
7	0	0	0	14	525
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

ただし，受験者数が少なく，真の群数が大きい場合には推奨群数に関して保守的な傾向にあることがわかる。特に受験者数が 100 人のときには，真の群数 5・6・7 群が，200 人のときには 6・7 群において真の群数と推奨される群数の食い違いが大きくなっていた。5 群であれば 200 人ほど，6 群の場合には 300 人以上から，安定して真の群数を推奨する傾向が見いだせるものの，7 群の場合には受験者数 500 人であっても真の群数と同じ値を推奨する割合は 5 割となっている。一般的に，BIC は AIC と比較してパラメタに関して倹約的な傾向にある。このことが今回の AIC と BIC の推奨頻度の差となって表れた一因であろう。

表 2.4 は分割方法を等間隔とした場合のシミュレーションにおける AIC の推奨群数の度数表である。真の群数が 3・4 群と少ない場合には，受験者数が 200 人を超えると，群数が多めに見積もられる傾向にあることがわかる。受験者数が多い場合には，群数に関するパラメタを過分に増やしてしまう可能性に留意する必要があるだろう。

表 2.5 は BIC の場合の推奨群数の頻度表である。BIC では，3・4 群の場合には，いずれの受験者数であっても概ね真の群数と同じ値を推奨する結果となった。しかし，真の群が 5 群以上の場合は，真の群数を推奨するには 100 人では人数が不足し，特に 6 群と 7 群ではより多くの受験者を必要とする可能性が示唆された。

表 2.4 等間隔データ AIC 推奨群数の頻度

群数	100	200	300	400	500
真の群数 (3)					
2	0	0	0	0	0
3	907	693	334	86	20
4	93	307	663	912	970
5	0	0	3	2	1
6	0	0	0	0	6
7	0	0	0	0	3
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (4)					
2	0	0	0	0	0
3	0	0	0	0	0
4	799	525	274	164	62
5	192	340	289	185	99
6	9	135	433	638	788
7	0	0	4	13	32
8	0	0	0	0	15
9	0	0	0	0	4
10	0	0	0	0	0
真の群数 (5)					
2	0	0	0	0	0
3	0	0	0	0	0
4	3	0	0	0	0
5	844	531	314	197	118
6	147	402	491	387	273
7	6	64	167	295	354
8	0	3	28	114	239
9	0	0	0	7	16
10	0	0	0	0	0
真の群数 (6)					
2	0	0	0	0	0
3	0	0	0	0	0
4	246	0	0	0	0
5	80	0	0	0	0
6	632	625	369	205	101
7	42	340	492	505	397
8	0	35	131	261	372
9	0	0	8	28	115
10	0	0	0	1	15
真の群数 (7)					
2	0	0	0	0	0
3	0	0	0	0	0
4	280	0	0	0	0
5	662	56	0	0	0
6	24	3	0	0	0
7	33	661	430	264	149
8	1	269	517	593	591
9	0	11	52	131	222
10	0	0	1	12	38

表 2.5 等間隔データ BIC 推奨群数の頻度

群数	100	200	300	400	500
真の群数 (3)					
2	0	0	0	0	0
3	1000	1000	1000	1000	999
4	0	0	0	0	1
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (4)					
2	110	0	0	0	0
3	10	0	0	0	0
4	880	1000	1000	1000	988
5	0	0	0	0	12
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (5)					
2	1	0	0	0	0
3	997	638	9	0	0
4	2	27	2	0	0
5	0	335	989	1000	1000
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (6)					
2	27	0	0	0	0
3	973	715	86	2	0
4	0	285	913	878	400
5	0	0	1	11	11
6	0	0	0	109	589
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
真の群数 (7)					
2	0	0	0	0	0
3	997	224	0	0	0
4	3	776	987	834	478
5	0	0	13	166	522
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

表 2.6 および表 2.7 は等人数法および等間隔法で群分けを行った場合の 1000 回のシミュレーションにおける，AIC，BIC による推奨群数の平均値と分散である。等人数法の場合，AIC に関しては真の群数が 3 群かつ受験者数が相対的に多いときには，推奨群数も大きくなる傾向にある。また，真の群数が 7 群，かつ受験者数が比較的少ない場合には真の群数よりも推奨群数が小さくなる傾向にあることが分かる。BIC については，真の群数が比較的小さい場合，AIC における結果と同様に真の群数を推奨することが可能であった。また，真の群数が 7 群の場合，設定した受験者数では一貫して真の群数よりも小さな群数に分割することが推奨される傾向も再確認された。ただし，受験者数が増えるに連れ，真の群数を推奨する傾向も窺うことが可能である。

等間隔法の場合も，等人数法と同様の傾向が確認されるものの，真の群数の周りでの散らばりが，等人数法よりも大きい傾向が認められた。

表 2.6 情報量規準による推奨群数の平均値

等人数データ平均値					
群＼人数	100	200	300	400	500
AIC(3)	3	3	3.003	3.066	4.044
AIC(4)	4	4	4	4	4.008
AIC(5)	5	5	5	5	5
AIC(6)	5.989	6	6	6	6
AIC(7)	6.391	6.999	7	7	7
BIC(3)	3	3	3	3	3
BIC(4)	4	4	4	4	4
BIC(5)	3.092	5	5	5	5
BIC(6)	3.002	3.271	5.323	5.998	6
BIC(7)	3.007	3.888	4.027	4.517	6.04
等間隔データ平均値					
群＼人数	100	200	300	400	500
AIC(3)	3.093	3.307	3.669	3.916	4.002
AIC(4)	4.210	4.610	5.167	5.500	5.851
AIC(5)	5.156	5.539	5.909	6.347	6.762
AIC(6)	5.470	6.410	6.778	7.115	7.546
AIC(7)	4.813	7.176	7.624	7.891	8.149
BIC(3)	3	3	3	3	3.001
BIC(4)	3.770	4	4	4	4.012
BIC(5)	3.001	3.697	4.98	5	5
BIC(6)	2.973	3.285	3.915	4.227	5.189
BIC(7)	3.003	3.776	4.013	4.166	4.522

表 2.7 情報量規準による推奨群数の分散

等人数データ分散					
群＼人数	100	200	300	400	500
AIC(3)	0	0	0.009	0.194	2.044
AIC(4)	0	0	0	0	0.032
AIC(5)	0	0	0	0	0
AIC(6)	0.011	0	0	0	0
AIC(7)	0.651	0.001	0	0	0
BIC(3)	0	0	0	0	0
BIC(4)	0	0	0	0	0
BIC(5)	0.160	0	0	0	0
BIC(6)	0.002	0.200	1.006	0.002	0
BIC(7)	0.007	0.100	0.026	0.378	1.156
等間隔データ分散					
群＼人数	100	200	300	400	500
AIC(3)	0.084	0.213	0.228	0.081	0.072
AIC(4)	0.184	0.508	0.696	0.605	0.453
AIC(5)	0.150	0.395	0.585	0.891	0.992
AIC(6)	0.826	0.312	0.483	0.574	0.771
AIC(7)	0.410	0.510	0.345	0.432	0.501
BIC(3)	0	0	0	0	0.001
BIC(4)	0.397	0	0	0	0.012
BIC(5)	0.003	0.882	0.038	0	0
BIC(6)	0.026	0.204	0.080	0.398	0.954
BIC(7)	0.003	0.174	0.013	0.139	0.250

以上の結果より，提案手法が真の群数に応じて，有効的に群数を推奨できることが確認できた。

2.6　実データを用いた適用例

シミュレーションを通じて,「方法 I」で提案された手法が，特に AIC を用いた場合に真の群数を推奨可能であることが示された。そこで以下では実際のテストデータに対して提案手法を適用し，その結果および解釈を示す。群数選択法を適用する実データとして豊田（2012）に掲載されている「学力テスト 1」（項目数 50 問，受験者数 226 名）を用いた。なお，適用例における群への分割方法は等人数法を用いた。

表 2.8 に「学力テスト 1」の 50 問の各項目について AIC および BIC が最小となった群数を示した。表 2.8 から，シミュレーションを通じて示唆されたように，BIC は AIC よりも保守的な傾向にあることがわかる。表 2.9 にテスト全体における，2 群から 10 群の各群へ分割したときの AIC と BIC を示し，図 2.1 にテスト全体に対する，群数ごとの情報量規準の変化を示した。実線および左縦軸は AIC を示し，点線および右縦軸は BIC を示している。

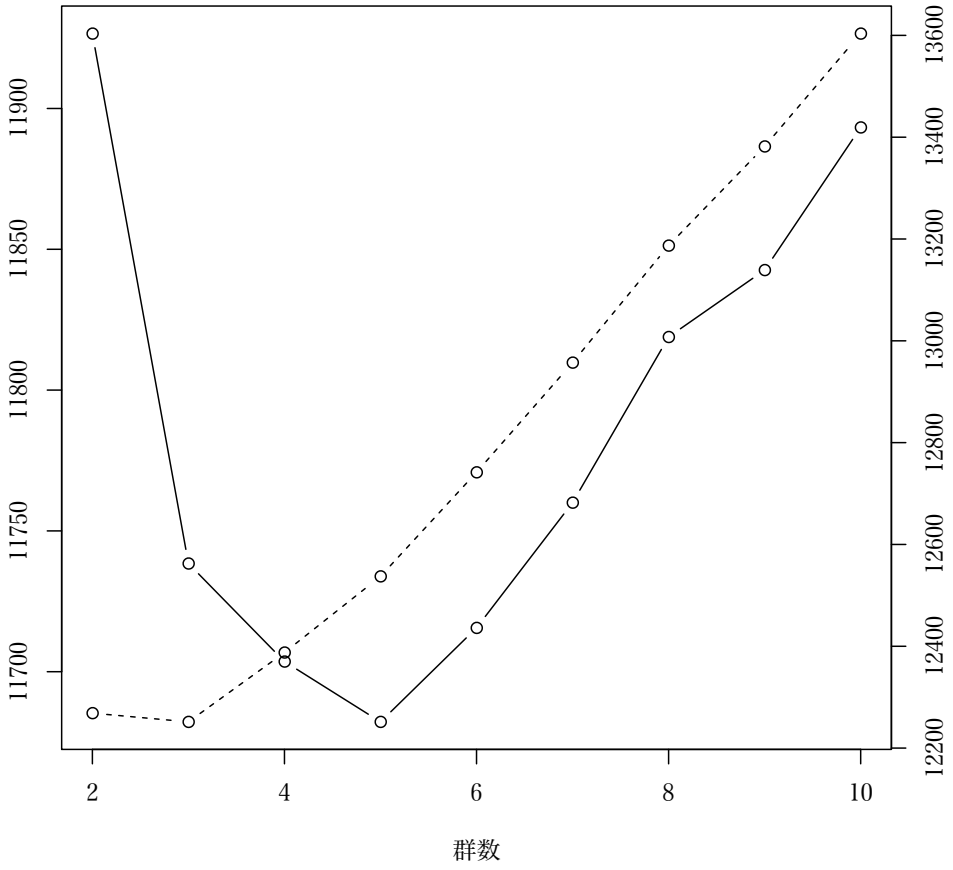


図 2.1 「学力テスト 1」全体における推奨分割数 (左軸:AIC; 右軸:BIC)

テスト全体での群数に関しては，AIC では 5 群が，BIC では 3 群が支持される結果となった。「学力テスト 1」について等人数で分割し，AIC を参照する場合には，これまで経験的に 5 群とされてきた項目特性図の群数について，AIC による推奨群数を傍証として，5 群分割とすることが可能である。BIC に基づいて全体の分割数を 3 群分割とする場合，下位群，中位群，上位群の傾向について分析することが可能となる。

表 2.8 各項目における情報量規準による推奨群数

項目	AIC	BIC	項目	AIC	BIC
1	2	2	26	8	3
2	6	2	27	3	2
3	2	2	28	5	3
4	2	2	29	7	5
5	2	2	30	3	3
6	5	2	31	3	2
7	3	2	32	5	4
8	9	3	33	4	4
9	6	3	34	5	3
10	2	2	35	4	2
11	2	2	36	5	2
12	9	2	37	2	2
13	3	2	38	8	3
14	6	3	39	2	2
15	5	2	40	4	4
16	5	2	41	4	2
17	10	3	42	5	2
18	4	3	43	4	3
19	10	4	44	4	2
20	7	3	45	4	4
21	5	2	46	6	3
22	8	4	47	2	2
23	3	3	48	3	3
24	5	2	49	5	3
25	9	3	50	4	4

表 2.9 テスト全体における群数ごとの情報量規準

群数	AIC	BIC
2	11926.60	12268.65
3	11738.42	12251.50
4	11703.61	12387.71
5	11682.19	12537.32
6	11715.56	12741.72
7	11760.04	12957.22
8	11818.86	13187.07
9	11842.62	13381.87
10	11893.27	13603.54

群数別項目特性図の比較　ここでは同一項目について，推奨群数と群数を変化させた場合で項目特性図を描き，これらを比較することで，群数選択法の特徴を示す。「学力テスト 1」に対して項目別の AIC による推奨分割数を算出し，3 群以上が推奨された項目から 3 項目を選び，それらの項目特性図とその解釈例を示す。

図 2.2 に「学力テスト 1」の第 34 項目について，AIC を用いて分割数を選択し，それに基づいて作成した項目特性図を示した。図 2.2 内左上は，2 群から 10 群の各分割数ごとの情報量規準の推移を示している。ここで実線および左縦軸は AIC を，点線および右縦軸は BIC を表している。図 2.2 内右上は AIC によって推奨された分割数を用いて作成した項目特性図である。図 2.2 内左下および右下はそれぞれ推奨分割数 ±1 の場合で作成した項目特性図を示した。

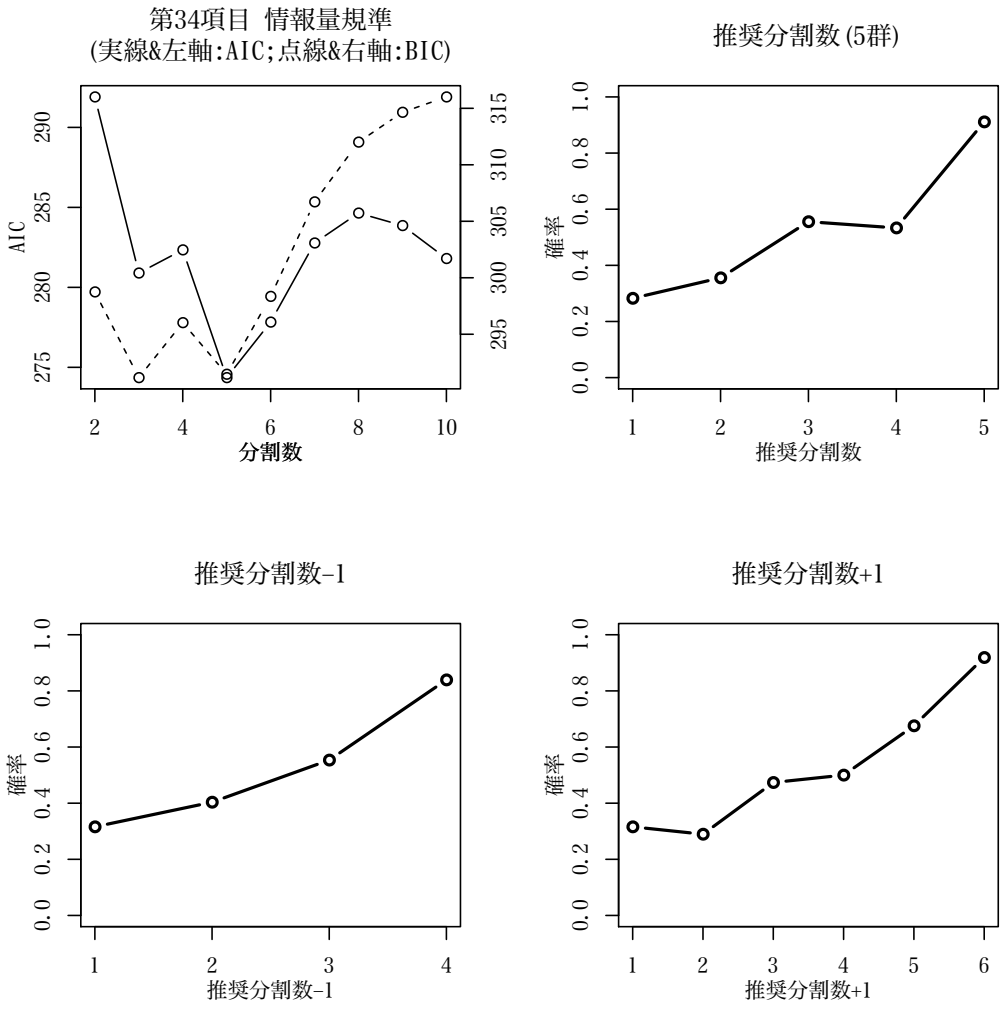


図 2.2 「学力テスト 1」第 34 項目における項目別推奨分割数

第 34 項目については AIC により 5 群が推奨された。分割数 5 ± 1 群について項目特性図を確認すると，5 群に分割した際に見られる，第 3 および第 4 群間で正答率が平坦となる表現が失われてしまっていることがわかる。このように，同一項目について，群数によって項目特性図の表現が異なり，分析者が表現の採用の是非に迷うような場合に,「方法 I」において算出した情報量規準の値を傍証として，項目特性図の群数を選択することが可能である。

図 2.3 には第 40 項目に対して群数選択法を適用した場合の項目特性図を示した。推奨群数は 4 群となった。項目特性図より，第 40 項目は高特性の受験者を識別することに適した項目であることがわかる。3 群での分割では正答率は全体的に右肩上がりの表現となっている。一方，5 群分割とした場合は，4 群と同

様に高特性者を識別することに適してはいるものの，第 3 群の正答率が落ち込んでいる。5 群分割の場合には，第 3 群の正答率下降の要因について，検討することが必要となろう。第 40 項目においては，情報量規準に基づいて，4 群分割とし，項目特性図の分類例における H 型項目として分類することが可能である。

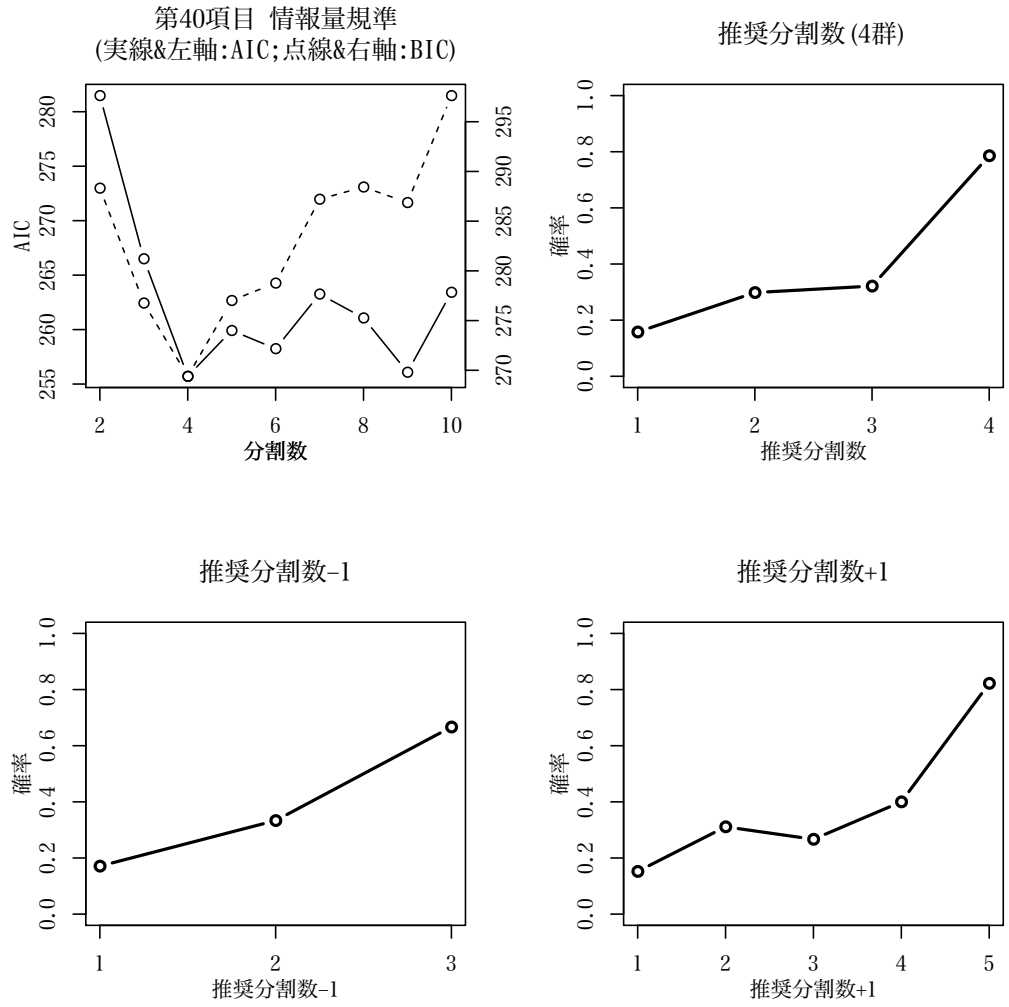


図 2.3 「学力テスト 1」第 40 項目における項目別推奨分割数

図 2.4 には第 45 項目において群数選択法を適用したときの項目特性図を示した。この項目では 4 群への分割が推奨された。5 群への分割は 4 群分割と表現が大変似通っている。ここでは各群の所属人数が比較的多くなり，より安定的な項目特性の表現を望むことが可能な 4 群への分割が推奨されたといえる。また 3 群での分割は表現が直線的になっており，図によって表現される項目特性

が異なる。このような場合に提案手法ではAICに基づいて4群分割を採用することが可能である。

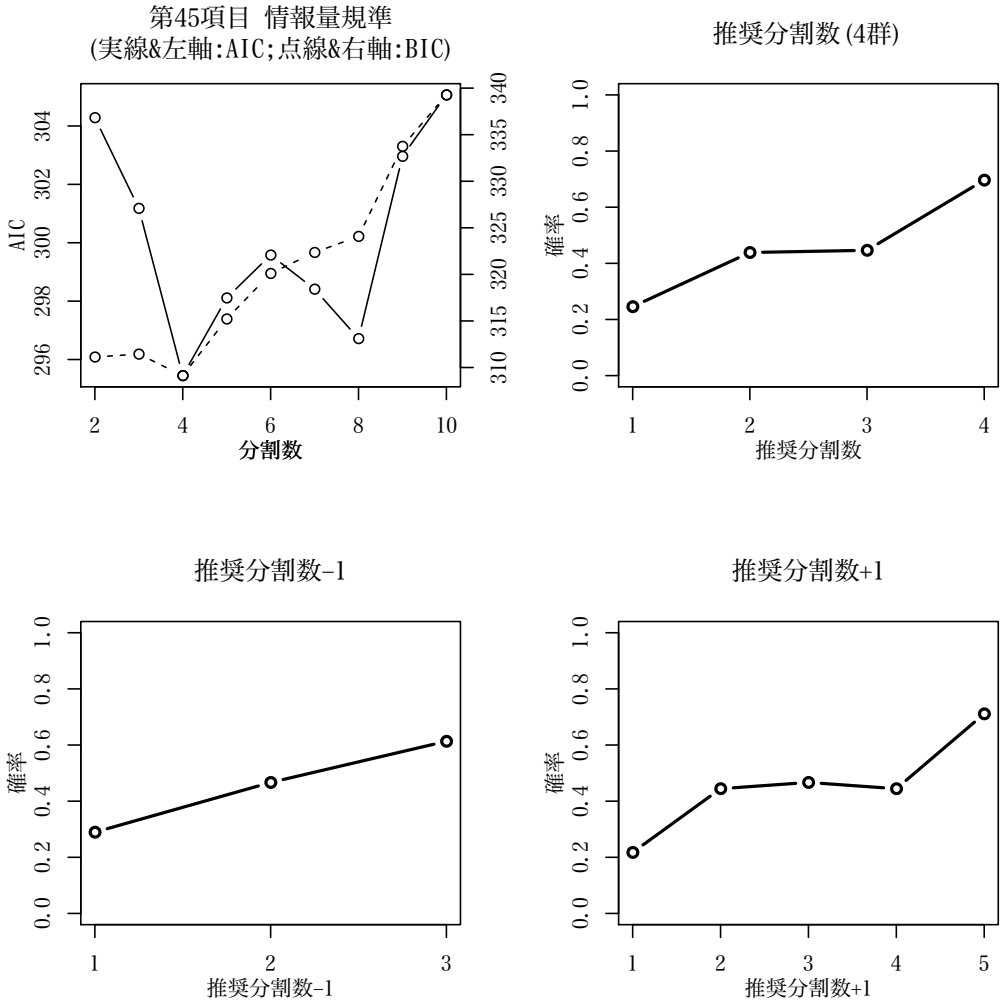


図2.4 「学力テスト1」第45項目における項目別推奨分割数

図2.5および図2.6に第38項目と第17項目の，群数選択法を適用した場合の項目特性図を示した。推奨群数はそれぞれ8群，10群となり，経験的に用いられる5群分割よりも多くなった。項目特性図を観察すると，図2.5では，第1群から第2群へと，一旦上昇した正答率が中程度の群においては下がり，上位群になるに連れて再度，上昇傾向に転じることが見てとれる。また図2.6においても同様に正答率が上下する様子が窺える。AICによって8群以上の群数が推奨されるような場合には，図2.5，図2.6において確認されるように，比較的複雑な項目特性を表現するために，推奨群数が多くなったものと推察される。

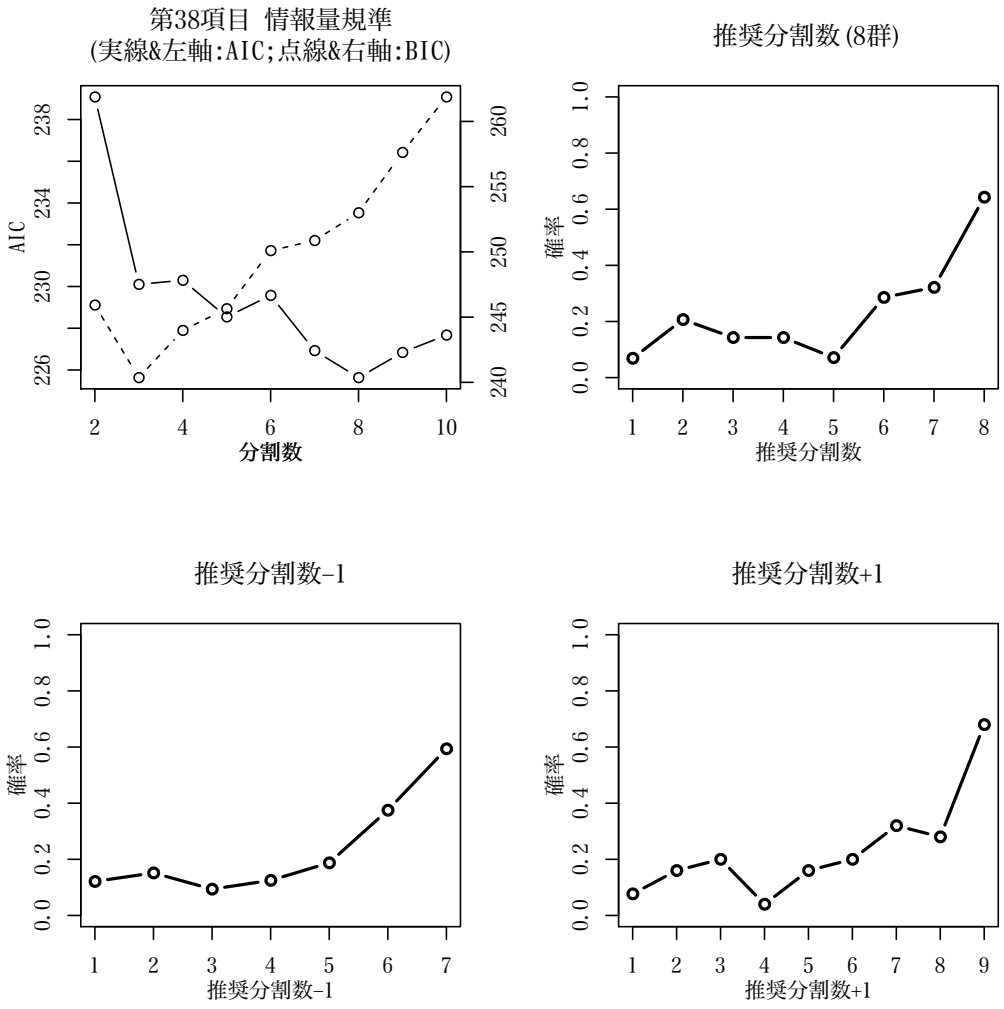


図 2.5 「学力テスト 1」第 38 項目における項目別推奨分割数

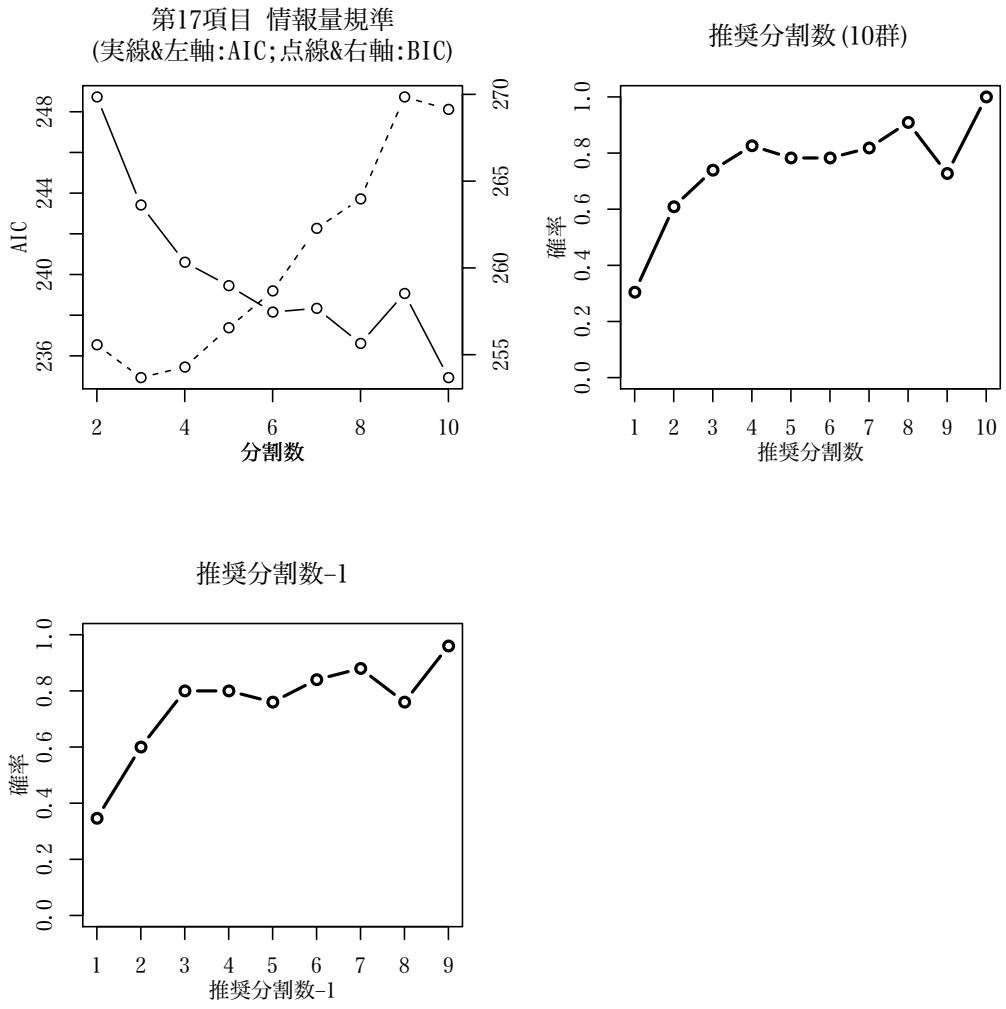


図 2.6 「学力テスト 1」第 17 項目における項目別推奨分割数

上記の適用例を通じて，AIC は分割数によって同一項目に関する項目特性図の表現が異なる場合に，群数選択基準を提供することが示された。また，不必要なまでに群数を増やすことで項目特性図の表現の安定性を失うことが避けられる可能性も示唆された。一方で，項目特性が複雑な形状を示す場合には，比較的受験者数が少数である場合にも，推奨群数が増える傾向にあることも示唆された。

2.6.1　全体データと抽出データの項目特性図の相似性

ここでは，あるデータセットから抽出されたデータを用い，提案手法（AIC）に基づいて作成された項目特性図が，抽出元のデータセット全体の項目特性を近似していることが示される。

比較用データとして日本経済新聞社と日本経済研究センターが主催する日経経済知力テスト (Test of Economic Sense and Thinking; 日経 TEST) の反応パタンデータ（項目数 100，受験者 5289 人）を用いた。5289 人という受験者数では，提案手法（AIC）を用いた場合，ほとんどの項目において 5 群以上が推奨された。ここでは 5289 人を全体データとし，この中から 400 人分の反応パタンを非復元無作為抽出し，抽出データを作成した。この抽出データについて提案手法を用いて項目特性図を作成し，当該項目特性図と全体データを 10 群（各群に属する受験者は 530 人弱）に分割した場合の項目特性図について比較した。全体データは十分に人数が多いため，10 群に分割しても表現は安定することが予想される。

抽出データおよび全体データの 100 項目に提案手法を適用した結果，AIC によって推奨された分割数の頻度を表 2.10 に示した。表 2.10 から 400 人では群数は減らした方が良く，5000 人強の受験者を擁する全体の場合は，経験的に 5 群に分割することや 10 群まで分割しても特性の表現が安定する傾向にあることが示唆された。

表 2.10 抽出・全体データにおける AIC 推奨分割数の度数

群数	2	3	4	5	6	7	8	9	10
抽出	11	12	18	20	11	11	6	7	4
全体	1	3	0	2	5	11	15	33	30

図 2.7 に「日経 TEST データ」第 22 項目における抽出データの項目特性図と全体データの項目特性図について示した。図 2.7 内，左上の図は分割数ごとの AIC の推移を示している。ここで実線および左縦軸は抽出データに関する AIC を，点線および右縦軸は全体データに関する AIC の推移をそれぞれ示している[1]。右上の図は抽出データにおける推奨分割数（AIC）による項目特性図を，左下は抽出データを 10 群に分割した場合の項目特性図を，右下は全体データを 10 群に分割した際の項目特性図をそれぞれ示している。

[1] ここで点線が示す意味が「学力テスト 1」への適用例とは異なっていることに注意してほしい。

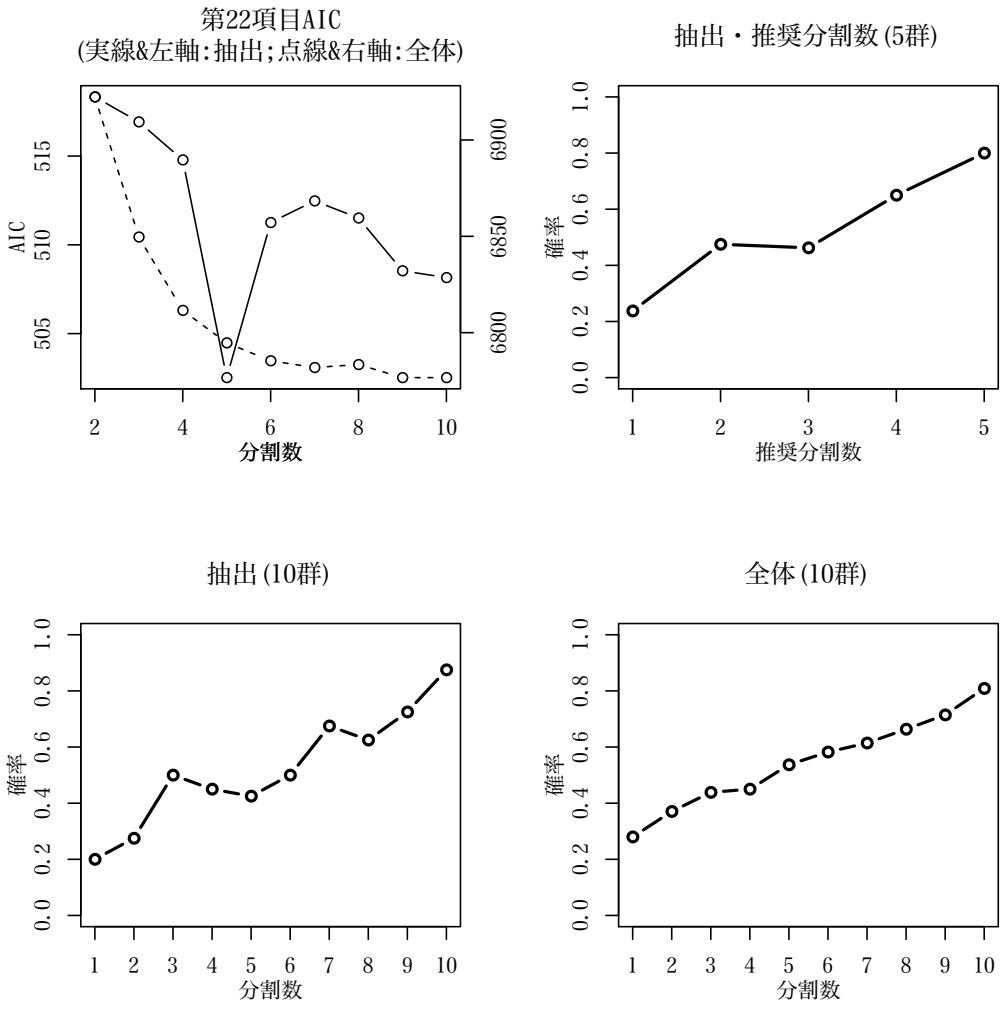


図 2.7 「日経 TEST データ」第 22 項目における項目特性図の相似性

第 22 項目の場合，直線の表現や，最下位群および最上位群における正答確率に注目すると，AIC による推奨分割数を用いた項目特性図は抽出（10 群）よりも全体データの項目特性図に似通っていると言える。よって第 22 項目における推奨分割数は理に適っていると言えるだろう。

図 2.8 に第 33 項目への抽出および全体データの項目特性図を示した。抽出データでの推奨分割数は 3 群となった。3 群分割では曲線表現による細やかな項目特性の考察を行うためには不十分ではあるものの，抽出データの下位，中位，上位群の関係と，全体データの特性傾向は近似しているといえる。第 33 項目においては，群数を減らして，安定した項目特性図の表現を得つつ，全体データに対応した傾向を知ることが可能な，3 群分割とすることにも利点があるといえ

るだろう。

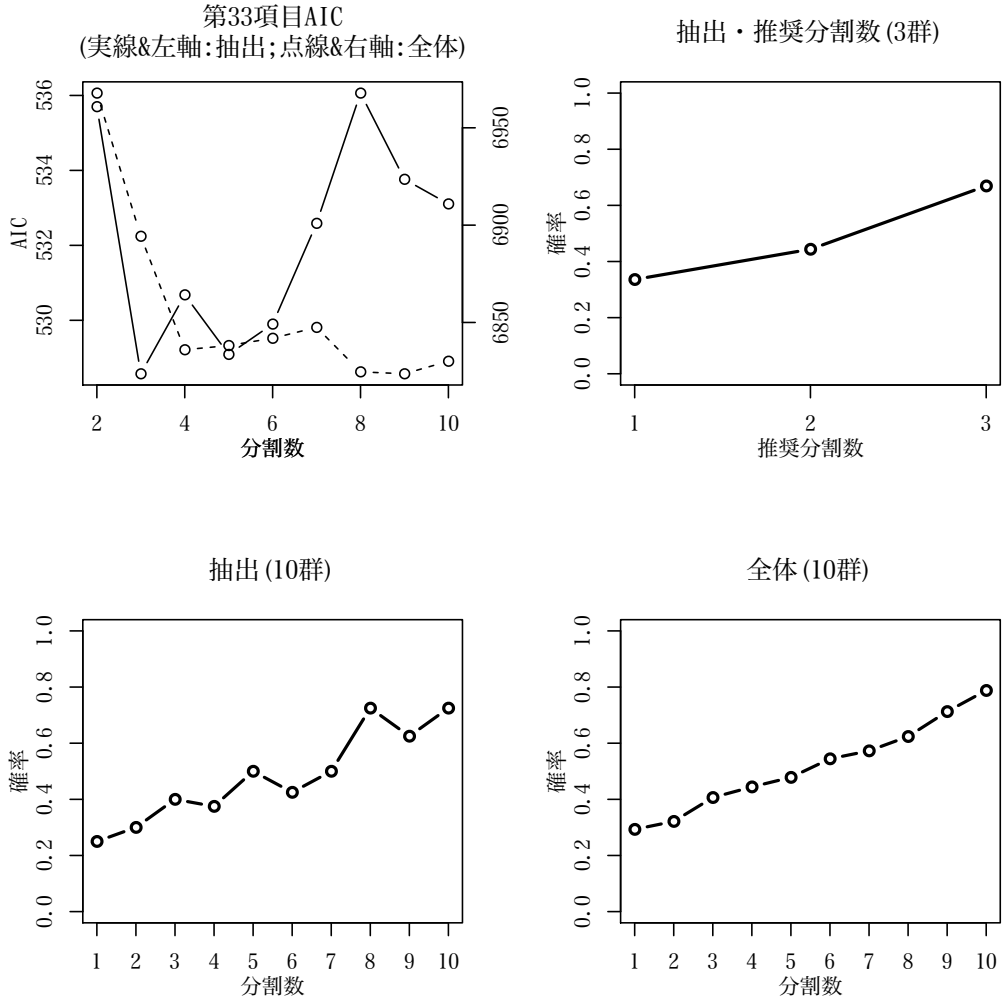


図 2.8 「日経 TEST データ」第 33 項目における項目特性図の相似性

図 2.9 に第 81 項目における抽出および全体データの項目特性図を示した。第 81 項目の場合，抽出データ推奨分割数，抽出（10 群），全体（10 群）で共通傾向にあるものの，中位以上の群に関する表現では，抽出データ推奨分割数と全体（10 群）で似通っている。また，推奨分割数は一般的に利用される 5 群を超過し，7 群まで増やされている。シミュレーションを通じて示唆されたように，受験者が少ない場合，提案手法の推奨分割数は減少傾向にある。しかし，第 81 項目への適用で 7 群分割が示唆されたことにより，提案手法は群を減らして表現を安定させるだけではなく，ときには表現を安定させるよりも，群数が増えてもデータ全体の構造に近い特性の表現を得ることが可能となる分割数を推奨

することが示唆された。「学力テスト 1」への適用例（表 2.8）において，5 群以上の群が推奨される項目があったことは，こうした理由があったものと推察される。

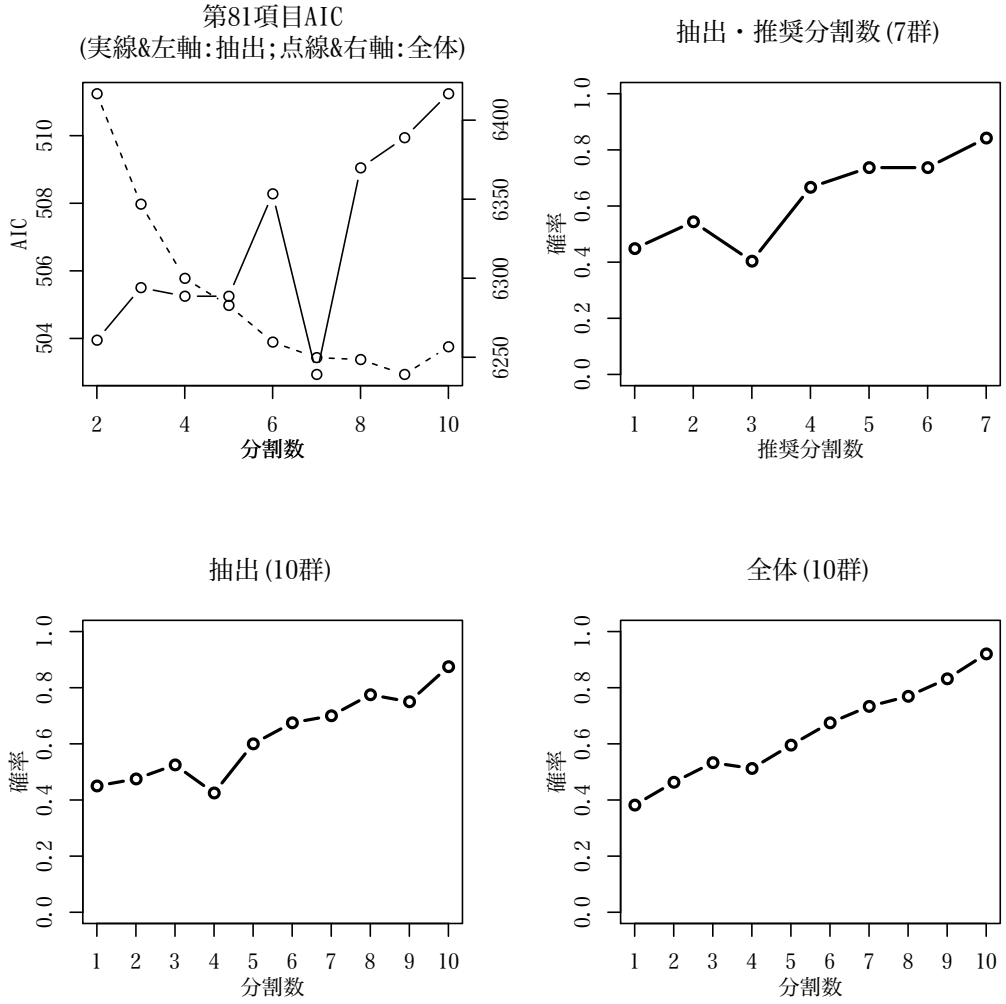


図 2.9 「日経 TEST データ」第 81 項目における項目特性図の相似性

上記の「日経 TEST データ」項目群に対する考察から，抽出データについて，提案手法を用いて項目特性図を作図した場合，当該項目特性図は全体データにおける群数を増やした場合の滑らかな項目特性図における特性の表現を近似する傾向にあることが示唆される結果となった。

2.7　研究 I 結論

情報量規準を用いた項目特性図の群数選択法は，一般に本試験よりも受験者数が遥かに少なくなる予備試験における項目分析を行う際に特に有用であると結論づけられる。また，群数の変更によって項目特性図の表現が変わり，項目特性の考察に影響し得る場合にも，情報量規準を傍証として，群数を選択することが可能である。

シミュレーションにおいて確認された各分割法と情報量規準の特徴をまとめる。分割法を等人数，かつ情報量規準を AIC とした場合，真の群数を的中させることが可能であった。受験者数が 200 人以下，真の群数 6 群以上の場合，真の群数よりも少なくなる傾向にあった。等人数かつ BIC では，受験者人数が少なく，真の群数が多い場合，真の群数と同じ値を推奨する傾向は低いことが認められた。等間隔かつ AIC では，200 名以上の受験者がいる場合，群数が過分に推奨され得ることに注意が必要である。等間隔かつ BIC では，受験者が少ない場合，大きな真の群数と同じ値を推奨する傾向が低いことは等人数かつ BIC の場合と共通の傾向にあるものの，等人数法よりもより多くの受験者数が必要とされる場合がある。

実際の適用場面では，情報量規準によって推奨される群数が 2 群や 3 群となる場合がある。受験者数が少ない場合には，群数を減らし，2 群や 3 群分割とすることで，安定した項目の性質を考察することが可能となる。しかし，2 群や 3 群ではきめ細やかな考察には力不足であるため，より詳細な性質を調べるために 4 群以上で項目特性図を作図することも考えられる。このような対処は経験的には重要な示唆を得られる可能性がある。ただし，情報量規準による推奨分割数以上の群数で描いた項目特性図の表現を，項目の絶対的な性質と見なすことには慎重な検討を要するだろう。

一方，受験者数が少ない場合には，滑らかな項目特性図の表現を得たくとも，項目特性図の安定性の見地から，経験的に 6 群以上の群数分割を行うことが躊躇われる場合があるだろう。情報量規準による群数選択法では，シミュレーションや適用例を通じて確認したように，受験者数が多くない場合であっても，6 群以上での分割が推奨され得るため，研究 I によって提案した方法 I で算出した情報量規準を，6 群以上で受験者を分割する際の傍証とすることが可能である。

項目特性図の作成において受験者を 5 群に分けるという経験的判断は，これまでの実務場面において効果を上げている。特に受験者が非常に多いテストに関しては情報量規準を用いた群数選択法の射程にはなく，経験的に 5 群に分割することで十分に項目特性の考察という目的を達成可能である。

本提案手法はテストにおける受験者数が少ない場合に特に効果を発揮する。例えば受験者数 200 名程度の予備試験と 5000 名程度の本試験があるような場合，予備試験の項目特性図に関しては提案手法を用いて作成することが有効であろう。本手法は，項目特性図作成における分割数の決定に傍証を与えることが可

能となる。特に同一項目において分割数ごとに表現される項目の特性が大きく異なるような場合には，分割数を選択する際の指針とすることが可能である。

第3章　誤答分析における項目特性図の作成方法（研究 II）

3.1　誤答分析における項目特性図の問題点

項目特性図の作図法　ここで，2 値以上の多肢選択式項目において，2 値に整理しない場合の項目特性図の作図法を概説する。

まず，テストにおける受験者の合計得点を昇順に並べ，任意の数の群（経験的に 5 群とされる）に分割する。次に分割した各群の，各項目のカテゴリごとの選択率を計算し，横軸に群，縦軸に確率を割り当てたグラフに選択率をプロットする。最後に各プロットを直線で結ぶ。

本章では正答選択肢の特性曲線を正答曲線，誤答選択肢の特性曲線を誤答曲線と呼称する。

3.1.1　正答分析

項目特性図を用いた正答率に関する分析は，項目の測定性能を調べるための基本的な項目分析の手法である。ある項目が受験者の実力が高い程，正答し易くなるという性質は，当該項目が適切に（分析者が想定している）受験者の実力を測定できているか，ということと密接な関係にある。項目特性図においては，高得点群である程，正答率が高い項目でなければ，一般に受験者能力の測定には性能不足，あるいは不適当な項目という判断を下すことになる。

3.1.2　誤答分析

誤答傾向に興味がある場合は，項目特性図上に正答曲線に加えて誤答曲線も描き，誤答分析を行うことが可能である。誤答分析も正答分析と同様に重要な分析であり（吉村, 2009; 豊田, 2012），正答分析によって項目の測定性能が担保された後は，誤答分析を行うことで，より詳細に，各選択肢と具体的な受験者群の特性を分析することが可能である。受験者の反応が複数のカテゴリ反応として整理される項目の場合，受験者がどの部分に躓き易かったのか，ということが誤答選択肢の選択確率という形で明示化される。Educational Testing Service

(1963) は受験者を成績ごとに 5 群に分割し，各群における選択肢別の選択者数を表に表す分析方法について論じている. また，赤根ら (2006) は選択肢別の選択率が，誤答が多い場合に受験者の理解度や項目の適否を検討することに有用であることを指摘している。選択率を視覚化した項目特性図もまた，項目の性質や，受験者の傾向を知るために有用な道具となる [1]。選ばれ易かった誤答選択肢が存在した場合，何故，当該誤答選択肢が選ばれたのか，受験者は如何なる誤った知識を適用してしまったのかを考察することで，テスト後の受験者の誤った知識の修正，更なる学習の発展に繋げていくことが期待できる。例えば龍岡・林 (2001) では，解答パターンを用いた誤認判断の方法を紹介している。また，麻柄 (2006) は種々の課題における誤った知識の修正方略について，具体的な事例を交えつつ，様々な側面から検討している。

誤答分析の結果，ある誤答選択肢は，他の誤答選択肢と比較して，特異に魅力的な選択肢として機能している等の知見が得られる。このような判断を下すためには，教科内容に関する知識を伴った高度な感性が要求される。感性に基づいて構成した仮説は，特性曲線に対して，以下に述べる操作を行うことで項目特性図に反映することができる。

例えば同様の方略，知識によって導かれる複数の誤答選択肢が存在している場合に，それらをまとめることが考えられる。あるいは，学習の特定の段階において誤用される方略，知識によって導かれる誤答に焦点を当てるため，それ以外の部分をまとめたい場合も想定可能である。これらの状況において，その妥当性が確認可能であれば，教科内容についての更なる理解の促進が期待できるであろう。

しかしながら，仮説を反映した項目特性図を解釈することは可能であっても，そこから得られた知見を検証する方法は未だ確立されていない。

項目特性図の他に，受験者個々人が各項目に如何なる反応を示したかを調べる道具として S-P 表（佐藤, 1998）がある。S-P 表も多肢選択形式の項目の誤答分析を行う上で有用な道具となる（佐藤, 1998, pp.51–57）。しかし S-P 表は受験者の人数や得点の範囲によって，大きな表サイズとなり得る。一方で，項目特性図は多くの受験者がおり，得点の幅が広範囲であっても視覚的な把握が容易である。また，項目特性図は分析対象項目以外の項目が多肢選択式以外の形式であっても，作成可能という利点が挙げられる。しかしながら，項目特性図における誤答選択肢の特性曲線は，迷わしとしての性質上，選択確率が似通い易く，結果的に図が煩雑化し易い。これでは項目特性図としての長所が生かされていない状況といえるだろう。煩雑化を避けるための簡単な方法としては，項目特性図における特性曲線の本数を減らすことが考えられる。例えば全く選ばれていない選択肢についてはそもそもプロットしないことや，一定の基準を設けて，基準以下の曲線は描画しないということが可能である。吉村 (2009) では

[1] 統計データの視覚化の重要性は山本・飯塚・藤野 (2013) によって指摘されている。

特性曲線によって図が煩雑化することを避けるため,（全群に渡って）選択確率10%未満のカテゴリについては表示しないという基準が示されている。しかし，どの程度の選択確率までが重要であるのかは，受験者数，テストの目的，項目の性質によって異なり得るため，一様な選択確率の基準を設けるのは困難な場合がある。

別の方法として，後述するヒューリスティックに基づいて選択肢間の平均選択率を用いることが考えられる。この方法では，すべての選択肢について考慮しつつ，特性曲線の本数を減らすことができる。本方法も項目特性図の視認性を確保し，誤答分析を行う上で有用となるだろう。

選択肢をまとめる際には，最終的に分析者の教科内容に関する知識と感性が重要となる。ただし項目特性図によっては，選択肢のまとめ方，すなわち仮説が複数考えられる場合がある。

項目特性に対する仮説が一意に定まらない場合や，分析者間の感性や仮説が対立する場合に，なぜ，当該の併合した，あるいは併合しなかった項目特性図を分析に用いるのかを統計的に根拠づける方法は広く知られていない。

本研究では，分析者が図 3.1(i) のような項目特性図の考察を通じて得た仮説を，項目特性図に反映する方法として，平均選択率を用いる。得られた仮説は単なる標本変動による見かけ上のものである場合や，他にも尤もらしい仮説（併合状況）が並立して得られる場合がある。このとき，情報量規準を用いたモデル比較によって，統計的な観点から仮説の採用を補強することが有効であろう。本論文では仮説に基づいて作成した併合，あるいは無併合の項目特性図をモデルと見做し，これらのモデルを比較するための情報量規準を算出する方法について提案する。副次的な効果として，図の視認性を確保することも可能となる。特性曲線の併合は項目ごとの仮説次第で様々な状況が考えられるものの，いくつかのヒューリスティックに従ってまとめることができる。以下では，誤答選択肢の特性曲線をまとめるヒューリスティックについて記述する。

誤答選択肢併合のためのヒューリスティック　ここで，誤答選択肢の特性曲線をまとめるヒューリスティックについて記述する。

まず，統合前のオリジナルの項目特性図を作図する。次に，描かれた折れ線から解釈される，出題内容やワーディングの特徴の可能性を見出す。

特定の誤答選択肢の，特異な形状から解釈される，出題内容やワーディングの特徴の可能性として，

H.1 最上位群の選択確率が高いなど，幹のワーディングの悪さから，当該選択肢にも正解の要素があった可能性

H.2 当該選択肢が一見して分かるほど的外れな内容であるために，選択候補に入らず，迷わしとして機能していなかった可能性

H.3 選択確率が高く，特定の能力レベルにおいて当該誤答選択肢が正答選択肢よりも魅力的に感じられた可能性

H.4 選択確率が高く，特定の能力レベルにおいて当該誤答選択肢が他の誤答選択肢よりも魅力的に感じられた可能性

H.5 複数の誤答選択肢が同様な折れ線を示し，互いに等確率で選ばれ，迷わしとして有効に機能している可能性

が考えられる。各誤答選択肢の特徴に加え，教科内容的な側面からも検討し，誤答選択肢をまとめる。

最後に，オリジナルの項目特性図やまとめ方の異なる複数の項目特性図に関して，見出された特徴の違いが標本誤差によるものか，それ以上のものであるのかを本章において提案した手法により確認する。

表 3.1 「学力テスト 1」第 7 問

自由面接による調査方法は

(a) 調査票に記載された質問に被調査者が回答する

(b) 政府，国際機関，マスコミ，企業，研究所が発表した統計資料・数字を 2 次的に比較検討する

(c) 調査者の質問に自由に回答した被調査者の発言内容をメモやレコーダーによって記録する

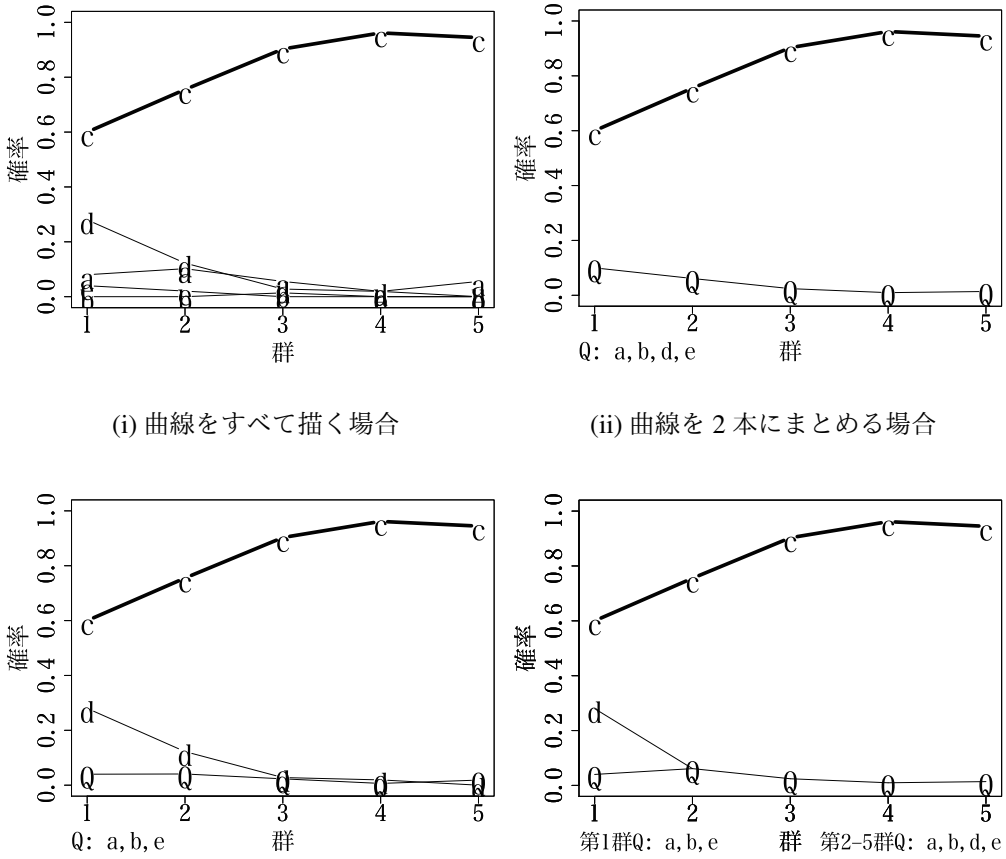
(d) 調査者自身が調査対象集団の活動に何らかの形で参加し，その活動を観察し，記録する

(e) 手紙・日記・自伝・新聞記事などに掲載された特定の社会事象の記録を内容的に分析し，調査者の興味の観点から結論を導く

正答 “c”

（豊田（2012, pp.235–238）より引用）

図 3.1 はある大学で実施された「統計的調査法」という講義の期末試験データ（豊田, 2012, p.113–114）である,「学力テスト 1」（受験者数 216 名（無効回答者除外前 226 名），項目数 50 個, 5 肢選択形式; 豊田, 2012, pp.235–238）における項目特性図の作成例である。誤答分析を行う際には，通常は図 3.1(i) のように，5 肢すべての選択肢に関して選択確率をプロットし，曲線を描画する。し



(i) 曲線をすべて描く場合

(ii) 曲線を 2 本にまとめる場合

(iii) 曲線を 3 本にまとめる場合

(iv) 第 1 群における特異な誤答傾向を保存したい場合

図 3.1 項目特性図（「学力テスト 1」・第 7 問）

かしながら，誤答曲線は互いに重なり合い，非常に煩雑である。この部分に関して，すべての選択肢について考慮しつつ，曲線の本数を減らしたい。教科内容的な知見が得られていたり，選択傾向が似通っていると判断される場合には，選択肢をまとめ，項目特性図上に描画する方法が考えられる。このまとめ方に関しても，検証を行いたい。ここでは誤答選択肢 “a”，“b”，“d”，“e”の選択確率はどれも似通っていると判断したものとする。そこで，すべての誤答選択肢の選択確率の平均値を新たにカテゴリ “Q”としてプロットした項目特性図が図 3.1(ii) である。これは，何れの誤答選択肢 “a”，“b”，“d”，“e”も同程度に選択され易いという仮定の状況を反映している。

しかしながら図 3.1(i) を観察すると，選択肢 “d”が，他の誤答選択肢とは特異に選択され易いという状況も想定され得る。この状況を表したのが図 3.1(iii) である。図 3.1(iii) では選択肢 “a”，“b”，“e”を統合し，平均値をプロットし，選択肢 “d”に関しては選択確率を元のままプロットしている。曲線は増えるものの，

5 肢すべての曲線を描くよりも視認性は高い。更に観察を進めると，特に第 1 群の受験者にとって，選択肢 “d”が特異に選択され易く，魅力ある誤答となっており，第 2 群以上の受験者に対しては，他の選択肢と似た傾向を示しているようである。図 3.1(iv) は，第 2 群から第 5 群までは 4 肢とも同程度に選択され易いと仮定した状況を反映している。図 3.1(iii) とは異なり，5 群すべてではなく，第 1 群のみ，誤答選択肢 “d”が特異に選択され易いという状況を表現している。

表 3.1 が第 7 問の項目内容であり，自由面接に関して正しいものを選択する問題である。正解は “c”であるが，第 1 群に配される受験者は特に選択肢 “d”が魅力的に感じられるようである。選択肢 “d”は参与観察に関する記述であるが，低得点群は調査における「面接」と「観察」を混同しているようである。当該受験者は両調査法の具体的な実施方法の違いについて，再度確認することが肝要であろう。

ここでは，ヒューリスティックの H.4 の可能性より，図 3.1(iv) の状況で誤答分析を行いたいものの，当該想定状況が他の項目特性図による想定状況よりも適切であることを示す根拠は広く見出されていない。どの選択肢に注目し，併合するのか，特異な特性曲線として描画するのかという判断は仮説の反映と見ることができるが，上記のようにこうした併合状況は複数考えられる。

本研究では，同一項目において，カテゴリ特性曲線の併合状況（モデル）が複数想定され得る場合に，分析者の仮定（ある特定のカテゴリ特性曲線の傾向が同等であると見做すことができるか）を確かめるための情報量規準による統計的指標を算出する方法を提案する。本提案手法を用いることで，効果的な誤答分析を行うための項目特性図を作成することが可能となる。本提案手法を用いることで，仮説を反映した項目特性図を統計的に比較し，採用の傍証とすることが可能となる。

3.2 項目特性図を用いた誤答分析の精緻化の提案（方法 II）

図 3.1 に示した 4 つの項目特性図は，受験者群の誤答状況に対する，並立する仮説を反映したモデルと捉えることができる。例えば分析者が誤答分析の結果，「誤答選択肢 “ d ” は特異に第 1 群に対して迷わしとして機能している」という仮説を得たときに，モデル（図 3.1(iv) ）を作成し，更なる項目分析に使用することが考えられる。しかしながら，前述した状況のように「選択肢 “ d ” は全群に渡って特異に魅力的である（図 3.1(iii) ）」や「選択肢 “ d ” の特異な見かけは標本変動であり迷わしとしての機能に誤答選択肢間の違いはない（図 3.1(ii)）」といった対立するモデルも考えられる。

もし，教科内容における先行研究や知識がある場合には，それらを傍証としてモデル図 1(iv) の採用を理由づけることも可能である。このとき，統計的にモ

デルを比較し，モデル採用の傍証として用いることができる統計的基準があれば便利である。

以下では，すべての群に渡って，特定の誤答選択肢が特異に機能している（その他の誤答選択肢は併合される）場合と，第 1 群のみで特異に誤答選択肢が機能する場合とに分けて，AIC と BIC を算出する方法を論じる。また後の節では具体的な項目に対する提案手法の適用例を示す。

3.2.1 すべてのカテゴリ特性曲線を描画する項目の場合

いま，J 個の項目によって構成されるテストに関して，第 $j\,(=1,2,\cdots,J)$ 番目の項目における受験者の選択結果が，C_j 個のカテゴリに整理されるものとする。例えば 5 肢選択形式の項目に対する受験結果を選択肢ごとに整理する場合，$C_j=5$ となる。また，無効解答を含めて整理する場合には $C_j=6$ となる。

ここで，項目 j に注目する。記号を簡素化するために添え字 j による区別を省略し，第 $g(=1,2,\cdots,G)$ 群[2] に属する $i\,(=1,2,\cdots,n_g)$ 番目の受験者の反応 x_{gi} が，$c\,(=1,2,\cdots,C)$ 番目のカテゴリとして観測される確率を p_{gc} とする。

このような受験者が n_g 人いるとき，複数の受験者の反応を並べたベクトル

$$\boldsymbol{x}_g=(x_{g1},\cdots,x_{gi},\cdots,x_{gn_g})'$$

を得る確率は，各カテゴリの反応が独立であるものとして，

$$\Pr(\boldsymbol{x}_g|\boldsymbol{p}_g)=\prod_{c=1}^{C}p_{gc}^{n_{gc}} \tag{3.1}$$

となる。ここで $\boldsymbol{p}_g$ は

$$\boldsymbol{p}_g=(p_{g1},\cdots,p_{gc},\cdots,p_{gC})'$$

であるような母数ベクトルであり，n_{gc} はカテゴリ c を選択した人数である。さらにベクトル

$$\boldsymbol{x}_1,\cdots,\boldsymbol{x}_g,\cdots,\boldsymbol{x}_G$$

を観測する確率は

$$\Pr(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_g,\cdots,\boldsymbol{x}_G|\boldsymbol{p}_1,\cdots,\boldsymbol{p}_g,\cdots,\boldsymbol{p}_G)=\prod_{g=1}^{G}\prod_{c=1}^{C}p_{gc}^{n_{gc}}$$

[2]本節，及び次節では群数を g によって任意の数として表現しているが，本研究では誤答分析が焦点であるため，適用例ではすべて，一般的に用いられる $G=5$ として作成する。

である。項目を表す添え字 j による区別を導入すると，(3.1) 式は

$$\Pr(\boldsymbol{x}_{jg}|\boldsymbol{p}_{jg}) = \prod_{c=1}^{C_j} p_{jgc}^{n_{jgc}} \tag{3.2}$$

と再表現される。項目 j における母比率の最尤推定量は

$$\hat{p}_{jgc} = \frac{n_{jgc}}{n_{jg}}$$

であり，これを母数ベクトル $\boldsymbol{p}_{jg}$ の要素として配することで，最大対数尤度を

$$\begin{aligned} \log l_{\max} &= \log l(\hat{\boldsymbol{p}}_{j1}, \cdots, \hat{\boldsymbol{p}}_{jg}, \cdots, \hat{\boldsymbol{p}}_{jG}) \\ &= \sum_{g=1}^{G} \sum_{c=1}^{C_j} n_{jgc} \times \log \hat{p}_{jgc} \end{aligned} \tag{3.3}$$

と導くことができる。ここで

$$\sum_{c=1}^{C_j} \hat{p}_{jgc} = 1$$

という制約があるため，

$$\hat{\boldsymbol{P}}_j = (\hat{\boldsymbol{p}}_{j1}, \cdots, \hat{\boldsymbol{p}}_{jg}, \cdots, \hat{\boldsymbol{p}}_{jG})$$

に含まれる自由度は $G \times (C_j - 1)$ 個となる。

(3.3) 式を用いて，項目 j に関して，群数 G で描く項目特性図の AIC 及び BIC，

$$\mathrm{AIC}_j = -2 \log l_{\max} + 2G(C_j - 1) \tag{3.4}$$

$$\mathrm{BIC}_j = -2 \log l_{\max} + G(C_j - 1) \log(n_j) \tag{3.5}$$

を算出可能である。ここで

$$n_j = \sum_{g=1}^{G} \sum_{c=1}^{C_j} n_{jgc}$$

である。モデル間で情報量規準の値が相対的に小さいモデルが，データへの当てはまりが良いと判断することが可能である（坂元, 石黒, 北川, 1983）。

3.2.2 特定のカテゴリが併合可能な項目の場合

いま，テストを構成する，ある項目 $j(j = 1, 2, \cdots, J)$ について，当該項目の回答が C_j 個の選択肢カテゴリに整理される状況を想定する。ここで，選択確率

が所与の下での各選択肢が観測される確率は (3.2) 式と同様に表現される。添字 j と g による区別はここでは非本質的であるため，表記を簡潔にするために以後，(3.2) 式について，項目に関する添字 j と群に関する添字 g による区別を省略し，

$$\Pr(\boldsymbol{x}|\boldsymbol{p}) = \prod_{c=1}^{C} p_c^{n_c} \tag{3.6}$$

と再表現する。いま C 個のカテゴリ中，任意の $M(M \leq C)$ 個のカテゴリを含む集合を定義し，当該集合内のカテゴリ間の選択確率が等しく p_Q として仮定されるものとする。つまり $p_{C-M+1} = p_{C-M+2} = \cdots = p_C = p_Q$ である [3]。すると (3.6) 式は

$$\Pr(\boldsymbol{x}|\boldsymbol{p}) = \prod_{c=1}^{C-M} p_c^{n_c} \times p_Q^{n_Q} \tag{3.7}$$

と再表現される [4]。ただし

$$\begin{aligned} p_Q =& \frac{(1 - \sum_{c=1}^{C-M} p_c)}{M} \\ n_Q =& \sum_{c=(C-M+1)}^{C} n_c, \quad \sum_{c=1}^{C-M} p_c + M p_Q = 1 \end{aligned} \tag{3.8}$$

である。(3.7) 式に (3.8) 式を代入し対数尤度を求めると，

$$\log L = \sum_{c=1}^{C-M} n_c \log p_c + n_Q \log \left(\frac{1 - \sum_{c=1}^{C-M} p_c}{M} \right)$$

となり，これを p_c で偏微分し，右辺を 0 と置くと

$$\frac{\partial \log L}{\partial p_c} = \frac{n_c}{p_c} - \frac{n_Q}{1 - \sum_{c=1}^{C-M} p_c} = 0$$

である。ここで (3.8) 式を利用して変形し，その結果を定数 T と置くと，

$$\frac{n_c}{p_c} = \frac{n_Q}{M p_Q} = T \tag{3.9}$$

[3] 本節では選択確率が等しいと仮定されないカテゴリについて $c = 1, \cdots, M$ とし，選択確率が等しいと仮定されるカテゴリを $c = C - M + 1, \cdots, C$ とする。

[4] (3.6) 式について，選択確率が等しいとは仮定されないカテゴリと，選択確率が互いに等しいと仮定されるカテゴリの積として再表現している。

となる。ここで，

$$
\begin{aligned}
n &= \sum_{c=1}^{C-M} n_c + n_Q \\
&= \sum_{c=1}^{C-M} \frac{n_c}{p_c} p_c + \frac{n_Q}{Mp_Q} Mp_Q \\
&= T(\sum_{c=1}^{C-M} p_c + Mp_Q) \\
&= T
\end{aligned} \tag{3.10}
$$

であるため，定数 $T = n$ である。添字 j と g による区別を再導入すると，(3.9) 式より最尤推定量は

$$\hat{p}_{jgc} = \frac{n_{jgc}}{n_{jg}}, \ (c = 1, \cdots, C_j - M_{jg}), \tag{3.11}$$

$$\hat{p}_{jgQ} = \frac{n_{jgQ}}{M_{jg} n_{jg}} \tag{3.12}$$

と導かれる。(3.11) 式および (3.12) 式を用いることで，最大対数尤度は

$$
\begin{aligned}
\log l_{\max} &= \log l(\hat{p}_{jg1}, \cdots, \hat{p}_{jgc}, \cdots, \hat{p}_{(C_j - M_{jg})}, \hat{p}_{jgQ}) \\
&= \sum_{g=1}^{G} \left(\sum_{c=1}^{C_j - M_{jg}} n_{jgc} \log \hat{p}_{jgc} + n_{jgQ} \log \hat{p}_{jgQ} \right)
\end{aligned} \tag{3.13}
$$

と構成される。各項目の AIC 及び BIC は (3.13) 式を用いて，

$$\mathrm{AIC}_j = -2 \log l_{\max} + 2 \sum_{g=1}^{G} (C_j - M_{jg}) \tag{3.14}$$

$$\mathrm{BIC}_j = -2 \log l_{\max} + \sum_{g=1}^{G} (C_j - M_{jg}) \times \log(n_j) \tag{3.15}$$

となり，算出可能となる。

3.3 実データを用いた適用例（適用例 II）

本節では前節で提案した方法を，実際のテストデータにおける項目特性図作成に適用し，その結果と解釈を示す。

適用例として用いた項目は，そのいずれもが誤答分析から，複数の仮説が立てられ，結果としてヒューリスティックに基づいて複数の項目特性図が作成で

きる。なお適用例で示した項目の何れにおいてもヒューリスティック H.1 から H.5 の範囲で扱うことが可能であった。これら多様なパターンを示す項目特性図，すなわちモデルに対して提案手法を適用し，モデル比較を情報量規準に基づいて行えるようになることを示す。

3.3.1 統計的調査法テスト（学力テスト 1）データへの適用

表 3.2 「学力テスト 1」第 7 問・各項目特性図の情報量規準

モデル	図 3.1(i)	図 3.1(ii)	図 3.1(iii)	図 3.1(iv)
群毎のカテゴリ数	5,5,5,5,5	2,2,2,2,2	3,3,3,3,3	3,2,2,2,2
パラメタ数	20	5	10	6
対数尤度	−106.520	−123.766	−116.983	−119.307
AIC	253.04	257.532	253.966	250.615
BIC	320.545	274.408	287.718	270.866

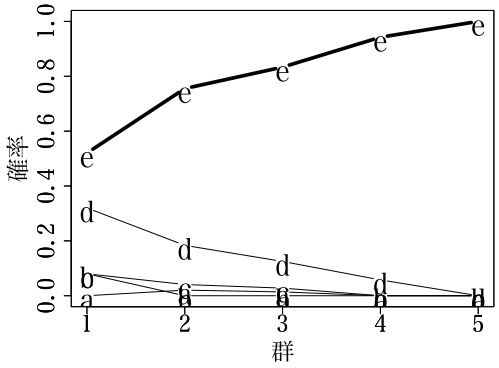
前節で述べた方法を用いることで，表 3.2 のような「学力テスト 1」第 7 問の各項目特性図の情報量規準が算出可能となる。情報量規準によると図 3.1(iv) のモデルが最も低い値を示していた。分析者の観察から得られた仮説が並立する状況において，本手法は選択肢“d”の特異性に注目して誤答分析を進めていく際の傍証を与えることが可能となる。

表 3.2 中のパラメタ数は，図 3.1(i) から図 3.1(iii) については (3.4) 式，(3.5) 式の第 2 項における $G(C_j-1)$ から算出される。例えば図 3.1(i) の場合，$G=5, C_j=5$ であるため，$G(C_j-1)=20$ となる。一方で一部の特性曲線が併合される図 1(iv) の場合は (3.14) 式，(3.15) 式の第 2 項における $\sum_{g=1}^{G}(C_j-M_{jg})$ より，群別カテゴリ数に基づいて，$(5-3)+3(5-4)=6$ と算出される。

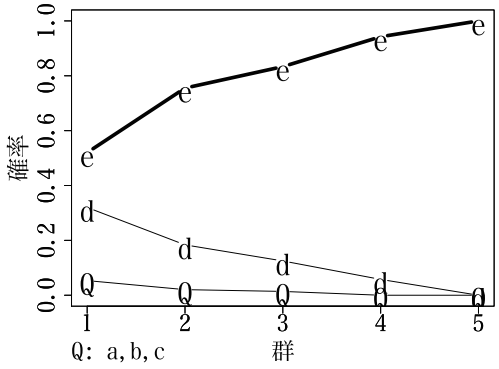
表 3.3 「学力テスト 1」第 8 問

興味の対象となる年齢・世代の最も若年層から標本を選び，その標本を時間とともに追跡する調査は
(a) 発見型調査　(b) 統制調査　(c) 断面的調査　(d) 横断的調査　(e) 縦断的調査
（正答）“e”

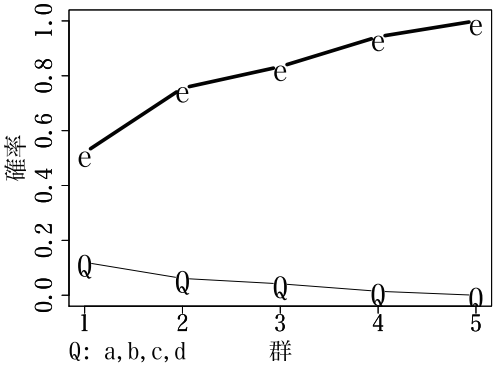
（豊田 (2012, pp.235–238) より引用）



(i) 曲線をすべて描く場合



(ii) 曲線が 3 本にまとめられる場合



(iii) 曲線が 2 本にまとめられる場合

図 3.2 項目特性図（「学力テスト 1」・第 8 問）

表 3.3 に第 8 問の項目内容を，図 3.2 に項目特性図を，表 3.4 に提案手法の適用結果をそれぞれ示した。第 8 問の正答は “e”「縦断的調査」である。図 3.2(i) から，すべての群に渡り，選択肢 “d”「横断的調査」が他の誤答選択肢よりも特異に選択される傾向にあることが見受けられる。ヒューリスティックの H.4 の状況が想定可能である。各モデルにおける情報量規準を比較すると，図 3.2(ii) が推奨される結果となり，上記の観察が支持された。

選択肢 “d”を選択した受験者は，調査名称を誤って記憶する，縦断と横断の区別が曖昧となっている状態であることが推察される。選択肢 “d”が第 4 群においても選択される傾向にあることから，比較的高特性の受験者であっても横断と縦断を取り違えている可能性が示唆された。誤答者は横断と縦断の概念的な違いと，言葉による定義の対応関係を再確認する必要があるだろう。

表 3.5 に第 30 問の項目内容を，図 3.3 に項目特性図を，表 3.6 に提案手法の適用結果をそれぞれ示した。第 30 問における選択肢 “a”から “d”はその何れもが調査における回答が不完全な状態で得られている状況である。そのため，“e”は一

表 3.4 「学力テスト 1」第 8 問・各項目特性図の情報量規準

モデル	図 3.2(i)	図 3.2(ii)	図 3.2(iii)
群毎のカテゴリ数	5,5,5,5,5	3,3,3,3,3	2,2,2,2,2
パラメタ数	20	10	5
対数尤度	−116.221	−120.615	−142.561
AIC	272.442	261.231	295.121
BIC	339.948	294.984	311.998

表 3.5 「学力テスト 1」第 30 問

調査票を回収した後に行う編集（エディティング）で修正を要するのは

(a) 白紙，実質的な回答のない調査票
(b) 回答欄ではなく選択肢にチェックしている回答
(c) ろ過的質問に矛盾している回答
(d) 存在しない選択番号あるいは数値の回答
(e) 上の 4 つは修正ではなく無効回答とする
（正答）“b”

（豊田 (2012, pp.235–238) より引用）

見して尤もらしい選択肢であり，実際に多くの受験者が誤答している。しかしながら，“b”のみは調査者側で修正を施しても不適切とは見なされない。ヒューリスティックの H.2，および H.3 の状況が想定される。

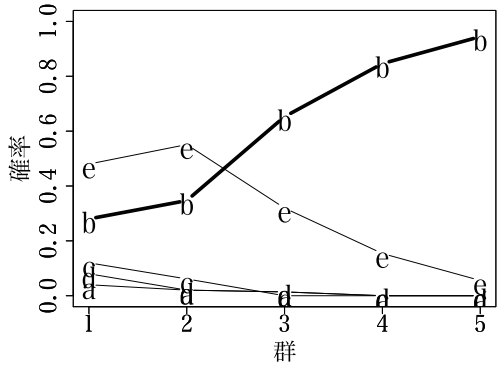
第 30 問では，表 3.6 より，図 3.3(ii) のモデルが支持された。非常に魅力的な誤答が存在する問題ではあるが，第 2 群以降は高得点群になるほど“e”の選択率も減少傾向にあることから，過度な迷わしとはならずに，どれだけテストに対して準備していたかを測ることが可能な項目であったと言えるだろう。

3.3.2 PISA データへの適用

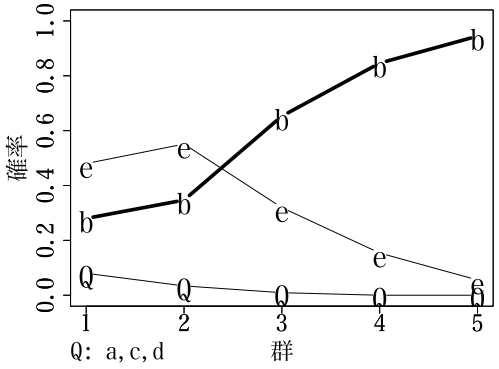
本節では「OECD[5] 生徒の学習到達度調査」(Programme for International Student Assessment, PISA; 国立教育政策研究所, 2004) の 2003 年に実施されたテストデータ用い，適用例を示す。

国立教育政策研究所 (2004) は，PISA 調査を「義務教育修了段階の 15 歳児が

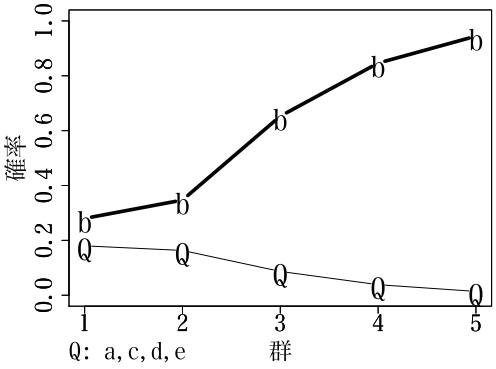
[5]Organisation for Economic Co-operation and Development（経済協力開発機構）



(i) 曲線をすべて描く場合



(ii) 曲線が 3 本にまとめられる場合



(iii) 曲線が 2 本にまとめられる場合

図 3.3 項目特性図（「学力テスト 1」・第 30 問）

持っている知識や技能を，実生活の様々な場面で直面する課題にどの程度活用できるかどうかを評価（特定の学校カリキュラムがどれだけ習得されているかを見るものではない）」する調査であると述べている。

PISA は国際的な調査であるが，本研究においては日本の調査結果のみを抽出して用いることとした。2003 年版 PISA の日本における受験者は 4707 人であった。また，PISA 調査は読解力，数学的リテラシー，科学的リテラシーを主要 3 分野としているが，2003 年の調査では数学的リテラシーが重点的調査対象とされていた。本研究においては数学的リテラシーの調査に用いられた 90 項目の内，解答形式が多肢選択形式であり，なおかつ項目内容が公開されている項目から 3 項目を適用対象とした。調査は 14 種類のブックレット形式で実施され，各ブックレットに含まれる項目，及び項目数は異なる。よって以下に示す適用例は項目間で受験者数と当該項目が含まれるブックレットにおける満点が異なる。表 3.7 に，情報量規準によるモデル比較を行う項目と，各項目が含まれるブックレット，及びその受験者，満点（配点を 1 点とした）を示す。

表 3.6 「学力テスト 1」第 30 問・各項目特性図の情報量規準

モデル	図 3.3(i)	図 3.3(ii)	図 3.3(iii)
群毎のカテゴリ数	5,5,5,5,5	3,3,3,3,3	2,2,2,2,2
パラメタ数	20	10	5
対数尤度	−163.634	−165.710	−235.582
AIC	367.268	351.420	481.163
BIC	434.774	385.173	498.04

表 3.7 適用項目が含まれるブックレットの受験者

項目	49	56	60
ブックレット	13	3	6
受験者	351	351	328
満点	23	35	36

結果と解釈

本項では PISA データへの適用結果，及び解釈を示す。PISA では数学的な内容に取り組むために必要な能力，認知的活動を「能力クラスター」という分類を用いて説明している。ここで分類名と定義を国立教育政策研究所 (2004, p.33) より引用する。[**再現クラスター**]:「比較的よく見慣れた，練習された知識の再現を主に要する問題を解く能力」。[**関連付けクラスター**]:「再現クラスターの上に位置づくもので，やや見慣れた場面，または，見慣れた場面から拡張され発展された場面において，手順がそれほど決まりきってはいない問題を解く能力」。[**熟考クラスター**]:「関連付けクラスターのさらに上に位置づくもので，洞察，反省的思考，関連する数学を見つけ出す創造性，解を生み出すために関連する知識を結びつける能力」。

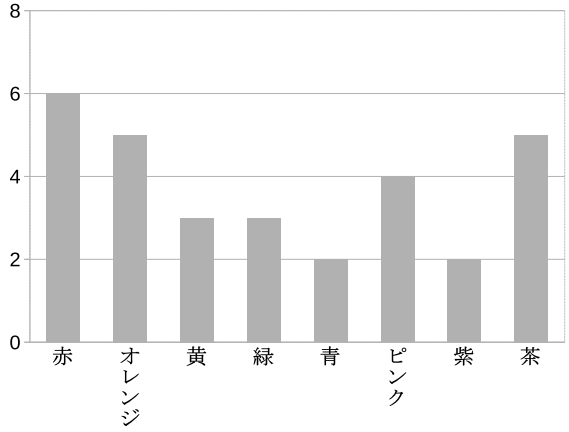
表 3.8 には 2003 年 PISA 調査において用いられた「いろいろな色のキャンディ」（以下「キャンディ」, ブックレット 13，受験者 351 名，23 点満点）と題された項目の内容を示した。

項目「キャンディ」の正答は “B”であるが，図 3.4(i) を観察すると，得点が低い群では誤答も比較的多く見受けられ，特に第 1 群では 40%超の受験者が誤答選択肢 “C”を選んでいることが示されている。また，誤答 “C”は他の群においても相対的に選ばれ易く，魅力ある誤答選択肢として機能している様子を窺うことができる。誤答 “A”は実際の計算結果として算出可能ではあるものの，それは黄色と緑のキャンディに関する確率である。誤答 “D”，“C”は問で与えられ

表 3.8 「いろいろな色のキャンディ」

いろいろな色のキャンディ

明さんにお母さんがバッグからキャンディを 1 個取るように言いました。明さんはキャンディを見ることができません。バッグの中のキャンディの色ごとの数は下のグラフの通りです。



明さんが赤いキャンディを取る確率はいくらですか。

A 10 %　B 20 %　C 25 %　D 50 %

（正答）“B”

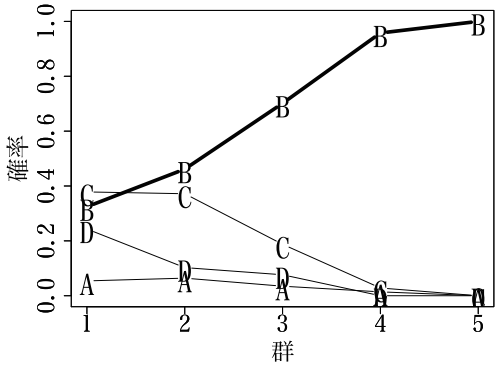
（国立教育政策研究所 (2004, p.321) より引用。字体，グラフ位置等の変更は本論文著者によるもの）

た数値から直接的に算出することは難しい値が答えとして設定されている。

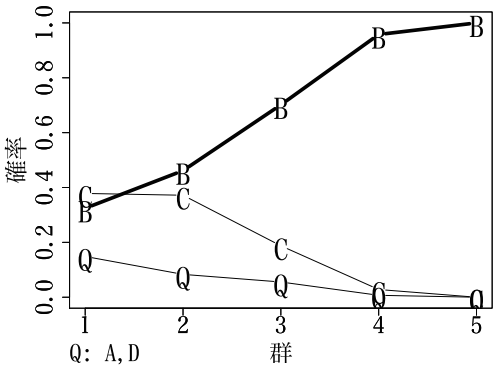
項目「キャンディ」に正答するためには，グラフを正しく読み取り，かつ読み取った値を用いて比率の計算を行うことが要求される。誤答 “C”を選んだ受験者は，計算間違いを犯したり，あるいは赤いキャンディが全体の半分はなく，10%では低過ぎると見当をつけた，もしくは単純に 20%台の選択肢が 2 つあるので何れかであるものと推測したと考えられる。

項目「キャンディ」は「再現クラスター」に分類される能力が要求される項目である。また，数学が用いられる状況として,「私的」: 生徒の日々の活動に直接関係する文脈（国立教育政策研究所, 2004, p.34）が想定されている。誤答した受験者は基本的な計算手順や，手順の練習量について再度確認し，練習問題と同様の状況に対して計算手順を適用する訓練を積む必要があるだろう。

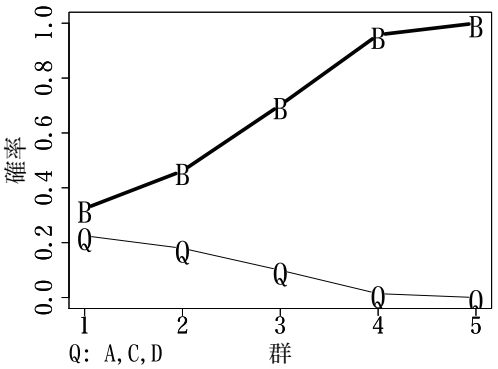
項目「キャンディ」においては選択肢 “A”，“D”は似た傾向を示す，魅力の薄



(i) 曲線をすべて描く場合



(ii) 曲線が 3 本にまとめられる場合



(iii) 曲線が 2 本にまとめられる場合

図 3.4 項目「キャンディ」・ブックレット 13 における項目特性図

い迷わしとして併合し，選択肢 “C”は全群に渡って特に魅力ある迷わしとして機能しているという状況が想定される。ここではヒューリスティックの H.3，H.4 が想定可能である。

この状況と，その他の状況について情報量規準を算出，比較する。図 3.4 と表 3.9 にブックレット 13（2003 年 PISA 調査）に含まれている項目「キャンディ」への提案手法の適用結果を示した。表 3.9 からは，図 3.4(ii) で表現される状況が推奨されることとなった。このことは選択肢 “C”が魅力的な誤答選択肢として機能しているという見立てを支持するものであるといえよう。

表 3.10 に「地震に関する問題」（以下,「地震」，ブックレット 3，受験者 351 名，35 点満点）の項目内容を示した。

誤答選択肢 “A”は問題で提示された数字を尤もらしく用いて計算しているが，

表 3.9 項目「キャンディ」・ブックレット 13 における各項目特性図の情報量規準

モデル	図 3.4(i)	図 3.4(ii)	図 3.4(iii)
群毎のカテゴリ数	4,4,4,4,4	3,3,3,3,3	2,2,2,2,2
パラメタ数	15	10	5
対数尤度	−274.611	−279.216	−301.513
AIC	579.222	578.431	613.026
BIC	637.134	617.039	632.33

20 年の内の 3 分の 2 の期間を問題としているわけではないことに注意しなければならない。誤答選択肢 “B”は 2 分の 1 を区切りとした白黒思考に囚われてしまい，誤答してしまうものと推察される。ただし，2 文目以降は,「だから」という演繹的な表現を除き，ゼットランド市という架空の都市であることを無視すると，こと日本においてはある種正しい状況であるともいえる。低得点群の受験者は,「地震」の問題を架空の都市における論理学的設問として理解することができず，身近な，現実世界における日本の問題として思考してしまっている可能性も考えられる。誤答選択肢 “D”は文章の正誤に関係なく，地震が起きるか否かを言明している「地質学者の言葉の意味」としては不適当である。

「地震」は「熟考クラスター」に分類される能力が要求される項目であり，最上位の能力が必要となる。また，数学が用いられる状況として,「科学的」, つまり，より抽象的な文脈で，技術的な過程，理論的な場面，明らかに数学的な問題についての理解に関連（国立教育政策研究所, 2004, p.34）した状況が想定されている。低特性の受験者は文章題の読み取り方や，確率の扱い方を基礎から確認することが重要であろう。

図 3.5(i) を観察すると，何れの誤答選択肢についても似た傾向を示しているものの，第 1 群においては選択肢 “B”の選択率が特異な傾向を示している。選択肢 “B”の文章内における尤もらしい確率の比較が，第 1 群に属する，即ち確率的な状況の判断が苦手な受験者の選択を誘因したものと推察され得る。ここではヒューリスティックの H.3 が想定される。

図 3.5 と表 3.11 にブックレット 3 に含まれている項目「地震」への提案手法の適用結果を示した。表 3.11 から，AIC によって図 3.5(ii) のモデルが推奨された。また，BIC では AIC の結果とは異なり，倹約的に，図 3.5(iii) のモデルが支持された。図 3.5(iii) でも第 1 群における誤答確率が上昇している様子は見受けられるものの，何れの誤答が選ばれ易かったのかは不明となる。第 1 群において “B”とその他の選択肢間の確率の差が大きければ，AIC と同様の結果を得たものと思われる。情報量規準間で推奨結果が異なる場合はあるものの，“B”が内容的に低得点群の受験者に選ばれ易かったという状況を表現することができる

表 3.10 「地震に関する問題」

地震

地震と地震の頻度についてのドキュメンタリー番組が放送されました。番組では地震を予知できるかどうかについても議論が交わされました。

ある地質学者は次のように言いました。「今後 20 年以内にゼットランド市で地震が起きる確率は 3 分の 2 だ」

この地質学者の言葉の意味を一番よく反映しているのは次のどれですか。

A $\frac{2}{3} \times 20 = 13.3$。だから，いまから 13 年から 14 年の間にゼットランド市ではいつか地震が起きる。

B $\frac{2}{3}$ は $\frac{1}{2}$ より大きい。だから，20 年の間にゼットランド市ではいつか必ず地震が起きる。

C 今後 20 年の間にゼットランド市で地震が起きる確率は，地震が起きない確率より大きい。

D 地震がいつ起きるかはだれも確信できないので，何が起きるかを予言することはできない。

（正答）“C”

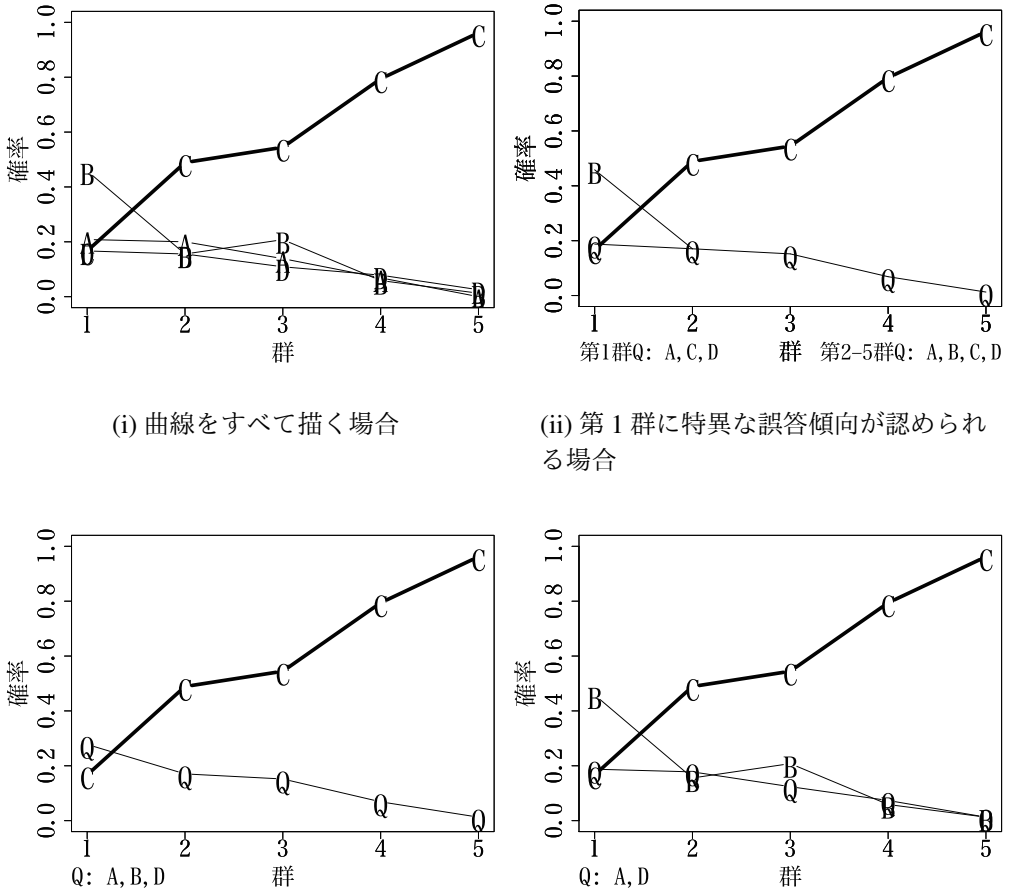
（国立教育政策研究所 (2004, p.327) より一部組み替えて引用）

図 3.5(ii) のモデルが支持されたこと，そして図 3.5(iv) や図 3.5(i) のモデルが支持されたわけではないことは手法の有用性の傍証となるだろう。

表 3.12 に「スケートボードに関する問 2」（以下「スケート」，ブックレット 6，受験者 328 名，36 点満点）の項目内容を示した。

図 3.6 と表 3.13 にブックレット 6 に含まれている項目「スケート」への提案手法の適用結果を示した。項目「スケート」は組み合わせの問題である。受験者はここには示していない,「スケートボードに関する問 1 」の時点で「デッキ」「車輪セット」「金具セット」「トラックセット」の部品があれば一つのスケートボードを組み立てられるという情報を既に与えられている。よってここでは単純な組み合わせの問題を問われている。正答に至るためにはすべての部品の種類数を掛け合わせる ($3 \times 2 \times 2 \times 1 = 12$)，あるいは樹形図を描き，数え上げるといった作業が必要となる（国立教育政策研究所, 2004, p.103）。

一方で，すべての部品の種類数を足し合わせてしまったり (B:$3+2+2+1 = 8$)，問題文を見落とす等，何らかの理由によって，ある部品の種類を抜いて組み合



(i) 曲線をすべて描く場合

(ii) 第 1 群に特異な誤答傾向が認められる場合

(iii) 曲線が 2 本にまとめられる場合

(iv) 曲線が 3 本にまとめられる場合

図 3.5 項目「地震」・ブックレット 3 における項目特性図

わせの計算を行ってしまう (“A”:$3\times2(\times1)=6, 3+2+1=6$) ことで，誤答 “A”，“B”に至るものと考えられる。しかしながら誤答選択肢 “C”は $(3+2)\times2=10$ とする等，出題の意図から離れ，加法と乗法を組み合わせた演算を施さなければ辿り着くことはできない。

図 3.6(i) を観察すると，選択肢 “A”，“B”は似た傾向を示しているが，特に不人気な誤答選択肢として “C”を見出すことができる。低得点者に対する迷わしとしてはやや複雑な計算手順によって選ばれ難くなっていたのであろう。

以上の項目「スケート」に対する誤答傾向の観察から，H.5 を想定して，選択肢 “A”，“B”を併合し，特異な，魅力のない誤答選択肢として “C”，正答選択肢 “D”の 3 本の曲線に整理して項目特性図を描くことが適当であろう。

各項目特性図の情報量規準を算出し，比較したところ，AIC により，図 3.6(ii) の状況で項目特性図を作成することが推奨され，誤答分析で得られた知見が支

表 3.11 項目「地震」・ブックレット 3 における各項目特性図の情報量規準

モデル	図 3.5(i)	図 3.5(ii)	図 3.5(iii)	図 3.5(iv)
群毎のカテゴリ数	4,4,4,4,4	3,2,2,2,2	2,2,2,2,2	3,3,3,3,3
パラメタ数	15	6	5	10
対数尤度	−294.957	−298.389	−300.360	−296.738
AIC	619.914	608.778	610.72	613.476
BIC	677.826	631.943	630.024	652.084

表 3.12 「スケートボードに関する問 2」

スケートボードに関する問 2

この店にはデッキ 3 種類，車輪セット 2 種類，金具セット 2 種類があります。トラックのセットは 1 種類しかありません。

浩二さんが組み立てられるスケートボードは何種類ですか。

A 6　B 8　C 10　D 12

（正答）“D”

（国立教育政策研究所 (2004, p.103) より一部組み替えて引用）

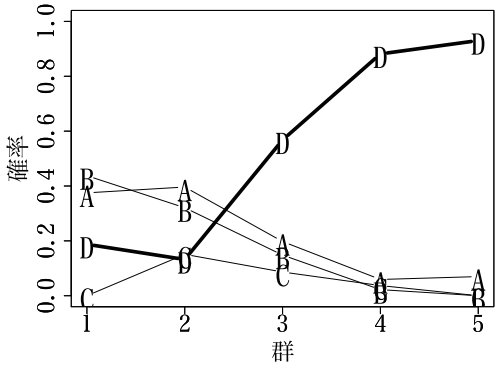
持される結果となった。

「スケート」は「再現クラスター」に分類される能力が要求され,「私的」な状況に関する項目である。誤答者は組み合わせの計算方法から再確認することが効果的であろう。

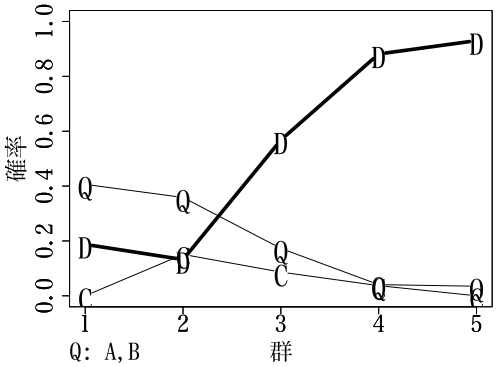
3.4 シミュレーション 研究 II による確認

本節では，これまで示した適用例が適切なものであったのかを，シミュレーションを通じて検討する。ただし，すべてのカテゴリ数，曲線のまとめ方について網羅的にシミュレーションを行うことは困難であるため，ここでは主に適用例で示されたモデルについて，項目特性図における曲線の本数を適切に推奨し得ることを確認する。

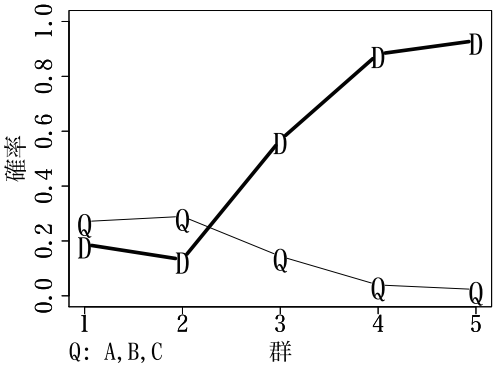
シミュレーションでは 4 種類のモデルを設定した。順番に，その内 1 つを真のモデルとし，1000 回の繰り返しの中で，情報量規準によって真のモデルが他のモデルに対して，適切に推奨される回数を確認した。以下で各モデルの項目



(i) 曲線をすべて描く場合



(ii) 特異に魅力のない誤答選択肢がある場合



(iii) 誤答曲線はすべて同様の傾向を示す場合

図 3.6 項目「スケート」・ブックレット 6 における項目特性図

特性図の状況と項目発生方法について述べる。なお受験者数を 350 名，項目形式を多肢選択形式とし，選択肢数を 4 とした。

モデル 1 全群に渡り，すべての誤答選択肢は同じ傾向を示していると見做せるモデル

曲線 1 正答曲線 1 本

曲線 2 3 つの誤答選択肢の選択率が等しいと見做した曲線 1 本

モデル 2 特定の誤答選択肢が全群に渡って同じ傾向を示していると見做せる（ここでは選択肢 3,4 とする）モデル

曲線 1 正答曲線 1 本

表 3.13 項目「スケート」・ブックレット 6 における各項目特性図の情報量規準

モデル	図 3.6(i)	図 3.6(ii)	図 3.6(iii)
群毎のカテゴリ数	4,4,4,4,4	3,3,3,3,3	2,2,2,2,2
パラメタ数	15	10	5
対数尤度	−253.192	−256.987	−267.962
AIC	536.385	533.974	545.925
BIC	593.28	564.973	559.397

曲線 2 特異な誤答曲線 1 本

曲線 3 2 本の誤答曲線を等しいと見做し，併合した曲線 1 本

モデル 3 第 2 群から第 5 群はすべての誤答選択肢は同じ傾向を示していると見做せるが，第 1 群において特別に特異と見做せる誤答選択肢が存在するモデル

曲線 1 正答曲線 1 本

曲線 2 第 2 群から 5 群まで，3 つすべての誤答選択肢を等しいと見做し，併合した曲線 1 本

曲線 3 第 1 群においてのみ，特異な誤答選択肢が 1 つ存在し，残りの 2 つの誤答選択肢が等しいと見做せる場合

モデル 4 すべての群において，すべての誤答選択肢が特異であるモデル

- 真なる構造の設定無し（すべての選択肢について，曲線を描画する）

項目数は 30 とし，内 1 項目が上記のモデルに従っている状況を仮定した。項目特性図作成時の分割群数は 5 群に固定し，選択肢 1 を正答選択肢とした。

3.4.1 各モデルにおけるシミュレーションデータの発生

シミュレーション・モデル 1

シミュレーション 1 では，モデル 1 に基づいて，選択率に関する構造を仮定した項目が 30 項目中 1 項目ある状況を想定し，当該項目に対して第 2.2 節で述べた方法で情報量規準の算出を行い，他のモデルとの比較を行った。シミュレーション 2 から 4 も同様の手順で行った。

表 3.14 各群の一様乱数による正答率発生範囲

群	始点	終点
第 1 群	0.00	0.20
第 2 群	0.15	0.40
第 3 群	0.35	0.60
第 4 群	0.55	0.80
第 5 群	0.75	1.00

モデル **1** の項目データ発生方法

過程 [A.1] 正答 1: 表 3.14 に基づき，一様乱数によって各群の正答率を発生させる。

過程 [A.2] 誤答 2–4: 誤答選択肢全体の選択率（1– 正答率）を 3 等分し，各誤答選択肢の選択率として設定する。

シミュレーション・モデル **2**

モデル **2** の項目データ発生方法

過程 [B.1] 正答 1: 表 3.14 に基づいて，一様乱数から正答率を発生させる。誤答選択肢全体の選択率（1– 正答率）を算出する。

過程 [B.2] 誤答 2: 0 から全体誤答選択率までの範囲で，一様乱数を発生させ，各群における特異な誤答曲線の選択率として設定する。

過程 [B.3] 誤答 3–4: {［1–（正答率+特異な誤答率）］/2} を計算し，各群の互いに等しいと見做し得る誤答選択率として設定する。

シミュレーション・モデル **3**

モデル 3 の項目データ発生方法　適用例の状況に基づき，特異な選択肢を設定する群を第 1 群とした。

過程 [C.1] 第 1 群・正答 1: シミュレーション 2（過程 B.1）に基づき，選択率を設定する。

過程 [C.2] 第 1 群・誤答 2–4: シミュレーション 2（過程 B.2 と B.3）に基づき，選択率を設定する。

過程 [C.3] 第 2–5 群・正答 1: 過程 C.1 を再度実行する。

過程 [C.4] 第 2–5 群・誤答 2–4: モデル 1 における誤答選択肢データの発生手順（過程 A.2）に基づき，選択率を設定する。

シミュレーション・モデル **4**

モデル **1** から **3** を仮定しない項目データの発生方法

過程 [D.1] 正答 1: 表 3.14 に基づき各群の正答率を発生させる。（1− 正答率）で全体誤答率を算出する。

過程 [D.2] 誤答 2: 0 から（1− 正答率）の範囲で選択率を設定する。

過程 [D.3] 誤答 3: 0 から［1−（正答率 1 ＋誤答率 2）］ の範囲で選択率を設定する。

過程 [D.4] 誤答 4:［1−（正答率 1 ＋誤答率 2 ＋誤答率 3）］を誤答選択肢 4 の選択率として設定する。

ここでは 30 項目の内，選択肢が併合可能な項目（モデル 1, 2, 3）が特に存在していない場合を想定し，シミュレーション 1 から 3 と同様の情報量規準の比較を行った。

3.4.2 受験者の項目反応データの生成

0 から 1 の範囲で一様乱数を群ごとの受験者数分，生成する。上記の過程 A から過程 D に基づいて発生させた各項目の，各群の各選択肢に対する選択率の累積確率と，受験者の一様乱数を比較する。各群の選択肢ごとの範囲の何れに受験者の一様乱数が所属しているかに基づき，受験者のカテゴリ反応データを生成した。例えば過程 B によって設定された，第 1 群における選択肢 1 から 4 までの選択率が $(0.163, 0.620, 0.108, 0.108)'$ である場合には，累積確率は $(0.163, 0.784, 0.892, 1.000)'$ となる。そして，各受験者の一様乱数と比較し，0 以上，0.163 未満ならば選択肢 1 を，0.163 以上 0.784 未満ならば選択肢 2 を選択したものとして各受験者の反応を決定する。選択肢 4 は上限として 1.000 までを含めた範囲 $[0.892, 1.000]$ で受験者の反応を決定している。表 3.15 に各モデルのシミュレーションデータについて，各過程で発生させた項目の内訳を示した。表 3.15 の項目構成に則った項目反応シミュレーションデータ（受験者 350 人，30 項目のサイズ 350×30）を 1000 回分発生させ，シミュレーションを行った。

表 3.15 シミュレーションテストデータの項目構成

	過程 A	過程 B	過程 C	過程 D
モデル 1	1	0	0	29
モデル 2	0	1	0	29
モデル 3	0	0	1	29
モデル 4	0	0	0	30

表 3.16 シミュレーション結果：（括弧内は BIC による結果）

\ 真の構造	モデル 1	モデル 2	モデル 3	モデル 4
モデル 1	917 (1000)	20 (172)	172 (314)	0
モデル 2	69 (0)	886 (828)	96 (0)	0
モデル 3	—	—	716 (686)	—
モデル 4	14 (0)	94 (0)	16 (0)	1000 (1000)
的中率	0.917 (1.00)	0.886 (0.828)	0.716 (0.686)	1.00 (1.00)

3.4.3 シミュレーション結果

表 3.16 に 4 種類のシミュレーションについて，情報量規準 (AIC, BIC) による推奨頻度の結果を示した。括弧内の数値は BIC の推奨頻度を表している。モデル 3 については，実際の適用場面においてモデル 2 が仮定された場合であっても必ずしも見出され得るわけではないため，真のモデルとして設定される場合以外には比較対象としていない。何れのモデルの場合も，試行数の過半数以上，真の構造を正しく推奨することが可能であることが示された。

モデル 1 の場合は全試行の内 90%以上が正しく推奨された。モデル 1 が真の構造として設定されている場合に，モデル 2 やモデル 4 が推奨されたのは，選択肢間の確率の差が，同等と見做すには大き過ぎたためであろう。モデル 2 の場合，全試行中 87%が正しく推奨された。モデル 2 が真の構造である場合には，選択肢間の確率の差が大き過ぎる，あるいは小さ過ぎる場合にモデル 1 やモデル 4 が推奨されたと推察される。モデル 3 の場合，72%が正しく推奨された。他のモデルと比較すると，低い推奨精度となった。特に，モデル 1 が推奨される傾向が見受けられる。これは第 1 群のみが特異であるという限定的な状況においては，当該群における確率の差が明確でない場合は，すべての誤答選択肢が等しいと見なせるモデル 1 が推奨され易くなるものと解釈可能である。また併合可能な選択肢が設定されていないモデル 4 の場合には，全試行に渡って通常の項目特性図を描くことが推奨された。BIC においても AIC と同様の傾向が確

認できるが，BIC の場合はパラメタ項に関してより保守的となることが示唆された。

3.5 研究 II 結論

実データへの適用例，シミュレーションを通じて，提案手法が，誤答分析において，併合可能であると判断可能なモデルを適切に推奨し得ることが確認された。ただし BIC は保守的な傾向にあることも同時に示唆された。また，適用例において確認されたように，妥当ではあるものの，AIC と BIC は必ずしも同一の項目特性図を推奨するわけではない。このことは情報量規準を用いたモデル選択の限界として不可避の事態であり，留意する必要があるが，今回の適用例においては例え異なるモデルが推奨されたとしても，解釈可能な範囲内であり，手法の有用性に大きく影響を与えるものではないと考えられる。

提案手法は，仮説を反映した項目特性図の優位性を自動的に付与するものではない。情報量規準による項目特性図（モデル）選択において，最終的な判断は分析者の教科内容に対する知識に基づいて行われるべきである。

本研究で用いた併合のための 5 つのヒューリスティックは，経験的に導き出されたものである。適用例で示したように，様々な特性曲線のパターンを示す項目に対して適用可能ではあるものの，適用例で扱った項目以外の誤答分析を行った際には，これらに加えて有用なヒューリスティックが更に見出される可能性は存在する。その場合には，提案手法の適切さについての更なる検討が必要となるだろう。

平均選択率を用いることで，描画する特性曲線を選択するのではなく，すべての選択肢について考慮した状態で特性曲線の本数を減らしつつ，より特徴的な誤答選択肢を抽出して項目特性図による誤答分析を行うことが可能となる。整理された状況で，細やかな項目分析，誤答分析が可能となり，受験者，特に低得点者への効果的な学習指導へと繋げることが期待できる。

この際，併合状況が仮説として複数考えられる場合に，本稿において提案した手法を用いることで，なぜその中から当該項目特性図を用いて，項目特性を表現したのかを，教科内容的観点に加えて統計的な観点からも述べることができるようになる。

第4章　ベイズ統計学とモンテカルロサンプリング

本章では，次章において用いられるベイズ統計学の枠組みを用いた母数推定方法について概説する。

4.1　同時確率・条件付確率・周辺化

4.1.1　同時確率・条件付確率

いま，2枚のコインを投げ，どちらの面が上となるかを記録する試行を考える。このとき，2つの事象が同時に起こる確率である，同時確率 (joint probability) は確率の乗法定理より，

$$\Pr(A \cap B) = \Pr(B)\Pr(A|B) = \Pr(A|B)\Pr(B)$$

によって表される。

$\Pr(A|B)$ は，事象 B が与えられたときの，事象 A が観察される確率を示しており，これを条件付確率 (conditional probability) と呼ぶ。条件付確率は

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

によって与えられる。

ここで，2枚のコインはどのような順番で，あるいは同時に投げようとも，互いの試行結果に影響を及ぼしあうことはない。事象 B という条件の下で事象 A が生起する確率は $\Pr(A|B) = \Pr(A)$ で与えられ，事象 A の生起確率に一致する。このとき，事象 A と事象 B は独立であるという。

互いに独立な事象 A と B の同時確率は各事象同士の生起確率の積を用いて，$\Pr(A \cap B) = \Pr(B)\Pr(A)$ と表現される。

4.1.2　周辺化

いま，手元に2枚のコイン，3枚のコイン，4枚のコインのセット S_1, S_2, S_3 があるものとする。この内1つのセットを無作為に選び，選ばれたセットのコ

インを投げてどちらの面が観察されるかを調べる試行を考える。このとき，投げたコインすべてにおいて表面 H_N（N は表面が出たコインの数）が観察される確率は，選ばれるセットに依存し，

$$\Pr(H_2|S_1), \quad \Pr(H_3|S_2), \quad \Pr(H_4|S_3)$$

となる。

いま，任意のセットについて，コインを投げ，すべて表面が観察される確率を調べるとき，セットが選ばれる確率自体も考慮しなければならない。しかしながら，セットが選ばれる確率には興味はない。このとき，興味の対象外であるセットというパラメタは局外パラメタと呼ばれる。

セットが選択される確率と，各セットにおいてすべての試行で表面が観察される確率の積和を計算することにより，局外パラメタを確率分布から消去することが可能である。この操作を周辺化 (marginalization) と呼ぶ。また，周辺化によって得られた分布を周辺分布 (marginal distribution) と呼ぶ。変数が連続型の場合には積和に代えて局外パラメタについて積分を行うことで同様の結果を得ることができる。

4.2 ベイズの定理

事象 B が与えられた下で，事象 A が観察される確率を

$$\Pr(A|B) = \frac{\Pr(B|A) \times \Pr(A)}{\Pr(B)} \tag{4.1}$$

と表現する。あるいは同等に，A が仮説であり，B が実際に得られるデータを表すものとし，データ B が得られた下での仮説 A が成り立つ確率と捉えることもできる。

(4.1) 式はベイズの定理 (Bayes' theorem) と呼ばれ，ベイズ統計学の根幹をなす確率表現である。ここで，$\Pr(A)$ は実際に手元にデータが得られる前の，仮説 A に対する事前情報，あるいは信念の強さを示す主観的な確率を表している。このことから $\Pr(A)$ は事前確率とも呼ばれる。

$\Pr(B|A)$ は仮説 A の下でデータが得られる尤度を表している。これは，仮説 A の下において，データ B が実際に得られる確率を表している。$\Pr(B)$ は $\Pr(B|A)$ を周辺化した，データが得られる確率を表している。

一方で，(4.1) 式の $\Pr(A|B)$ は実際にデータ B が手元に得られた後に，更新された仮説 A についての信念の強さを表す主観確率である。このことから $\Pr(A|B)$ は事後確率と呼ばれる。

(4.1) 式は個別の事象から，確率変数および分布を扱う場合へと拡張することが可能である。確率変数 Y と Y に関する未知パラメタ θ に関して，ベイズの定

理は

$$f(\theta|Y) = \frac{f(Y|\theta) \times f(\theta)}{f(Y)} \tag{4.2}$$

と表現される。このとき (4.2) 式の $f(\theta)$ は実際のデータが得られる以前の信念としての θ が従っている分布を表し，事前分布 (prior distribution) と呼ばれる。また，$f(Y|\theta)$ はパラメタ θ に関する尤度である。$f(Y)$ は $f(Y|\theta)$ に関する周辺尤度であり（渡部, 1999, p.16），規格化定数とも呼ばれる。

$f(\theta|Y)$ は実際にデータが得られた後に更新された θ についての分布を表しており，事後分布 (posterior distribution) と呼ばれる。

4.3 ベルヌイ分布，2項分布，階層ベイズモデル，ベータ分布

本節では照井 (2010, p.22-26)，古谷 (2008, p.16) に基づき，第 5 章で用いる 2 項分布，及びベータ分布を導入する。

4.3.1 ベルヌイ試行・ベルヌイ分布・2項分布

いま，硬貨を投げて表裏を当てるという試行を行うことを考える。この試行結果を y で表すとすると，成功 ($y = 1$) もしくは失敗 ($y = 0$) の 2 値となる。このとき，成功確率が p であるような試行をベルヌイ試行 (Bernoulli trial) という。ベルヌイ試行の確率関数は

$$\Pr(y|p) = p^y(1-p)^{1-y} \tag{4.3}$$

であり [1]，(4.3) 式をベルヌイ分布 (Bernoulli distribution) と呼ぶ。ここで N 回の独立なベルヌイ試行結果を $\boldsymbol{y} = (y_1, \ldots, y_N)'$ としたとき，これらの同時確率は

$$\Pr(\boldsymbol{y}|p) = \Pr(y_1, \ldots, y_N|p) = \prod_{i=1}^{N} y_i(1-p)^{1-y_i} \tag{4.4}$$

$$= p^{\sum_{i=1}^{N} y_i}(1-p)^{n-\sum_{i=1}^{N}} \tag{4.5}$$

と表される。ここで，N 回の独立なベルヌイ試行 $y_1, \ldots, y_n$ を行ったとき，任意の順番で n 回成功したかは，

$$n = y_1 + y_2 + \cdots + y_N = \sum_{i=1}^{N} y_i \tag{4.6}$$

[1] ここで $\Pr(\cdot)$ と p はそれぞれ別個の確率を表している記号であることに注意されたい

で与えられる。(4.6) 式の事象が観察される確率は

$$\Pr(\boldsymbol{y}|p,N) = \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{N-n} \tag{4.7}$$

という確率関数（尤度関数）によって表現される。(4.7) 式は 2 項分布 (binominal distribution) と呼ばれる。ここで

$$\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!(N-n)!} \tag{4.8}$$

であり，N 回試行中，どのような組み合わせで n 回の成功事象が観察されたかを表している。その尤度関数は事象の組み合わせに関わらず,

$$\Pr(n|p,N) \propto p^n (1-p)^{N-n} \tag{4.9}$$

と与えられる。ここで $\propto$ は比例を意味する。

4.3.2 階層ベイズモデル

あるモデルにおける，任意のパラメタ p の事前分布に対して，更なる事前情報が仮定されるものとして，その事前情報を規定するパラメタ θ を用いて階層的に

$$\begin{aligned} \boldsymbol{y}|p &\sim \Pr(\boldsymbol{y}|p) \\ p|\theta &\sim \Pr(p|\theta) \\ \theta &\sim \Pr(\theta|\theta_0) \end{aligned}$$

という関係性が想定されるとき，当該モデルを階層モデル (hierarchical model) と呼ぶ。階層的な事前分布 $\Pr(\theta)$ のパラメタである θ_0 は特に超パラメタ (hyper parameter) と呼ばれる。超パラメタの値は既知のものとして扱われることもあるが，無情報的な事前分布が仮定されることもある。

例えばいま，生徒の学力調査を行うことを想定しよう。ここで生徒は各々が所属する，学校単位の属性から影響を受け，学校は複数地域から抽出され，所属地域単位から影響を受け，各地域は当該の国の文化等から影響を受けていると考えられる場合，この構造は階層的であるといえる。階層モデルはデータが得られる状況，事前情報の影響関係を柔軟に表現することができる有用な道具である。

上記階層モデルの同時事後分布は

$$\begin{aligned}\Pr(p,\theta|\boldsymbol{y}) =& \frac{\Pr(p,\theta,\boldsymbol{y})}{\Pr(\boldsymbol{y})} \\ =& \frac{\Pr(\boldsymbol{y}|p,\theta)\Pr(p,\theta)}{\Pr(\boldsymbol{y})} \\ =& \frac{\Pr(\boldsymbol{y}|p)\Pr(p|\theta)\Pr(\theta)}{\Pr(\boldsymbol{y})}\end{aligned}$$

となる。周辺事後分布 $\Pr(p|\boldsymbol{y})$ は θ に関して積分消去を行うことで

$$\begin{aligned}\Pr(p|\boldsymbol{y}) =& \int \Pr(p,\theta|\boldsymbol{y})d\theta \\ =& \int \frac{\Pr(\boldsymbol{y}|p)\Pr(p|\theta)\Pr(\theta)}{\Pr(\boldsymbol{y})}d\theta\end{aligned}$$

と表される。上式について，更に規格化定数 $\Pr(\boldsymbol{y})$ を無視すれば，

$$\Pr(p|\boldsymbol{y}) \propto \Pr(\boldsymbol{y}|p)\Pr(p|\theta)\Pr(\theta)$$

となる (照井, 2010, p.11)。

3 階層以上の場合も同様に構造化することができる。階層構造は理論的には無限に仮定することが可能ではあるが，実際場面では 2 ないし 3 階層までを仮定することが一般的である。

4.3.3 ベータ分布・共役性

前項までに導入した 2 項分布に関するパラメタ p の階層的な事前分布として，ベータ分布 (beta distribution)，

$$\Pr(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}, \quad \alpha,\beta>0 \tag{4.10}$$

を仮定する。ベータ分布はパラメタとして α,β を持ち，これらは形状パラメタ (shape parameter) と呼ばれる。α を形状パラメタ，β を尺度パラメタ (scale parameter) と呼び分ける場合もある。ここでは形状パラメタ α,β が超パラメタとしての役割を担っている。(4.10) 式中の $\Gamma(\cdot)$ はガンマ関数を表しており，

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy, \quad \alpha>0 \tag{4.11}$$

である (豊田, 2006, p.159)。

ベータ分布の平均と分散はそれぞれパラメタ α, β を用いて

$$\begin{aligned} E(p) =& \int_0^1 p \Pr(p) dp \\ =& \frac{\alpha}{\alpha + \beta} \end{aligned} \tag{4.12}$$

および

$$\begin{aligned} \text{Var}(p) =& \int_0^1 p^2 \Pr(p) dp - E(p)^2 \\ =& \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned} \tag{4.13}$$

と与えられる。このとき事後分布は規格化定数を無視すると，

$$\begin{aligned} \Pr(p|\boldsymbol{y}) \propto& \Pr(p) \Pr(\boldsymbol{y}|p) \\ \propto& p^{\alpha-1}(1-p)^{\beta-1} p^{\sum_{i=1}^n y_i} (1-p)^{n-\sum_{i=1}^n y_i} \\ \propto& p^{(\alpha+\sum_{i=1}^n y_i)-1} (1-p)^{(\beta+n-\sum_{i=1}^n y_i)-1} \end{aligned} \tag{4.14}$$

となり，事後分布もまた，ベータ分布に従っていることが分かる（照井, 2010, pp.22-24）。データが得られ，情報が更新された後の事後分布が，更新前の分布と同じ分布族に従っているような事前分布のことを特に（自然）共役事前分布 ((natural) conjugate prior distribution) と呼ぶ（繁桝, 1985, p.45; 渡部, 1999, p.85）。

4.4 マルコフ連鎖モンテカルロ法

(4.2) 式における $\Pr(Y)$ を得るためには，

$$\int \Pr(Y|\theta) \Pr(\theta) d\theta \tag{4.15}$$

を計算する必要があるが，これはしばしば困難を伴うものとなる。モデルの記述力を上げるために多くの母数を取り上げ，複雑な事前分布と尤度を構成すると，それだけ高次の積分計算が求められるようになるためである。

比較的簡易なモデルであっても，多重積分を行う必要が生じるため，電子計算機が十分に発達するまでは推定の負担が小さくはなかった。また，積分が解析的に解ける保証もないため，有用なモデルであることが見込まれても，解を求めることができないことがあった。これらの問題を軽減するために，従来のベイズ推定においては，共役事前分布を用いた事後分布の導出が行われてきた。しかしながら，推定の負担を軽減するために事前分布を選択するという行為は本来の分析目的とは何ら関係がない。

歴史的に，分析者の事前知識を導入することは客観性に欠けるという批判が行われることがあったが，更に共役性のある事前分布を便宜上の理由から仮定することもまた，事前情報を表現するという事前分布の妥当性に欠ける行為であるとして批判の対象となることもある。

また，すべてのモデルについて共役性が保障される事前分布が適用可能というわけではないために，分析の方向性が限定されてしまうことも，ベイズ推定が実際に利用されることを妨げる一因となった。このため，ベイズ統計学は事前情報を組み込み，データを得た後に，この情報が事後分布という形式で更新されるという興味深い方法ではあるものの，主としてベイズ統計学，および推定法それ自体を対象とする理論的な側面からの研究が多かった。

しかしながら，近年はマルコフ連鎖モンテカルロ (markov chain monte carlo; MCMC) 法と呼ばれる，目的とする確率分布から直接サンプリングを行う方法が広まるに連れ，ベイズ統計学の応用的な広まりを見せている。

マルコフ連鎖モンテカルロ法は，事後分布を数値的にシミュレートすることで高次元積分を回避し，推定量の分布を直接調べることを可能とする。従来は電子計算機による計算コストが高い手法であったために，複雑なモデルへの適用は控えられてきた。しかし近年の電子計算機の能力発達に伴い，複雑なモデルへの適用が増えている。

マルコフ連鎖モンテカルロ法を用いることで，分析者は高次積分の解を解析的に求めることが可能かどうかに注意を払う必要なく，パラメタの事後分布を求めることが可能となる。また，事前分布の選択も，共役性が保たれるか否かという，研究において非本質的な性質を気にすることなく，事前情報の反映という本来の役割に基づいて―選択した事前分布が本当に適切であるかどうかということには依然として注意を払わなければならないものの―選択することが可能となる。

本節以降では，マルコフ連鎖モンテカルロ法の概要について，豊田 (2008a), 森村・高橋 (1979) に基づいて述べた後，マルコフ連鎖モンテカルロ法と同様に，サンプリングアルゴリズムの一種であるハイブリッドモンテカルロ法について概説する。

4.4.1 マルコフ連鎖

いま，時点 $t = 1, 2, \cdots$ について，各時点における状態 (確率過程, stochastic process) を実現値とする確率変数を $x^{(t)}$ と表す。また，$x^{(t)}$ のとりうるすべての状態を含む集合を S とする。この集合 S は状態空間 (state space) と呼ばれる。いま，条件付確率

$$\Pr(x^{(t+1)} = j | x^{(1)} = i_1, \cdots, x^{(t)} = i) \tag{4.16}$$

を考える。ここで (4.16) 式は，時点 t において状態が $i(i \in S)$ にある事象が，次の $t+1$ 時点において j となる確率である。この条件付確率 (4.16) 式が，

$$\mathrm{Pr}(x^{(t+1)} = j | x^{(1)} = i_1, \cdots, x^{(t)} = i) = \mathrm{Pr}(x^{(t+1)} = j | x^{(t)} = i) \tag{4.17}$$

という性質をすべての t について満たすとき，確率過程 (4.17) 式はマルコフ性を持つといい，マルコフ連鎖 (markov chain) と呼ばれる。マルコフ性を持つ確率過程では，任意の時点 t において，$x^{(t)}$ が状態 i にあるとき，次の時点 $t+1$ において，$x^{(t+1)}$ が状態 j へと推移する確率は現在の状態 $x^{(t)}$ のみに依存することとなる。つまり，$x^{(t)}$ が所与の下では，$t-1$ 以前の時点の状態 $x^{(1)}, \cdots, x^{(t-1)}$ とは独立となる。

また，どの時点 t においても，推移の確率は同じであるという仮定を推移確率の定常性といい，定常な推移確率をもつマルコフ過程は斉時的であるといわれる。

いま，(4.17) 式より，ある時点 t から次の時点 $t+1$ に推移する条件付確率を

$$p_{ij} = \mathrm{Pr}(x^{(t+1)} = j | x^{(t)} = i)$$

と表現する。この p_{ij} を要素とする行列は

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I1} & p_{I2} & \cdots & p_{IJ} \end{bmatrix}, \quad I = J \tag{4.18}$$

と表される。(4.18) 式は推移確率行列 (transition probability matrix) と呼ばれる。あるいはマルコフ連鎖の推移を決定する中核を担うことから，推移核 (transition kernel) とも呼ばれる。なお (4.18) 式中の各要素 p_{ij} は確率であることから，次の状態への推移，あるいは現在の状態に留まる場合も含めて，

$$0 \leq p_{ij} \leq 1, \quad \sum_{j=1}^{J} p_{ij} = 1$$

という制約が存在する。

4.4.2 マルコフ連鎖と不変分布

時点 t において状態 i にある確率を

$$\pi_i^{(t)} = \mathrm{Pr}(x^{(t)} = i) \tag{4.19}$$

とする。同様にすべての状態について表すと

$$\boldsymbol{\pi}^{(t)} = (\pi_1^{(t)}, \pi_2^{(t)}, \ldots, \pi_I^{(t)})' \tag{4.20}$$

となり，(4.20) 式は時点 t における状態確率分布 (state probability distribution) と呼ばれる。推移が始まる前の初期状態の分布は初期分布と呼ばれる。ここで，状態確率分布における初期分布と，推移確率行列は，

$$\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(1)}\boldsymbol{P} \tag{4.21}$$

という関係を有している。更に，同様の関係を用いて，$\boldsymbol{\pi}^{(3)}$ は，$t-1$ 時点の状態確率分布を用いて，

$$\boldsymbol{\pi}^{(3)} = \boldsymbol{\pi}^{(2)}\boldsymbol{P} \tag{4.22}$$

となる。つまり，任意の時点 t について，

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(t-1)}\boldsymbol{P} \tag{4.23}$$

である。ここで (4.21) 式の結果を用いると，

$$\boldsymbol{\pi}^{(1)}\boldsymbol{P}\boldsymbol{P} = \boldsymbol{\pi}^{(3)} \tag{4.24}$$

と表すことも可能である。(4.22) 式の関係は時点 t について，

$$\boldsymbol{\pi}^{(t)}\boldsymbol{P} = \boldsymbol{\pi}^{(t+1)} \tag{4.25}$$

と表現可能である。同様に，確率 $\boldsymbol{\pi}^{(t+1)}$ については，(4.21) 式，(4.24) 式の関係から，$\boldsymbol{\pi}^{(1)}$ に対して，$\boldsymbol{P}$ を必要な回数分，繰り返し掛けることによって，

$$\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(1)}\boldsymbol{P}^t \tag{4.26}$$

と表すことができる。(4.26) 式から，$t+1$ 時点の状態確率分布 $\boldsymbol{\pi}^{(t+1)}$ を求めることが可能となる。

以上の結果から，t における状態確率分布のベクトルと，その推移確率行列との積を計算することで，次の時点 $t+1$ における状態となる確率を求めることが可能となる。

4.4.3 エルゴード性

MCMC では，まずエルゴード的なマルコフ連鎖を構成することを目指す。ここでは，エルゴード性を満たすための条件について述べる。

既約的

マルコフ連鎖の状態 i が有限回の推移 t において，状態 j へと到達する確率がゼロではなく，$P_{ij}^{t} > 0$ のとき，状態 i は状態 j に到達可能であるという。さらに，$P_{ji}^{t\prime} > 0$ も成り立つならば，状態 i と状態 j は互いに到達可能であるという。マルコフ連鎖の状態空間 S の要素がすべて互いに到達可能であるとき，既約的 (irreducible) という。既約的なマルコフ連鎖では，状態空間全体が 1 つの既約な集合となっている。既約な集合は，状態が含まれている集合のうち，集合内のある状態 i から，状態 j に対して推移を繰り返すことで到達することができるが，集合に含まれていない状態 k へは推移することはないという性質をもつ。

再帰的

状態 i から推移を開始し，再び状態 i となる最小の推移（時点）回数を T_i とする。このとき，$\Pr(T_i < \infty) = 1$，つまり有限回の推移回数で再び同一の状態となることが確実であるなら，状態 i は再帰的 (recurrent) であるといわれる。再帰的な状態 i に関して $E[T_i] < \infty$ であるとき，正再帰的であるといわれる。既約的なマルコフ連鎖が正再帰的であるとき，任意の状態は推移の間に幾度も到達されることとなる。

周期的・非周期的

既約的なマルコフ連鎖において，状態 i から状態 j へと到達する推移回数の規則性の有無により，連鎖の性質が周期的か非周期的かに分類される。周期 (period) とは，連鎖の各状態が同一の繰り返しになる時点間の長さのことである。例えば，3 つの状態 i_1, i_2, i_3 について，状態 i_1 から状態 i_2 へ，また，状態 i_2 から状態 i_3 へは到達可能であるが，その逆は不可能であるようなマルコフ連鎖があるとする（この場合各推移確率は 1 である）。このとき，任意の状態から出発し，同じ状態に到達するまで 3 回の推移が必要となる。ここで，i_1 から i_2 へ到達するための必要推移回数は 1 であり，その他の個別の推移に関しても必ず 1 回で辿り着くことが可能であるが，周期的な性質を持つ連鎖のうち，最大回数を当該マルコフ連鎖の周期として扱うこととなる。よって，当該マルコフ連鎖の周期は 3 であり，周期 3 の周期的マルコフ連鎖と呼ばれる。周期的マルコフ連鎖には，個別の状態のみならず，状態を含む集合間への到達に関して，規則性がある場合も含まれる。周期的なマルコフ連鎖は，推移を何度繰り返しても，初期状態の影響が残り続ける。一方で，すべての状態について周期が 1 であるようなマルコフ連鎖は非周期的 (aperiodic) であるといわれる。

エルゴード的なマルコフ連鎖

あるマルコフ連鎖が既約的，正再帰的，非周期的であるとき，当該マルコフ連鎖はエルゴード性 (ergodicity) を満たし，エルゴード的マルコフ連鎖と呼ばれる。エルゴード的マルコフ連鎖は，状態空間の部分集合におけるすべての状態 i, j に対し，

$$\lim_{t\to\infty} P_{ij}^t = \pi(j) \tag{4.27}$$

が成立する。つまり，エルゴード的なマルコフ連鎖は，いかなる状態を初期状態として推移を開始しても，推移を $t \to \infty$ と繰り返したとき，分布

$$\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi} \tag{4.28}$$

に収束する。(4.28) 式は，推移確率行列 $\boldsymbol{P}$ に従って状態が推移したとしても，全体の分布としては $\boldsymbol{\pi}$ であることを示している。もし，推移を開始する時点での分布が

$$\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi} \tag{4.29}$$

である場合，(4.23) 式および (4.28) 式から，

$$\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(T)} = \boldsymbol{\pi} \tag{4.30}$$

となり，$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}$ へと収束する。この分布 $\boldsymbol{\pi}$ を推移確率行列 $\boldsymbol{P}$ を持つマルコフ連鎖の不変分布 (invariant distribution)，または定常分布 (stationary distribution) と呼ぶ。

4.4.4 モンテカルロ法

乱数を，特定の分布や標本から生成するとき，この方法をモンテカルロ発生 (monte carlo generation) という。

確率変数の列 $\{x^{(t)}\}(t = 1, 2, \ldots)$ と確率変数 x について，$\epsilon > 0$ に対して

$$\lim_{t\to\infty} \Pr(|x_t - x| \geq \epsilon) = 0 \tag{4.31}$$

または，同等に

$$\lim_{t\to\infty} \Pr(|x_t - x| < \epsilon) = 1 \tag{4.32}$$

であるならば，$\{x^{(t)}\}$ は x に確率収束するという。

いま，確率変数の列 $\{x^{(t)}\}$ が，相互に独立であり，それぞれ平均 μ，分散 $\sigma^2 < \infty$ をもつ分布に従っているとする。このとき，$\{x^{(t)}\}$ は独立同分布に従う (independent and identically distributed; iid) という。これらの標本平均 $\bar{x}^{(t)} = t^{-1}\sum_{i=1}^{t} x^{(t)}$ については，大数の弱法則と呼ばれる定理

$$\lim_{t\to\infty} \Pr(|\bar{x}^{(t)} - \mu| < \epsilon) = 1 \tag{4.33}$$

が成立する。大数の弱法則 (4.33) 式より，$t \to \infty$ のとき，標本平均は期待値 μ に確率収束し，期待値が標本平均で近似できることになる。また，任意の連続関数 $h(\cdot)$ によって x を変換した場合も，大数の法則は成立する。

いま，連続関数 f における積分

$$\int f(\boldsymbol{x})d\boldsymbol{x} \tag{4.34}$$

を求めたいとする。もし f の積分を解析的に求めることが困難ならば，数値解析によって解くこととなる。このとき，数値解析の方法の一つとして，モンテカルロ法を用いることが可能である。積分の対象である関数を $f(\boldsymbol{x})$ とする。もし，独立に同一の分布 $f(\boldsymbol{x})$ に従う乱数 $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ を生成することができるならば，積分は

$$\int_a^b f(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{T}\{f(\boldsymbol{x}^{(1)}) + \cdots + f(\boldsymbol{x}^{(T)})\} \tag{4.35}$$

のように近似することが可能である。大数の弱法則から，$T \to \infty$ であるとき，平均による近似は対象の積分の値に確率収束することが保証される。このように，乱数を用いて，積分を数値的に求める方法はモンテカルロ積分と呼ばれる。乱数のサンプリング法として，棄却サンプリングや重点的サンプリングが提案されている。

モンテカルロ法は，特定の分布や標本から乱数を発生させたい場合に用いられる。しかしながら，生成したい分布が複雑であり，特定するのが難しいような場合には，サンプリングが困難であるという性質を持つ。

4.4.5 マルコフ連鎖モンテカルロ法

マルコフ連鎖において，状態はそれぞれ条件付確率の分布に従って推移するものの，推移を繰り返していく間に，いずれの状態確率分布も，同一の不変分布に従うようになるという性質をもつ。モンテカルロ法において乱数のサンプリングが困難である場合に，マルコフ連鎖の性質を用いることが有効である。

マルコフ連鎖モンテカルロ法は，不変分布が目標分布となるように，エルゴード的なマルコフ連鎖を構成し，状態の推移を繰り返すことによって，やがて目標

分布へと到達し，そこからの乱数発生を可能とする。マルコフ連鎖モンテカルロ法におけるアルゴリズムの代表的なアルゴリズムとして，メトロポリス・ヘイスティングス (metropolis–hastings; MH) アルゴリズムや，多重ブロック (multiple-block)MH アルゴリズム，ギブスサンプラー (gibbs sampler) が挙げられる。各アルゴリズムと不変分布との関係性については豊田 (2008, pp.9-20) に見ることができる。

MCMC のアルゴリズムでは，目標分布 $\pi(x)$ から直接的にサンプリングを行うことが困難な場合に，まず目標分布を近似しつつ容易にサンプリングできる分布を提案分布 (proposal distribution) として採用しサンプリングを行う。そこに提案分布と目標分布との違いを詳細釣り合い条件 (detailed balance condition) と呼ばれる条件が満たされるように修正操作を含めることで，最終的に目標分布からのサンプリングを可能とする。

ここでは豊田 (2008) の議論を基として詳細釣り合い条件を導入する。

詳細釣り合い条件

詳細釣り合い条件は，マルコフ連鎖が不変分布に収束することに関する十分条件である。詳細釣り合い条件は，状態空間に属する全ての x について，

$$\pi(x^{(t)})\Pr(x^{(t+1)}|x^{(t)}) = \pi(x^{(t+1)})\Pr(x^{(t)}|x^{(t+1)}) \tag{4.36}$$

が満たされている場合に成り立つ。このとき，ある状態 $x^{(t+1)}$ から状態 $x^{(t)}$ へ反転して推移する状況を全ての $x^{(t)}$ について周辺化すると 1 となる。そのため，詳細釣り合い条件は，

$$\int \pi(x^{(t)})\Pr(x^{(t+1)}|x^{(t)})dx^{(t)} = \pi(x^{(t+1)}) \tag{4.37}$$

であることを意味する。(4.37) 式は時点 t から $t+1$ に遷移した場合に，$x^{(t+1)}$ が従う分布は $x^{(t)}$ が従っていた分布と同様に $\pi(\cdot)$ となることを示している。

いま，提案分布を $q(x|x^{(t)})$ とする。x および $x^{(t)}$ は任意で良いため，提案分布が詳細釣り合い条件を満たしていない状況は一般的に

$$\pi(x^{(t)})q(x|x^{(t)}) > \pi(x)q(x^{(t)}|x) \tag{4.38}$$

のように仮定できる。(4.38) 式は釣り合いが崩れ，$x^{(t)}$ は候補の x へと多く推移するが，x から $x^{(t)}$ への推移は少ないという状況を表している。

この非対称性を修正するために，確率 $\alpha(x|x^{(t)})$ を導入し，$x^{(t)}$ から x への推移の量を

$$\Pr(x|x^{(t)}) = q(x|x^{(t)}) \times \alpha(x|x^{(t)}) \tag{4.39}$$

のように，$x^{(t)}$ から x へ推移する釣り合いがとれるよう調整する。

(4.39) 式左辺が詳細釣り合い条件を満たすようにするには，x から $x^{(t)}$ への推移は少ないので，調節のために $\alpha(x^{(t)}|x)$ を 1 として

$$\alpha(x|x^{(t)}) = \frac{\pi(x)q(x^{(t)}|x)}{\pi(x^{(t)})q(x|x^{(t)})}$$

のようにする。この $\alpha(x^{(t)}|x)$ を用いて，

$$\begin{aligned}\pi(x^{(t)})q(x|x^{(t)})\alpha(x|x^{(t)}) &= \pi(x)q(x^{(t)}|x)\alpha(x^{(t)}|x) \\ &= \pi(x)q(x^{(t)}|x)\end{aligned}$$

とする。

最終的に，詳細釣り合い条件が満たすために採択確率 $\alpha(\cdot|\cdot)$ について

$$\alpha(x|x^{(t)}) = \begin{cases} \min\left[\frac{\pi(x)q(x^{(t)}|x)}{\pi(x^{(t)})q(x|x^{(t)})}, 1\right] & \text{分母} > 0 \\ 1 & \text{分母} = 0 \end{cases}$$

に従い提案分布からの候補を採用しつつサンプリングを行う。

4.5 ハイブリッド・モンテカルロ法

本節では次章で用いられる推定方法である，ハイブリッドモンテカルロ (hybrid Monte Carlo; HMC; Duane, Kennedy, Pendleton & Roweth, 1987, Neal, 1996) 法について概説する。

ハイブリッドモンテカルロ法はサンプリングアルゴリズムに関するパラメタ設定に注意を要するものの，適切なパラメタ設定の下では，従来の MCMC 法におけるギブスサンプリングよりも不変分布への収束が早まるという利点を有している。

また，ギブスサンプリングの場合には，すべての母数に関して，全条件付き事後分布を導出し，それぞれサンプリングの難易度に応じて，いくつかのブロックに分割してサンプリングアルゴリズムを変更するという作業が求められることがある。しかしながら HMC 法の場合には，母数に関する事後分布の微分が与えられていればよく，左記のような調整は求められない。

4.5.1 ハミルトン関数

HMC 法では，ハミルトン力学 (Hamiltonian dynamics) を利用してマルコフ連鎖を構成し，目標分布からのモンテカルロ法を用いたサンプリングを企図する。

いま，仮想的な物体の位置に関する変数ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n)'$ の値が仮想的な時間 t の経過により変化するものとする。ここで位置変数ベクトルの各要素に対応する運動量変数ベクトル[2] $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ を導入する。位置変数と運動量変数による結合空間は位相空間と呼ばれる。

物体の運動に関して，位置エネルギー $U(\boldsymbol{x})$ は位置変数ベクトル $\boldsymbol{x}$ によって与えられ，運動エネルギー $K(\boldsymbol{\delta})$ は運動量変数ベクトル $\boldsymbol{\delta}$ によって与えられる。エネルギー保存則から，系 (system) の全エネルギーは位置エネルギーと運動エネルギーの和

$$H(\boldsymbol{x}, \boldsymbol{\delta}) = U(\boldsymbol{x}) + K(\boldsymbol{\delta}) \tag{4.40}$$

により与えられる。(4.40) 式はハミルトン関数 (Hamiltonian function; Neal, 2011)，もしくはハミルトニアン（松本, 2004）と呼ばれる。ハミルトン関数を利用することで，1 次連立微分方程式

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial \delta_i}, \quad \frac{d\delta_i}{dt} = -\frac{\partial H}{\partial x_i} \tag{4.41}$$

が得られる。(4.41) 式はハミルトンの運動方程式と呼ばれ，(4.41) 式を解くことで，物体の運動の変化を知ることが可能となる。

4.5.2 リープフロッグ法を用いた数値積分

ハミルトンの運動方程式で与えられる (4.41) 式は，実際には解析的に解くことは困難であるため，数値積分を利用して解を近似することになる。数値積分法の 1 つとしてリープフロッグ (leap-frog) 法が挙げられる。リープフロッグ法では，初期値 $\boldsymbol{x}^{(0)}, \boldsymbol{p}^{(0)}$ を定め，以下の更新式

$$\delta_i(t + \epsilon/2) = \delta_i(t) - \frac{\epsilon}{2}\frac{\partial U}{\partial x_i}[x(t)], \tag{4.42}$$

$$x_i(t + \epsilon) = x_i(t) + \epsilon\frac{\delta_i(t + \epsilon/2)}{m_i} \tag{4.43}$$

$$\delta_i(t + \epsilon) = \delta_i(t + \epsilon/2) - \frac{\epsilon}{2}\frac{\partial U}{\partial x_i}[x(t + \epsilon)] \tag{4.44}$$

に基づき，運動量変数をステップサイズ $\epsilon/2$ として半ステップ更新した後にステップサイズ ϵ による位置変数の更新を行う。これを $L = T/\epsilon$ 回繰り返し，時点 T における位置変数ベクトルと運動量変数ベクトル $\boldsymbol{x}^{(T)}, \boldsymbol{p}^{(T)}$ を得る。

なお，ハミルトン力学には，時間反転性 (reversibility)，ハミルトニアン H の不変性 (conservation of the Hamiltonian)，位相空間の体積保存 (volume preservation) という性質がある。これら 3 つの性質を崩さずに連鎖の更新を行う必要がある。リープフロッグ法は，体積保存性を保ったまま数値積分を行うことが可能である。

[2] 以下では位置変数ベクトル，運動量変数ベクトルをともに位置変数，運動量変数とも呼ぶ。

時間反転性　任意の時点 t から d の分だけ経過した時点 $t+d$ において，速度 $\boldsymbol{\delta}$ を $-\boldsymbol{\delta}$ と反転させると，往路と同じ軌跡を通り，元の時点 t へと戻る性質を時間反転性という。時点 t における状態

$$(\boldsymbol{x}(t), \boldsymbol{\delta}(t)) \tag{4.45}$$

から，時点 $t+s$ における状態

$$(\boldsymbol{x}(t+s), \boldsymbol{\delta}(t+s)) \tag{4.46}$$

へは写像の関係にあり，逆写像が存在する。また，写像

$$g : (\boldsymbol{x}', \boldsymbol{\delta}') = g(\boldsymbol{x}, \boldsymbol{\delta}) \tag{4.47}$$

に関して，

$$(\boldsymbol{x}, -\boldsymbol{\delta}) = g(\boldsymbol{x}', \boldsymbol{\delta}') \tag{4.48}$$

が成立する。

ハミルトニアン H の不変性　力学的エネルギー保存保存法則から，ハミルトニアン H の不変性が導かれる。ハミルトン力学によって提案される候補点をメトロポリス法によって採択するかを決定する際に，H が不変であれば採択確率は必ず 1 となる。ただし，実際には数値積分を行うため，必ずしも H が不変とはならない。

位相空間の体積保存　体積保存は写像 g のヤコビ行列の行列式が 1 になることと等しい (Neal, 2011)。この性質は，ハミルトン力学の枠組みにおいて時間を発展させたときに，位相空間の形状は変化しても，その体積は変化しないということである (ビショップ, 2012, p.265)。

4.5.3　ハミルトン力学を利用したモンテカルロサンプリング

ここで位置変数 $\boldsymbol{x}$ と運動量変数 $\boldsymbol{\delta}$ の位相空間における同時分布を

$$p(\boldsymbol{x}, \boldsymbol{\delta}) = \frac{1}{Z_H} \exp[-H(\boldsymbol{x}, \boldsymbol{\delta})/\iota] \tag{4.49}$$

$$= \frac{1}{Z_H} \exp[-U(\boldsymbol{x})/\iota] \exp[-K(\boldsymbol{\delta})/\iota] \tag{4.50}$$

と定義する。

$$p(\boldsymbol{z}) = \frac{1}{Z} \exp[-E(\boldsymbol{x})/\iota] \tag{4.51}$$

による表現はカノニカル分布 (canonical distribution) と呼ばれる。ここで $E(\boldsymbol{x})$ はエネルギーを表す。Z は正規化定数[3]，ι は系の温度 (temperature of the system; Neal, 2011, p.123) と呼ばれ，ここでは $\iota = 1$ とする。位置変数 $\boldsymbol{x}$ がパラメタとなるため，その事後分布が目標分布となる。いま事前分布 $\pi(\boldsymbol{x})$ と尤度関数 $L(\text{Data}|\boldsymbol{x})$ により，パラメタ $\boldsymbol{x}$ の事後分布が $\pi(\boldsymbol{x}|\text{Data}) \propto L(\text{Data}|\boldsymbol{x})\pi(\boldsymbol{x})$ と表されるとすると，位置エネルギーの事後分布を

$$U(\boldsymbol{x}) = -\log[L(\text{Data}|\boldsymbol{x})\pi(\boldsymbol{x})] \tag{4.52}$$

のように定義できる。また，運動エネルギーは

$$K(\boldsymbol{\delta}) = \sum_{i=1}^{n} \frac{\delta_i^2}{2m_i} \tag{4.53}$$

と定義される。ここで m は質量に対応するパラメータである（松本, 2004, p.149, Neal, 2011, p.114）。(4.52) 式と (4.53) 式を用いて同時分布 (4.50) 式から乱数を発生させることで，HMC 法を用いた MCMC 標本を得ることが可能となる。

4.5.4 HMC 法のアルゴリズム

ここで，HMC 法を用いたサンプリングにおけるアルゴリズムを示す。HMC 法における繰り返しは大別して 2 つのステップから成り立っている。

1. $\boldsymbol{x}$ の初期値 $\boldsymbol{x}^{(0)}$，ステップサイズ ϵ，および更新回数 L を定める。
2. 時点 $t = 1, 2, \ldots, T$ において，以下の手順を繰り返す。
   (a) 運動量変数に関する初期値 $\boldsymbol{\delta}^{(0)}$ を多変量正規分布 $N(\boldsymbol{0}, \boldsymbol{I}_n)$ より発生させる。
   (b) $\boldsymbol{x}^{(t-1)}$ と $\boldsymbol{\delta}^{(0)}$ を初期値とし，リープフロッグ法により，候補点 $\tilde{\boldsymbol{x}}$，$\tilde{\boldsymbol{\delta}}$ を求める。
   (c) 採択確率

   $$\alpha = \min\{1, \exp[-H(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\delta}}) + H(\boldsymbol{x}^{(t-1)}, \boldsymbol{\delta}^{(0)})]\} \tag{4.54}$$

   を計算する。
   (d) 一様乱数 u を $\text{Uniform}(0, 1)$ から発生させ，$\boldsymbol{x}^{(t)}$ を以下に基づいて決定する。

$$\boldsymbol{x}^{(t)} = \begin{cases} \tilde{\boldsymbol{x}} & u \leq \alpha \text{の場合} \\ \boldsymbol{x}^{(t-1)} & \text{それ以外の場合} \end{cases} \tag{4.55}$$

[3] $p(\boldsymbol{z})$ の和もしくは積分を 1 に調整するために導入される (Neal, 2011, p.123)。

なおアルゴリズム 2(a) は，エルゴード的なマルコフ連鎖を構成するために必要な手順である。また，アルゴリズム 2(c) は，H の不変性が，数値積分を行ったことによる誤差が原因で満たされないことを防ぐための手順である。この更新手順において，誤差によるバイアスを取り除くことで，不変分布からのサンプリングが保たれることとなる。このステップにおいて，詳細釣合条件が満たされることとなる [4]。

なお，事後統計量を算出する際には，同時分布 $p(\boldsymbol{x}, \boldsymbol{\delta})$ から乱数を発生させ，運動量変数に関する乱数を破棄し，位置変数に関するサンプルのみを利用する。

4.5.5 連鎖の収束

本項ではマルコフ連鎖の判断基準について，Gelman & Rubin (1992)，Gelman (1996)，豊田 (2008) に基づき，指標の構成方法および推測方法に関して概説する。

マルコフ連鎖モンテカルロ法およびハイブリッドモンテカルロ法を用いた推測では，構成，推移した連鎖が目標分布へと到達しているのか，そして目標分布の全範囲を十分に被覆できているのかを確認することが肝要である。

連鎖を更新しサンプリングを行った結果，得られる連鎖の構成要素は，不変分布からの無作為標本となっていることが期待される。不変分布到達後の連鎖の構成要素には，初期値の影響が残っておらず，不変分布からの無作為標本と見なせる状態に至ったとき，連鎖が収束したという。

ただし，更新開始初期の時点では，分析者が恣意的に設定した初期値の影響が残っており，連鎖が未だ目標分布へと到達していないことが予想される。

例え最新の構成要素が不変分布からの無作為標本であったとしても，初期値からの影響が残る連鎖の構成要素をも推測に用いることは，誤った結論を導くことに繋がりかねない。これに対処するためには，十分な更新回数を経た後の連鎖から，初期の構成要素を取り除き，不変分布からの無作為標本となっている部分のみを用いて推測を行う方法が採られる。

この破棄される更新期間をバーンイン期間 (burn-in period) という。パラメタ推定には，このバーンイン期間以後の構成要素を用いて行う。なお，バーンイン期間はウォームアップ期間 (warm-up period) と呼称される場合もある。

連鎖の初期要素を破棄するという方略自体は単純かつ効果的ではあるものの，適切なバーンイン期間を予め定める統一的な方法および基準は未だ存在しない。そのため，実際に利用場面では，分析者が任意に連鎖の更新期間を定め，かつ設定したバーンイン期間に基づき，連鎖の初期要素を破棄することとなる。

その後，残りの構成要素が収束状態に至っているか否かを判定することで，推測に利用する連鎖が収束していることを事後的に確認する。

[4] 詳細釣合条件が成立することの証明は Neal (2011) において確認することができる

収束判定の方法には様々なものが提案されているが，連鎖が収束していることを明確に判定できる方法は未だ見出されていない。

収束判定指標 R

Gelman & Rubin (1992) は，複数の連鎖を発生させ，それらの分散を比較に用いることで連鎖の収束を判定する方法およびその指標 $\hat{R}$ を提案した。なお $\hat{R}$ は estimated potential scale reduction (EPSR; Gelman & Rubin, 1992) とも呼ばれる。本指標に関しては，Gelman & Rubin(1992) においては，$n \to \infty$ となるに従い収縮する，保守的な事後分布の尺度という表現も見られる。ここでは Gelman & Rubin(1992)，Gelman(1996) および豊田 (2008) を元に収束判定指標 $\hat{R}$ の算出方法について述べる。

いま，初期値を任意に設定し，パラメタの標本を発生させたマルコフ連鎖を，K 本構成するものとする。パラメタ θ に関する，事後分布からのサンプリング標本を用いて，

$$[\theta_k^{(1)}, \theta_k^{(2)}, \ldots, \theta_k^{(T)}]', \quad k = 1, \ldots, K \tag{4.56}$$

のように K 本構成する。ここで T は連鎖の更新期間を表す。これらの K 本の連鎖それぞれから，初期の m 回分の標本をバーンイン期間として破棄するものとする。ここで，破棄後の $T - m$ 個の標本を含む残りの連鎖が，不変分布へと収束しているかどうかを判定したい。

まずは系列内平均 (with-in sequence mean) $\mathrm{Var}_B(\theta)$ を

$$\widehat{\mathrm{Var}}_B(\theta) = \frac{T-m}{K-1} \sum_{k=1}^{K} (\overline{\theta}_k^{(.)} - \overline{\theta}_.^{(.)})^2 \tag{4.57}$$

のように求める。(4.57) 式は異なる連鎖間における θ の値の散らばりを与える。次に系列内分散 (within-sequence variance) $\mathrm{Var}_W(\theta)$ の推定値を

$$\widehat{\mathrm{Var}}_W(\theta) = \frac{1}{K} \sum_{k=1}^{K} s_i^2 \tag{4.58}$$

$$= \frac{1}{K} \sum_{k=1}^{K} \frac{1}{T-m-1} \sum_{t=1}^{T-m} (\theta_k^{(t)} - \overline{\theta}_.^{(.)})^2 \tag{4.59}$$

$$= \frac{1}{K(T-m-1)} \sum_{k=1}^{K} \sum_{t=1}^{T-m} (\theta_k^{(t)} - \overline{\theta}_.^{(.)})^2 \tag{4.60}$$

のように求める。(4.60) 式は連鎖内における値の散らばりの推定値を表している。

なお，式中の $\overline{\theta}_k^{(\cdot)}$ は k 番目の連鎖に含まれる m 個を破棄した後の全要素の平均

$$\overline{\theta}_k^{(\cdot)} = \frac{1}{T-m}\sum_{t=1}^{T-m}\theta_{kt} \tag{4.61}$$

を表している。また，$\overline{\theta}_{\cdot}^{(\cdot)}$ は K 本の連鎖ごとの K 個の全要素平均を表し，

$$\overline{\theta}_{\cdot}^{(\cdot)} = \frac{1}{K}\sum_{k=1}^{K}\overline{\theta}_{k\cdot}^{(\cdot)} \tag{4.62}$$

である。

これら分散成分 $\widehat{\mathrm{Var}}_B(\theta)$, $\widehat{\mathrm{Var}}_W(\theta)$ を用いて，Gelman & Rubin (1992) では目標分布のパラメタ θ の真の分散 $\mathrm{Var}(\theta)$ を推定する方法を述べている。まずはじめに

$$\widehat{\mathrm{Var}}(\theta) = \frac{T-m-1}{T-m}\widehat{\mathrm{Var}}_W(\theta) + \frac{1}{T-m}\widehat{\mathrm{Var}}_B(\theta) \tag{4.63}$$

を構成する。ここで (4.63) 式が $\widehat{\mathrm{Var}}(\theta)$ の不偏推定量となるのは，K 本の連鎖が目標分布へと収束している場合である。$\mathrm{Var}(\theta)$ は過剰な散らばりに対する保守的な推定量 (Gelman & Rubin, 1996) とされる。K 本の連鎖のうち，非収束の連鎖がある場合は，連鎖間の同一要素の値が異なり得るため，$\widehat{\mathrm{Var}}_B(\theta)$ が過剰に推定される。このことから，(4.63) 式の値も真の $\mathrm{Var}(\theta)$ より大きくなり得る。

推定量 (4.63) 式の $\widehat{\mathrm{Var}}(\theta)$ と (4.60) 式の $\widehat{\mathrm{Var}}_W(\theta)$ は $\mathrm{Var}(\theta)$ に対する，異なる側面からの推定方法を提供している。そこで，これら分散成分の推定値を用い，真の分散 $\mathrm{Var}(\theta)$ に対する推測指標として $\widehat{\mathrm{Var}}(\theta)$ と $\widehat{\mathrm{Var}}_W(\theta)$ を用いて，

$$\hat{R} = \sqrt{\frac{\widehat{\mathrm{Var}}(\theta)}{\widehat{\mathrm{Var}}_W(\theta)}} \tag{4.64}$$

のように比 $\hat{R}$ を構成する。

もし，このとき連鎖の更新回数が不十分であり，推測に用いる連鎖の内に非収束の連鎖が含まれている場合には，目標分布の全範囲からのサンプリングが十分に行われていないことが予想されるため，$\widehat{\mathrm{Var}}_W(\theta)$ の値は真値 $\mathrm{Var}(\theta)$ に対して過少に推定されることとなる。

一方で，推測に用いた連鎖が全て同一の不変分布に収束している場合には，$\widehat{\mathrm{Var}}_W(\theta)$ の値は，真の分散 $\mathrm{Var}(\theta)$ に等しくなることが予想される。

もしすべての連鎖が完全に収束しているならば，(4.64) 式を構成する要素はどちらも $\mathrm{Var}(\theta)$ に一致し，このとき $\hat{R}=1$ となる。よって $\hat{R}$ が十分に 1 に近け

ればマルコフ連鎖は収束しているものと判断し，そうでないならば連鎖は非収束であると判断可能である。

Gelman (1996) は，シミュレーションによる $\hat{R}$ の検討を行い，$\hat{R}$ が 1.2 以上の場合には，連鎖の非収束が示唆されたことから，$\hat{R}$ の値が 1.2，もしくは 1.1 よりも小さければ連鎖は収束したと見なす判断基準を提案した。

なお，構成した連鎖が 1 本の場合は，要素を複数に分割し，それらの要素集合を擬似的に構成された複数連鎖と見なすことで，上述の手順を適用することで収束判定を行うことが可能である。

その他の収束判定指標として Geweke (1992) による方法，Heidelberger & Weltch (1983) による方法，Raftery & Lewis (1992a, 1992b) による方法が提案されている。

第5章　同一項目に複数の項目特性図が作成可能な場合の分析方法（研究 III）

5.1　研究 III 目的

テスト実施会場が広い地域に分かれている場合や，多数の受験者数が見込まれるために，同一時刻，会場でのテストの実施が困難となる場合には，ブックレット (booklet, BL) 方式によってテストが実施されることがある。ブックレット方式とは，同一の特性を測定するための項目群を複数の小冊子（ブックレット）に分割して配置し，複数の受験者群にブックレットを配布し，項目に対して回答を求める形式である。項目内容の漏洩からテストを守るために，ブックレットごとの構成項目は異なる場合が一般的である。つまり，配布されるブックレットが異なる受験者は異なる項目集合に対して回答を行う。

これら構成項目が異なり得るブックレット間の受験者群の特性（能力）を比較可能とする方法は，ブックレット間で同一の受験者にテストを実施する共通受験者法か，もしくは各ブックレットに共通の項目を含ませる共通項目法である。

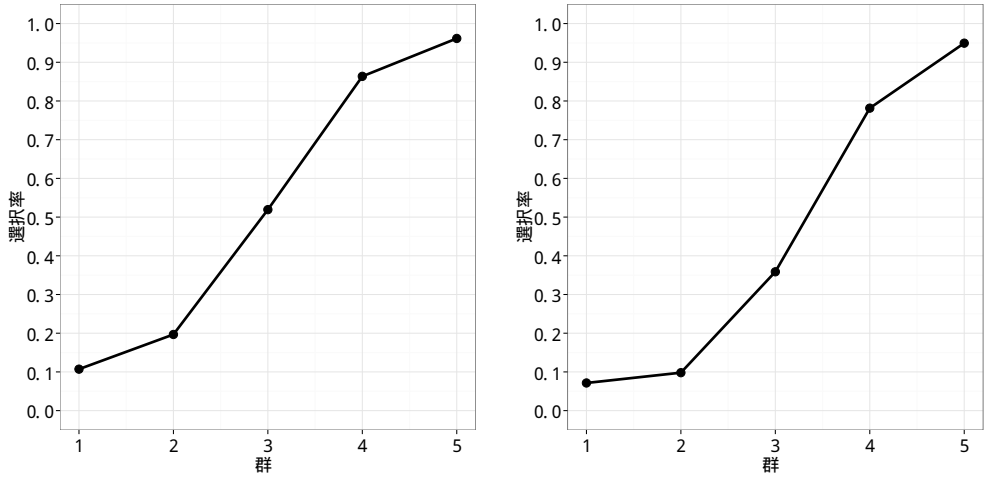
共通受験者法の場合，複数のブックレットへ回答する人員を手配する必要がある。また，ブックレット数が多くなると，共通受験者の負担も増し，疲労によるパフォーマンスの低下の可能性について考慮，対策する必要が生じる。

共通項目法では複数のブックレット内に，共通の項目を配置することとなるが，項目分析を行う場合に，ブックレットに対応した数の分析指標が算出されることとなる。これは項目特性図も例外ではなく，共通項目についてはブックレット数分の項目特性図が描画可能となる。同一項目の項目特性図であっても，受験者群が異なれば，特性曲線の表現も異なるものとなり得る。

例えば，図 5.1 は，PISA 2003 における項目「輸出」のブックレットごとの項目特性図である [1]。ここでは例示のため 5 群分割としている。なお，項目「輸出」の項目内容は付録付録に示した。項目「輸出」はそれぞれ項目数，受験者数の異なる 4 種類のブックレットに含められている。本研究においては，便宜上，共通項目が含まれるブックレットを区別するための番号を，1 から 4 に振り

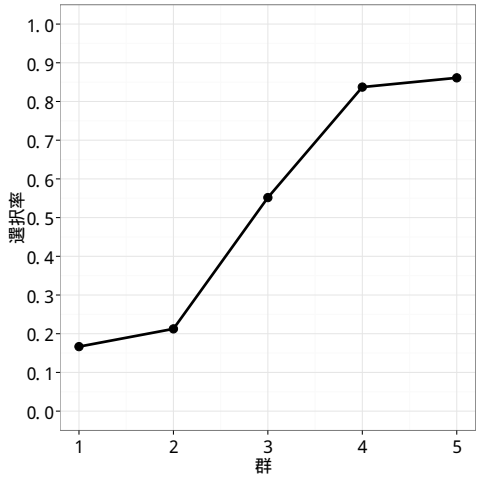
[1] 各ブックレットについては 1 問につき配点を 1 点とした。

直している。

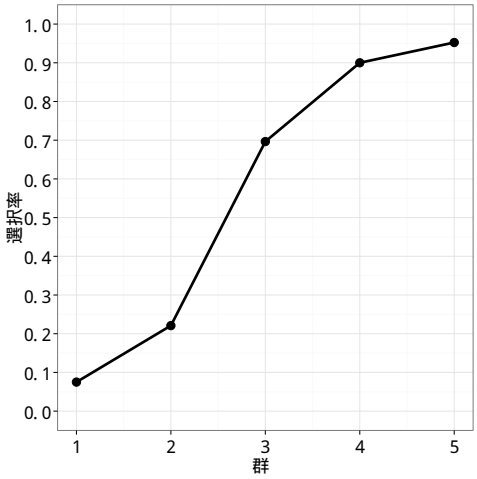


(i) BL 1

(ii) BL 2



(iii) BL 3



(iv) BL 4

図 5.1 ブックレット別項目特性図（PISA 2003 項目「輸出」）

図 5.1 を観察すると，項目「輸出」は下位群，高位群の識別力はやや低く，中位群の識別に適した項目であることが示唆される。しかしながら，ブックレットごとの項目特性図は互いに似た傾向を示してはいるものの，同時にブックレット間で特性表現の揺らぎが生じている様子も観察される。

このように，同一項目であっても，ブックレット間で項目特性の表現が異なり得る。共通項目が含まれるブックレットが多い場合には，1 つの項目についてすべてのブックレットの表現を確かめることは煩雑である。更に，ブックレット

間で共通項目の項目特性図の表現が大きく異なるような場合には，項目特性の安定性が疑われることとなり，安定性それ自体を吟味することも求められるであろう。しかしながら，多くのブックレットの表現を加味しつつ項目特性を解釈することは困難である。このような場合に，どのブックレットの表現を項目分析の対象として採用し，解釈すべきか，明確な基準は広く知られてはいない。

理想的な状況では，同程度の特性を有する複数の受験者集団に対して，共通項目を含んだブックレットを用いてテストを実施すると，例えブックレットが異なったとしても，ブックレットが割り当てられた受験者群間の背後に仮定される分布が互いに同一であるならば，受験者数が増加するにしたがい，項目特性図の表現も互いに似通うようになるはずである。すなわち，この場合にはブックレットごとの，項目特性図の表現の違いは標本誤差のみによるものということができる。

しかしながら，実際の状況ではブックレットの横軸，つまり群分けはブックレット得点によって定められ，ブックレット得点は，ブックレット間で互いに異なり得る（項目特性図作成対象となっている当該項目以外の）他の項目群に対する正誤反応およびその背後に想定される項目特性によって定められる。

よって，項目特性図の横軸にはそれぞれ異なる項目特性の影響が含まれていると考えられるために，各ブックレットごとの項目特性図の表現の異なりは，純粋に標本誤差のみの影響とは見なすことが難しい状況にある。

上記の理由から，ブックレットごとの項目特性図の表現は互いに完全に異なったものを表していると一概に見なすことは難しい。一方で，ブックレットごとの項目特性図における表現は，標本誤差以外はすべて同じであると見なすこともまた困難である。

研究 Ⅲ では共通項目について複数の項目特性図が作成され得る場合に，異なる受験者集団に対する，ブックレット別項目特性図の表現の違いを加味しつつ，同時に統一的な観点から項目特性図を作成し，項目分析に用いる方法を提案する。なお，本研究では正答分析に焦点を当てることとする。

5.2 階層ベイズモデルを用いた統一的な項目特性図作成法の提案（方法 Ⅲ）

同程度の特性を有する複数の受験者集団に対して，共通項目を含んだブックレットを用いてテストを実施すると，例えブックレットが異なったとしても，同じ分布となるという仮定の下，階層ベイズによる群ごとの 2 値反応データを用いたモデル化を行い，超パラメタを用いた解釈方法について提案する。

5.2.1 モデル

任意の項目が複数の B 個のブックレットに共通項目として含まれている状況を想定する。なお，ここでは 1 つの項目のみに着目するため，項目に関する添え字は省略する。

いま，ある項目の，第 b 番目 $(b = 1, 2, \ldots, B)$ のブックレットにおける，第 g 群 $(g = 1, 2, \ldots, G)$ に所属する，第 i 番目 $(i = 1, 2, \ldots, N_{bg})$ の受験者の 0-1 の 2 値反応を x_{bgi} と表す。つまり，

$$\boldsymbol{x}_{bg} = (\boldsymbol{x}_{bg1}, \boldsymbol{x}_{bg2}, \ldots, \boldsymbol{x}_{bgN_{bg}}) \tag{5.1}$$

である。

更に，個々の第 g 群における正答人数を n_{bg} と表す。ここで $n_{bg} = \sum_{i=1}^{N_{bg}} x_{bgi}$ である。また，個々の第 g 群における正答反応 $(x_{bgi} = 1)$ の比率（正答率）を p_{bg} と表す。つまり，

$$p_{bg} = \frac{n_{bg}}{N_{bg}} \tag{5.2}$$

である。すると，ブックレット b，群 g において，反応ベクトル $\boldsymbol{x}_{bg}$ から求められる正答数 n_{bg} が観察される確率は，正答率 p_{bg} が所与の下で

$$f(n_{bg}|N_{bg}, p_{bg}) = \left(\begin{array}{c} N_{bg} \\ n_{bg} \end{array} \right) p_{bg}^{n_{bg}} (1 - p_{bg})^{(N_{bg} - n_{bg})} \tag{5.3}$$

と与えられる。(5.3) 式についてベイズ推定を行うものとする。

p_{bg} の事前分布には

$$p_{bg} \sim Beta(\alpha_g, \beta_g) \tag{5.4}$$

とし，ベータ分布を仮定する。また，ベータ分布の形状パラメタ α_g および β_g の事前分布として，

$$\alpha_g \sim Uniform(0, 100) \tag{5.5}$$

$$\beta_g \sim Uniform(0, 100) \tag{5.6}$$

のように一様分布を仮定する。p_{bg} に関して周辺化すると，

$$f(n_{bg}|\alpha_g, \beta_g) = \int f(\boldsymbol{n}_{bg}|p_{bg}) f(p_{bg}|\alpha_g, \beta_g) dp_{bg} \tag{5.7}$$

となる。ブックレット，群ごとの正答数を要素とするベクトルを

$$\boldsymbol{n} = (n_{11}, \ldots, n_{1G}, n_{21}, \ldots, n_{2G}, n_{B1}, \ldots, n_{BG}) \tag{5.8}$$

とすると，周辺尤度は

$$L(\boldsymbol{n}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{g=1}^{G}\prod_{b=1}^{B} f(n_{bg}|\alpha_g,\beta_g) \tag{5.9}$$

となり，パラメタの同時事後分布は

$$f(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{n}) \propto L(\boldsymbol{n}|\boldsymbol{\alpha},\boldsymbol{\beta}) f(\boldsymbol{\alpha}) f(\boldsymbol{\beta}) \tag{5.10}$$

となる。

p_{bg} の全ブックレットの群別平均値 p_g は事前分布としてベータ分布を仮定しているため，α_g および β_g の推定値を用いて，

$$\hat{p}_g = \frac{\hat{\alpha}_g}{\hat{\alpha}_g + \hat{\beta}_g} \tag{5.11}$$

により与えられる。

また，$\hat{p}_g$ の散らばりの程度は，ベータ分布の分散公式を用いて，

$$\hat{\sigma}^2_{p_g} = \frac{\hat{\alpha}_g \hat{\beta}_g}{(\hat{\alpha}_g + \hat{\beta}_g)^2(\hat{\alpha}_g + \hat{\beta}_g + 1)} \tag{5.12}$$

により算出される。またその標準偏差は

$$\hat{\sigma}_{p_g} = \sqrt{\hat{\sigma}^2_{p_g}} \tag{5.13}$$

により与えられる。

p_g はブックレットに依存しない第 g 群の正答率として解釈することができる。換言すれば，p_g を用いることで，ブックレットごとに異なる項目特性図の背後に仮定される，共通した観念的な項目特性図を描くことが可能となる。

p_g を用いた項目特性図を項目分析の対象とすることで，同一項目に対するブックレット間の違いを加味しつつ，統一的な視点から解釈を行うことができる。

なお，本研究では，群への分割方法は群間が等間隔となるように，パーセンタイル点を用いて群分けを行う。また，群数は 5 とした。

5.2.2 ハイブリッドモンテカルロ法による推定

本研究では，前述の階層ベイズモデルについて，HMC 法によるパラメタ推定を行う。HMC 法の実行にはソフトウェア Stan (Stan Development Team, 2014b) および統計解析環境 R (Ihaka & Gentleman, 1996) 上における Stan のインターフェースパッケージ RStan (Stan Development Team, 2014a) を用いた。

連鎖の構成数は 4 とした。また，各連鎖の更新期間は 2200 回とし，このうち初めの 200 回をバーンイン (burn-in) 期間（あるいはウォームアップ (warm-up) 期間）として破棄した。なお間引きについては間隔を 1 とし，行わないものとした。以上の設定より，各連鎖におけるバーンイン期間後のサンプル 8000 (4×2000) 個を用いて事後統計量を構成した。

5.3 結果と考察

本節では HMC 法によって求められた事後統計量を示し，それらの推定結果について考察を行う。

5.3.1 適用例 PISA 2003 項目「輸出」

まず，PISA 2003 項目「輸出」に対する提案手法の適用結果を示す。

項目「輸出」データ

表 5.1 および表 5.2 に項目「輸出」に関して，推定に用いるデータを示した。表 5.1 は 5 群に分けた際の，各群の所属人数を表している。また，表 5.2 は当該群内において正答した受験者数を表している。

表 5.1 項目「輸出」 ブックレット・群別受験者数

	BL1	BL2	BL3	BL4
g_1	28	14	36	40
g_2	61	51	80	86
g_3	104	92	87	112
g_4	110	119	86	70
g_5	52	79	36	21
N_b	355	355	325	329

表 5.2 項目「輸出」 ブックレット・群別正答者数

	BL1	BL2	BL3	BL4
g_1	3	1	6	3
g_2	12	5	17	19
g_3	54	33	48	78
g_4	95	93	72	63
g_5	50	75	31	20

表 5.3 項目「輸出」 p_{bg} 推定結果

p_{bg}	平均	SD	$\hat{R}$
p_{11}	0.120	0.042	1.000
p_{21}	0.115	0.046	1.000
p_{31}	0.138	0.041	1.000
p_{41}	0.106	0.036	1.001
p_{12}	0.196	0.039	1.000
p_{22}	0.158	0.039	1.000
p_{32}	0.204	0.036	1.000
p_{42}	0.209	0.035	1.000
p_{13}	0.525	0.041	1.000
p_{23}	0.429	0.050	1.000
p_{33}	0.545	0.044	1.000
p_{43}	0.639	0.047	1.000
p_{14}	0.852	0.028	1.000
p_{24}	0.803	0.030	1.000
p_{34}	0.836	0.032	1.000
p_{44}	0.866	0.032	1.000
p_{15}	0.939	0.026	1.000
p_{25}	0.937	0.023	1.000
p_{35}	0.901	0.034	1.000
p_{45}	0.930	0.033	1.000

表 5.4 項目「輸出」 $\hat{p}_g$ および $\hat{p}_{bg}$

群	$\hat{p}_g$	$\hat{\sigma}^2_{p_g}$	$\hat{\sigma}_{p_g}$	p_{1g}	p_{2g}	p_{3g}	p_{4g}
1	0.124	0.002	0.040	0.120	0.115	0.138	0.106
2	0.195	0.002	0.046	0.196	0.158	0.204	0.209
3	0.535	0.006	0.075	0.525	0.429	0.545	0.639
4	0.836	0.002	0.044	0.852	0.803	0.836	0.866
5	0.922	0.001	0.034	0.939	0.937	0.901	0.930

項目「輸出」推定結果

表 5.3 に p_{bg} の推定結果となる事後統計量を示した。HMC における収束判定のために，Gelman & Rubin (1992) において提案された収束判定指標 $\hat{R}$ を用いた。判定にあたっては Gelman (1996) を参考に，$\hat{R}$ が 1.1 よりも小さい値であれば，当該推定値に関する連鎖は収束したものと判断した。

項目「輸出」に関しては，すべての推定値に関して $\hat{R}$ が 1.1 以下であったため，連鎖は収束したものと判断し，分析結果の解釈に用いることとした。表 5.4 の 2 列目から 4 列目には，(5.11) 式，(5.12) 式および (5.13) 式によって算出される推定値 $\hat{p}_g$，$\hat{\sigma}^2_{p_g}$ および $\hat{\sigma}_{p_g}$ を配した。また，5 列目以降はブックレットごとの p_{bg} の推定値を示している。

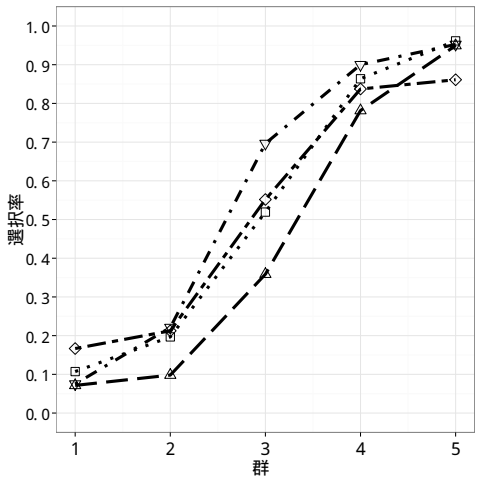
図 5.2(i) は観測データの選択比率（正答率・正答反応率）を用いて作成した項目特性図である。また，図 5.2(ii) の実線は表 5.4 に示した事後統計量を用いて作成した（観念的な）推定項目特性図である。図 5.2(ii) にはブックレットごとの正答反応確率の推定値 $\hat{p}_{bg}$ を用いた特性曲線についても点線で描画している。図 5.2(iii) は $\hat{p}_g$ を用いた特性曲線のみを描いた推定項目特性図である。

図 5.2(i) と図 5.2(ii) を比較すると，推定項目特性図では，選択率から直接作図される項目特性図によって表現され得る特徴（項目特性）に関しても，表現し得ていることがうかがえる。下位（低特性）群において一旦正答率が下降する様子や，中位（中特性）群において，ブックレット間の特性曲線の乖離の大きさも推定項目特性図によって表現されている。また，上位（高特性）群では特性曲線の傾きが小さくなる傾向も表現されている。

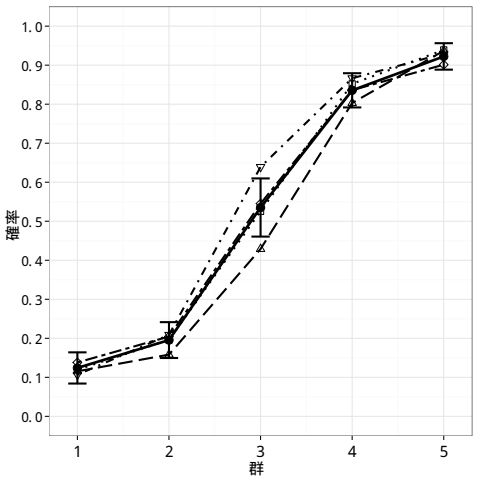
推定項目特性図上には，$\hat{p}_g$ の正答率の揺れの程度を表す標準偏差 $\hat{\sigma}_{p_g}$ を描画している。標準偏差の範囲から，$\hat{p}_g$ がどの程度ブックレット間で安定した指標であるのかも確認することが可能である。図 5.2(iii) における $\hat{p}_g$ および $\hat{\sigma}_g$ を用いた特性曲線はブックレットごとの特性曲線の特徴を統合的に表現したものと見なせるであろう。

図 5.2(iii) を用いた項目分析より，項目「輸出」は主に中特性者を識別することに適した項目であること，低特性者および高特性者については識別力が低い項目であることが分かる。その一方で，中位群に関しては $\hat{\sigma}_{p_g}$ の値が比較的大きい傾向にあるため，受験するブックレットが異なると，同じ群に所属していも正答率にばらつきが生じる傾向にあることが示唆された。

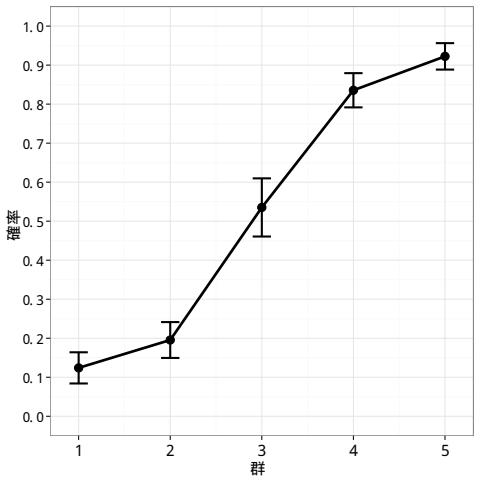
低特性者および高特性者に関しては背後に仮定される能力が中特性者と比較して，安定していることが考えられるため，こうした傾向が項目特性に表れたものと解釈される。



(i) 項目特性図



(ii) 推定項目特性図 (BL 別特性曲線付き)



(iii) 推定項目特性図

図 5.2 項目特性図と推定項目特性図（PISA 2003 項目「輸出」）

5.3.2　適用例 PISA 2003 項目「キャンディ」

項目「キャンディ」データ

表 5.5 および表 5.6 に項目「キャンディ」に関して，推定に用いるデータを示した。

表 5.5 項目「キャンディ」 ブックレット・群別受験者数

	BL1	BL2	BL3	BL4
g_1	17	30	29	49
g_2	64	61	115	96
g_3	87	89	114	114
g_4	130	117	86	71
g_5	54	56	10	21
N_b	352	353	354	351

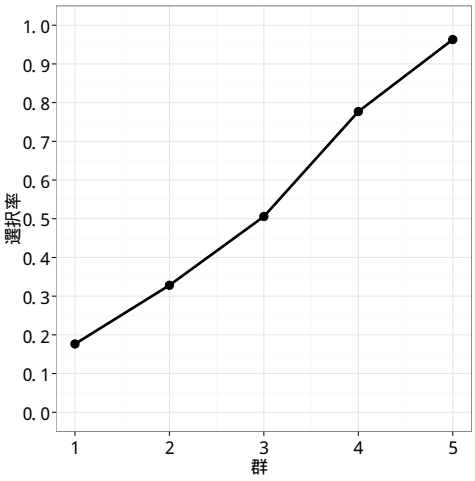
表 5.6 項目「キャンディ」 ブックレット・群別正答者数

	BL1	BL2	BL3	BL4
g_1	3	10	5	15
g_2	21	22	45	47
g_3	44	50	77	87
g_4	101	97	79	68
g_5	52	54	10	21

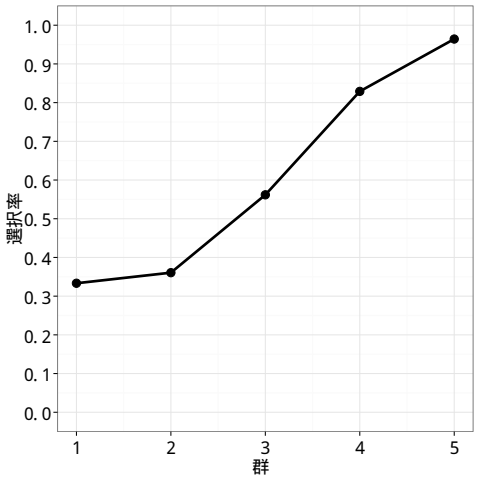
項目「キャンディ」推定結果

図 5.3 は PISA 2003 における項目「キャンディ」のブックレットごとの項目特性図である。

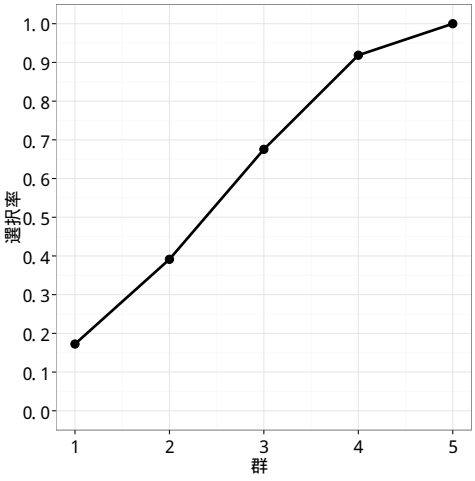
これらの項目特性図を観察すると，図 5.3(i) は特性曲線が下位群から上位群まで直線状に表現されており，受験者を識別することに適した項目であることが分かる。一方で，図 5.3(ii) では下位群について識別力が低い様子が伺える。図 5.3(iii) と図 5.3(iv) は双方ともに同様の特性曲線の形状を示しており，特に上位群の識別力がやや低い傾向にあることが見て取れる。一方で図 5.3(iv) では第 1



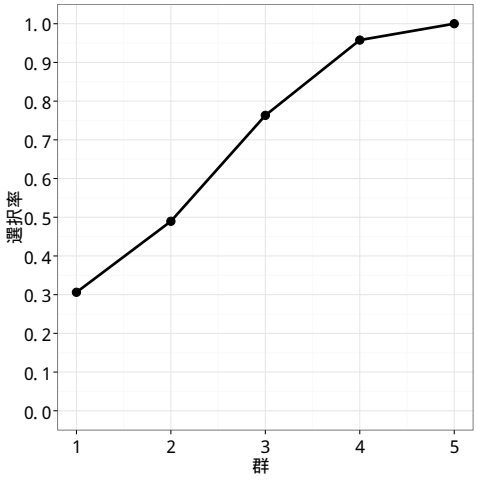
(i) BL 1



(ii) BL 2



(iii) BL 3



(iv) BL 4

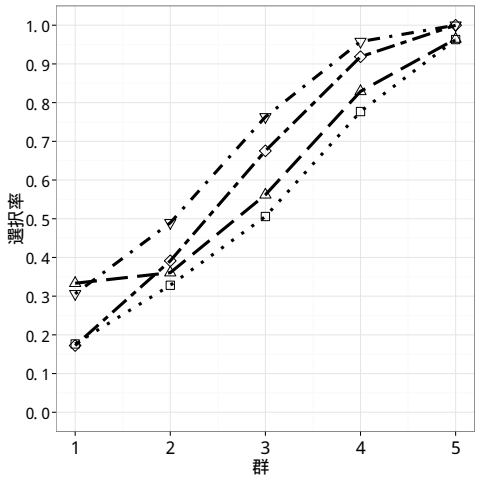
図 5.3 項目特性図（PISA 2003 項目「キャンディ」）

表 5.7 項目「キャンディ」 p_{bg} 推定結果

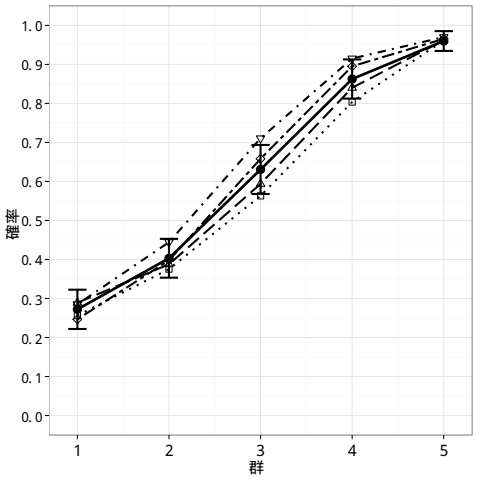
p_{bg}	平均	SD	$\hat{R}$
p_{11}	0.256	0.058	1.000
p_{21}	0.289	0.054	1.000
p_{31}	0.247	0.055	1.000
p_{41}	0.285	0.048	1.000
p_{12}	0.375	0.044	1.000
p_{22}	0.388	0.044	1.001
p_{32}	0.398	0.037	1.001
p_{42}	0.444	0.040	1.000
p_{13}	0.563	0.045	1.001
p_{23}	0.594	0.043	1.000
p_{33}	0.658	0.038	1.000
p_{43}	0.711	0.039	1.000
p_{14}	0.804	0.032	1.000
p_{24}	0.840	0.029	1.000
p_{34}	0.895	0.029	1.000
p_{44}	0.915	0.030	1.000
p_{15}	0.961	0.021	1.000
p_{25}	0.962	0.020	1.000
p_{35}	0.965	0.027	1.000
p_{45}	0.970	0.023	1.000

表 5.8 項目「キャンディ」 $\hat{p}_g$ および $\hat{p}_{bg}$

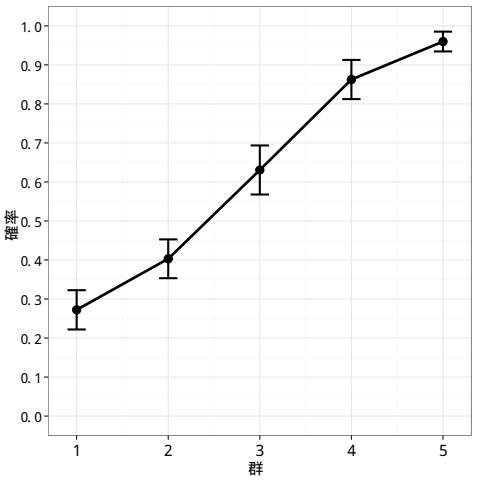
群	$\hat{p}_g$	$\hat{\sigma}^2_{p_g}$	$\hat{\sigma}_{p_g}$	p_{1g}	p_{2g}	p_{3g}	p_{4g}
1	0.272	0.003	0.050	0.256	0.289	0.247	0.285
2	0.403	0.002	0.050	0.375	0.388	0.398	0.444
3	0.631	0.004	0.063	0.563	0.594	0.658	0.711
4	0.863	0.003	0.050	0.804	0.840	0.895	0.915
5	0.960	0.001	0.025	0.961	0.962	0.965	0.970



(i) 項目特性図



(ii) 推定項目特性図 (BL 付き)



(iii) 推定項目特性図

図 5.4 項目特性図と推定項目特性図（PISA 2003 項目「キャンディ」）

群の正答率が 0.3 以上あり，図 5.3(ii) と同様にやや平易，もしくは当て推量による正答の可能性が示唆される。

項目「キャンディ」について,「方法 Ⅲ」において提案した階層ベイズモデルおよび推定を行った結果の推定値を表 5.7 と表 5.8 に掲載する。

表 5.7 より，すべての推定値について $\hat{R}$ の値が 1.1 以下となったため，連鎖は収束したものと判断した。表 5.8 から，上位群については $\hat{p}_g$ の値がブックレット間で相対的に安定していることが分かる。また，第 3 群の正答率の揺れが比較的大きいことが表れている。

図 5.4 に，項目「キャンディ」における $\hat{p}_g$ を用いて作成した推定項目特性図を示した。図 5.4(i) と図 5.4(ii) を観察すると，図 5.4(i) の場合には，各群の正答率はブックレット間で安定していない傾向が窺える。一方で図 5.4(ii) の場合には，ブックレット間の正答率の散らばりも加味しつつ，観念的な特性曲線を推定できている様子が見られる。

推定項目特性図から，項目「キャンディ」もまた，全体的に受験者をよく識別することが可能であり，その中でも中位群の識別に適した項目であることが示唆された。ただし項目「輸出」と同様に，中位群の推定値の幅がやや広いこと，上位群と比較して，下位群の識別はやや推定の幅が広いことに留意する必要があるだろう。

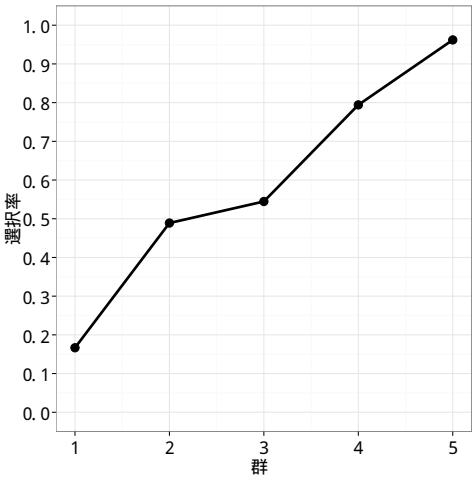
5.3.3 適用例 PISA 2003 項目「地震」

項目「地震」データ

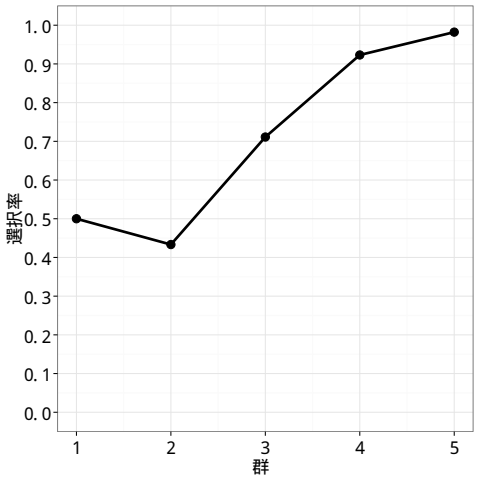
表 5.9 および表 5.10 に項目「地震」に関して，推定に用いるデータを示した。

表 5.9 項目「地震」 ブックレット・群別受験者数

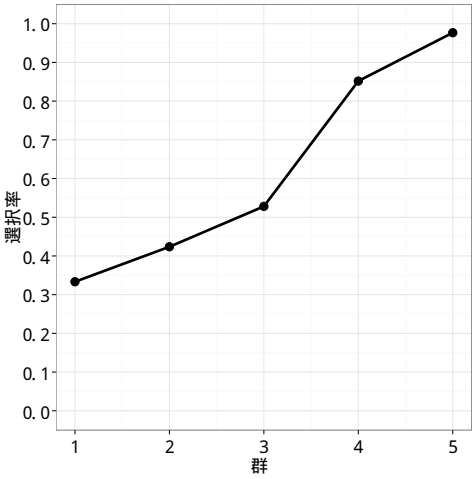
	BL1	BL2	BL3	BL4
g_1	24	26	21	21
g_2	45	60	59	34
g_3	101	90	89	89
g_4	102	117	135	83
g_5	79	56	43	50
N_b	351	349	347	277



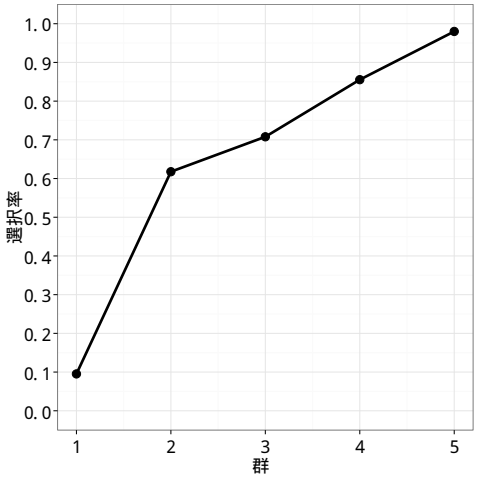
(i) BL 1



(ii) BL 2



(iii) BL 3



(iv) BL 4

図 5.5 項目特性図（PISA 2003 項目「地震」）

表 5.10 項目「地震」 ブックレット・群別正答者数

	BL1	BL2	BL3	BL4
g_1	4	13	7	2
g_2	22	26	25	21
g_3	55	64	47	63
g_4	81	108	115	71
g_5	76	55	42	49

項目「地震」推定結果

図 5.5 は PISA 2003 における項目「地震」の項目特性図である。図 5.5(i) から図 5.5(iv) を比較，観察すると，項目「地震」は上位群を識別することに適しているものの，下位群についてはブックレット間で項目特性が大きく異なり得ることが分かる。

4 つの項目特性図を観察すると，共通して中位群以上の識別は安定して行える項目であることが示唆される。

一方で，低特性者の識別に関してはブックレット間で解釈結果が異なり得る表現となっている。図 5.5(i)，図 5.5(iii) からは，低特性者を敏感に識別可能であることが示唆される。しかし，図 5.5(ii) では下位群の識別については不適当であることが，また，図 5.5(iv) からは，下位群の識別力が比較的高く，また，上位群についても識別可能であるものの，その識別力は比較的低いことが示唆される。

ブックレット間で解釈が大きく異なり得るために，下位群の識別には慎重を期さねばならないだろう。

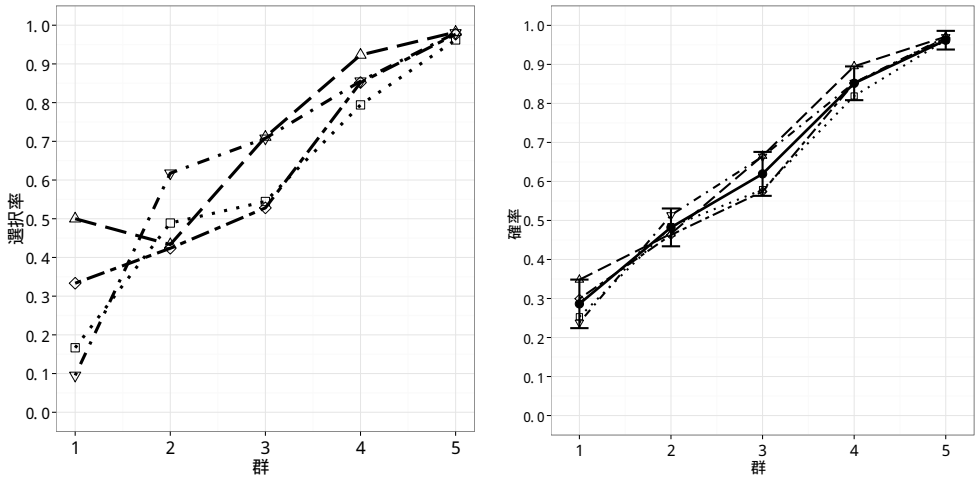
表 5.11 より，収束判定指標 $\hat{R}$ はすべての推定値について 1.1 以下であったため，連鎖は収束したものと判断した。

図 5.6 に正答選択率から直接作成した項目特性図と，推定項目特性図を示した。図 5.6(i) から，上位群の特性曲線は比較的安定しているものの，下位群については個々の特性曲線もやや不安定であり，なおかつブックレット間での正答率の揺れも大きいことが分かる。この特性表現からは，下位群を識別するための出題項目としては不適当であるという判断を下さざるを得ないであろう。

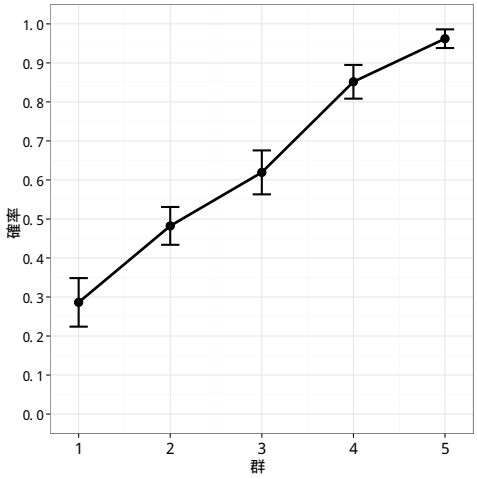
一方，図 5.6(ii) を観察すると，上位群の正答率の安定の程度は同様に観察されるものの，下位群についても特性曲線が右肩上がりとなる様子が見られ，下位群の識別にも使用可能な項目であることが示唆された。ただし，下位群についてはブックレット間の正答率の揺れが比較的大きい傾向にあるため，下位群に対するワーディングの方法が適切であったのかを検討する必要があるだろう。ただし，低特性群は高特性群と比較すると，当て推量といった受験者の行動の

表 5.11 項目「地震」 p_{bg} 推定結果

p_{bg}	平均	SD	$\hat{R}$
p_{11}	0.253	0.063	1.000
p_{21}	0.348	0.067	1.000
p_{31}	0.299	0.065	1.000
p_{41}	0.238	0.067	1.001
p_{12}	0.484	0.048	1.000
p_{22}	0.466	0.045	1.001
p_{32}	0.463	0.046	1.000
p_{42}	0.514	0.053	1.000
p_{13}	0.579	0.040	1.000
p_{23}	0.666	0.042	1.000
p_{33}	0.574	0.043	1.000
p_{43}	0.664	0.041	1.000
p_{14}	0.819	0.031	1.000
p_{24}	0.895	0.025	1.000
p_{34}	0.852	0.027	1.000
p_{44}	0.853	0.031	1.000
p_{15}	0.962	0.017	1.000
p_{25}	0.971	0.017	1.000
p_{35}	0.968	0.019	1.000
p_{45}	0.970	0.018	1.000



(i) 項目特性図　　(ii) 推定項目特性図 (BL 付き)



(iii) 推定項目特性図

図 5.6 項目特性図と推定項目特性図（PISA 2003 項目「地震」）

表 5.12 項目「地震」$\hat{p}_g$ および $\hat{p}_{bg}$

群	$\hat{p}_g$	$\hat{\sigma}^2_{p_g}$	$\hat{\sigma}_{p_g}$	p_{1g}	p_{2g}	p_{3g}	p_{4g}
1	0.286	0.004	0.062	0.253	0.348	0.299	0.238
2	0.482	0.002	0.048	0.484	0.466	0.463	0.514
3	0.619	0.003	0.056	0.579	0.666	0.574	0.664
4	0.852	0.002	0.043	0.819	0.895	0.852	0.853
5	0.962	0.001	0.024	0.962	0.971	0.968	0.970

結果，特性曲線が不安定となっていることも考えられるため，ある程度の正答率の揺れは許容され得るだろう。

本例のように，提案手法を用いて作成した観念的な項目特性図を項目分析の対象とすることで，通常の項目特性図を観察するだけでは得られない知見を得ることが可能となる。

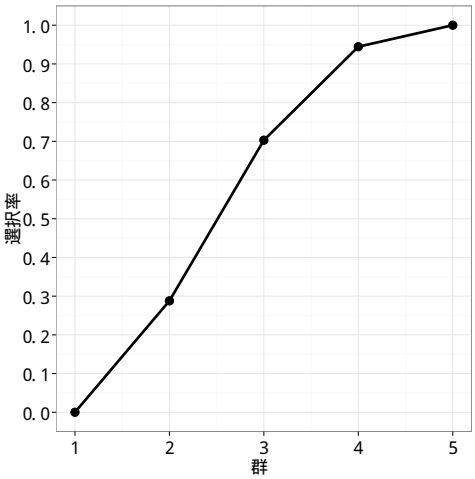
5.3.4 適用例 PISA 2003 項目「スケート」

項目「スケート」データ

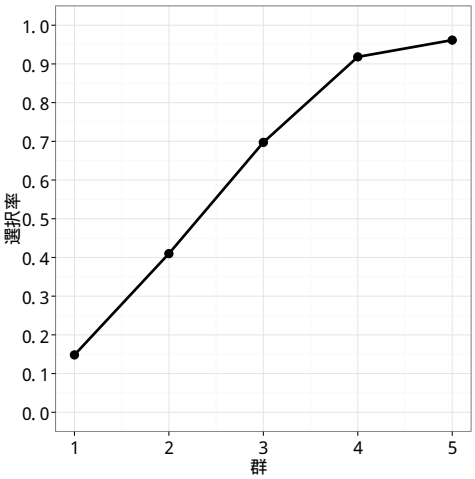
表 5.13 および表 5.14 に項目「スケート」に関して，推定に用いるデータを示した。

表 5.13 項目「スケート」 ブックレット・群別受験者数

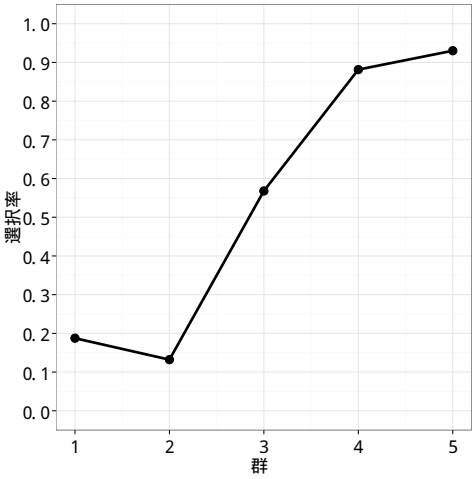
	BL1	BL2	BL3	BL4
g_1	17	27	16	21
g_2	66	61	53	40
g_3	101	109	81	102
g_4	126	110	135	95
g_5	36	52	43	86
N_b	346	359	328	344



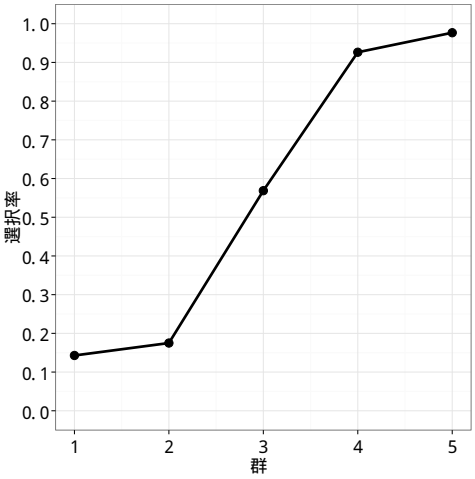
(i) BL 1



(ii) BL 2



(iii) BL 3



(iv) BL 4

図 5.7 項目特性図（PISA 2003 項目「スケート」）

表 5.14 項目「スケート」 ブックレット・群別正答者数

	BL1	BL2	BL3	BL4
g_1	0	4	3	3
g_2	19	25	7	7
g_3	71	76	46	58
g_4	119	101	119	88
g_5	36	50	40	84

項目「スケート」推定結果

図 5.7 は PISA 2003 項目「スケート」のブックレットごとの項目特性図である。4 つの項目特性図を観察すると，共通して中位群の識別力が高く，また上位群の識別力はやや低いことが分かる。また，図 5.7(i) と図 5.7(ii) では下位群についても，識別に適した項目であることが示唆される一方で，図 5.7(iii) と図 5.7(iv) からは，下位群については不適当，もしくは識別力が非常に低い項目であることが示唆される。

本項目に関しても，ブックレット間で項目特性の解釈に違いが認められる。よって提案手法を適用することで，正答分析を統一的な観点から実行可能となることが期待される。

項目「スケート」に対して，提案手法の適用結果を示す。

図 5.8(i) はこれまでの適用例と同様，観測データの反応比率を用いて直接的に作成した項目特性図である。また，図 5.8(ii) は表 5.16 による事後統計量を用いて作成した推定項目特性図である。図 5.8(ii) より，項目「スケート」は上位群に関しては識別力が低い傾向にあり，中位群に関しては識別力が高いことが示唆された。また，これらの推定選択率の安定の程度は比較的高いことも分かる。

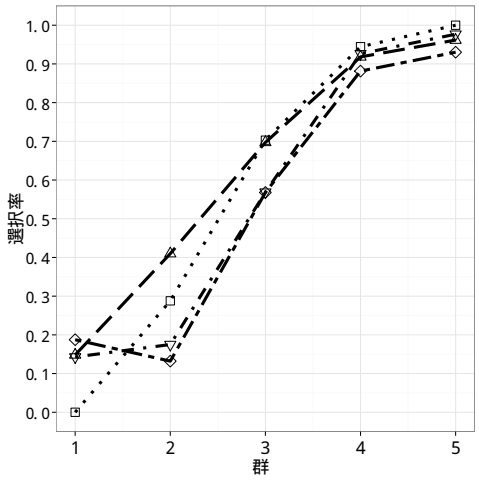
下位群に関しては低い識別力があることが示唆されるものの，不安定度合いが相対的に大きい傾向にあることが示唆された。

5.4 研究 III 結論

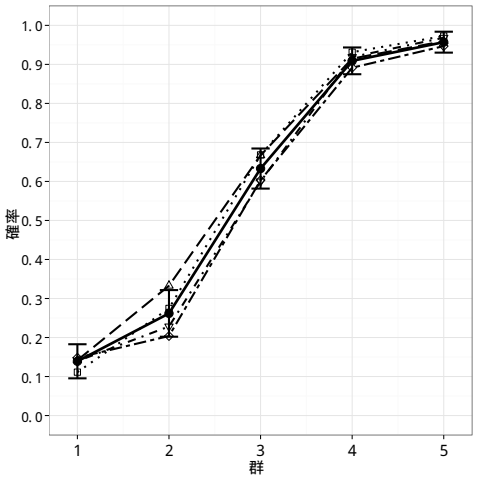
本研究 III では，複数ブックレット間に含まれる共通項目について，ブックレット間の項目特性図の表現の違いを加味した項目分析を可能とする方法について提案した。階層ベイズモデリングを用いた分析によって，観念的な項目特性図を作成し，項目分析の対象とすることで，ブックレットごとの項目特性図の比較分析のみからは得ることのできない知見を得られることが示唆された。本提

表 5.15 項目「スケート」 p_{bg} 推定結果

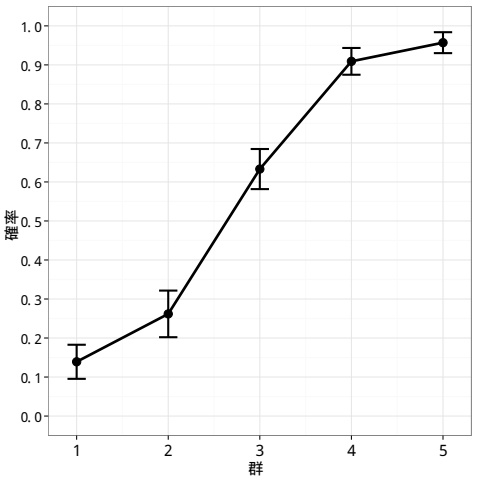
p_{bg}	平均	SD	$\hat{R}$
p_{11}	0.111	0.049	1.001
p_{21}	0.142	0.046	1.000
p_{31}	0.148	0.052	1.000
p_{41}	0.140	0.048	1.000
p_{12}	0.275	0.044	1.000
p_{22}	0.333	0.051	1.001
p_{32}	0.204	0.046	1.000
p_{42}	0.228	0.049	1.000
p_{13}	0.668	0.038	1.000
p_{23}	0.667	0.037	1.000
p_{33}	0.603	0.041	1.000
p_{43}	0.600	0.039	1.001
p_{14}	0.931	0.020	1.000
p_{24}	0.915	0.022	1.000
p_{34}	0.891	0.023	1.000
p_{44}	0.919	0.024	1.000
p_{15}	0.973	0.020	1.000
p_{25}	0.960	0.021	1.000
p_{35}	0.947	0.024	1.001
p_{45}	0.968	0.016	1.000



(i) 項目特性図



(ii) 推定項目特性図 (BL 付き)



(iii) 推定項目特性図

図 5.8 項目特性図と推定項目特性図（PISA 2003 項目「スケート」）

表 5.16 項目「スケート」$\hat{p}_g$ および $\hat{p}_{bg}$

群	$\hat{p}_g$	$\hat{\sigma}^2_{p_g}$	$\hat{\sigma}_{p_g}$	p_{1g}	p_{2g}	p_{3g}	p_{4g}
1	0.139	0.002	0.044	0.111	0.142	0.148	0.140
2	0.262	0.004	0.060	0.275	0.333	0.204	0.228
3	0.633	0.003	0.051	0.668	0.667	0.603	0.600
4	0.909	0.001	0.034	0.931	0.915	0.891	0.919
5	0.957	0.001	0.027	0.973	0.960	0.947	0.968

案手法は，共通項目の項目特性図を用いた項目分析のために有効な方法となるであろう。

第6章　総合考察

研究 I，研究 II，研究 III を通じて，多肢選択形式項目の項目特性図作成方法に関して，3 つの観点から精緻化を行う方法を論じてきた。

研究 I では，項目特性図の群数選択に関する統計的な基準を提案した。統計データを視覚的に把握することは，非常に重要な方略である。項目分析を視覚的に行うことが可能となる項目特性図は，統計分析を専門的に行う者のみならず，必ずしも統計分析に通暁しているわけではない，各科目の専門家，項目作成者にとっても，重要な道具となる。項目特性図から項目特性の評価結果について視覚的に得ることで，更なる項目内容の改善に役立てることが可能となる。

一方で，項目特性図による特性表現は必ずしも一意に定まるものではなく，受験者の群数次第で，表現は変化し得る。一般的には 5 群分割とされて作成されてきており，有効に活用することができている。しかしながら，何れの表現で正答分析（項目の測定性能の調査）を行うべきかの統計的な基準は広く知られておらず，分析者が判断に迷うような場合であっても，恣意的に決定せざるを得ず，何故そのような判断を下したのかは，あくまで経験的に説明することしかできなかった。

また，項目分析が活用される場面はテストの準備段階であることが多い。例えば実際のテスト運営では非常に多くの受験者が見込まれるような場合であっても，テスト作成段階では比較的，極少数の受験者に対するテスト実施結果を通じて，項目分析を行わなければならない。準備段階であっても，十分に多くの受験者を用意することが可能である場合には，経験的に群数を選択し，項目特性の解釈を行っても大きな問題とはならない。しかしながら準備段階と本番で受験者数に大きな差が認められる場合には，両者の項目特性図を同じ群数で作成することは特性表現の大きな食い違いを生じさせる可能性がある。

受験者数が少ない場合には，群数が多いとき，項目特性図の表現は不安定になるため，本試験に項目特性を適切に反映することができていない可能性が残る。この場合には群数を減らし，各群の所属人数を増やすことで安定性を増すことが期待できるが，この場合にもどのように群数を選択すべきかの基準は知られていない。

研究 I による方法は，項目特性図を用いた正答分析を行う際の群数選択の一助となることが期待される。特性表現を安定させつつ，全体データの傾向を表現するための群数を決定する際の傍証として用いることが可能である。

研究Ⅱでは誤答分析のための項目特性図精緻化の方法を提案した。正答分析により，基本的な測定性能が保障された項目については，誤答分析を行うことで，受験者の教科教育を行うための知見を得ることができる。

一方で，誤答選択肢の特性曲線は，互いに似通いやすいという性質があり，図による視覚的な把握性が阻害され得るという問題が存在する。また，誤答分析によって解釈された誤答傾向が標本誤差による影響であるのかどうかを検証する方法も広く知られていない。このため，分析結果に基づいて，教科教育方針を立てたとしても，基準に照らし合わせて有効であるか否かを議論することはできなかった。

このとき，特異な傾向が認められる誤答選択肢以外の特性曲線について平均選択率を計算することでまとめ，改めて項目特性図を作成することで，誤答選択肢に関して効果的に特性を示す図を作成することが可能となる。また，まとめ方について，研究Ⅱによる提案手法を用いて情報量規準を算出し，比較することで，特性曲線の併合に用いた仮説が妥当であるかどうかを検証するための傍証とすることが期待される。

研究Ⅲでは，ブックレット形式におけるテスト共通項目に関する，項目特性図作成方法について提案した。

ブックレット形式のテストでは，ブックレット間に共通する項目を含めることで，統計的に互いに比較可能な状態へと導くことができる。このとき，共通項目に関してはブックレットの数だけ，項目特性図が作成可能である。

例え同一項目であっても，標本誤差およびブックレット構成項目の違いから，項目特性はブックレットごとに変化し得る。このような場合に，何れの表現を項目特性として採用すべきか参照可能な基準は広く知られてこなかった。

研究Ⅲにおいて提案した手法は，互いに全く異なるとも，全く同一であるとも見なすことの難しいこれらのブックレット間項目特性図に関して，階層ベイズモデルを用いることで複数のブックレットの背後に共通して仮定される観念的な項目特性図を作成することを可能とする。また，同時に，当該項目特性図の安定性について分析することも可能となる。

観念的な項目特性図を解釈対象とすることで，ブックレット間項目特性図選択の恣意性を回避し，統一的な視点から項目分析を行えることが期待される。

上記の研究における提案手法は本稿を通じて個別に項目特性図に適用されてきた。しかしながらこれらの方法は状況に応じて組み合わせて適用することも可能であり，より柔軟に項目特性図を用いた項目分析を行う状況において役立つことが期待される。

付 録A　項目「輸出」内容

以下に，2003 年 PISA 調査で使用された,「輸出に関する問２」（以下「輸出」）の項目内容を示した（ブックレット 7，受験者 325 名，23 点満点）。

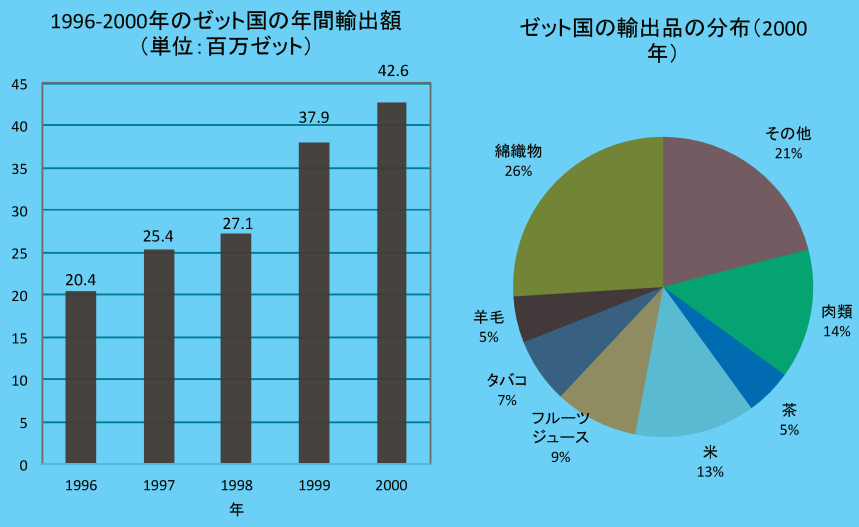


図 A.1 第 42 項目項目内容（国立教育政策研究所（2004, p.91）より引用）

表 A.1 「輸出に関する問２」

輸出に関する問２

2000 年にゼットランド国が輸出したフルーツジュースの金額はいくらでしたか。

A　1.8 百万ゼット

B　2.3 百万ゼット

C　2.4 百万ゼット

D　3.4 百万ゼット

E　3.8 百万ゼット

（正答）“E”

（国立教育政策研究所（2004, pp.92-93）より引用）

付 録B　研究 Ⅲ Stanコード

研究 Ⅲ において使用した Stan コードを以下に示す。

```
data{
    int<lower=0, upper=10> G; //G数
    int<lower=0> B; //BL数
    int<lower=0> Nbg[G, B]; //BL別G別人数
    int<lower=0> Succ[G,B];
}
parameters{
    real<lower=0, upper=1> p_bg[G, B];
    real<lower=0, upper=100> alpha_g[G];
    real<lower=0, upper=100> beta_g[G];
}
model{
    for(b in 1:B){ // BL
        for(g in 1:G){ //G数
        p_bg[g, b] ~ beta(alpha_g[g], beta_g[g]);
        Succ[g,b] ~ binomial(Nbg[g, b], p_bg[g, b]);
        }
    }
}
generated quantities{
    real<lower=0, upper=1> p_g[G];
    real<lower=0> sigb[G];
    // 平均と分散
    for(g in 1:G){
        p_g[g] <- alpha_g[g]/(alpha_g[g] + beta_g[g]);
        sigb[g] <- (alpha_g[g] * beta_g[g])/
            ((alpha_g[g] + beta_g[g])^2 *
                (alpha_g[g] + beta_g[g] + 1));
    }
}
```

引用文献

Akaike, H. (1973). Information theory and an extention of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *2nd International Symposium on Information Theory*., pp.267-281., Akadimiai Kiado.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, E., K. Tanabe, & G. Kitagawa. (Eds.), *Selected papers of Hirotugu Akaike*., pp.199-214., New York: Springer.

赤池弘次・甘利俊一・北川源四郎・樺島祥介・下平英寿 (著) 室田一雄・土谷隆 (編) (2007). 赤池情報量規準 AIC—モデリング・予測・知識発見— 共立出版

赤根敦・伊藤圭・林篤裕・椎名久美子・大澤公一・柳井晴夫・田栗正章 (2006). 識別指数による総合試験問題の項目分析. 大学入試センター研究紀要, **35**, pp.19-47. (Akane, A., Ito, K., Hayashi A., Shiina, K., Osawa, K., Yanai, H., & Taguri, M. (2006). Item analysis of non-curriculum-based ability test based upon discrimination index. *Research Bulletin of NCUUE*, **35**, 19-47.)

甘利俊一 (2007). 赤池情報量規準 AIC—その思想と新展開, 赤池弘次・甘利俊一・北川源四郎・樺島祥介・下平英寿 (著) 室田一雄・土谷隆 (編) 赤池情報量規準 AIC—モデリング・予測・知識発見— 共立出版, pp.52-78.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer-Verlag New York. (C.M. ビショップ (著) 元田浩・栗田多喜夫・樋口知之・松本裕治・村田昇 (監訳) (2012). パターン認識と機械学習 (下)—ベイズ理論により統計的予測, 丸善出版)

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In Brennan, R. L.(Ed.) *Educational measurement, Fourth edition*. Praeger, pp.307-353.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference, Second Edition*. Springer-Verlag New York.

Croon, M. (2002). Ordering the classes. In J. A. Hagenaars & A. L. McCutcheon (Eds.) *Applied latent class analysis*., pp.137-162., Cambridge University Press.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, **195**(2), pp.216-222.

Educational Testing Servece (1963). Multiple-choice questions: A close look.

古谷知之 (2008). ベイズ統計データ分析—R & WinBUGS— 朝倉書店

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S.

Richardson & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*., pp.131-143., London: Chapman and Hall.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, pp.457-511.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernerdo, J. O. Berger, A. P. Dawid & A. F. Smith (Eds.), *Bayesian Statistics 4* (pp.169-193). Oxford, NY: Oxford University Press.

Heidelberger, P. & Welch, P. D. (1983). Simulation run length control in the presence of an initial transit. *Operations Research*, **31**, pp.1109-1144.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. J. Comp. Graph. Stat. 5:299-314. Available via `http://www.R-project.org`.

池田央 (1973). 心理学研究法 8 テスト II 東京大学出版会

池田央 (1992). テストの科学―試験にかかわるすべての人に― 日本文化科学社

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, **30**(1), pp.17-24.

菊地賢一 (1999). 項目反応理論を用いた設問解答率分析図の評価, 大学入試センター研究紀要, **29**, pp.1-8. (Kikuchi K. (1999). Evaluation of quintile item response chart using item response theory. *Research Bulletin of NCUUE*, **29**, 1-8.)

北川源四郎 (2007). 情報量規準と統計的モデリング, 赤池弘次・甘利俊一・北川源四郎・樺島祥介・下平英寿 (著) 室田一雄・土谷隆 (編) 赤池情報量規準 AIC―モデリング・予測・知識発見― 共立出版, pp.79-109.

小西貞則・北川源四郎 (2004). 情報量基準 朝倉書店

国立教育政策研究所 (編) (2004). 生きるための知識と技能 (2) —OECD 生徒の学習到達度調査 (PISA) ・2003 年調査国際結果報告書. ぎょうせい

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**(1), pp.79-86.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

麻柄啓一 (編) (2006). 学習者の誤った知識をどう修正するか—ル・バー修正ストラテジーの研究. 東北大学出版会

松本隆 (2004). ハミルトニアン・モンテカルロ法によるベイズ的学習と予測 In 松本隆, 石黒真木夫, 乾敏郎, 田邉國士 (著) 階層ベイズモデルとその周辺 ―時系列・画像・認知への応用― 岩波書店 pp.145-157.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and Ability. In *Educational measurement* (3rd ed.). New York:

American Council on Education and Macmillan. Linn, R. L. (Ed.). (池田央・藤田恵璽・柳井晴夫・繁桝算男 (訳編) (1992). 教育測定学原著第 3 版（下） (pp.3-54) C.S.L. 学習評価研究所)
森村英典・高橋幸雄 (1979). マルコフ解析 日科技連
Neal, R. M. (1996). *Bayesian learning for neural networks*. Lecture Notes in Statistics 118, NY: Springer–Verlag.
Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In *Handbook of Markov chain Monte Carlo*. Brooks, S., Gelman, A., Galin L. Jones G. L. & Meng, X-L., (Eds.). Chapman & Hall/CRC, pp.113-162.
日経 TEST http://ntest.nikkei.jp/ アクセス日時:2014.04.19 05:51
OECD Programme for International Student Assessment (PISA) http://www.oecd.org/pisa/ アクセス日時:2013.05.11 21:39
大津起夫 (2006). 大規模テストデータの簡易集計と視覚化のためのツール. 日本行動計量学会大会発表論文抄録集 **34**, 26-27.
Raftery, A. E. & Lewis, S. (1992a). How many iterations in the Gibbs sampler? In J. M. Bernerdo, J. O. Berger, A. P. Dawid & A. F. Smith (Eds.), *Bayesian statistics 4* (pp.763-773). Oxford, NY: Oxford University Press.
Raftery, A. E. & Lewis, S. (1992b). One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science*, **7**, pp.493-497.
坂元慶行・石黒真木夫・北川源四郎 (1983). 情報量統計学 共立出版
佐藤隆博 (1998). コンピュータ処理による S–P 表分析の活用法—学習指導の個別対応のために—. 明治図書.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.
繁桝算男 (1985). ベイズ統計入門 東京大学出版会
清水留三郎 (1983). 共通一次学力試験における解答の分析について, 大学入試フォーラム, **1**, pp.36-37.
Shojima, K. (2008). Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*, 07-03.
荘島宏二郎 (2010). ニューラルテスト理論—学力を段階評価するための潜在ランク理論— 植野真臣・荘島宏二郎（共著） 学習評価の新潮流 (第 4 章, pp.83-111) 朝倉書店
Stan Development Team (2014a). RStan: the R interface to Stan, Version 2.4. http://mc-stan.org/rstan.html.
Stan Development Team (2014b). Stan: A C++ Library for Probability and Sampling, Version 2.4. http://mc-stan.org.
龍岡菊美・林篤裕 (2001). 個人の潜在的知識ステートを診断する統計的方法

論. 計測と制御, **40**, pp.561-567
照井伸彦 (2010). R によるベイズ統計分析 朝倉書店
Hogg, R. V., McKean, J., & Craig, A. T. (2012). *Introduction to mathematical statistics, Seventh edition*, Pearson. (豊田秀樹 (監訳) (2006). 数理統計学ハンドブック （原著第 6 版） 朝倉書店)
豊田秀樹 (2012). 項目反応理論［入門編］【第 2 版】 朝倉書店
渡部洋 (1999). ベイズ統計学入門 福村出版
山本義郎・飯塚誠也・藤野友和 (2013). 統計データの視覚化 共立出版
吉村宰 (2009). 大学入試センターにおけるテストデータベースによる項目分析　植野真臣・永岡慶三 (共編). e テスティング (第 8 章) 培風館, pp.167-190.