項目応答理論（IRT）

- 研究では、学習能力や性格特性など、直接観測できない潜在的な特性を扱うことがあります。IRT モデルは、観測できる個々の項目がどのように潜在特性に関連し、またこれらの項目グループ全体がどのように潜在特性に関係しているのかを調べます。
- IRT モデルは、一次元の確認的因子分析（CAF）のバイナリおよびカテゴリカルな結果への拡張モデルや、一般化線形混合効果モデルの特殊なケースとして考えることができます。

この例題集でできること

- バイナリデータに 1PL モデルと 2 PL モデルをフィットさせる。
- 項目応答曲線（ICC）とテスト特性曲線（TCC）を作成する。
- カテゴリカルデータに段階反応モデル（GRM）をフィットさせる。
- 項目情報関数（IIF）、境界特性曲線（BCC）、項目情報関数グラフ（CCC）を作成する。

バイナリ IRT モデル

- この例では、バイナリデータを使って IRT 分析を行い、irt コマンドの事後評価機能を紹介します。
- De Boeck and Wilson (2004)による、数学の能力に関するデータを使用します。生徒はテストを受け、正解した場合 1、不正解だった場合 0 と記録されます。
- 下記のコマンドで例題用のサンプルデータ「`masc1`」を入手し、1 行目から 5 行目までの内容を確認します。

```
. use https://www.stata-press.com/data/r16/masc1
. list in 1/5
```

下記の表が表示されます。

	q1	q2	q3	q4	q5	q6	q7	q8	q9
1.	1	1	1	0	0	0	0	1	0
2.	0	0	1	0	0	0	0	1	1
3.	0	0	0	1	0	0	1	0	0
4.	0	0	1	0	0	0	0	0	1
5.	0	1	1	0	0	0	0	1	0

- この表は、生徒 1 が問い q1、q2、q3、q8 に正解し、生徒 2 が q3、q8、q9 に正解したことを表しています。生徒 3 以降も同様に、「1」が正解したことを表しています。
- この検定の目的は、生徒の数学の能力を評価し、たとえば優・良・可のようなグループに生徒を分けることです。
- 各生徒の総合点は分かっていますが、総合点は試験の構造に依存するという問題点があります。もし試験が簡単な問題で構成されていたら、ほとんどの生徒は優の評価になるでしょう。逆に、難しい問題ばかりであれば、多くの生徒が可の評価になります。
- モデルがデータにフィットする際に、IRT には、測定誤差を除いてパラメーター推定が不変であるという利点があります。能力の推定値は試験に依存せず、項目パラメーターはグループに依存しなくなります。
- 下記のコマンドで、1 パラメータロジスティックモデル（1PL モデル）をバイナリデータ q1 から q9 にフィットさせます。

```
. irt 1pl q1-q9
```

```
Fitting fixed-effects model:

Iteration 0:   log likelihood = -4275.6606
Iteration 1:   log likelihood = -4269.7861
Iteration 2:   log likelihood = -4269.7825
Iteration 3:   log likelihood = -4269.7825

Fitting full model:

Iteration 0:   log likelihood = -4153.3609
Iteration 1:   log likelihood =  -4142.374
Iteration 2:   log likelihood = -4142.3516
Iteration 3:   log likelihood = -4142.3516

One-parameter logistic model                    Number of obs     =        800
Log likelihood = -4142.3516
```

		Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
	Discrim	.852123	.0458445	18.59	0.000	.7622695	.9419765
q1	Diff	-.7071339	.1034574	-6.84	0.000	-.9099066	-.5043612
q2	Diff	-.1222008	.0963349	-1.27	0.205	-.3110138	.0666122
q3	Diff	-1.817693	.1399523	-12.99	0.000	-2.091994	-1.543391
q4	Diff	.3209596	.0976599	3.29	0.001	.1295498	.5123695
q5	Diff	1.652719	.1329494	12.43	0.000	1.392144	1.913295
q6	Diff	.6930617	.1031842	6.72	0.000	.4908243	.8952991
q7	Diff	1.325001	.1205805	10.99	0.000	1.088668	1.561335
q8	Diff	-2.413443	.1691832	-14.27	0.000	-2.745036	-2.08185
q9	Diff	-1.193206	.1162054	-10.27	0.000	-1.420965	-.965448

- 出力された表をみると、最初の行に困難度の識別パラメーター（Descrim）があります。1 PL モデルでは、このパラメーターは全ての項目で共有されます。0.85 という推定値

は、この項目が完全な識別能力があるわけではないことを示しています。つまり、特定の困難度の推定値付近では、異なる能力を持つ 2 人が項目に対して似たような応答をする可能性があります。

- 残りの行では、各項目に対して困難度のパラメーターの推定値（Diff）が表示されています。項目の困難度のスペクトラムが広範囲に渡っていることがわかります。最低値は q8 の $\hat{b}_8$=-2.41 で、最高値は q5 の $\hat{b}_5$=1.65 となっています。

- 結果を指定した順番で並び替えるために estat report コマンドを使います。この例では、困難度が低い項目から高い項目の順番に並べます。

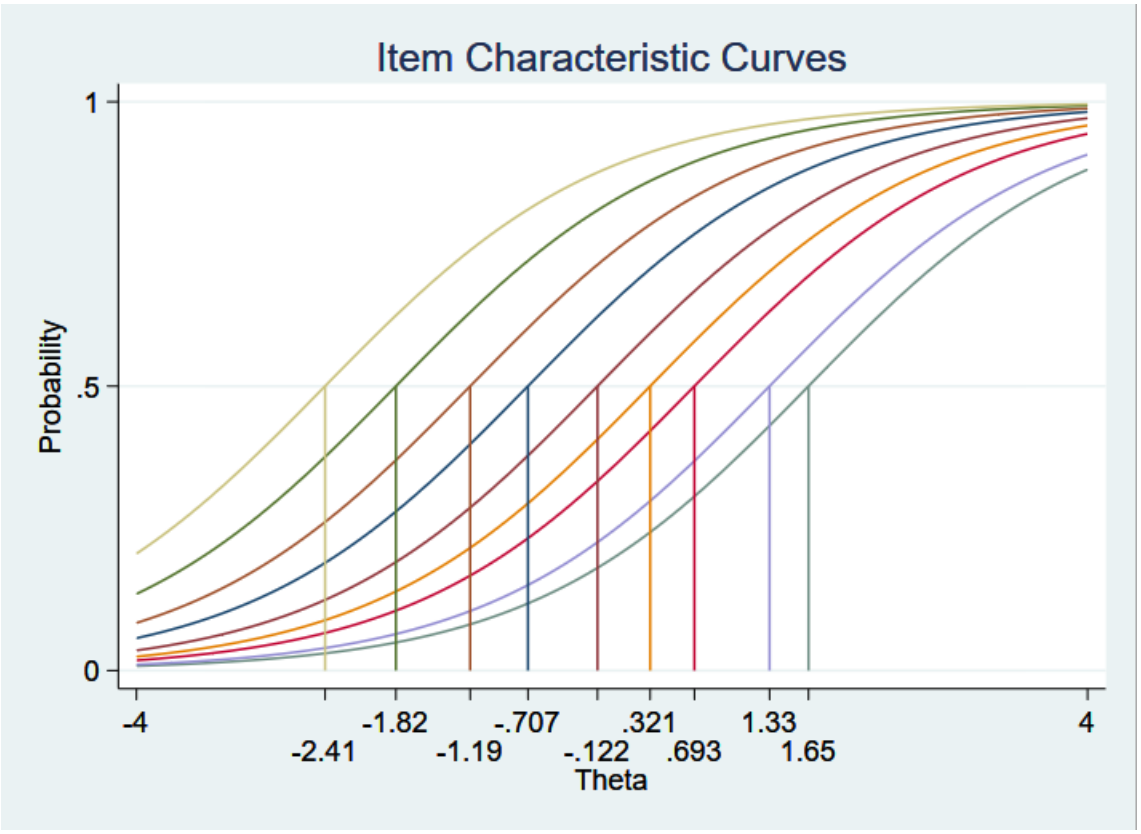
```
. estat report, sort(b) byparm
```

```
One-parameter logistic model                    Number of obs     =        800
Log likelihood = -4142.3516
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Discrim	.852123	.0458445	18.59	0.000	.7622695	.9419765
Diff						
q8	-2.413443	.1691832	-14.27	0.000	-2.745036	-2.08185
q3	-1.817693	.1399523	-12.99	0.000	-2.091994	-1.543391
q9	-1.193206	.1162054	-10.27	0.000	-1.420965	-.965448
q1	-.7071339	.1034574	-6.84	0.000	-.9099066	-.5043612
q2	-.1222008	.0963349	-1.27	0.205	-.3110138	.0666122
q4	.3209596	.0976599	3.29	0.001	.1295498	.5123695
q6	.6930617	.1031842	6.72	0.000	.4908243	.8952991
q7	1.325001	.1205805	10.99	0.000	1.088668	1.561335
q5	1.652719	.1329494	12.43	0.000	1.392144	1.913295

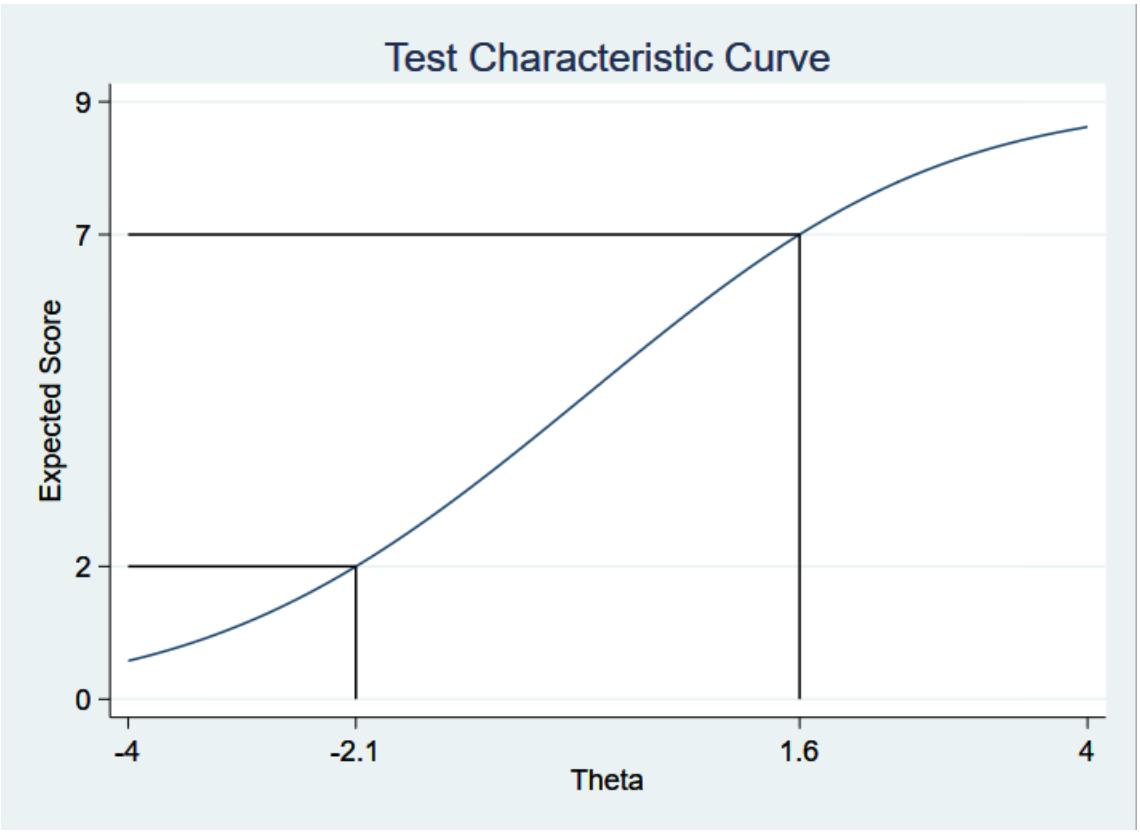
- 困難度のスペクトラムを表すために、項目応答曲線（ICC）を描きます。

```
. irtgraph icc, blocation legend(off) xlabel(,alt)
```



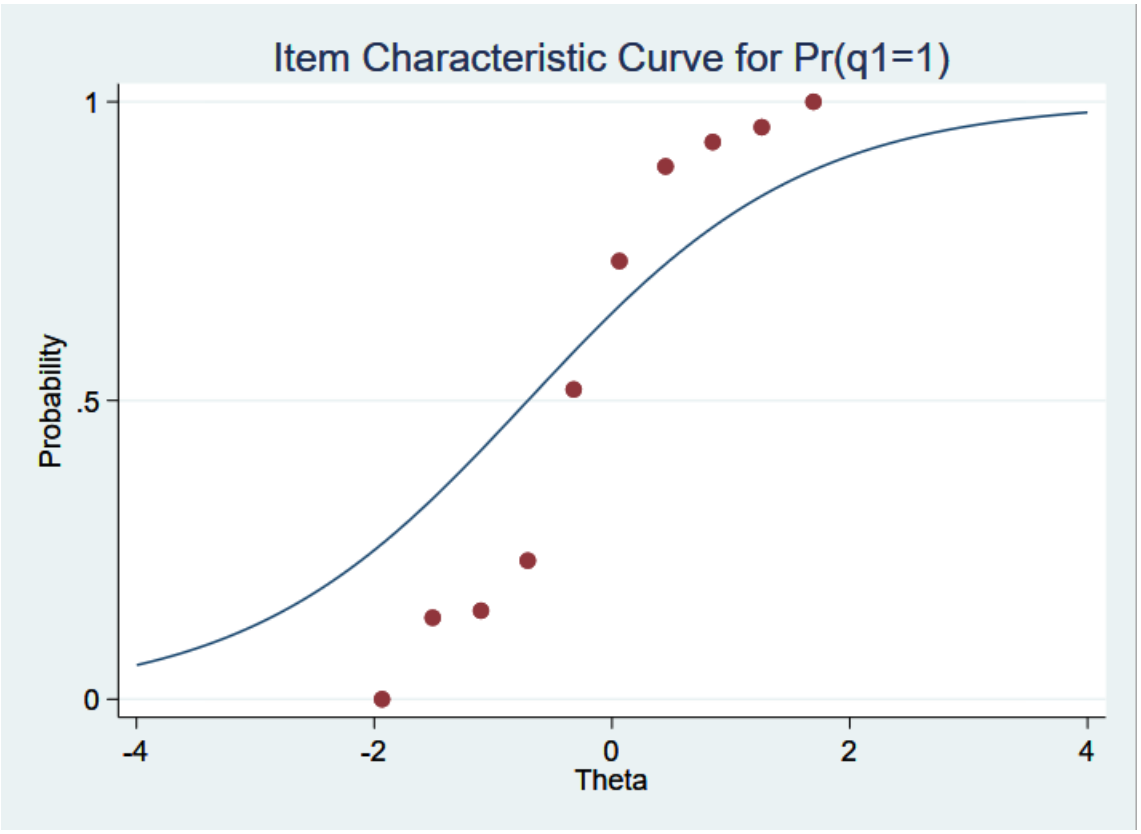
- 潜在能力 θ に対する、各項目の正答確率を表しています。1 PL モデルでは、各項目の50%確率は、困難度パラメーターの推定値に対応します。
- 確率の合計を計算すると、テスト全体で期待される点数がわかります。この潜在特性に対する期待される点数をプロットした図は、テスト特性曲線（TCC）と呼ばれます。
- TCC を作成するには、`irtgraph tcc` コマンドを使います。`scorelines(2 7)`オプションは、期待される点数 2 と 7 に対応するプロットに線を引きます。

```
. irtgraph tcc, scorelines(2 7)
```



- 推定される TCC によると、期待される点数が 2 の場合の潜在特性は-2.1、 7 の場合は 1.6 です。
- IRT の不変性プロパティは、モデルがデータに適合する場合のみ保持されます。フィットを確認する非公式な方法のひとつは、ICC に経験的比率 empirical proportions を重ね合わせることです。予測された ICC が経験的比率に沿っている場合、モデルがフィットしていると考えられます。
- 経験的比率を計算するには、潜在特性を予測し、潜在特性によって項目を collapse します。そして、`intgraph icc` コマンドに `addplot( )`オプションを使用して、ICC に重ねます。

```
. predict Theta, latent
. collapse q*, by(Theta)
. irtgraph icc q1, addplot(scatter q1 Theta)
```



- モデルの中の全項目について、経験的確率が ICC にフィットしていないことが分かります。この項目については、2PL モデルを適用した方が良いかもしれません。
- 2 PL モデルを適用する前に、ここまでの推定を onep として保存しておきます。

```
. estimates store onep
```

- 2 PL モデルを適用するには、下記のコマンドを入力します。

```
. use https://www.stata-press.com/data/r16/masc1, clear
. irt 2pl q1-q9
```

```
Fitting fixed-effects model:

Iteration 0:   log likelihood = -4275.6606
Iteration 1:   log likelihood = -4269.7861
Iteration 2:   log likelihood = -4269.7825
Iteration 3:   log likelihood = -4269.7825

Fitting full model:

Iteration 0:   log likelihood = -4146.9386
Iteration 1:   log likelihood = -4119.3568
Iteration 2:   log likelihood = -4118.4716
Iteration 3:   log likelihood = -4118.4697
Iteration 4:   log likelihood = -4118.4697

Two-parameter logistic model                    Number of obs     =        800
Log likelihood = -4118.4697
```

		Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
q1							
	Discrim	1.615292	.2436467	6.63	0.000	1.137754	2.092831
	Diff	-.4745635	.074638	-6.36	0.000	-.6208513	-.3282757
q2							
	Discrim	.6576171	.1161756	5.66	0.000	.4299171	.885317
	Diff	-.1513023	.1202807	-1.26	0.208	-.3870481	.0844435
q3							
	Discrim	.9245051	.1569806	5.89	0.000	.6168289	1.232181
	Diff	-1.70918	.242266	-7.05	0.000	-2.184012	-1.234347
q4							
	Discrim	.8186403	.1284832	6.37	0.000	.5668179	1.070463
	Diff	.3296791	.1076105	3.06	0.002	.1187663	.5405919
q5							
	Discrim	.8956621	.1535128	5.83	0.000	.5947825	1.196542
	Diff	1.591164	.2325918	6.84	0.000	1.135293	2.047036
q6							
	Discrim	.9828441	.147888	6.65	0.000	.6929889	1.272699
	Diff	.622954	.1114902	5.59	0.000	.4044373	.8414708
q7							
	Discrim	.3556064	.1113146	3.19	0.001	.1374337	.5737791
	Diff	2.840278	.8717471	3.26	0.001	1.131685	4.548871
q8							
	Discrim	1.399926	.233963	5.98	0.000	.9413668	1.858485
	Diff	-1.714416	.1925531	-8.90	0.000	-2.091814	-1.337019
q9							
	Discrim	.6378452	.1223972	5.21	0.000	.3979512	.8777392
	Diff	-1.508254	.2787386	-5.41	0.000	-2.054571	-.9619361

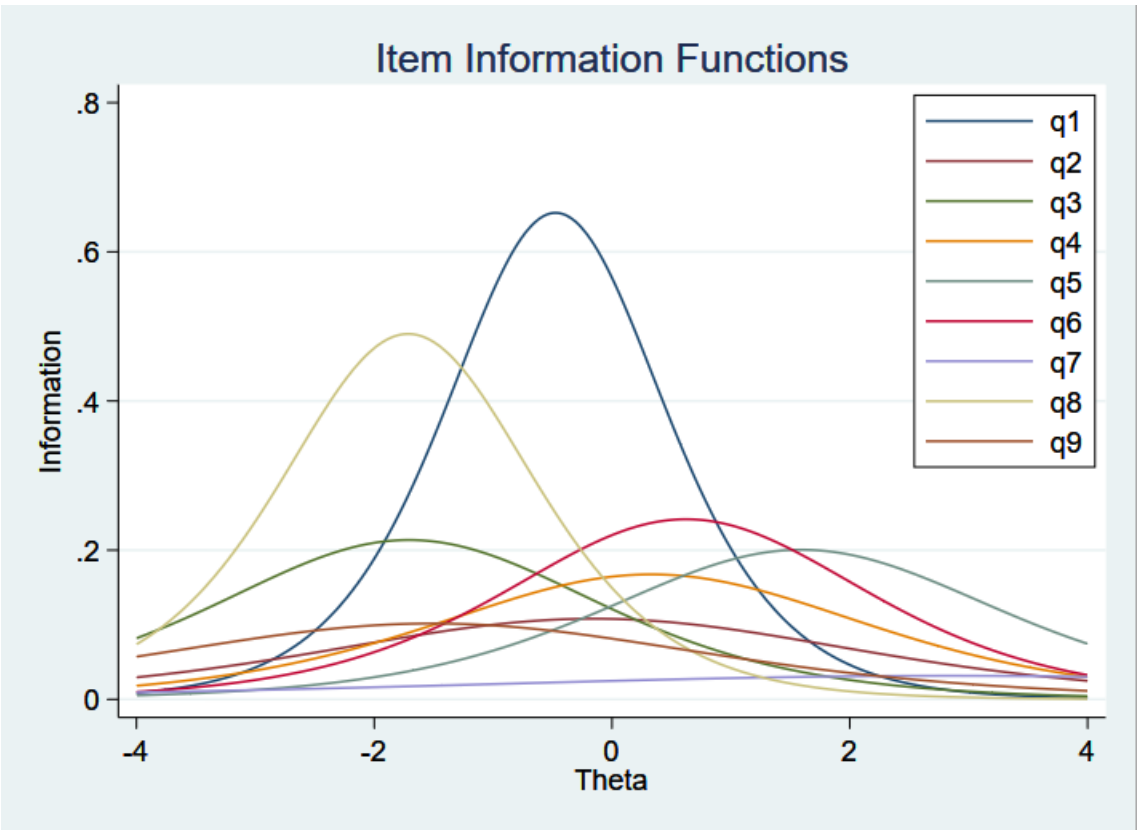
- ２PL モデルでは、各項目ごとに識別パラメーターがあります。１PL モデルでは、全ての項目で共通の識別パラメーターの値が 0.85 と推定されました。
- 出力結果の表によると **q1** の識別パラメーターは 1.62 と推定されており、経験的確率のグラフの傾きが大きいことによりフィットしているといえます。
- １PL モデルは２PL モデルにネストされているので、尤度比検定を実行することでどのモデルがより適しているかを調べることができます。

```
. lrtest onep .
```

```
Likelihood-ratio test                                 LR chi2(8)  =     47.76
(Assumption: onep nested in .)                        Prob > chi2 =    0.0000
```

- 有意水準がほぼゼロなので、識別パラメーターが各項目ごとに分かれていることを示しています。
- ２PL モデルで、潜在特性を推定するための項目の情報量をプロットすることができます。潜在特性に対する項目情報のグラフは、項目情報関数（IIF）と呼ばれます。
- モデルで全項目の IIF を出力するには、**irtgraph iif** コマンドを使います。

```
. irtgraph iif, legend(pos(1) col(1) ring(0))
```

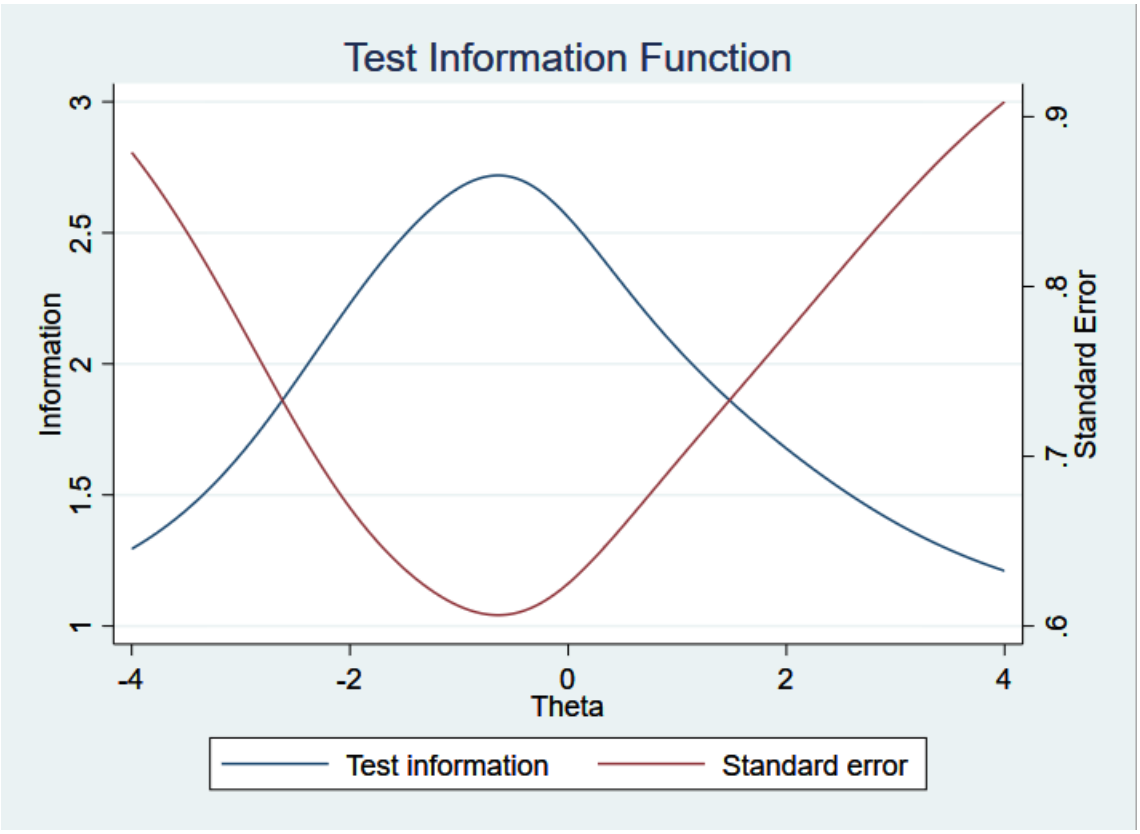


- ２PL モデルの場合、IIF には単峰性と対称性があります。IIF の高さ（困難度パラメーターの項目の情報の量）は、各項目の推定された識別力に比例します。項目 q1 と q8 は、

最も大きな識別力があり、急な傾きを持っています。

- 尺度の信頼性を調べるテスト情報関数（TIF）を得るために、IIF の和を求めます。

```
. irtgraph tif, se
```



- このテストは、θ=-0.5 の位置にいる人に最大の情報量を提供することが分かります。θ が-0.5 から外れると、TIF の標準誤差が大きくなり、提供される情報量が少なくなります。
- バイナリモデルには、1 PL モデルと 2 PL モデル以外にも 3 PL モデルがあります。3 PL モデルは、推測の可能性に対応することによって、2 PL モデルに追加されます。

例 2　カテゴリカルな IRT モデル

- カテゴリカルな IRT モデルには、順序付きの応答モデルと順序なしの応答モデルが含まれます。ここでは、順序付きの段階反応モデル（GRM）を扱います。GRM はカテゴリカルな結果に対応するための、2 PL モデルの拡張モデルです。
- モデルを描くために、Zheng and Rabe-Hesketh (2007)のデータを使用します。このデータには ta1 から ta5 までの 5 項目のアンケートの質問があり、慈善団体に対する信念と信頼を測っています。
- 反応は、強く同意する(0)、同意する(1)、同意しない(2)、全く同意しない(3)の 4 段階

です。高い点数であるほど、不信感が強いことを表しています。5 つの質問のデータを見てみましょう。

```
. use https://www.stata-press.com/data/r16/charity
. list in 1/5, nolabel
```

	ta1	ta2	ta3	ta4	ta5
1.	.	2	1	1	.
2.	0	0	0	0	0
3.	1	1	2	0	2
4.	1	2	2	0	1
5.	.	1	1	1	1

- 最初の行を見ると、この回答者は ta1 と ta5 には無回答、ta2 には 2、ta3 と ta4 には 1 と回答しています。irt コマンドは、観測値のない項目を尤度計算から除外しますが、他の欠測していない項目は残します。モデルから欠測値のある人のデータ全てを除外するには、listwise オプションを使用します。
- 下記のコマンドで GRM をフィットさせます。

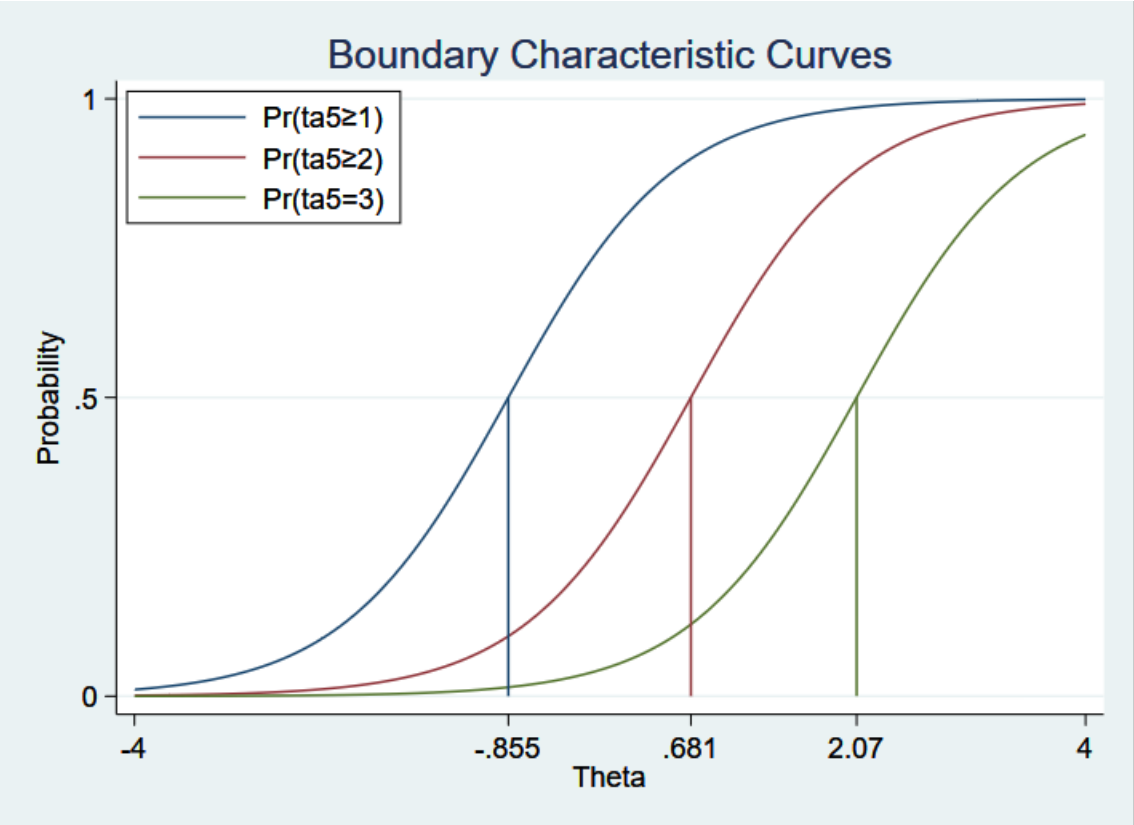
```
. irt grm ta1-ta5
```

```
Fitting fixed-effects model:

Iteration 0:   log likelihood = -5559.6414
Iteration 1:   log likelihood = -5473.9434
Iteration 2:   log likelihood = -5467.4082
Iteration 3:   log likelihood = -5467.3926
Iteration 4:   log likelihood = -5467.3926

Fitting full model:

Iteration 0:   log likelihood = -5271.0634
Iteration 1:   log likelihood = -5162.5917
Iteration 2:   log likelihood = -5159.2947
Iteration 3:   log likelihood = -5159.2791
Iteration 4:   log likelihood = -5159.2791

Graded response model                           Number of obs     =        945
Log likelihood = -5159.2791
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ta1						
Discrim	.907542	.0955772	9.50	0.000	.7202142	1.09487
Diff						
>=1	-1.540098	.1639425			-1.861419	-1.218776
>=2	1.296135	.1427535			1.016343	1.575927
=3	3.305059	.3248468			2.668371	3.941747
ta2						
Discrim	.9434675	.0967483	9.75	0.000	.7538444	1.133091
Diff						
>=1	-1.661331	.167878			-1.990366	-1.332296
>=2	.0068314	.082222			-.1543208	.1679836
=3	2.531091	.2412513			2.058247	3.003935
ta3						
Discrim	1.734201	.1554383	11.16	0.000	1.429548	2.038855
Diff						
>=1	-1.080079	.0835119			-1.243759	-.9163983
>=2	1.016567	.0796635			.8604297	1.172705
=3	2.232606	.1497814			1.93904	2.526172
ta4						
Discrim	1.93344	.1857629	10.41	0.000	1.569351	2.297528
Diff						
>=1	-.3445057	.0578468			-.4578833	-.2311282
>=2	1.466254	.0983823			1.273428	1.65908
=3	2.418954	.162392			2.100672	2.737237
ta5						
Discrim	1.42753	.1263962	11.29	0.000	1.179798	1.675262
Diff						
>=1	-.8552358	.0833158			-1.018532	-.6919399
>=2	.6805315	.07469			.5341418	.8269211
=3	2.074243	.1538858			1.772632	2.375853

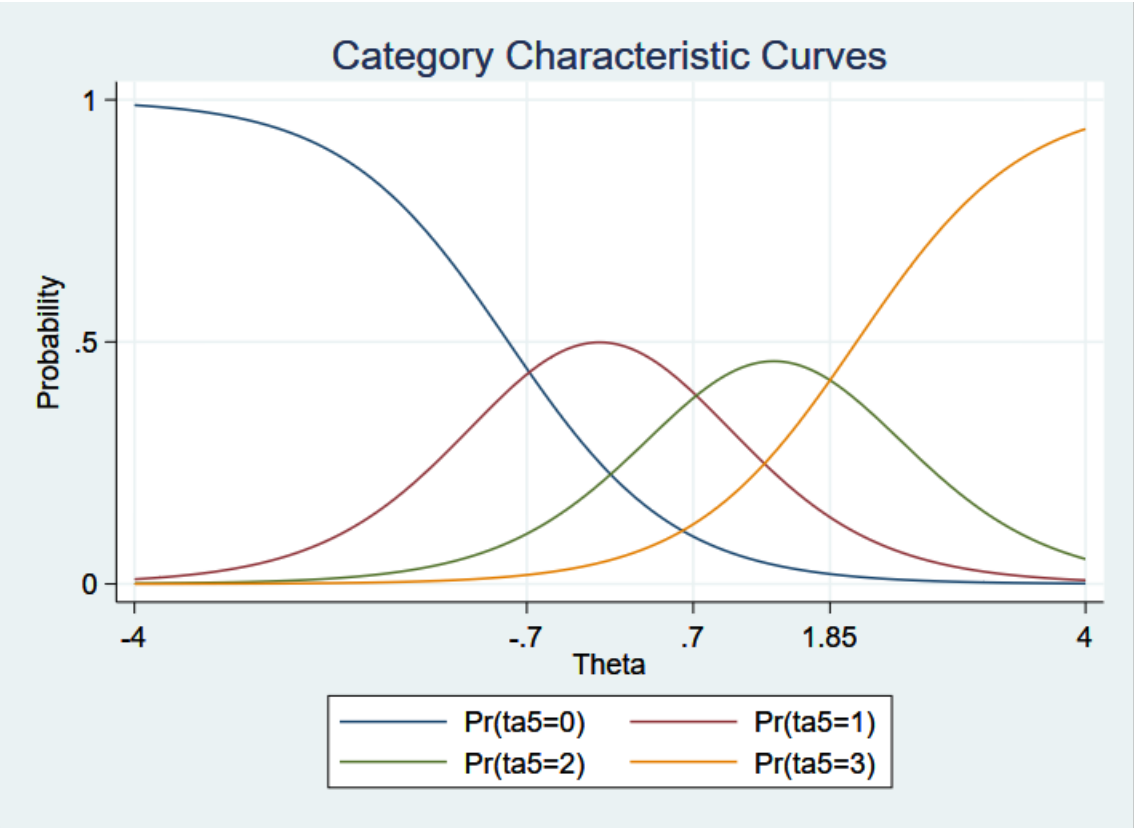
- GRM は累積確率で算出されるため、推定されたカテゴリの困難度は、特定の困難度と等しい能力を持つ人物が、指定されたカテゴリ以上の困難度に応答する確率が 50%であることを示しています。
- 例えば、ta5 において、θ=-0.86 の人が 0 と回答する確率は 50%であり、1 以上と回答する確率も同じく 50%です。θ=0.68 の人は 0 または 1 と回答する確率は 50%で、2 または 3 と回答する確率も 50%です。 θ=2.07 の人は、2 以下と回答する確率は 50%で、3 と回答する確率も 50%です。
- irtgraph icc コマンドを使って、これらの確率をグラフにします。ここでは、ta5 について推定されるカテゴリの困難度と確率をプロットします。
- GRM では、各カテゴリの中間確率は推定されたカテゴリの困難度に相当します。

```
. irtgraph icc ta5, blocation legend(pos(11) col(1) ring(0))
```



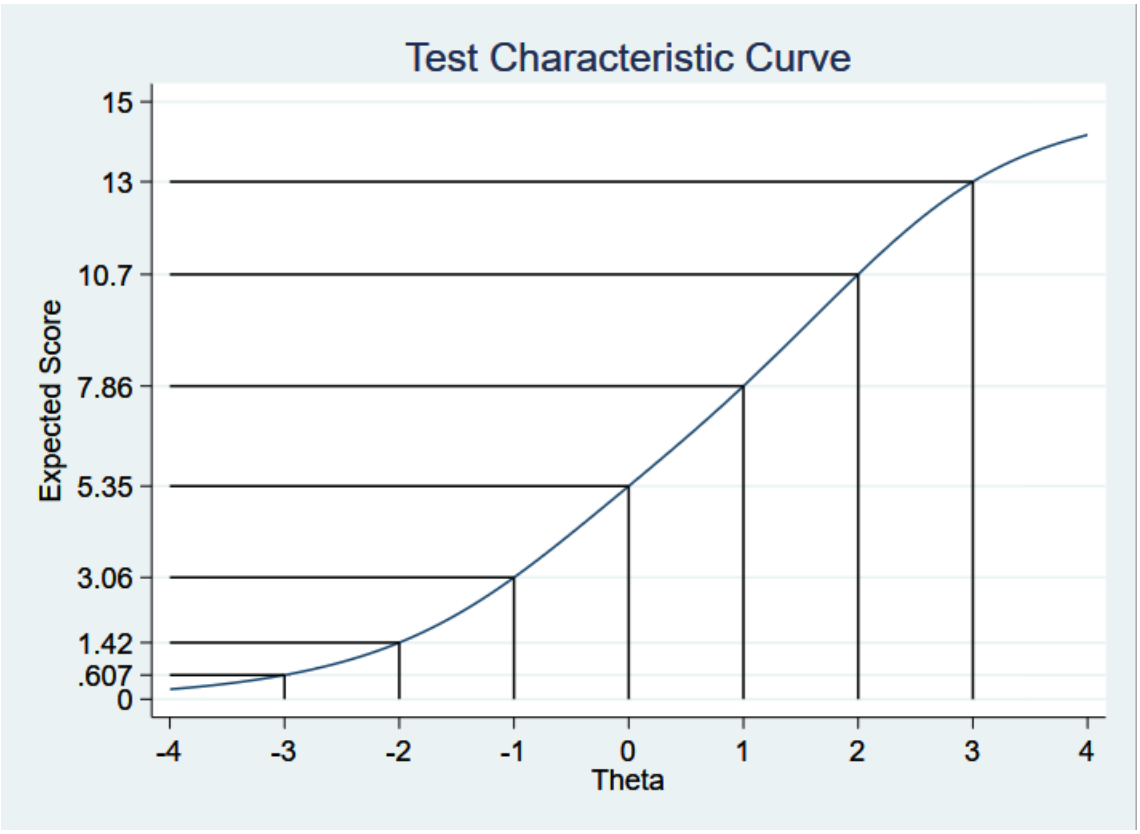
- カテゴリカルな項目で特性曲線を描いた場合は、境界特性曲線（BCC）と呼びます。
- カテゴリ k を選択する回答者の確率をプロットすることもできます。カテゴリカルな項目の場合、これは項目情報関数グラフ（CCC）と呼ばれます。

```
. irtgraph icc ta5, xlabel(-4 -.7 .7 1.85 4, grid)
```



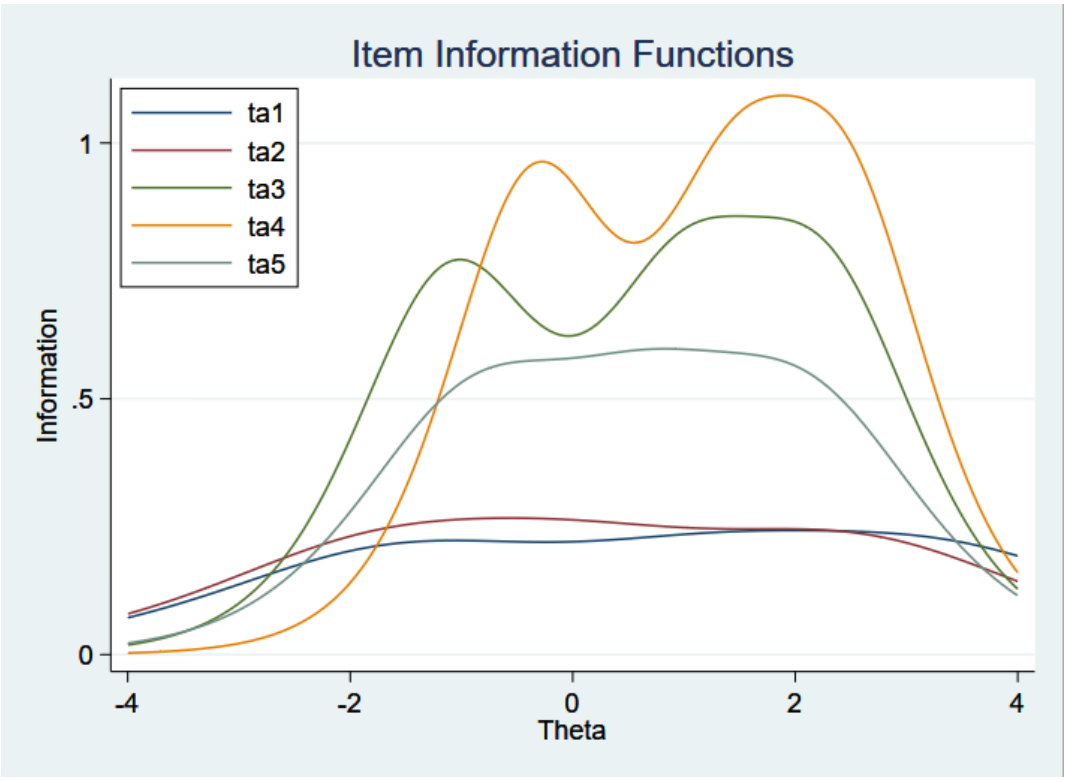
- カテゴリが交差する点は、あるカテゴリから隣のカテゴリに遷移することを示しています。
- つまり、不信感のレベルが低い回答者（約 θ =-0.7 以下の回答者）は `ta5` について最初のカテゴリ（0：強く同意する）をほとんど選択する傾向にあります。中間の θ =-0.7 から θ =0.7 の人は、`ta5` について 2 番目のカテゴリ（1：同意する）を選択する傾向にあります。
- 最初の例のように、全体のテスト特性関数をプロットできます。

```
. irtgraph tcc, thetalines(-3/3)
```



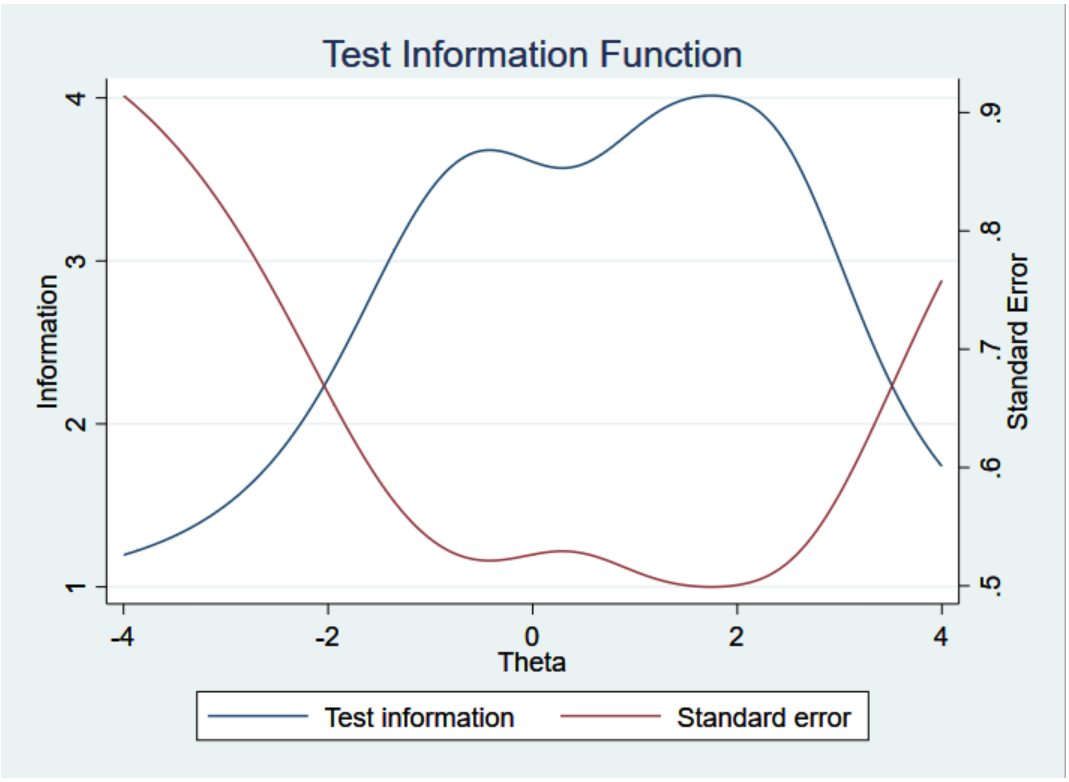
- 項目は 5 つあり、各スコアは最小値 0 から最大値 3 までなので、全体のスコアは 0 から 15 までの範囲になります。
- θ の値の違いによるスコアの違いを、`irtgraph iif` コマンドでプロットします。
- θ =-3 以下の場合、期待されるスコアは 1 以下です。これは、θ =-3 以下に該当する回答者が、全ての項目で 0 を選択する可能性が最も高いことを示しています。
- カテゴリカルな項目の場合、項目情報関数は単峰性や対称性を示しません。何故なら、各カテゴリは独自の情報を持っていて、それぞれが異なる範囲で最高値を取る可能性があるからです。下記のグラフで確認しましょう。

```
. irtgraph iif, legend(pos(11) col(1) ring(0))
```



- テスト情報関数は各 IIF の和なので、このプロットにも山と谷があります。

```
. irtgraph tif, se
```



- 上記の例では、GRM を実行しました。`irt` コマンドは、カテゴリカルな反応に対して次のようなモデルをサポートしています。
  - `irt nrm`：標準応答モデル（NRM）
  - `irt pcm`：partial credit モデル（PCM）
  - `irt rsm`：rating scale モデル（RSM）
- バイナリ IRT またはカテゴリカル IRT モデルに加えて、`irt` コマンドでさらに別のモデルを適用することもできます。項目をサブセットし、全体に対して単一のキャリブレーションを行うためのモデル `irt hybrid` があります。