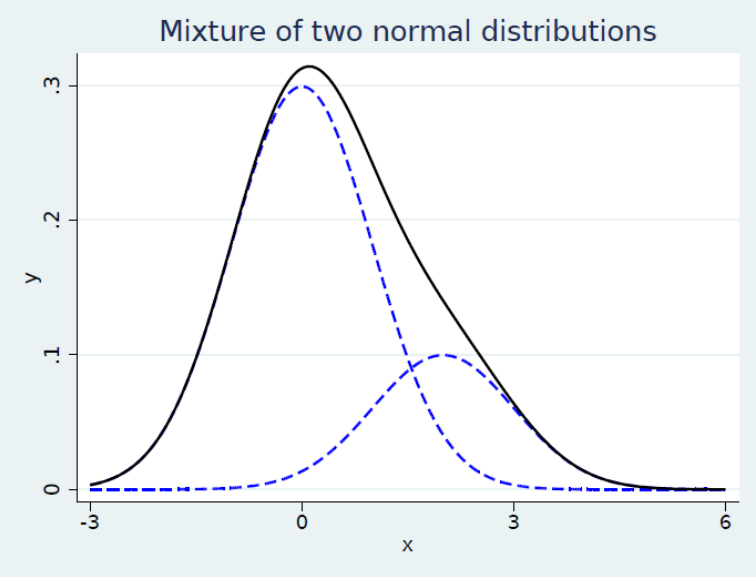
有限混合モデル(FMM)

はじめに

- 有限混合モデル(FMM: Finite Mixture Models)の主な概念は、観測されたデータがいくつかの非観測な部分母集団から構成されているとするものです。下図に例を示します。
    - 黒の実線：集団全体として観測された分布
    - 青の破線：上記の分布を構成する 2 つの部分母集団の分布密度（非観測）



観測された分布は正規分布に近いですが、負のデータよりも正のデータの方が多く、やや非対称な形状となっています。この非対称性は、分布が 2 つの正規分布の混合であるために発生しています。右側の分布によって、全体としての分布が右方向に歪められています。FMM を使用して、これら 2 つの分布の平均と分散、およびこれらの分布が全体に占める割合を推定できます。

- FMM は柔軟性があるため、観測値の分類、クラスタリングの調整、および非観測な分散不均一性のモデル化などの目的のために、さまざまな分野で広く使用されています。
- 分散が等しい正規分布の混合を使用して、任意の連続分布を近似できます。これにより、FMM は多峰性データ、歪みのあるデータ、非対称データをモデル化するための有用なツールとされています。

✧ 本文では次の用語を使用します。

「クラス」「グループ」「タイプ」「成分」	非観測な部分母集団の意
「クラス確率」「成分確率」	混合中のある成分に帰属する確率

※「クラス確率」については、「混合重み」「混合比率」という用語を使用している文献もあります。

✧ 本文中のコマンドをコピーし、Stata のコマンドウィンドウに貼り付けて実行できます。全ての操作のコマンドは、do ファイル fmm.do にまとめられています。

✧ Stata のコマンドウィンドウで「`help` *コマンド名*」を実行すると、各コマンドのヘルプを確認できます。適宜ご利用ください。

有限混合モデル

FMM は複数の確率密度関数を組み合わせた確率モデルです。FMM の場合、観測された応答yはg種類の異なるクラス$f_1, f_2, \ldots, f_g$からそれぞれ$\pi_1, \pi_2, \ldots, \pi_g$の比率でもたらされるものと仮定されます。最も単純な形式では、g個の成分からなる混合モデルの密度関数は次の通りです。

$$f(y) = \sum_{i=1}^{g} \pi_i f_i(y|x'\beta_i)$$

ここで、π_iはi番目のクラスに対する確率$(0 \le \pi_i \le 1, \sum \pi_i = 1)$を、$f_i(\cdot)$は$i$番目のクラスのモデルにおける応答変数の条件付き確率密度関数を意味します。

一方、`fmm` コマンドは潜在クラスに対する確率を多項ロジスティック分布によってモデル化します。i番目の潜在クラスに対する確率は次のように表現されます。

$$\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{g} \exp(\gamma_j)}$$

ここで、γ_iはi番目の潜在クラスに対する線形予測値です。デフォルトでは、1 番目の潜在クラスはベースレベルとみなされるため、$\gamma_1 = 0,\ \exp(\gamma_1) = 1$となります。

正規分布の混合　―　FMM の例題

メキシコの 1872 年のイダルゴ切手は、その時代の切手に特徴的なさまざまな種類の紙に印刷されました。コレクターにとっては、切手の紙が厚いほど価値があります。FMM を使用して、厚紙を使用した切手である確率を予測できます。

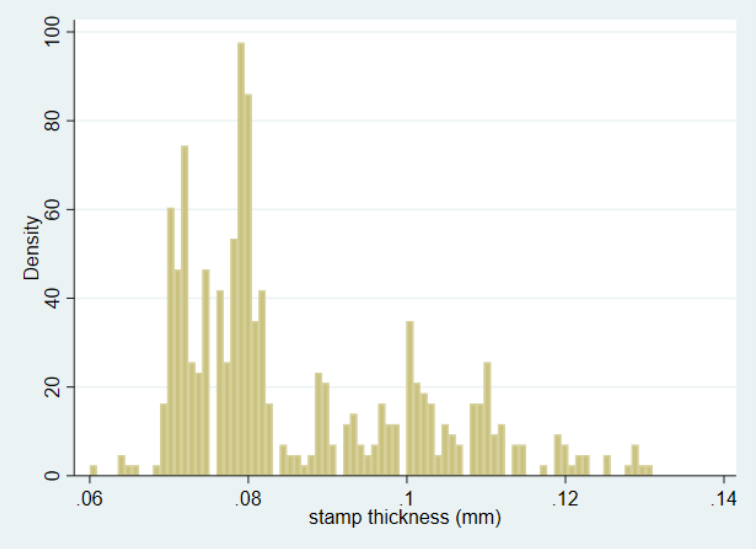
サンプルデータ stamp.dta は、485 枚の切手の厚さのデータです。厚さは 1000 分の 1 ミリメートル単位まで記録されています。

サンプルデータをインポートして、切手の厚さのヒストグラムを作成します。Stata のコマンドウィンドウで次のコマンドを実行します。

```
use https://www.stata-press.com/data/r17/stamp, clear

histogram thickness, bin(80)
```

結果は次のようになります。ヒストグラムは、データには少なくとも二峰性があることを示唆しています。



ここでは Izenman and Sommer (1988)にしたがって、それぞれが独自の平均と分散を持つ 3 つの正規分布の混合をデータにフィットさせます。また、各分布が全体の密度に寄与する割合を推定します。3 つの分布は、切手が印刷された 3 つの異なるタイプの紙（厚紙、中紙、薄紙）を表すと考えることができます。モデルは次の通りです。

$$f(y) = \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2) + \pi_3 N(\mu_3, \sigma_3^2)$$

各クラスに属する確率は、多項ロジスティック回帰を使用して推定されます。

$$\pi_1 = \frac{1}{1 + \exp(\gamma_2) + \exp(\gamma_3)}$$

$$\pi_2 = \frac{\exp(\gamma_2)}{1 + \exp(\gamma_2) + \exp(\gamma_3)}$$

$$\pi_3 = \frac{\exp(\gamma_3)}{1 + \exp(\gamma_2) + \exp(\gamma_3)}$$

ここでγ_iは多項ロジットモデルの切片です。デフォルトでは、最初のクラスがベースとして扱われるため、$\gamma_1 = 0$です。

このモデルをフィットする場合の Stata のコマンドは次の通りです。

```
fmm 3: regress thickness
```

3 つの成分が混在しているため、`fmm 3:` と入力します。続けて `regress thickness` と入力して、各成分の線形回帰モデルにフィットするように `fmm` に指示します。共変量がない場合、`regress` は、各成分の Gaussian（正規）密度の平均と分散を推定します。
（⇒`help fmm regress` を参照）

上記コマンドの実行結果は次の通りです。

```
. fmm 3: regress thickness

Fitting class model:

Iteration 0:   (class) log likelihood =  -532.8249
Iteration 1:   (class) log likelihood =  -532.8249

Fitting outcome model:

Iteration 0:   (outcome) log likelihood =  1949.1228
Iteration 1:   (outcome) log likelihood =  1949.1228

Refining starting values:

Iteration 0:   (EM) log likelihood =  1396.8814
Iteration 1:   (EM) log likelihood =  1404.8995
Iteration 2:   (EM) log likelihood =  1412.4626
Iteration 3:   (EM) log likelihood =  1416.9678
Iteration 4:   (EM) log likelihood =  1419.0044
Iteration 5:   (EM) log likelihood =  1419.0582
Iteration 6:   (EM) log likelihood =  1417.9719
Iteration 7:   (EM) log likelihood =  1416.4213
Iteration 8:   (EM) log likelihood =  1414.8176
Iteration 9:   (EM) log likelihood =  1413.3462
Iteration 10:  (EM) log likelihood =  1412.0695
Iteration 11:  (EM) log likelihood =   1410.992
Iteration 12:  (EM) log likelihood =  1410.0961
Iteration 13:  (EM) log likelihood =  1409.3574
Iteration 14:  (EM) log likelihood =  1408.7518
Iteration 15:  (EM) log likelihood =  1408.2578
Iteration 16:  (EM) log likelihood =  1407.8564
Iteration 17:  (EM) log likelihood =  1407.5315
Iteration 18:  (EM) log likelihood =  1407.2694
Iteration 19:  (EM) log likelihood =  1407.0695
Iteration 20:  (EM) log likelihood =  1406.9013
note: EM algorithm reached maximum iterations.

Fitting full model:

Iteration 0:   log likelihood =  1516.5252
Iteration 1:   log likelihood =  1517.1348  (not concave)
Iteration 2:   log likelihood =  1517.8203  (not concave)
Iteration 3:   log likelihood =   1518.153
Iteration 4:   log likelihood =  1518.6491
Iteration 5:   log likelihood =  1518.8474
Iteration 6:   log likelihood =  1518.8484
Iteration 7:   log likelihood =  1518.8484
```

Finite mixture model Number of obs = 485
Log likelihood = 1518.8484

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.Class	(base outcome)					
2.Class						
_cons	.6410696	.1625089	3.94	0.000	.3225581	.9595812
3.Class						
_cons	.8101538	.1493673	5.42	0.000	.5173992	1.102908

Class: 1
Response: thickness
Model: regress

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
thickness						
_cons	.0712183	.0002011	354.20	0.000	.0708242	.0716124
var(e.thick~s)	1.71e-06	4.49e-07			1.02e-06	2.86e-06

Class: 2
Response: thickness
Model: regress

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
thickness						
_cons	.0786016	.0002496	314.86	0.000	.0781123	.0790909
var(e.thick~s)	5.74e-06	9.98e-07			4.08e-06	8.07e-06

Class: 3
Response: thickness
Model: regress

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
thickness						
_cons	.0988789	.0012583	78.58	0.000	.0964127	.1013451
var(e.thick~s)	.0001967	.0000223			.0001575	.0002456

- 出力には、4 つの反復ログが表示されます。最初の 3 つは、初期値を取得するのにフィットしたモデル用です。混合モデルでは、適切な初期値を見つけるのが難しいことがよくあります。fmm は、初期値を指定・計算するためのさまざまなオプションを提供しています。
- 最初に出力されている表は、推定されたクラス確率を多項ロジスティックスケールで示しています。これらの推定値は次のように確率に変換できます。

$$\pi_1 = \frac{1}{1 + \exp(0.64) + \exp(0.81)} \approx 0.19$$

$$\pi_2 = \frac{\exp(0.64)}{1 + \exp(0.64) + \exp(0.81)} \approx 0.37$$

$$\pi_3 = \frac{\exp(0.81)}{1 + \exp(0.64) + \exp(0.81)} \approx 0.44$$

- estat lcprob コマンドを使用すると、上記の確率と、関連する標準誤差および信頼区間を計算できます（⇒help fmm estat lcprob を参照）。

```
. estat lcprob

Latent class marginal probabilities                     Number of obs = 485

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      [95% conf. interval]
-------------+----------------------------------------------------------------
       Class |
           1 |   .1942968   .0221242       .1545535    .2413428
           2 |   .3688746   .0286318       .3147305    .4265356
           3 |   .4368286    .027885        .383149     .49203
------------------------------------------------------------------------------
```

- fmm の出力の残り 3 つの表は、各正規分布の推定平均と分散を示しています。
- 結果から混合密度は、平均、分散、およびクラス確率の最尤推定値とともに、次の式で与えられます。

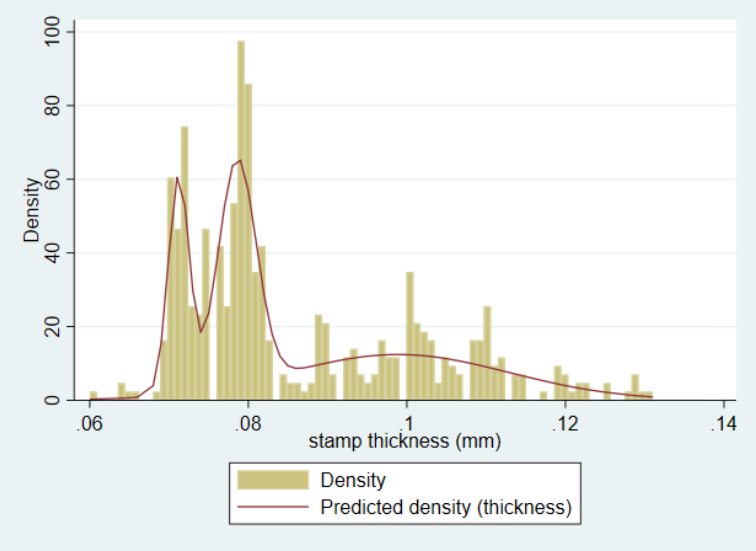
$$0.19 \times N(0.071, 0.0000017) + 0.37 \times N(0.079, 0.0000057) + 0.44 \times N(0.099, 0.0001967)$$

上記の式は切手の厚さの予測密度を示します。これを切手の厚さの経験的分布に対してプロットします。はじめに predict コマンドで予測値の新規変数 den を作成します（⇒help fmm postestimation を参照）。次に変数 den の折れ線と変数 thickness のヒスト

グラムを重ね描きします。

```
predict den, density marginal

histogram thickness, bin(80) addplot(line den thickness)
```



- 分散が小さい最初の 2 つの成分は経験的分布の左側をモデル化しています。
- 一方、分散がはるかに大きい 3 番目の成分は経験的分布の右側のロングテールをカバーしていることがわかります。

クラスメンバーシップの事後確率の予測を使用して、各切手の各クラスに属する確率を評価できます。はじめに **predict** コマンドで事後潜在クラス確率の新規変数 **pr1**、**pr2**、**pr3** を作成します（⇒**help fmm postestimation** を参照）。これらの変数はワイルドカードを使用して **pr*** と書くことができます。次に変数 **pr*** のフォーマットを固定長数値形式、全桁数 4 桁、少数桁数 3 桁に変更します（⇒**help format** を参照）。最後に変数 **thickness** と変数 **pr*** の 1 行目の内容を表示します。

```
predict pr*, classposteriorpr

format %4.3f pr*

list thickness pr* in 1, abbreviate(10)
```

1.

thickness	pr1	pr2	pr3
.06	0.000	0.000	1.000

結果から、データセットの 1 行目に記録されている切手の場合、紙タイプのクラス 3（厚紙）に属する確率は 1 です。

モデルには共変量がないため、事後確率は特定の厚さの切手で同じであり、次のようになります。

thickness	pr1	pr2	pr3
.06	0.000	0.000	1.000
.064	0.000	0.000	1.000
.065	0.001	0.000	0.999
.066	0.026	0.000	0.974
.068	0.723	0.001	0.276
.069	0.915	0.001	0.083
.07	0.960	0.002	0.037
.071	0.965	0.007	0.028
.072	0.937	0.026	0.037
.073	0.789	0.134	0.076
.074	0.335	0.525	0.140
.075	0.038	0.838	0.123
.076	0.002	0.910	0.088
.077	0.000	0.930	0.070
.078	0.000	0.936	0.064
.079	0.000	0.930	0.070
.08	0.000	0.912	0.088
.081	0.000	0.871	0.129
.082	0.000	0.788	0.212
.083	0.000	0.635	0.365
.084	0.000	0.406	0.594
.085	0.000	0.185	0.815
.086	0.000	0.060	0.940
.087	0.000	0.015	0.985
.088	0.000	0.003	0.997
.089	0.000	0.001	0.999
.09-.131	0.000	0.000	1.000

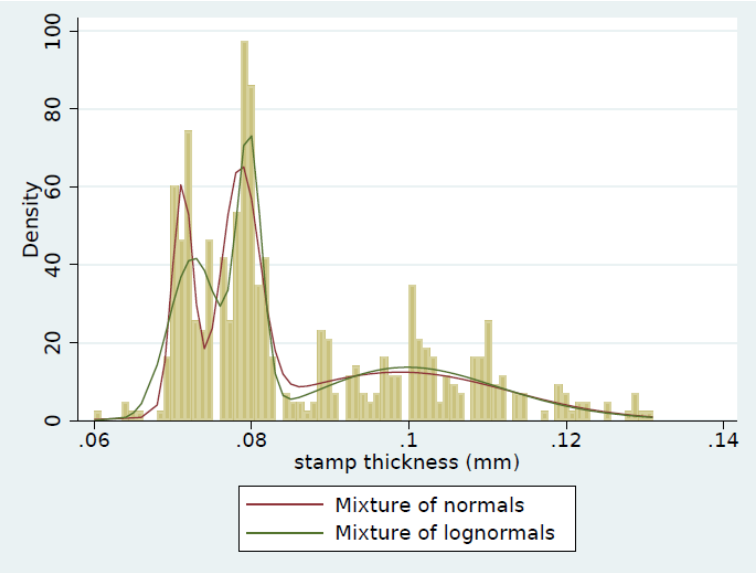
3 番目の混合成分の分散は比較的大きいため、1～4 行目の薄い切手は、最終的に紙タイプのクラス 3（厚紙）に誤って分類されています。切手の厚さを負にすることはできないため、対数正規分布など、正の実数直線のみをサポートする密度を使用すると、モデルの適合度を向上させることができます。

対数正規分布でフィットする場合のコマンドは次の通りです（⇒help fmm glm を参照）。

```
fmm 3: glm thickness, family(lognormal)
```

（結果省略）

正規分布の混合からの予測密度と、対数正規分布の混合からの予測密度を重ねてプロットすると次のようになります。



対数正規分布の混合により、1～4 行目の薄い切手は紙タイプのクラス 1（薄紙）に正しく分類されます。

thickness	pr1	pr2	pr3
.06	.889	0	.111
.064	.992	0	.008
.065	.994	0	.006
.066	.996	0	.004
.068	.997	0	.003
.069	.997	0	.003
.07	.996	0	.004