PROC KDE及びPROC DISCRIMによる分布の重なり具合（OVL）の推定

奥山 ことば
萬有製薬株式会社
臨床統計部

ブリッジング試験における問題

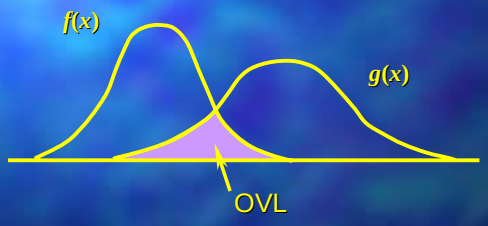
- 「類似性」の評価指標
  - ＯＶＬが分布の類似性の指標として提唱
  - ＰＫデータのブリッジング試験の評価指標(2002/1)
    - ノンパラメトリック密度推定を適用してOVLを算出
  - OVLの性能評価
- 両地域で症例数が大きく異なる
- 複数の評価指標（例　AUC、C_{max}）

発表の流れ

- OVLとは
- 1変量におけるOVL
  - パラメトリック、ノンパラメトリック
  - KDEプロシジャ、DISCRIMプロシジャ
  - 数値計算Ⅰ　;様々な分布、平均差と分散比
  - 数値計算Ⅱ　;サンプルサイズが異なる場合
- 2変量におけるOVL
  - パラメトリック、ノンパラメトリック
  - KDEプロシジャ、DISCRIMプロシジャ
  - 数値計算Ⅲ　;2変量正規分布におけるOVL
- まとめ

OVLとは？

- OVL (Overlapping Coefficient)
  分布の重なりの程度を表す指標

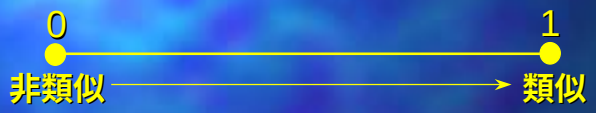


OVLの概念図

OVLとは？

- **OVLの定義式**

$$OVL = \int \min[\, f(x), g(x)]dx$$



– OVLは、0～1の数値をとる統計量

– 単調な変数変換に対し、OVLは不変

OVLとは？

- 分布の密度推定法の選択によって影響を受ける
- パラメトリックOVLは、研究が進んでいる
- ノンパラメトリックOVLは、始ったばかり
  - Stine and Heyse(2001)[12]
  - ノンパラメトリックな密度推定法は、研究が進んでいる

1変量におけるOVL

- パラメトリック法（理論式）
  - 等分散

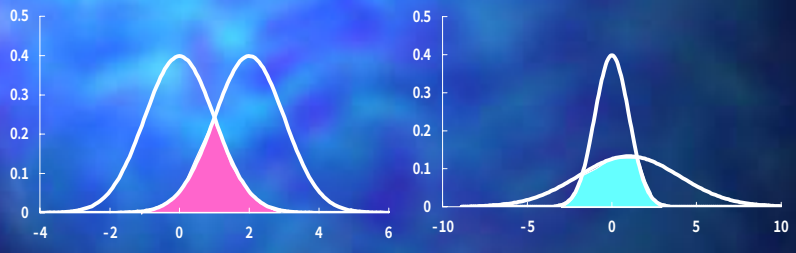
$$OVL = 2\Phi\left(-\left|\mu_f - \mu_g\right| / 2\sigma\right)$$

  - 不等分散

$$OVL = \Phi\left((a-\mu_f)/\sigma_f\right) + \Phi\left((b-\mu_g)/\sigma_g\right) - \Phi\left((a-\mu_g)/\sigma_g\right) - \Phi\left((b-\mu_f)/\sigma_f\right) + 1$$

$$(a,b) = (\sigma_g^2 - \sigma_f^2)^{-1}[(\mu_f\sigma_g^2 - \mu_g\sigma_f^2) \pm \sigma_f\sigma_g\{(\mu_f-\mu_g)^2 + 2(\sigma_g^2-\sigma_f^2)\log(\sigma_g/\sigma_f)\}^{1/2}]$$

パラメトリックOVL

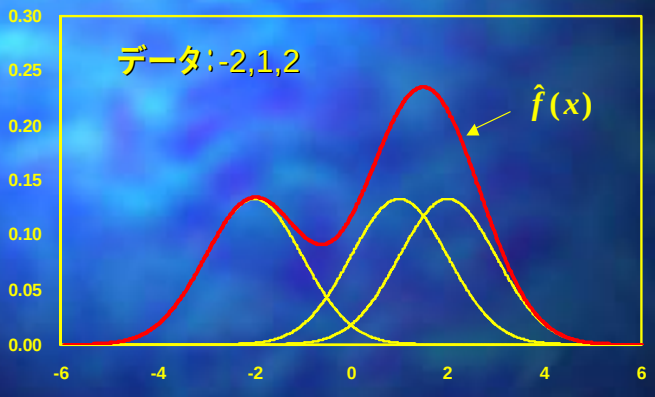


ノンパラメトリックOVL

- ノンパラメトリック法
  - カーネル密度推定が代表的

    各観察値に、各観測値を中心とする密度関数を当てはめ、それらの確率密度の和（重合せ）

ノンパラメトリックOVL



カーネル密度推定の例（n=3）

ノンパラメトリックOVL

- カーネル密度推定の定義式

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

x_i ：観測値 $(i=1,2,\cdots,n)$

n ：症例数

$K(y)$ ：カーネル関数（標準正規分布など）

h ：平滑パラメータ／バンド幅

ノンパラメトリックOVL

例）カーネル関数が標準正規分布$N(0,1^2)$の場合

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right) \qquad N(x_i, h^2)$$

$$= \sum_{i=1}^{n}\frac{1}{n}\frac{1}{\sqrt{2\pi}h}\exp\left\{-(x - x_i)^2 / 2h^2\right\}$$

各観測値の密度関数$N(x_i,h^2)$からxにおける確率密度に、それぞれ$1/n$をかけ、和をとったもの

ノンパラメトリックOVLの使用上の注意

- 影響力

  平滑パラメータ(h)の選択　>>　カーネル関数($K(y)$)の選択

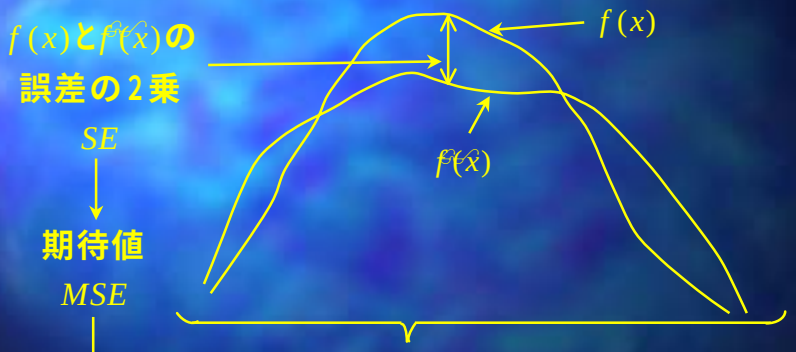
- 探索的試験
  - 様々な平滑パラメータを当てはめて、分布の事前情報に適合するような平滑パラメータを選択

ノンパラメトリックOVLの使用上の注意（つづき）

- 検証的試験（ブリジング試験など）
  - 平滑パラメータの恣意的選択が可能

    例　OVLを最大にするものを選択可能
  - 平滑パラメータの**算出式/選択アルゴリズム**を事前にプロトコル等に規定しておくべき
  - 最良な**算出式/選択アルゴリズム**を知る必要性

推定の良さの評価指標



推定の良さの評価指標

- AMISE最小化規準

$$MSE = Var_f[\hat{f}(x)] + \left\{Bias_f[\hat{f}(x)]\right\}^2$$

$$AMISE = \frac{R(K)}{nh} + \frac{h^4 \sigma_K^4 R(f'')}{4}$$

$$h_{AMISE} = \left[\frac{R(K)}{\sigma_K^4 R(f'')}\right]^{1/5} n^{-1/5}$$

$K(y)$の選択で決まる

推定する部分

$$R(g) = \int \{g(x)\}^2 dx$$

$$\sigma_K^2 = \int y^2 K(y) dy$$

平滑パラメータの推定方法

- SNR法
  - 真の分布として正規分布を想定し $R(f'')$ を推定
- OS法
  - 分布によっては、平滑化し過ぎてしまう傾向
- SROT法
  - Silverman(1986)[11]が考案した経験的方法
- SJPI法
  - Sheather and Jones (1991) [10] が考案
  - $R(f'')$ 推定の方法論で、現在のところ最良の方法

KDEプロシジャ

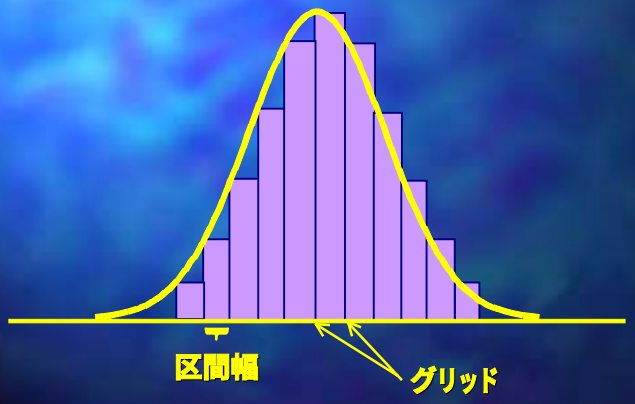
- バージョン8より使用可能
- カーネル関数
  - Normalカーネル
- 平滑パラメータ
  - 4つの推定方法 (Method=オプション)
    - SNR法 (デフォルト)、OS法、SROT法、SJPI法
  - 任意値 (BWM=値)

入力データセット;d1

群	測定値
g	val
1	-1.17057
1	-1.82856
◖	◖
2	0.09523
2	-1.22060
◖	◖

積分とグリッド・区間幅

グリッド数 ＝ 区間数 ＋ 1



プログラム例1

- 積分範囲-6～6を100区間 (グリッド101)
- グリッドごとの密度推定値を得る

```
PROC KDE DATA= d1 METHOD= SNR
     GRIDL= -6 GRIDU= 6 NGRID= 101
     OUT = test;
     BY g;
     VAR val;
RUN;
```


DISCRIMプロシジャ

- バージョン8以前でも使用可能
- 多群、多変量の分布を推定可能
- パラメトリック法(Method = Normal)
  - 等分散;1次判別を利用　(pool = yes)
  - 不等分散;2次判別を利用　(pool = no)

DISCRIMプロシジャ（つづき）

- ノンパラメトリック法(Method = npar)
  - カーネル関数(Kernel=オプション)
    - Uniform、Normal、Epanechnikov、Biweight、Triweight
  - 平滑パラメータ(r=値)
    - 任意値

平滑パラメータの指定方法（DISCRIMプロシジャ）

- 各群で平滑パラメータを共通とする
  - 問題なし
- 各群で平滑パラメータを共通としない
  - DISCRIMプロシジャを群毎に平滑パラメータを変えて2回実行
- DISCRIMにおける平滑パラメータ
  - $h = \sigma r$

プログラム例（テストデータ）

- グリッドごとの密度推定値を得るため作成
- 積分範囲-6～6を100区間（グリッド数101）

```
DATA test;
  DO val= -6 TO 6 BY 0.12;
    OUTPUT;
  END;
RUN;
```


プログラム例2

- パラメトリック法;正規分布、等分散

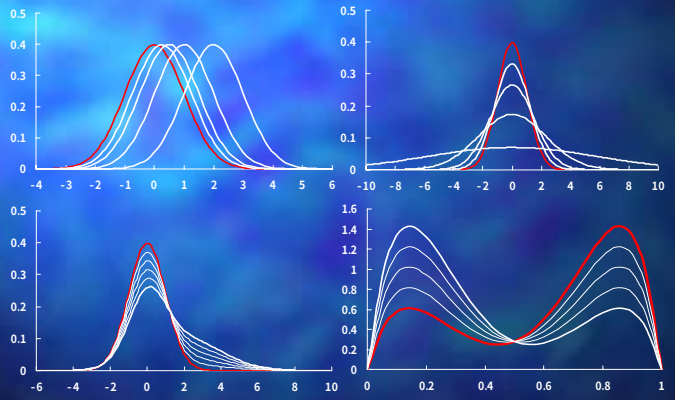
```
PROC DISCRIM DATA= d1
     TEST= test TESTOUTD= test1
     METHOD=NORMAL POOL=YES DISTANCE;
  CLASS g;
  VAR val;
RUN;
```

プログラム例3

- ノンパラメトリック法;Normalカーネル、SNR法

```
PROC DISCRIM DATA= d1
     TEST= test TESTOUTD= test1
     METHOD=NPAR KERNEL=NORMAL R= &r
     POOL=NO DISTANCE;
  CLASS g;
  VAR val;
RUN;
```


数値計算 I
Stine and Heyse(2001)[12]



数値計算 I
Stine and Heyse(2001)[12] (つづき)

- 密度推定法(18通り)
  - パラメトリック (理論式、DISCRIM)
    - 等分散、不等分散
  - ノンパラメトリック (KDE、DISCRIM)
    - KDE プロシジャ; 4つの平滑パラメータと Normalカーネル
    - DISCRIMプロシジャ;2つの平滑パラメータと5つの カーネルの組合せ

数値計算 I
Stine and Heyse(2001)[12] (つづき)

- サンプル100／群
- グリッド数101
- 繰り返し1000回
- 精度の評価 Bias (偏り) の平均、標準偏差

Bias = OVL推定値 − OVL真値

数値計算 I
結果 (表1，2) と考察

- 設定1 (平均値の異なる正規分布)
  - パラメトリック法 (等分散) が最適
  - 他の手法も悪くない
- 設定2 (標準偏差の異なる正規分布)
  - パラメトリック法 (不等分散) が最適
  - 次に、ノンパラメトリック法が良い
  - パラメトリック法 (等分散) は過大評価

数値計算 I
結果 (表1，2) と考察 (つづき)

- 設定3 (正規分布 vs 混合正規分布)
  - パラメトリック法 (等分散) が最適
  - 次にノンパラメトリック法が良い
  - パラメトリック法 (不等分散) は過小評価
- 設定4 (混合ベータ分布 vs 混合ベータ分布)
  - ノンパラメトリック法が良いが・・・
  - SJPI法はかなり過大評価

数値計算Ⅰ
結果（表1，2）と考察（つづき）

- パラメトリック法（設定1～4）
  - 理論式　　DISCRIMプロシジャから算出
- 「類似性」を判断する基準の設定が重要
  - 0.7以上？、0.8以上？・・・
  - プロトコルに事前に規定

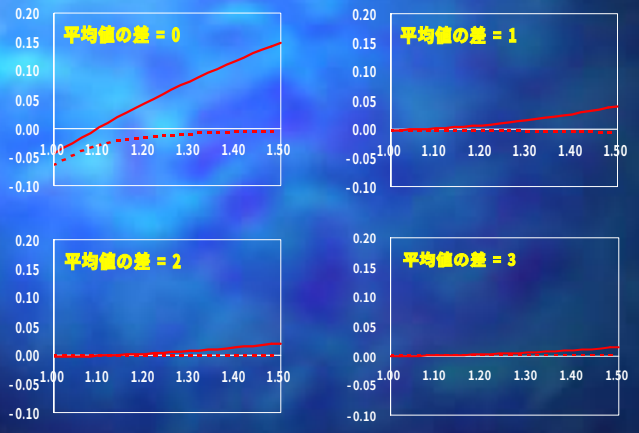
数値計算Ⅰ
不等分散性の影響

- 設定5　　$N(0,1^2)$　vs　$N(\mu, \sigma^2)$

  $\mu = 0, 0.5, \ldots, 2.5, 3$　(0.5刻み)

  $\sigma = 1, 0.02, \ldots, 1.48, 1.5$　(0.02刻み)
- 密度推定法（理論式）
  - パラメトリック法（等分散、不等分散）の2通り
- 他の条件は設定1～設定4と同じ

不等分散性の影響
結果 (図4)



縦軸：Bias平均値、横軸：標準偏差の比σ/1

――― ：パラメトリック法 (等分散)、 - - - - ：パラメトリック法 (不等分散)

不等分散性の影響
考察

- 分散比に対して、平均の差が大きい場合
  - 等分散性を仮定したパラメトリック手法でも
    十分な推定精度が保たれる (100例/群)

数値計算II
サンプルサイズが異なる場合

- $N(0,1^2)$ vs $N(0,\sigma^2)$; σ = 1.228, 1.51, 2.31
- 密度推定法（7通り）
  - パラメトリック（理論式）
    - 等分散、不等分散
  - ノンパラメトリック（KDE、DISCRIM）
    - KDEプロシジャ；4つの平滑パラメータとNormalカーネル
    - DISCRIMプロシジャ；SROT法とNormalカーネル
- グリッド数51、繰り返し2000回

数値計算II
サンプルサイズが異なる場合(つづき)

- サンプルサイズ(n_1:n_2)

1:1	2:1	4:1	8:1
120:120	120:60	120:30	120:15
60:60	60:30	60:15	
30:30	30:15		
15:15			

- 精度の評価　　Bias（偏り）の平均、標準偏差

数値計算II
結果(表3)と考察

- サンプルサイズの比が大きくなるか、サンプルサイズの減少と共に、推定精度が落ちる

多変量への拡張

- 主要変数が複数ある場合、それら変数の同時分布が興味の対象となる
  - 例　AUCとC_{max}の同時分布
- 各パラメータを独立に比較するより、有用な比較となる場合がある

ノンパラメトリック法

- 多変量における密度推定関数

$$\hat{f}(\mathbf{x}) = \frac{1}{n\|\mathbf{H}\|}\sum_{i=1}^{n} K_p\left[\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\right]$$

x_i　：p次元ベクトル(i=1,2,･･･,n)

$\mathbf{H}$　：平滑パラメータ行列(p×p)

平滑パラメータ行列の選択

- 平滑パラメータ行列の要素の数；$p(p+1)/2$
  - $\mathbf{H}= hI$
    各変数が独立、基準化した値に最適
  - $\mathbf{H}= diag(h_1,h_2,\cdots,h_p)$
    各変数が独立のとき最適（2変量の時、ロバスト）
  - $\mathbf{H}= h_0 \Sigma^{1/2}$
    多変量正規分布のとき最適
  - $\mathbf{H}= h_0\, diag(\Sigma^{1/2})$
    上記の対角要素、各変数が独立のとき最適

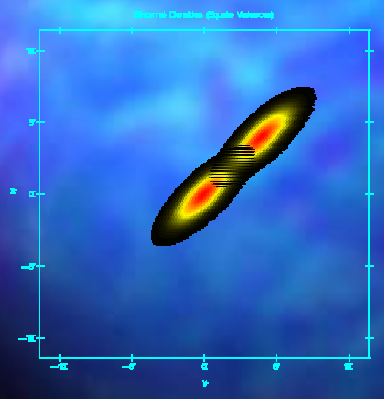
DISCRIMプロシジャ

- 多変量パラメトリックOVLを解析的に求めるのは非常に容易
- 多変量ノンパラメトリック法
  - $\mathbf{H} = h\mathbf{I}$ ；Metric=Identity　R= h
  - $\mathbf{H} = diag(h_1, h_2, \cdots, h_p)$ ；実行できない
  - $\mathbf{H} = h_0 \sum^{1/2}$ ；Metric=Full　R= h_0
  - $\mathbf{H} = h_0\, diag(\sum^{1/2})$ ；Metric=Diagonal　R= h_0

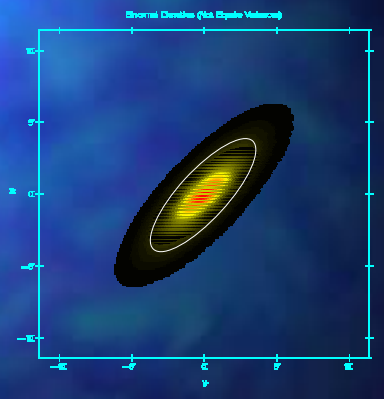
KDEプロシジャ

- 2変量まで推定可能
- 平滑パラメータ行列（SNR法、任意値）
  - $\mathbf{H} = n^{-1/6} diag(\sigma_1, \sigma_2)$ ；Method= SNR
  - $\mathbf{H} = diag(h_1, h_2)$ ；BWM= h_1　h_2

数値計算III
2変量におけるOVL



設定1



設定2

数値計算III
2変量におけるOVL（つづき）

- 密度推定法（6通り）
  - パラメトリック（DISCRIM）
    - 等分散、不等分散
  - ノンパラメトリック（KDE、DISCRIM）
    - KDE ；SNR法
    - DISCRIM ；Normal,Uniform,Epanechnikov カーネルの各AMISE最小化基準
- サンプルサイズ100/群、グリッド51×51、繰り返し1000回
- 精度の評価指標　Biasの平均と標準偏差

数値計算III
結果（表4）と考察

- 設定1
  - パラメトリック法（等分散）が最適
  - パラメトリック法では100例/群でも良い
- 設定2
  - パラメトリック法（不等分散）が最適
  - 100例/群では、どの推定法も1変量の100例/群での精度より劣る

数値計算III
結果（表4）と考察（つづき）

- 500例/群だと、精度が1変量の100例/群の場合と同様の結果が得られた
- 多変量の場合のノンパラメトリック密度推定法の研究は、1変量より遅れているため、今後の展開に注目していく必要がある

まとめ

- KDEプロシジャ、DISCRIMプロシジャを用い、OVLを推定する様々な方法を示した
- 正規性、等分散性などの条件が満たされない場合のOVLの推定精度を評価した
- 前提条件が崩れた場合を考慮し、手法を選択する必要がある

ご清聴、ありがとうございました。

okuymakb@banyu.co.jp