Multivariate Statistical Inference on Random Truncation Data

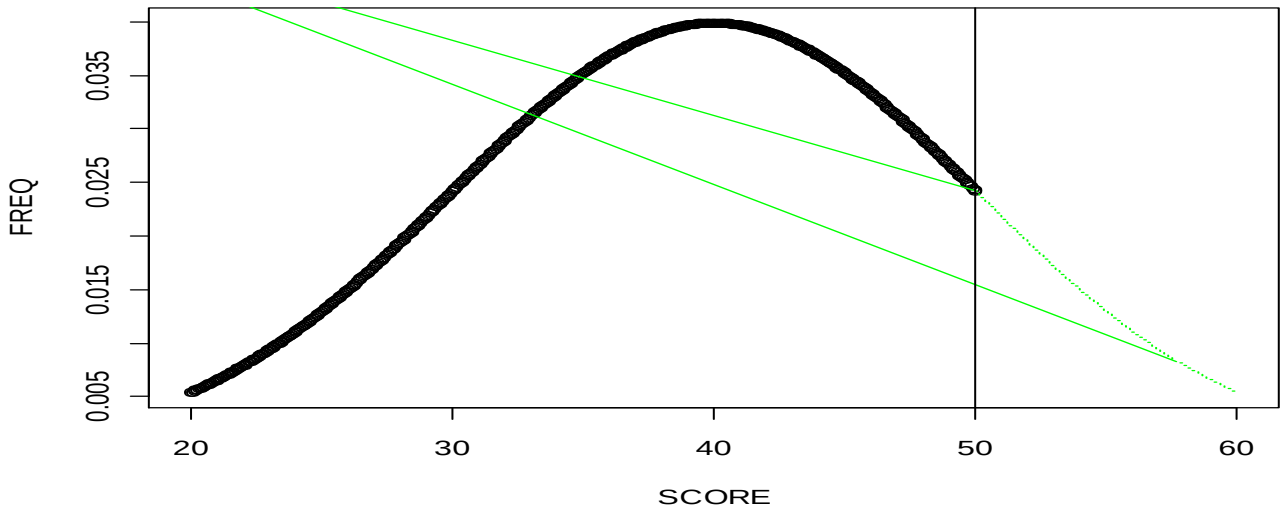
Presented on August, 7-10, 2008, 統計関連合大会 統計関連合大会，慶応義塾大学

江村剛志, Takeshi EMURA　（国立交通大学　統計学研究所）
今野良彦, Yoshihiko KONNO　（日本女子大学　理学部）

This work is currently published as

T. Emura, Y. Konno, Multivariate Normal Distribution Approaches for Dependently Ttruncated Data, Statistical Papers 53 (2012), 133-149

(on 2012/2/18, Takeshi EMURA)



発表の流れ

先行研究の紹介

- 切断正規分布とCohen (1959) の方法
- 左切断データとLynden-Bell (1971) の方法

新たな手法の提案

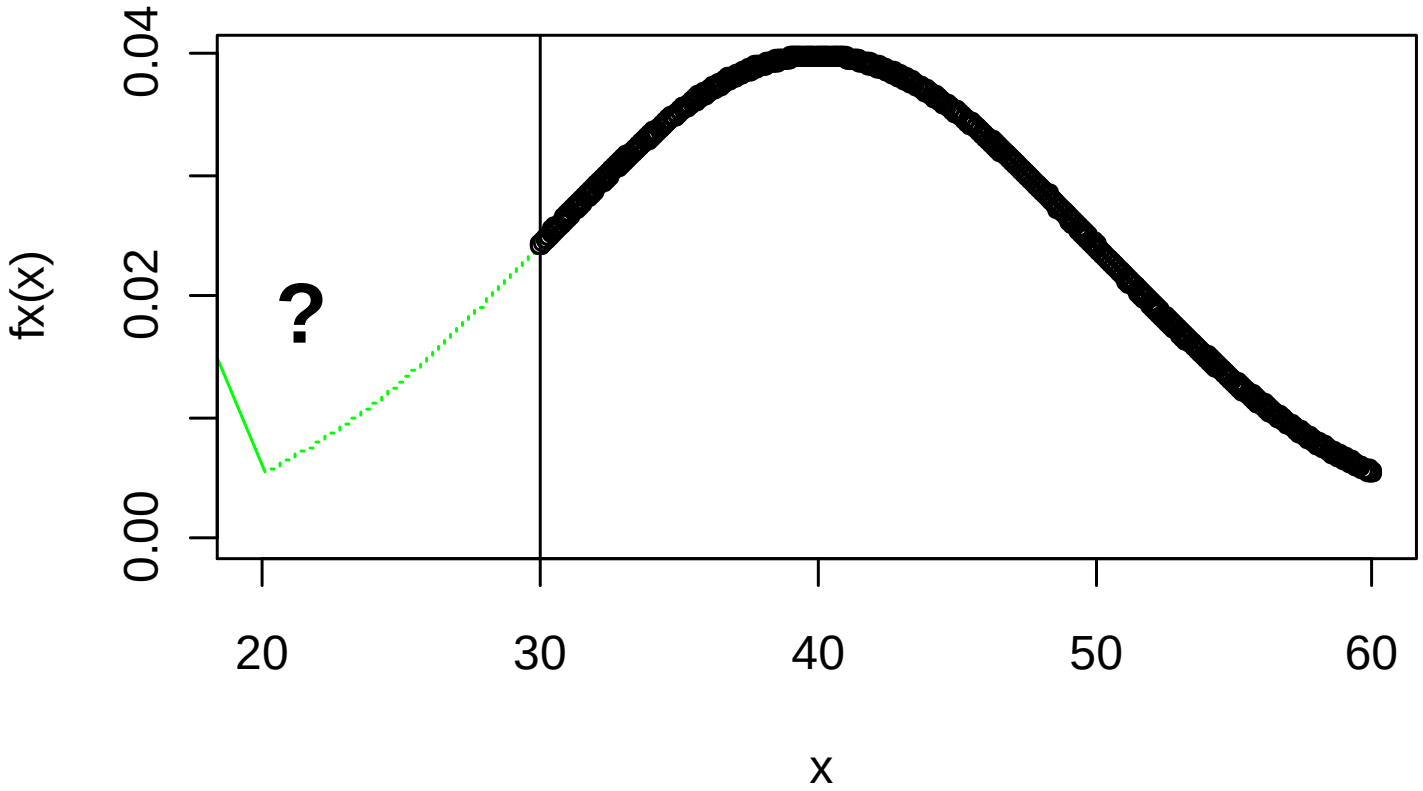
- ２次元正規分布によるアプローチ
- 切断による情報損失の解析
- 2-stage course placement system への応用
- 数値実験によるLynden-Bell の方法との比較

研究背景

- 固定切断データ（Cohen 1959, 1960）：

  $\{X_j; j=1,\ldots,n\}$ ただし $l \le X_j$

  カットオフ値 l は既知



研究背景；固定切断データ

モデル；　$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2}$

$l \le X$　の下で条件付密度を最大化

- 母数の推定（Cohen, 1959, 1960）

$(\hat{\mu}_X, \hat{\sigma}_X^2)$　＊要数値解法

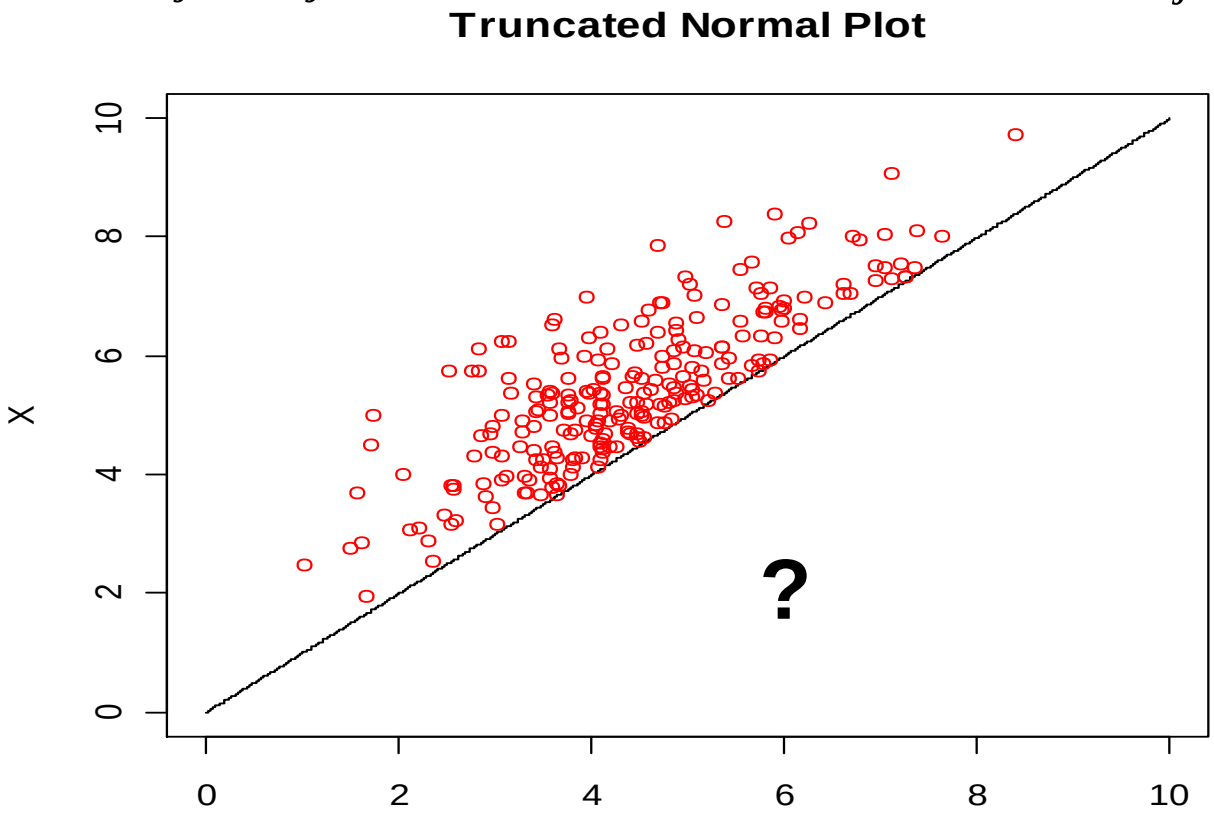
- Inclusion probability (Hansen & Zeger 1980)

$$\hat{c} = \hat{\Pr}(l \le X) = 1 - \Phi\left(\frac{l - \hat{\mu}_X}{\hat{\sigma}_X}\right)$$

研究背景

- 左切断（Left-truncation）データ:

$$\{(L_j, X_j)(j=1,\ldots,n)\} \text{ ただし } L_j \le X_j$$



L~N(5,2), X~N(5,2), Cov(L,X)=0.5

研究背景；独立左切断データ

記号

切断前確率変数 (L^O, X^O)

- $L^O > X^O$ ならば切断　（何も観測されない）
- $L^O \leq X^O$ ならば観測可能

$\Rightarrow$　$(L^O, X^O) = (L_j, X_j)$　とおく

切断後観測値 ;$\{(L_j, X_j)(j = 1, \ldots, n)\}$　$L_j \leq X_j$

研究背景；独立左切断データ

モデル；$L^O \perp X^O$、

L^O, X^Oの分布形は仮定しない

- 切断前分布関数の推定(Lynden-Bell, 1971)

$$\hat{F}_X(t) = \hat{\Pr}(X^O \le t)$$

- Inclusion Probability (He and Yang, 1998)

$$\hat{c} = \hat{\Pr}(L^O \le X^O)$$

＊左切断データのみを使用して推測；

$$\{(L_j, X_j)(j = 1, \ldots, n)\} \qquad L_j \le X_j$$

研究背景；独立性検定

仮定； $H_0 : L^O \perp X^O$

- ケンドール τ を応用した検定 (Tsai 1990)
  $\tau(L^O, X^O) = 0$ under H_0

  統計量 : $\hat{\tau}$
- U-statistics を応用しTsai の統計量の漸近分散を得る (Martin & Betensky 2005)
- H_0 が成立しない多くの応用例が存在；Tsai, Martin & Betensky

研究背景;まとめ

	Parametric	Nonparametric
1・固定左切断 $l \le X_j$ l: 既知	Xの分布の推定； Cohen (1959, 60) Hansen & Zeger (1980)その他多数	文献なし
2・独立左切断 $L_j \le X_j$ L , X: 独立	文献少数	Xの分布の推定； Lynden-Bell (1971) He and Yang (1998), その他多数
3・従属左切断 $L_j \le X_j$ L , X: 従属	文献少数 提案する方法（1，2を特別な場合として含む）	Tsai (1990) Martin & Betensky (2005)など少数

２次元正規分布によるアプローチ

- 切断前確率変数のモデル

$$\begin{pmatrix} L^O \\ X^O \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_L \\ \mu_X \end{pmatrix}, \begin{bmatrix} \sigma_L^2 & \sigma_{LX} \\ \sigma_{LX} & \sigma_X^2 \end{bmatrix} \right)$$

未知パラメータ； $\boldsymbol{\theta}' = (\mu_X, \mu_L, \sigma_X^2, \sigma_L^2, \sigma_{LX})$

- データ

$$\{(L_j, X_j); j = 1, \ldots, n\} \quad \text{subject to} \quad L_j \le X_j$$

- Inclusion probability

$$c(\boldsymbol{\theta}) = \Pr(L^O \le X^O) = \Phi\left(\frac{\mu_X - \mu_L}{\sqrt{\sigma_X^2 + \sigma_L^2 - 2\sigma_{LX}}} \right)$$

尤度法

- 条件付尤度 $f(l,x \mid L^O \le X^O) = \begin{cases} \dfrac{f(l,x)}{\Pr(L^O \le X^O)} & l \le x \\ 0 & l > x \end{cases}$
- データ $\{(L_j, X_j); L_j \le X_j, j = 1,\ldots,n\}$ の尤度

$$L(\boldsymbol{\theta}) = \left(\frac{1}{c(\boldsymbol{\theta}) \cdot 2\pi\sqrt{\sigma_L^2\sigma_X^2 - \sigma_{LX}^2}} \right)^n \exp\left[-\frac{1}{2}\sum_j D_j^2(\boldsymbol{\theta}) \right]$$

ここで $$c(\boldsymbol{\theta}) = \Pr(L^O \le X^O) = \Phi\left(\frac{\mu_X - \mu_L}{\sqrt{\sigma_X^2 + \sigma_L^2 - 2\sigma_{LX}}} \right)$$

$$D_i^2(\boldsymbol{\theta}) = \frac{\sigma_X^2(L_i - \mu_L)^2 - 2\sigma_{LX}(L_i - \mu_L)(X_i - \mu_X) + \sigma_L^2(X_i - \mu_X)^2}{(\sigma_L^2\sigma_X^2 - \sigma_{LX}^2)}$$

尤度法

- 対数尤度

$$l = -n\log\{c(\boldsymbol{\theta})\} - \frac{n}{2}\log\{(2\pi)^2\} - \frac{n}{2}\log(\sigma_L^2\sigma_X^2 - \sigma_{LX}^2)$$

$$-\frac{1}{2}\sum_j D_j^2(\boldsymbol{\theta})$$

- スコア関数 $\dfrac{\partial l}{\partial \boldsymbol{\theta}} = -\dfrac{n}{c(\boldsymbol{\theta})}\dfrac{\partial c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_i \mathbf{U}_i(\boldsymbol{\theta})$

- 最尤推定値

$$\mathbf{U}_i(\boldsymbol{\theta}) = \frac{1}{\sigma_L^2\sigma_X^2 - \sigma_{LX}^2}\begin{bmatrix} \sigma_X^2(L_i-\mu_L) - \sigma_{LX}(X_i-\mu_X) \\ -\sigma_{LX}(L_i-\mu_L) + \sigma_L^2(X_i-\mu_X) \\ -\sigma_X^2/2 + \sigma_X^2 D_i(\boldsymbol{\theta})/2 - (X_i-\mu_X)/2 \\ -\sigma_L^2/2 + \sigma_L^2 D_i(\boldsymbol{\theta})/2 - (L_i-\mu_L)/2 \\ \sigma_{LX} - \sigma_{LX} D_i^2(\boldsymbol{\theta}) + (L_i-\mu_L)(X_i-\mu_X) \end{bmatrix}$$

$$\hat{\boldsymbol{\theta}};\quad \mathbf{0} = -\frac{n}{c(\boldsymbol{\theta})}\frac{\partial c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_i \mathbf{U}_i(\boldsymbol{\theta})$$

切断による情報損失

$$\boldsymbol{\theta} = (\mu_L, \mu_X)' \quad \begin{bmatrix} \sigma_L^2 & \sigma_{LX} \\ \sigma_{LX} & \sigma_X^2 \end{bmatrix} \quad ; \text{既知とおく}$$

- $\boldsymbol{\theta}$ の推定に関するFisher 情報量

$$I\{c(\boldsymbol{\theta})\} = \begin{bmatrix} \sigma_L^2 & \sigma_{LX} \\ \sigma_{LX} & \sigma_X^2 \end{bmatrix}^{-1} - \frac{w\{c(\boldsymbol{\theta})\}}{\sigma_L^2 + \sigma_X^2 - 2\sigma_{LX}} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

切断による情報損失分

- 単調性 $c' > c \quad \Rightarrow \quad I(c') - I(c);$ 半正定値

  ＊w(c); 非負単調減少関数

Two-stage course placement system
Schiel & Harmston (2000)への応用例

- 各生徒に対するScreening test（S）と Placement test（P）の分布；

$$\begin{pmatrix} \text{Screening test}\,(S^O) \\ \text{Placement test}\,(P^O) \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_S \\ \mu_P \end{bmatrix}, \begin{bmatrix} \sigma_S^2 & \sigma_{SP} \\ \sigma_{SP} & \sigma_P^2 \end{bmatrix} \right)$$

設定；Screenig test で基準点(K)以下の場合、
Placement test を受験

- **Placement test 受験者のみのデータ；**

$$\{(S_j, P_j);\ j = 1, \ldots, n\} \ \text{subject to} \ S_j \le K(\text{定数})$$

＊成績の悪い生徒がサンプルに入りやすい

＊(P)受験者からSの値をさかのぼって調査

Two-stage course placement system
Schiel & Harmston (2000)への応用例

- サンプリング基準の書き換え

$$S^O \le K(\text{定数})$$

$$\Leftrightarrow P^O - S^O \ge P^O - K$$

$$\Leftrightarrow X^O \ge L^O$$

ここで、$L^O = P^O - K$ 、$X^O = P^O - S^O$

- 観測データ；

$\{(L_j, X_j); j = 1,\ldots,n\}$ subject to $L_j \le X_j$

- 最尤推定値を変換

$$(\hat{\mu}_X, \hat{\mu}_L, \hat{\sigma}_X^2, \hat{\sigma}_L^2, \hat{\sigma}_{LX}) \rightarrow (\hat{\mu}_S, \hat{\mu}_P, \hat{\sigma}_S^2, \hat{\sigma}_P^2, \hat{\sigma}_{SP})$$

数値実験（センター試験モデル）

- 2008年度、国語（X）と英語（Y）の結果

$$\begin{pmatrix} X^O \\ Y^O \end{pmatrix} \sim N\left(\begin{bmatrix} 60.82 \\ 62.63 \end{bmatrix}, \begin{bmatrix} 19.64^2 & (19.64)(1681)\rho_{\mathrm{LX}} \\ (19.64)(1681)\rho_{\mathrm{LX}} & 16.81^2 \end{bmatrix} \right)$$

公的入手不可能；0, 0.25, 0.50, 0.75　を設定

- 設定；ある大学の合格基準；$X^O + Y^O \geq 120$
- <u>合格者のみのデータ</u>

$$\{(X_j, Y_j); j = 1, \ldots, n\} \quad \text{subject to} \quad X_j + Y_j \geq 120$$

- Inclusion probability

$$c(\boldsymbol{\theta}) = \Pr(X^O + Y^O \geq 120)$$

数値実験（センター試験モデル）

- 合格基準の書き換え

$$X^O + Y^O \geq 120$$

$$\Leftrightarrow X^O \geq 120 - Y^O$$

$$\Leftrightarrow X^O \geq L^O, \quad \text{where } L^O = 120 - Y^O$$

- 観測データ；

$$\{(L_j, X_j); j = 1, \ldots, n\} \quad \text{subject to} \quad L_j \leq X_j$$

$$\text{where} \quad L_j = 120 - Y_j$$

- 最尤推定値 $\hat{\boldsymbol{\theta}}' = (\hat{\mu}_X, \hat{\mu}_L, \hat{\sigma}_X^2, \hat{\sigma}_L^2, \hat{\sigma}_{LX})$

国語の母平均60.82の推定値

数値実験（センター試験モデル）

- 国語と英語間の独立性を仮定する場合

$$\rho_{XY} = 0 \quad \Rightarrow \quad H_o : L^O \perp X^O$$

1) H_o の下で、Lynden-Bell 推定量 $\hat{F}_X(x)$ を適用

$$\hat{\mu}_X^{NP} = \int_{-\infty}^{\infty} x d\hat{F}_X(x)$$

国語の母平均60.82の推定値

2) $\sigma_{LX} = 0; \text{fix}$ の下での最尤推定値

$$\hat{\boldsymbol{\theta}}' = (\hat{\mu}_X^0, \hat{\mu}_L^0, \hat{\sigma}_X^{0\,2}, \hat{\sigma}_L^{0\,2})$$

国語の母平均60.82の推定値

数値実験（3つの方法の比較）

$\rho_{XY}=0$ の下でのMLE

MLE

Lynden-Bell

推定目標 μ_X =60.82			$\hat{\mu}_X^0$	$\hat{\mu}_X$	$\hat{\mu}_X^{NP}$
1) Not correlated ρ_{XY} =0.00,	n=100	Mean	60.761	59.854	60.799
		MSE	7.721	45.495	14.540
	n=200	Mean	60.806	60.438	60.837
		MSE	3.900	17.173	7.139
2) Moderately correlated ρ_{XY} =0.50,	n=100	Mean	68.436	59.559	68.879
		MSE	60.986	76.347	67.872
	n=200	Mean	68.386	60.183	68.803
		MSE	58.636	25.723	65.237

数値実験（まとめ）

- $\rho_{XY}=0$ の場合

  不必要な相関パラメータを推定したことにより、$\hat{\mu}_X$ の効率は著しく減少

  当然 $\hat{\mu}_X^o$　$\hat{\mu}_X^{NP}$ の推定精度は高い

- $\rho_{XY}\geq 0.50$ の場合

  相関パラメータを入れることにより、$\hat{\mu}_X$ の一致性が保証され、高い推定効率

  一致性の欠落により、$\hat{\mu}_X^o$　$\hat{\mu}_X^{NP}$ の推定精度は低い

全体のまとめ

- 独立条件 $L^o \perp X^o$ の成立しない場合での推定法を提案
- 利点；

  ２次元正規分布に基づく簡単な方法

  一致性、漸近正規性、漸近有効性

  Fisher情報量による切断効果の理解
- 欠点；

  正規性の仮定が必要

  相関が小さい場合にはLynden-Bellより劣る