

変化点問題，丸められたデータの問題

西山 陽一

統計数理研究所

April 13, 2011

統計数理セミナー

講演の構成

1. 変化点問題

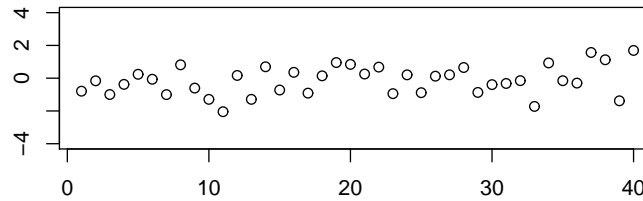
- 1.1. パラメトリック変化点問題
- 1.2. ノンパラメトリック変化点問題

2. 丸められたデータの問題

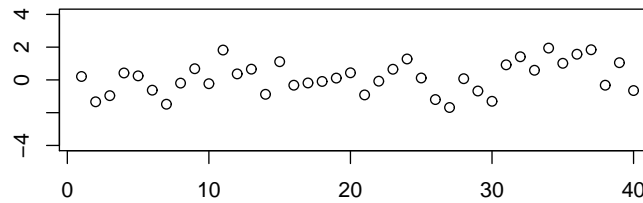
- 2.1. 丸められたデータによる適合度検定
- 2.2. 丸められたデータによる Z -推定

1. 変化点問題

H0 (no change point)



H1 (a change point)



上下の図の違いが判別できますか？

大学院時代の友人がPMDAで統計解析の仕事をしています。

『安全部に移ってから、仕事で変化点問題を扱っていて、

1. ある特定の薬剤でのある特定の種類の副作用の発生について、医療機関からの報告を随時受け付ける；
2. 毎週の副作用発生数が強度 λ のポアソン分布に従うと仮定する；
3. ある週以降に副作用の発生数が増加している（ λ が増加方向に変化した）と判断されたら、「何かまずいことが生じているかもしれない」というアラートを発して医学的な検討を開始する；

というシステムを作ろうとしています』

『副作用の発生時刻の間隔の長さが指数分布に従うとし、ある時点からその平均が小さくなったかどうかを検定するというアプローチも考えています』

1.1. パラメトリック変化点問題

測度空間 $(\mathcal{X}, \mathcal{A}, \mu)$, 確率密度 $f(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}^d$.

独立データ $\{X_i\}_{i=1, \dots, n}$ について :

H_0 : 全ての $i = 1, \dots, n$ について $\theta = \theta_0$.

H_1 : ある異なる θ_0, θ_1 と , ある $u_* \in (0, 1)$ が存在して ,
 $i \leq u_* n$ に対しては $\theta = \theta_0$, $i > u_* n$ に対しては $\theta = \theta_1$.

- 「尤度比の方法」
- 「Fisher-Score Process 法」 我々は推奨

「尤度比の方法」が Csörgő & Horváth (1997) の本に書いてある :

$$\Lambda_{n,k} = \frac{\sup_{\theta} \prod_{i=1}^n f(X_i, \theta)}{\sup_{\theta, \vartheta} \prod_{i=1}^k f(X_i, \theta) \prod_{i=k+1}^n f(X_i, \vartheta)}$$

とおくと, H_0 のもとで

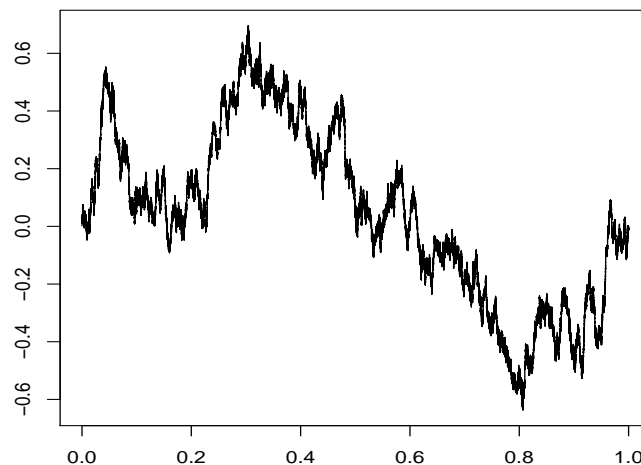
$$\max_{1 \leq k \leq n} \frac{k}{n} \left(1 - \frac{k}{n}\right) (-2 \log \Lambda_{n,k}) \rightarrow^d \sup_{u \in [0,1]} \sum_{p=1}^d |B^{\circ,p}(u)|^2.$$

ただし $(B^{\circ,1}, \dots, B^{\circ,d})^{\top}$ は独立な標準ブラウン橋のベクトル . 標準ブラウン橋とは平均ゼロの確率過程であって共分散構造が

$$EB^{\circ,p}(u_1)B^{\circ,p}(u_2) = u_1 \wedge u_2 - u_1u_2$$

であるもの .

標準ブラウン橋 $\{B^\circ(u), u \in [0, 1]\}$ のサンプルパス.



$$EB^\circ(u) = 0, \quad EB^\circ(u)B^\circ(v) = u \wedge v - uv.$$
$$P\left(\sup_{u \in [0,1]} |B^\circ(u)| \leq x\right) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 x^2).$$

しかしながら、「尤度比の方法」は：

- 帰無仮説のもとでの証明は短くない（15 ページ）。
- 対立仮説のもとでの議論がなされていない。
- 確率過程への拡張が容易であるかどうか、私にははっきりしない。

「Fisher-Score Process 法」が Horváth & Parzen (1994) によって提案されている： $l(x, \theta) = \log f(x, \theta)$ において

$$U_{n,k} = \frac{1}{n} \sum_{i=1}^k \dot{l}(X_i, \hat{\theta}_n)$$

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^n \dot{l}(X_i, \hat{\theta}_n) \dot{l}(X_i, \hat{\theta}_n)^\top$$

(ただし $\hat{\theta}_n$ は最尤推定量) と定義して

$$S_n = n \max_{1 \leq k \leq n} U_{n,k}^\top \hat{I}_n^{-1} U_{n,k}$$

を提案すると, H_0 のもとで

$$\mathcal{S}_n \rightarrow^d \sup_{u \in [0,1]} \sum_{p=1}^d |B^{\circ,p}(u)|^2.$$

- Horváth & Parzen (1994) の証明は 8 ページだったが, 我々は 1 ページの証明を与える .
- Horváth & Parzen (1994) は対立仮説のもとでの議論をしていないが, 我々は議論する .
- 確率過程への拡張が容易であるから, 私は「Fisher-Score Process 法」を推奨する .

「Fisher-Score Process 法」の H_0 のもとでの簡単な証明 .

$$V_n(u, \theta) := \frac{1}{n} \sum_{i=1}^n (1\{i \leq nu\} - u) \dot{l}(X_i, \theta), \quad (u, \theta) \in [0, 1] \times \Theta$$

とおくと , 簡単な計算により , 次が成立することがわかる :

$$S_n = n \sup_{u \in [0, 1]} V_n(u, \hat{\theta}_n)^\top \hat{I}_n^{-1} V_n(u, \hat{\theta}_n).$$

このことは $u = k/n$ とおいたときの

$$U_{n,k} = V_n(k/n, \hat{\theta}_n)$$

という関係から明らかである .

第1のポイントは

$$\sup_{u \in [0,1]} |\sqrt{n}(V_n(u, \hat{\theta}_n) - V_n(u, \theta_0))| \xrightarrow{p} 0$$

を示すところである。実際、そのことは

$$\begin{aligned} & \sqrt{n}(V_n(u, \hat{\theta}_n) - V_n(u, \theta_0)) \\ &= \frac{1}{n} \sum_{i=1}^n (1\{i \leq nu\} - u) \ddot{l}(X_i, \tilde{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_0) \end{aligned}$$

からほぼ明らかであろう。

一方, Donsker の定理より, 「確率場 $u \rightsquigarrow \sqrt{n}V_n(u, \theta_0)$ は $\ell^\infty([0, 1])$ 空間の中で $u \rightsquigarrow I(\theta_0)^{1/2}B^\circ(u)$ に弱収束する」ことが言える. よって連続写像定理により

$$u \rightsquigarrow nV_n(u, \hat{\theta}_n)^\top \hat{I}_n^{-1} V_n(u, \hat{\theta}_n)$$

は

$$u \rightsquigarrow (I(\theta_0)^{1/2}B^\circ(u))^\top I(\theta_0)^{-1} (I(\theta_0)^{1/2}B^\circ(u))$$

すなわち

$$u \rightsquigarrow \sum_{p=1}^d |B^{\circ,p}(u)|^2$$

に弱収束する. $\sup_{u \in [0,1]}$ をとって結論を得る.

「Fisher-Score Process 法」の H_1 のもとでの挙動。
記号

$$I_{\theta_0}(\theta) := E_{\theta_0} \dot{l}(X, \theta) \dot{l}(X, \theta)^\top$$

$$u_* E_{\theta_0} \dot{l}(X, \theta) + (1 - u_*) E_{\theta_1} \dot{l}(X, \theta) = 0 \quad \text{の解 } \theta = \theta_*$$

を導入すると,

$$2\mathcal{S}_n \geq n(v_*^\top I_*^{-1} v_* - o_P(1)) - O_P(1)$$

が成り立つ。ただし

$$I_* = u_* I_{\theta_0}(\theta_*) + (1 - u_*) I_{\theta_1}(\theta_*)$$

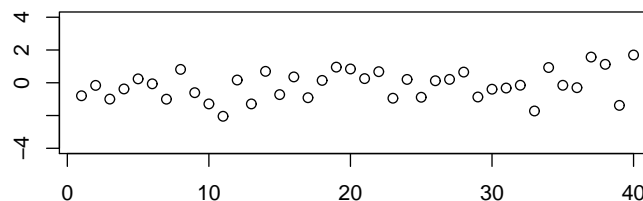
は正定値行列であり,

$$v_* = u_*(1 - u_*)(E_{\theta_0} \dot{l}(X, \theta_*) - E_{\theta_1} \dot{l}(X, \theta_*))$$

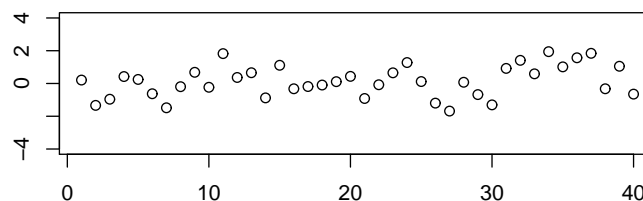
はゼロではないベクトルである。従って, 検定は一致性もつ。

シミュレーション：（図は $n = 40$ ）

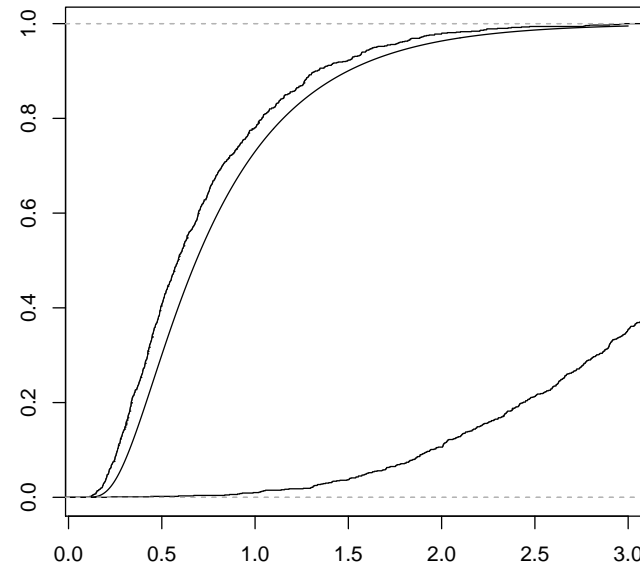
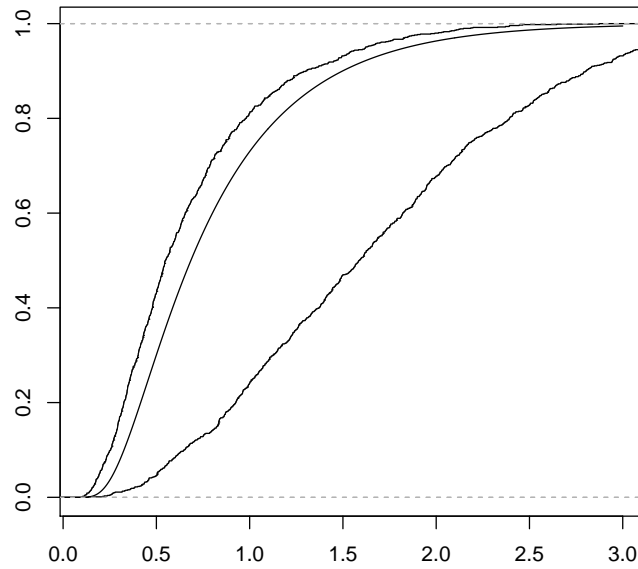
H0 (no change point)



H1 (a change point)



H_0 は , $X_1, \dots, X_n \sim N(0, 1)$,
 H_1 は , $X_1, \dots, X_{[3n/4]} \sim N(0, 1)$, $X_{[3n/4]+1}, \dots, X_n \sim N(1, 1)$.



$n = 40$ (左) と $n = 100$ (右) について,
 H_0 と H_1 のもとでの実験結果と漸近分布を重ねて描いた.

「Fisher-Score Process 法」の確率過程への拡張 1 . 拡散過程モデル

$$X_t = X_0 + \int_0^t S(X_s, \theta) ds + \int_0^t \sigma(X_s) dW_s$$

を考える .

$$U_{T,t} = \frac{1}{T} \int_0^t \frac{\dot{S}(X_s, \theta)}{\sigma(X_s)^2} (dX_s - S(X_s, \theta) ds) \Bigg|_{\theta = \hat{\theta}_T},$$

$$\hat{I}_T = \frac{1}{T} \int_0^T \frac{\dot{S}(X_s, \hat{\theta}_T) \dot{S}(X_s, \hat{\theta}_T)^\top}{\sigma(X_s)^2} ds,$$

$$\mathcal{S}_T = T \sup_{0 \leq t \leq T} U_{T,t}^\top \hat{I}_T^{-1} U_{T,t}.$$

この方法は Lee, Nishiyama & Yoshida (2006) より優れていると思う .

「Fisher-Score Process 法」の確率過程への拡張 2 . 拡散過程モデル

$$X_t = X_0 + \int_0^t S(X_s, \alpha) ds + \int_0^t \sigma(X_s, \beta) dW_s$$

を考える . 離散観測 $0 = t_0^n < t_1^n < \dots < t_n^n$ において , $\Delta_n = \max_{1 \leq i \leq n} |t_i^n - t_{i-1}^n|$ とおき ,

$$n\Delta_n^2 \rightarrow 0, \quad t_n^n \rightarrow \infty$$

を仮定する .

H_0 : α, β とともに変化しない .

H_1^α : α は変化する (β は問わない) .

H_1^β : β は変化する (α は問わない) .

の検定について , 同様の統計量が提案できる .

「Fisher-Score Process 法」の確率過程への拡張 3 .
 条件付き強度 $s \rightsquigarrow \lambda_s(\theta)$ をもつ計数過程 $s \rightsquigarrow N_s$ について ,

$$U_{T,t} = \frac{1}{T} \int_0^t \frac{\dot{\lambda}_s(\hat{\theta}_T)}{\lambda_s(\hat{\theta}_T)} (dN_s - \dot{\lambda}_s(\hat{\theta}_T) ds),$$

$$\hat{I}_T = \frac{1}{T} \int_0^T \frac{\dot{\lambda}_s(\hat{\theta}_T) \dot{\lambda}_s(\hat{\theta}_T)^\top}{\lambda_s(\hat{\theta}_T)} ds$$

とにおいて

$$\mathcal{S}_T = T \sup_{0 \leq t \leq T} U_{T,t}^\top \hat{I}_T^{-1} U_{T,t}.$$

特にポアソン過程の場合 (1) :

両側仮説 (つまり $H_1 : \lambda_0 \neq \lambda_1$) の場合には

$$\sqrt{S_T} = \frac{1}{\sqrt{N_T}} \sup_{t \in [0, T]} \left| \frac{t}{T} N_T - N_t \right|$$

とおくと, H_0 のもとで

$$\sqrt{S_T} \xrightarrow{d} \sup_{u \in [0, 1]} |B^\circ(u)|.$$

H_1 のもとで

$$\sqrt{S_T} \geq \sqrt{T} \left\{ \frac{u_*(1-u_*)|\lambda_0 - \lambda_1|}{\sqrt{u_*\lambda_0 + (1-u_*)\lambda_1}} - o_P(1) \right\} - O_P(1).$$

特にポアソン過程の場合 (2) :

片側仮説 (つまり $H_1 : \lambda_1 > \lambda_0$) の場合には

$$\mathcal{T}_T = \frac{1}{\sqrt{N_T}} \sup_{t \in [0, T]} \left(\frac{t}{T} N_T - N_t \right)$$

とおくと, H_0 のもとで

$$\mathcal{T}_T \rightarrow^d \sup_{u \in [0, 1]} B^\circ(u).$$

H_1 のもとで

$$\mathcal{T}_T \geq \sqrt{T} \left\{ \frac{u_*(1 - u_*)(\lambda_1 - \lambda_0)}{\sqrt{u_*\lambda_0 + (1 - u_*)\lambda_1}} - o_P(1) \right\} - O_P(1).$$

1.2. ノンパラメトリック変化点問題

1次元の独立データ $\{X_i\}_{i=1,\dots,n}$ について：

H_0 ：全ての $i = 1, \dots, n$ について，同一の連続分布 F に従う．

H_1 ：ある $u_* \in (0, 1)$ が存在して， $i \leq u_*n$ は F_0 に従い， $i > u_*n$ は F_1 に従う．ただし，ある $x \in \mathbb{R}$ に対して $F_0(x) \neq F_1(x)$ ．

• この問題を， F, F_0, F_1 の情報なしに，ノンパラメトリックに考察する．

$\{X_i\}_{i=1,\dots,n}$ の順位統計量を $\{R_i\}_{i=1,\dots,n}$ とし，次を提案する：

$$D_n = \frac{1}{\sqrt{n}} \max_{1 \leq i, j \leq n} \left| \sum_{q=1}^i 1\{R_q \leq j\} - \frac{ij}{n} \right|.$$

- H_0 のもとで ,

$$D_n \rightarrow^d \sup_{0 \leq s, t \leq 1} |B^\circ(s, t)|,$$

ただし B° は平均ゼロの正規確率場であって共分散構造が

$$EB^\circ(s_1, t_1)B^\circ(s_2, t_2) = (s_1 \wedge s_2 - s_1s_2)(t_1 \wedge t_2 - t_1t_2)$$

であるもの (standard Brownian pillow と呼ばれています) .

- H_1 のもとで ,

$$D_n \geq \sqrt{n} \left\{ u_*(1 - u_*) \sup_x |F_0(x) - F_1(x)| - o(1) \right\} - O_P(1).$$

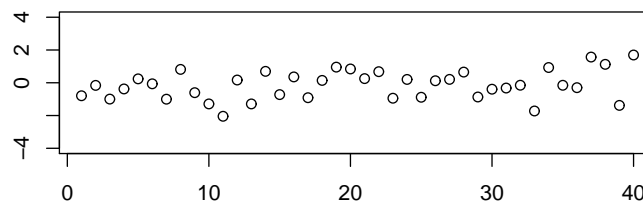
重要な注意 : 我々の統計量は , H_0 のもとで , 漸近的分布不変であるだけでなく , n を止めても分布不変である .

統計量の導出：

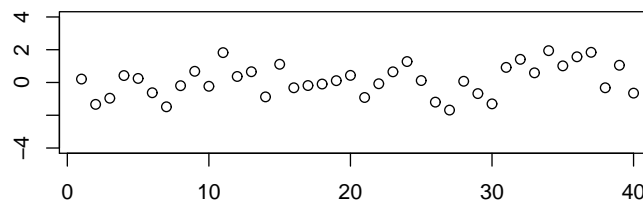
$$\begin{aligned}
 & \sup_{u \in [0,1]} \sup_{x \in \mathbb{R}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (1\{i \leq un\} - u) 1\{X_i \leq x\} \right| \\
 &= \max_{1 \leq i \leq n} \sup_{x \in \mathbb{R}} \frac{1}{\sqrt{n}} \left| \sum_{q=1}^i 1\{X_q \leq x\} - \frac{i}{n} \sum_{q=1}^n 1\{X_q \leq x\} \right| + O\left(\frac{1}{\sqrt{n}}\right) \\
 &= \max_{1 \leq i, j \leq n} \frac{1}{\sqrt{n}} \left| \sum_{q=1}^i 1\{X_q \leq X_{(j)}\} - \frac{i}{n} \sum_{q=1}^n 1\{X_q \leq X_{(j)}\} \right| + O\left(\frac{1}{\sqrt{n}}\right) \\
 &= \max_{1 \leq i, j \leq n} \frac{1}{\sqrt{n}} \left| \sum_{q=1}^i 1\{R_q \leq j\} - \frac{ij}{n} \right| + O\left(\frac{1}{\sqrt{n}}\right).
 \end{aligned}$$

シミュレーション：（図は $n = 40$ ）

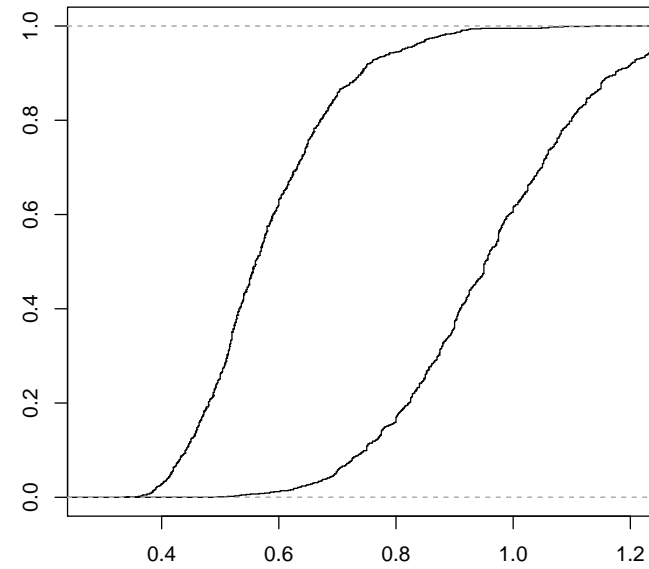
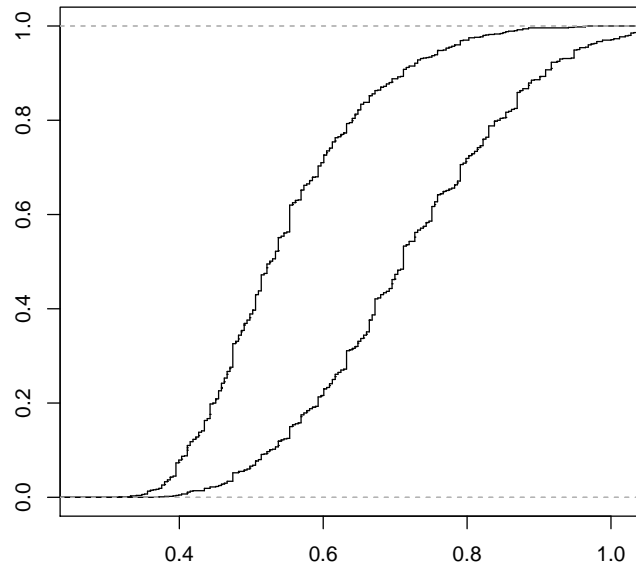
H0 (no change point)



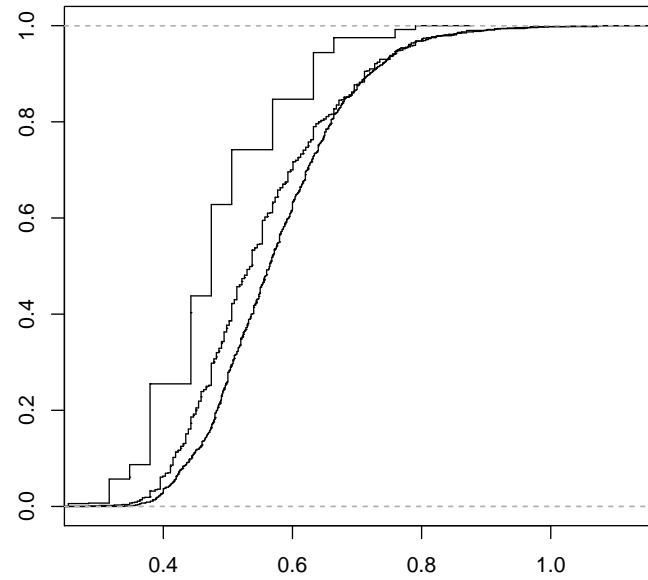
H1 (a change point)



H_0 は , $X_1, \dots, X_n \sim N(0, 1)$,
 H_1 は , $X_1, \dots, X_{[3n/4]} \sim N(0, 1)$, $X_{[3n/4]+1}, \dots, X_n \sim N(1, 1)$.



$n = 40$ (左) と $n = 100$ (右) について,
 H_0 と H_1 のもとでの実験結果を並べて描いた.



H_0 のもとでの 10,000 回のシミュレーションを
 $n = 10, 40, 100$ について行った .

2. 丸められたデータの問題

$(\mathcal{X}, \mathcal{A}, P)$ は確率空間であるとする．各 $n \in \mathbb{N}$ に対し，可算な \mathcal{A} -可測非交和分割，すなわち $\mathcal{X} = \bigcup_k A_k^n$ が与えられたとする．各集合より，点 $a_k^n \in A_k^n$ が選ばれているものとする．

- 発生した本来のデータ $\{Z_i\}_{i=1, \dots, n}$ が

$$Z_i^{n*} = a_k^n, \quad Z_i \in A_k^n \text{ のとき}$$

という具合に，丸められて観測されている状態を考える．

- メッシュおよび丸められた真の分布を

$$\delta_n = \sup_k P(A_k^n),$$
$$P^{n*}(dz) = \sum_k P(A_k^n) \delta_{a_k^n}(dz)$$

によって定義する．

2.1. 丸められたデータによる適合度検定

$$D_n = \sup_{x \in \mathbb{R}} n^{1/2} |\widehat{F}_n(x) - F_0(x)|$$

$$D_n^* = \sup_{x \in \mathbb{R}} n^{1/2} |\widehat{F}_n^*(x) - F_0(x)|$$

$$D_n^{**} = \sup_{x \in \mathbb{R}} n^{1/2} |\widehat{F}_n^*(x) - F_0^{n*}(x)|$$

定理 . 分布関数 $x \mapsto F_0(x)$ は連続とする .

- (i) もし $n\delta_n^2 \rightarrow 0$ ならば D_n^* の収束が成立する .
- (ii) もし $\delta_n \rightarrow 0$ ならば D_n^{**} の収束が成立する .

では一番基本的な例に戻ろう。

- F_0 は $[0, 1]$ 上の一様分布とする。
- $\{Z_1, \dots, Z_n\}$: i.i.d. with F_0 .
- $\delta_n = 0.01$. $\{Z_1^{n*}, \dots, Z_n^{n*}\}$ は δ_n まで丸められたデータとする。

$$Z_1 = 0.67774205 \quad Z_1^{n*} = 0.68$$

$$Z_2 = 0.81124449 \quad Z_2^{n*} = 0.81$$

...

$$Z_n = 0.61694806 \quad Z_n^{n*} = 0.62$$

- \hat{F}_n は $\{Z_1, \dots, Z_n\}$ の経験分布関数とする。
- \hat{F}_n^* は $\{Z_1^{n*}, \dots, Z_n^{n*}\}$ の経験分布関数とする。

- 連続観測の場合 , Kolmogorov-Smirnov 統計量

$$D_n = \sup_{x \in [0,1]} n^{1/2} |\hat{F}_n(x) - F_0(x)|$$

は

$$\sup_{u \in [0,1]} |B^\circ(u)|$$

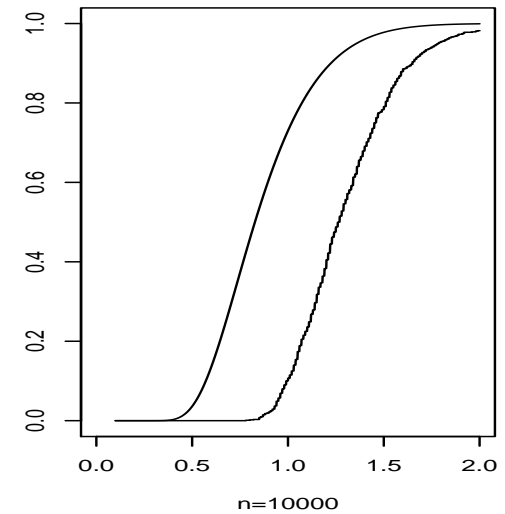
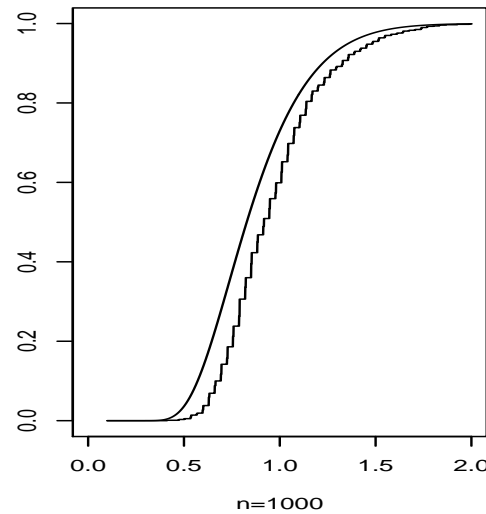
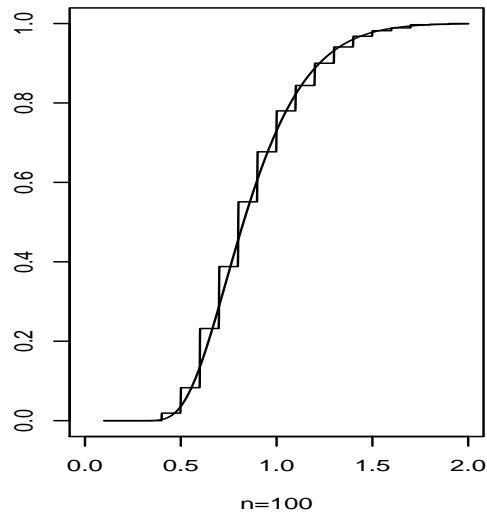
に分布収束する .

- 我々の結果によると , もし $n\delta_n^2 \rightarrow 0$ ならば検定統計量

$$D_n^* = \sup_{x \in [0,1]} n^{1/2} |\hat{F}_n^*(x) - F_0(x)|$$

も同じ極限に分布収束する .

D_n^* のシミュレーションを $n = 100, 1000, 10000$ および $\delta_n = 0.01$ で行った結果が次の図である。

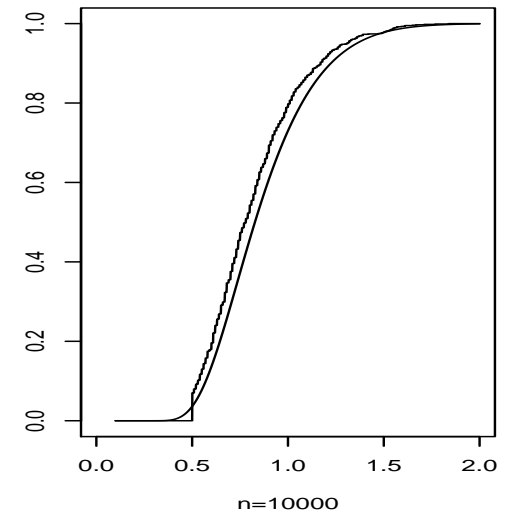
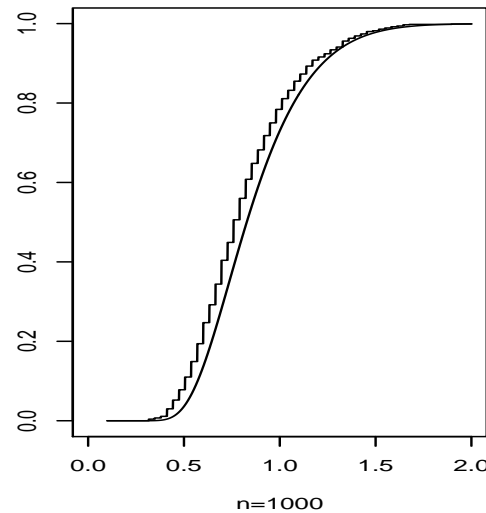
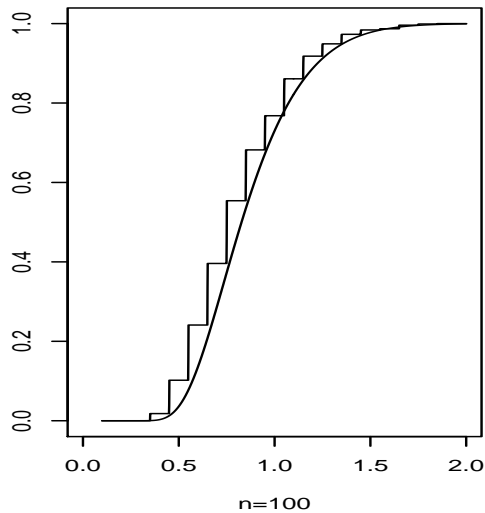


n が大きいとき $\sup_{u \in [0,1]} |B_u^\circ|$ の分布 (実線) への収束は悪い。これは、**仮定 $n\delta_n^2 \rightarrow 0$ が成り立っていない**からである。

そこで検定統計量

$$D_n^{**} = \sup_{x \in \mathbb{R}} n^{1/2} |\hat{F}_n^*(x) - F_0^{n*}(x)|$$

を用いる．これは， $\delta_n \rightarrow 0$ という緩い条件のもとで連続観測の場合と同じ極限に分布収束する．



古典的な方法によると：この事実は，確率過程

$$Y_n(x) = n^{1/2}(\widehat{F}_n^*(x) - F_0^{n*}(x))$$

の弱収束を示すことによって得られる．ただしそれは古典的な緊密性判定条件

$$E|Y_n(x) - Y_n(x')|^p \leq K|x - x'|^{1+r} \quad \text{for some } p, K, r > 0$$

では証明できない．なぜなら F_0^{n*} は離散分布だからである．

Billingsley (1968, page 133) は，この判定条件を，ある右連続関数 H に対する

$$\begin{aligned} P(|Y_n(x) - Y_n(x_1)| \geq \lambda, |Y_n(x_2) - Y_n(x)| \geq \lambda) \\ \leq \frac{1}{\lambda^{2\gamma}} [H(x) - H(x_1)]^\alpha [H(x_2) - H(x)]^\alpha \end{aligned}$$

に置き換えているが，関数 H は全ての n に共通にとらなければならない．

現代的な方法によると：関数によって添え字づけられた経験過程

$$Y_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ f(Z_i) - \int f(z) P_0^{n*}(dz) \right\}, \quad f \in \mathcal{F} \subset L_2(P_0)$$

に対するブラケティング・エントロピー条件

$$\int_0^1 \sup_{n \in \mathbb{N}} \sqrt{\log N_{[\]}(\varepsilon, \Psi, L_2(P_0^{n*}))} d\varepsilon < \infty$$

でよい．幸いにしてブラケティング数 $N_{[\]}(\varepsilon, \mathcal{F}, L_2(Q))$ は Q に依存せずに押さえられる．実際， $\Psi = \{1_{(-\infty, x]} : x \in \mathbb{R}\}$ のとき，ある定数 $K > 0$ が存在して **任意の** 確率測度 Q に対し

$$\log N_{[\]}(\varepsilon, \mathcal{F}, L_2(Q)) \leq K \left(\frac{1}{\varepsilon} \right)$$

が成り立つ（例えば van der Vaart and Wellner (1996) を見よ．）

2.2. 丸められたデータによる Z -推定

丸められたデータに対する経験測度を次のように定義する .

$$\mathbb{P}_n^*(dz) = \frac{1}{n} \sum_{i=1}^n \delta_{Z_i^{n*}}(dx) = \frac{1}{n} \sum_{i=1}^n \sum_k 1_{\{Z_i \in A_k^n\}} \delta_{a_k^n}(dx).$$

有界な $\Theta \subset \mathbb{R}^d$ に対し, $\{\psi_\theta(\cdot); \theta \in \Theta\}$ が与えられたとし, 推定方程式

$$\mathbb{P}_n^* \psi_\theta = 0$$

の解 $\hat{\theta}_n$ の挙動を調べよう .

定理．適当な条件のもとで，もしも $n^{1/2}P^{n*}\psi_{\theta_0} = O(1)$ ならば

$$\begin{aligned} & n^{1/2}(\hat{\theta}_n - \theta_0 - (P\dot{\psi}_{\theta_0})^{-1}P^{n*}\psi_{\theta_0}) \\ & \rightarrow^d N(\mathbf{0}, (P\dot{\psi}_{\theta_0})^{-1}P\psi_{\theta_0}\psi_{\theta_0}^\top)((P\dot{\psi}_{\theta_0})^{-1})^\top). \end{aligned}$$

• 従って，通常の理論と同様の結論を得るためには

$$n^{1/2}P^{n*}\psi_{\theta_0} \rightarrow 0$$

が成り立てばよい．

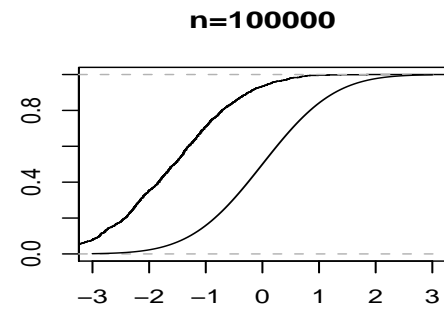
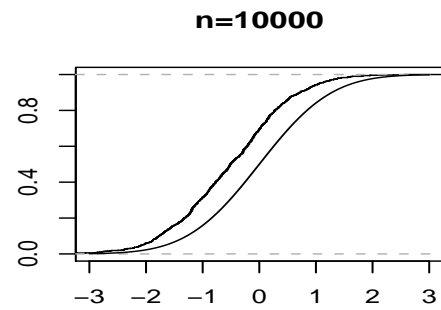
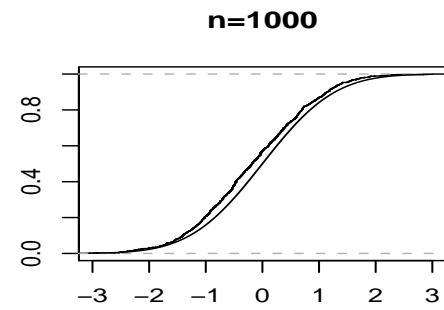
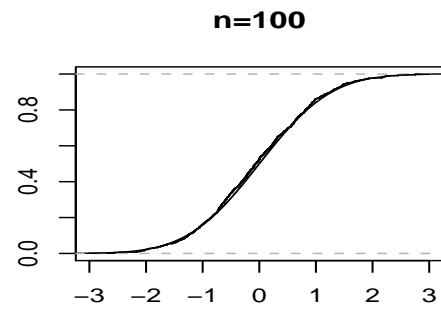
• この条件は，“ $n\delta_n^2 \rightarrow 0$ ” ならば概ね満たされる．この結果は，常識的な直観に合致する．

• しかし，幸いなことに，**中点サンプリングスキーム（四捨五入）**のもとでは“ $n\delta_n^4 \rightarrow 0$ ” ならば満たされる．この結果は，私には意外性のある嬉しい誤算だった．

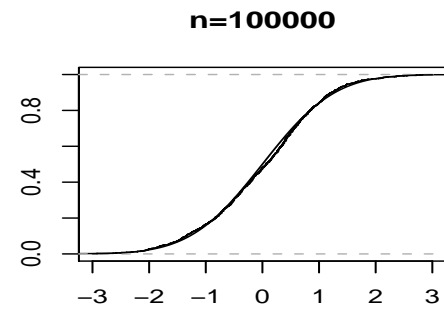
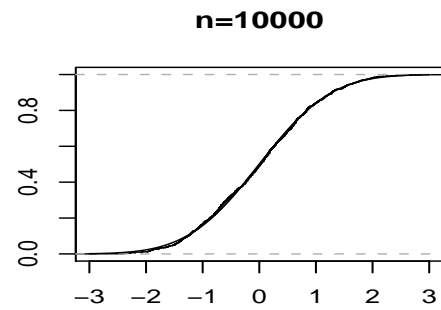
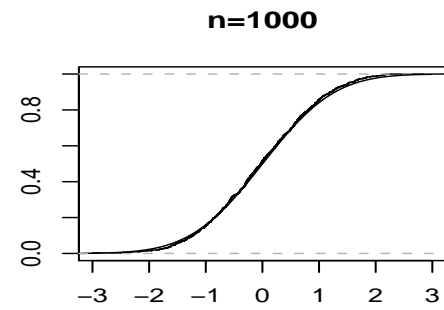
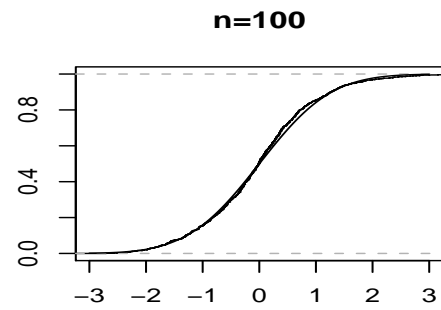
シミュレーション 1 : 正規分布 $N(\theta, 1)$ における平均の推定 .

$\delta_n = 0.01$ は固定し , $n = 100, 1000, 10000, 100000$ について実験を行った .

	$n = 100$	$n = 1,000$	$n = 10,000$	$n = 100,000$
$n\delta_n^2$	0.01	0.1	1	10
$n\delta_n^4$	0.000001	0.00001	0.0001	0.001



切り捨ての場合, n が大きいとき良くない.



四捨五入の場合 .

シミュレーション 2 : t -検定 .

正規分布 $N(\theta_0, \sigma^2)$ からの連続観測 , および , $\delta_n = 0.01$ に対応する四捨五入観測 , 切り捨て観測が得られた場合に , t -統計量

$$T_n = \frac{\sqrt{n}(\bar{Z}_n - \theta_0)}{\hat{\sigma}_n},$$

ただし

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i, \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n |Z_i - \bar{Z}_n|^2,$$

が帰無仮説のもとでどう挙動するかを調べるために , $m = 100,000$ 回の実験を行った . シミュレーションの真値は $(\theta_0, \sigma^2) = (0, 1)$ とした . 以下の表では , 比 $\#\{T_n > t_{n-1, \alpha}\} / m$ を $n = 11, 41, 121, 241$ と $\alpha = 0.050, 0.025$ について記してある .

	連続観測	四捨五入	切り捨て
$n = 11$	0.0505	0.0498	0.0494
$n = 41$	0.0500	0.0494	0.0478
$n = 121$	0.0504	0.0493	0.0459
$n = 241$	0.0509	0.0513	0.0420

$$\alpha=0.050$$

切り捨ては n が大きいとき良くない。

	連続観測	四捨五入	切り捨て
$n = 11$	0.0251	0.0254	0.0241
$n = 41$	0.0254	0.0258	0.0237
$n = 121$	0.0243	0.0247	0.0213
$n = 241$	0.0259	0.0246	0.0205

$$\alpha=0.025$$

切り捨ては n が大きいとき良くない。

References

- [1] Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley, New York.
- [2] Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, New York.
- [3] Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **23** 277-281.
- [4] Doob, J.L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20** 303-403.
- [5] Horváth, L. and Parzen, E. (1994). Limit theorems for Fisher-score change processes. *In: Change-point Problems, (Edited by Carlstein, E., Müller H.-G. and Siegmund, D.)* IMS Lecture Notes – Monograph Series **23** 157-169.
- [6] Lee, S., Nishiyama, Y. and Yoshida, N. (2006). Test for parameter change in diffusion processes by cusum statistics based on one-step estimators. *Ann. Inst. Statist. Math.* **58** 211-222.
- [7] Nishiyama, Y. (2000a). Weak convergence of some classes of martingales with jumps. *Ann. Probab.* **28** 685-712.
- [8] Nishiyama, Y. (2000b). *Entropy Methods for Martingales*. CWI Tract **128** Centrum voor Wiskunde en Informatica, Amsterdam.

- [9] Nishiyama, Y. (2008). Donsker's theorem for discretized data. *J. Japan Statist. Soc.* **38** 505-515.
- [10] Nishiyama, Y. (2009). Two sample problem for rounded data. *J. Japan Statist. Soc.* **39** 233-238.
- [11] Nishiyama, Y. (2011). On Z -estimation by rounded data. *J. Statist. Plann. Inference* **141** 287-292.
- [12] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.