統計科学・可視化・マイニング

統計数理研究所　データ科学研究系
田村義保
第3回データ科学ワークショップ（科学情報学研究会）
2010年11月25日
筑波大学計算科学研究センター

目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究例
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

古い計算機の話し(1)

統計数理44巻(1996)の特集「計算統計学の発展」

『手回し計算機、リレー式計算機の時代から計算機は統計科学の発展に多大な影響を与えて来た。岩崎も指摘しているように(1992)、Karl Pearsonの伝記には卓上計算器の置かれた机に座って仕事をしているPearsonの写真が収められている。また、1956年に統計数理研究所にFACOM128というリレー式万能計算機が導入されているが、1957年に発行された統計数理研究所彙報第5巻には、この計算機を用いて計算を行った論文が6編掲載されている。』

古い計算機の話し(2)

コンピュータが計算機と呼ばれた時代
（株式会社アスキー2005年12月）

『日本初の商用計算機として作られたのがFACOM128である。・・・記憶容量は180語、演算速度も、加減算で0.15秒、乗算は0.15～0.4秒、・・・。FUJICと同じ1956年に完成し、1号機は文部省統計数理研究所に、2号機は・・・納入された。』

古い計算機の話し（3）

年譜 1　統計数理研究所メインフレーム・スパコンの変遷

1954	FACOM415A（相関係数専用機）
1956	FACOM128（リレー計算機）
	物理乱数発生機（放射能）
1963	HIPAC103（名称 TSKIII、パラメトロン計算機、ALS-1000 とのハイブリッド）
	物理乱数発生機（ダイオード熱雑音、12,000 ビット／秒）
1971	HITAC8500 改良型（主記憶 524KB）／HITAC8300（主記憶 64KB）
	物理乱数発生機（ダイオード熱雑音、200KB／秒）
1972	HITAC8700（主記憶 524KB）／HITAC8400（主記憶 131KB）
1978	HITAC M180（主記憶 2MB）
1981	HITAC M200H（主記憶 10MB）
1985	HITAC M280H（主記憶 24MB）
1989	HITAC M682H（主記憶 256MB、内蔵アレイプロセッサ）
	HITAC M660D（主記憶 64MB）
	物理乱数発生機（ダイオード熱雑音、1.5MB／秒）
1993	HITAC M680H（主記憶 256MB）
	HITAC M680H（主記憶 256MB、内蔵ベクトル演算機構）
1994	HITAC M880/180（主記憶 512MB）
	HITAC S3600/120（主記憶 256MB、拡張記憶 2GB）
1996	HITAC S3800/162（主記憶 1GB、拡張記憶 4GB、4GFlops）
1996	IBM SP2（48 ノード、主記憶 12GB）
1999	HITAC SR8000（20 ノード 主記憶 8GB/ノード 8Gflops/ノード）
	物理乱数発生装置（日立製：ダイオード熱雑音、25MB/秒、6 枚)
	HP9000V2250（16CPU、主記憶 12GB、PA-RISC8500(240MHz)）

古い計算機の話し（4）

年譜 1　統計数理研究所メインフレーム・スパコンの変遷
(続き)

2000	Origin2000(64CPU、主記憶 48GB、R12000(300MHz)) （400MHz にクロックアップ 2000 年 8 月） 統計計算開発環境支援システムを 16CPU にして計 80CPU で運用
2004	SGI Altix3700（主記憶 1920GB ピーク 5.2Gflops×256） 物理乱数発生装置（東芝製：ダイオード熱雑音、133MB/秒、6 枚) NEC SX6(主記憶 128GB ピーク 9GFlops×12） Hitachi SR11000（主記憶 128GB ピーク 6.8GFlpos×64）
2006	HP XC4000 (計算ノード　ProLiant DL 145G2 128 ノード (OPteron 2.6GHz × 2 主記憶 640GB） SGI Prism（Itanium2 1.5GHz ×16、主記憶 32GB、グラフィックカード 8 枚） 物理乱数サーバー　8 ノード （ノード構成：Opteron 2.4GHz 2CPU、主記憶 4GB、物理乱数ボード 4 枚）
2010	Fujitsu SPARC Enterprise M9000(2 システム　64CPU(256 コア)、主記憶 2TB、 ピーク 2949GFlops と　24CPU(96 コア),主記憶 1TB、ピーク 1105GFlops) Fujitsu PRIMERGY RX200S5(360 ノード（2880 コア)、主記憶　48GB×160+24GB ×200、ピーク　93.76GFlops×360)

古い計算機の話し（5）

年譜 2　統計数理研究所特殊用途計算機器の変遷

1971	特殊入出力装置（HITAC10、HITAC8500 と接続しアナログ量の入出力用）
1974	ハイブリッド計算機（アナログ部分 S-300、デジタル部分 HITAC10）
1975	遠隔入出力制御装置（YHP2000MX、リモートステーション用）
1980	視聴覚的情報検索システム（AICOM-C6、大量時系列データ解析用）
1982	領域クラスタリングシステム（MINC11/23）
1986	アナログシミュレータ（EAI1000）
1987	ワークステーション（DN570）
	カラーグラフィックス・インテリジェントシステム化拡張装置（DN3000）
1988	ワークステーション（DN4000、SUN3/50）
	大量時系列データ情報検索システム（SUN3/260）
	統計エキスパートシステム（EXPLORER2）
1989	超高度物理乱数発生装置
	情報管理用システム（IBM9370/20）
1990	高速演算サーバーシステム（SUN4/370）
	情報網管理用システム（POWER STATION 520）
	（IBM 9370/20 を 9370/30 にバージョンアップ）
1991	パラレル演算装置（TITAN3000(4CPU)）
1992	時空現象可視化装置（TITAN3000V(2CPU)）
	探索的データ解析支援システム（SparcStation2）
	情報網拡張システム（CISCO ルータ）
1993	統計メタ情報ナビゲーション・システム（SparcStation2）
	情報網拡張システム（S4/10）
	大容量光磁気記憶システム（EPOCH2）
1994	統計データサーバーシステム（TOSHIBA UX3000）
	時系列解析支援システム（IBM RS/6000-3BT）
1995	統計情報資源統合化ハイパーネットワーク・システム
	（SparcStation5）
1996	統計情報資源サーバーシステム（SparcStation20）
1997	統計動画像システム（Onyx 2CPU R10000）
1998	統計計算開発環境支援システム（Origin2000 4CPU R10000）
1999	統計的データ解析支援システム
	（Enterprise450 UltraSparc 400MHz 4CPU 主記憶 2GB）

古い計算機の話し（6）

年譜 2　統計数理研究所特殊用途計算機器の変遷
（続き）

2001	ブートストラップシステム（2006 年 1 月　32 台へ縮小）
	SGI1200（PentiumIII（800MHz）、512MB メモリ、物理乱数ボード）100 台
	Origin3400（R12000（400MHz）プロセッサ x 8、8GB メモリ
	SGI1450（PentiumIII Xeon（700MHz） x 4、2GB メモリ）
	磁気ディスク（RAID5 145.6GB）
2001	高精度高速グラフィックシステム
	Onyx3200（R12000（400MHz）プロセッサ x 4、2GB メモリ
	TP9100 ストレージシステム（73GB ハードディスク x 24）
	DMS-B9 オートローダ（DTF 2 ドライブ）
2002	Soralis クラスターシステム
	富士通 PrimePower1 4 台
2010	物理乱数発生システム(2010 年 3 月)
	4.2GHz IBM Power6 x1socket / 8GB(4GB x2)
	146.8GB SAS ドライブ x2 / 1000BASE-T x2
	日立製作所製 物理乱数発生器(200MB/s)
2010	物理乱数サーバーシステム(2010 年 7 月)
	Altix UV100　2 セット
	(6 コア　Intel Nehalem-EX 2.66GHz　x 16
	Numalink5，主記憶 512GB，システムディスク 300GB SAS x 2)
	SGI Infinite Storage220　1 セット
	(2RAID コントローラ，1TB SATA x 16)
	ランダムマスター　RMHe　2 セット
	(PCI-Express　4x　400MB/s，PowerEdge R710（X5560 2.8GHz　x　2)
2010	計算統計学支援システム(2010 年 7 月)
	CRAY XT6m
	(12 コア AMD Opteron 88CPU(44Nodes)，主記憶 1.375TB)
	磁気ディスクアレイ　LSI4998 Storage System

古い計算機の話し（7）

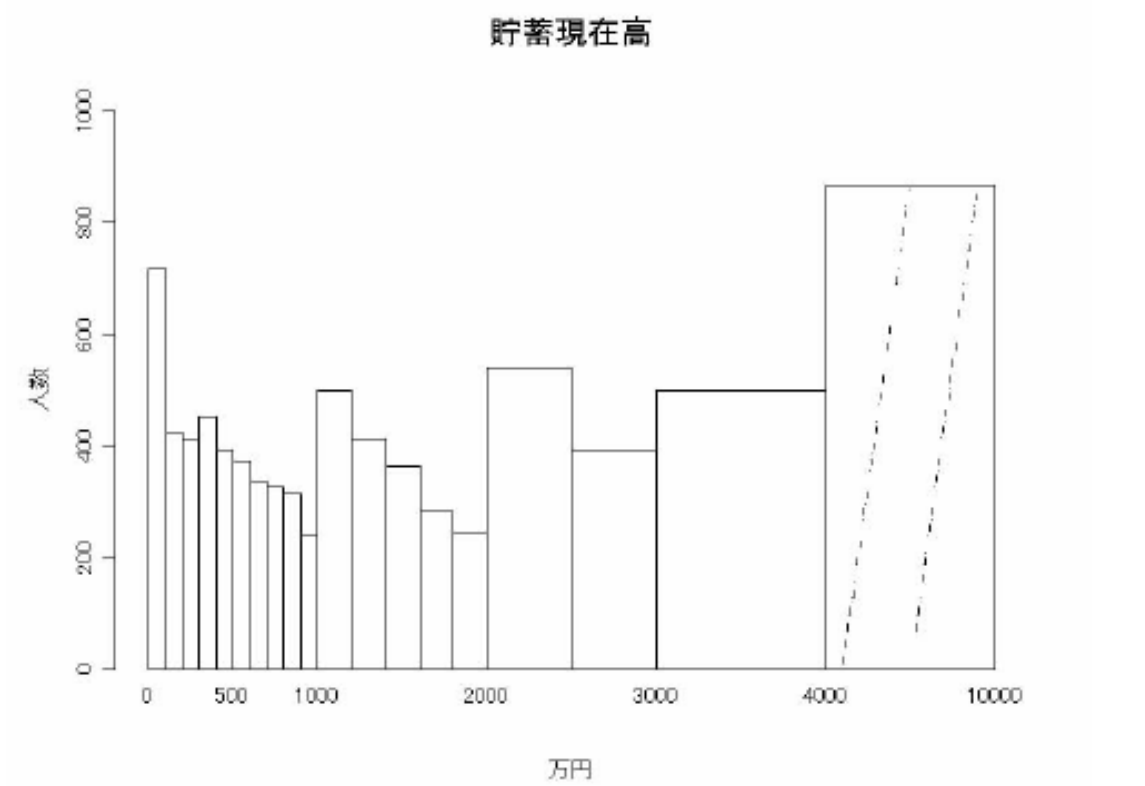
年譜 3　統計数理研究所ネットワーク環境の変遷

1973	大型機 TSS サービス開始
	3 台：1973～1979
	8 台：1979～1981
	30 台以上：1981～（34 回線まで同時使用可能）
	30 台以上：1989～（72 回線まで同時使用可能）
1985	大型機直結端末サービス
	（2020 を使用、1989 年からは 2020E 及び 2020E/32）
1988	イーサネットによる所内 LAN（リピータによる接続）
	JUNET と接続（慶應義塾大学理工学部との公衆回線による接続、1992 年まで）
1989	N1 との接続（東京工業大学と 9600BPS 専用線による接続）
	BITNET との接続（東京理科大学と 9600BPS 専用線による接続、1994 年 3 月まで）
1990	TISN との接続（東京大学理学部と 64KB 専用線による接続、1996 年 3 月まで）
1992	所内 LAN 拡張（ルータを使用し多セグメント化）
1994	統計科学計算機システムの端末のイーサネット接続
1995	SINET との接続（512KB 専用線による接続、1995 年 8 月 17 日より）
	N1 も同回線による接続に変更
1996	計算統計学支援システムの各システム間の FDDI 接続
1997	一部のワークステーションをスイッチで 100Mbps 接続
1998	幹線を 1Gbps（光、レイヤー3 スイッチ 4 台）
	支線を 100Mbps にし、研究室に情報コンセント
	(1998 年 12 月に工事、1999 年 1 月から稼働)
1999	SINET との接続が 1.5Mbps(1999 年 10 月 27 日)
2002	SINET との接続を 100Mbps(2002 年 7 月)
2003	Super　SINET との接続 2.4Gbps(2003 年 10 月 1 日)
2007	SINET3 へ変更 2.4Gbps(2007 年 4 月 19 日)
2009	幹線を 10Gbps、支線を 1Gbps

目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究例
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

階級（万円）			度数	累積相対度数 (%)
0	～	100	717	8.85
100	～	200	425	14.10
200	～	300	410	19.16
300	～	400	453	24.75
400	～	500	394	29.61
500	～	600	374	34.23
600	～	700	336	38.37
700	～	800	327	42.41
800	～	900	315	46.30
900	～	1,000	242	49.28
1,000	～	1,200	500	55.46
1,200	～	1,400	412	60.54
1,400	～	1,600	365	65.05
1,600	～	1,800	284	68.55
1,800	～	2,000	246	71.59
2,000	～	2,500	541	78.26
2,500	～	3,000	394	83.13
3,000	～	4,000	500	89.30
4,000	～	10,000	867	100.00


貯蓄現在高
人数
0
200
400
600
800
1000
0
500
1000
2000
3000
4000
10000
万円

ヒストグラムはその面積と度数（あるいは相対度数）を比例するように描く必要がある．したがって，階級数を k として，各階級の度数を f_j, $j=1,\cdots,k$ ，階級の幅を Δ_j $j=1,\cdots,k$ ，データ数を n とする時，ヒストグラムは

$$\frac{f_j}{n\Delta_j} \quad \text{あるいは} \quad \frac{f_j}{\Delta_j}$$

が高さになるように描く必要がある．

目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究例
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

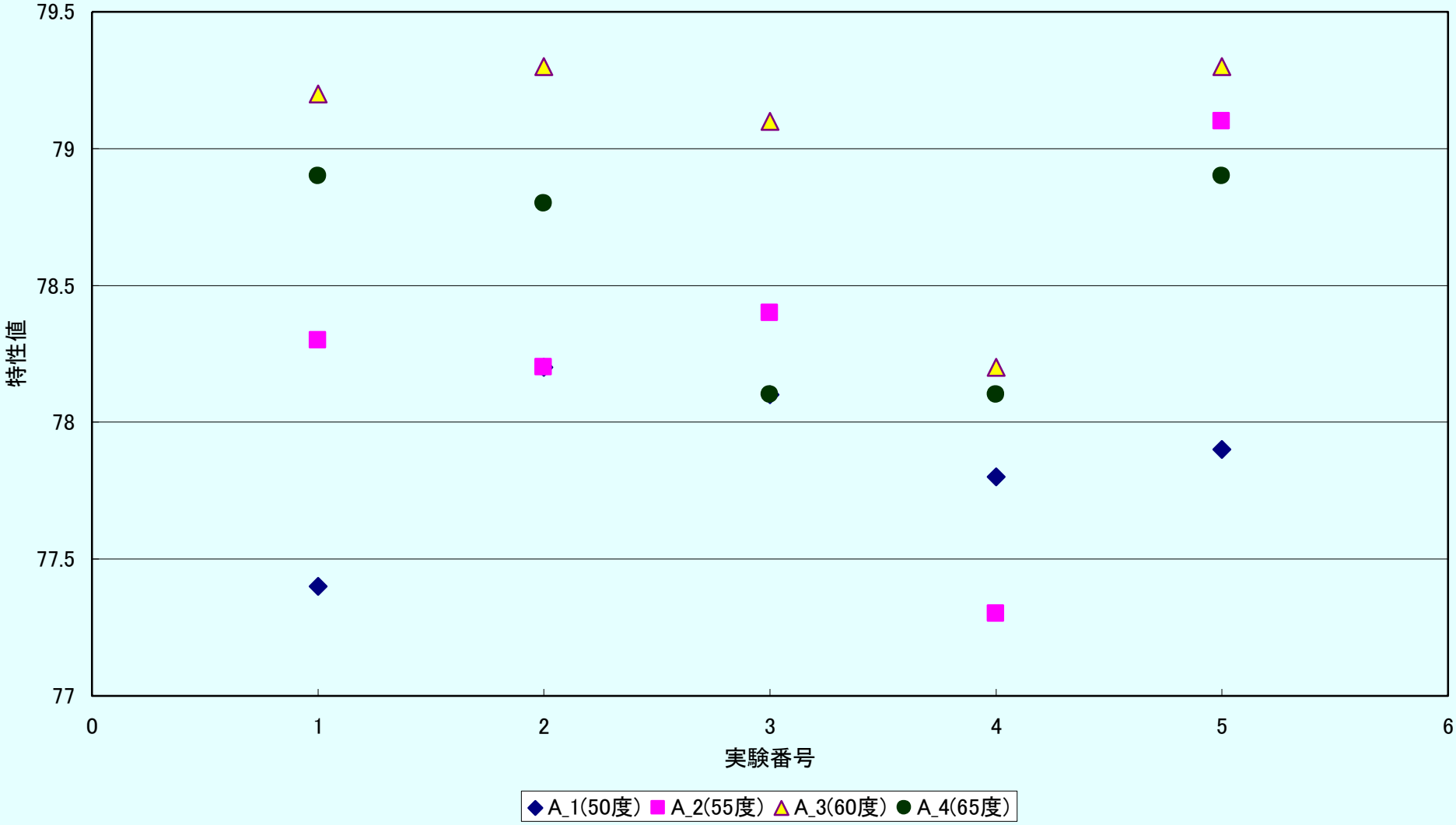
仮説検証型からモデル選択型へ(1)
分散分析の場合

説明に用いるデータ

1元配置のデータの例（東京大学出版会「自然科学の統計学」

反応温度	A_1(50度)	A_2(55度)	A_3(60度)	A_4(65度)		
1	77.4	78.3	79.2	78.9		
2	78.2	78.2	79.3	78.8		
3	78.1	78.4	79.1	78.1		
4	77.8	77.3	78.2	78.1		
5	77.9	79.1	79.3	78.9		
計	389.4	391.3	395.1	392.8	総計	1568.9
平均	77.88	78.26	79.02	78.56	総平均	78.43

1元配置データ



一元配置分散分析について

Level	A_1	A_2	$\cdots$	A_j	$\cdots$	A_k
1	x_{11}	x_{21}	$\cdots$	x_{j1}	$\cdots$	x_{k1}
2	x_{12}	x_{22}	$\cdots$	x_{j2}	$\cdots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		x_{jn_j}		$\vdots$
$\vdots$	x_{1n_1}	$\vdots$				$\vdots$
		x_{2n_2}				$\vdots$
						x_{kn_k}
Total	T_1	T_2	$\cdots$	T_i	$\cdots$	T_k
Mean	$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_j$	$\cdots$	$\bar{x}_k$

モデル

$$x_{ji} = \mu + \alpha_j + \varepsilon_{ji}$$

平均

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ji}}{n_j}$$

$$\bar{x} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ji}}{N} \qquad (N = \sum_{j=1}^{k} n_j)$$

変動の分割

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\bar{x})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)^2+\sum_{j=1}^{k}n_j(\bar{x}_j-\bar{x})^2$$

誤差変動の分散の推定値

$$S_E^2=\frac{SS_E}{N-k}=\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\bar{x}_j)^2}{N-k}$$

要因による変動の分散の推定値

$$S_A^2=\frac{SS_A}{k-1}=\frac{\sum_{j=1}^{k}n_j(\bar{x}_j-\bar{x})^2}{k-1}$$

検定統計量

$$F = \frac{S_A^2}{S_E^2}$$

分散分析表

Source	SS	df	variance	F
treatment	$SS_A = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$	$k-1$	$S_A^2 = \frac{SS_A}{k-1}$	S_A^2/S_E^2
error	$SS_E = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2$	$N-k$	$S_E^2 = \frac{SS_E}{N-k}$	
total	$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - \bar{x})^2$	$N-1$		

この例の分散分析表

source	SS	df	variance	F
treatment	3.482	3	1.160667	5.13002
error	3.62	16	0.22625	
total	7.102	19		

P値

0.01123806

モデル選択による方法

モデル1:要因ごとに平均と分散が異なる

$$
\begin{aligned}
l(\mu_1,\mu_2,\mu_3,\mu_4,\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2) \;=\; & -\frac{n_1+n_2+n_3+n_4}{2}\log 2\pi \\
& -\frac{n_1}{2}\log\sigma_1^2-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2 \\
& -\frac{n_2}{2}\log\sigma_2^2-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n_2}(x_{2i}-\mu_2)^2 \\
& -\frac{n_3}{2}\log\sigma_3^2-\frac{1}{2\sigma_3^2}\sum_{i=1}^{n_3}(x_{3i}-\mu_3)^2 \\
& -\frac{n_4}{2}\log\sigma_4^2-\frac{1}{2\sigma_4^2}\sum_{i=1}^{n_4}(x_{4i}-\mu_4)^2
\end{aligned}
$$

モデル選択による方法(続き)

モデル2:要因ごとに平均が異なる

$$
\begin{aligned}
l(\mu_1,\mu_2,\mu_3,\mu_4,\sigma^2) \;=\; & -\frac{n_1+n_2+n_3+n_4}{2}\log 2\pi \\
& -\frac{n_1}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2 \\
& -\frac{n_2}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(x_{2i}-\mu_2)^2 \\
& -\frac{n_3}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_3}(x_{3i}-\mu_3)^2 \\
& -\frac{n_4}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_4}(x_{4i}-\mu_4)^2
\end{aligned}
$$

モデル選択による方法(続き)

モデル3:平均も分散も同じ

$$
\begin{aligned}
l(\mu_1,\mu_2,\mu_3,\mu_4,\sigma^2) \;=\; & -\frac{n_1+n_2+n_3+n_4}{2}\log 2\pi \\
& -\frac{n_1}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_{1i}-\mu)^2 \\
& -\frac{n_2}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(x_{2i}-\mu)^2 \\
& -\frac{n_3}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_3}(x_{3i}-\mu)^2 \\
& -\frac{n_4}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n_4}(x_{4i}-\mu)^2
\end{aligned}
$$

モデル選択による方法(続き)

モデル1:推定値とAIC

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n_j} x_{ji}}{n_j}, \qquad \hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ji} - \hat{\mu}_j)^2}{n_j}, \qquad j = 1, \cdots, k$$

$$\begin{aligned} \mathrm{AIC}(\hat{\mu}_1, \cdots, \hat{\mu}_k, \hat{\sigma}_1^2, \cdots, \hat{\sigma}_k^2) &= \sum_{j=1}^{k} n_j (\log 2\pi + 1) \\ &+ \sum_{j=1}^{k} n_j \log \hat{\sigma}_j^2 \\ &+ 2 \times (2 \times k) \end{aligned}$$

モデル選択による方法(続き)

モデル2:推定値とAIC

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n_j} x_{ji}}{n_j}, \qquad \hat{\sigma}^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\hat{\mu}_j)^2}{\sum_{j=1}^{k} n_j}, \qquad j = 1, \cdots, k$$

$$\begin{aligned} \mathrm{AIC}(\hat{\mu}_1, \cdots, \hat{\mu}_k, \hat{\sigma}^2) &= \sum_{j=1}^{k} n_j(\log 2\pi + 1) \\ &+ \sum_{j=1}^{k} n_j \log \hat{\sigma}^2 \\ &+ 2 \times (k+1) \end{aligned}$$

モデル選択による方法(続き)

モデル3:推定値とAIC

$$\hat{\mu} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ji}}{\sum_{j=1}^{k} n_j}, \qquad \hat{\sigma}^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ji}-\hat{\mu})^2}{\sum_{j=1}^{k} n_j}$$

$$\begin{aligned} \mathrm{AIC}(\hat{\mu},\hat{\sigma}^2) &= \sum_{j=1}^{k} n_j(\log 2\pi + 1) \\ &+ \sum_{j=1}^{k} n_j \log \hat{\sigma}^2 \\ &+ 2\times(2) \end{aligned}$$

モデル選択による方法(続き)

- モデル1のAIC　-20.81878
- モデル2のAIC　-24.12999
- モデル3のAIC　-16.70712

AICが最小のモデルを選ぶというのが、情報量規準を使う時の指導原理

（AIC最小化＝カルバック･ライブラ情報量最小化）

結論：モデル2の4つの温度条件で平均は異なるが分散は等しいというモデルを選ぶ。

モデル選択による方法(発展)

例えば、分散をすべての水準で異なっているとはせずに、温度が低い方と高い方で、2つの分散を持つようにするというモデルも考慮する。

さらなる発展

温度の水準が4つあることが分かっているとして解析しているが、その数は分からないが、複数の温度水準で作られているとして、水準数を見つけようにも発展できる。この場合、どのデータがどの水準のものかは分からないで、事後的にどの水準のものかも見つけるようにすることができる。

仮説検証型からモデル選択型へ(2)
カテゴリカルデータ解析の場合

説明に用いるデータ

(国民性の調査の質問)

	1.To male	2.To female	Total
1.Male	749	83	832
2.Female	445	636	1081
Total	1194	719	1913

独立性の検定による分析

期待度数

	1.To male	2.To female	Total
1.Male	519.2933	312.7067	832
2.Female	674.7067	406.2933	1081
Total	1194	719	1913

検定統計量

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

この問題の場合のカイ二乗値

$$\chi^2 = 478.4209$$

検定の結論：この値は、自由度1のカイ二乗分布の上側5%点の値3.84より大きいので、独立性は棄却され男女によって答え方に差があることになります。

AICを使ったモデル選択

$$\sum_{i_1=1}^{c_1}\sum_{i_2=1}^{c_2} p(i_1,i_2)=1$$

$$\sum_{i_1=1}^{c_1}\sum_{i_2=1}^{c_2} n(i_1,i_2)=n$$

多項分布モデル

$$P(\{n(i_1,i_2)\}|\{p(i_1,i_2)\})=\frac{n!}{\prod_{i_1=1}^{c_1}\prod_{i_2=1}^{c_2} n(i_1,n_2)!}\prod_{i_1=1}^{c_1}\prod_{i_2=1}^{c_2} p(i_1,i_2)^{n(i_1,n_2)}$$

対数尤度

$$l(\{p(i_1,i_2)\})=\log\{\frac{n!}{\prod_{i_1=1}^{c_1}\prod_{i_2=1}^{c_2} n(i_1,n_2)!}\}+\sum_{i_1=1}^{c_1}\sum_{i_2=1}^{c_2} n(i_1,n_2)\log p(i_1,i_2)$$

モデル0

$$p(i_1, i_2) = \theta(i_1, \cdot)\theta(\cdot, i_2)$$

制約条件

$$\sum_{i_1=1}^{c_1} \theta(i_1, \cdot) = \sum_{i_2=1}^{c_2} \theta(\cdot, i_2) = 1$$

推定値

$$\widehat{\theta}(i_1, \cdot) = \frac{n(i_1, \cdot)}{n}$$

$$\widehat{\theta}(\cdot, i_2) = \frac{n(\cdot, i_2)}{n}$$

モデル1

$$p(i_1, i_2) = \theta(i_1, i_2)$$

制約条件

$$\sum_{i_1=1}^{c_1} \sum_{i_2=1}^{c_2} \theta(i_1, i_2) = 1$$

推定値

$$\widehat{\theta}(i_1, i_2) = \frac{n(i_1, i_2)}{n}$$

AIC=-2・最大対数尤度＋2・パラメータ数

モデル0のAIC

$$AIC(0) = -2[C_1 + \sum_{i_1=1}^{c_1}\sum_{i_2=1}^{c_2} n(i_1,i_2)\log\frac{n(i_1,\cdot)n(\cdot,i_2)}{n^2}] + 2(c_1+c_2-2)$$

モデル1のAIC

$$AIC(1) = -2[C_1+\sum_{i_1=1}^{c_1}\sum_{i_2=1}^{c_2} n(i_1,i_2)\log\frac{n(i_1,i_2)}{n}]+2(c_1c_2-1)$$

ただし

$$C_1 = \log\frac{n!}{\prod_{i_1=1}^{c_1}\prod_{i_2=1}^{c_2} n(i_1,n_2)!}$$

この例の場合のAIC

$$
\begin{aligned}
AIC(0) &= -2(1194\log 1194 + 749\log 749 + 832\log 832 + 1081\log 1081 \\
&\quad -2\times 1913\log 1913) + 2(2+2-2) \\
&= 5156.27
\end{aligned}
$$

$$
\begin{aligned}
AIC(1) &= -2(749\log 749 + 445\log 445 + 83\log 83 + 636\log 636 \\
&\quad -1913\log 1913) + 2(2\times 2-1) \\
&= 4630.29
\end{aligned}
$$

このデータの場合、AIC(1)の方がAIC(0)より小さくなるので性別と回答の仕方に関係があると結論できます。

説明に用いるデータ2

(国民性の調査の質問)

		1.To male	2.To female	Total
1.Male	1.Under 29	165	22	187
	2.Over 30	584	61	645
2.Female	1.Under 29	113	125	238
	2.Over 30	332	511	843
Total		1194	719	1913

モデル1：性別のみで説明できると考える

$$\begin{aligned}\mathrm{AIC}(I_1;I_2) &= (-2)\sum_{i_1,i_2} n(i_1,i_2)\log\frac{n\cdot n(i_1,i_2)}{n(i_1)n(i_2)} + 2(c_1-1)(c_2-1)\\ &= (-2)\left\{749\log\frac{1913\times749}{1194\times832}+\cdots+636\log\frac{1913\times636}{719\times1081}\right\}+2(2-1)(2-1)\\ &= -526.08\end{aligned}$$

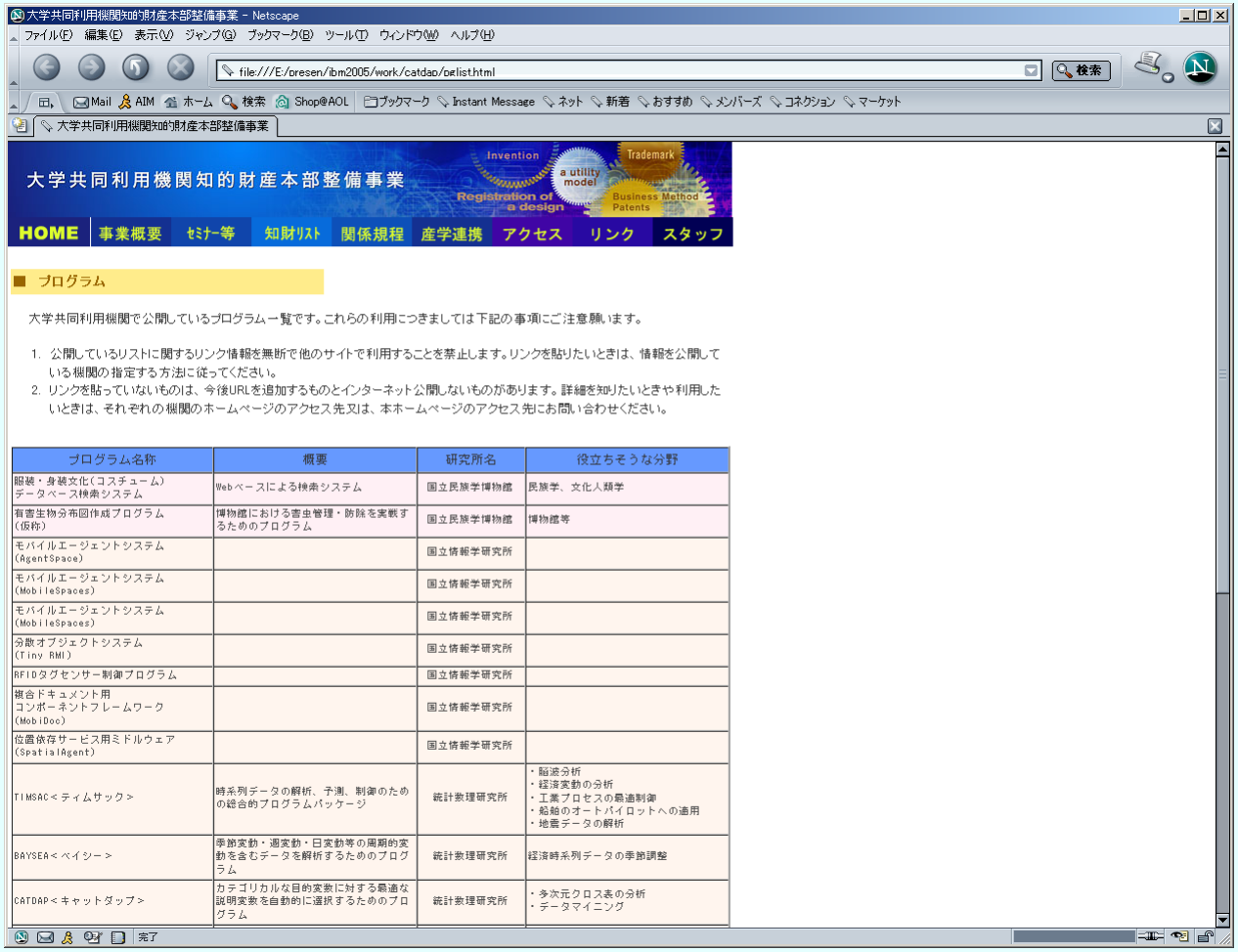
モデル2：性別と年齢で説明できると考える

$$\begin{aligned}\mathrm{AIC}(I_1;I_2,I_3) &= (-2)\sum_{i_1,i_2,i_3} n(i_1,i_2,i_3)\log\frac{n\cdot n(i_1,i_2,i_3)}{n(i_1)n(i_2,i_3)} + 2(c_1-1)(c_2c_3-1)\\ &= (-2)\left\{165\log\frac{1913\times165}{1194\times187}+\cdots+511\log\frac{1913\times511}{719\times843}\right\}+2(2-1)(2\times2-1)\\ &= -527.89\end{aligned}$$

モデル3：年齢のみで説明できると考える

$$\begin{aligned}\mathrm{AIC}(I_1;I_3) &= (-2)\sum_{i_1,i_3} n(i_1,i_3)\log\frac{n\cdot n(i_1,i_3)}{n(i_1)n(i_3)} + 2(c_1-1)(c_3-1)\\ &= (-2)\left\{278\log\frac{1913\times278}{1194\times425}+\cdots+572\log\frac{1913\times572}{719\times14881}\right\}+2(2-1)(2-1)\\ &= -0.11\end{aligned}$$

大学共同利用機関知的財産本部
http://www.chizai.nii.ac.jp/pglist.html



CATDAP
カテゴリカルな目的変数に対する最適な説明変数を自動的に選択するためのプログラム

[fpr 2100] 小講演　データマイニング雑感

http://www.nuis.ac.jp/~mat/fpr/fpr2001/0210.html

豊田@早大心理です.
Keizo Hori さんは書きました:
>堀@香川大学経済学部です。
>豊田さんの小講演データマイニングと認知科学にでました。
>CATDAP02だと全体を一望しようとするけど，aidやTree
>では逐次型。全体一望型では
>処理の爆発的な増大が起こる。変数の数にもよりますが，ま，
>2元（もとの分割を入れて3
>元）くらいだとやってやれなくないのではとも思うのですが。

金を儲ける為の知見（書いててちょっとハズカシイ）は，非常に 深いクロスの底の，ごくごくわずかなセルの中で見つかります． CATDAPは，伝統的な統計学の中で生まれた手法なので 最適性（探索した中では一番良い）を重視します．だから 深いクロスの底の，ごくごくわずかなセルの知見には到達しません． 一方，決定木は他にもっといい解があるかもしれない，という 懸念には頓着しません．成功しそうな知見に早くたくさんヒット することが主たる役割です．それが組み合わせ爆発に対抗する（現時点における）もっとも有効な方策なのだと思います． その潔さは惚れ惚れするほどです.

ここは心理学研究の基礎メーリングリストに投稿された過去の記事を掲載しているページです。

目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

田村のこれまでの研究例

- 非定常時系列解析によるGPTデータの解析
- 多変量時系列モデルを用いた呼吸系の解析
- マルコフ連鎖モデルによる肺がん死亡数シミュレーション
- 白血病治療効果
- 歯学と相関分析
- 歯学と判別分析
- 歩行の時系列解析
- 蓮の花弁の形の解析・ネズミの下顎形状
- 脳の活動と血中ヘモグロビン濃度

ライフサイエンスとの出会い(1-1)

Ann. Inst. Statist. Math.
39 (1987), Part B, 227-241

AN APPROACH TO THE NONSTATIONARY PROCESS ANALYSIS

YOSHIYASU-HAMADA TAMURA

(Received May 22, 1985; revised July 30, 1986)

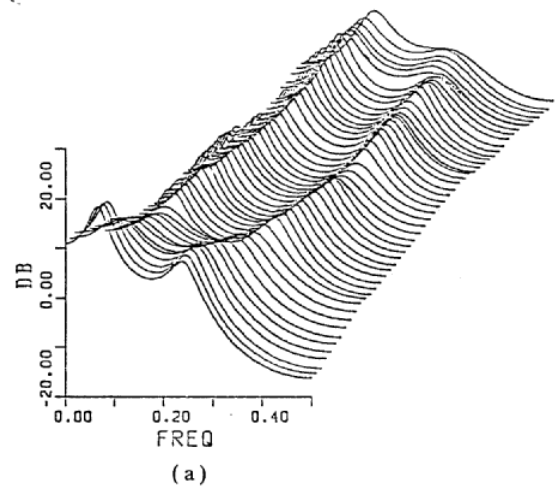
Summary

A Bayesian approach to nonstationary process analysis is proposed. Given a set of data, it is divided into several blocks with the same length, and in each block an autoregressive model is fitted to the data. A constraint on the autoregressive coefficients of the successive blocks is considered. This constraint controls the smoothness of the temporal change of spectrum as shown in Section 2. A smoothness parameter, which is called a hyper parameter in this article, is determined with the aid of the minimum ABIC (Akaike Bayesian Information Criterion) procedure. Numerical examples of our procedure are also given.

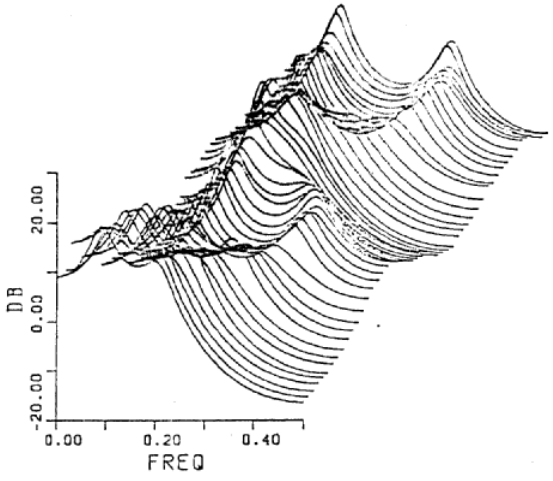
非定常時系列解析の論文。パワースペクトルの時間変化の滑らかさを制約条件として、時変スペクトルを推定している。

ライフサイエンスとの出会い(1-2)

推定された時変スペクトル。



(a)



(b)

Fig. 2. Estimated changing spectrum obtained with (a) $M=4$, $\eta=4.2$ for model 1, (b) $M=4$, $\eta=2.2$ for model 2.

ライフサイエンスとの出会い(1-3)

応用例:GPTデータ(月に1回測定)

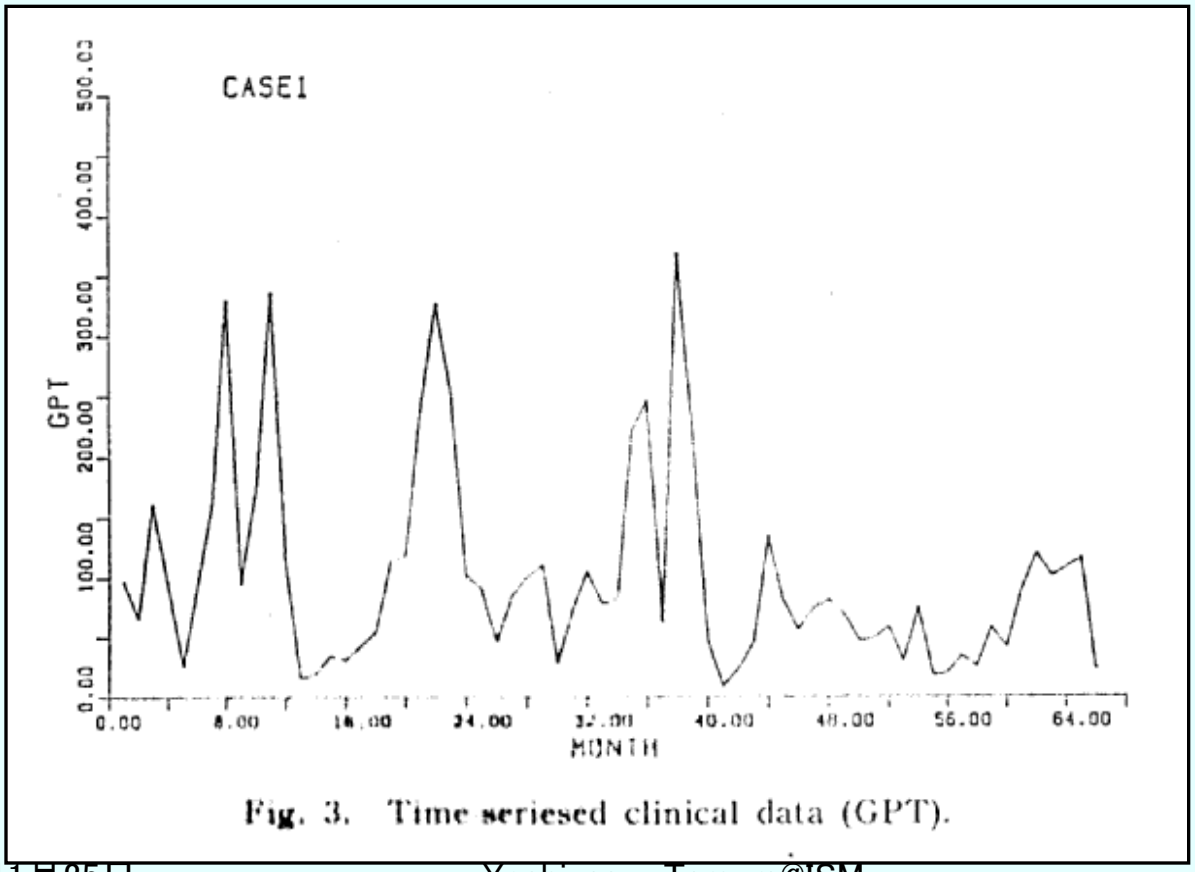


Fig. 3. Time seriesed clinical data (GPT).

ライフサイエンスとの出会い(1-4)

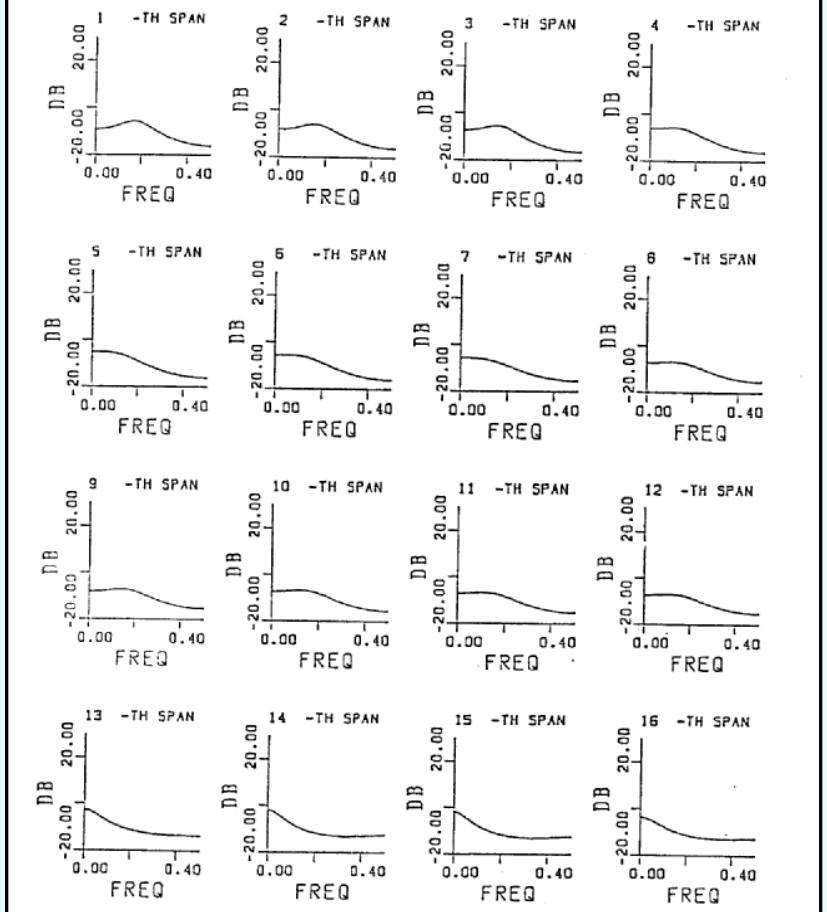


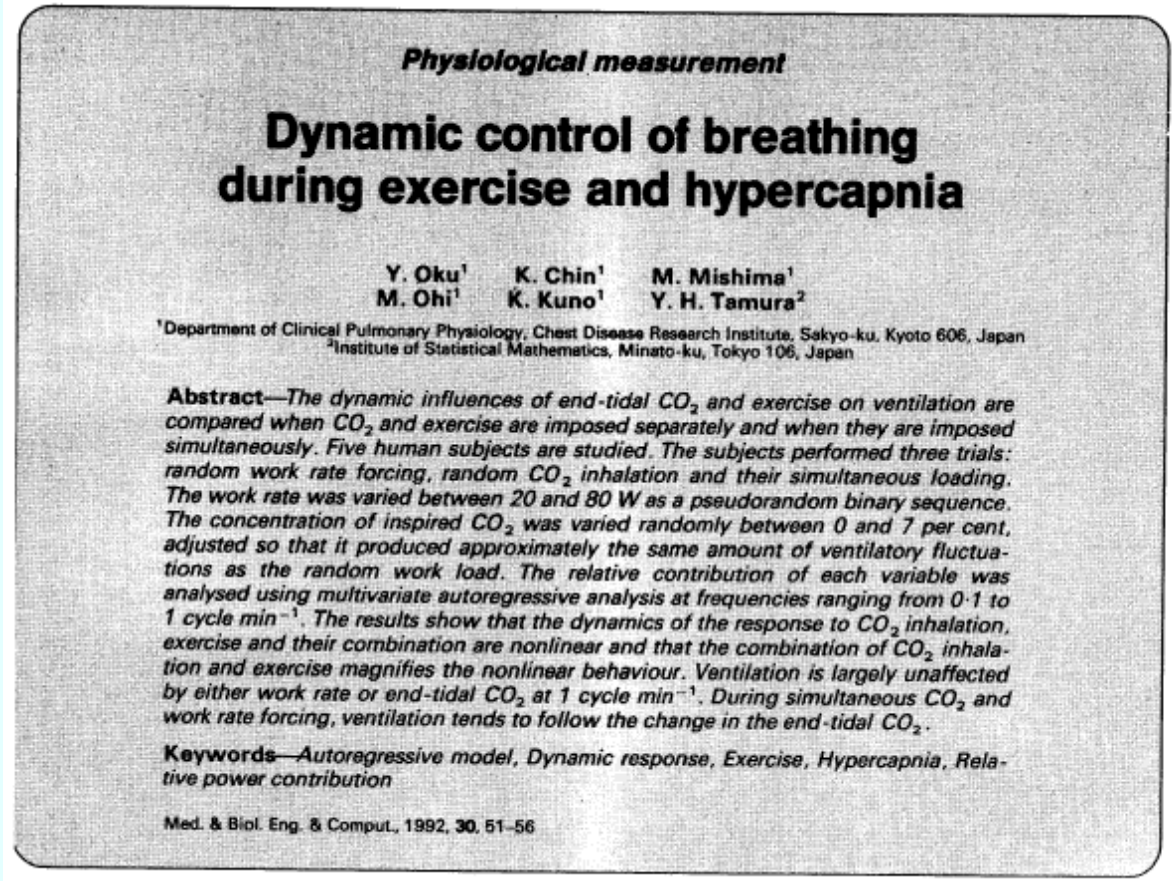
Fig. 4. Estimated changing spectrum for log-transformed GPT data.

応用例：GPTデータの時変スペクトル

13番目のスパン以降、スペクトルのピークがなくなっているのが分かる。

実際には40ヶ月目くらいに、この方は肝硬変の診断を受けている。

ライフサイエンスとの出会い(2-1)



Physiological measurement

Dynamic control of breathing during exercise and hypercapnia

Y. Oku[1] K. Chin[1] M. Mishima[1]
M. Ohi[1] K. Kuno[1] Y. H. Tamura[2]

[1]Department of Clinical Pulmonary Physiology, Chest Disease Research Institute, Sakyo-ku, Kyoto 606, Japan
[2]Institute of Statistical Mathematics, Minato-ku, Tokyo 106, Japan

Abstract—*The dynamic influences of end-tidal CO_2 and exercise on ventilation are compared when CO_2 and exercise are imposed separately and when they are imposed simultaneously. Five human subjects are studied. The subjects performed three trials: random work rate forcing, random CO_2 inhalation and their simultaneous loading. The work rate was varied between 20 and 80 W as a pseudorandom binary sequence. The concentration of inspired CO_2 was varied randomly between 0 and 7 per cent, adjusted so that it produced approximately the same amount of ventilatory fluctuations as the random work load. The relative contribution of each variable was analysed using multivariate autoregressive analysis at frequencies ranging from 0·1 to 1 cycle min^{-1}. The results show that the dynamics of the response to CO_2 inhalation, exercise and their combination are nonlinear and that the combination of CO_2 inhalation and exercise magnifies the nonlinear behaviour. Ventilation is largely unaffected by either work rate or end-tidal CO_2 at 1 cycle min^{-1}. During simultaneous CO_2 and work rate forcing, ventilation tends to follow the change in the end-tidal CO_2.*

Keywords—*Autoregressive model, Dynamic response, Exercise, Hypercapnia, Relative power contribution*

Med. & Biol. Eng. & Comput., 1992, **30**, 51–56

呼吸系の分析のために多変量時系列モデルを用いている。プログラムとして、TIMSACを用いている。

ライフサイエンスとの出会い(2-2)

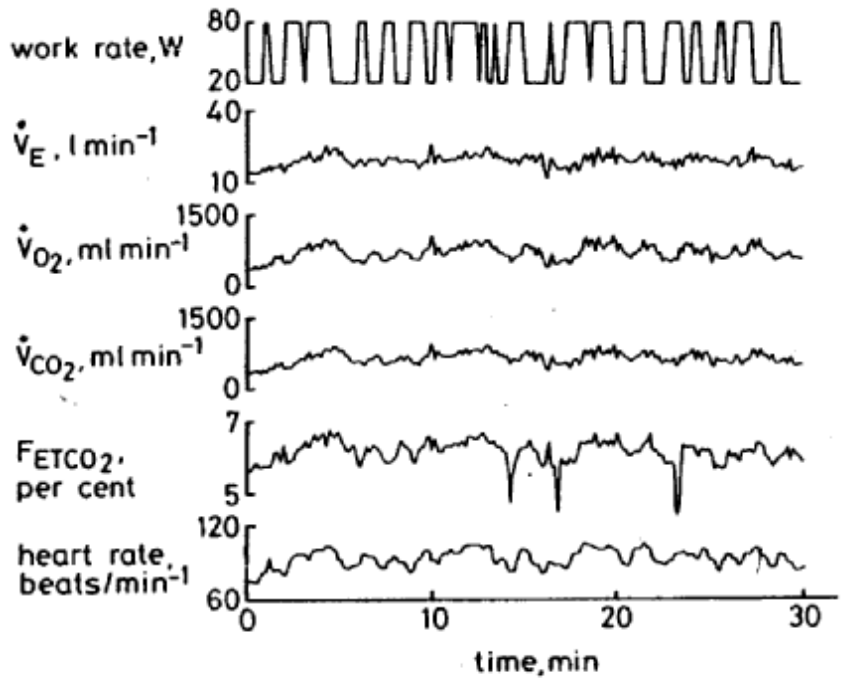


Fig. 3 *Changes in ventilation and gas exchange variables in a sample run during random work rate forcing*

測定データ

サンプリング時間30秒

ライフサイエンスとの出会い(2-3)

解析結果：赤池のパワー寄与率

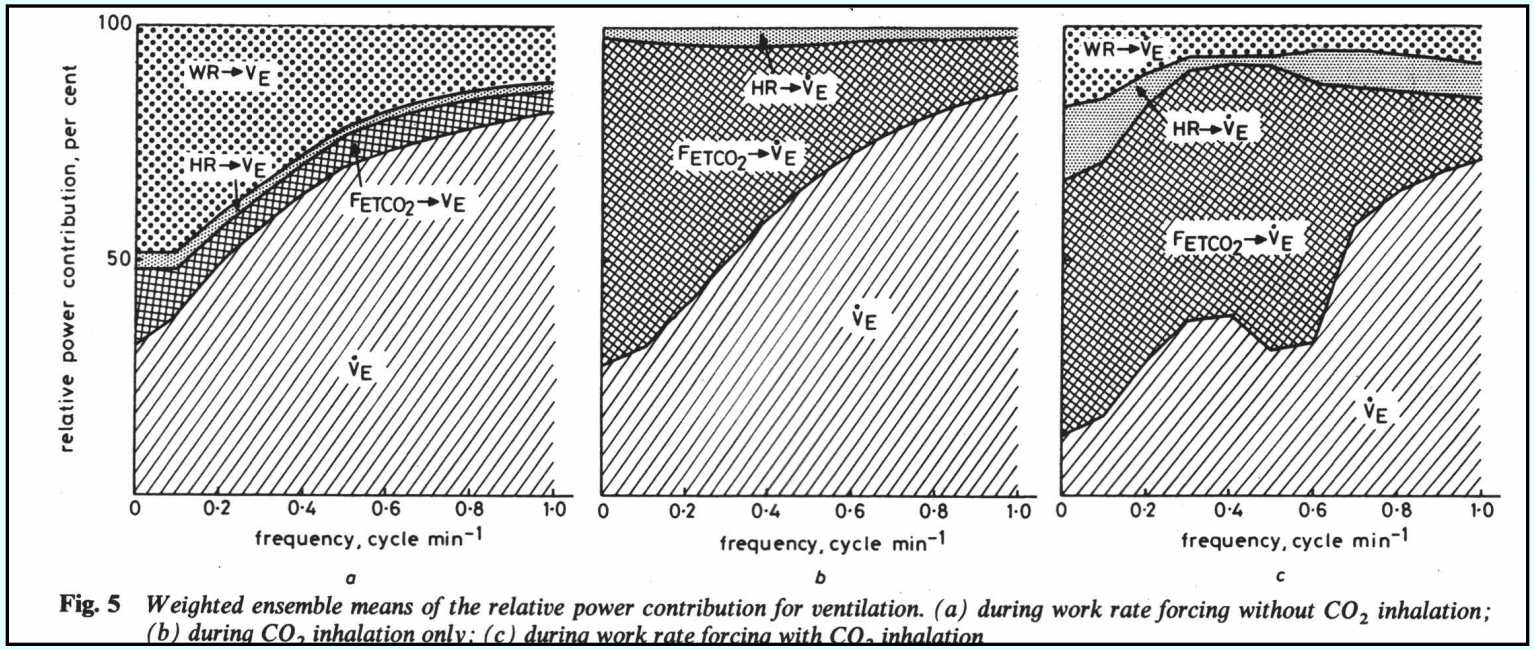


Fig. 5 *Weighted ensemble means of the relative power contribution for ventilation. (a) during work rate forcing without CO_2 inhalation; (b) during CO_2 inhalation only; (c) during work rate forcing with CO_2 inhalation*

TIMSACについて(1)

APPLICATIONS OF TIMSAC

Yoshiyasu Tamura
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu Minato-ku Tokyo

1. Introduction

The TIMSAC(TIMe Series Analysis and Control) program packages have been developed in the past 20 years at the Institute of Statistical Mathematics. Many of the programs within the TIMSAC series were developed to provide procedures for analysing practical applications. The series of TIMSAC are composed of the following four packages.

- original TIMSAC (TIMSAC-72)
- TIMSAC-74
- TIMSAC-78
- TIMSAC-84

TIMSACはTIMe Series Analysis and Control program package の略でセメントキルンの解析を行うために1972年に開発された。ここで説明する他にWindows版もある。SAS、IMSLにも関数やライブラリとして入っている。R版もある。

TIMSACについて(2)

2. TIMSAC-72

The programs in TIMSAC-72 are listed below.

AUTCOR	:Autocovariance and autocorrelation compuation by direct method
MULCOR	:Multiple covariance and correlation computation by direct method
FFTCOR	:Auto and/or cross correlation via FFT
AUSPEC	:Power spectrum estimation by Blackman-Tukey type procedure
MULSPE	:Cross spectrum estimation by Blackman-Tukey type procedure
SGLFRE	:Frequency response function computation(single input)
MULFRE	:Multiple frequency response function computation(multiple inputs)
FPEAUT	:FPE computation for uni-variate autoregressive model
FPEC	:FPE computation for control system model or multivariate autoregressive model
MULNOS	:Relative power contribution computation
DECONV	:Impulse response computation
RASPEC	:Rational spectrum computation(uni-variate)
MULRSP	:Rational spectrum computation(multi-variate)
OPTDES	:Optimal Controller Design
OPTSIM	:Optimal Controller Simulation
WNOISE	:White Noise Simulation

TIMSACについて(3)

3. TIMSAC-74

The programs in TIMSAC-74 are listed below.

CANARM	:Canonical correlation analysis(uni-variate)
AUTARM	:Automatic AR-MA model fitting(uni-variate)
COVGEN	:Covariance generation from gain function
CANOCA	:Canonical correlation analysis(multi-variate)
MARKOV	:Automatic AR-MA model fitting(multi-variate)
PRDCTR	:Prediction by AR-MA model
SIMCON	:Optimum controller design and simulation
NONST	:Locally stationary AR model fitting(uni-variate)
PWDPLY	:Power spectrum plot(PLOT10 compatible)
FRDPLY	:Frequency response function plot
THIRMO	:Third order moment computation
BISPEC	:Bi-spectrum computation

TIMSACについて(4)

4. TIMSAC-78

The programs in TIMSAC-78 are listed below.

UNIMAR	:AR model fitting(uni-variate, minimum AIC method)
UNIBAR	:AR model fitting(uni-variate, Bayesian method)
BSUBST	:Subset regression analysis by a model linear in parameters (uni-variate, Bayesian methos)
MULMAR	:AR model fitting(multi-variate, minimum AIC method)
MULBAR	:AR model fitting(multi-variate, Bayesian method)
PERARS	:Periodic autoregression(uni-variate, minimum AIC method)
MLOCAR	:Locally stationary AR model fitting(uni-variate, minimum AIC method)
BLOCAR	:Locally stationary AR model fitting(uni-variate, Bayesian method)
MLOMAR	:Locally stationary AR model fitting(multi-variate, minimum AIC method)
BLOMAR	:Locally stationary AR model fitting(multi-variate, Bayesian method)
NADCON	:Noise adaptive controller
EXSAR	:AR model fitting(uni-variate, exact maximum likelihood method)
XSARMA	:AR-MA model fitting(uni-variate, exact maximum likelihood method)

TIMSACについて(5)

5. TIMSAC-84

The programs in TIMSAC-84 are listed below.

BAYSEA	:Bayesian seasonal adjustment
BAYTAP	:Bayesian tidal data analysis
DECOMP	:Time series decomposition into components by the Bayesian approach
LOCCAR	:Locally constant AR model
TVCAR	:Time varying coefficient AR model
NONSPA	:Nonstationary spectrum analysis by minimum ABIC procedure
MULCON	:Multiple time series analysis by simulated prediction and control
SNDE	:Stochastic nonlinear differential equation model
ADAR	:Amplitude dependent AR model
EPTREN	:Exponential polynomial or Fourier series modeling of trend and :cycle of Poisson intensity
LINLIN	:Linearly self-exciting process with trend and linear input
PGRAPH	:Graphic point process analysis
LINSIM	:Simulation of the point process identified by LINLIN
SIMBVH	:Simulation of a bivariate Hawkes point process
PTSPEC	:Point process spectrum by direct Fourier transform

ライフサイエンスとの出会い(3-1)

Evaluation of Cancer Prevention Strategies by Computerized Simulation Model: Methodological Issues

Naohito Yamaguchi,[1] Yoshiyasu Tamura,[2] Tomotaka Sobue,[3] Suminori Akiba,[4] Megu Ohtaki,[5] Yasumasa Baba,[2] Shoichi Mizuno,[1] and Shaw Watanabe[1]

[1]Division of Epidemiology, National Cancer Center Research Institute, Tokyo, Japan; [2]Institute of Statistical Mathematics; [3]Radiation Effect Research Foundation; [5]Research Institute for Nuclear Medicine and Biology, Hiroshima University

A computerized simulation model developed to evaluate the potential impact of primary and secondary prevention is discussed from methodologic perspectives. In the simulation model, named CANSAVE (Cancer Strategy Analysis and Validation of Effect), the natural history of cancer was modeled as a Markovian stochastic process from cancer-free state to death. The lung cancer death rate among Japanese males was projected for 50 years to the year 2041. The simulation showed that the age-adjusted death rate would increase and reach a peak of 166 per 100,000 in 1989 and then decrease to 148 in 2003. It then shows a tendency to increase again, up to 255 in 2028. This change may be attributed to a lower smoking initiation rate among those born in the 1930s. Promotion of mass screening programs exhibits a more prompt effect than antismoking efforts, but the reduction in annual deaths is expected to be only 11%, even if a 100% participation is realized by the year 2000. The reduction in smoking initiation rate, on the other hand, begins to show a visible effect very slowly. It was predicted that a 1% annual reduction in smoking initiation rate would result in a 20% decrease in the number of deaths in 2041. The smoking cessation program is in the middle with regard to promptness. The predicted reductions in lung cancer deaths in 2041 were 13, 47, and 66%, respectively, when the annual smoking cessation rate was increased from 0.46% (present status) to 1, 3, and 5%. The combined application of all three preventive measures seems essential to realize the most effective reduction in lung cancer deaths. — Environ Health Perspect 102(Suppl 8):67–71 (1994)

Key words: cancer, lung, smoking, screening, simulation, prevention, evaluation, epidemiology

肺がん患者数の予測シミュレーション

（マルコフ連鎖を用いた確率モデル）

ライフサイエンスとの出会い(3-2)

シミュレーション結果：年齢層別死亡率

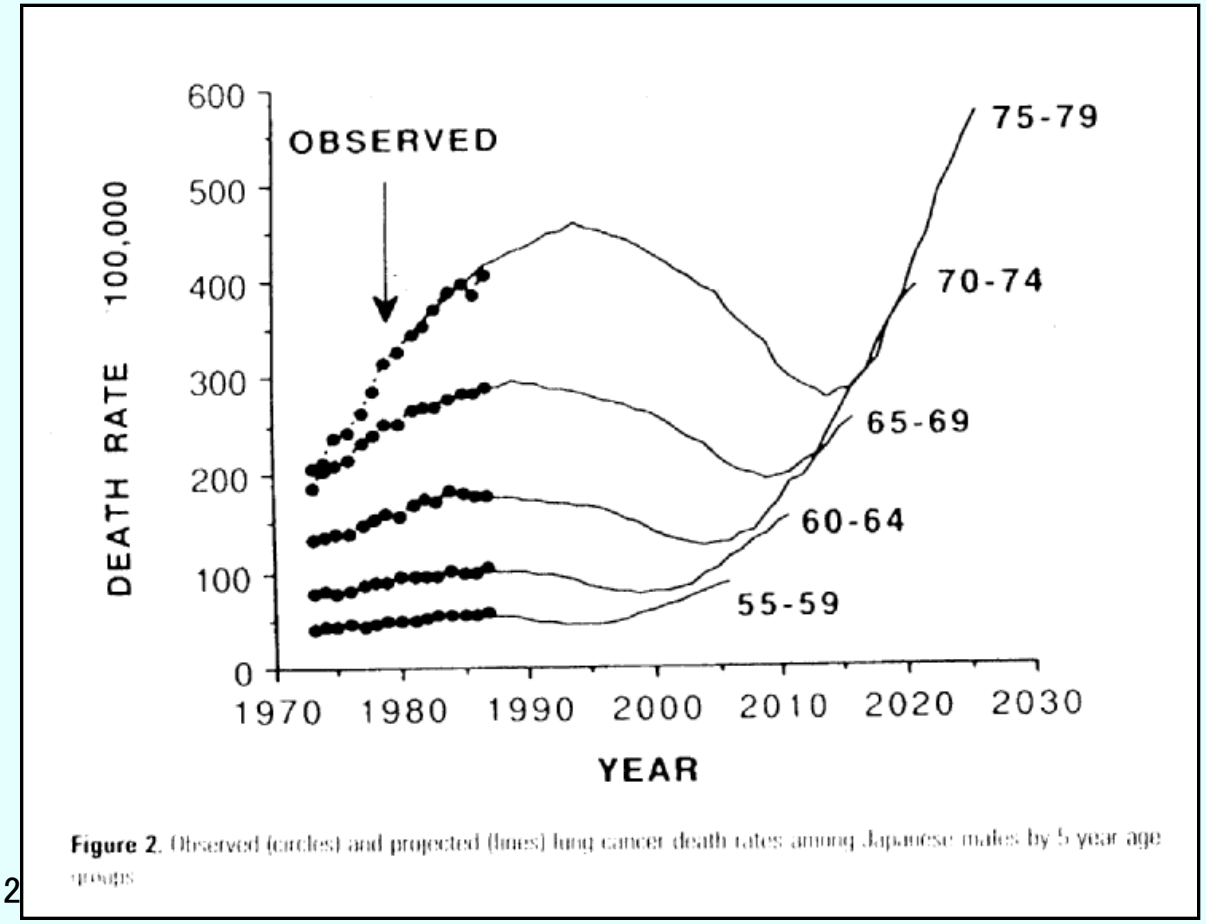
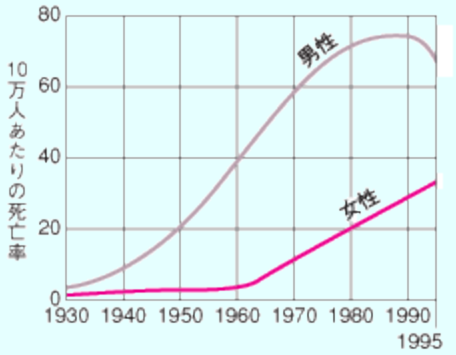


Figure 2. Observed (circles) and projected (lines) lung cancer death rates among Japanese males by 5 year age groups

万有製薬:メルクマニュアル医学百科
http://mmh.banyu.co.jp/mmhe2j/sec04/ch057/ch057a.html



各種の癌の中で、肺癌は男女ともに最も死亡率の高い癌です。肺癌による死亡者数は、喫煙者の増加にともなって増えています。データによる2001年の肺癌 による推定死亡者数は、15万7000人以上に上ります。このうち約9万人が男性、約6万7000人が女性です。この数字は、すべての癌による全死亡者数 の約28%に相当します。

ライフサイエンスとの出会い(4-1)

ニューラルネットワークと時系列解析
Neural Networks and Time Series Analysis

田村義保
Yoshiyasu Tamura

統計数理研究所　統計データ解析センター
東京都港区南麻布４－６－７
Statistical Data Analysis Center, The Institute of Statisitcal Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106 JAPAN

Abstract

Recently many researchers, who are not statisticians, are willingly to use so-called neural networks in stead of statistical data analysis in order to analyse their data. A few of statisticians take interest in neural networks. In this manuscript whether neural networks and statistical data analysis conflict each other or the one of them complements the other are discussed. For example the number of leukemic cells are predicted with the aid of nonlinear time series model and neural network. Results of both methods are compared.

白血病の治療効果の予測

時系列モデル

VS.

ニューラルネットワーク

ライフサイエンスとの出会い(4-2)

5．おわりに

時系列解析の3手法とニューラルネットワークの解析で、v_{51}からv_{95}までの予測値と実際の値との差の二乗の和、すなわち残差二乗和を計算したのが次の表である。

モデル名	残差二乗和
自己回帰モデル	1.728982
ベイズ型非定常モデル	0.1904905
カルマンフィルタ	0.3147818
ニューラルネットワーク	0.2855617

予測値を求める際に、自己回帰モデルでは、データは既知のものとしている。また、すでに述べたようにベイズ型非定常モデルのあてはめでは、全てのデータを用いている。カルマンフィルタを用いた非定常モデル、ニューラルネットワークを用いた解析では、51番目以降のデータは未知として、予測値を用いて次々に予測している。このようなこともあって、ベイズ型非定常モデルの残差二乗和が最小になっているのかもしれないが、予測の良さという点から見ると、このモデルが最良であると言える。しかし、すでに述べたように、このモデルの推定はデータにオーバーフィットする結果を与えている。また、薬の効果と

ライフサイエンスとの出会い(4-3)

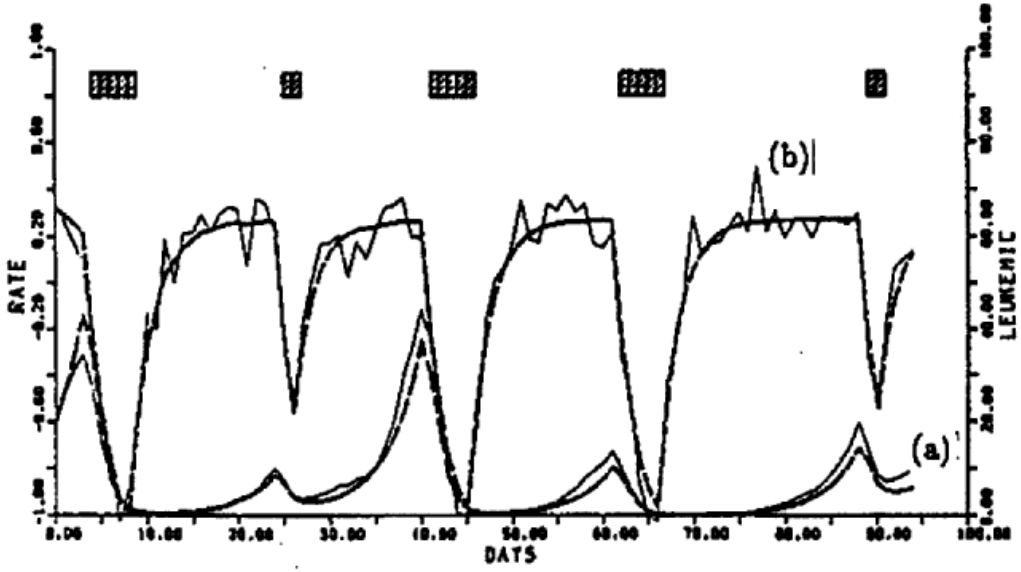


図1. 白血病細胞数の変化

ライフサイエンスとの出会い(5-1)

斜状骨欠損および歯周組織のX線学的検討

池島 厚　田村義保*
日本大学松戸歯学部　放射線学教室
*統計数理研究所　統計計算開発センター研究開発第一部

キーワード：斜状骨欠損，歯科用X線写真，歯根膜腔，歯槽硬線，相関分析

要旨：咬合性外傷等による垂直性骨吸収の一型である斜状骨欠損52歯のX線写真を抽出して，その斜状骨欠損の歯軸に対する傾斜度，歯根膜腔の幅，歯槽硬線の厚さ，歯槽骨と歯槽硬線間とのコントラストを調査した。その後，斜状骨欠損の歯軸に対する角度を対象として，歯根膜腔の幅，歯槽硬線の厚さおよび歯槽骨と歯槽硬線とのコントラストのそれぞれを相関分析した結果，

1．斜状骨欠損の歯軸に対する角度と歯根膜腔の幅とでは，相関のある部位が認められなかった。

2．斜状骨欠損の歯軸に対する角度と歯槽硬線の厚さとでは，下顎における近心から遠心方向に傾斜した症例群の近心歯槽骨頂部で負の相関が，また遠心から近心方向に傾斜した症例群の遠心歯槽頂部で相関が認められた。

3．斜状骨欠損の歯軸に対する角度とコントラストでは，相関があっても部位的な特徴はなかった。

以上より咬合性外傷における斜状骨欠損では，その傾斜角度と歯槽硬線の幅は密接な関係のあることが判明した。

ライフサイエンスとの出会い(5-2)

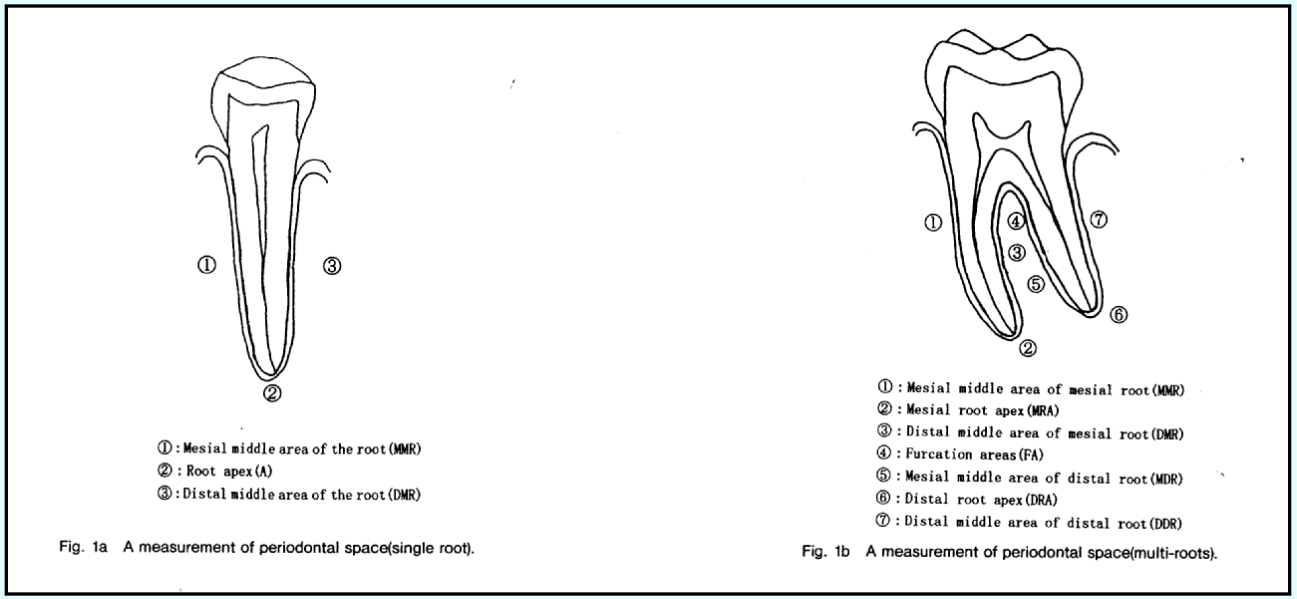


Fig. 1a A measurement of periodontal space(single root).

Fig. 1b A measurement of periodontal space(multi-roots).

ライフサイエンスとの出会い(6-1)

Journal of Oral Science, Vol. 44, No. 1, 13-17, 2002

Original

Differential diagnosis between dentigerous cyst and benign tumor with an embedded tooth

Atsushi Ikeshima and Yoshiyasu Tamura[§]

Department of Radiology, Nihon University School of Dentistry at Matsudo, Chiba 271-8587

[§]Research and Development 1, Center for Development of Statistical Computing,
The Institute of Statistical Mathematics, Tokyo 106-8569

(Received 16 October 2000 and 21 February 2002)

ライフサイエンスとの出会い(6-2)

Abstract: It has been generally recognized that the radiological appearances of cysts and tumors related to an embedded tooth are similar. However, based on their clinical experience, Abrams et al. pointed out that there was a difference between the two lesions at the attachment point to the embedded tooth. To investigate this difference, we conducted a study employing the radiographs of patients who visited Nihon University Dental Hospital at Matsudo and were pathologically defined as having a cyst or tumor. Using radiographs of these patients, we investigated the attachment point to the embedded tooth, and expressed the results as the proportion of the attachment point to the embedded tooth root length. The study was carried out in 100 patients with cysts (87 dentigerous cysts and 13 odontogenic keratocysts), and 27 patients with benign tumors (24 ameloblastomas and three adenomatoid odontogenic tumors). Prior to treatment based on the numerical results, the distribution of the results was examined. Thus, we evaluated several methods of examining the distributions, and found the best method to be discriminant analysis. The results showed that the discriminated boundary value (from the cemento-enamel junction) was 0.38 for the embedded tooth root length. The cases showing a boundary value of less than 0.4 for the cemento-enamel junction were judged to be cysts, and those showing a value of 0.4 or more were judged to be benign tumors. The rate of misjudgement was 28% in the cyst group and 33.3% in the benign tumor group. (J. Oral Sci. 44, 13-17, 2002)

Key words: differential diagnosis; embedded tooth; cyst; benign tumor; discriminant analysis.

Correspondence to Dr. Atsushi Ikeshima, Department of Radiology, Nihon University School of Dentistry at Matsudo, 2-870-1, Sakaecho-Nishi, Matsudo, Chiba 271-8587, Japan
Tel: +81-047-360-9530
Fax: +81-047-360-9418
E-mail address: ikeshima@mascat.nihon-u.ac.jp

Introduction

Dentigerous cysts and some benign tumors with embedded teeth show similar images on the radiographs(1-4). These benign tumors are ameloblastoma and adenomatoid odontogenic tumor without calcified bodies on the radiograph. Based on their clinical experience, Abrams et al.(5) pointed out a difference between cysts and benign tumors with an embedded tooth: the attachment point to the embedded teeth. Generally, radiographs of the dentigerous cysts and the benign tumors with an embedded tooth have similar appearances in terms of the contents and the boundaries of the lesions(6-8). The treatment plan is basically different for dentigerous cysts(9,10) and benign tumors(11,12). Therefore, it is important to distinguish between dentigerous cysts and benign tumors on the radiographs. For the above reasons, we attempted to make a differential diagnosis between dentigerous cysts and benign tumors with an embedded tooth. A study of this problem is underway(13-15).

In this study on dentigerous cysts and benign tumors with an embedded tooth, the number of diseases and cases has been on the increase compared with our last report(15).

ライフサイエンスとの出会い(6-3)

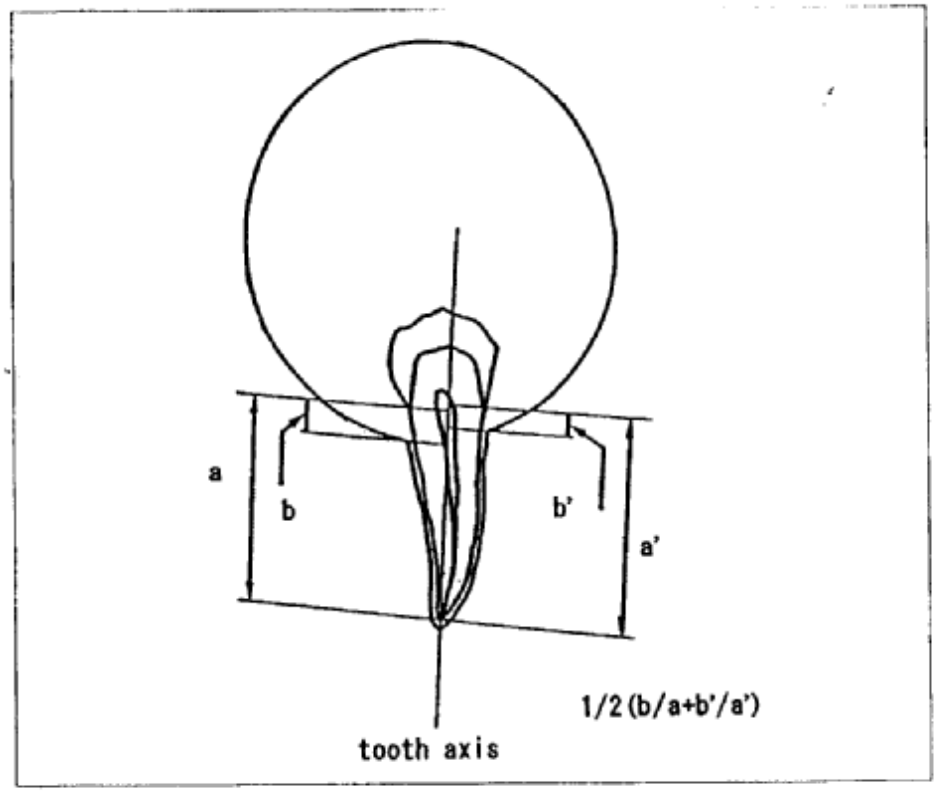


Fig. 1 An example of the measurement. (A single root)

ライフサイエンスとの出会い(6-4)

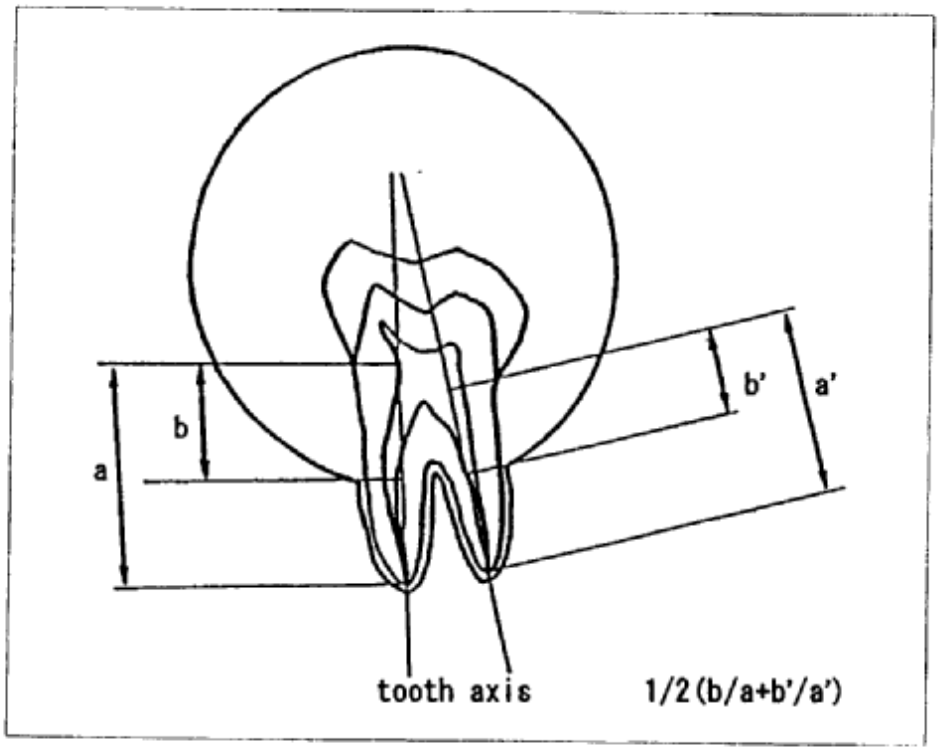


Fig. 2 An example of the measurement. (A multi-root)

ライフサイエンスとの出会い(6-5)

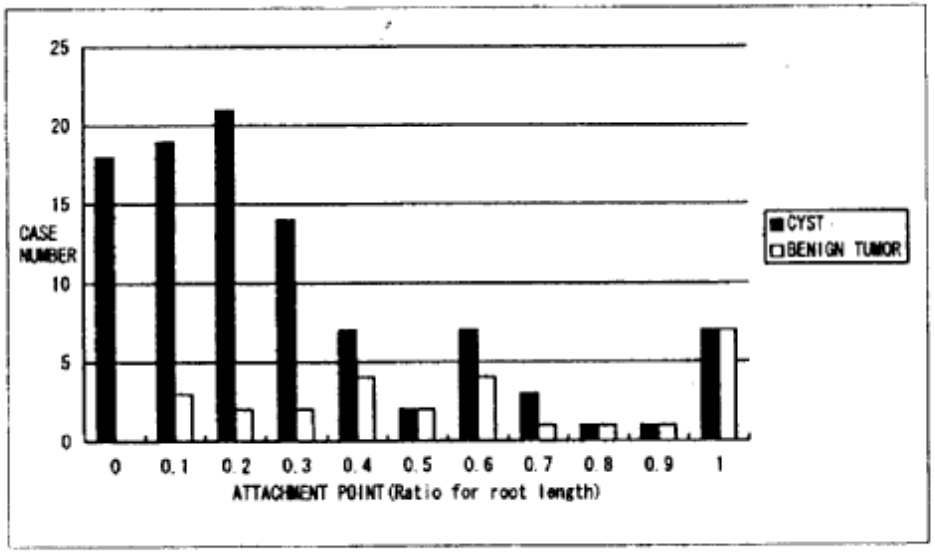


Fig. 3 Distribution of the data.

測定結果のヒストグラム

Attachment Pointにより二つの病気を判別する。

ライフサイエンスとの出会い(7-1)

体型による適合クッションの推定
—ファジィ推論モデルと重回帰モデルによる検討—*

三家礼子**，藤巻吾朗**，田村義保***，野呂影勇****

Research on seating comfort has been often conducted in the field of ergonomics. Consideration of differences in the attributes of a person who wishes to sit when he or she chooses a cushion is important. Therefore, focus has been placed on an attribute of body type and a model to predict an appropriate cushion for an individual has been proposed. The relationship between body size and seating comfort has already been in preceding research, and an item for evaluation that is extensively related to overall cushion evaluation is an item related to hardness. In the past, multiple regression was used in a model to indicate the relationship between items for evaluation as expressed in cushion characteristics and overall evaluation as expressed by seating comfort. In the current research, fuzzy logic using non-linearity has been proposed. An experiment for verification was conducted; results indicated a high correlation between predicted values and measured values, and the validity of the system itself was noted. Further progress in research indicated that the model would be highly superior with addition of dummy variables to multiple regression, which is inappropriate, in a conventional subjective evaluation model. Evaluation and examination of both systems were also conducted.

ライフサイエンスとの出会い(8-1)

阪神地区における女子の飲行動の年齢階層別比較

Comparison of behaviors on drinking by age group among females of 20-75 years of age

早川史子[*1]・前田昭子[*2]・岡崎章子[*1]・石津陽子[*1]・猪口智子[*1]・南　幸[*3]
中森正代[*4]・田村義保[*5]
([*1]滋賀県立大学・[*2]相愛女子短期大学・[*3]幾央大学・[*4]徳島大学・[*5]文部科学省数理統計研究所)

ライフサイエンスとの出会い(8-2)

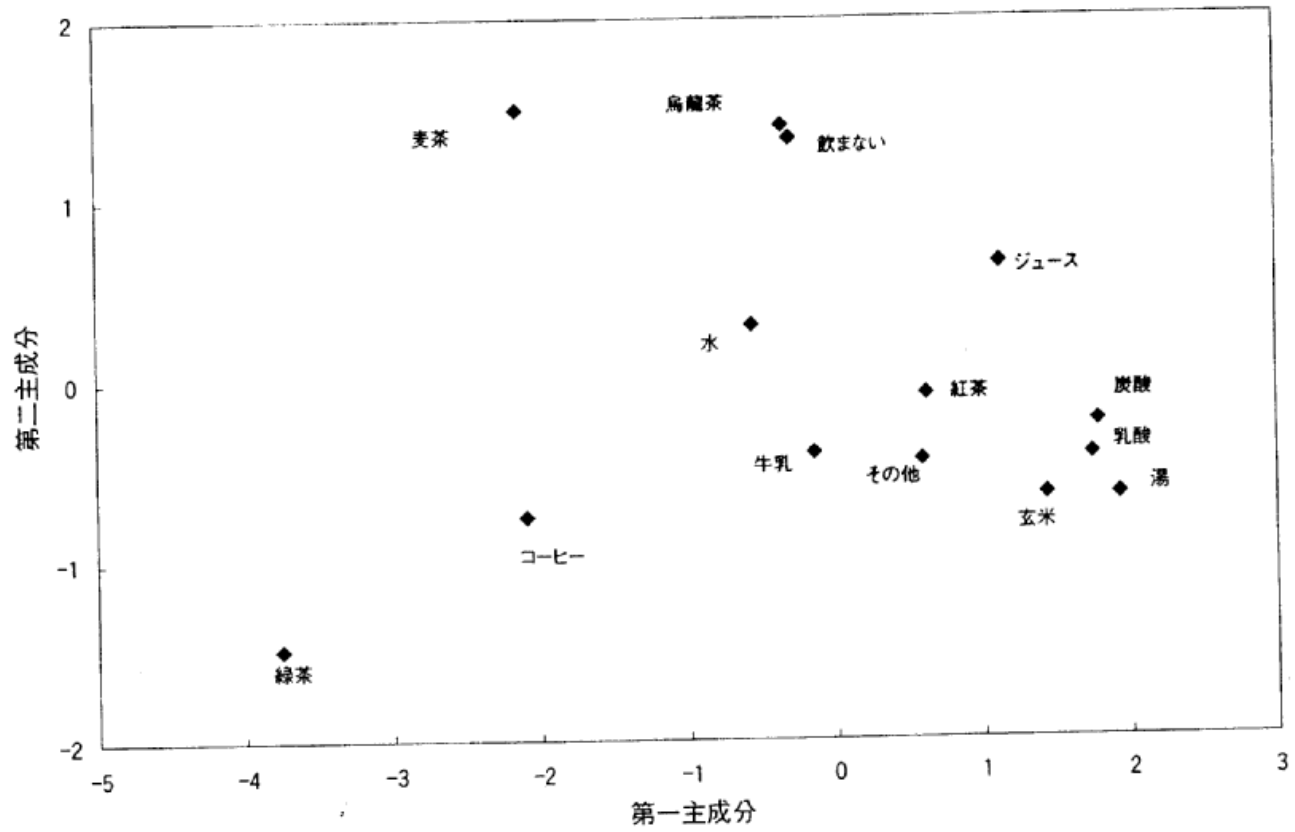


図1　一日の飲み物の飲用率の主成分プロット

Analysis of walking improvement with dynamic shoe insoles, using two accelerometers

Yuriko Tsuruoka[a,*], Yoshiyasu Tamura[a], Ryosuke Shibasaki[b], Masako Tsuruoka[b]

[a]*Center for Development of Statistical Computing, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan*

[b]*Center for Spatial Information Science, University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan*

P型フーリエ記述子を用いた蓮の花弁に関する定量的な分析

鄭　澤宇・田村　義保

本研究ではP型フーリエ記述子の方法を用い，蓮花弁の形状を定量的にデータ化し，統計学的手法を駆使して様々な分析を行った。開曲線にも有効なP型フーリエ記述子を用いて植物器官の形態を定量的に分析する初めての研究である。蓮花弁先端部分の輪郭線の特徴を自動的に抽出して，統計的な手法で品種を識別する方法を提案している。

自動化を実現するために，花弁のデジタル写真から輪郭線を抽出し，P型フーリエ記述子を計算するためのソフトウェアを開発した。輪郭線のコーディング方法はFreemanチェーンを使用している。抽出したP型フーリエ記述子に対し，主成分分析を行い，必要な主成分得点を用いてSVM及び2次元判別手法によって品種の分類を行った。認識率の評価はジャックナイフ法により行い，その結果において，5種類の蓮に関する正答率は78%に達した。

最近研究(2-2)

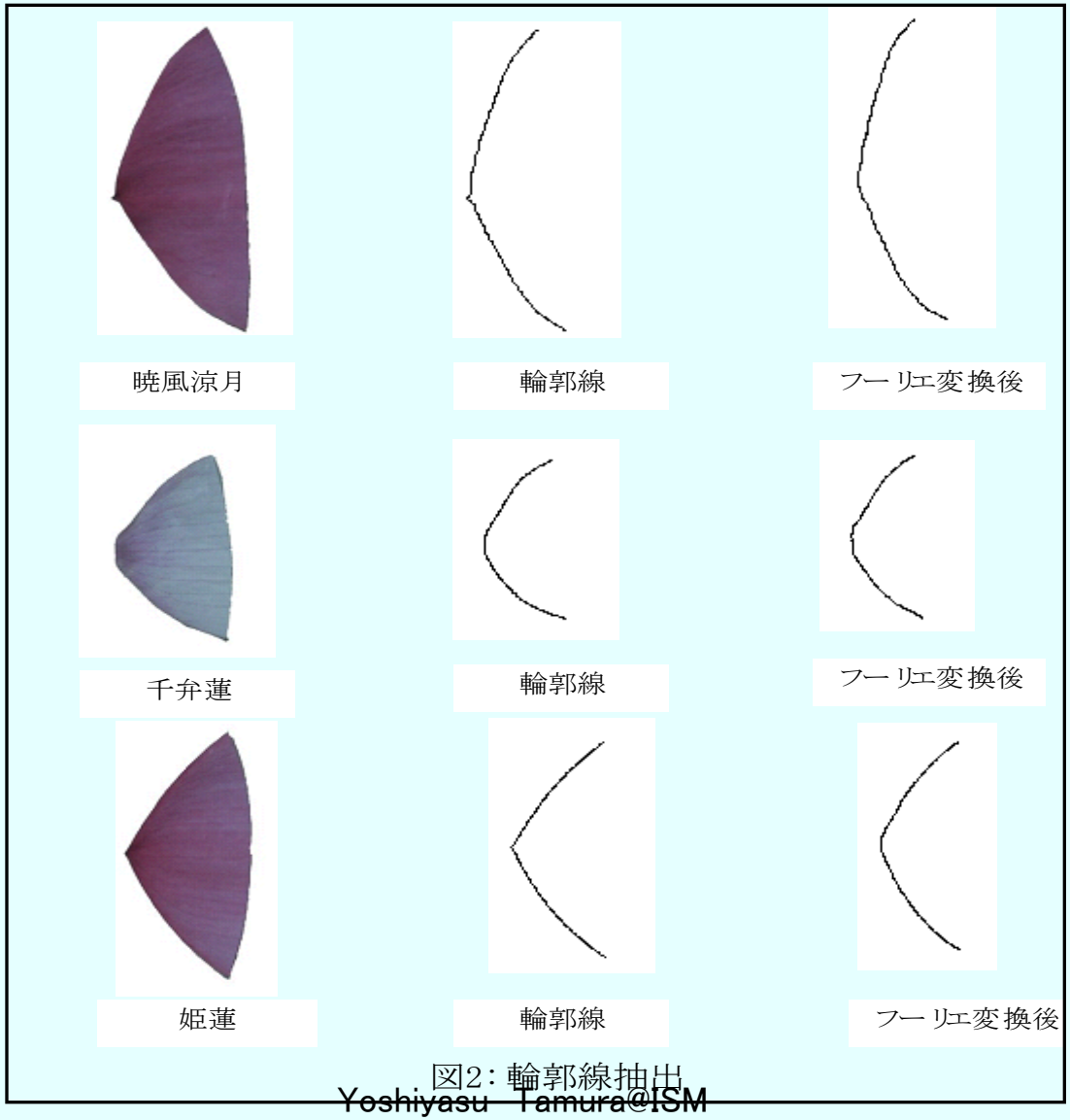


図2：輪郭線抽出

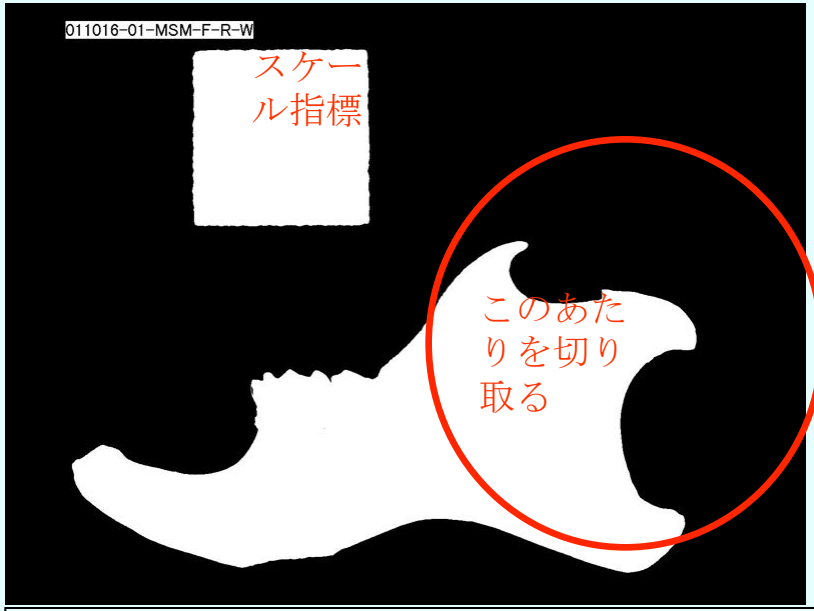
表3: ジャックナイフによる認識率

線形、二次、RBFはそれぞれ線形カーネルを用いたSVM、二次元判別、=0.02のRBFカーネルを用いたSVM判別方法である。係数は輪郭線抽出に用いられたP型フーリエ記述子の係数である。累積寄与率は主成分分析の累積寄与率である。その累積寄与率以上である累積寄与率を得られる最も少ない主成分を使用して判別分析を行う。P型フーリエ記述子の係数によって累積率と主成分の関係は表2に示している。

記述子(個)					
累積寄与率	16	32	48		
	認識率(%)	認識率(%)	認識率(%)		
認識方法	線形	二次		線形	二次
	RBF	線形	二次	RBF	
80%	68	66	70	68	70
	66	66	70	72	
85%	76	72	74-	68	76
	68	68	76	68	
90%	76	72	74	72	78
	72	70	52	70	

記述子(個)					
累積寄与率(%)	64	96	128		
	認識率(%)	認識率(%)	認識率(%)		
認識方法	線形	二次	RBF	線形*	二次
	RBF	線形	二次	RBF	
80%	66	68	68	70	72
	70	74	46	68	
85%	72	72	72	72	64
	70	70	26	66	
90%	74	46	70	68	-
	70	68	-	70	

ネズミの下顎骨形状と遺伝系統



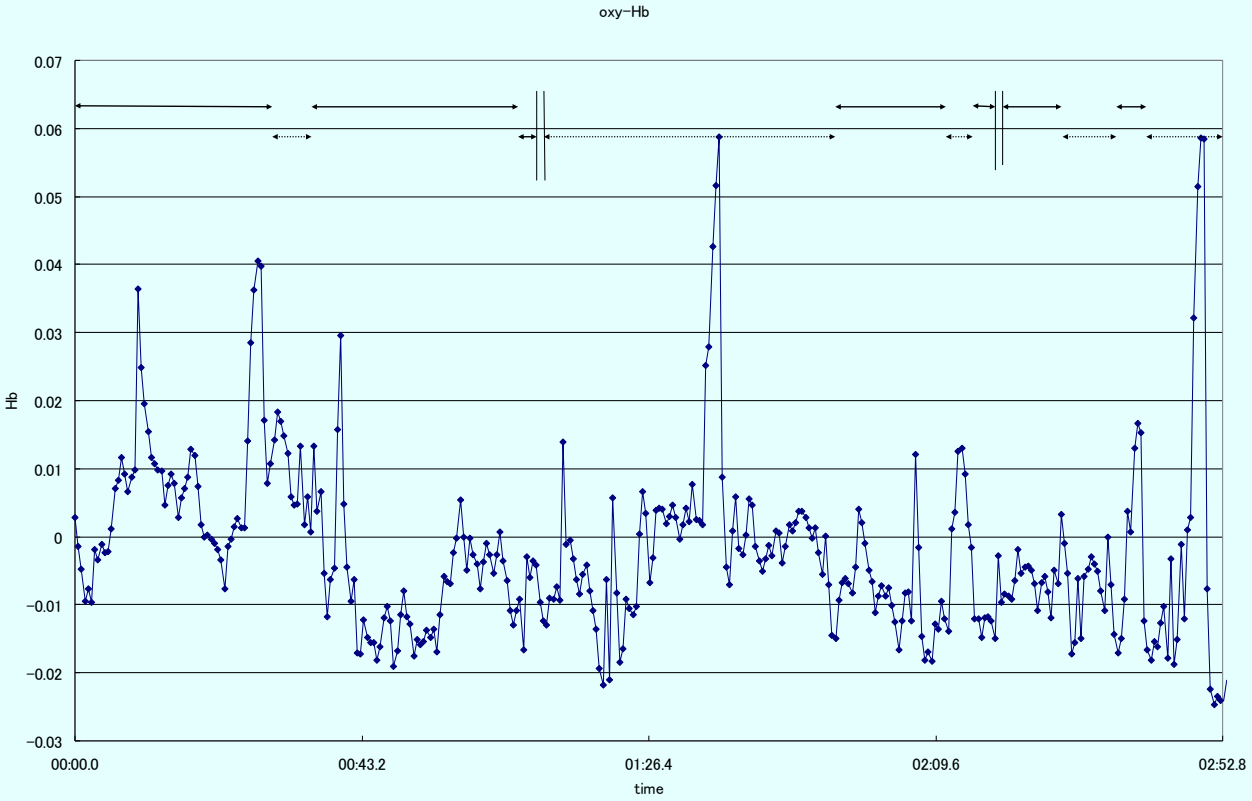
国立遺伝学研究所：城石俊彦・細谷理樹
統計数理研究所　：田村義保・鄭　澤宇

研究目的：ネズミの下顎骨形状の遺伝系統による差異を見出す。

研究方法：国立遺伝学研究所で撮影された下顎骨の形状をp型フーリエ記述子により分析し、遺伝子のどの部分が形状決定に関係しているかを見出す。

新規性：従来は閉曲線の解析のために開発された楕円型フーリエ記述子を用いていた。このために、下顎骨全体をフーリエ記述子により数量化する必要があり、種間の差異がない部分も解析に含める必要があった。P型フーリエ記述子は開曲線を扱えるために、一部分のみを抽出して解析できる。一部分のみを発生学上の諸問題を考慮して自動的に抽出するための画像処理方法と抽出した部分を自動的にコード化し、p型フーリエ記述子を計算する方法を開発した。遺伝学と統計科学の融合が研究を進めるのに不可欠である。

脳血流中のヘモグロビン濃度の変化



目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

データマイニング(1)
weblio　MBA用語集

- 統計学等のデータ解析の技法を駆使して、大量のデータを網羅的に分析することにより、ビジネス上意味のある知識を発見する手法。
- データ・マイニングの例としては、あるスーパーの購買履歴データを分析することで、週末に同じ人が紙おむつと缶ビールを同時に買う傾向があることが分かった、などが挙げられる。サンマと大根、パスタとミートソースのような、人間が容易に見当のつけられる関係ではなく、大量のデータをコンピュータで分析して初めて分かるような発見を得ることを、特にデータ・マイニングと称する。
- マーケティングの需要予測、販促、店頭マーチャンダイジング等の分野において主に取り入れられてきたが、株価予測、人的資源管理等の分野にも活用が期待されている。

データマイニング(2)
weblio　マーケティング用語集

- 大量の販売データなどを様々な統計手法により分析することで、消費者の購買傾向や商品間の関連性を見つけ出す手法またはプロセスのこと。
  データマイニングには様々な分析手法が存在しますが、マーケティングの観点から最も重要視するポイントは、顧客の特性や傾向分析と併買傾向分析があげられます。
  近年、市販の様々なツールが提供されるようになりましたが、分析して得られた仮説を元に、どのようなプロモーション戦略を構築するかが重要な課題となります。

データマイニング(3)
weblio　広告用語辞典

- 膨大なデータベースからビジネスに有益なパターンやルールを発見する手法。多量のデータを鉱山に見立て、その中から価値ある情報を採掘（mining)するのでこう呼ばれる。特定のモデルを仮定した多変量解析と異なり、特定のビジネス上の知識やルールを獲得するために行われることなどから、使われる手法は様々である。データ・マイニングの専用ツールには、決定木I（CHAID,C5,CART)、ニューラル・ネットワーク、自己組織化マップ（コホネン・ネットワーク）、リンク分析、K-meansクラスター分析、アソシエーション・ルール、記憶ベース推論などの手法が、データのビジュアル化ツールや複数の結果の評価ツールとともに実装されているという。

データマイニング(4)
weblio　IT用語辞典バイナリ

- **データマイニング**とは、データベースに蓄積されている大量のデータから、統計や決定木などを駆使して、マーケティングに必要な傾向やパターンなどの隠された規則性、関係性、仮説を導き出す手法のことである。マイニング（mining）とは「採鉱」を意味するもので、いわば眠っている金脈を探り当てることになぞらえられている。
- 一見無秩序・無関係に見えるデータの山も、着眼点次第では各事項の間に有益な連関が見えてくることが少なくない。例えば、ある商店ではサングラスを買う人の多くが一緒にガムを買っているという事実が見つかるかも知れない。あるいは他の店舗では、曇りがちの日には生魚の売り上げが伸びているかもしれない。こうした連関を実績として見出すことによって、サングラスの陳列棚の近くにガムを配置する、とか、雲の多い日に鮮魚のセールを実施するとか、効果的なマーケティングを行うことができる。
- データマイニングは、データベースの発展を中心とした情報技術の向上によって盛んに行われるようになった手法であるといえる。元となるデータが多ければ多いほど、処理作業は膨大なものになるが、実証性は高くなる。既存データを専用のデータベースに取り込んで意思決定に活用するシステムはデータウェアハウスと呼ばれるが、データウェアハウスは一個のデータマイニングツールであるといえる。

データマイニング(5)
weblio　OR辞典(1)

概要

- データベースに蓄えられた多量のデータから，機械学習（machine learning）や統計的手法（statistical method）を用いてデータの中に含まれる知識を発掘する手法をいう．知識発見プロセスとしての，データ獲得，選択，前処理，変換，知識発見アルゴリズムの適用，解釈，評価といった一連のサイクルを指す．獲得した知識に基づく意思決定が目的であり，データ収集，発掘，評価といった人間と計算機の共同作業を伴う知識マネジメントとして捉えられる．

詳説

- **データマイニング** (data mining)は，データベース (data base) [5] に蓄えられた大量の生データに対して，機械学習 (machine learning)に関連する複数の手順を用いる戦略により，データに内在する規則性 (regularity)，制約 (constraint)，ルール (rule)などを効率よく求める研究である．なお，データベースからの知識発見 (KDD: knowledge discovery in databases)とも呼ばれ，知識発見 (knowledge discovery)に関わる多数の学習アルゴリズムが，人工知能だけでなくデータベースや統計学の側面を含めて研究されている．まず，ノイズや例外を含み疎な構造をもつことも多い生データを対象としたデータマイニングに共通する知識発見の手順を(1)～(6)に簡単に示す[1]．

データマイニング(6)
weblio　OR辞典(2)

概要

- データベースに蓄えられた多量のデータから，機械学習（machine learning）や統計的手法（statistical method）を用いてデータの中に含まれる知識を発掘する手法をいう．知識発見プロセスとしての，データ獲得，選択，前処理，変換，知識発見アルゴリズムの適用，解釈，評価といった一連のサイクルを指す．獲得した知識に基づく意思決定が目的であり，データ収集，発掘，評価といった人間と計算機の共同作業を伴う知識マネジメントとして捉えられる．

詳説

- **データマイニング** (data mining)は，データベース (data base) [5] に蓄えられた大量の生データに対して，機械学習 (machine learning)に関連する複数の手順を用いる戦略により，データに内在する規則性 (regularity)，制約 (constraint)，ルール (rule)などを効率よく求める研究である．なお，データベースからの知識発見 (KDD: knowledge discovery in databases)とも呼ばれ，知識発見 (knowledge discovery)に関わる多数の学習アルゴリズムが，人工知能だけでなくデータベースや統計学の側面を含めて研究されている．まず，ノイズや例外を含み疎な構造をもつことも多い生データを対象としたデータマイニングに共通する知識発見の手順を(1)～(6)に簡単に示す[1]．

データマイニング(5)
weblio　OR辞典(1)

概要

- データベースに蓄えられた多量のデータから，機械学習（machine learning）や統計的手法（statistical method）を用いてデータの中に含まれる知識を発掘する手法をいう．知識発見プロセスとしての，データ獲得，選択，前処理，変換，知識発見アルゴリズムの適用，解釈，評価といった一連のサイクルを指す．獲得した知識に基づく意思決定が目的であり，データ収集，発掘，評価といった人間と計算機の共同作業を伴う知識マネジメントとして捉えられる．

詳説

- **データマイニング** (data mining)は，データベース (data base) [5] に蓄えられた大量の生データに対して，機械学習 (machine learning)に関連する複数の手順を用いる戦略により，データに内在する規則性 (regularity)，制約 (constraint)，ルール (rule)などを効率よく求める研究である．なお，データベースからの知識発見 (KDD: knowledge discovery in databases)とも呼ばれ，知識発見 (knowledge discovery)に関わる多数の学習アルゴリズムが，人工知能だけでなくデータベースや統計学の側面を含めて研究されている．まず，ノイズや例外を含み疎な構造をもつことも多い生データを対象としたデータマイニングに共通する知識発見の手順を(1)～(6)に簡単に示す[1]．

データマイニング(6)
weblio　OR辞典(2)

【手順】

- (1)対象となるデータに対する既知の性質(背景知識)を利用してデータ収集を行い, データベースやデータウェアハウス (data warehouse)に格納する. (2)データに対する選択操作を前処理として行う. この段階はデータクリーニングと呼ばれる. (3)実装を前提とする制約のもとでデータの次元低減などによる変形操作を行う. (4)データマイニングを行うアルゴリズムを実行する. (5)導出された記述の解釈, ならびに, 記述の妥当性の検証を後処理として行う. (6)最終的な記述が評価され, 知識となる.
- 手順(4)のアルゴリズムで求まる知識の表現法によって, データの統計的解析 (statistical analysis of data)とデータの論理的解析 (logical analysis of data)の二種類に大きく分類される.
- データの論理的解析の一種である決定木 (decision tree)を図1に示す
. なお, 決定木を求めるアルゴリズムとしてID3 [4] などが知られており
, エントロピーやMDL(minimum description length)基準が記述を選択する際に用いられる.

データマイニング(6)
weblio　OR辞典(2)

【手順】

- (1)対象となるデータに対する既知の性質(背景知識)を利用してデータ収集を行い, データベースやデータウェアハウス (data warehouse)に格納する. (2)データに対する選択操作を前処理として行う. この段階はデータクリーニングと呼ばれる. (3)実装を前提とする制約のもとでデータの次元低減などによる変形操作を行う. (4)データマイニングを行うアルゴリズムを実行する. (5)導出された記述の解釈, ならびに, 記述の妥当性の検証を後処理として行う. (6)最終的な記述が評価され, 知識となる.
- 手順(4)のアルゴリズムで求まる知識の表現法によって, データの統計的解析 (statistical analysis of data)とデータの論理的解析 (logical analysis of data)の二種類に大きく分類される.
- データの論理的解析の一種である決定木 (decision tree)を図1に示す
. なお, 決定木を求めるアルゴリズムとしてID3 [4] などが知られており
, エントロピーやMDL(minimum description length)基準が記述を選択する際に用いられる.

データマイニング(7)
IBM東京基礎研究所

データ解析技術
Data Analytics

- IBMの研究部門が1993年に提案した、データベース内のアイテム集合に対する相関規則の発見手法は、マーケティングの分野に数理的手法を持ち込んだ最初の成功例のひとつでした。その後、この種のアプローチは、データマイニングと呼ばれるようになりました。IBMの研究部門は、この分野の創始以来、一貫して先導的な役割を果たしています。

  データマイニングは、機械学習、統計学、人工知能、オペレーションズリサーチ、といった分野と相互作用をしながら

  、実世界に現れるあらゆる形態のデータか ら意味のある知見を取り出す技術、いわば「知識発見工学」として、計算機科学の分野で大きな存在感を持つようになりました。

  IBM東京基礎研究所のデータ解析グループは、データマイニングにおけるIBM研究部門の伝統の一翼を担っています。我々は、1990年代後半には相関規則の発見手法の拡張において顕著な研究成果を挙げたほか、2000年代前半には構造データからの機械学習の分野において、先駆者としての名を歴史に刻んでいます。

  我々のグループの特色は、IBM研究部門の全世界的ネットワークを活用しながら、日本市場の実問題に即した課題に取り組んでいることです。例えば、日本の製造業の技術水準は明らかに世界最高レベルにありますが、そこで必要になってくるのはセンサーデータからの知識発見技術です。また、サービス業においても、日本市場が要求する品質水準は世界的にもおそらくトップクラスであり、精緻なビジネスデータ解析技術が必要とされています。

  2007年、IEEE ICDM (International Conference on Data Mining)という権威ある国際会議で催されたデータマイニングコンテストにおいて、我々のチームが挙げた成果は、この分野での研究活動と、世界の先端をいくお客様との協業プロジェクトを通して培われた実問題への深い洞察力とが相まった結果と言えるでしょう(詳細はこちらをご覧下さい)。

ビジュアルデータマイニング(1)
http://di-ver-si-ty.seesaa.net/article/8903201.html

- 現在、大量のデータが蓄積されています。
  しかし、そのなかから価値のあるものを見つけることは困難です。
  コンピュータにより自動的に価値のある情報を発見する、
  いわゆるデータマイニング技術はその解決法の1つですが、
  コンピュータにマイニング結果を提示されても、
  どうしてそうなったのか理解することが難しく、
  またマイニング結果が意味のないものであることはしばしばあります。

  情報可視化することにより、データマイニングの過程において、
  人間が直接入り込むことができます。
  コンピュータの膨大な記憶容量、高速な計算能力に、
  人間の（現在のコンピュータには真似できない）柔軟性、創造性、常識、推論能
  力、直感能力、また、視覚的認知能力を組み合わせることにより、
  それまで発見できなかったものを、発見できる可能性が生まれます。

  このコンピュータと人間の優れた部分を、情報可視化技術により統合し、
  マイニングすることを、ビジュアルデータマイニングと呼びます。

ビジュアルデータマイニング(2)
http://di-ver-si-ty.seesaa.net/article/8903201.html

- ビジュアルデータマイニングは、
  データに対してあまり知識がなく、マイニングのゴールがあいまいな時、
  よりその効果を発揮できます。
  人間とコンピュータが試行錯誤を繰り返しながらマイニングできるためです。
  また、そのマイニング結果は、
  人間が可視化結果を見て得られたものであり、直感的に理解できます。

  しかし、情報検索の場合には、
  通常のディレクトリ型検索エンジンの方が、
  可視化したものより、
  早く欲しいものを見つけられたという実験結果が出ています。
  ただ、上でも書いたように、目的地がはっきりしていない場合、
  つまり欲しい情報が明確に分からない場合は、有用です。

  ビジュアルデータマイニングプロセスは次の3つのステップで行われます。
  1．全体を見せ
  2．ズームやフィルタリングをして
  3．要求により詳細を見せる
  この3つをうまく合わせてやることが必要です。

ビジュアルデータマイニング(3)
http://di-ver-si-ty.seesaa.net/article/8903201.html

- 情報可視化技術は3つの基準で分類することができます。
  1．データのタイプ
  （1次元？2次元？多次元？時系列？グラフ？文書？）
  2．可視化法
  （2次元分布図？地図？アイコン？）
  3．インターフェースとゆがめ方
  （ズーム？射影？）
  データタイプや応用先によって
  適切な可視化法やインターフェースは
  もちろん異なります。

  通常の可視化研究はただ「〇〇データを△△法で可視化してみました」
  で終わっているものが多いです。
  どのようなデータでどのような目的の場合、
  どのような可視化法とインターフェースが適しているか、
  一般的に当てはまる体系をつくることが、まず必要だと思われます。

  さらなる発展のために、従来の機械学習や統計の技術と
  より密接に統合する必要があると考えられます。

ビジュアルデータマイニング(4)
http://di-ver-si-ty.seesaa.net/article/8903201.html

- また、ajaxやflashが普及し、普通の人でも可視化アプリケーションをつくり、Webでサービスを行うことができる環境ができてきました。
  まずサービスしてみて、多くの人に使ってもらい、
  そのなかで情報可視化技術が発展していくのが、一番の近道だと思います。

  さまざまな情報可視化技術は
  "Information Visualisation: Tutorial Notes"
  http://www.iicm.edu/ivis/ivis.pdf
  で紹介されています。

  日本語のWebサイトとしては
  私のブックマーク:情報可視化とデジタルミュージアム
  http://www.ai-gakkai.or.jp/jsai/journal/mybookmark/16-2.html
  Visualization
  http://pitecan.com/bib/Visualization.html
  などがあります。

目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

乱数表とコイン

（北川敏男先生が中心になり乱数表を作成したと聞いている。）

結構面白い

数理解析研究所講究録 1257 巻 2002 年 260-274

第二次世界大戦期に於ける日本人数学者の戦時研究

放射線式（昭和30年代，部品）

（やはり、最初は放射線。経年劣化の問題はどのように考えていたのか？）

放射線利用物理乱数
発生装置部品　1956年
FACOM128で利用

乱数発生装置（メインフレーム用:200kB/s）

（研究所に入った時はこれだった。午前と午後で出力に変化があった？）

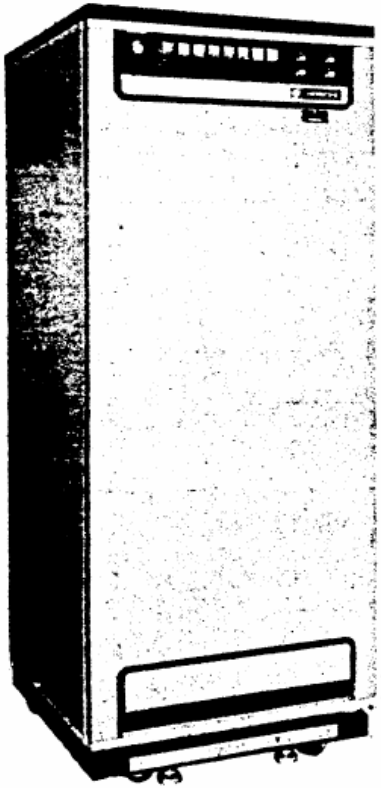


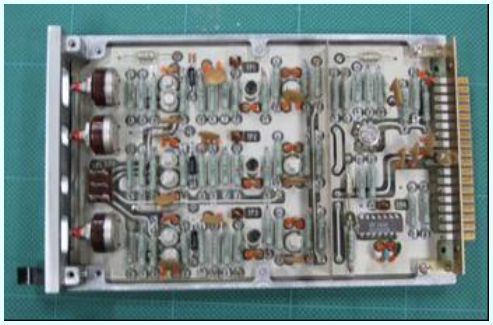
Figure 1. Exterior view of the random number generator

第3代物理乱数
発生機部品　1971年
発生速度　200KB/s
HITAC8300から
HITAC M280Hで利用

1963年にダイオード雑音の初代開発。(12,000ビット／秒）HIPAC103(パラメトロン)にインストールしたものがあり。

乱数発生装置（メインフレーム用：1.5MB/s）

（日立と共同開発した。物理乱数発生装置に関わった最初。1989年）

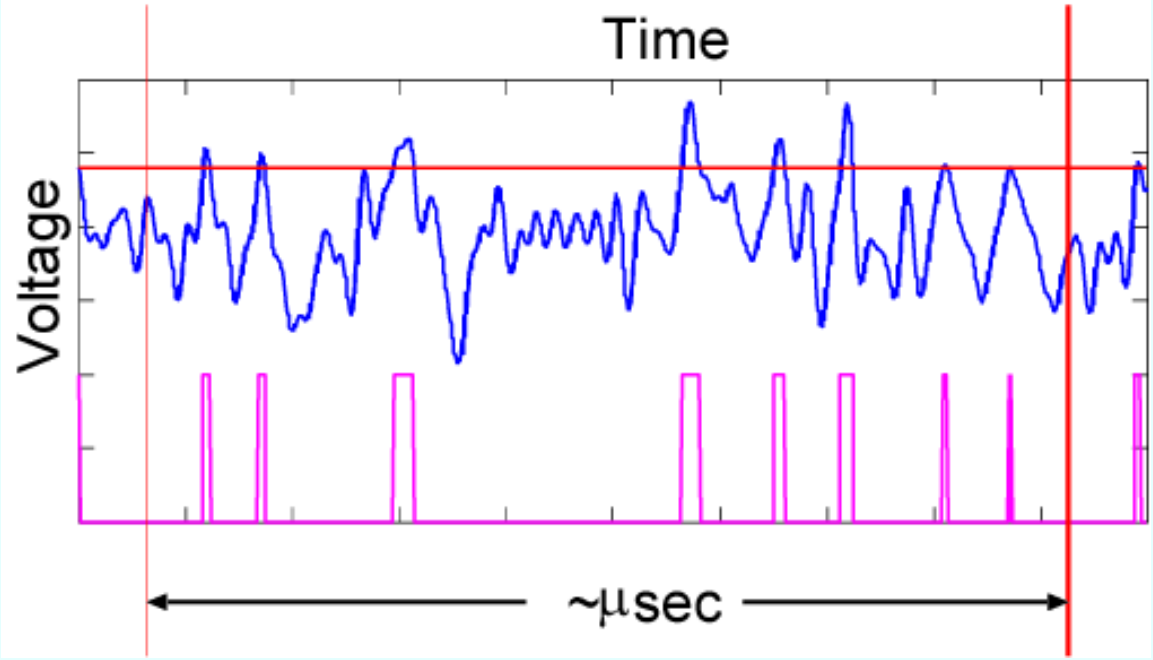


第4代物理乱数発生機部
ダイオード熱雑音　1989年
発生速度　1.5MB/s
M682H、M880/180、S3800/162
で利用

目次

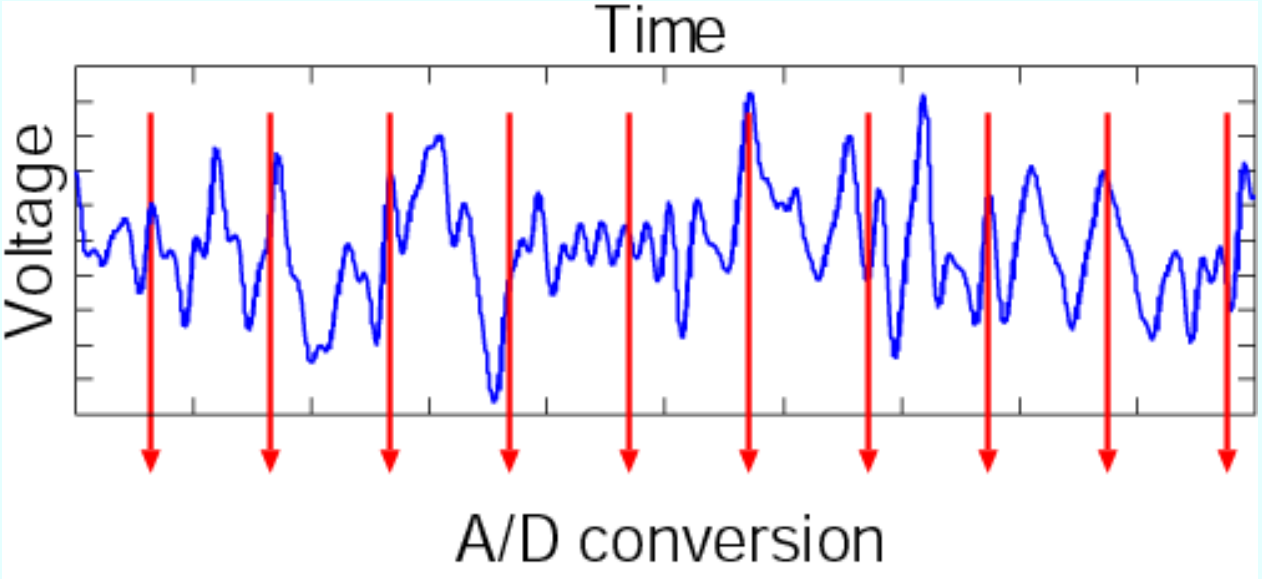
- 1997年以前の物理乱数発生装置
- 発生方法の見直し
- 1998年以降の物理乱数発生装置
- 融合研究による研究
- 今回導入した物理乱数発生装置
- 今後の発展及びまとめ

デジタル化の方法(旧)



- 信号をパルスに変換(=1bit A/D変換)
- 単位時間内のパルスをカウント
- 遅い 1.5Mbyte/sec/channel

デジタル化の方法(新)

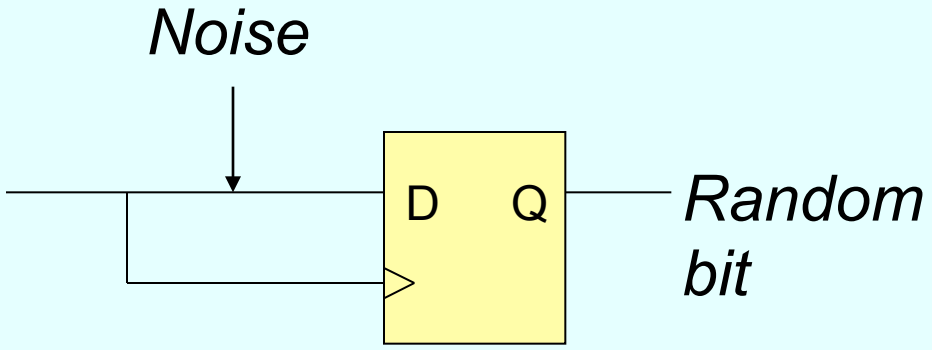


- 信号を直接デジタル化(8bit A/D変換)
- 10/100Mサンプル/秒
- サンプルあたり4bitを乱数として使えば、5～50Mbyte/秒

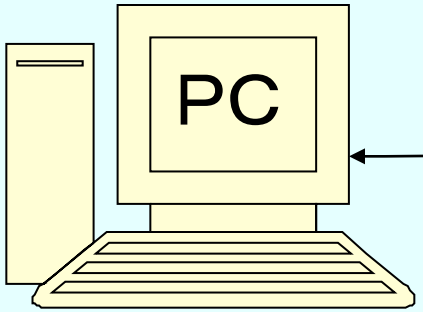
その他増幅法

- FDKの方法
- デジタル回路の閾値付近での敏感性

  =高い増幅度を利用
- 信号をそのまま増幅して1ビットA/D変換していることに相当
- 2つの入力での相対的な入力を見ており、閾値自体の系統的変動に強い



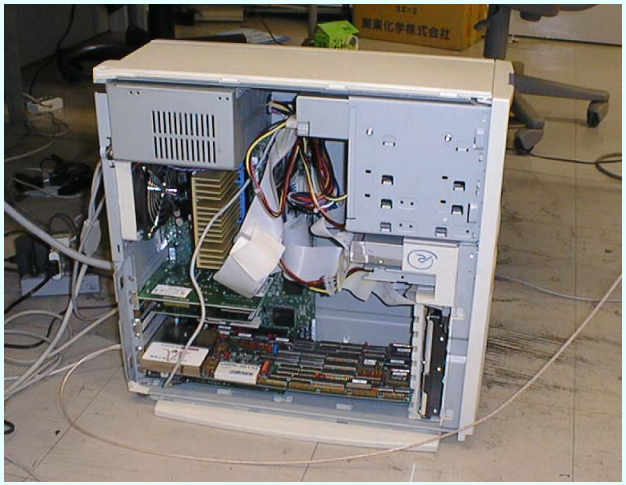
物理乱数発生システム



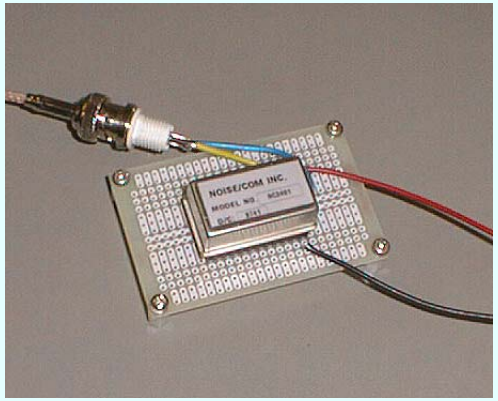
A/D変換
DAS-4100

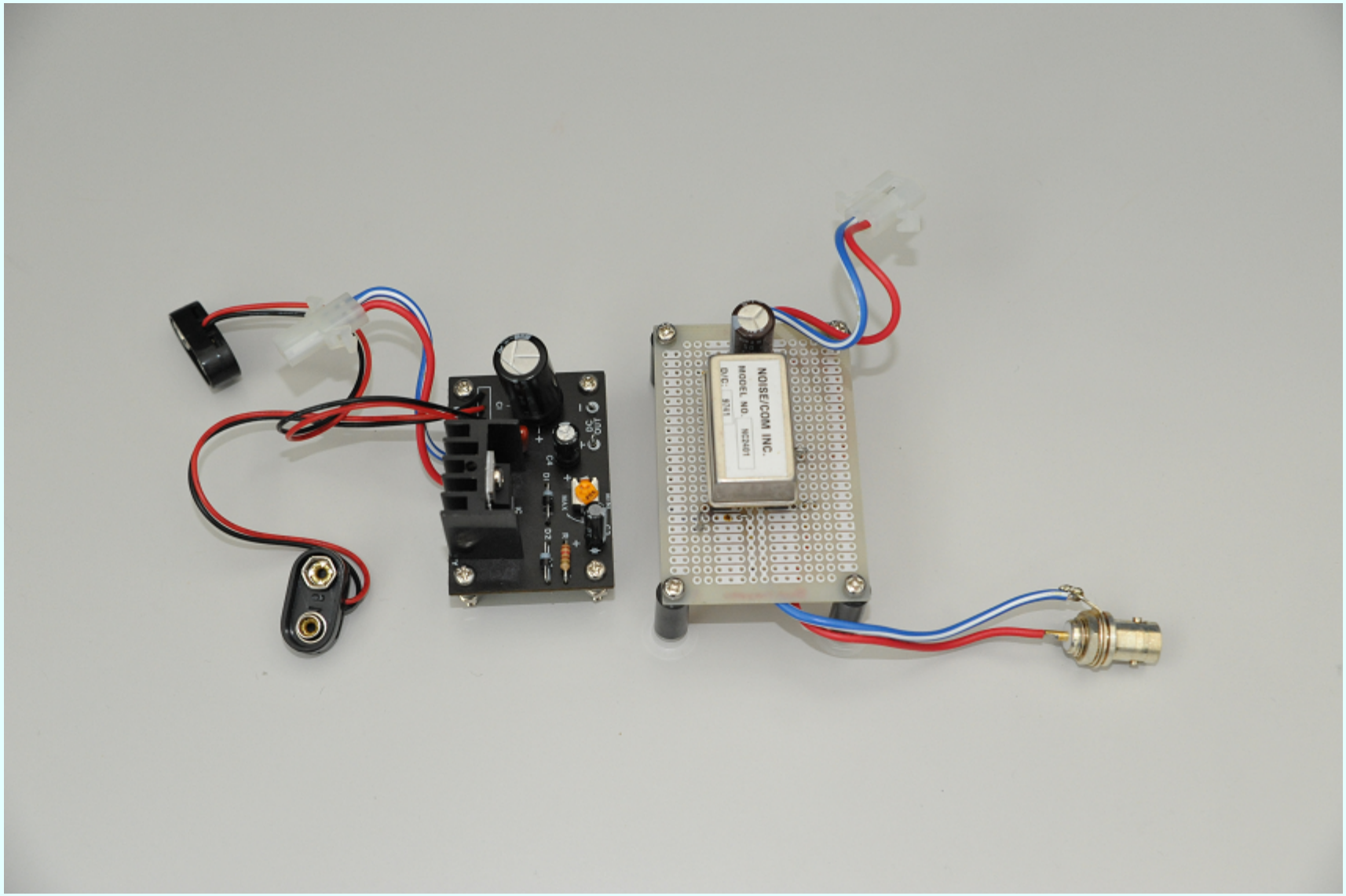
64Mbyte/sec 8bit

ノイズ源
NC2401



JISの
乱数表

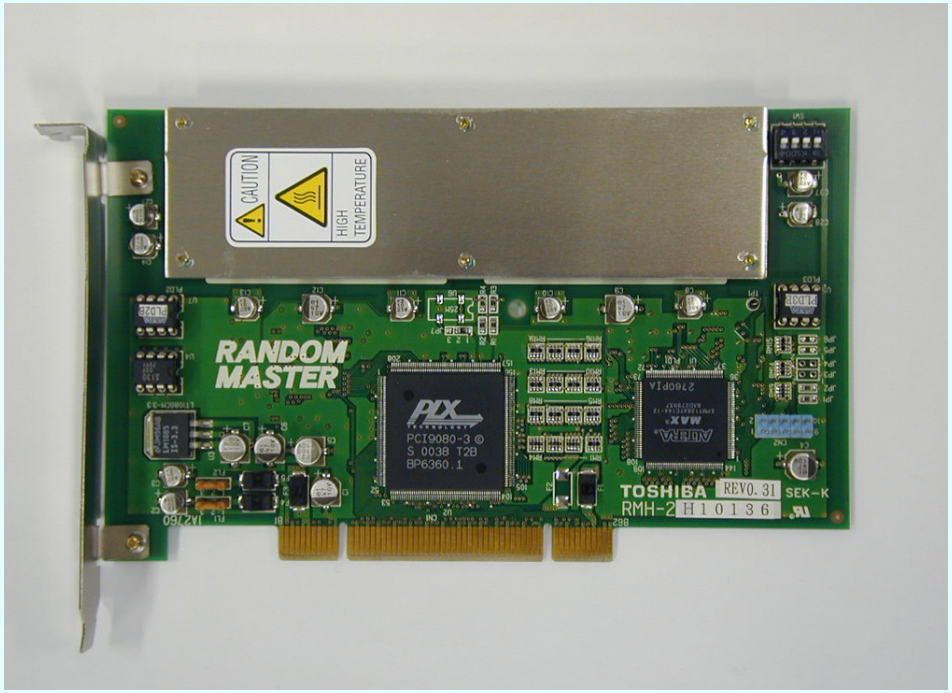



NOISE/COM INC.
MODEL NO.

目次

- 1997年以前の物理乱数発生装置
- 発生方法の見直し
- 1998年以降の物理乱数発生装置
- 融合研究による研究
- 今回導入した物理乱数発生装置
- 今後の発展及びまとめ

ハーフPCIボード（東芝：32MB/sec max）
2001年ブートストラップで利用



(1)乱数源はダイオードの熱雑音
(2)A/D変換＋一様分布にする
know-how
（特許3034516）

東芝の特許

【請求項1】 特許3034516

ノイズ信号を出力するノイズ源と、前記ノイズ信号を交流結合により直流分を除去しつつ増幅するAC結合増幅手段と、前記AC結合増幅手段により増幅された 増幅ノイズ信号を量子化するA／D変換手段と、前記A／D変換手段により変換されたディジタル値を構成する下位Nビットと、予め与えられたNビットの一様 乱数表より順次取り出された数値との排他的論理和を求めてNビットの乱数データを出力する補正処理手段とを具備することを特徴とする物理乱数発生装置。

PCIボード（日立：25MB/sec max）
1999年SR8000で利用

(1)乱数源はダイオードの熱雑音
(2)A/D変換＋一様分布にする
know-how
(特許第3680091号)

日立との特許

【請求項1】 特許3680091

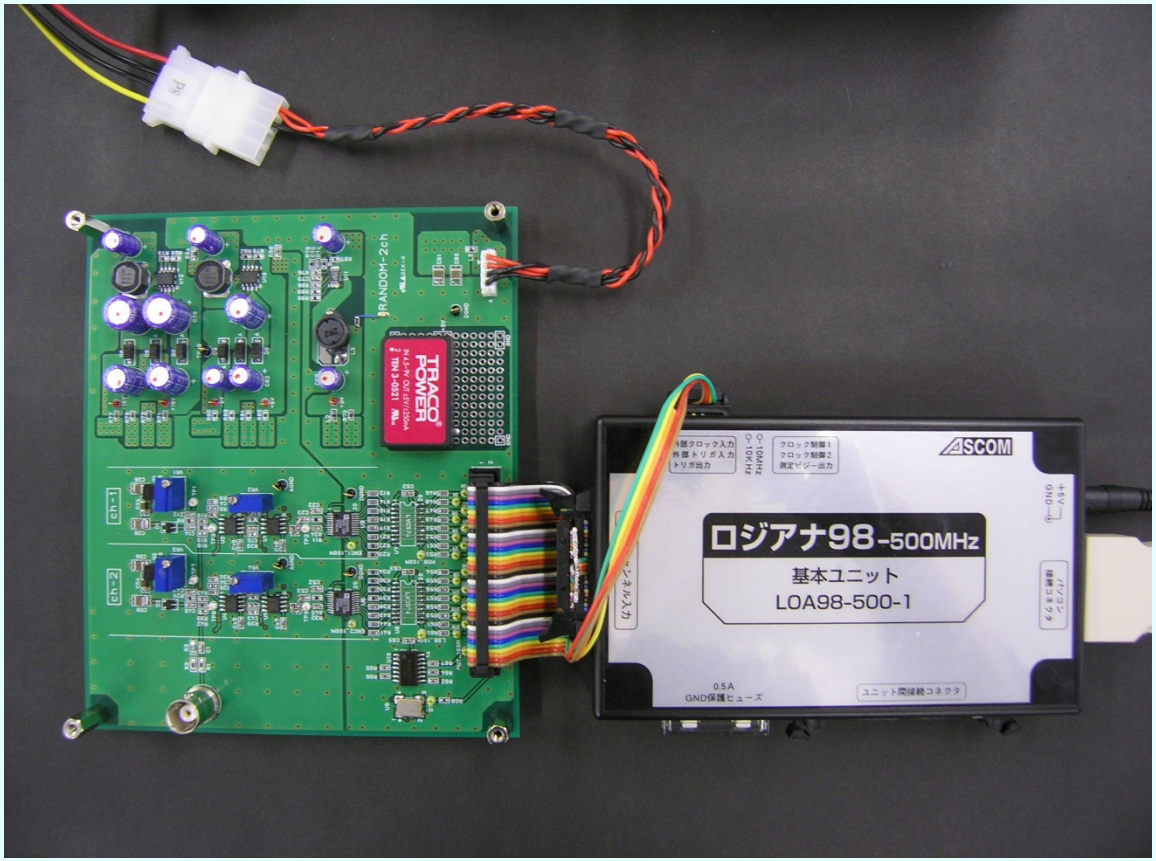
時間軸方向に規則性のない物理量を発生する物理量発生手段と、前記物理量を2ビット以上の第1のディジタルデータに変換するデータ変換手段と、前記第1の ディジタルデータの少なくとも一部を乱数データとして外部に出力する外部出力インタフェースと、を含む乱数生成装置であって、
0と1が不規則に変化しながら均等な比率で現れる反転トリガ信号を生成する反転トリガ信号生成手段と、
前記データ変換手段と前記外部出力インタフェースとの間に設けられ、前記第1のディジタルデータの少なくとも一部のビットと、前記反転トリガ信号との排他 的論理和をとり、前記第1のディジタルデータの代わりに第2のディジタルデータとして前記外部出力インタフェースに出力するデータ反転手段と、を含み、
前記反転トリガ信号生成手段は、任意の乱数発生手段から時系列に入力されるディジタルの乱数データと、自身の出力との排他的論理和をとり、前記反転トリガ信号として出力する反転トリガ更新論理と、前記反転トリガ更新論理から出力される前記反転トリガ信号をクロック信号に同期してラッチして前記データ反転手段および前記反転トリガ更新論理に出力する反転トリガ信号ラッチ手段とを含むことを特徴とする乱数生成装置。

PCI-Xボード（東芝：**133**MB/sec max）
2004年Altix3700で利用

目次

- 1997年以前の物理乱数発生装置
- 発生方法の見直し
- 1998年以降の物理乱数発生装置
- 融合研究による研究
- 今回導入した物理乱数発生装置
- 今後の発展及びまとめ

H17年度開発のボード(TED)
(発生方法の再検討)



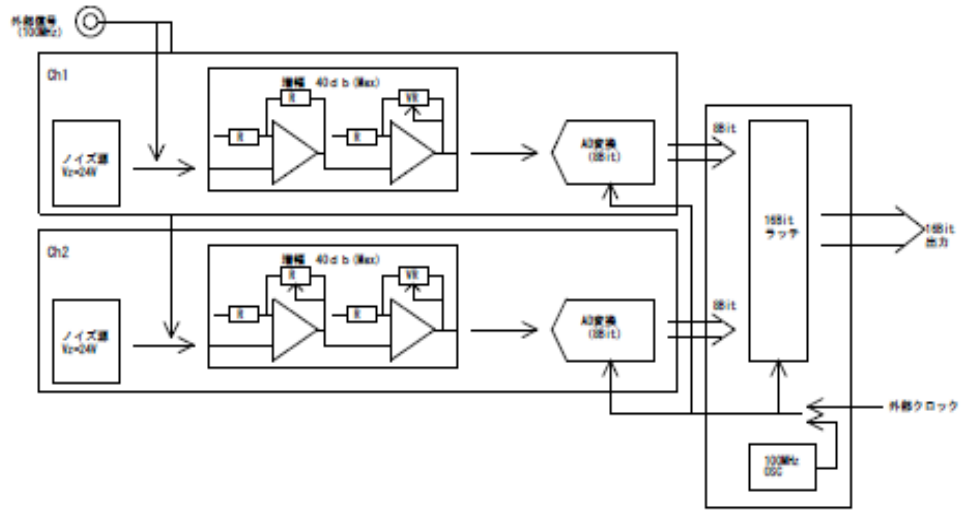
H17年度開発のボード(TED)の回路図

【構成】

・AD 変換ボード

ノイズ源＋増幅部＋AD 変換部を一つのチャンネルとし、２チャンネルを一枚の基板に搭載する。AD 変換の出力は 8bit とし２チャンネルを同一クロックでラッチし 40Pin のコネクタにデジタルデータとして出力する。

AD 変換ボードの構成図



H17年度開発のボード(TED)の仕様

【AD 変換ボードの各部の仕様】

・ノイズ源

ノイズ源：	定電圧ダイオード Vz=24V
電源：	定電流源（1mA　可変）

・増幅部

段数：	２段（固定増幅＋可変増幅）
増幅度：	20db～40db 可変

・AD 変換部

ビット数：	８Bit
最大周波数：	100MHz

・電源

電圧：	5V
コネクタ：	内臓ドライブ用電源コネクタ

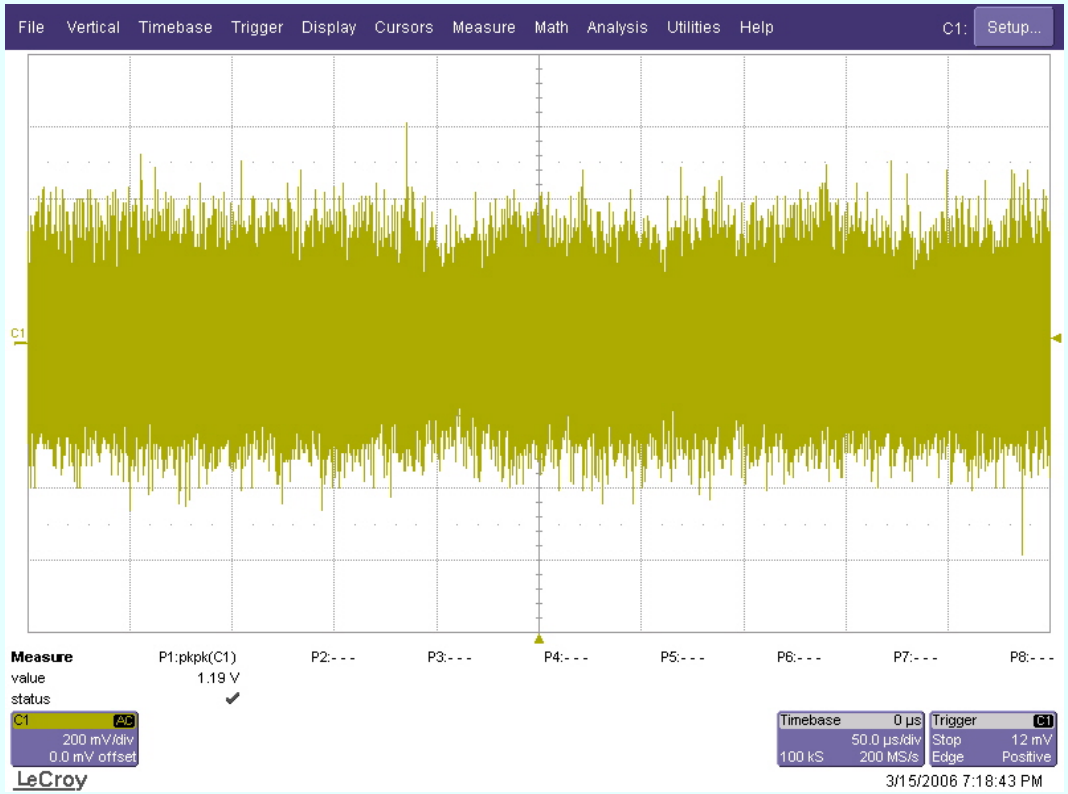
・基板サイズ

100mm×200mm

【納品物】

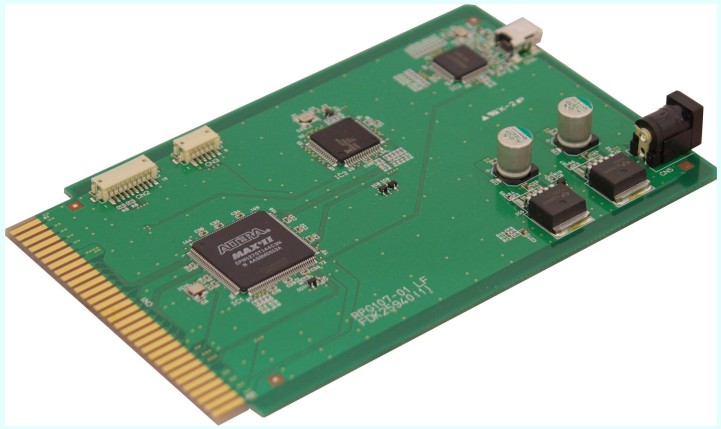
・AD 変換ボード：	基板単体　１枚
・計測器：	ロジアナ 98-500MHz（TPD）　１セット

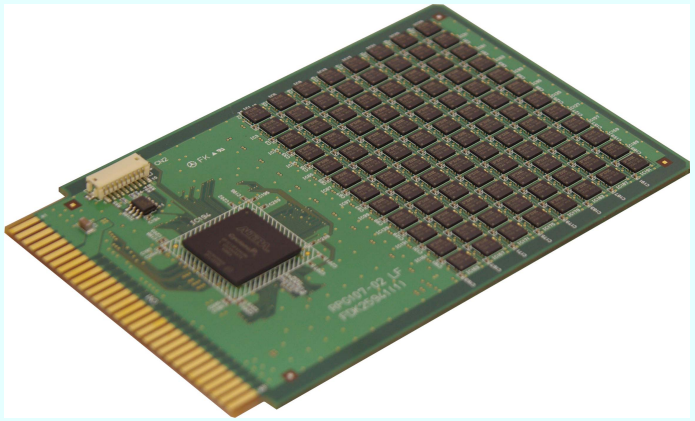
H17年度開発のボード(TED)の信号例



H17年度開発の小型装置(FDK)

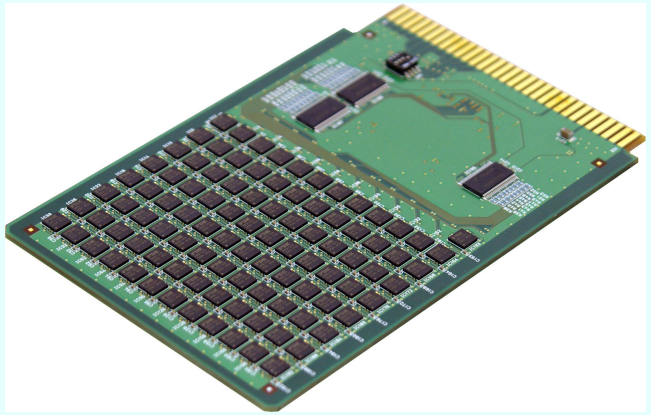
H17年度開発の小型装置(FDK)のボード





メインボード

サブボード(表)



サブボード(裏)

H17年度開発の小型装置(FDK)の仕様

ハードウェア仕様	
インターフェース	USB1.1／2.0準拠
ポートタイプ	USB Mini-Bタイプ
乱数源	真性乱数生成IC　RPG100B
平均転送速度	24MB/s
乱数品位	FIPS 140-2相当
電源電圧	付属AC-DCアダプタより給電
電源電流	未定
動作温度範囲	5から45℃
動作湿度範囲	20から80%(結露なきこと)
保存温度範囲	-10から50℃
保存湿度範囲	20から85%(結露なきこと)
外形寸法	160x70x100mm
質量	未定
ケース材質	アルミニウム

H17年度開発の小型装置(FDK)と他の製品との比較

既存品との比較　物理乱数発生装置(USB　Box型)仕様比較(2006年3月2日現在)

国名／会社名／URL	製品名	サイズ	発生方法	速度
スウェーデン／Protego http://www.protego.se/index.htm	R210	32.3x18.6x72.2mm	熱雑音	0.08Mbps
	R220	32.3x18.6x72.2mm	熱雑音	0.39Mbps
	R230	32.3x18.6x72.2mm	熱雑音	2.04Mbps
米国／Comscire http://comscire.com/Home/	QNG Model J1000KU	不明	熱雑音	1Mbps
ハンガリー／Random http://random.com.hr/index.html	HG416 2006年1月10日	25 x 60 x 80 mm	熱雑音	16Mbps
	HG424	25 x 60 x 80 mm	熱雑音	24Mbps
	HG432	25 x 60 x 80 mm	熱雑音	32Mbps
ドイツ／True-Random http://www.true-random.com/	rw3usb	8 x 5,5 x 2,5 cm	熱雑音	1.5Mbps
ハンガリー Mario Stipcevic, Ph. D. http://qrbg.irb.hr/	QRBG121	55 x 65 x 90 mm	LEDの量子効果	12Mbps
ドイツ／WESTPHAL ELECTRONIC http://www.westphal-electronic.com/	ZRANDOM USB	15X15X3cmくらい (写真からの推測)	熱雑音	40kbps (生モード) 220kbps (XORモード)
	PRG400	115x80x35mm	熱雑音	115kbps (vonNeumann) 800kbp (raw mode)
スイス／GAP http://www.gapoptic.unige.ch/Prototypes/QRNG/default.asp	Quantum Random Numbers Generator	68 x 150 x 188 mm	光ビームスプリット	100kbps
日本／Zixsys http://www.zixsys.com/index.html	RPG001	35x42x12mm	微弱α粒子	記載なし
日本／FDK http://www.fdk.co.jp/	Random Streamer 24M (共同開発中)	160x70X100mm	熱雑音	24MB/s

H19年度開発の装置(TED)
(アナログ部のみ　発生方法の再検討)

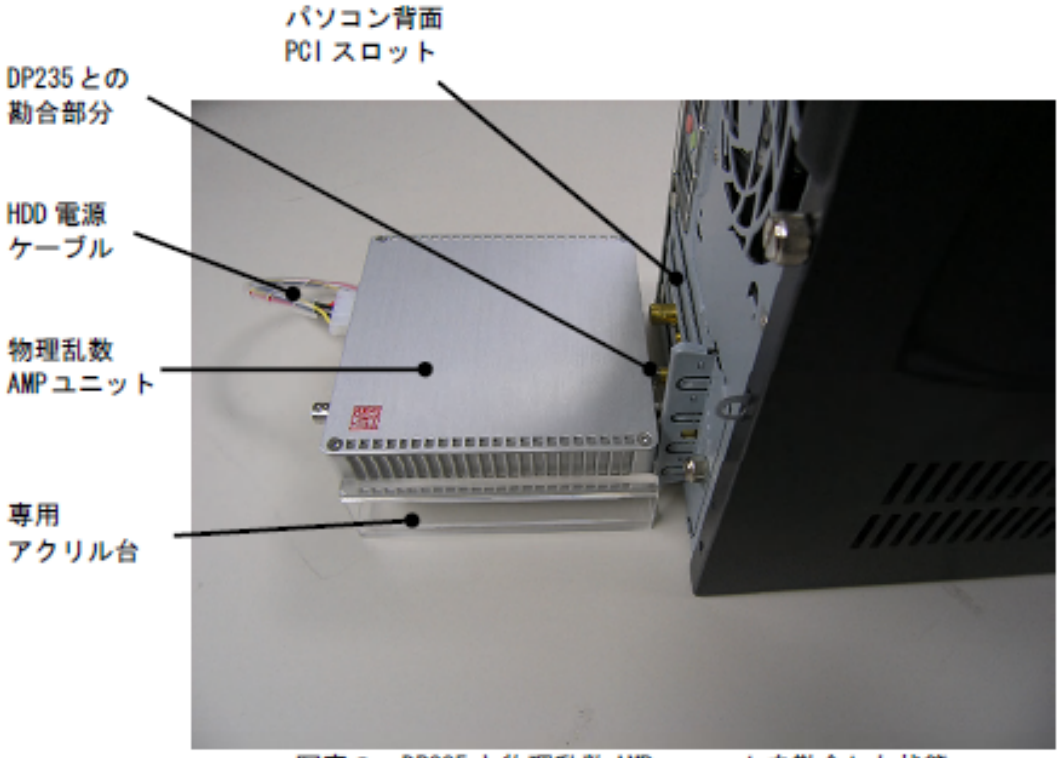
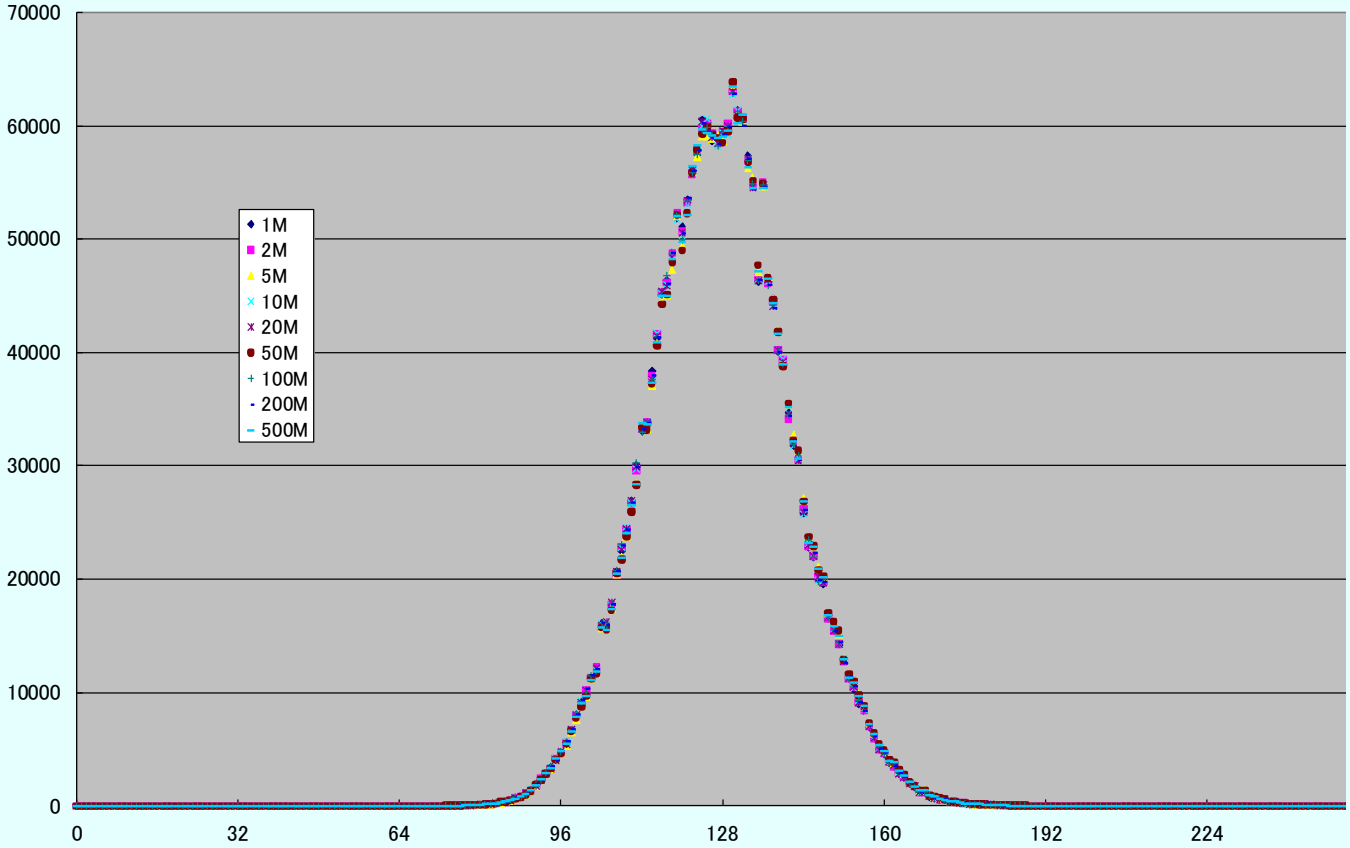
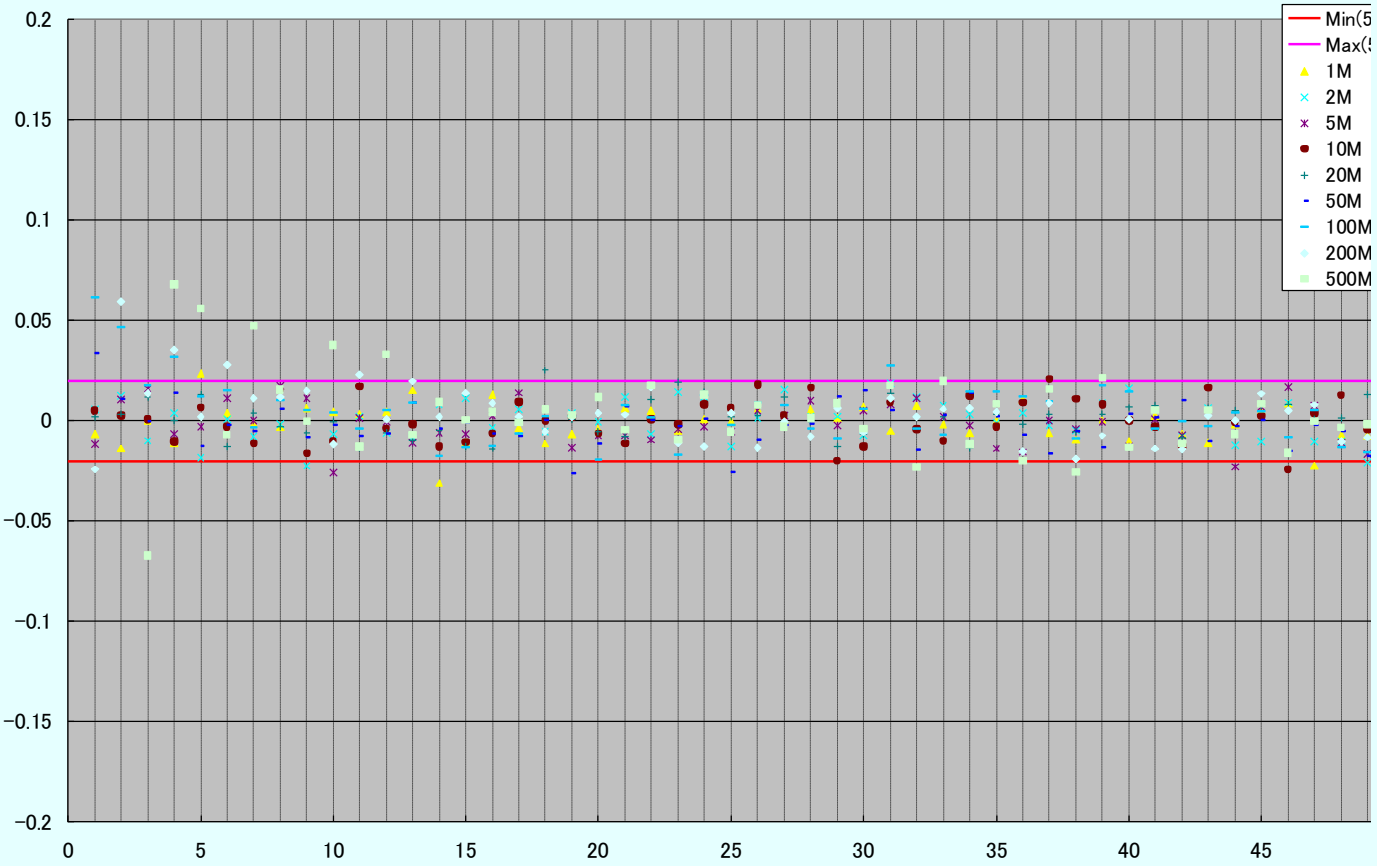


写真２　DP235 と物理乱数 AMP ユニットを勘合した状態

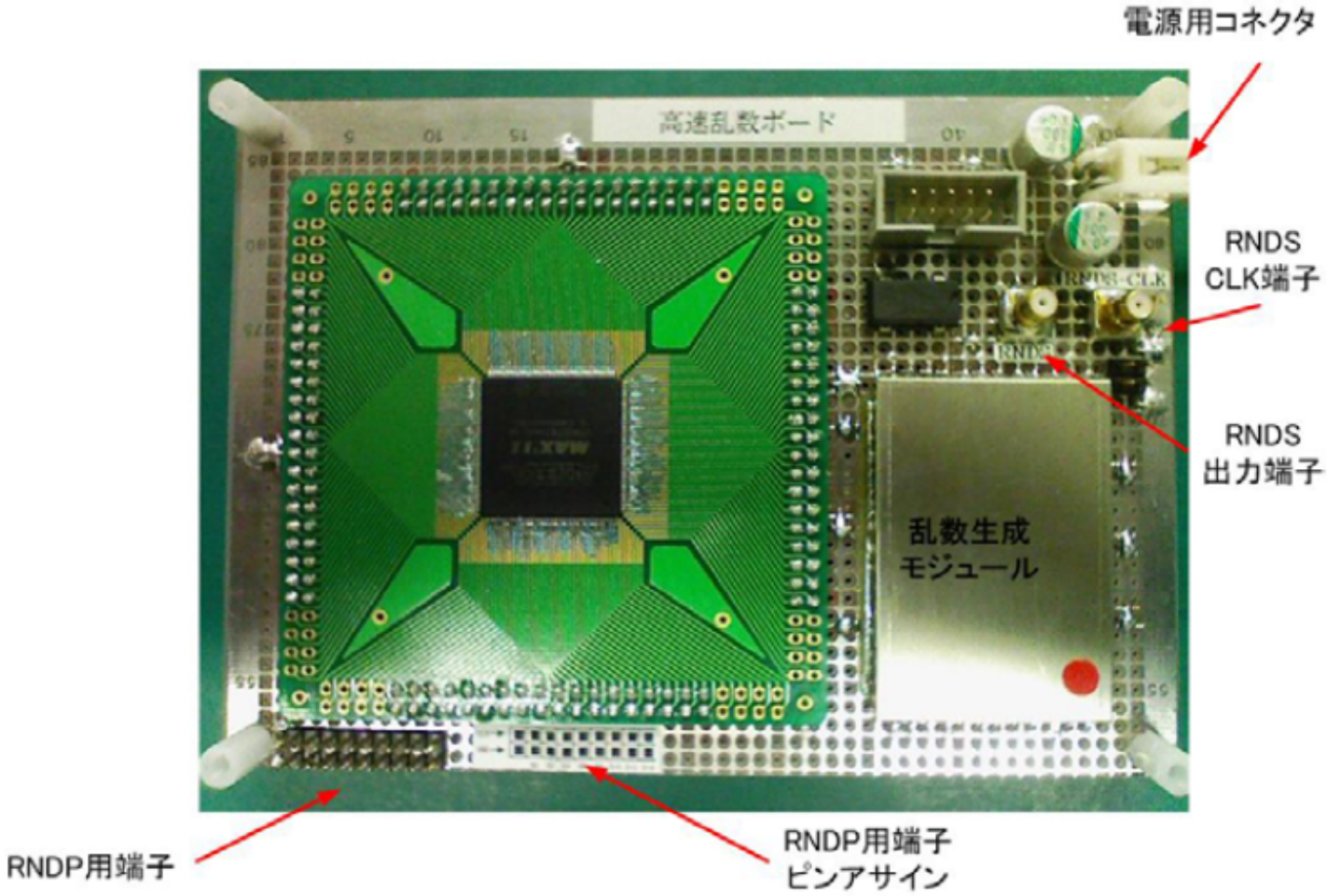
H19年度開発の装置(TED)
ヒストグラム



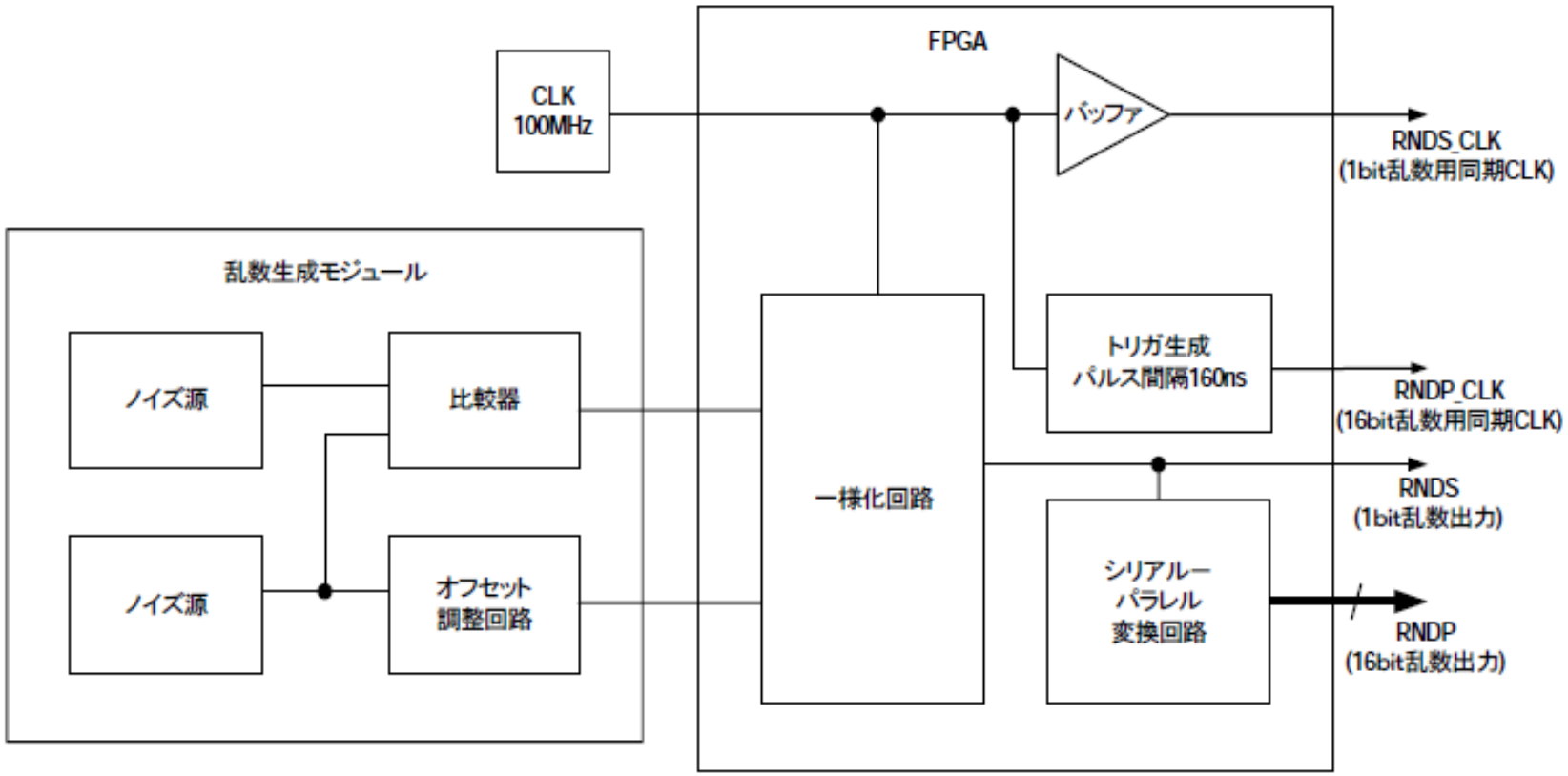
H19年度開発の装置(TED)
自己相関



H19年度にFDKと共同研究したボード



H19年度にFDKと共同研究したボード　ブロック図



H19年度にFDKと共同研究したボードに関係した特許

FDKとの特許

【請求項1】 特許公開 2008－165500

2組のノイズ源と、
これらのノイズ源より発生する各々ノイズの直流分を除去して増幅する第1増幅回路と第2増幅回路と、
前記第1増幅回路と前記第2増幅回路の各出力電圧を比較して0または1を出力する比較回路と、
当該比較回路の出力を基準クロックに同期して保持し、物理乱数を得るフリップ・フロップと、
前記比較回路の一方の入力に接続されて、当該比較回路のオフセット電圧を設定するオフセット電圧回路と、
前記オフセット電圧回路を制御する一様化フィードバック回路とで構成されることを特徴とする物理乱数発生器。

その他の物理乱数発生装置(1)

内田淳史(埼玉大学) et. al.

半導体レーザカオスを用いて乱数生成

- 数10Gbpsを超える速度も可能
- NISTの検定は通っている
- 現時点ではオンライン利用は不可
- 現時点の装置は大きいが小型化も可能

その他の物理乱数発生装置(2)

福島章雄（産総研） et. al.

磁気抵抗素子 の スピン注入磁化反転を用いて発生

基板サイズ 65 x 90 mm

生成レートは 8 kbit/sec

USBポートに差すだけで動作

目次

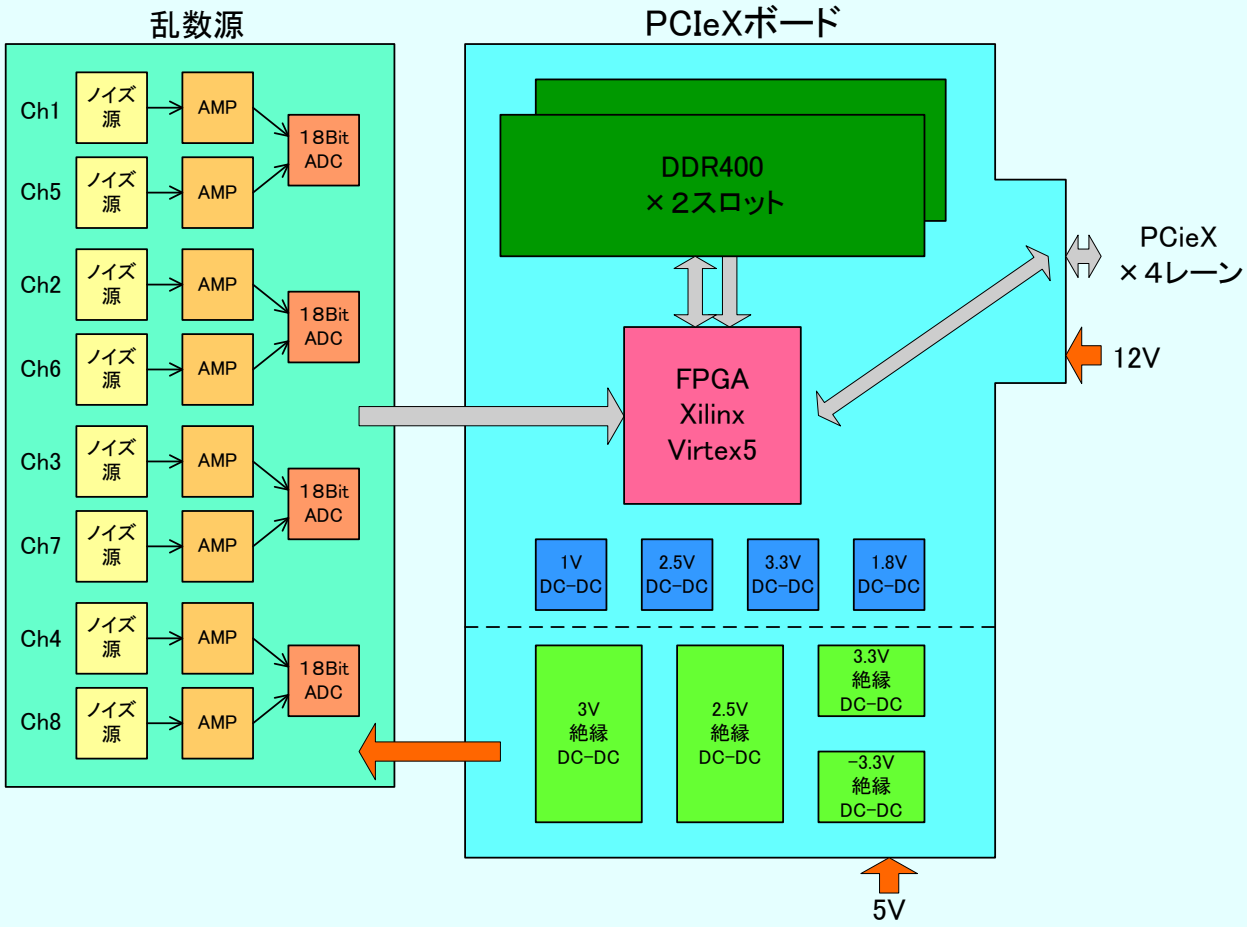
- 1997年以前の物理乱数発生装置
- 発生方法の見直し
- 1998年以降の物理乱数発生装置
- 融合研究による研究
- 今回導入した物理乱数発生装置
- 今後の発展及びまとめ

TED製物理乱数ボードの特徴

- 乱数源にツェナーダイオードを使用
  - ・ツェナーダイオードの雪崩降伏現象により発生するノイズを利用する事で コストを軽減
- 乱数源とデジタル回路を分離
  - ・乱数源（アナログ）をデジタル（FPGA）とは別のユニットにし、 電源回路も分ける事で、デジタルノイズの影響を軽減
  - ・シールドケースで外来ノイズの影響を軽減
  - ・乱数源をユニット化しているため、乱数源のみ別途開発して交換が可能
- 400MB/sの速度で乱数を発生
  - ・乱数源を4ch(+4ch)実装、100MspsでAD変換する事で、400MB/sを実現
  - ・大容量高速FPGAを使用
- PCIeX＋2GByteのバッファメモリ
  - ・PCIeX×4レーンのインタフェースと、1GByte×2のバッファメモリにより 乱数データを高速で転送
- FPGAにて柔軟な回路を構成
  - ・01補正処理の有無、マトリックス回路の設定等をレジスタで設定が可能

TED製物理乱数ボードの全体の構成

- 全体構成図



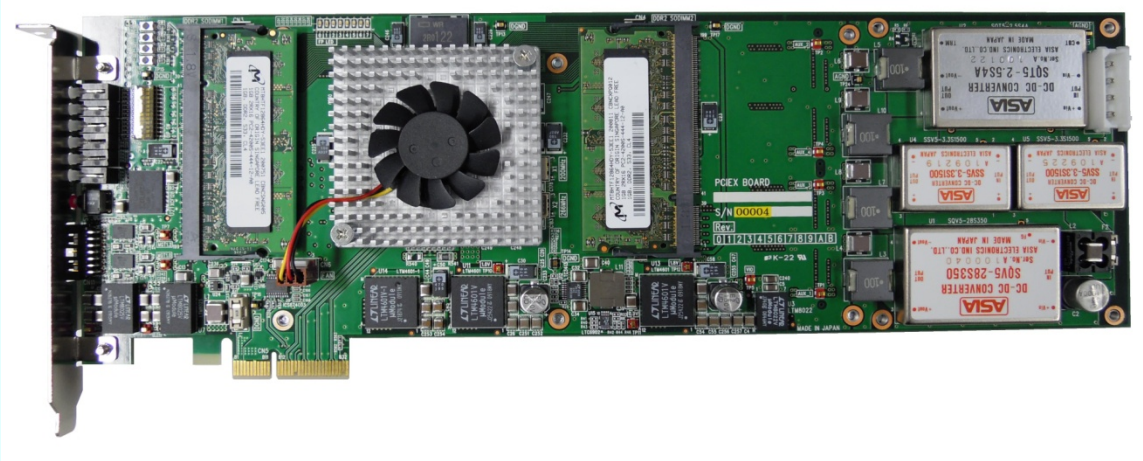
TED製乱数ボードの緒元

- 乱数源ユニット
  - ツェナーダイオード：Vz=24V
  - Ch数：8cH
  - AMP増幅度：45dB
  - AMP帯域：1MHz～50MHz
  - ADC:16Bit/Vin=1Vp-p/125Msps
  - サンプリング速度：100Msps
  - ユニットサイズ：178×98.4×12.7

- PCIeXボード
  - インタフェース：PCIeX ×4レーン
  - FPGA：Virtex5(Xilinx) FX130T-2以上
  - バッファメモリ：DDR2-400 1GB×2スロット
  - 基板サイズ：111.15×312 (PCIeX Add-in Card Standard Height, full length)
- 全体
  - 転送速度：412MB/s（DELL 690にて、基板一枚で、2GBを転送した場合）
  - 総重量：1300g
  - 占有スロット数：2スロット

TED製物理乱数ボードの外観

- 部品面

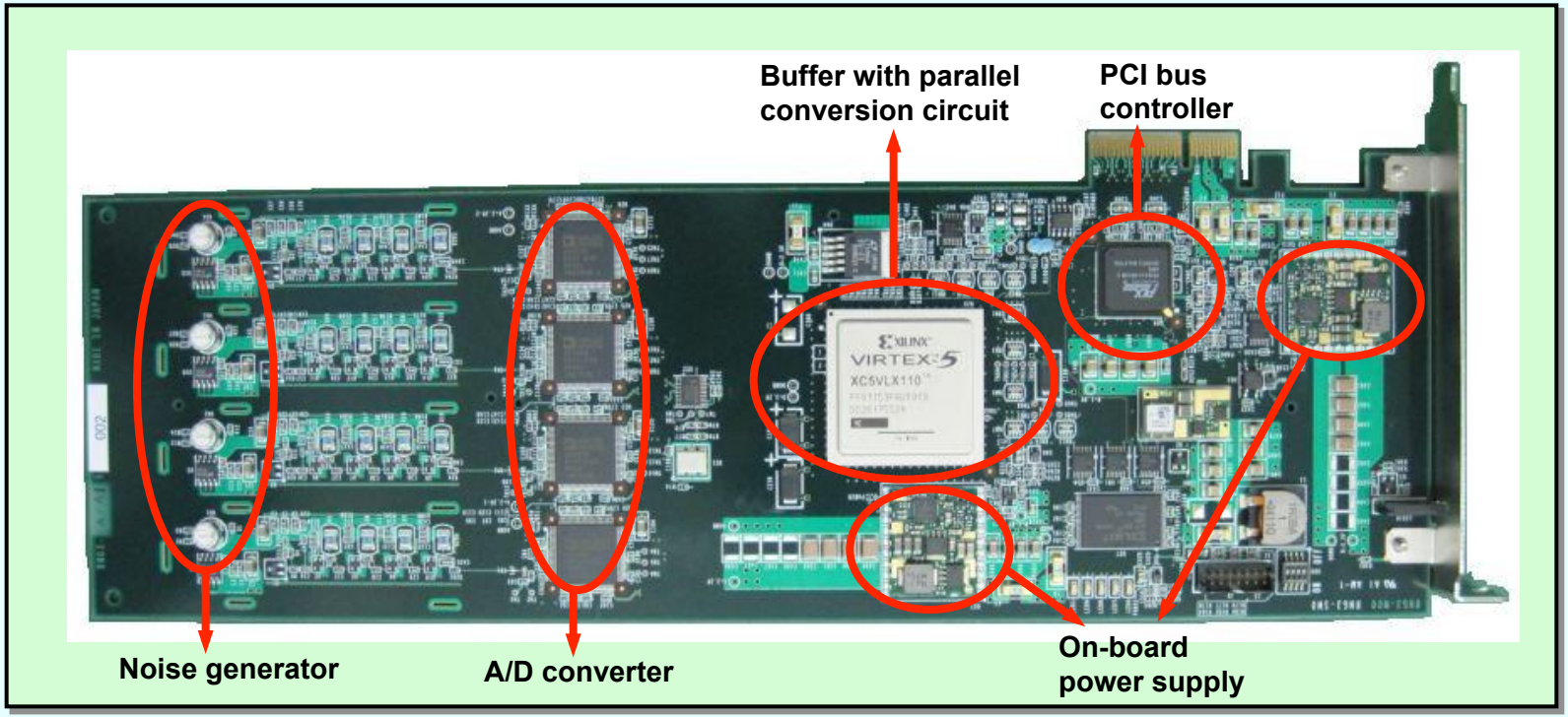


- 上部

日立製物理乱数ボードの特徴

- 乱数発生ノイズ源
  - 白色ノイズ（熱雑音）
- デジタル化法
  - ADコンバータ
- 発生速度
  - 200MB/sを実現
- インタフェース
  - PCIeX×4レーン（Gen2)
- 一様化方法
  - 積算反転方式

日立製物理乱数ボードの外観

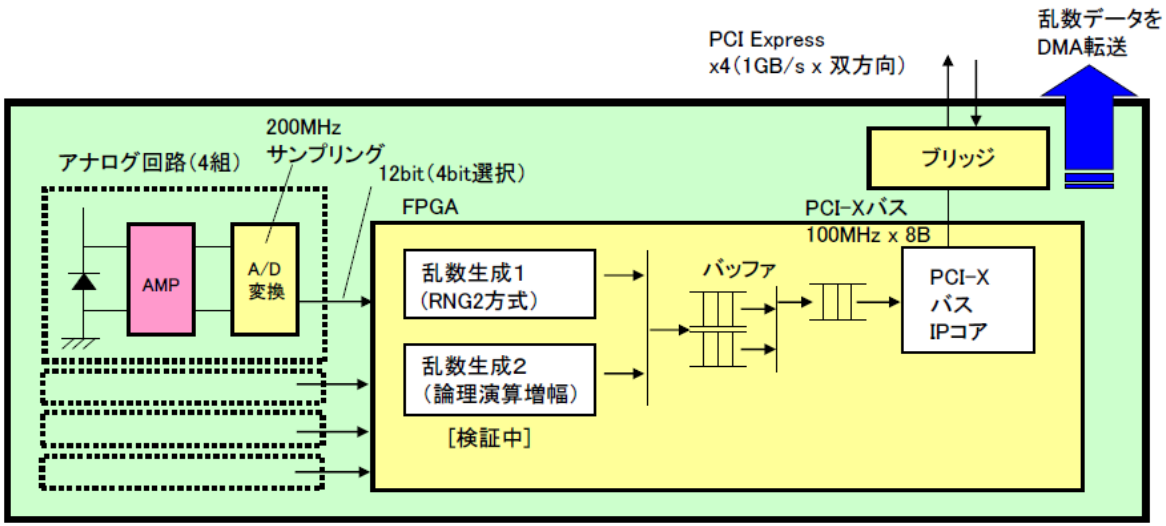


- 乱数カード・ホストI/F: PCI-Exp 4x
- 物理乱数発生源: 白色ノイズ(熱雑音)
- 理論乱数生成性能 : 200MB/s/カード, 800MB/s/ノード

日立製物理乱数ボードのブロック図

新(第三世代)物理乱数カードの論理構成図

HITACHI
Inspire the Next



- ダイオード電圧の熱雑音をA/Dコンバータでサンプリング
- 独立した4系統の生成源を統合し乱数生成
- 新規の乱数生成方式(スループット増強)を試行、検証中

uVALUE

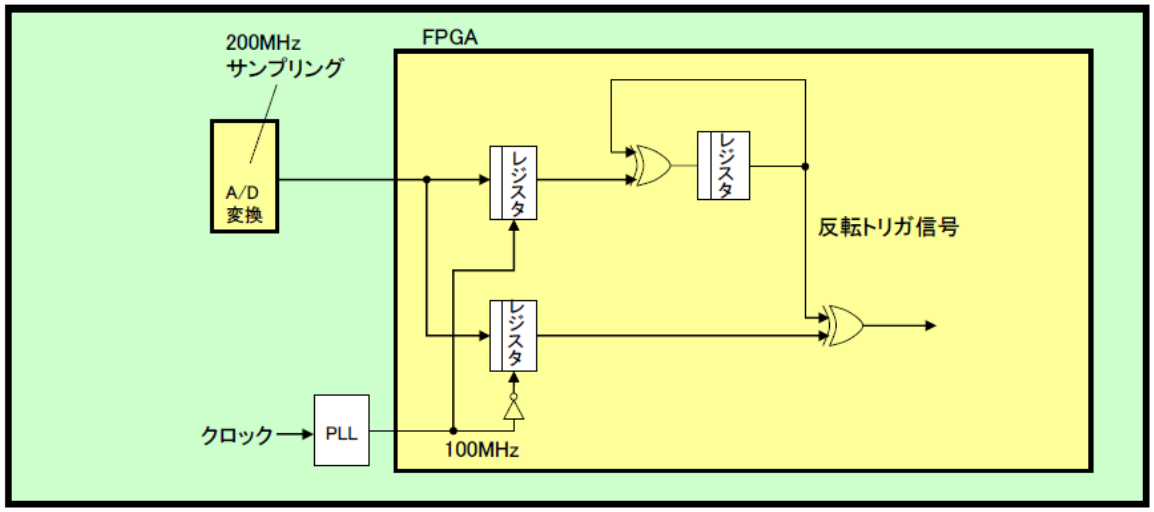
Hitachi Confidential

© Hitachi, Ltd. 2009. All rights reserved.

日立製物理乱数ボードの一様化法

積算反転論理構成図

HITACHI
Inspire the Next



- A/Dコンバータの出力を2系列に分割
- 1系列から0と1が等確率で出現する反転トリガ信号生成
- 残り1系列の信号を反転トリガ信号により反転

uVALUE
Hitachi Confidential
© Hitachi, Ltd. 2010. All rights reserved.

日立製物理乱数ボードの検定結果

評価結果

HITACHI
Inspire the Next

	成功数	一様性
1次元度数検定	合格	合格
8次元度数検定	合格	合格
連検定	合格	合格
累積和検定	合格	合格
ユニバーサル統計量検定	合格	合格
近似エントロピー検定 (m=8)	合格	合格
テンプレートマッチ検定	合格	合格

- 表は10,000,000ビット乱数1,000個に対して検定を行った結果
- 成功数：1,000個中、p_value<0.01だった試行が1～19回なら合格
- 一様性：1,000個のp_valueの分布を検定、結果のp_value<0.01なら合格

uVALUE

Hitachi Confidential

© Hitachi, Ltd. 2010. All rights reserved.

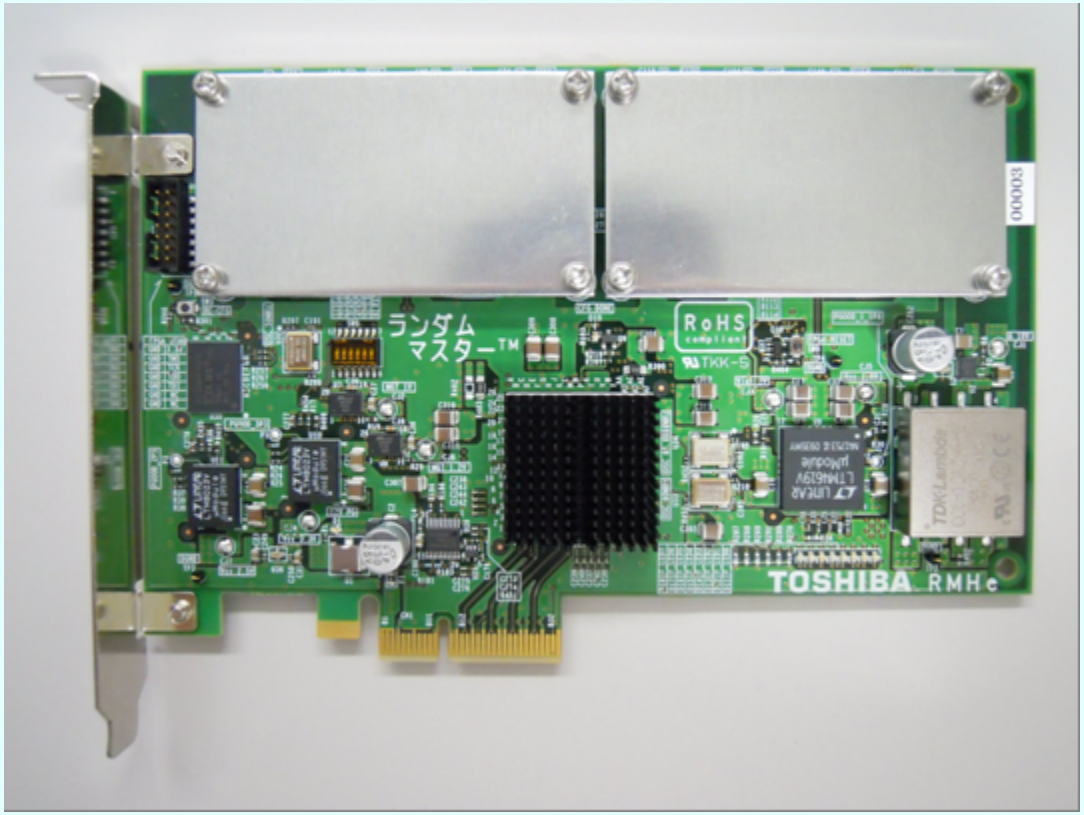
0

東芝製物理乱数ボードの特徴

- 乱数発生ノイズ源
  - ・ダイオードのノイズ
- デジタル化法
  - ・ADコンバータ
- 発生速度
  - ・640MB/sを実現
- インタフェース
  - ・PCIeX（Gen1)
- 一様化方法
  - ・ADCの下位2ビットを利用。排他的論理和による補正

From Toshiba 1/2

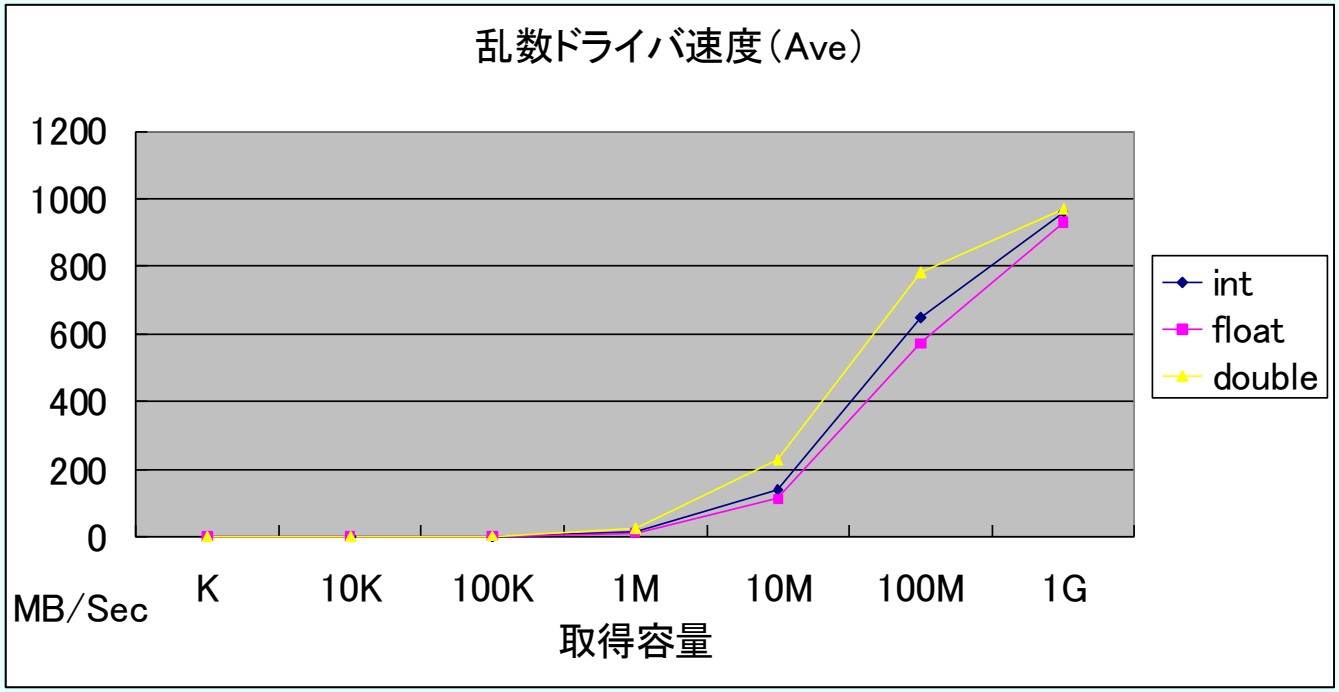
東芝製物理乱数ボードの外観



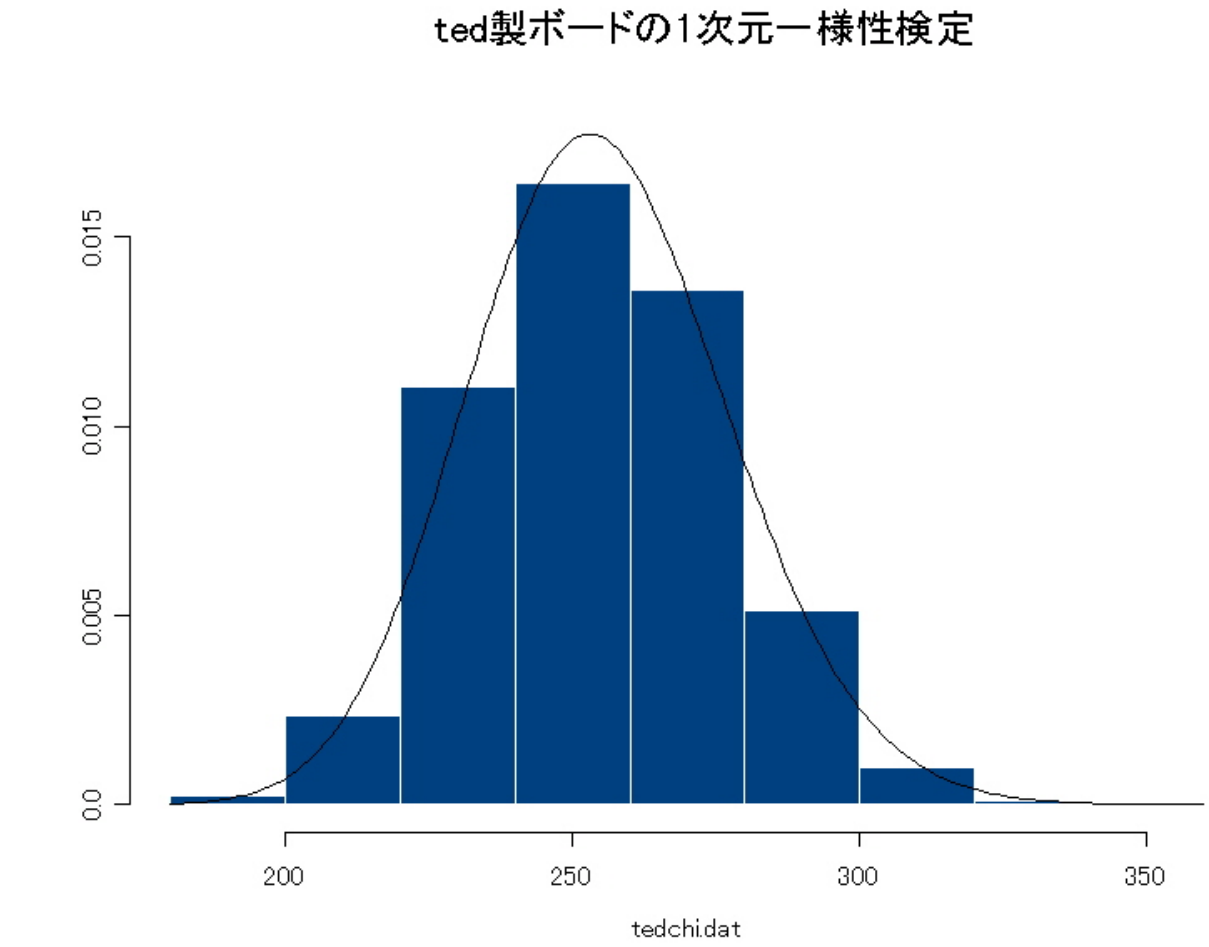
東芝製物理乱数ボードの速度テスト

AVE

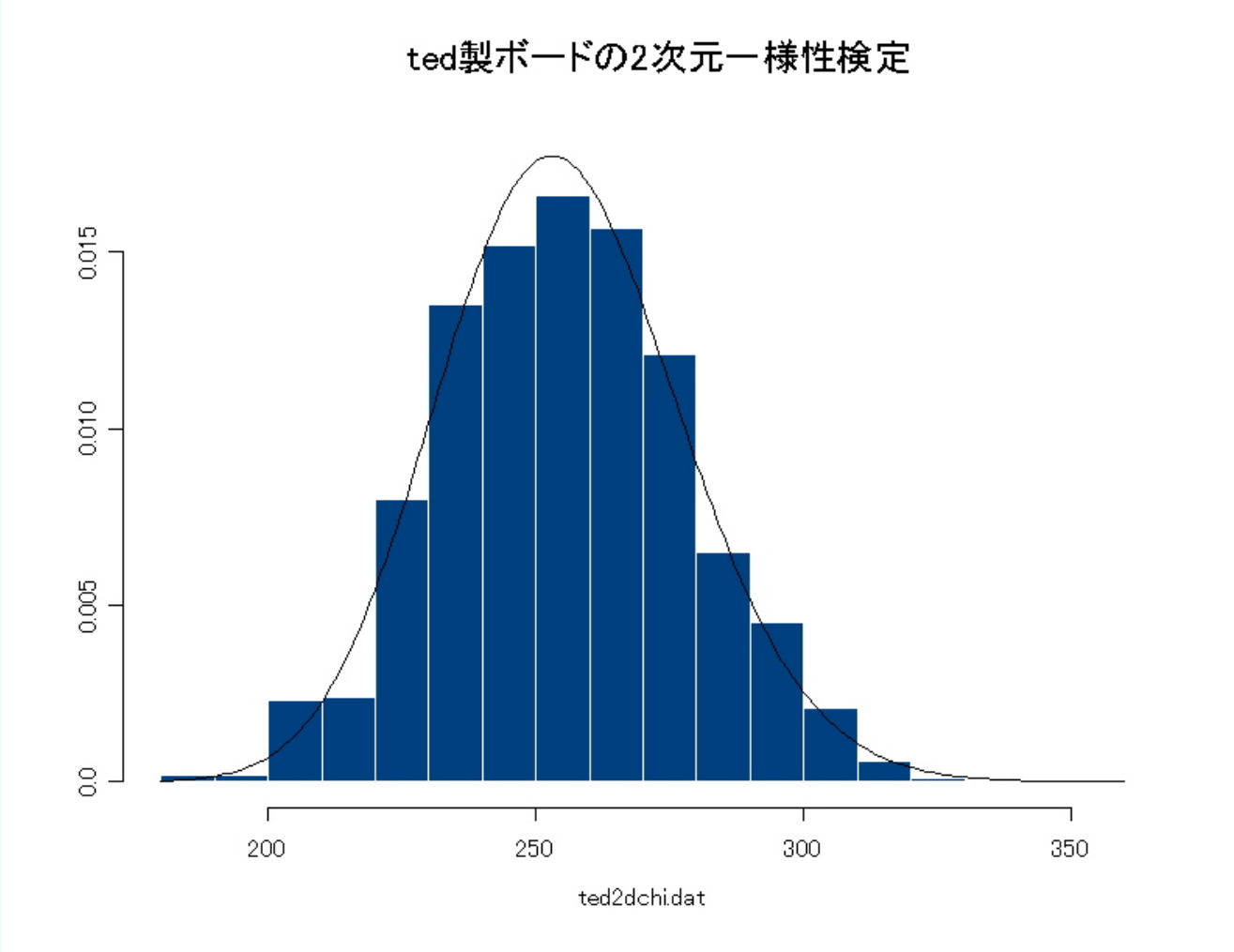
取得容量	K	10K	100K	1M	10M	100M	1G
int	0.015111	0.153967	1.50806	15.340399	137.855734	649.0476642	960.1546933
float	0.012543	0.127686	1.242016	12.145653	112.6769771	572.0063996	929.6535819
double	0.026982	0.2695	2.715492	26.758782	228.6656806	783.1205234	971.2566447



TED製ボードの検定結果(1)



TED製ボードの検定結果(2)

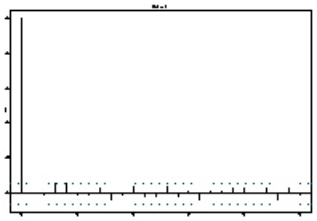
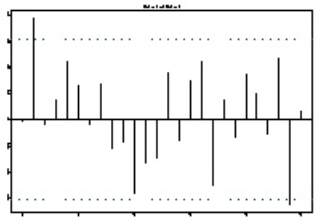
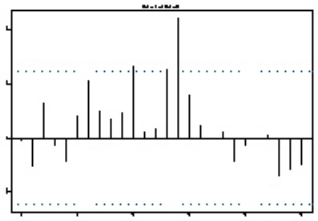
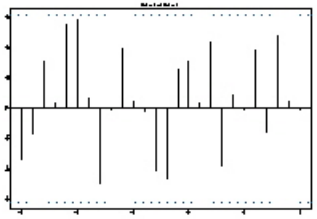
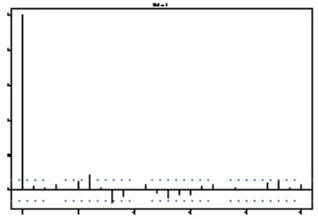
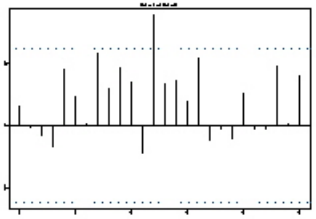
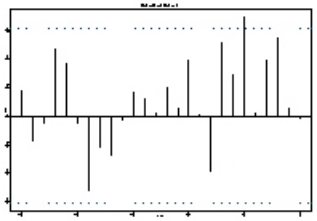
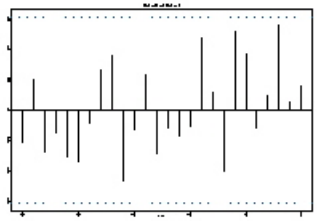
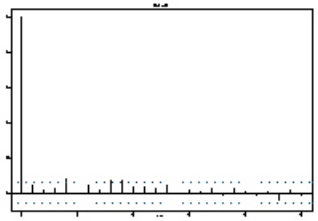


TED製ボードの検定結果(3)

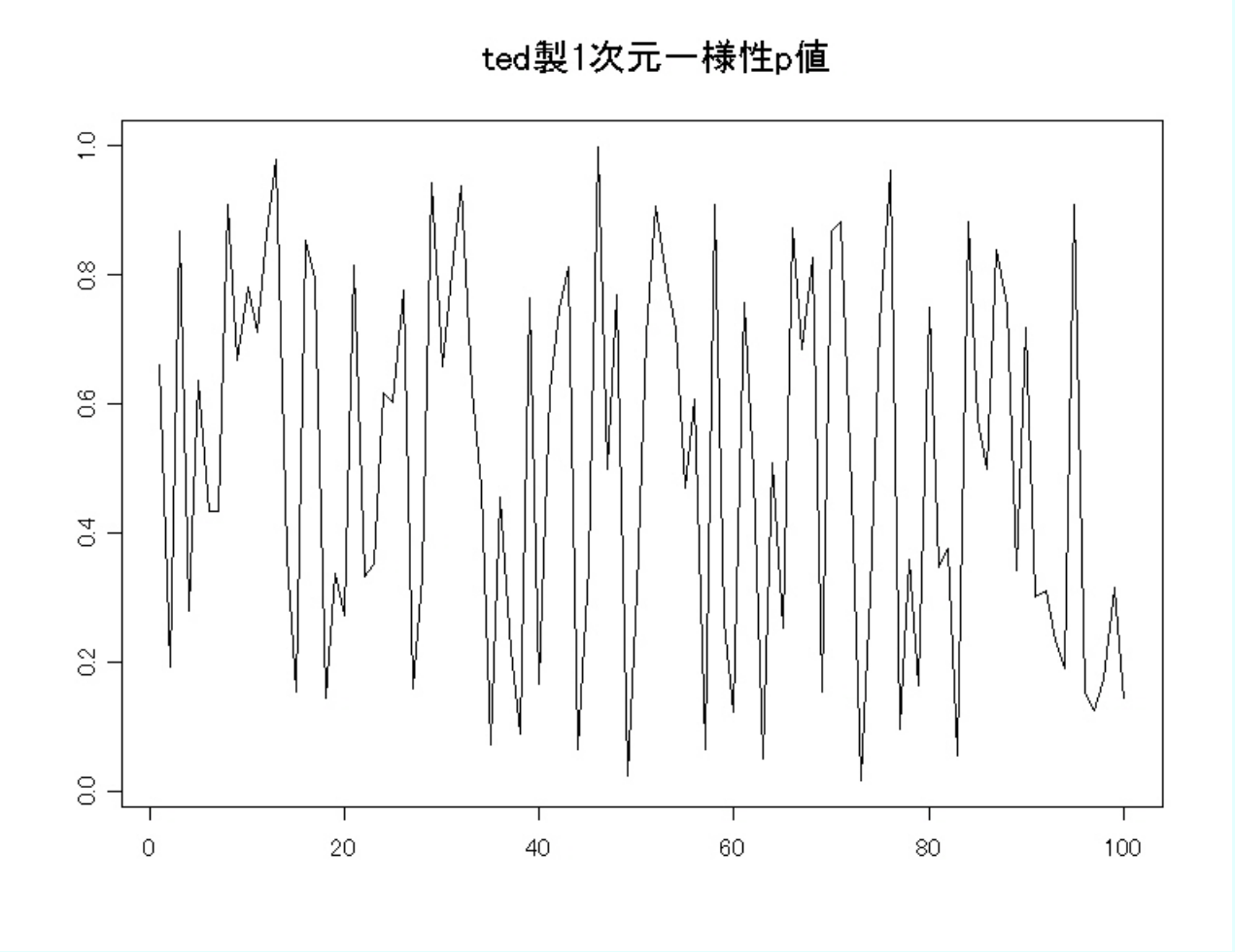
```
> sort(tedpv.dat)[1:100]
  [1] 0.00003310109 0.00136779212 0.00295652937 0.00435845450 0.00445732791 0.00523753168 0.00578221870
  [8] 0.00584580591 0.00594237522 0.00878435777 0.01032300698 0.01215928691 0.01259768771 0.01456676789
 [15] 0.01493243476 0.01789888360 0.02314992195 0.02336768663 0.02347723173 0.02391986351 0.02529126692
 [22] 0.02600201154 0.02760109480 0.02785444545 0.02941677344 0.03091409534 0.03204342696 0.03204342696
 [29] 0.03306018689 0.03350425227 0.03410419966 0.03643567144 0.03856370578 0.03873164203 0.03873164203
 [36] 0.04114797100 0.04387558038 0.04406265459 0.04558327922 0.04655561771 0.04694933900 0.04714723191
 [43] 0.04834914673 0.05040855712 0.05167852742 0.05189272535 0.05232331030 0.05363269919 0.05429739626
 [50] 0.05429739626 0.05474426322 0.05542020077 0.05702393068 0.05702393068 0.05914123791 0.06083181829
```

```
> sort(ted2dpv.dat)[1:100]
  [1] 0.002939425 0.004955480 0.005534091 0.005534091 0.007574863 0.009650645 0.010269840 0.011150505
  [9] 0.012221075 0.012471022 0.012471022 0.013248279 0.013381910 0.013381910 0.013789986 0.014932435
 [17] 0.014932435 0.015230769 0.015688112 0.016157494 0.016477229 0.018514786 0.021983340 0.022401516
 [25] 0.025058101 0.026002012 0.026485330 0.027224775 0.029550284 0.029550284 0.029818858 0.029818858
 [33] 0.030089515 0.030362265 0.030362265 0.030637121 0.031757854 0.032331180 0.032621127 0.032913280
 [41] 0.032913280 0.033504252 0.034407571 0.034713223 0.036276161 0.039239142 0.040270897 0.040619829
 [49] 0.040619829 0.041325290 0.041325290 0.042766978 0.042766978 0.044627866 0.046164638 0.048551908
 [57] 0.050199416 0.050618416 0.051040292 0.053193284 0.054968822 0.055874604 0.057722728 0.058665287
```

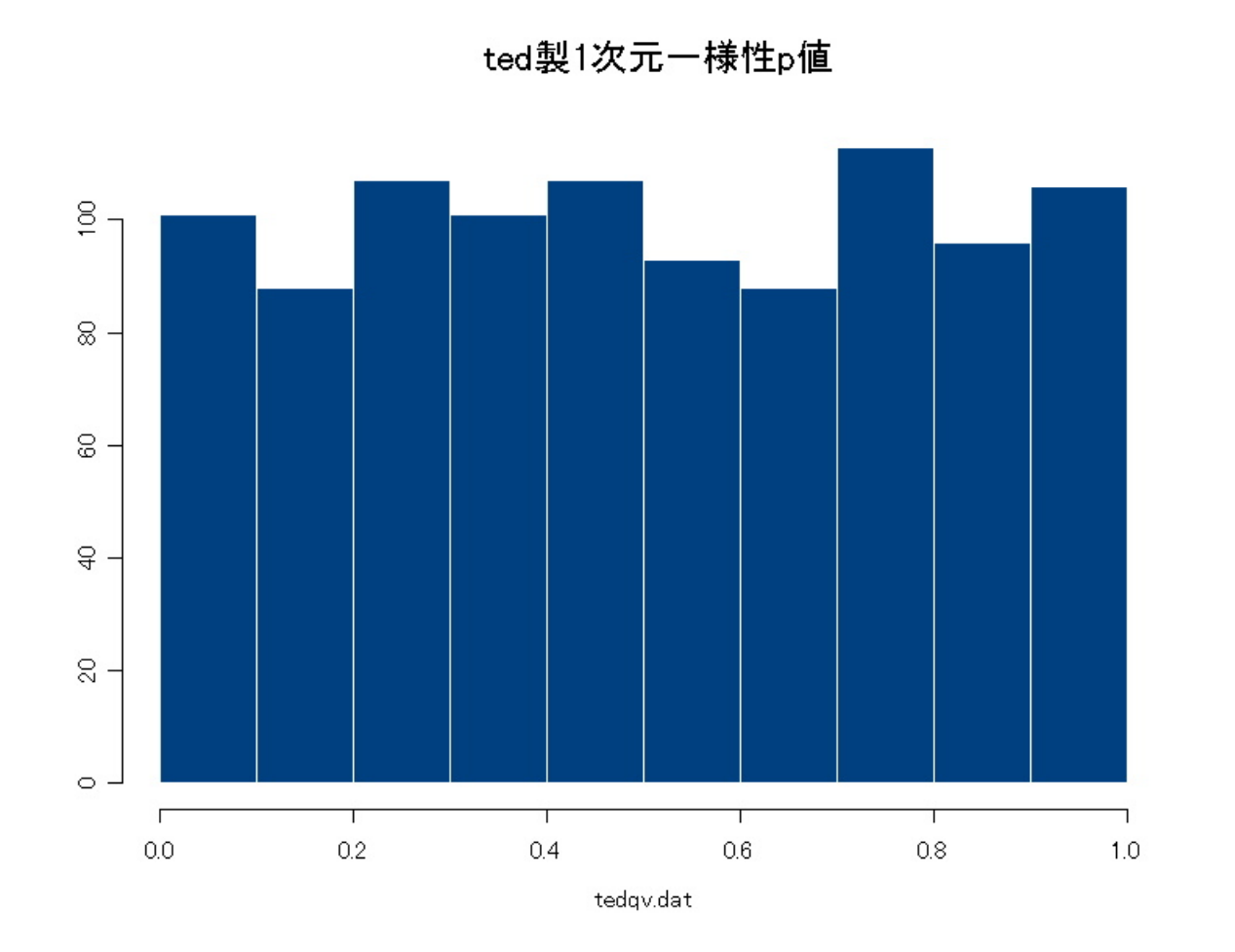
TED製ボードの検定結果(4) 自己相関

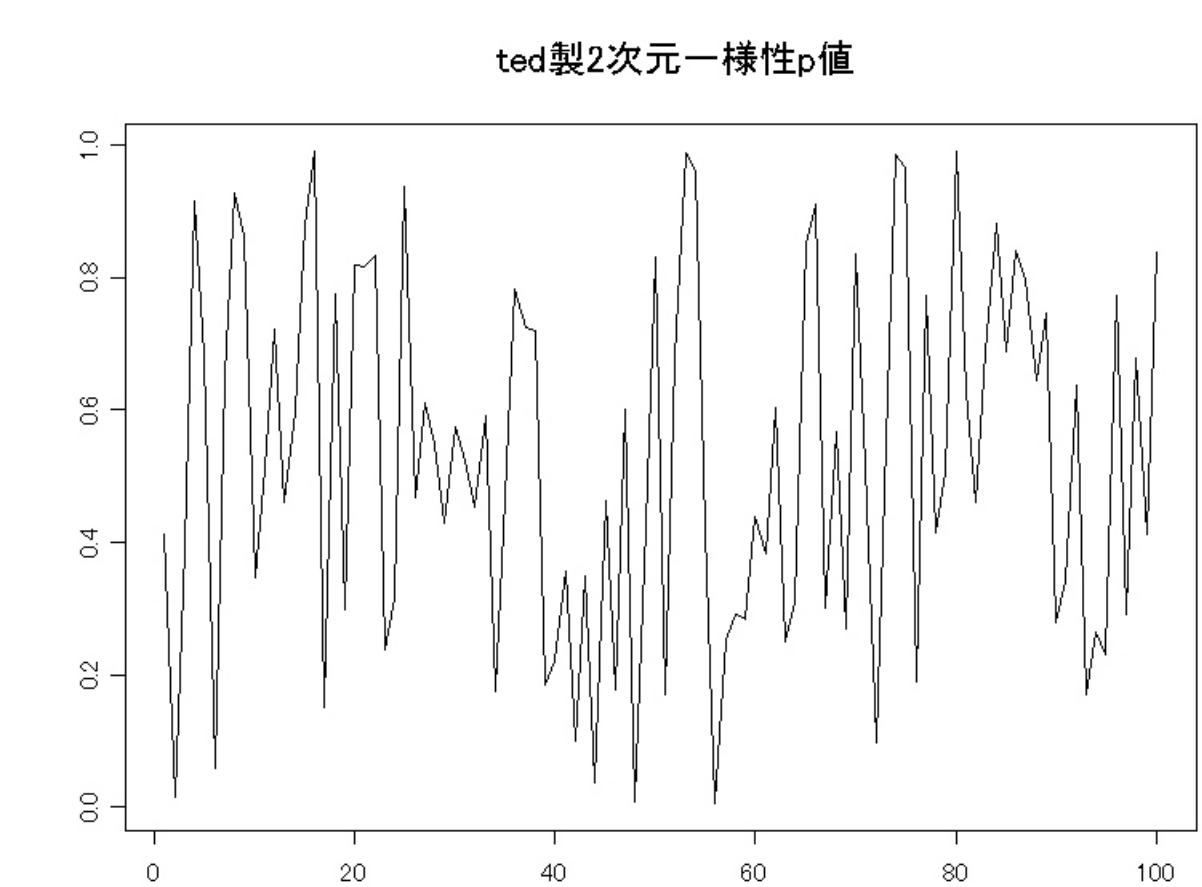
TED製ボードの検定結果(5)　p値の変化



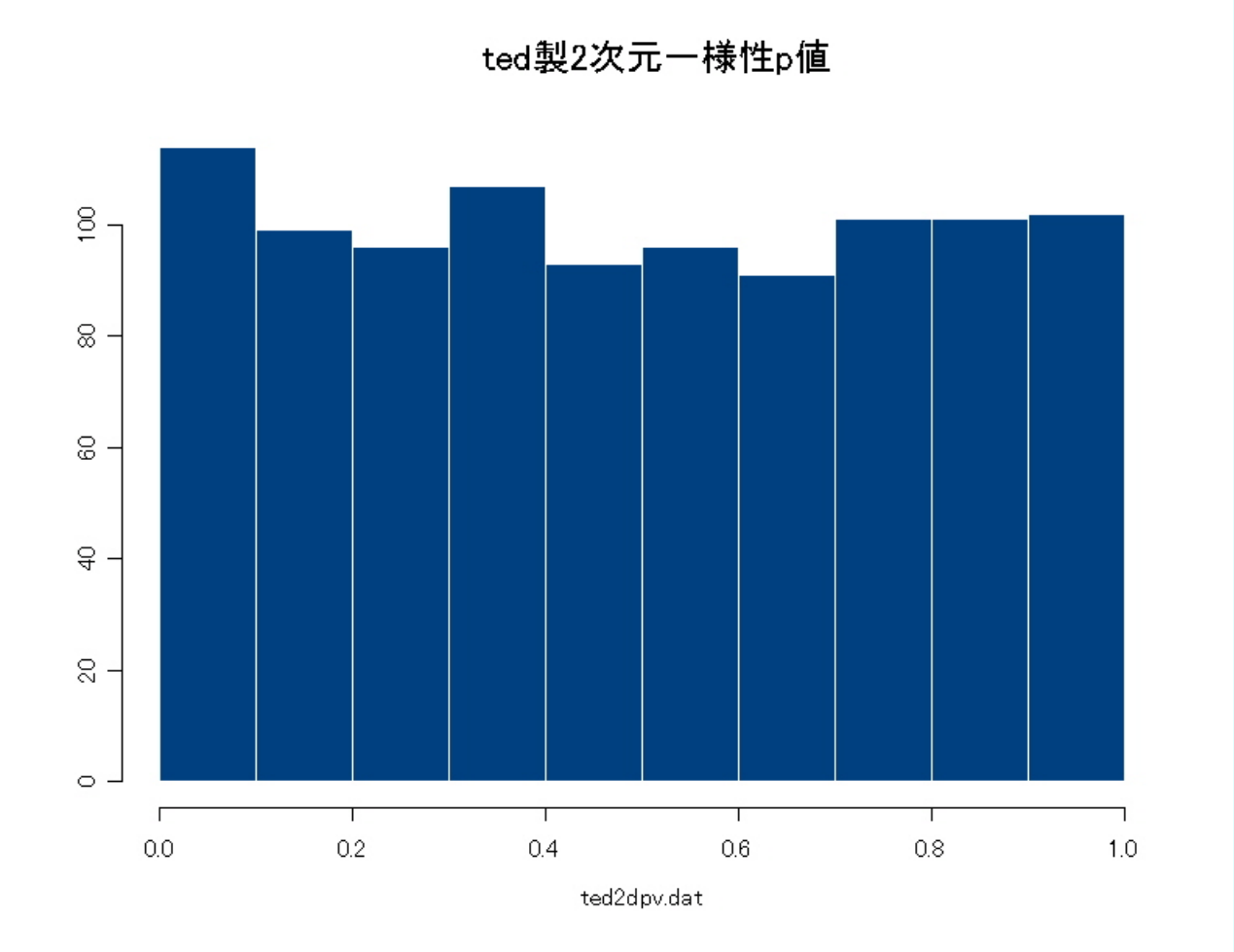
TED製ボードの検定結果(6)　p値の変化



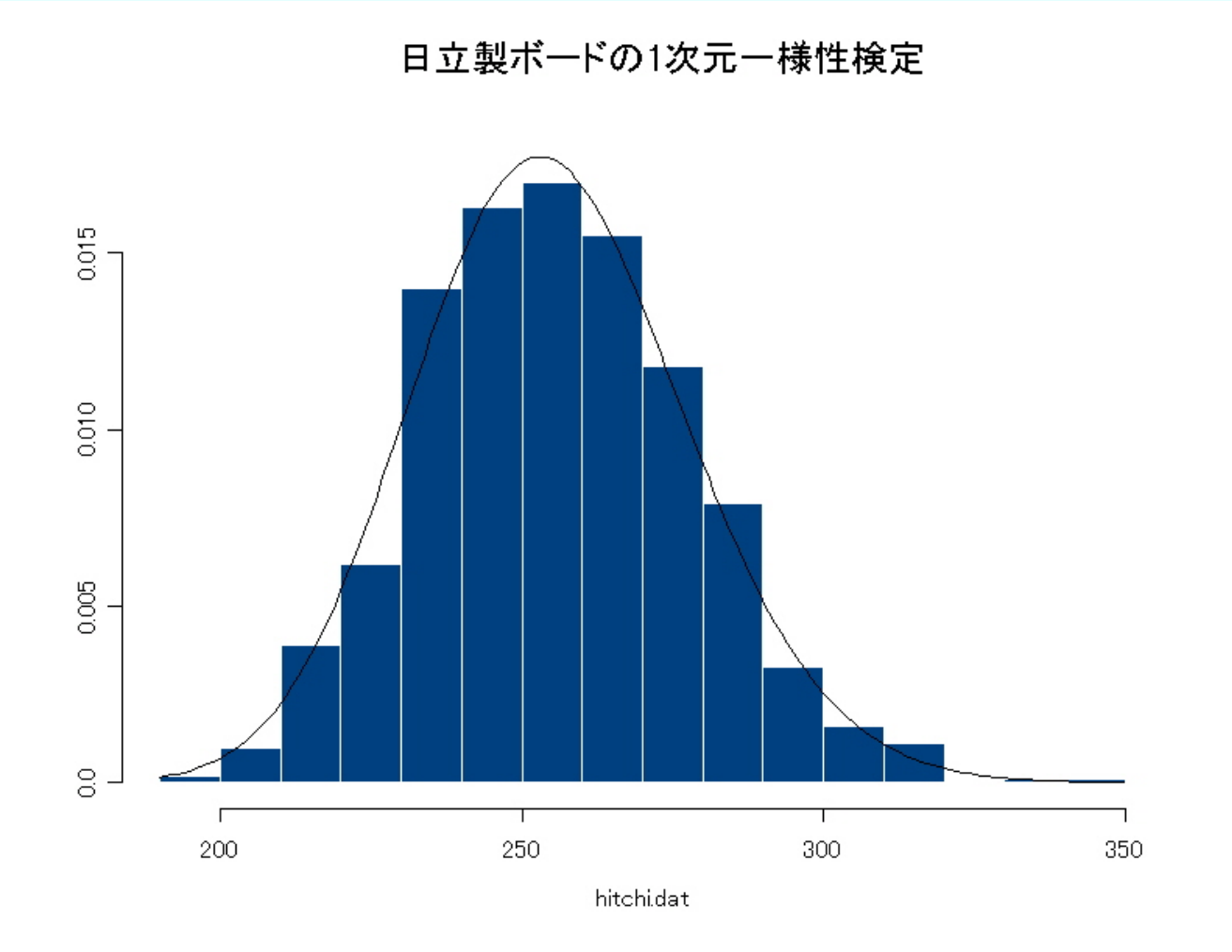
TED製ボードの検定結果(7)　p値の変化



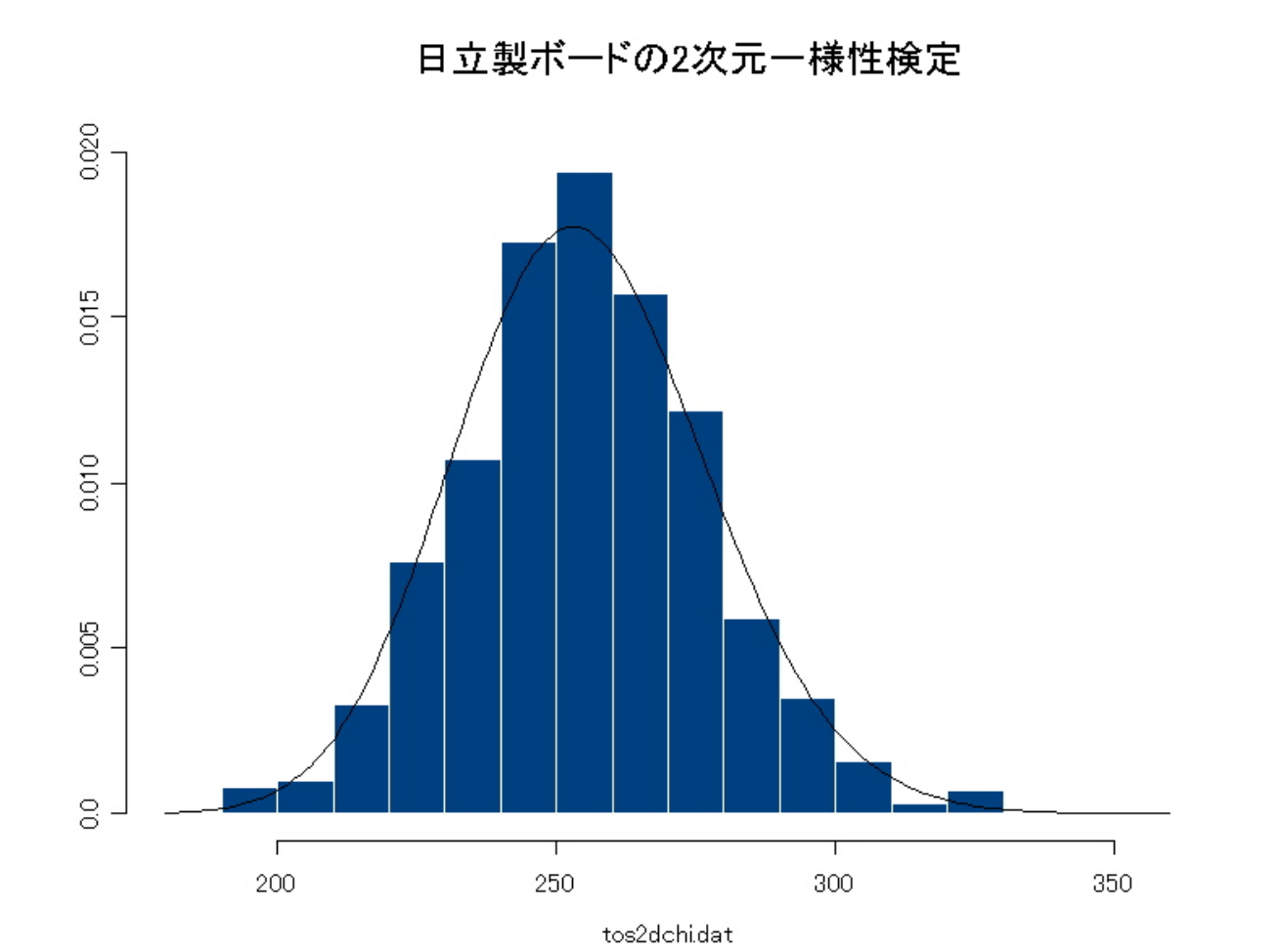
TED製ボードの検定結果(8)　p値の変化



日立製ボードの検定結果(1)



日立製ボードの検定結果(2)

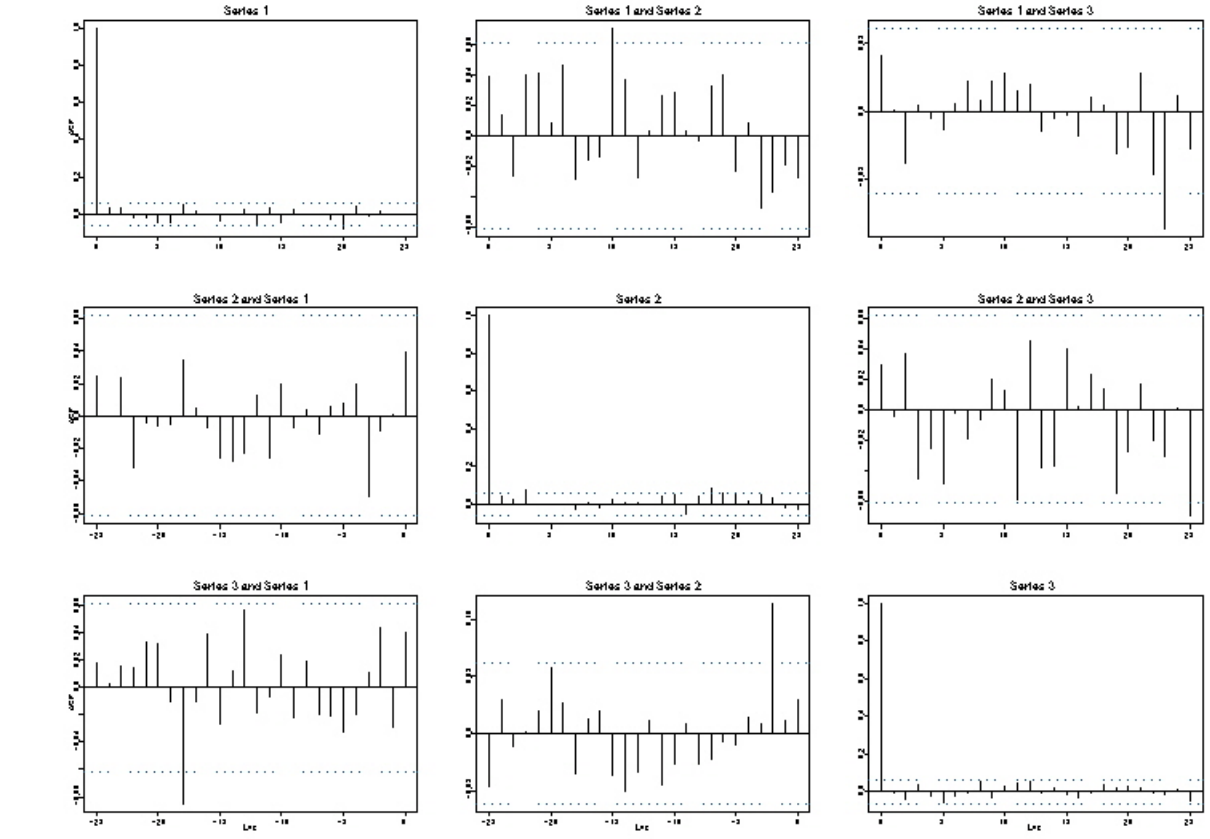


日立製ボードの検定結果(3)

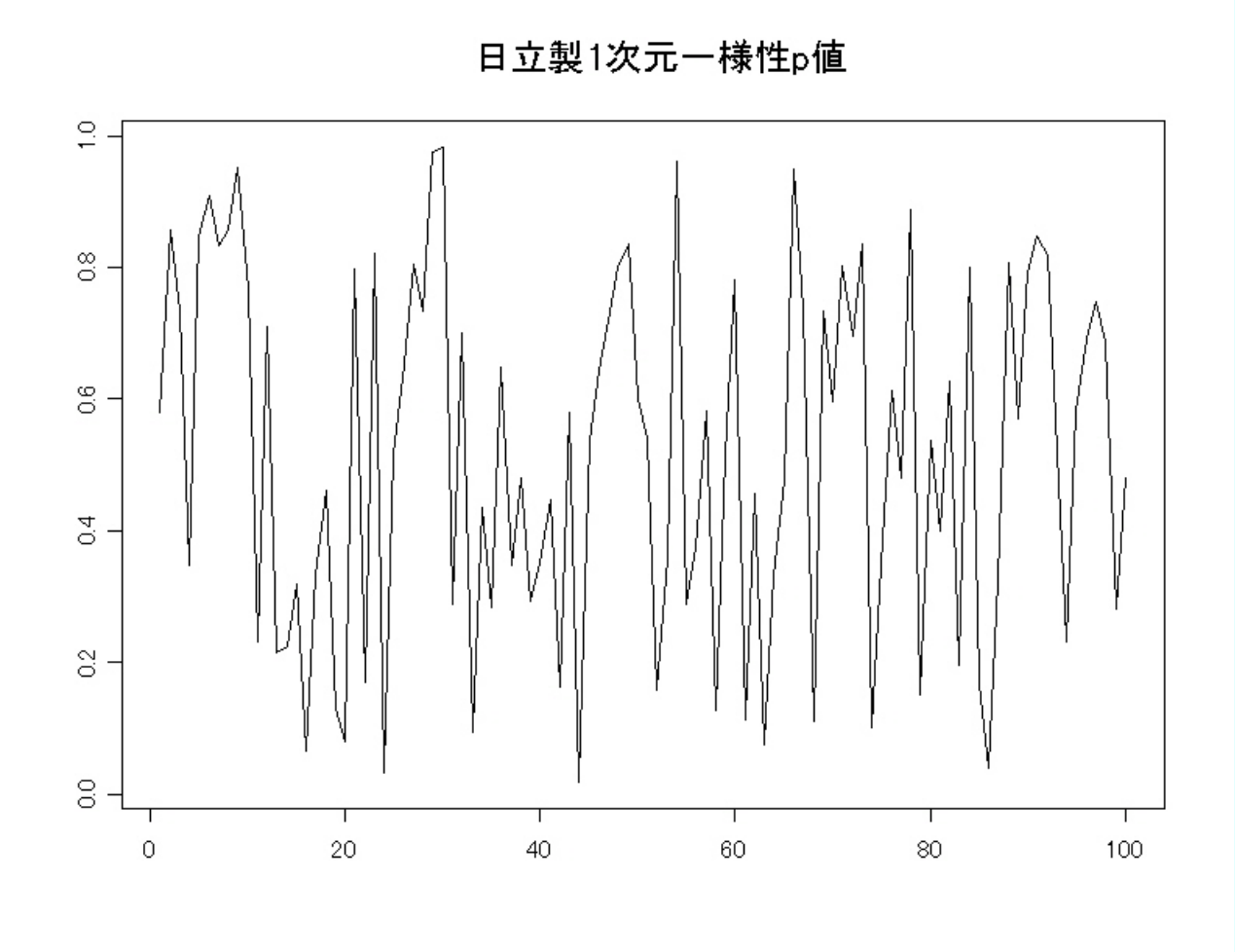
```
> sort(hitpv.dat)[1:100]
  [1] 0.0002823703 0.0003846547 0.0043829836 0.0046873101 0.0047135197 0.0053839581 0.0059748847
  [8] 0.0061733463 0.0069530971 0.0075345832 0.0080735677 0.0086469564 0.0104300775 0.0108683331
 [15] 0.0110368626 0.0110935540 0.0129845556 0.0131158304 0.0143511795 0.0153818950 0.0156110595
 [22] 0.0163166735 0.0173009861 0.0179857572 0.0200882015 0.0221915744 0.0238085350 0.0247118258
 [29] 0.0251744512 0.0277275221 0.0284965572 0.0298188585 0.0311932009 0.0323311804 0.0323311804
 [36] 0.0323311804 0.0339533659 0.0341041997 0.0382296689 0.0382296689 0.0397522120 0.0411479710
 [43] 0.0424026774 0.0448176044 0.0485519078 0.0485519078 0.0493699906 0.0504085571 0.0508289933
 [50] 0.0512523122 0.0514650569 0.0521076524 0.0521076524 0.0527568253 0.0545204551 0.0561029433
```

```
> sort(hit2dpv.dat)[1:100]
  [1] 0.001921909 0.002016215 0.002646688 0.003078844 0.005534091 0.007820640 0.010059610 0.012471022
  [9] 0.012854446 0.013789986 0.014639263 0.015534340 0.015688112 0.015843223 0.015843223 0.015999680
 [17] 0.016802507 0.017640462 0.018160612 0.018336951 0.019613766 0.022191574 0.022613177 0.023258584
 [25] 0.023697655 0.024369686 0.025526300 0.025526300 0.026485330 0.026485330 0.029019331 0.030089515
 [33] 0.030362265 0.030914095 0.031193201 0.031474450 0.032913280 0.035331425 0.035644000 0.040619829
 [41] 0.041681847 0.043503416 0.043875580 0.044250392 0.045776386 0.046949339 0.047345815 0.048147086
 [49] 0.049369991 0.051465057 0.052756825 0.053193284 0.053193284 0.054520455 0.054968822 0.055420201
```

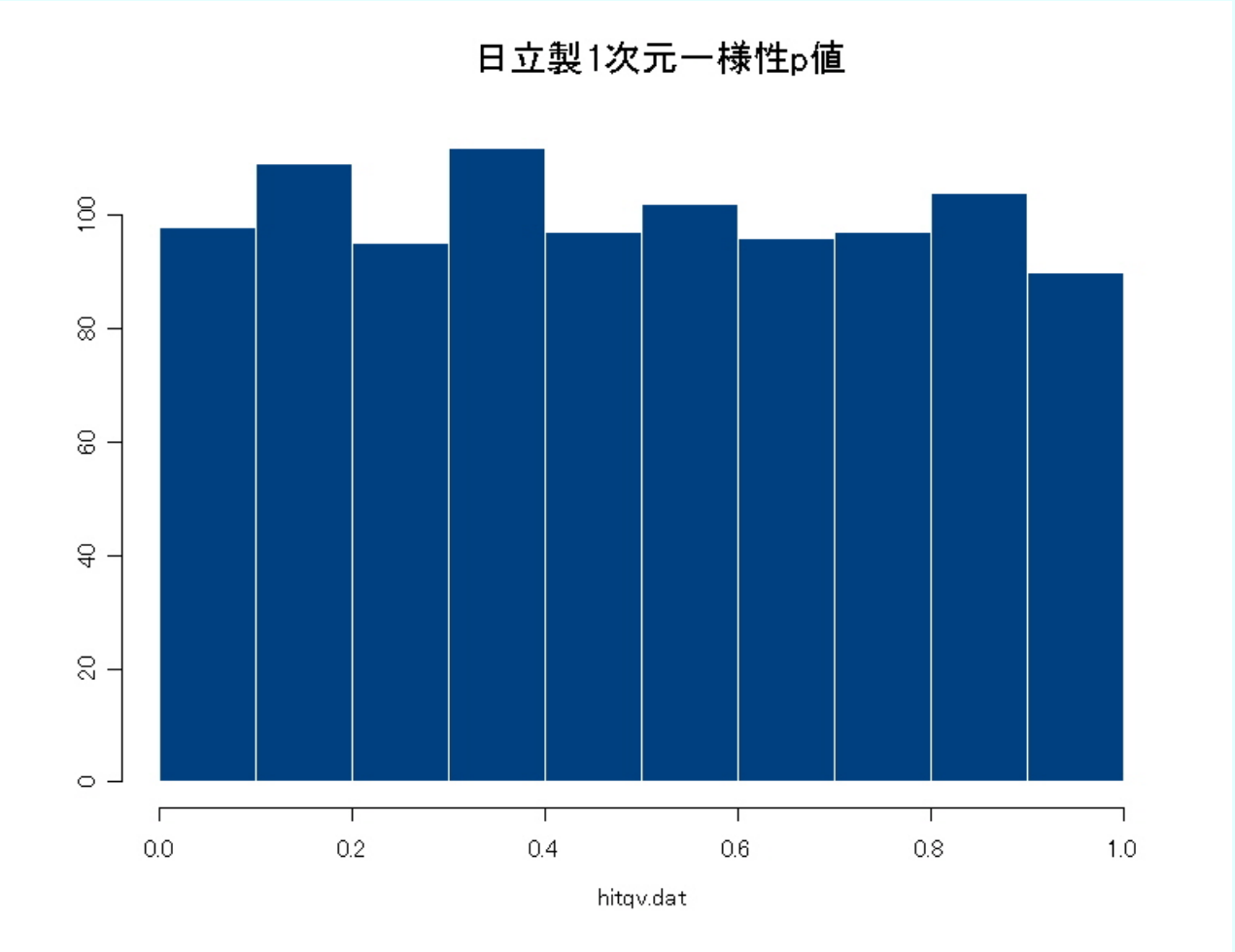
日立製ボードの検定結果(4) 自己相関



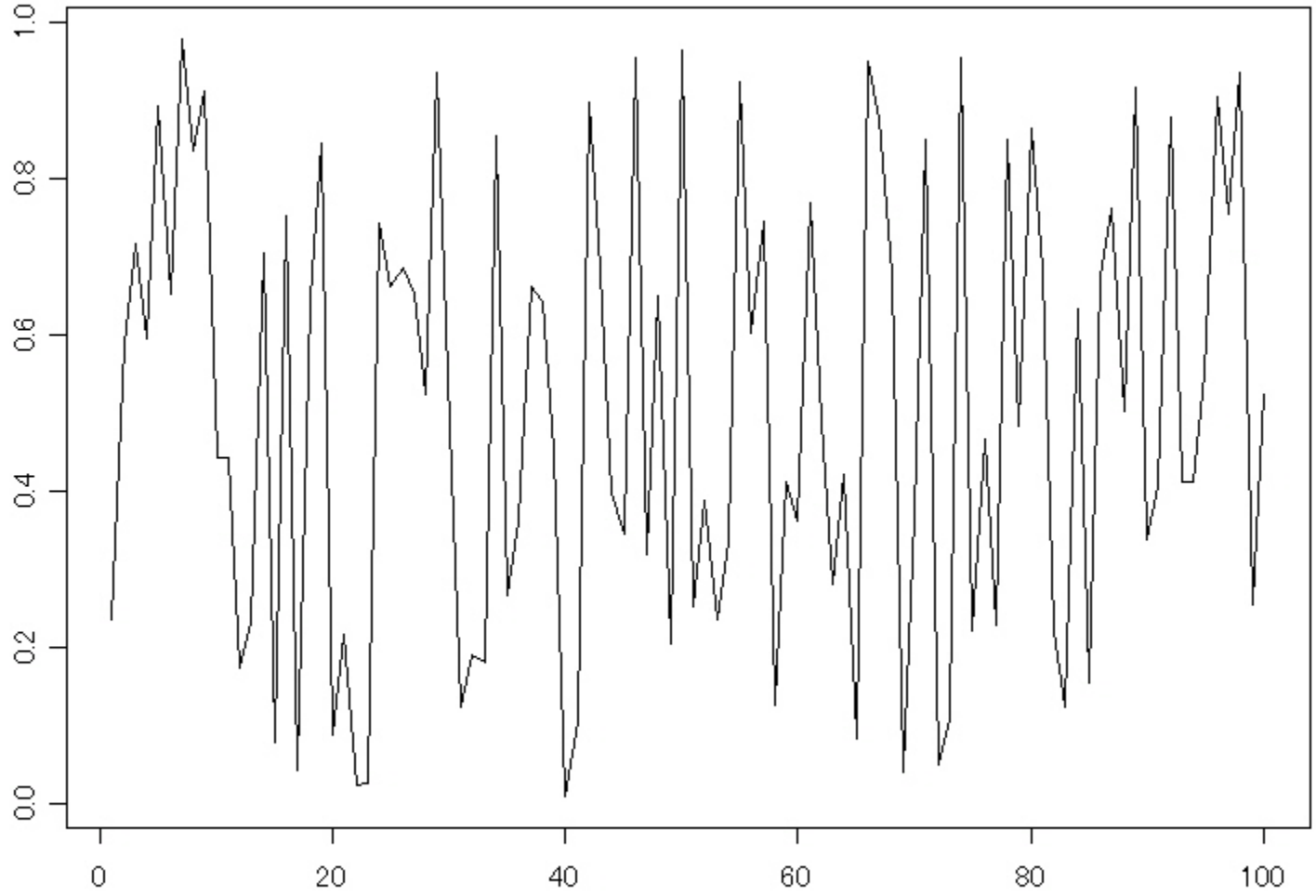
日立製ボードの検定結果(5) p値の変化



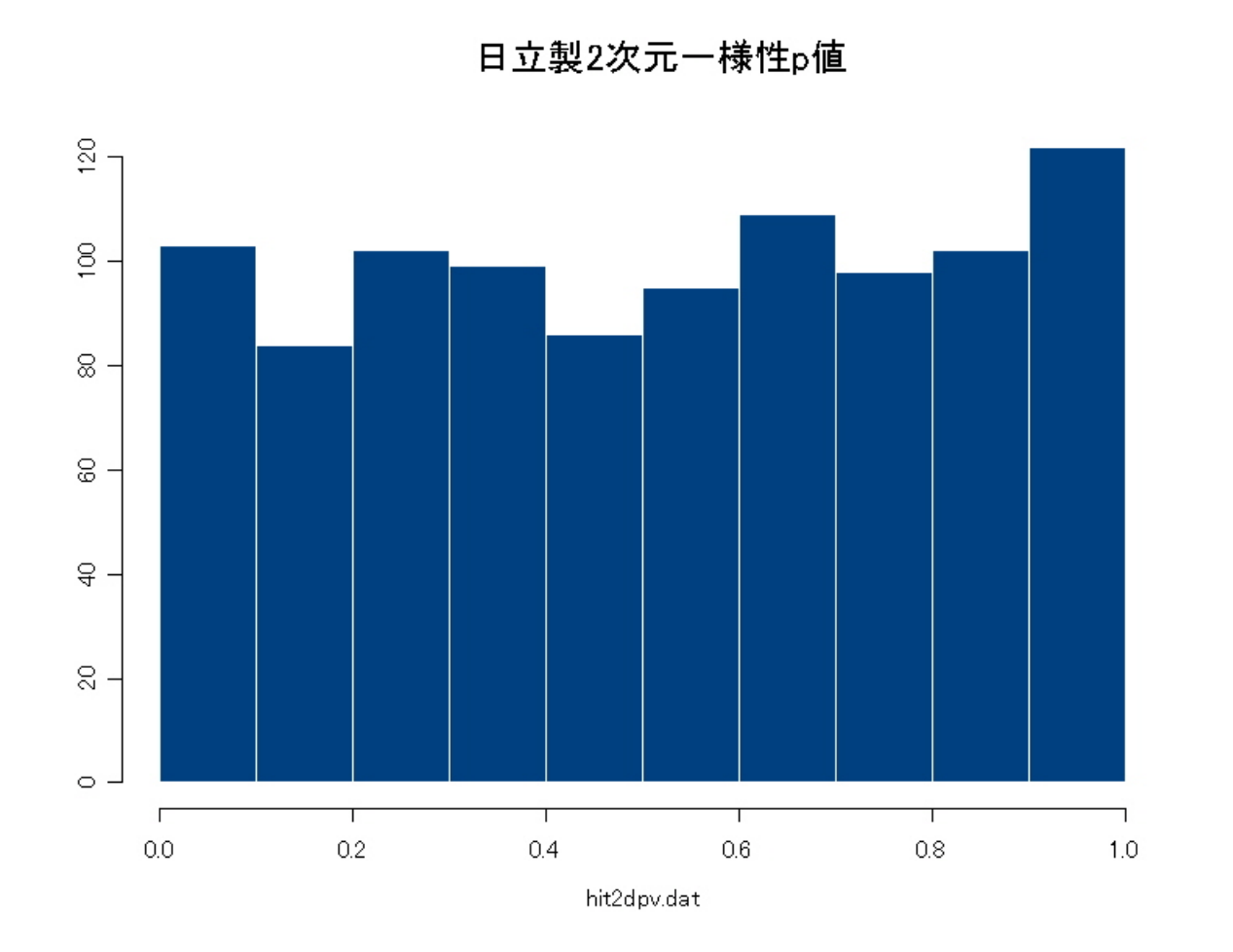
日立製ボードの検定結果(6) p値の変化



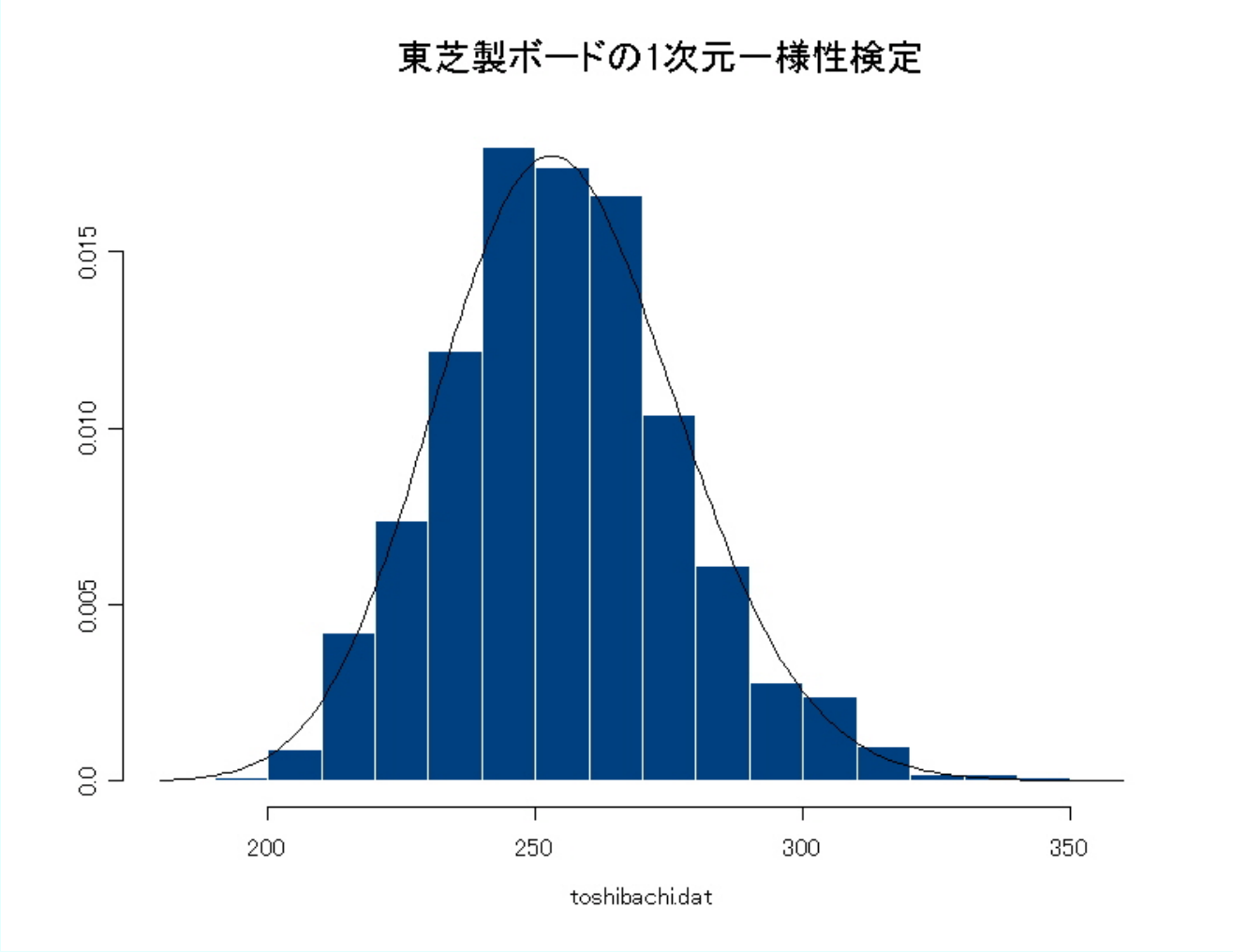
日立製2次元一様性p値



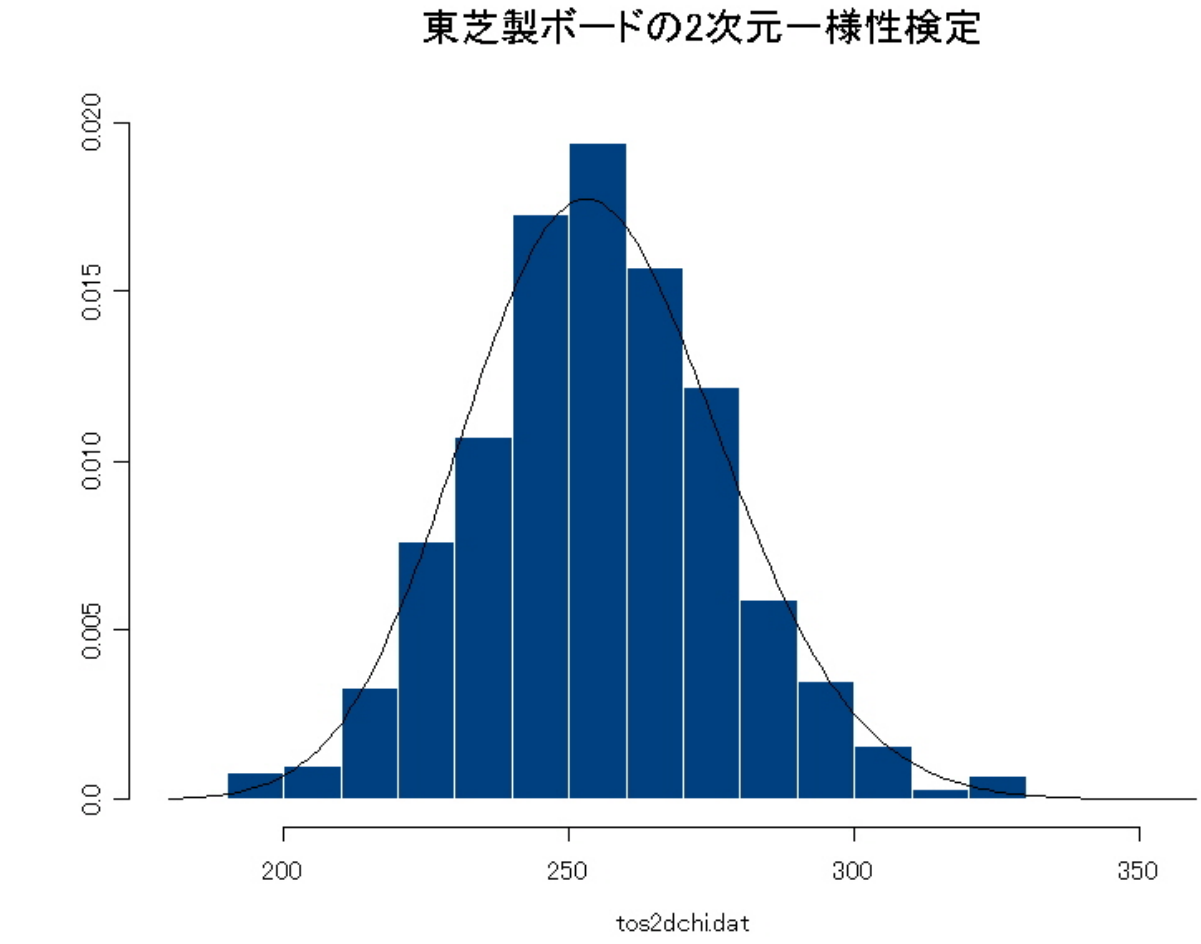
日立製ボードの検定結果(8) p値の変化



東芝製ボードの検定結果(1)



東芝製ボードの検定結果(2)

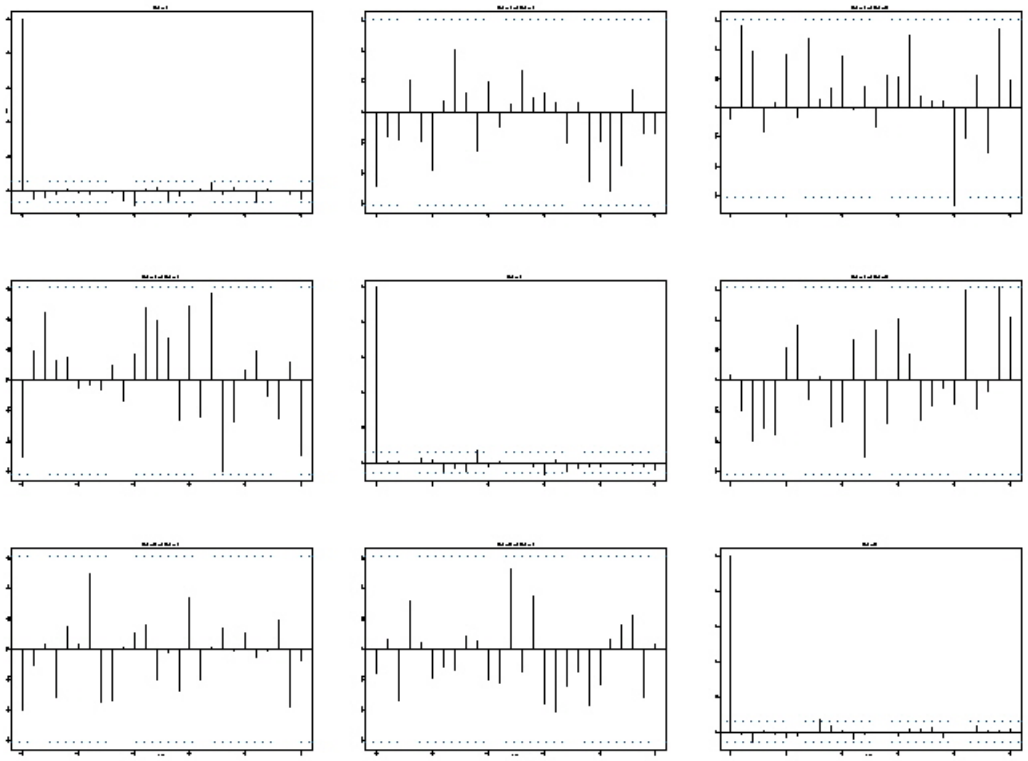


東芝製ボードの検定結果(3)

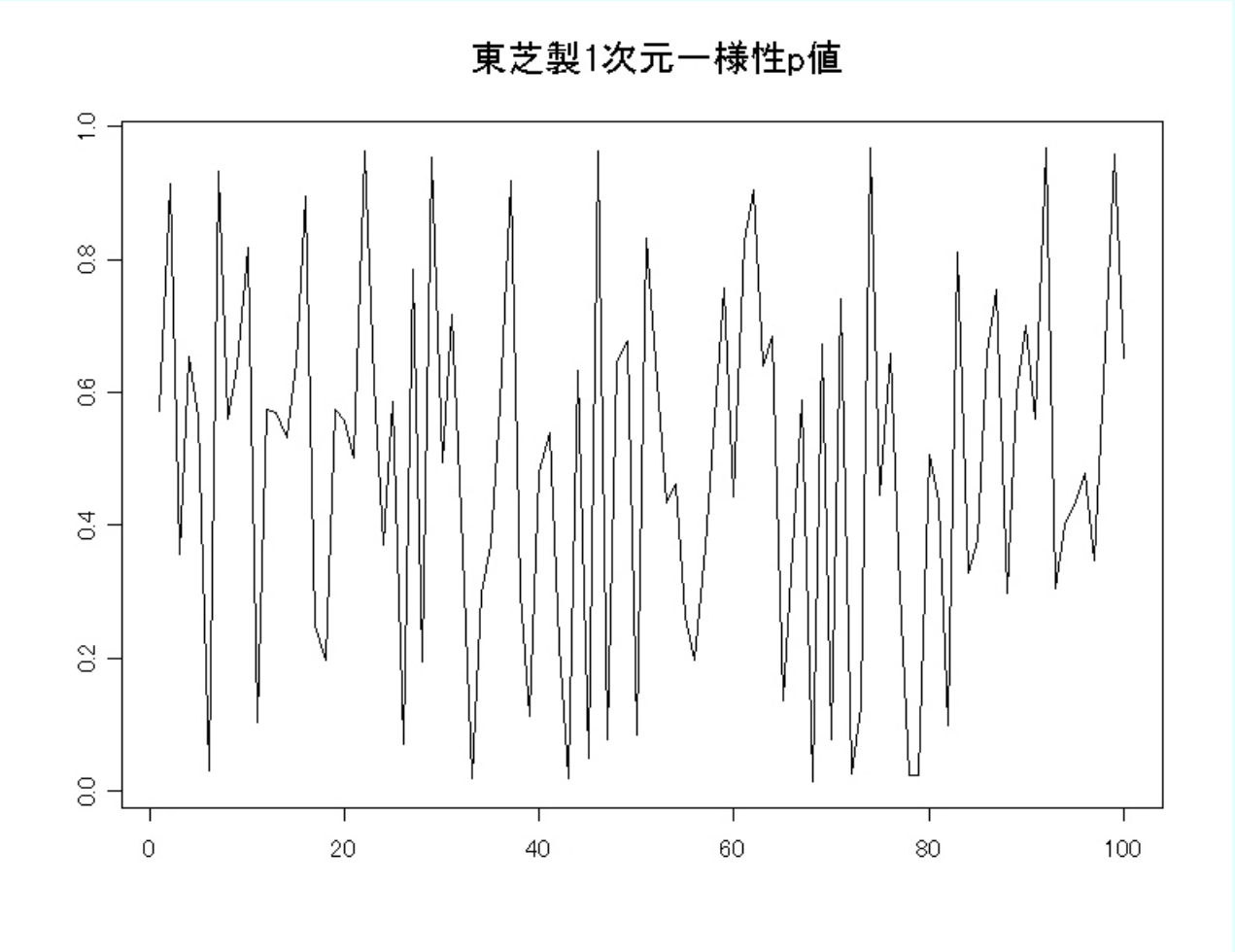
```
> sort(toshibapv.dat)[1:100]
  [1] 0.0001722135 0.0004630741 0.0008293569 0.0019334701 0.0028386935 0.0040967787 0.0048196978
  [8] 0.0059423752 0.0079461993 0.0082462457 0.0087383413 0.0098532313 0.0099559429 0.0100596099
 [15] 0.0103230070 0.0109242531 0.0110368626 0.0131158304 0.0131819074 0.0151556973 0.0175550509
 [22] 0.0188749813 0.0190573627 0.0198023554 0.0198023554 0.0209674027 0.0210671266 0.0220872446
 [29] 0.0224015162 0.0225071310 0.0228265677 0.0239198635 0.0249422139 0.0250581007 0.0254085493
 [36] 0.0256445197 0.0263637819 0.0266073588 0.0271003173 0.0311932009 0.0316158817 0.0327669268
 [43] 0.0344075706 0.0361172401 0.0362761609 0.0369177572 0.0372421258 0.0375688919 0.0383963819
 [50] 0.0385637058 0.0435034162 0.0473458152 0.0473458152 0.0475450906 0.0483491467 0.0495762793
 [57] 0.0540750851 0.0558746035 0.0563320442 0.0565619079 0.0581924551 0.0586652872 0.0603448467
```

```
> sort(tos2dpv.dat)[1:100]
  [1] 0.001286059 0.001508672 0.001831699 0.001876302 0.001968540 0.002495896 0.003008393 0.005782219
  [9] 0.007737929 0.009751468 0.012854446 0.014209023 0.014351180 0.017640462 0.017985757 0.018160612
 [17] 0.019802355 0.019802355 0.021167262 0.021776803 0.023697655 0.024143870 0.025058101 0.025526300
 [25] 0.026242714 0.026729871 0.027981866 0.028238208 0.028756928 0.029019331 0.030089515 0.030914095
 [33] 0.032331180 0.032331180 0.034713223 0.034713223 0.035958907 0.038229669 0.039239142 0.039924481
 [41] 0.042402677 0.042402677 0.042766978 0.045008013 0.046164638 0.048551908 0.048959538 0.048959538
 [49] 0.049369991 0.049783279 0.050199416 0.051892725 0.052756825 0.053193284 0.058192455 0.060587936
```

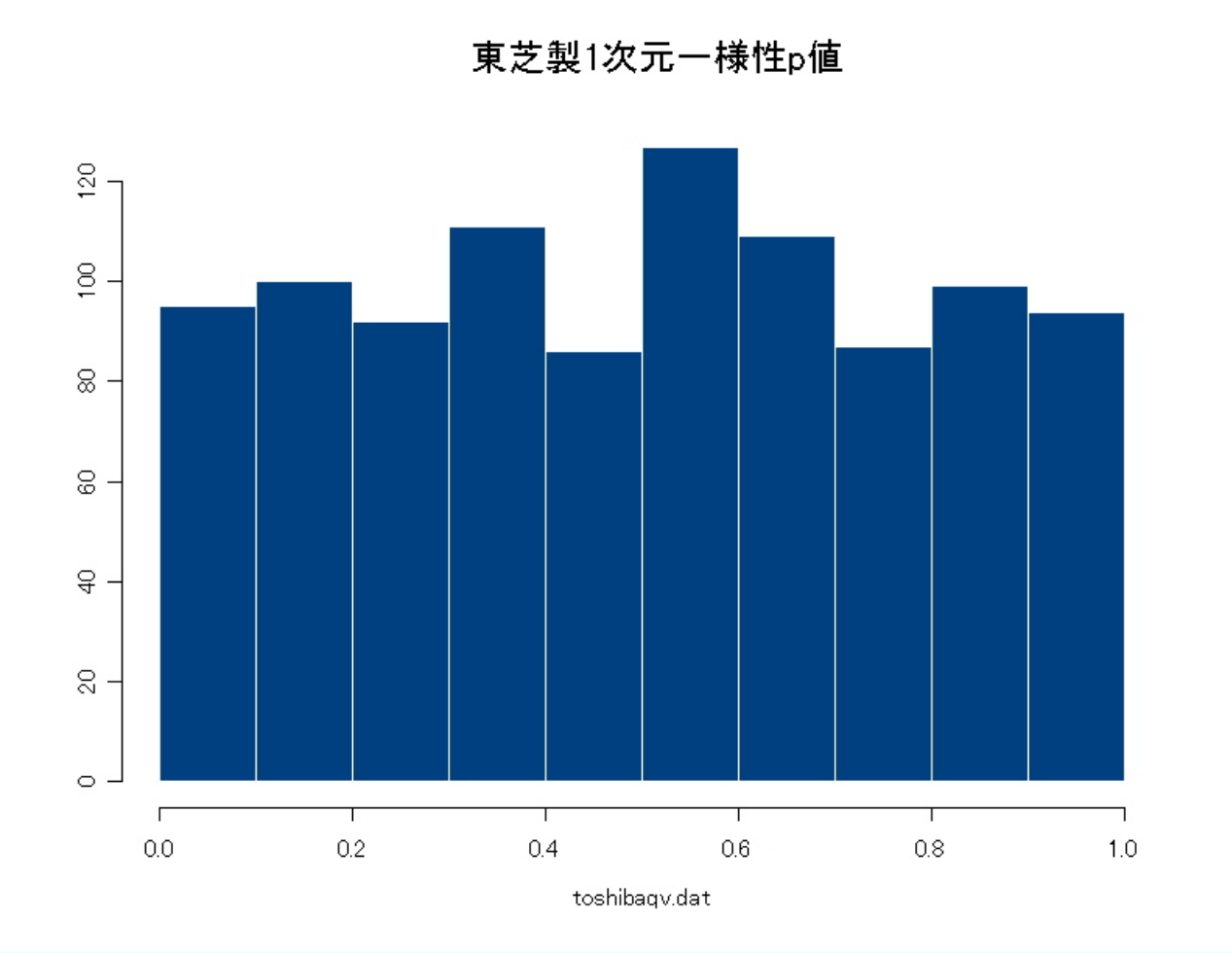
東芝製ボードの検定結果(4) 自己相関



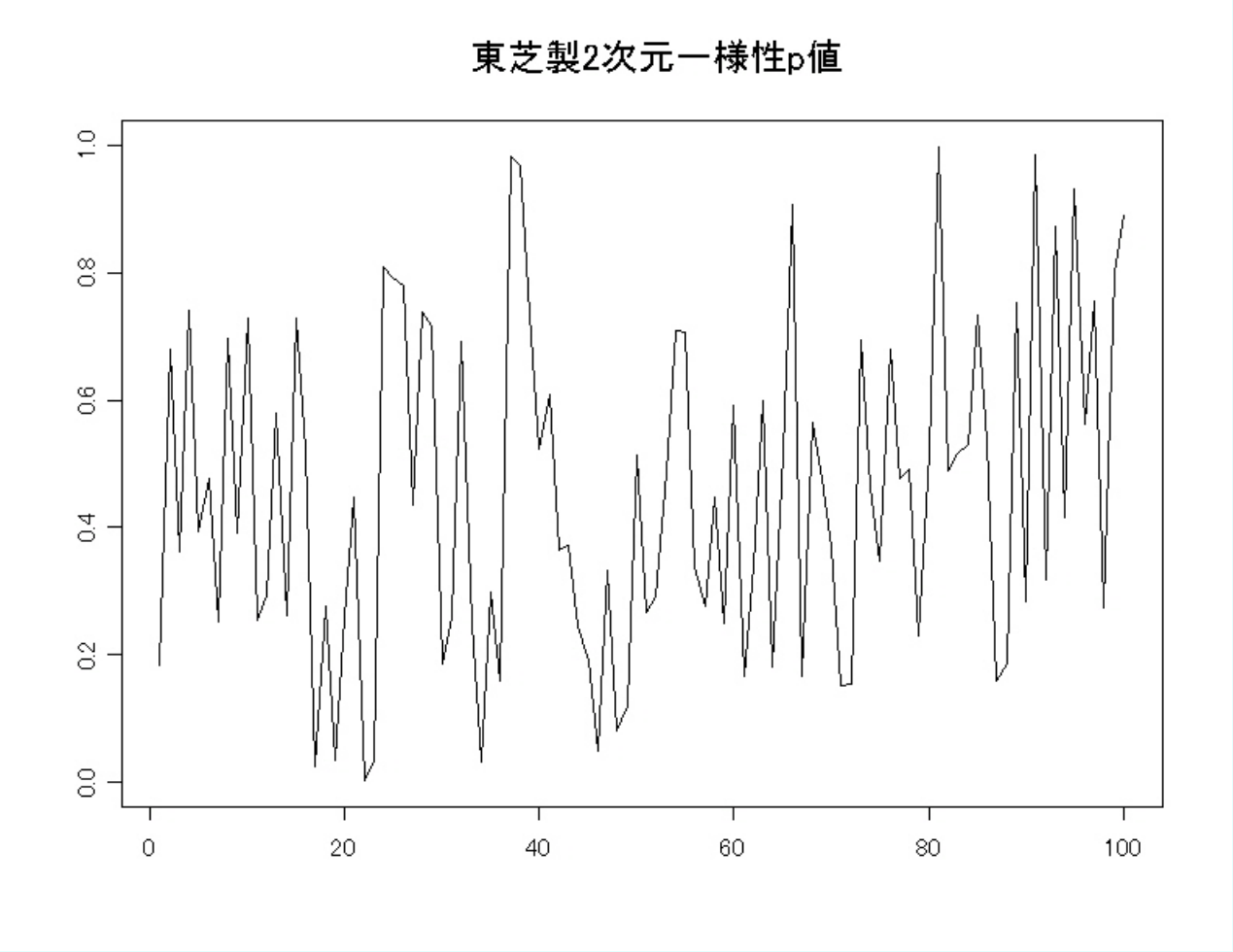
東芝製ボードの検定結果(5) p値の変化



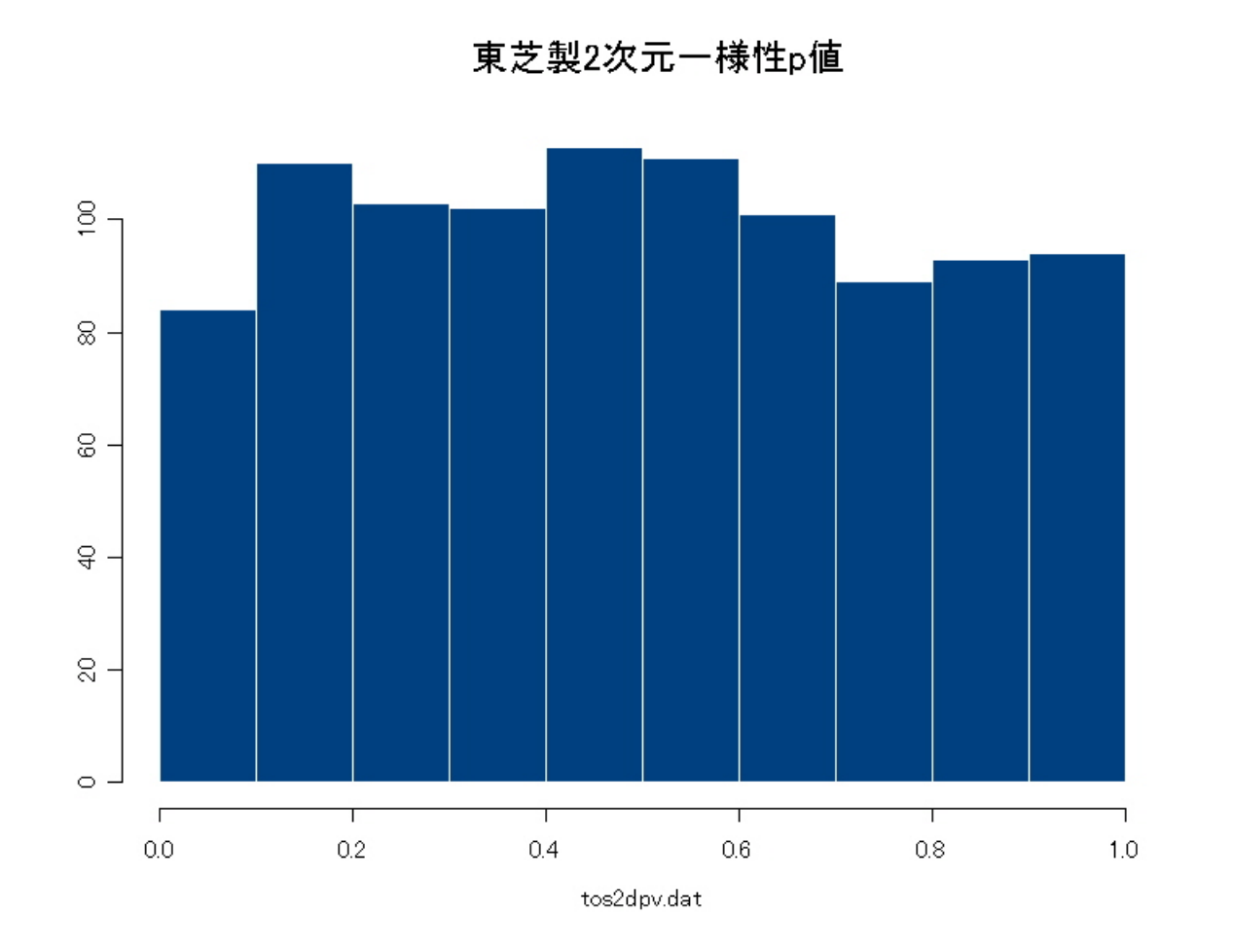
東芝製ボードの検定結果(6) p値の変化



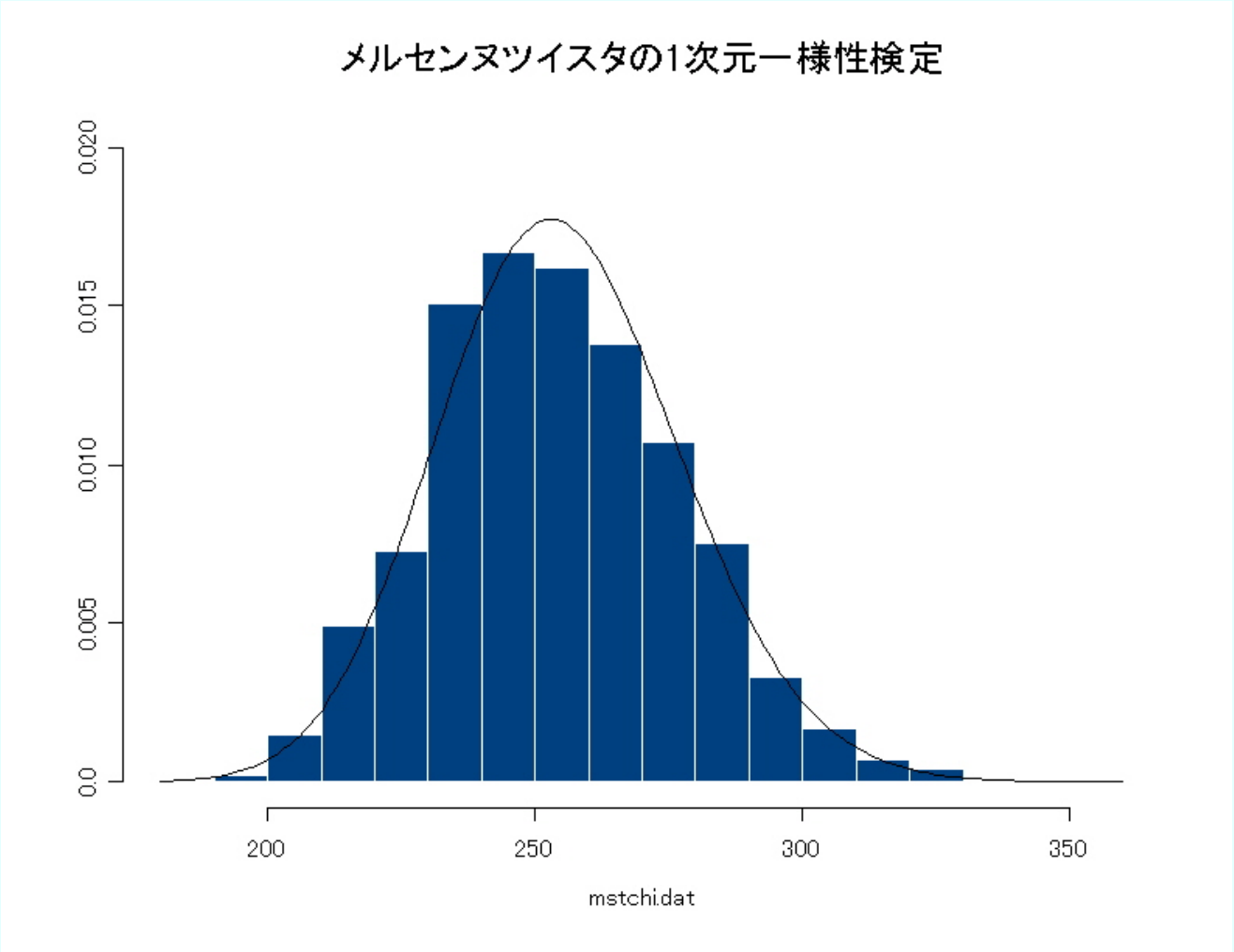
東芝製ボードの検定結果(7) p値の変化



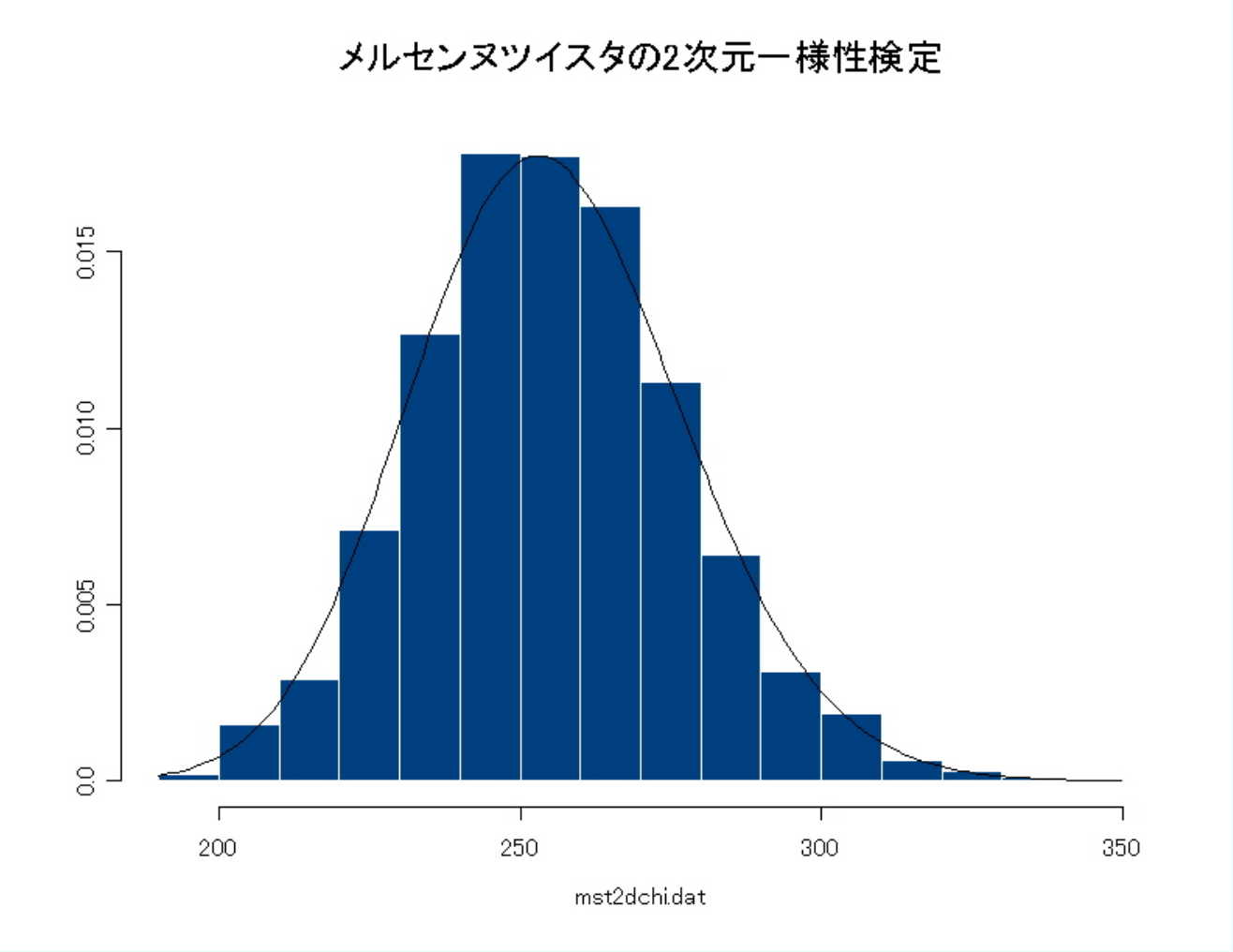
東芝製ボードの検定結果(8) p値の変化



メルセンヌツイスターの検定結果(1)



メルセンヌツイスターの検定結果(2)

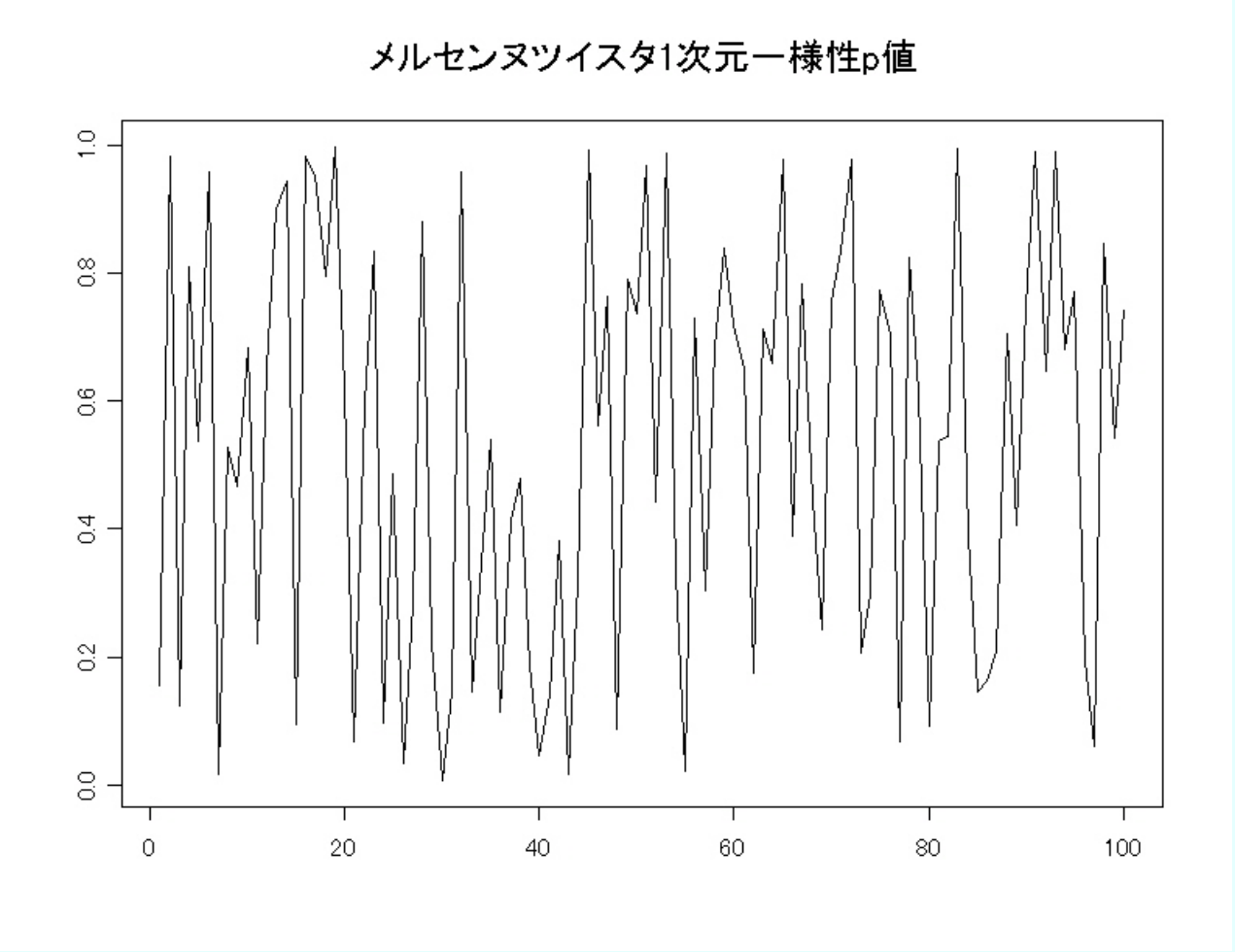


メルセンヌツイスターの検定結果(3)

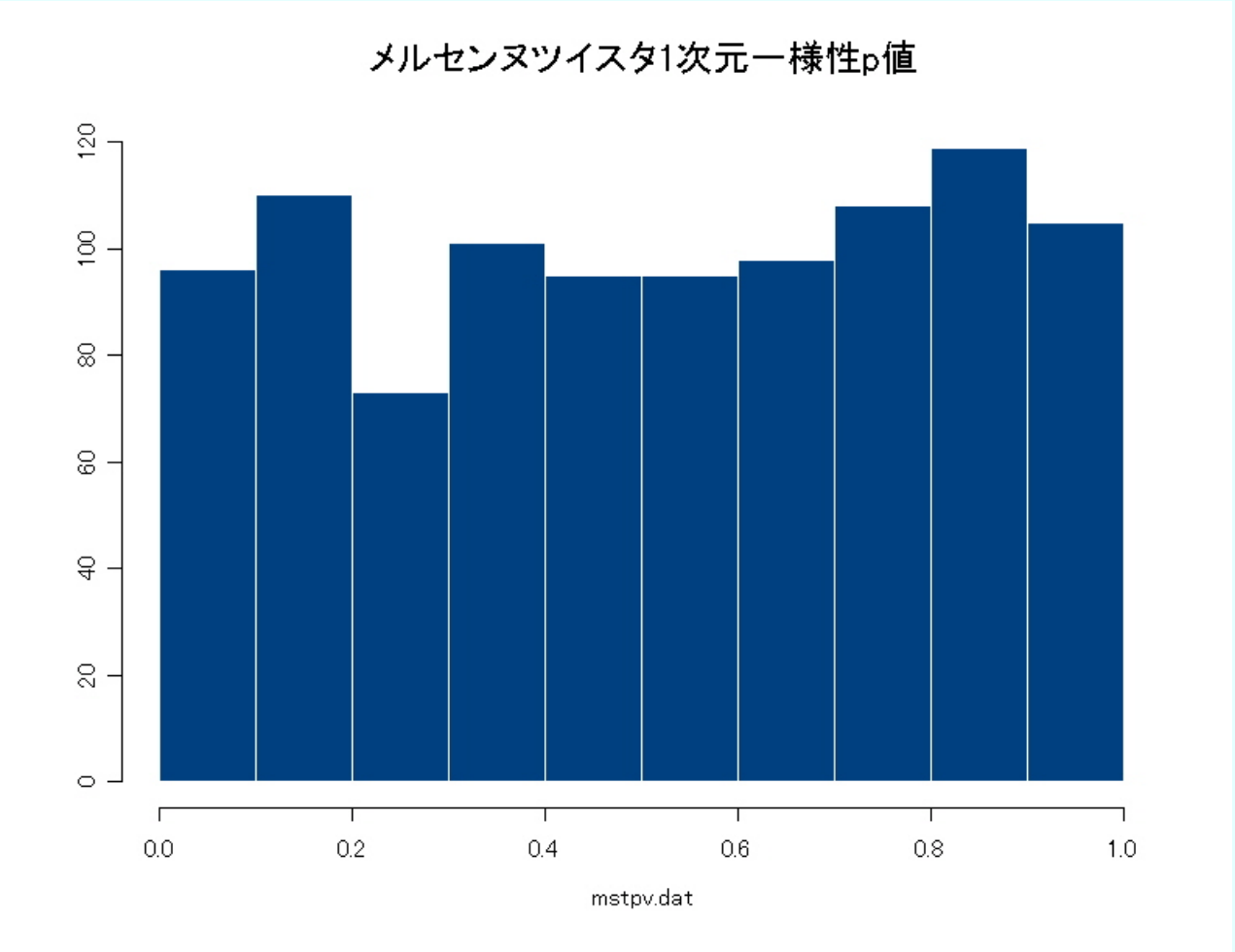
```
> sort(mstpv.dat)[1:100]
  [1] 0.001445455 0.002381005 0.003008393 0.003395482 0.004143262 0.006343267 0.006482237 0.006805030
  [9] 0.006878704 0.008970604 0.009700939 0.011380918 0.011674886 0.012221075 0.014279945 0.015080952
 [17] 0.015306168 0.016157494 0.017640462 0.019520060 0.020184279 0.020572599 0.021267811 0.022087245
 [25] 0.024711826 0.025882374 0.025882374 0.026607359 0.027854445 0.030089515 0.032043427 0.032766927
 [33] 0.033355672 0.034255601 0.035331425 0.037242126 0.039069358 0.039239142 0.039752212 0.041325290
 [41] 0.041861087 0.042402677 0.043318323 0.043689168 0.044062655 0.045583279 0.046164638 0.049576279
 [49] 0.051040292 0.053193284 0.055420201 0.055874604 0.056561908 0.057256093 0.058665287 0.059141238
```

```
> sort(mst2dpv.dat)[1:100]
  [1] 0.0008949491 0.0016230478 0.0016629266 0.0034941366 0.0054735879 0.0059100261 0.0070282152
  [8] 0.0083338226 0.0092566110 0.0094517907 0.0109242531 0.0119755840 0.0133819101 0.0137899863
 [15] 0.0137899863 0.0166391705 0.0166391705 0.0166391705 0.0179857572 0.0188749813 0.0209674027
 [22] 0.0221915744 0.0226131769 0.0228265677 0.0230416996 0.0232585838 0.0250581007 0.0264853297
 [29] 0.0274751622 0.0290193307 0.0311932009 0.0320434270 0.0329132795 0.0341041997 0.0341041997
 [36] 0.0350211705 0.0356439997 0.0356439997 0.0365957732 0.0372421258 0.0385637058 0.0392391417
 [43] 0.0399244810 0.0402708972 0.0402708972 0.0442503924 0.0450080135 0.0450080135 0.0453908492
 [50] 0.0473458152 0.0481470862 0.0523233103 0.0536326992 0.0536326992 0.0540750851 0.0545204551
 [57] 0.0558746035 0.0640752171 0.0645863560 0.0645863560 0.0651007738 0.0671914984 0.0677225089
```

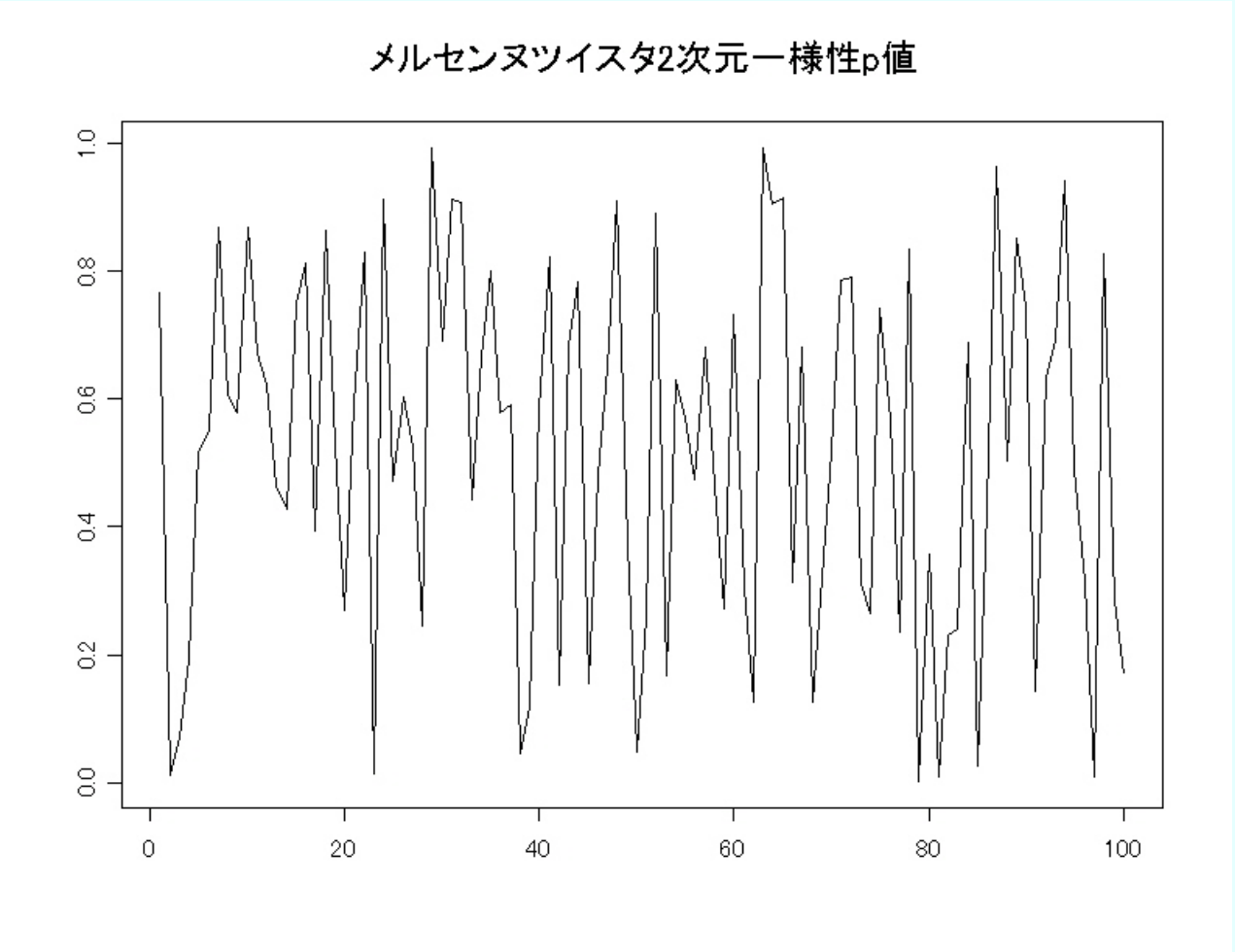
メルセンヌツイスターの検定結果(4) p値の変化



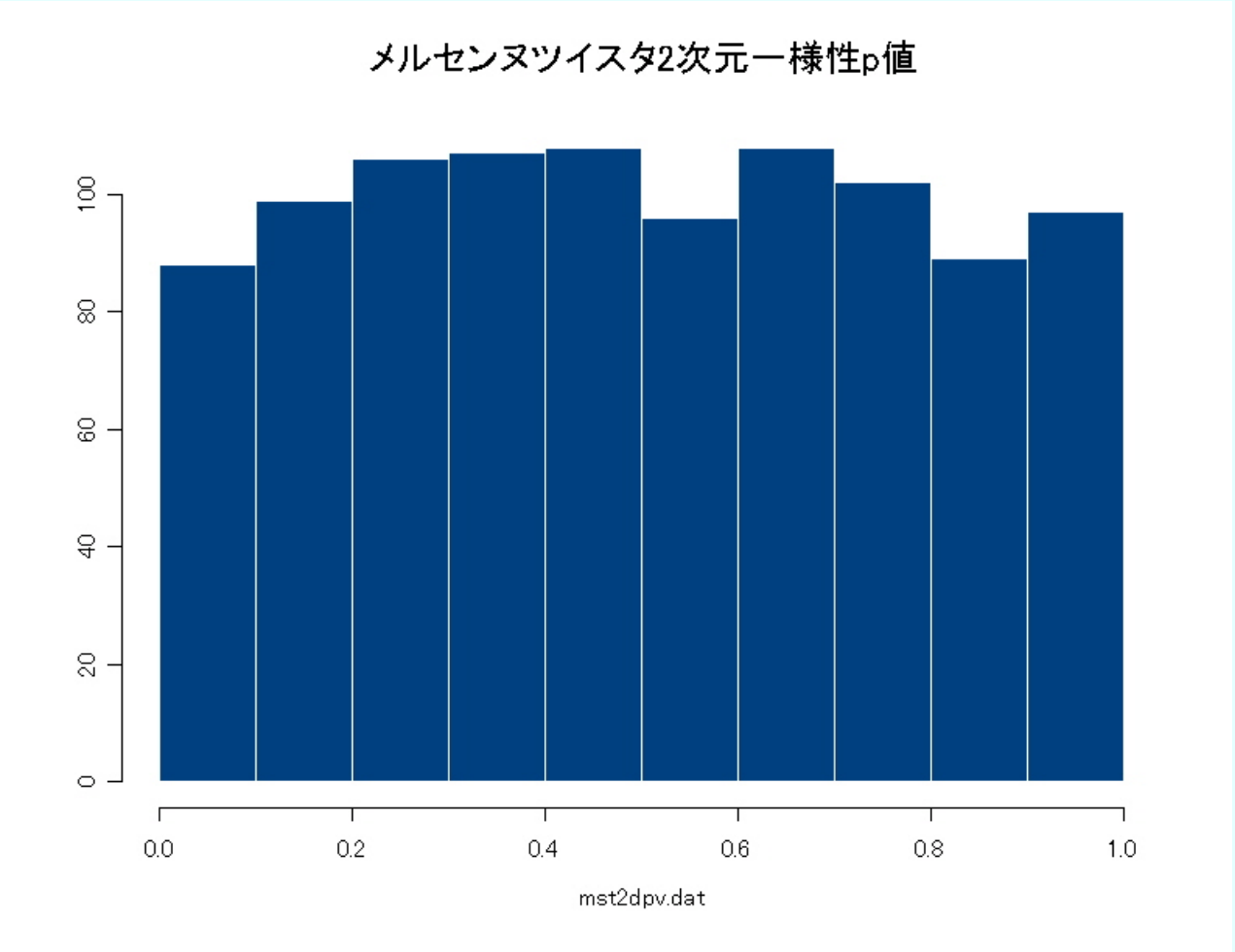
メルセンヌツイスターの検定結果(5) p値の変化



メルセンヌツイスターの検定結果(6) p値の変化



メルセンヌツイスターの検定結果(7) p値の変化



目次

1. 統計数理研究所の計算機
2. ヒストグラムの描き方
3. 仮説検証型からモデル選択型へ
4. 田村が関係した研究
5. データマイニング、可視化
6. 物理乱数について
7. 産学官連携会議（2004年6月）のポスター

産学官連携会議（2004年6月）のポスター

•アナログデータ処理とNeuron MOSによる移動無線・高速データ通信（瀧澤由美・統計数理研究所・予測制御研究系）

•効果的マイクロマーケティング活動実現のための販売データの高度活用（佐藤忠彦・（財）流通経済研究所（総合研究大学院大学・統計科学専攻・2004年3月修了））

•RBF-ARXモデルを基にする非線形システムへの統計的モデリングと制御（彭輝 ・統計数理研究所・客員教授）

•マイクロマーケティングへの統計科学からの接近
（樋口知之・統計数理研究所・予測発見戦略研究センター）

産学官連携会議（2004年6月）のポスター(続き)

•非侵襲的脳機能計測-人間の脳機能解明へ向けて-
（山下宙人・統計数理研究所・PD研究員）

•ICAのMEGへの適用
（池田思朗・統計数理研究所・調査実験解析研究系）

•音声に含まれる様々な情報の利用―新しい統計手法に基づく話者認識―（松井知子・統計数理研究所・統計計算開発センター）

•金融時系列分析においてモンテカルロフィルタが有効な例
（佐藤整尚・統計数理研究所・予測制御研究系）