統計的データ解析と乱数の利用

統計数理研究所
統計計算開発センター
田村義保

国立天文台

2002年12月12日

アウトライン

1. 数理統計学から計算統計学
    - ブートストラップ法の登場
    - マルコフ連鎖モンテカルロ法の登場
    - モンテカルロフィルタ（パーティクルフィルタ）の登場
2. 乱数の利用
    - 擬似乱数
    - 準乱数
    - 物理乱数
3. さらなる発展へ
    - グリッドコンピューティング
    - 高速通信網
    - テラフロップスからペタフロップス

1. 数理統計学から計算統計学

ブートストラップ法の登場

例えば、データ

$$x_1, x_2, \cdots, x_n$$

に対して母平均 μ の信頼区間 95%の区間推定は

$$\bar{x} - t_{0.025}(n-1)s/\sqrt{n} \leq \mu \leq \bar{x} + t_{0.025}(n-1)s/\sqrt{n}$$

となる。ただし、$\bar{x}$ は標本平均、s^2 は不偏標本分散、$t_{0.025}(n-1)$ は自由度 $n-1$ の t 分布の上側 2.5%点である。この式は、母集団が正規分布すると仮定することにより得られる。

スタンフォード大学の Efron 教授が 1978 年にリサンプリング法の一つとしてブートストラップ法を提唱した。

ブートストラップ法での推定は与えられたデータから復元抽出法により、m 組の標本を作り、それぞれの平均を求める。

$$\begin{array}{llll}
\text{標本}1 & x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)} & \rightarrow & \text{標本平均}1 \quad \bar{x}^{(1)} \\
\text{標本}2 & x_1^{(2)}, x_2^{(2)}, \cdots, x_n^{(2)} & \rightarrow & \text{標本平均}2 \quad \bar{x}^{(2)} \\
 & \cdots\cdots\cdots & & \\
\text{標本m} & x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)} & \rightarrow & \text{標本平均m} \quad \bar{x}^{(m)}
\end{array}$$

$m = 1000$ の場合には、$\bar{x}^{(1)}, \bar{x}^{(2)}, \cdots, \bar{x}^{(1000)}$ の小さい方から、26 番目の値を $\bar{x}_{(26)}$、975 番目の値を $\bar{x}_{(975)}$ とすると

$$\bar{x}_{(26)} \leq \mu \leq \bar{x}_{(975)}$$

で信頼区間 95%の区間推定を行うことができる。（もっと正確には、下限は 25 番目の値と 26 番目の値にそれぞれ 1 と 39 の重みを与えた重みつき平均である。上限は 975 番目と 976 番目に 39 と 1 の重みをつけて平均する。）母集団の形を仮定しなくて良いことや平均のような簡単な統計量ではなくもっと複雑な統計量にも用いることができるという利点がある。復元抽出のために乱数を用いる。1000 個の標本を作るのに 1000 台の計算機があれば並列計算することが可能で計算時間は 1/1000 近くに短縮できる。

マルコフ連鎖モンテカルロ法の登場

もともと、マルコフ連鎖モンテカルロ法（MCMC）は統計物理や画像処理で使われていた手法である。1990 年頃からベイズ統計学においてパラメータを推定するために用いられるようになっている。

ギブスサンプラー

未知パラメータを $\theta = (\theta_1, \cdots, \theta_m)$、データを Y、事後分布を $\pi(\theta|Y)$ とする。

1. 初期値 $(\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_m^{(0)})$ を用意する。
2. $\theta_1^{(1)}$ を $\pi(\theta_1|\theta_2^{(0)}, \cdots, \theta_m^{(0)}, Y)$ から発生させる。
3. $\theta_2^{(1)}$ を $\pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \cdots, \theta_m^{(0)}, Y)$ から発生させる。
4. $\theta_3^{(1)}$ を $\pi(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \cdots, \theta_m^{(0)}, Y)$ から発生させる。同様に、$\theta_4^{(1)}, \cdots, \theta_m^{(1)}$ を発生させる。
5. 一般に $\theta^{(i)} = (\theta_1^{(i)}, \cdots, \theta_m^{(i)})$ が得られたら
   (a) $\theta_1^{(i+1)}$ を $\pi(\theta_1|\theta_2^{(i)}, \cdots, \theta_m^{(i)}, Y)$ から発生させる。
   (b) $\theta_2^{(i+1)}$ を $\pi(\theta_2|\theta_1^{(i+1)}, \theta_3^{(i)}, \cdots, \theta_m^{(i)}, Y)$ から発生させる。
   (c) $\theta_3^{(i+1)}$ を $\pi(\theta_3|\theta_1^{(i+1)}, \theta_2^{(i+1)}, \theta_4^{(i)}, \cdots, \theta_m^{(i)}, Y)$ から発生させる。同様に、$\theta_4^{(i+1)}, \cdots, \theta_m^{(i+1)}$ を発生させる。

ステップ 5 を繰り返し、$N \to \infty$ のとき、$\theta^{(N)}$ が事後分布の標本となる。

例　正規分布の場合

$$x_i, \cdots, x_n \sim \text{i.i.d.} N(\mu, \sigma^2)$$

$$\pi(\mu) \propto du$$

$$\pi(\sigma) \propto d\sigma/\sigma$$

事後分布

$$\pi(\mu, \sigma) \propto \sigma^{-n-1} \exp\left(\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right)$$

標本の発生

$$\mu|\sigma^2, x \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

$$\sigma^{-2}|\mu, x \sim \text{Gamma}\left(\frac{n}{2}, \frac{2}{\sum_{i=1}^n (x_i - \mu)^2}\right)$$

モンテカルロフィルタ（パーティクルフィルタ）の登場

状態空間表現

$$\begin{array}{ll} \text{システムモデル} & x_n = Fx_n + Gv_{n-1} \\ \text{観測モデル} & y_n = Hx_n + w_n \end{array}$$

ノイズがガウス分布の場合はカルマンフィルタの利用

一般化状態空間表現

$$\begin{aligned} x_n &\sim F(\cdot|x_{n-1}) \\ y_n &\sim H(\cdot|x_n) \end{aligned}$$

ただし

$$F(x_n|x_{n-1}) = p(x_n|x_{n-1}, \cdots, x_0, y_{n-1}, \cdots, y_1)$$

$$H(y_n|x_n) = p(x_n|x_n, \cdots, x_0, y_{n-1}, \cdots, y_1)$$

分布のモンテカルロ近似

$$\begin{array}{ll} \text{予測分布} & \{p_n^{(1)}, \cdots, p_n^{(m)}\} \sim p(x_n|Y_{n-1}) \\ \text{フィルタ分布} & \{f_n^{(1)}, \cdots, f_n^{(m)}\} \sim p(x_n|Y_n) \end{array}$$

2. 乱数の利用

擬似乱数

1. 平方採中法：$2m$ 桁の乱数の場合、x_{n+1} は $4m$ 桁となる x_n^2 の真ん中の $2m$ 桁とする。ほとんど使われないが、最近までは JIS 標準であった。
2. 線形合同法：$x_{n+1} = ax_n + c(\mathrm{mod}m)$ で発生させる。a、c の選び方が重要。しかし、どのように選んでも、有限の周期となる。$m = 10$、$x_0 = a = c = 7$ とすると

   $$7, 6, 9, 0, 7, 6, 9, 0, \cdots$$

   となる。$c = 0$ の場合は乗算合同法、$c \neq 0$ の場合は混合合同法と呼ばれている。
3. M 系列 (by 伏見)
4. メルセンヌツイスタ (by 松本)
5. 新しい擬似乱数 (by 三浦)

準乱数

一様性にのみに注目し、乱数の持つべきその他の性質については目をつむっている。いわゆるモンテカルロ積分を行うのに適しており、金融関係の人は好きであるように思える。

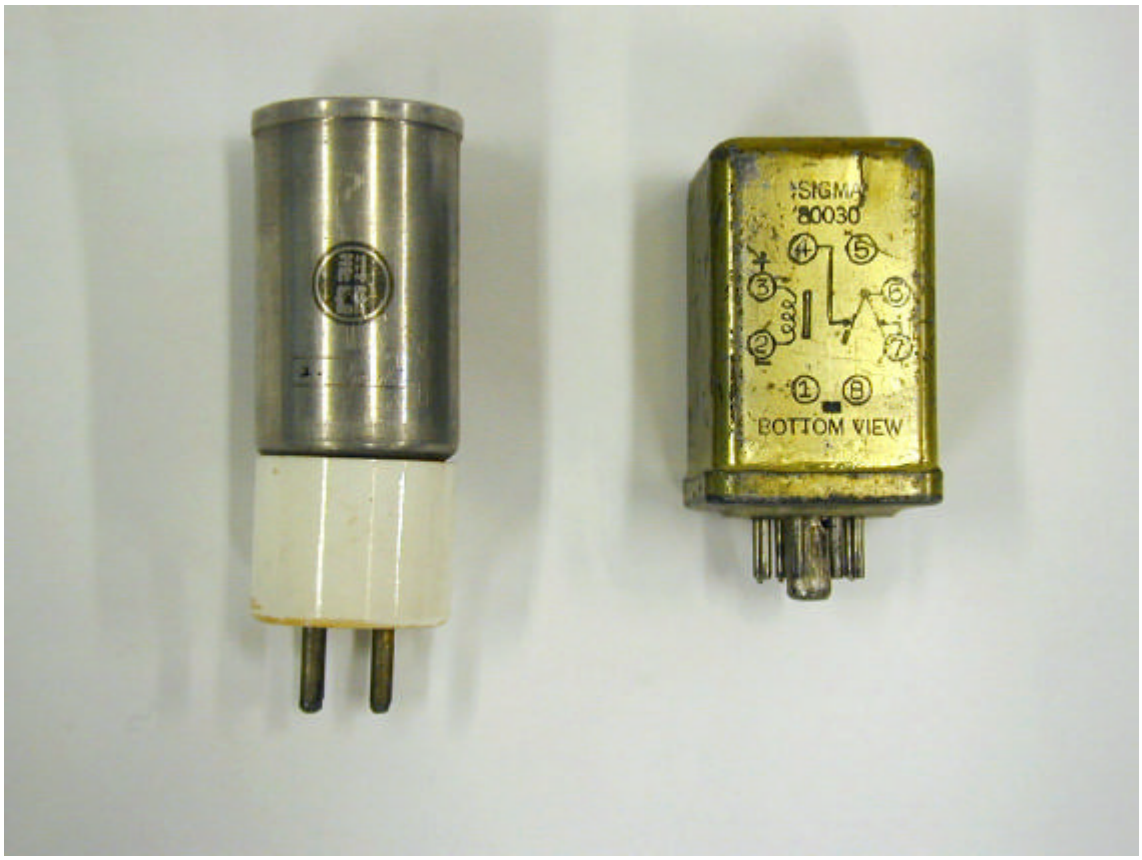
物理乱数

1. サイコロ・ルーレット・宝くじの円板
2. 例えば、0 から 99 の番号を書いた多くのコインを袋に入れて選び出す。
3. 放射線の計測
4. 散乱光の光電子増倍管での計測
5. 電気回路の熱雑音の増幅

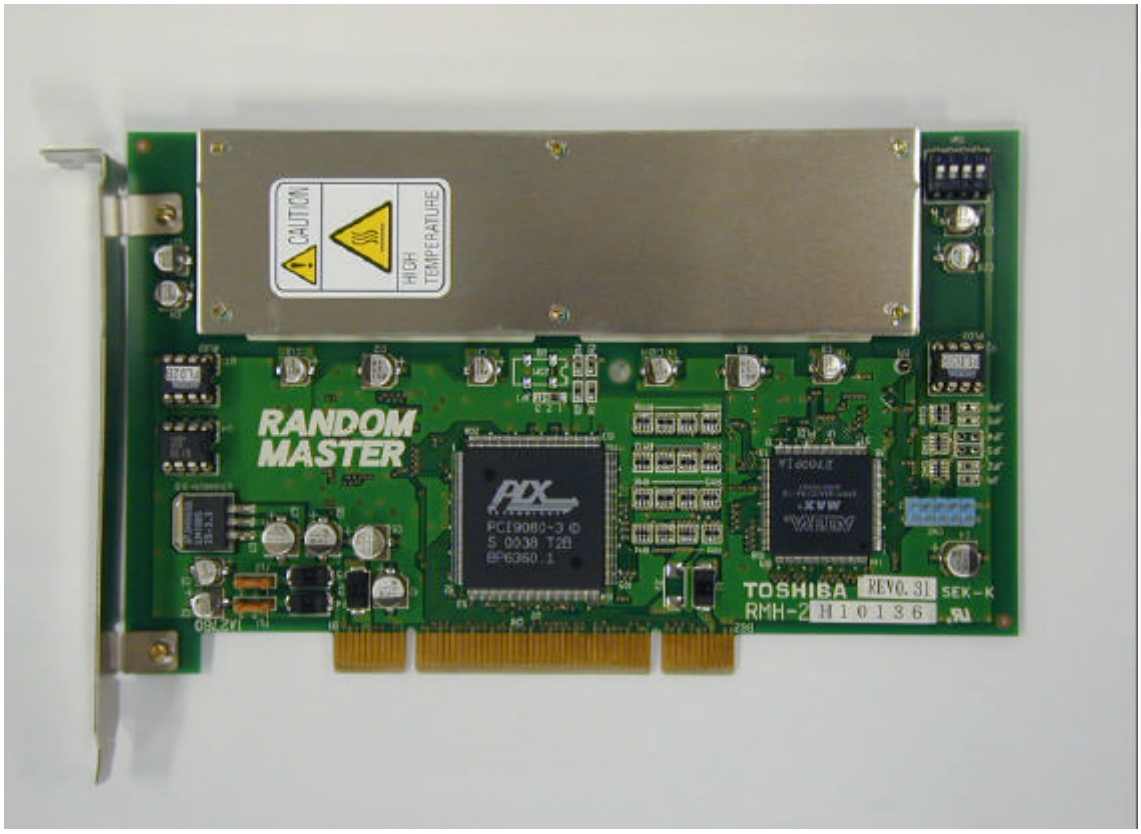
物理乱数の歴史

統計数理研究所が乱数表を作った時のコイン

統計数理研究所の初代の物理乱数発生装置（放射線を計数）の部品



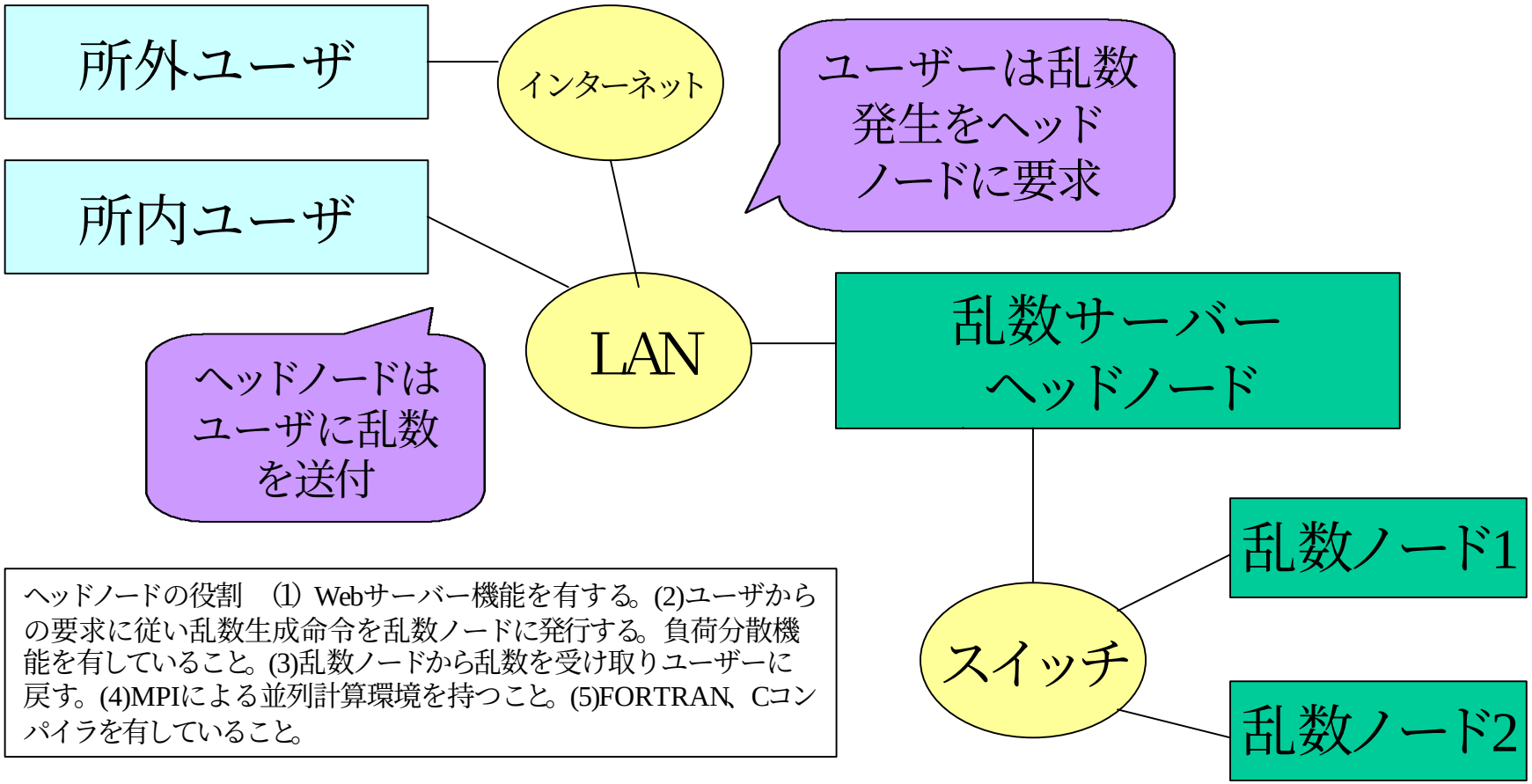
市販されている物理乱数発生ボード（電気回路の熱雑音を増幅。30MB/s の発生速度）



乱数提供サーバー仕様案(2002年10月２９日)

- 目的　所内外からWebブラウザを用いて、本システムにアクセスし、乱数をオンデマンド形式で獲得する。
- 位置づけ　本年度の導入はプロトタイプ開発を目的としているが、将来の本格運用に備えたソフトウェア開発を必要とする。
- 提供企業の条件　仕様に示すハードウェアとソフトウェアを全て1社が窓口となり提案すること。

乱数提供サーバー機能イメージ図



ヘッドノードの役割　⑴ Webサーバー機能を有する。(2)ユーザからの要求に従い乱数生成命令を乱数ノードに発行する。負荷分散機能を有していること。(3)乱数ノードから乱数を受け取りユーザーに戻す。(4)MPIによる並列計算環境を持つこと。(5)FORTRAN、Cコンパイラを有していること。

乱数ノードの役割　(1)東芝ランダムマスターによる乱数発生機能を持っていること。(2)MPIによる系列計算機能を有していること。(3)FORTRAN、Cで開発されたプログラムの実行環境を有していること。

3. さらなる発展へ

グリッドコンピューティング

計算グリッド、データグリッド、装置グリッドなどある。データグリッドがもっとものになりそうな気がしている。

高速通信網

現在、日本の先端研究機関は10Gbps（家庭のADSLの1000倍）の通信網で接続されている。グリッド研究のために用いられているが、40Gbpsでグリッド網を作ろうとしている計画も多数ある。

テラフロップスからペタフロップス

現時点で世界最高速の計算機は、40TFlopsの地球シミュレータである。ペタフロップスに向けて前進ちゅであるが、重要な点は高速通信機構（インターネットのような対外接続ではなく、CPU同士の接続）である。Cray、IBM、NEC、Fujitsuなどがしのぎを削っている。Pentium4を500,000台つなげばPFlopsになるが現実性はない。また、パソコンクラスタの場合は、台数を増やしても通信のボトルネックでそれほど速くならない。ブートストラップ法、モンテカルロフィルタ、MCMCは通信が少なくてすむのでパソコンクラスタ向きである。