プレイマイコン・シリーズ ❷
刀根薫監修

古林隆著

培風館

統計解析

プレイマイコン・シリーズ ❷

刀根薫監修

統計解析

古林隆著

培風館

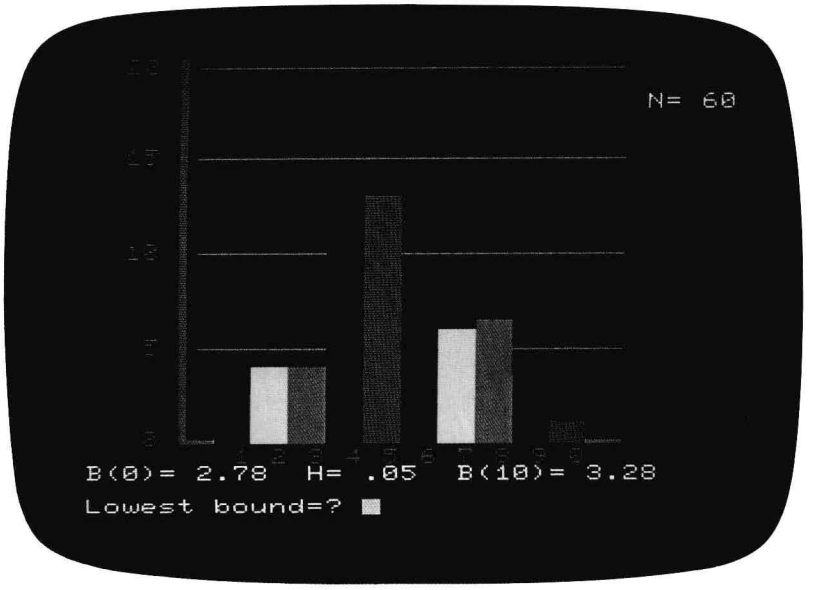

図1　ヒストグラム

（pp. 46～48 参照）

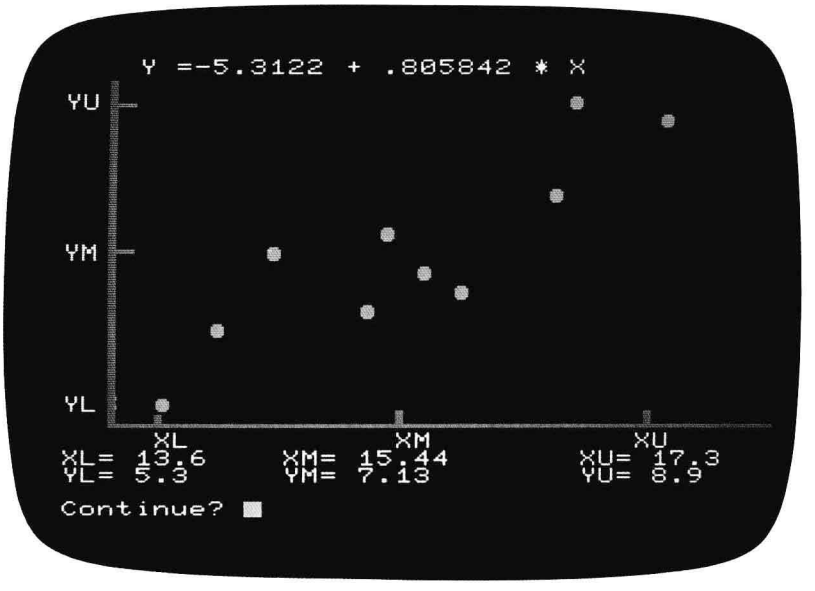

図2　2次元データと回帰直線

（pp. 156～159 参照）

刊行に際して

マイコン革命といわれるように，マイコンの登場以来，コンピュータは我々のより身近な存在となった。

私事にわたって恐縮ですが，私とコンピュータとの付き合いも，かれこれ四半世紀におよぶ。IBM の 650 (真空管) から始まって NEAC 2203, HITAC 301, FACOM 222, 230 などの国産機，UNIVAC 1107, IBM 360, 370, CDC 6600 などの外国機，そして大型から小型までの多数のコンピュータを使ってきた。バッチ処理，リモートバッチ，TSS と使い方もさまざまである。正しいコンピュータ利用の普及をめざして，NHK コンピュータ講座で，お茶の間に浸透させる努力もやってみた。最近では，スーパーコンピュータ CRAY-1 (LSI) を使って仕事をしている。たしかに，コンピュータはすばらしい性能をもつに至った。高速，大容量を目標に，日進月歩の革新が進んでいる。

ところで，このような技術進歩に支援されて，全く別方向の芽が吹き出したのである。カリフォルニアはシリコン・バレーの一角からあがったマイクロコンピュータの火の手は，あっという間に全世界を席巻する。革命である。もとより，革命を迎え入れる素地はあった。四半世紀におよぶコンピュータ利用の間に，その素地は着実に培われていたのである。

マイコンの理念はなんであろうか。それは**人間向き**ということに尽きよう。マイコンを使っていると，従来機を使う場合と違い，皮膚感覚で接していることに気づく。安心して間違えることができる。なに，間違えれば直せばよい。それも即座に。間違いやすいのは人間の性である。コンピュータ人間など願い下げと行こう。人間には個性があり，画一化をきらう。特に発想段階ではそうである。個性的な情報処理が要求される。構想を具体化していく，いわゆるモデル作りの段階では，人間とコンピュータの対話が必要であり，そのとき，高速や大容量といった性能は不要である。むしろ，必要に応じて常時呼び出し

が可能であり，修正操作が即時にできることの方が重要である。また，人間には視覚的な面が強く，そのような対応がコンピュータ側に要望されるが，マイコンのモニターテレビは，それにうってつけである。

こういったパーソナルな面の要求にマッチすればこそ，マイコンの爆発的な普及が実現しつつあるのだ。値段の安さももちろん一因ではある。そのうえ，家庭用の電源で間に合うし，省エネルギー時代にふさわしいコンピュータであるといえよう。

こうしてみてくると，マイコンは一種の文化ショックを現代社会に引き起こしながら，これまでのコンピュータとは異次元の世界で，新しい展開を始めているのに気づく。教育，研究，事務処理はもとより，経営，予測，生産，流通，行政，マスコミ，デザイン，レクリエーションといった各界に，急速に浸透して行くはずである。この小さな傑作は，それが始めて作られたシリコン・バレーのインベーダーゲーム用のオモチャから変身して，いまや，文明社会のすみずみまでインベードしつつある。気がついてみたら，あるいは気がつかないうちに，我々はその影響下に暮らしているということになるだろう。

さて，こういった魔力を活かすも殺すも，ソフトウェアしだいである。せっかくの魔法の杖を有効に使うために，各方面でソフトの整備が要望されている。

本シリーズを企画した目的も，上記の時代の流れに沿ったテキスト作りにある。

幸い，メーカー各社の御協力をいただき，友人達の快い御援助と培風館の御好意の下で，シリーズ発刊の運びとなった。感謝にたえない。

シリーズの性格上，はじめから巻数やテーマを限定することなく，いわばオープンエンドで，先駆的な仕事を紹介して行きたいと思う。

あたたかくて，血の通ったコンピュータの本を，世に送り出したいと願う次第である。

1981年 3月

刀根 薫

はしがき

本書は，マイコンを使って，基本的な統計解析を行なおうとする場合に必要である BASIC によるプログラムを提供するものであるが，単なるプログラム・ライブラリーの解説書ではない．マイコンを使いながら統計解析の手法を学習するのにも利用できるように，第1章で数理統計学における考え方を簡潔に解説するとともに，第3章以下の解析手法の解説においては，できるだけ実際的なデータの例を示し，それに手法を適用したときの結果を示した．

マイコンをすぐに利用できる方は，とにかくここに書かれたプログラムを入れて，実行していただきたい．従来のバッチ処理のプログラムと異なり，コンピュータとの会話を通して，統計解析手法の目的，使い方，制約(限界)が習得できることであろう．

これからは，ディスプレイ(テレビ受像機)に示されたヒストグラム等のグラフや分散分析表などの表を見ながら(見せながら)，統計解析を学習する(教育する)時代になるであろう．本書の読者が，そのような時代の先端を進まれることを期待している．

なお，本書では，BASIC の解説はほとんど行なっていない．BASIC についての予備知識のない読者は，本シリーズ第1巻『BASIC』を参照されたい．

最後に，数理統計学に関して，学生の頃からご指導いただいている東京理科大学 朝香鉄一教授(東京大学名誉教授)，電気通信大学 森口繁一教授(東京大学名誉教授)，名古屋大学 吉村功助教授および執筆の機会を与えて下さった埼玉大学 刀根薫教授に，厚く御礼申しあげます．また出版にあたり，いろいろお世話をいただいた培風館の牧野末喜氏，近藤勝子さんにも謝意を表する次第です．

1981年10月

著　者

目　次

プログラムのカタログ

1 統計解析の基礎知識

1.1 記述統計と推測統計

統計 (statistics) とは，いくつかの「もの」のある特性 (1 種類でなくともよい) の測定値または観測値の集まりといってよいであろう†．データもほとんど同じ意味で使われることがある．統計から，測定された「もの」全体の (集団的) 特徴を知ることができる．たとえば，国勢調査によって得られた膨大な統計から，日本の人口，世帯の大きさ (人数) の分布，年齢分布等々がわかる．プロ野球のある選手の打撃成績にしても，全試合の成績の統計から，本塁打，打率，打点等，その選手の全体的な打撃成績を表わす数値が得られる．また，ある会社で，新入社員 (数百人いるとしよう) に制服を与えるために全員の体形測定を行なったとする．一人ずつあわせるわけにはいかないので，全員をできるだけ似た体形の者が集まるように，いくつかのグループに分けることがある．このように，統計を解析して，測定の対象になった「もの」全体の特徴を表わす方法を研究することを**記述統計学**という．

これに対し，測定された「もの」が，測定される可能性のあったもの全体の中の一部であって，一部の測定値から，全体の特徴を知ろうとすることがある．たとえば，ある工場で製造された製品のロットからいくつか抜き取って，それらを検査するのは，ロット全体の不良率がどれぐらいであるかを知りたいためであるし，選挙の前に，有権者の中から何人かを選んで支持政党を尋ねるのは，有権者全体の中での割合を知りたいからである．また，初秋になると，稲作地では，いくつかの稲を選んで，それらの育ち具合を調べるのは，その地域全体の育ち具合を知りたいからであろう．(全部調べたら，収穫する分がなくなる．)

---

† ここでの「もの」は，測定 (観測，調査) の対象になりうるもののことである．したがって，いわゆる物体に限らず，人や地域のこともある．

全体のことを**母集団** (population) といい，それから測定するために選ばれた「もの」の集まりを**標本** (sample) という．標本は母集団の一部にすぎないから，標本から母集団の特徴を完全に知ることは，一般に不可能である．したがって，母集団の特徴を推測することになる．標本から母集団の特徴を推測することを**統計的推測** (statistical inference) といい，統計的推測の理論や手法を研究するのが**推測統計学**である．

推測統計学では，母集団から選ばれる標本を確率的現象とみなす．すなわち，標本は，いくつかの標本になりうるもの——**根元事象**という——の中から選ばれるが，すべての根元事象に対し，それが選ばれる(起こる)確率が与えられているものとする．一例をあげよう．ロットの中に 1000 個の製品がはいっていて，1 から 1000 までの番号がついているときに，2 けたの乱数 ab をひいて†，下 2 けたが ab である番号の製品 10 個を標本にすることにしよう．たとえば，$ab=25$ とすると，{25, 125, 225, 325, 425, 525, 625, 725, 825, 925} の(番号のついた)製品 10 個が標本である．このとき，下 2 けたが同じである 10 個の製品の組合せ(全部で 100 通りある)が根元事象であって，いずれも選ばれる確率は 1/100 である．また，3 けたの相異なる 10 個の乱数をひいて，それらを番号とする製品††を選んで標本にするならば，1000 個から 10 個を選ぶすべての組合せ (全部で ${}_{1000}C_{10}$ 通り存在する) が根元事象であって，いずれも選ばれる確率は等しい．すべての組合せが根元事象になるように標本を選ぶことを**ランダム抜取**または**ランダム・サンプリング** (random sampling) という．

母集団の特徴を推測する場合，標本そのものではなく，標本から得られる数値に関心がある場合が多い．たとえば，前出のロットから選ばれた 10 個の部品にしても，ロット (母集団) の不良率を推測するためには，10 個の部品のうちの不良品の数さえわかれば十分であって，不良品は何番であったかという内訳は必要ではない．(理論的な説明は非常に複雑である．直感的に理解していただきたい．) 根元事象に対応する数は**確率変数**であるから，実際に選ばれた標本から求められた数値は**確率変数の実現値**である．たとえば，ロットから選ばれた 10 個の中に，3 個不良品があったとすると，ロットから(ランダムに) 10 個選ぶとき，その中に含まれる不良品の数 X は確率変数であって，3 は X の実現値である．確率変数は確率分布をもっている．そしてその確率分布は，母集団(と標本の選び方)によって定まる．前出の X の分布は，

$$\Pr[X=x]=\frac{{}_{1000p}C_x \cdot {}_{1000(1-p)}C_{10-x}}{{}_{1000}C_{10}} \qquad (x=0, 1, \cdots, 10) \qquad (1.1)$$

である．ここで，p はロットの不良率であって，母集団の特徴を表わす未知定数であるが，これによって X の分布が表わされている．母集団またはそれに対

† 「2 けたの乱数をひく」とは，100 組の 2 けたの数字 (00, 01, ⋯, 99) の中から，どれも確率 1/100 で選ばれるようにして，1 組の 2 けたの数字を選ぶことをいう．
†† 乱数 000 は 1000 番に対応させる．

応する確率分布の特徴を表わす未知定数を**母数**または**パラメータ** (parameter) という．以上より，統計的推測とは，未知定数を含む確率分布(母集団に対応)に従う確率変数の実現値(標本に対応)から未知定数を推測すること，ということができる．

このように表わすと，サイコロをふったときにでた目の数のように，母集団が実在しないけれども，確率分布が対応づけられる統計は，標本とみて，統計的推測を行なうことができる．サイコロの場合，i $(i=1, 2, \cdots, 6)$ の目がでる確率を p_i とすると，$1, 2, \cdots, 6$ と記したカードを，それぞれ $Np_1, Np_2, \cdots, Np_6$ 枚(計 N 枚)用意して，そこから1枚抜くことは，サイコロを1回ふることに完全に対応するから，このような N 枚のカードからなる母集団を想定することができる．確率分布から想定できる母集団を**仮想母集団** (hypothetical population) という．

また，確率分布は，現実に完全にあっていなくても，解析しやすいものを用いることが多い．たとえば前出の例で，X の分布は，厳密には(1.1)で与えられるが，これを次の分布で近似することが多い．

$$\Pr[X=x]={}_{10}\mathrm{C}_x\, p^x (1-p)^{10-x} \qquad (x=0, 1, 2, \cdots, 10) \qquad (1.2)$$

(1.1)を**超幾何分布**，(1.2)を**二項分布**という．超幾何分布に比べて，二項分布のほうが，確率の計算が容易である．多少現実から離れていても，解析できる形，あるいは，解析しやすい形で表わすほうが実用的である．

なお，基本的な確率分布については次節で解説する．

1.2 基本的な確率分布

1.2.1 離散分布

基本的な離散分布を紹介し，その特徴を示す．なお確率関数(値 x をとる確率)を $p_r(x)$ で表わすことにする．(多次元分布のときは，x をベクトルとみなす．)

(1) **超幾何分布** (hypergeometric distribution)

$$p_r(x)=\frac{{}_M\mathrm{C}_x \cdot {}_{N-M}\mathrm{C}_{n-x}}{{}_N\mathrm{C}_n} \qquad (x=0, 1, 2, \cdots, n) \qquad (1.3)$$

ただし，N, M, n は正整数であり，$M<N, n<N$ とする†.

$$\text{平均}\ \frac{nM}{N}, \qquad \text{分散}\ \frac{nM}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n}{N}\right)$$

$N=100$, $M=30$, $n=10$ のときの確率関数のグラフを図1.1に示す．

特徴 A を有する「もの」M 個と A を有さない「もの」$(N-M)$ 個からなる母集団から，ランダムに n 個抜き取るとき，その中に A を有する「もの」

† $i<j$ のとき，${}_i\mathrm{C}_j=0$ とする．

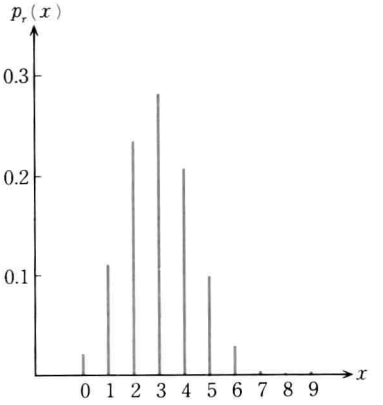

図 1.1　$N=100$, $M=30$, $n=10$ のときの超幾何分布

の数の分布が超幾何分布である．たとえば，ロットから抜き取られる製品の中の不良品の数の分布は超幾何分布である．

（2）　**二項分布**(binomial distribution)

$$p_r(x)={}_nC_x\,p^x(1-p)^{n-x} \qquad (x=0,1,2,\cdots,n) \qquad (1.4)$$

ただし，n は正整数であり，$0<p<1$ とする．

平均 np,　　分散 $np(1-p)$

この分布を $B(n,p)$ で表わす．また，p を**二項確率**(binomial probability)ということもある．

$B(10,\ 0.3)$ の確率関数のグラフを図 1.2 に示す．これは，図 1.1 のグラフに非常に近い．グラフでは，違いがわかりにくいので，確率を表 1.1 に示す．

事象 A が起こる確率が p である試行を独立に n 回行なうとき，事象 A

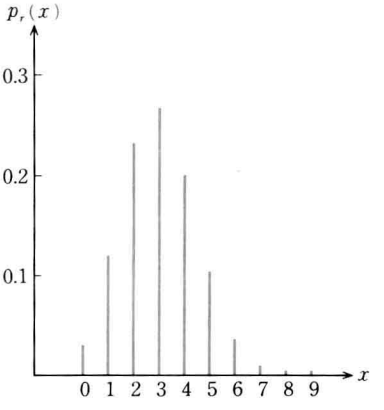

図 1.2　$B(10, 0.3)$ の確率関数

表 1.1 超幾何分布と二項分布の確率

x	超幾何分布 $N=100, M=30, n=10$	二項分布 $n=10, p=0.3$
0	0.023	0.028
1	0.113	0.121
2	0.237	0.233
3	0.281	0.267
4	0.208	0.200
5	0.100	0.103
6	0.031	0.037
7	0.006	0.009
8	0.001	0.001
9	0.000	0.000

が起こる回数の分布が $B(n, p)$ である．たとえば，正しいサイコロを n 回ふるとき，1の目がでる回数の分布は $B(n, 1/6)$ である．

（a） X_i $(i=1, 2, \cdots, n)$ は，確率 p で1をとり，確率 $(1-p)$ で0をとり，互いに独立であるとすると，それらの和 $\sum X_i$ の分布は $B(n, p)$ である．

（b） 超幾何分布(1.3)で，$M/N=p$（一定）として，$N\to\infty$ とすると，$B(n, p)$ に近づく．

したがって，大きさが N であるロットから n 個抜き取るとき，その中に含まれる不良品の数の分布は，N が n に比べてかなり大きければ，二項分布とみなしてよい．

（3） 多項分布(multinomial distribution)

k 個$(k\geqq 3)$の確率変数 $X_1, X_2, \cdots, X_k$ の(同時)確率が次式で与えられる分布．

$$p_r(x_1, x_2, \cdots, x_k)=\frac{n!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}$$

$$\left(x_i=0, 1, \cdots, n, \quad \sum x_i=n\right)$$

ただし，n は正整数であり，$0<p_i<1$, $\sum p_i=1$ とする．

X_i の期待値　　np_i

X_i の分散　　$np_i(1-p_i)$

X_i と X_j $(i\neq j)$ の共分散　$-np_ip_j$

1回の試行で，k 個の事象 $A_1, A_2, \cdots, A_k$ のいずれか1個だけが起こり，A_i が起こる確率は p_i であるとする．この試行を n 回行なうとき，$A_1, A_2, \cdots, A_k$ が起こる回数の同時分布が多項分布である．たとえば，サイコロを何回かふるとき，1の目，2の目，$\cdots$，6の目がでる回数の分布は，$k=6$ の多項分布である．大きさが N である有限母集団から n 個抜き取って，それぞ

れが k 個の特性 $A_1, A_2, \cdots, A_k$ のうちのどれを有しているかを調べるとき，それぞれの特性を有する「もの」の数の同時分布は，N が n に比べてかなり大きければ，多項分布とみなしてよい(例9.3参照)．

1.2.2 連続分布

基本的な連続分布を紹介し，その特徴を示す．なお，確率密度関数を $f(x)$ で表わす．(多次元分布のときは，x をベクトルとみなす．)

(1) **正規分布**(normal distribution)

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

ただし，$\sigma>0$ とする．

平均 μ,　　分散 σ^2

この分布を $N(\mu, \sigma^2)$ で表わす．密度関数のグラフを図1.3に示す．とくに，$N(0, 1)$ を**標準正規分布**という．

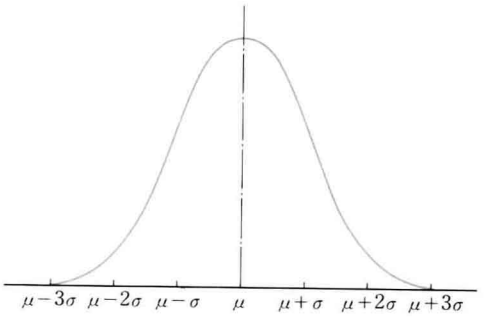

図 1.3　$N(\mu, \sigma^2)$ の密度関数

(a)　X が $N(\mu, \sigma^2)$ に従うとき，

$$U=\frac{X-\mu}{\sigma}$$

は $N(0, 1)$ に従う．

すなわち，X の累積分布関数と $N(0, 1)$ の累積分布関数の間には，次の関係が成り立つ．

$$P_N(u)=\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2}\right]dt$$

とおくと，

$$\Pr[X\leqq x]=P_N\left(\frac{x-\mu}{\sigma}\right) \tag{1.5}$$

なお，$N(0, 1)$ の上側累積確率 $1-P_N(u)$ を $Q_N(u)$ で表わすことにする．また，P $(0<P<1)$ に対して $Q_N(u)=P$ となる u を u_P で表わすことにす

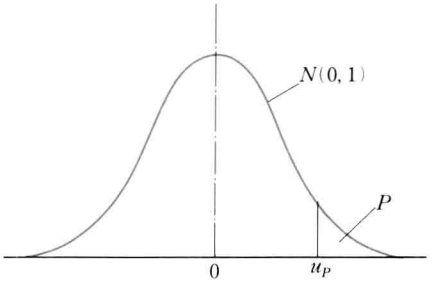

図 1.4 $N(0,1)$ の上側パーセント点

る(図 1.4). u_P を標準正規分布の**上側 100P パーセント点**という.

おもな u に対する $Q_N(u)$ および P に対する u_P を表 1.2 に示す. (計算法については, 第 4 章で述べる.)

表 1.2 $Q_N(u)$ と u_P

u	$Q_N(u)$	P	u_P
0	0.5000	0.05	1.645
0.5	0.3085	0.025	1.960
1	0.1587	0.01	2.326
1.5	0.0668	0.005	2.576
2	0.0228		
2.5	0.0062		
3	0.0013		

(b) X_i $(i=1, 2, \cdots, n)$ が互いに独立に $N(\mu_i, \sigma_i^2)$ に従うとき, $\sum a_i X_i$ (a_i は定数)は $N\left(\sum a_i\mu_i, \sum a_i^2\sigma_i^2\right)$ に従う.

(2) **カイ二乗分布** (Chi-square distribution または χ^2 distribution)

$$f(x)=\frac{1}{2\Gamma\left(\frac{f}{2}\right)}\left(\frac{x}{2}\right)^{\frac{f}{2}-1}e^{-\frac{x}{2}} \qquad (x>0)$$

ただし, $f>0$ である†.

平均 f, 分散 $2f$

f を**自由度** (degree of freedom; DF と略記することがある)という. $f=5$, 10 のときの密度関数のグラフを図 1.5 に示す. ($f\leqq 2$ のときは, 単調減少になる.)

(a) $U_1, U_2, \cdots, U_k$ が互いに独立に $N(0,1)$ に従うとき,

† $\Gamma(z)$ はガンマ関数であり,

$$\Gamma(z)=\int_0^\infty y^{z-1}e^{-y}\,dy \qquad (z>0)$$

である. z が正整数であるとき, $\Gamma(z)=(z-1)!$ である.