②

グラフィック端末を用いた統計教育

井　上　隆　勝，安　田　正　實，星　守
田　栗　正　章

Using Interactive Computer Graphics for Statistical Education

by

Takakatsu Inoue,* Masami Yasuda**,
Mamoru Hoshi* and Masaaki Taguri***

千葉大学工学部研究報告

第34巻　第２号

(1982年?)
57年度

JOURNAL
OF
FACULTY OF ENGINEERING
CHIBA UNIVERSITY
VOL. 34 NO. 2

Using Interactive Computer Graphics for Statistical Education

by

Takakatsu Inoue*, Masami Yasuda**, Mamoru Hoshi* and Masaaki Taguri***

SUMMARY

A new development of CHESS (CHiba Educational System for Statistics) is presented in view of the TSS graphic terminal, mainly written by FORTRAN. This is for benefit not only of the students but also of research members. CHESS system is still developing into fullness and with refinement.

1. INTRODUCTION

From the needs of data analysis in various fields, statistical education and computer education for university students and members of society must be made considerable effort to accomplish it. Although the hardware of computers have been advanced remarkably, curricula of the education in Japan are neither provided effectively nor promoted the adequate project. It is now important to develop a statistical education having deep relevance to the computer (see *Table 1.* It has a close relation between Statistics and Information Science). Our objective in this paper is to persuade it to plan a new enterprising material by the computer software in the university's course such as statistics, numerical analysis, mathematics, information science, etc.*(Fig.1)*. Also we intend to

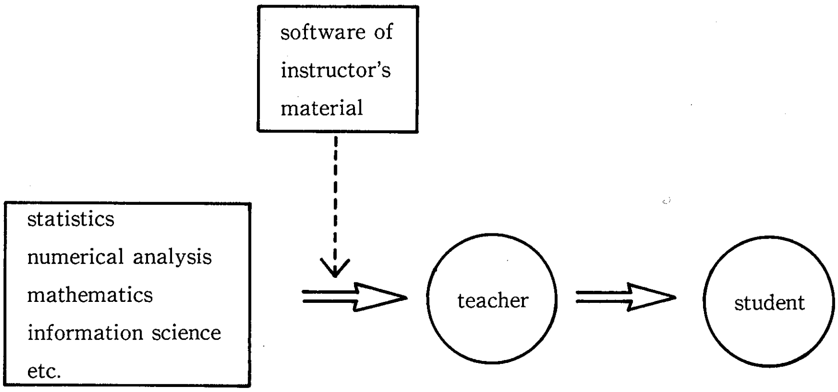


Fig.1 A plan of teaching.

A part of this paper was presented at the 49th Annual Meeting of Japan Statistical Society (July 1981, Kansai University, Osaka) and at the first International Conference On Teaching Statistics (ICOTS, August 1982, The University of Sheffield, The United Kingdom) .

* Faculty of Engineering.
** College of General Education.
*** Faculty of Science (Visiting Stanford University) .

show some illustrative examples in the statistics teaching. Similar attempt in Physics, Chemistry and Engineering is, for example, "Computer Simulation in University Teaching", North-Holland, 1981, Proceeding of the FEoLL Workshop, Paderborn, Germany, 1980.1., pp.28-30.

STATISTICS	MATHEMATICAL SCIENCE COMPUTER SCIENCE
Statistical Error { Measurement Error { Computational Error	Numerical Analysis { Error Theory Eigenvalue Singular Value Decomposition ....
Statistical Table	
Simulation	Programming Language and Library { FORTRAN, GPSS, REDUCE, Plot 10, APL, SPSS ....
Asymptotic Expansion	
Graphical Representation of Multi-dimentional Data	
Multivariate Analysis { Regression Analysis Discriminant Analysis Cluster Analysis	Pattern Recognition { Diagnosis Fingerprint, Phonetics Letter, Figure Remote Sensing
Statistical Map	
Prediction (Least Square Method)	Nonlinear Programming
Maximum Likelihood Method	
Statistical Data Base	Data Base

Table 1 . Relation of Statistics and Information Science.

2. METHOD OF DEVELOPMENT

Much use of computer as a tool for statistical education surely enhances the effectiveness of his or her teaching. At least, a teacher will obtain a picture and be able to show it, which is produced with a hard copy unit by the computer graphics, in the ordinary classroom for the course. To encourage to use the computer by teachers of statistics, they should be to bridge the gap between using the computer and achieving actually its purpose of notion or technique about the subjects. To achieve an adequate interface, three major points are important.

(i) COMPUTER SIMULATION: The concept of Probability and Statistics depends largely upon the experiment of trials, dice, card, chips, etc. But the development of trial is restricted physically. Collating the theoretical results with experimental values, we can be confirmed to explain it and lead to hold a new key for further study. Trial-and-error is also invaluable to learn. There may be cases which the experiment is not met with success. This question and discussion will by itself stimulate taking an interest in the content.

(ii) GRAPHICAL REPRESENTATION: The plot makes it easier to see how the things are going on. The explicit form of a probability distribution for various parameters, iterated plots for correlation and regression, the representation of multivariate data are excellent.

(iii) INTERACTIVE OPERATING: Spontaneity is the main advantage of TSS. A student who asks "What if " can be answered on the spot. Inspired idea is so good that one becomes the next term's preplanned one. This refines and clarifies the subject of the program. Enthusiastic the reaction is, the program must have flexibility to support these. So that the study expands its depth by the reaction and the theoretical considering.

Method : It is easy and cheap to use Micro or Personal computer nowadays. But rather than these, we have adapted ourselves to a large scale system of our university (see *Fig.*3). Because, it strengthen (1) Speed (2) Capacity of memory (3) Ability of the display (4) Program portability. The method aided by the computer is a slight different from an ordinally teaching. In addition to a kl talk or a discussion, the material retrieval is operated by a teacher using TSS. And the most important thin¥ is that he is sufficiently considerate of efforts to the response of the student and returning it back to them. If a student is eager for the subject to understand, the direct operating to the computer is permitted under TSS's configuration.

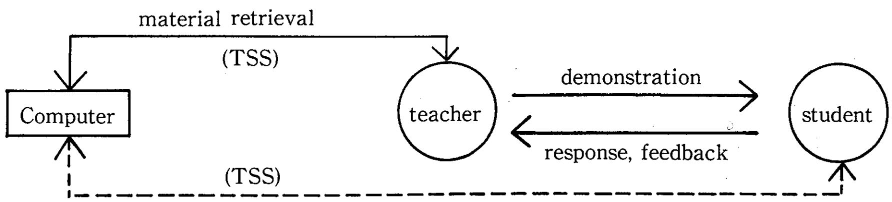


Fig.2 The method of Development.

Effects: The main effects are summarized and listed as follows.

(1) To understand the statistical notions, the simulation is important and the "live" experience attracts their attention. Comparing trial by hands, the result is complete enough. (e.g. generation of random number, sampling theory, sampling from probability distributions) (2) The visual aided instruction by the graphic display helps students to understand the subject quite effectively. (e.g., shape of probability distributions, correlation and regression) (3) Further usage of the computer is open to all of students. They have an intimate acquaintance with using the computer. This empirical operation is also useful in studying Computer Science. (4) For teachers or researchers, it may happen to meet with unexpected results. So they proceed to consider the cause and the detailed analysis. It may possibly find a new result by some chance.

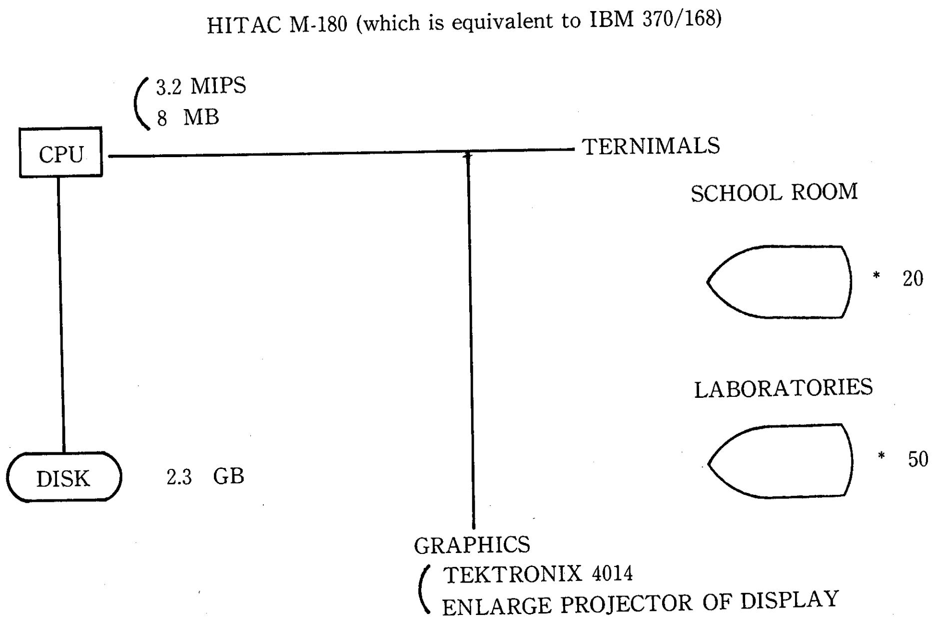


Fig.3 The computer system of our university.

3. ILLUSTRATIVE EXAMPLES

Some typical type of illustrative examples are summarized as follows.

(1) Confirm results of the theory by watching graphics.

→ Beta Distribution (Appendix 1), Central Limit Theorem (Appendix 4), Characteristics of Some Estimators (Appendix 5), Regression Analysis (Appendix 6).

(2) Find a case which is known by experience but is not sufficient.

→ Normal Approximation of Binomial Distribution (Appendix 2).

(3) Get useful and suggestive figures for textbook or in the classroom.

→ Central Limit Theorem, Regression Analysis.

(4) Should make detailed analysis of the interesting phenomena.

→ Computational Method for Variance (Appendix 7).

(5) Help to develop tools of statistical research.

→Air Polution Analysis of Chiba city in Japan (to appear).

4. DISCUSSION

Needless to say, this system needs entailing great expense of equipments so we can not induce in usual classrooms of a university. But its effect is surely astonishing to us. 1) If all the circumstances go well, we need High leveled intelligent graphic terminals. 2) And we should like to connect the audio-visual instruction system in order to demonstrate it for many students. 3) To enlarge the material so easy and speedy, developing a program software; easy-to-use, 3-dimensional, colour graphics are necessary to use. 4) Co-operating to use data base of instructional materials, especially an interesting data in the real world, surely makes it more attractive and students consider how to applicate their knowledge. 5) Also as a researcher's problem, developing a graphical method for representing multi-dimensional data is needed in these computer usage.

Appendix 0. Flow of CHESS system.

Fig.A0-1 is flow of CHESS starting command "chess". Sequentially we select Chapter, Section and Input-parameter or Data. For example, *Fig.A0-2* shows Welcome Message and contents of Section 5 in Chapter 7.

Appendix 1. Beta Distribution.

The density function for Beta distribution: $f(x)=\frac{1}{B(p,\ q)}x^{p-1}(1-x)^{q-1}$ $(p>0,\ q>0;\ 0\leqq x\leqq 1)$ shows various shape. In *Fig.A1-1,* we can see change of $f(x)$ according to the change of (1) p-value (q=constant), (2) q-value (p=constant) and (3) change of $f(x)$ under $p=q$.

Appendix 2. Normal Approximation of Binomial Distribution.

(1) Probability distribution of $B(n,\ p)$: $P\{X=x\}=\binom{n}{x}p^x(1-p)^{n-x}$ for $p=0.1$ (0.1) 0.9 in *Fig.A2-1* with fixed n. (2) Normal approximation in the case of $np>5$. The vertical lines in the following figures denote its binomial density and the curve is its fitting normal density. *Fig.A2-2-1, 2-2* is (i) $n=20$; $p=0.25$, 0.40 respectively and *Fig.A2-3* is (ii) $p=0.1$; $n=20$, 40, 100. As is known by experience that the conditon: $np>5$ is sufficient, but these figures say that it is not enoughly satisfied in some cases.

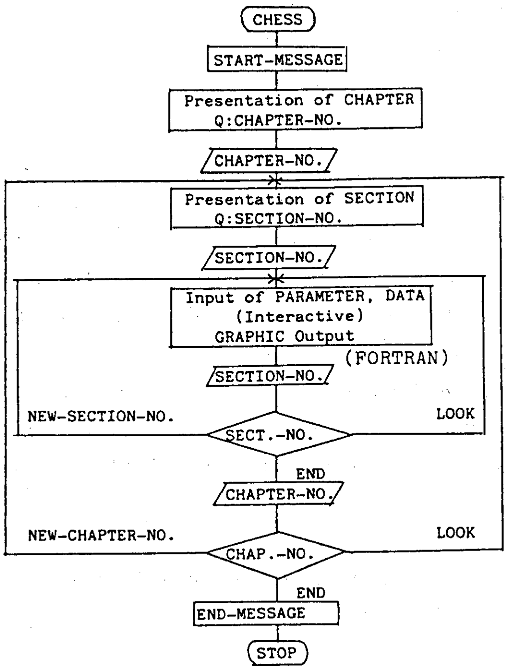


Fig.A0-1 Flow of CHESS system.

```
R>>CHESS

  *******              WELCOME TO                         82-04-27    *
  *     < CHIBA-UNIVERSITY EDUCATIONAL SYSTEM OF STATISTICS >   *******

  >-------------------------- CONTENTS ---------------------------<
        CHAPTER  0.  STATISTICAL PROBLEMS AND SOME PREPARATION
        CHAPTER  1.  SAMPLING THEORY
        CHAPTER  2.  DESCRIPTION OF SAMPLE
        CHAPTER  3.  THEORETICAL DISTRIBUTIONS
        CHAPTER  4.  SAMPLING DISTRIBUTIONS
        CHAPTER  5.  ESTIMATION
        CHAPTER  6.  TESTING HYPOTHESIS
        CHAPTER  7.  CORRELATION AND REGRESSION
        CHAPTER  8.  NONPARAMETRIC METHODS
        CHAPTER  9.  GRAPHICAL REPRESENTATION OF MULTIVARIATE DATA

Q : KEY-IN   CHAPTER-NO. =7

  >-------------------------- SECTIONS ---------------------------<
        CHAPTER  7.  CORRELATION AND REGRESSION
            SECTION 1. POPULATION CORRELATION COEFFICIENT
            SECTION 2. SAMPLE CORRELATION COEFFICIENT
            SECTION 3. REGRESSION MODELS
            SECTION 4. LEAST SQUARE METHODS
            SECTION 5. ESTIMATION OF REGRESSION COEFFICIENT
            SECTION 6. ANALYSIS OF VARIANCE
            SECTION 7. SELECTION OF REGRESSION MODELS

Q : KEY-IN   SECTION-NO. =5
```

Fig.A0-2 Contents of CHESS.

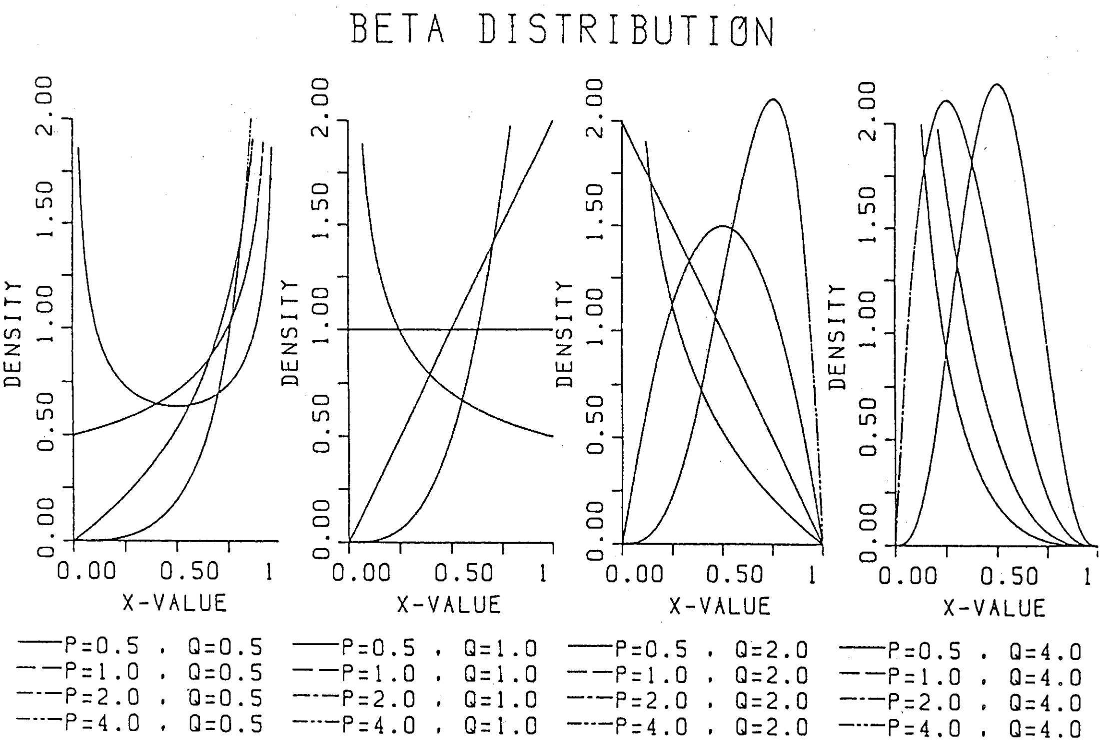


Fig.A1-1 Beta distributions with various parameters.

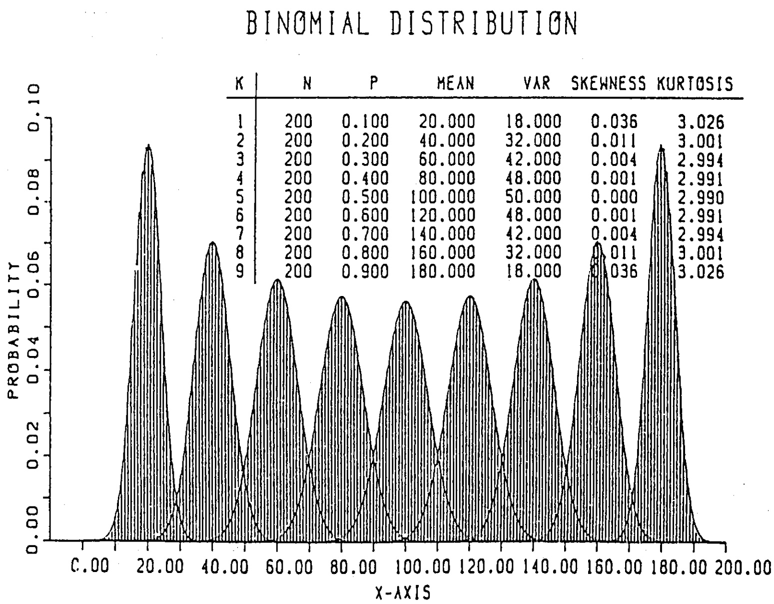


Fig.A2-1 Symmetric property in p of Binomial distribution.

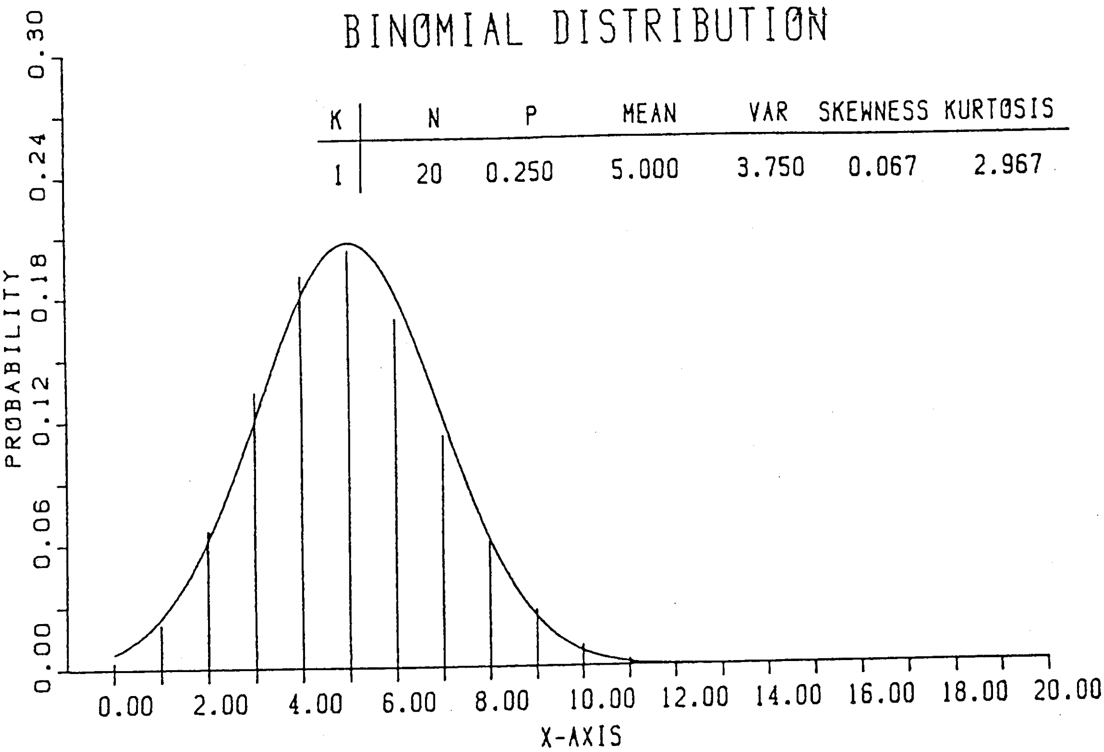


Fig.A2-2-1 Normal approximation of Binomial distribution ($np=5$).

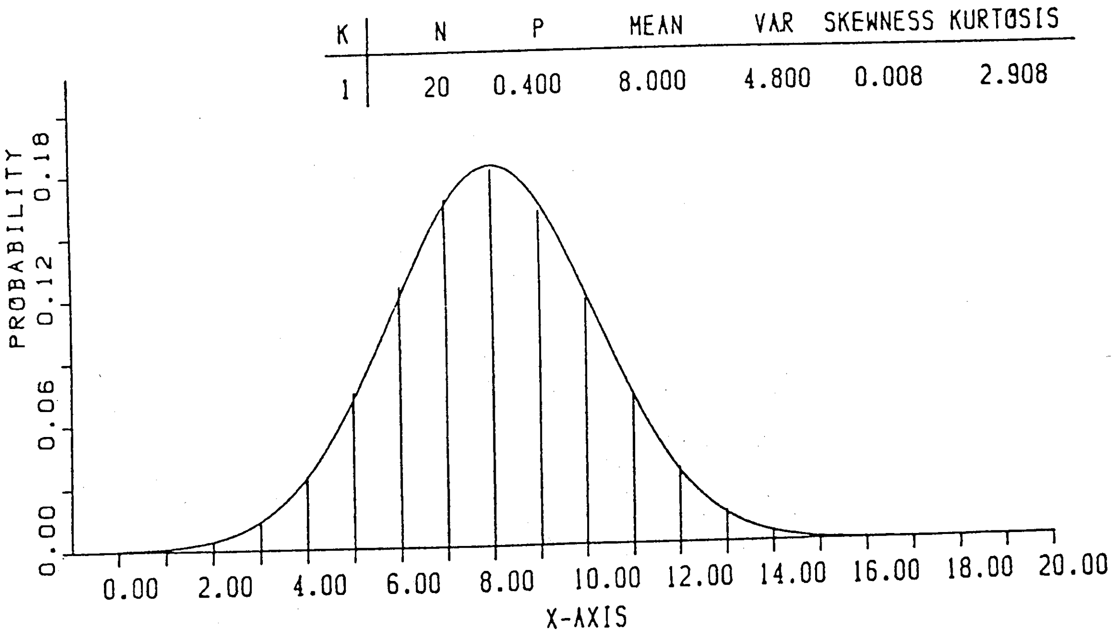


Fig.A2-2-2 Normal approximation of Binomial distribution ($np=8>5$).

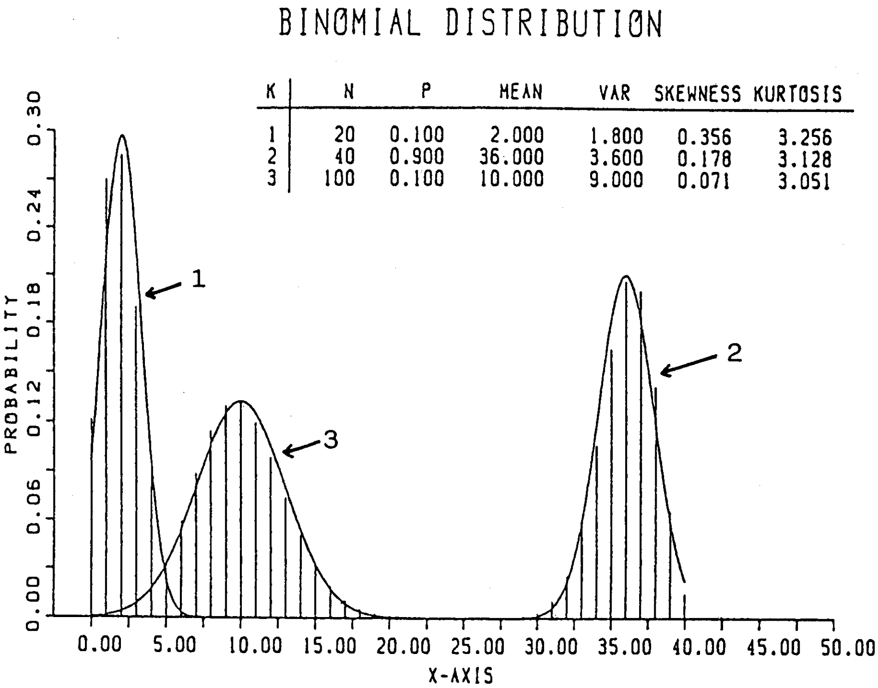


Fig.A2-3 Normal approximation of Binomial distribution.

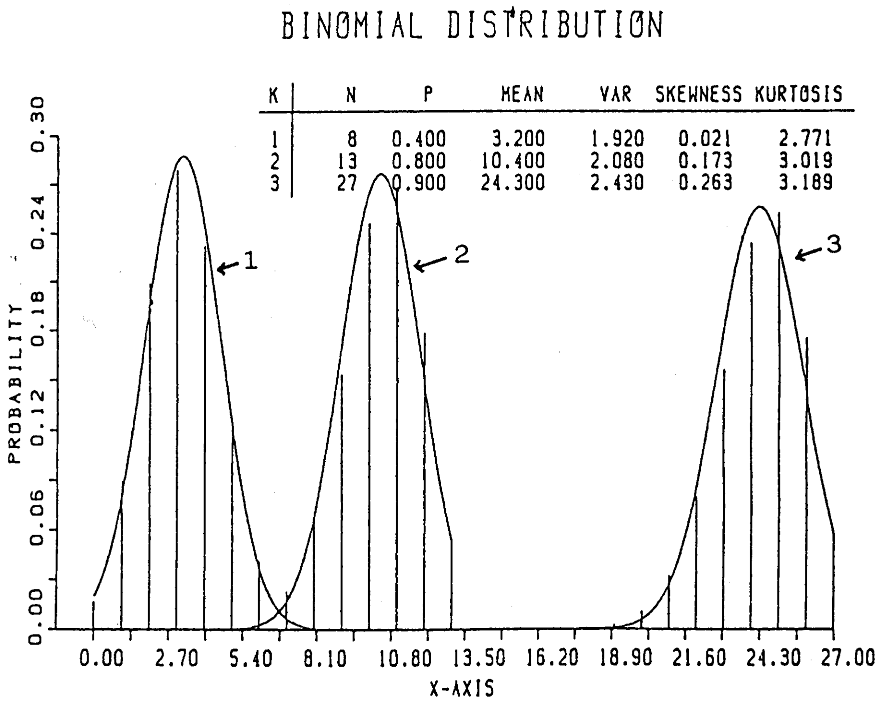


Fig.A2-4 Difference of N (np, npq) and B (n, P).

Appendix 3. Variety of Normal Distribution.

A slightly different distribution from Normal distribution is considered by changing as follows:

Creation of a p. d. f. $f(x)$ from $N(0, 1)$

$$f(x)=\phi(x)\{1+c_1P_1(x)+c_2P_2(x)+c_3P_3(x)+c_4P_4(x)\},$$

$\{P_i(x)\}$: Hermite Polynomial System

[Mean]	$E_f\{X\}=0$	$c_1=0$
[Variance]	$E_f\{X^2\}=1$	$c_2=0$
[Skewness]	$E_f\{X^3\}=\sqrt{\beta_1}$	$c_3=\sqrt{2\pi}\cdot\sqrt{\beta_1}$
[Kurtosis]	$E_f\{X^4\}=\beta_2$	$c_4=\sqrt{2\pi}(\beta_2-3)$

$$f(x)=\phi(x)\ \{1+\frac{1}{8}(\beta_2-3)-\frac{1}{2}\sqrt{\beta_1}x-\frac{1}{4}(\beta_2-3)x^2+\frac{1}{6}\sqrt{\beta_1}x^3+\frac{1}{24}(\beta_2-3)x^4\},$$

$$\equiv \phi(x) * c(x)$$

Condition on $\sqrt{\beta_1}$, β_2

$f(x)\geqq 0$ for $^\forall x \epsilon R^1$ $\quad \beta_2 \geqq 3;\ c(x^*)\geqq 0$ s.t. $\frac{d}{dx}c(x)\big|_{x=x^*}=0$

Fig.A3-1 is region where $f(x)$ becomes a p.d.f. and its graph is drawn in *Fig.A3-2-1, 2-2.* Generated random numbers from $f(x)$ is in *Fig.A3-3.*

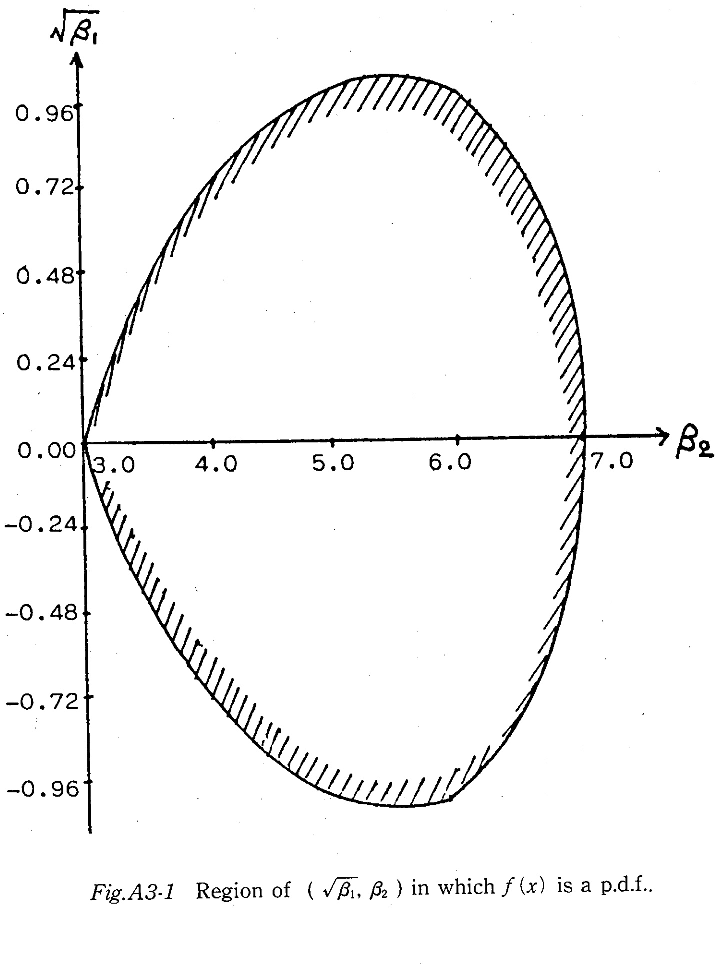


Fig.A3-1 Region of ($\sqrt{\beta_1}$, β_2) in which $f(x)$ is a p.d.f..

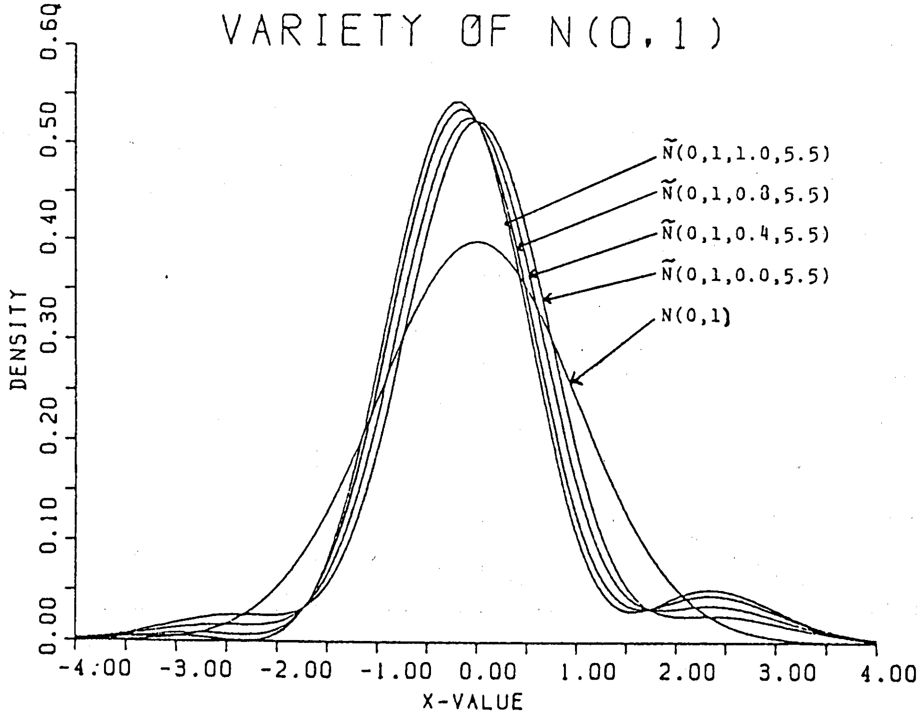


Fig.A3-2-1 Variety of N (0, 1) for several $\sqrt{\beta_1}$ (β_2=5. 5).

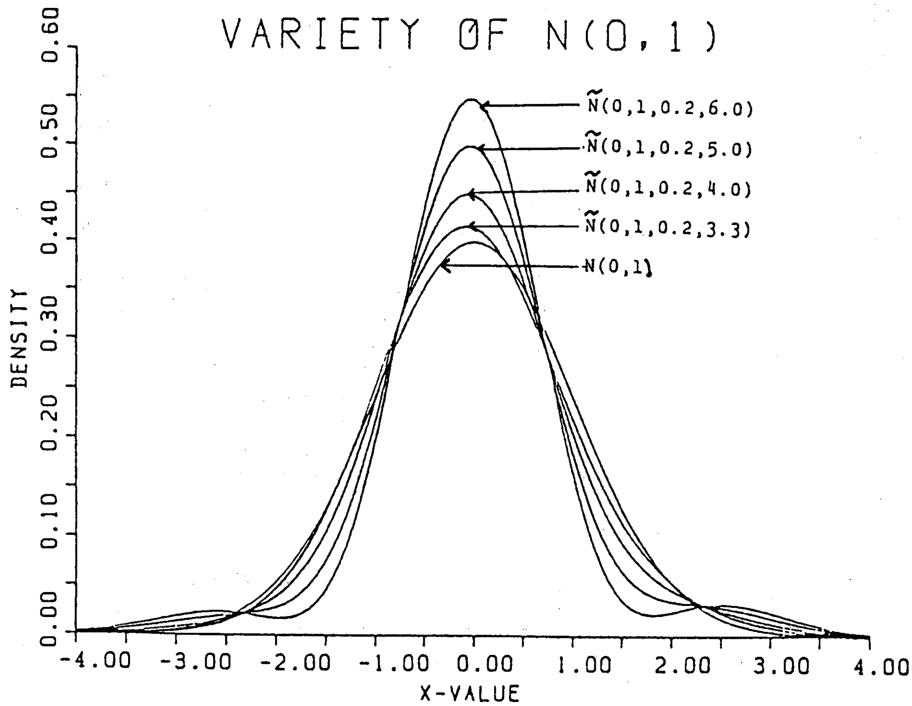
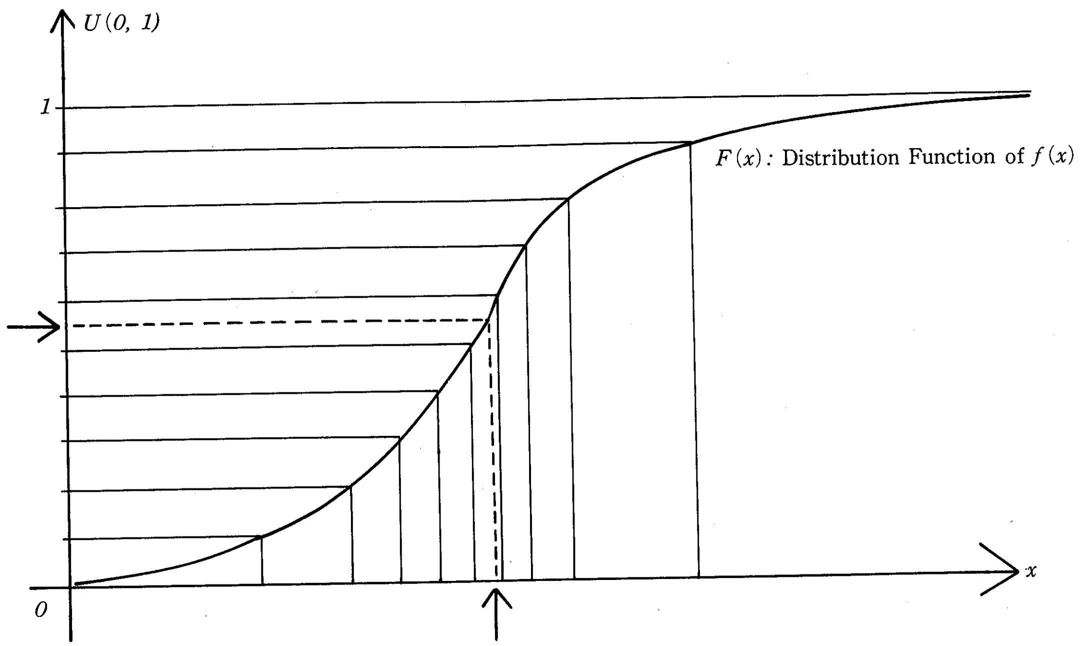


Fig.A3-2-2 Variety of N (0, 1) for several β_2 ($\sqrt{\beta_1}$=0.2).



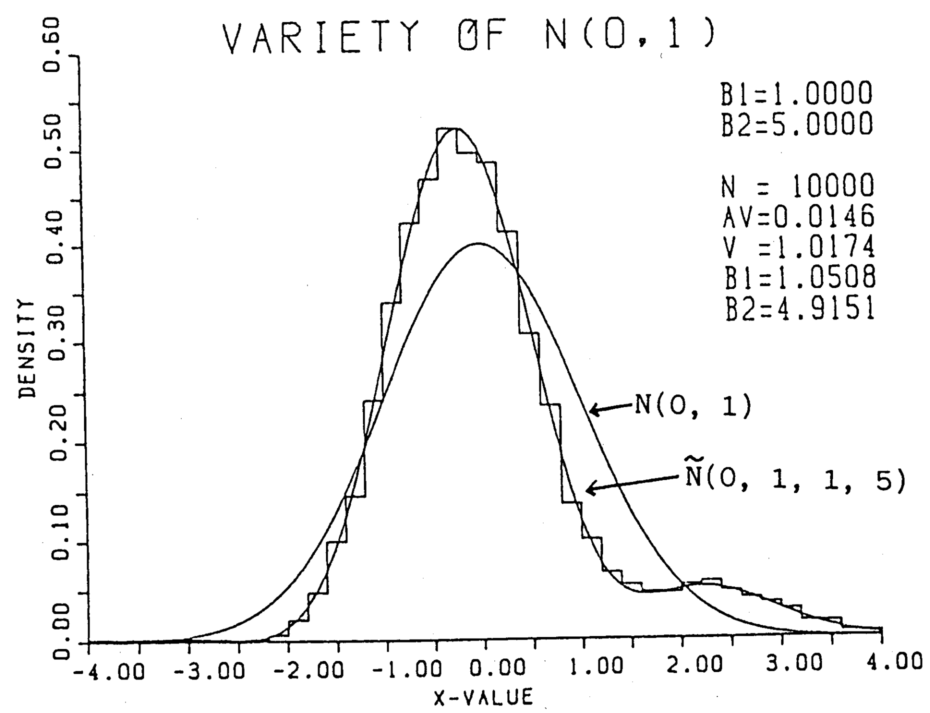


Fig.A3-3 Histogram of samples from $\tilde{N}$ (0, 1, 1, 5).

Appendix 4. Sampling Distribution; Central Limit Theorem.

$$X_1,\ X_2, \cdots\cdots,\ X_n \overset{i.i.d}{\sim} F(x:\mu,\ \sigma^2)$$

$$\rightarrow$$

$$\overline{X}=\frac{1}{n}(X_1+X_2+\ldots+X_n)\sim N(\mu,\ \frac{\sigma^2}{n})$$

Distributions

1°	$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	$(-\infty<x<\infty)$	: Normal Distribution
2°	$f(x)=\frac{1}{2}e^{-\vert x\vert}$	$(-\infty<x<\infty)$	: Doubly Exponential Distribution
3°	$f(x)=1$	$(-\frac{1}{2}<x<\frac{1}{2})$	: Uniform Distribution
4°	$f(x)=e^{-x}$	$(0<x<\infty)$	: Exponential Distribution
5°	$f(x)=\vert x\vert$	$(-1<x<1)$	: V-shape Distribution

Fig.A4-1 shows input parameters by the request and its results, histograms of random numbers generated by several types of distributions, are at *Fig.A4-2, A4-3* and *A4-4.*

```
   == CHAP-6 : SAMPLING DISTRIBUTION ==
?-0: # OF THEME
      THEMES    1 : POPULATION & SAMPLE
                2 : SAMPLING DISTRIBUTION
                3 : CENTRAL LIMIT THEOREM
?
2
       *** SAMPLING DISTRIBUTION ***
?-1: # OF DISTRIBUTION TYPE
     1. NORMAL    2. DOUBLY EXPONENTIAL    3. UNIFORM
     4. GAMMA     5. EXPONENTIAL           6. V-SHAPE

?
1
?-2: # OF CASES [K] & # OF EXPERIMENTS [M] (K < 5,M <= 10000)
?
2 5000
?-3: # OF SAMPLE [N(I):I=1,K] (N(I) < 6)
?
2 4
```

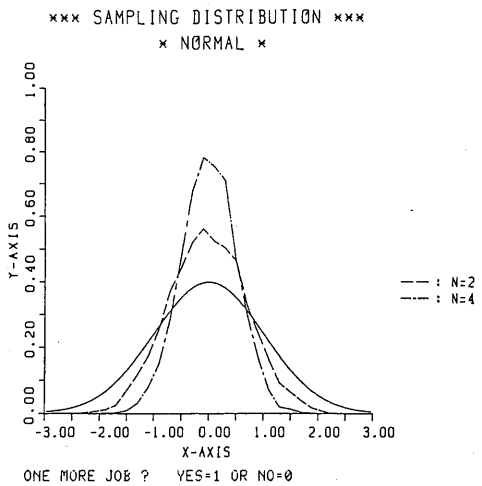


Fig.A4-1 Conversation and sampling distribution from N (0, 1).

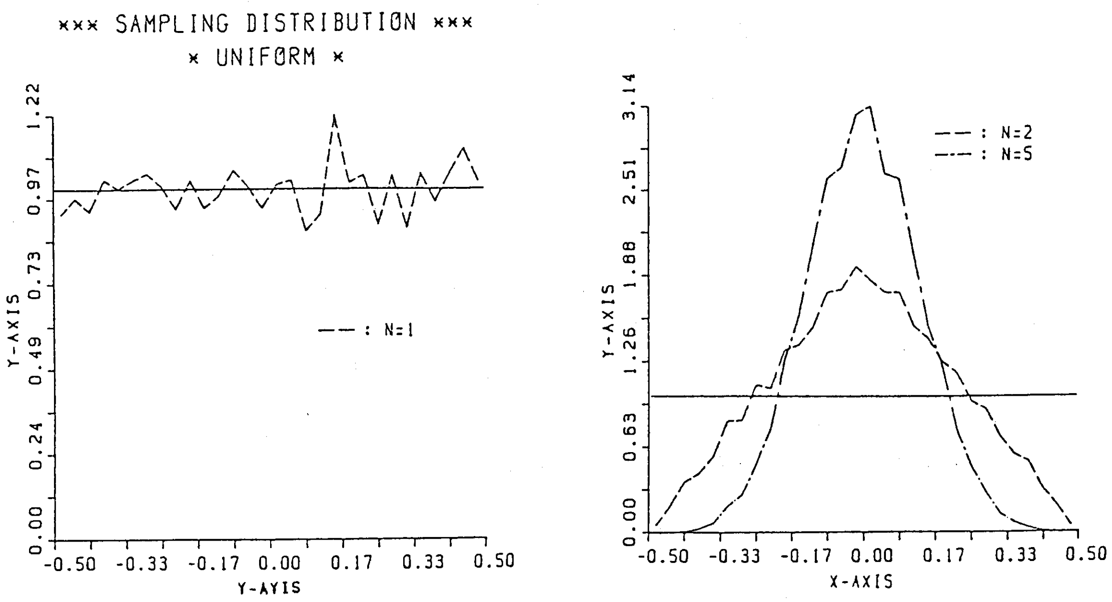


Fig. A4-2 Sampling distribution from Uniform : $U(-\frac{1}{2}, \frac{1}{2})$.

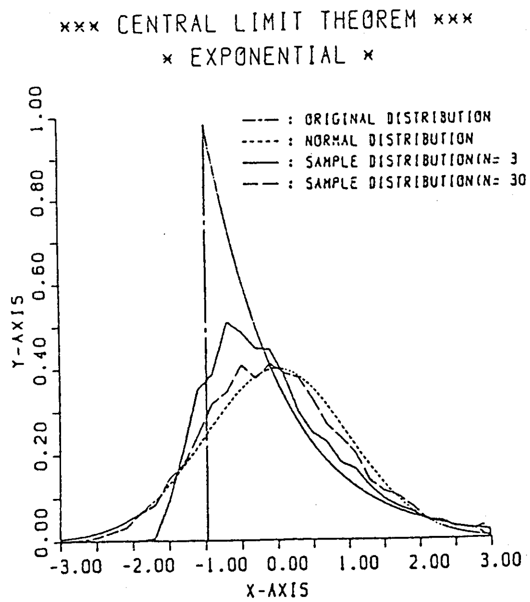


Fig.A4-3 Exponential case

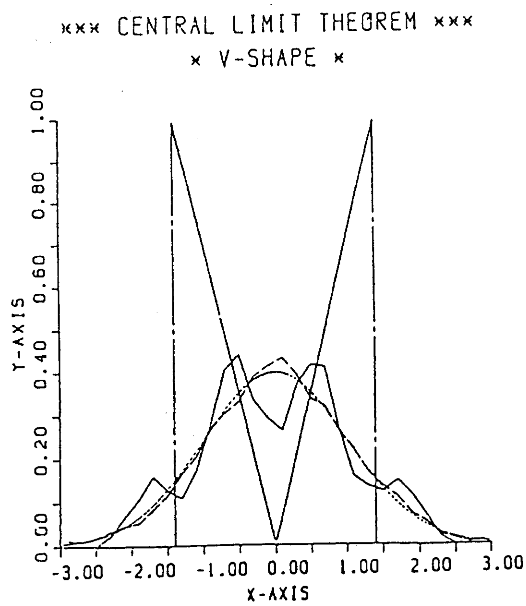


Fig.A4-4 V-shape case

Appendix 5. Characteristics of Some Estimators.

$(X_1, X_2 \cdots\cdots, X_n)$: Random Sample; n : Sample Size

Me : Median

Mo : Mode

$\overline{X} : = \frac{1}{n}\sum_{i=1}^{n} X_i$

$Trm : = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} X_{(i)}$

$Extr : = \frac{1}{2}(X_{(1)} + X_{(n)})$

$[X_{(i)}$: i-th Order Statistic$]$

$V_n : = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$

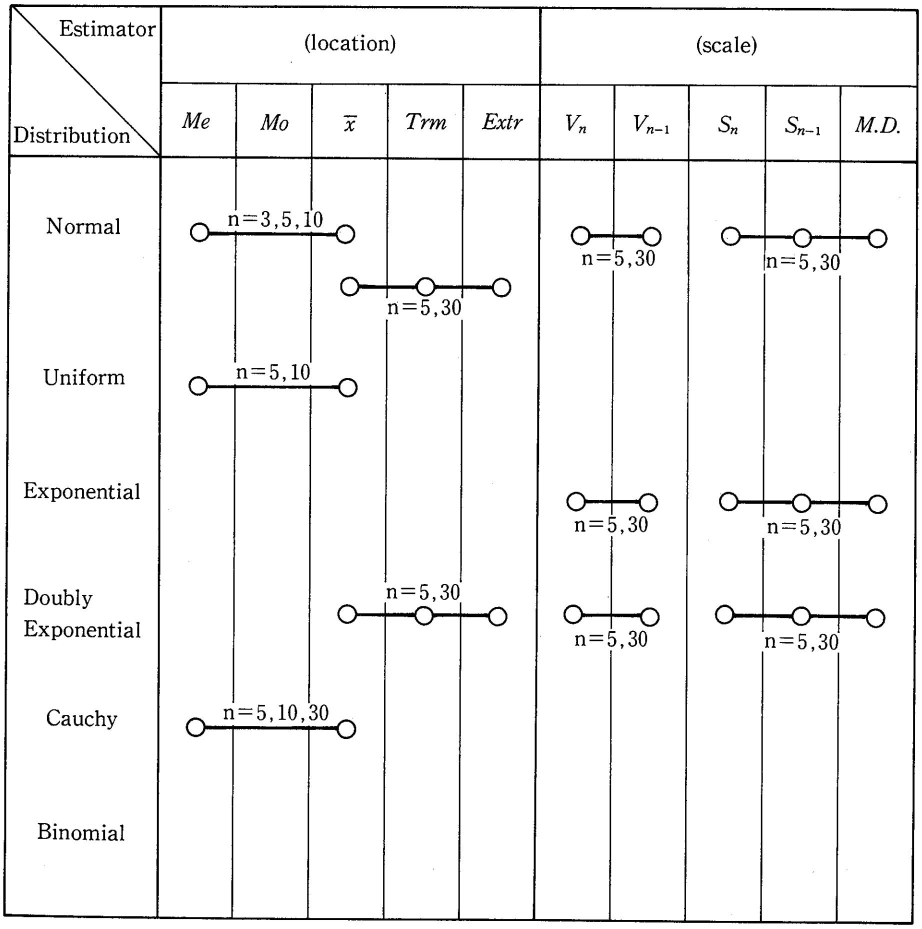
$V_{n-1} : = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$

$S_n : = V_n^{1/2}$

$S_{n-1} : = V_{n-1}^{1/2}$

$M.D. : = \frac{1}{n}\sum_{i=1}^{n} | X_i - \overline{X} |$

Fig.A5-2 is a comparison among the location estimators and *A5-3* is among the scale estimators.



```
********** MENU OF TYPE OF DISTRIBUTION **********

     < 1> ----- UNIFORMAL
     < 2> ----- NORMAL
     < 3> ----- BINOMIAL
     < 4> ----- EXPONENTIAL
     < 5> ----- CAUCHY
     < 6> ----- T
     < 7> ----- POISSON
     < 8> ----- BETA
     < 9> ----- LOGISTIC
     <10> ----- UNIFORM OF DISCRETE
     <11> ----- DOUBLE EXPONENTIAL

  WHICH TYPE OF DISTRIBUTION ?
?
2

INPUT "MU" AND "SIGMA**2"
?
0.0,1.0

HOW MANY STATISTICS DO YOU WANT? (MAX 3)
?
1

*********** MENU OF STATISTIC *****************

LOCATION ******
      < 1> ----- MEAN                          : MEAN
      < 2> ----- MEDIAN                        : ME
      < 3> ----- TRIMMED MEAN                  : TRIM
      < 4> ----- MODE                          : MO
      < 5> ----- EXTREMAL MEAN                 : EXTR
SCALE      ******
      < 6> ----- VARIANCE                      : V(N)
      < 7> ----- UNBIASED VARIANCE             : V(N-1)
      < 8> ----- SQUART ROOT OF <6>            : S(N)
      < 9> ----- SQUART ROOT OF <7>            : S(N-1)
      <10> ----- MEAN DEVIATION                : M.D.

NO. -- 1
  WHICH STATISTIC SELECT ?
?
1

*********** SIMULATION START **************

  INPUT SIZE OF SAMPLE (1<=SIZE<=1000)
?
5

  INPUT ITERATION OF SIMULATION (1<=ITERATION<=5000)
?
3000

  HOW MANY COPIES ?  (MAXIMUM 4)
?
0

  AUTO INTERVAL SET?  YES---<1> ; NO--- ANOTHER KEY
```

Fig.A5-1 Conversation of Chapter 5 (comparison of several estimators).

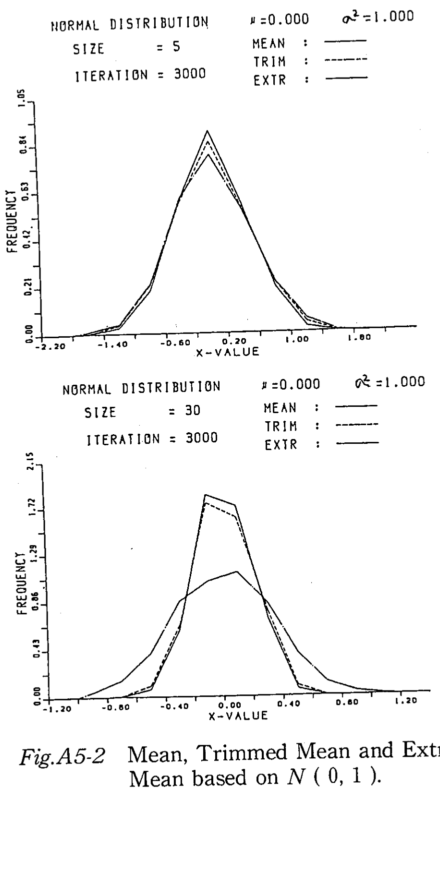


Fig.A5-2 Mean, Trimmed Mean and Extremal Mean based on N (0, 1).

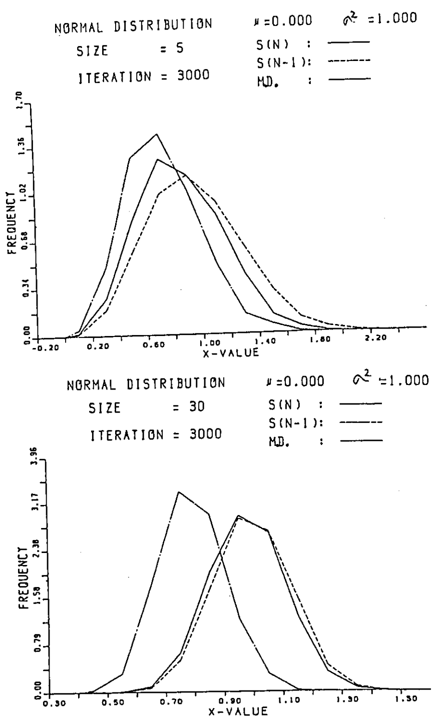


Fig.A5-3 S_n, S_{n-1} and Mean Deviation based on N (0, 1).

Appendix 6. Regression Analysis.

$$Y = X\beta + \varepsilon,\ \varepsilon \sim N(O_N, \sigma^2 I_N),\ \mathrm{rank}(X) = p$$

$$Y_{ij} = x_i^T\beta + \varepsilon_{ij} \quad (i=1,2\ \cdots\cdots,\ l\ :\ j=1,2\ \cdots\cdots,\ m)$$

$$X \equiv \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_l^T \end{bmatrix}$$

(1) $\overline{Y}_i \equiv \frac{1}{m}\sum_{j=1}^{m} Y_{ij} \sim N(x_i^T\beta, \frac{\sigma^2}{m})$

(2) $\hat{Y}_i \equiv x_i^T\hat{\beta} \sim N(x_i^T\beta, \frac{\sigma^2}{m}x_i^T(X^TX)^{-1}x_i)$; $\hat{\beta}$; Least Square Estimator of β

(3) $\hat{\sigma}^2 \equiv \frac{1}{n-p}(Y-\hat{Y})^T(Y-\hat{Y})$; $\hat{Y} = [\hat{Y}_1, \cdots, \hat{Y}_1, \cdots\cdots, \hat{Y}_l, \cdots \hat{Y}_l]^T$

$T_i' \equiv \hat{Y}_i / \{\frac{\hat{\sigma}^2}{m}x_i^T(X^TX)^{-1}x_i\}^{1/2}$

$T_i \equiv T_i' - x_i^T\beta / \{\frac{\hat{\sigma}^2}{m}x_i^T(X^TX)^{-1}x_i\}^{1/2} \sim t(n-p)$

Fig.A6-2-1 is plotted first data and its reregression line and *Fig.A6-2-2* is iterated 20 times' results. In *Fig. A6-3*, standardized values T_i of Y_i at different points 2 and 10 are shown. We will confirm surely that they reduce the same curve of t-distribution.

```
Q-1 : SELECT <1>-(SIMULATION) , <0>-(REAL-DATA)
? 1
Q-2 : # OF SAMPLE POINTS <L>
? 20
Q-2.1 : SAMPLE POINTS <X(I),I=1,L>
? 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Q-3 : SELECT ERROR-DIST. <1>-(NORMAL) , <0>-(NON-NORMAL)
? 1
Q-3.1 : INPUT < M , SIGMA2 , BETA0 , BETA1 , ISEND >
        M                 : # OF OBSERVATIONS AT THE SAME SAMPLE POINT
        SIGMA2            : VARIANCE OF RANDOM ERROR
        BETA0 , BETA1     : REGRESSION COEFFICIENTS
        ISEND             : # OF SIMULATIONS
? 2 32 4 2 100
```

Fig.A6-1 Conversation of Chapter 7 (estimation of regression coefficient).

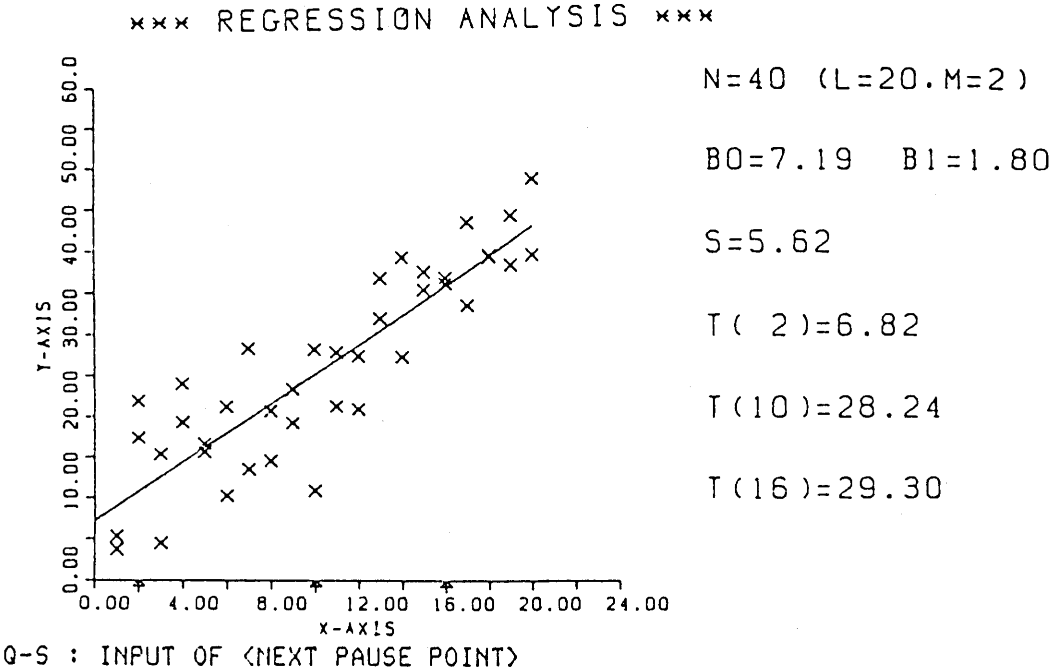


Fig.A6-2-1 The first data and its regression line.

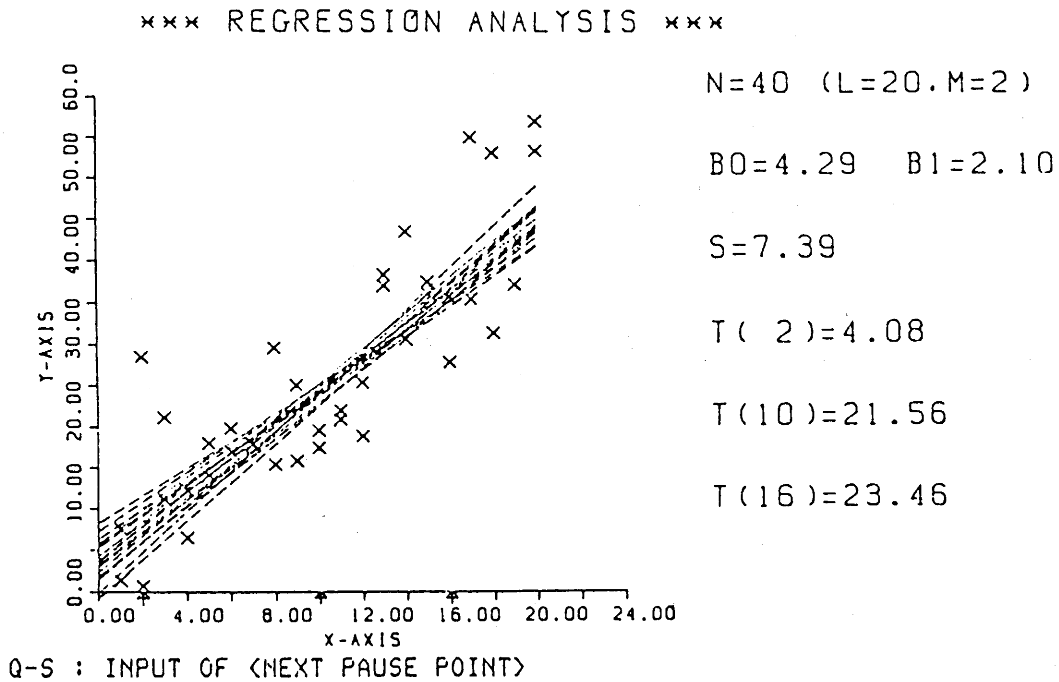


Fig.A6-2-2 The 20th data and iterated 20 regression lines.

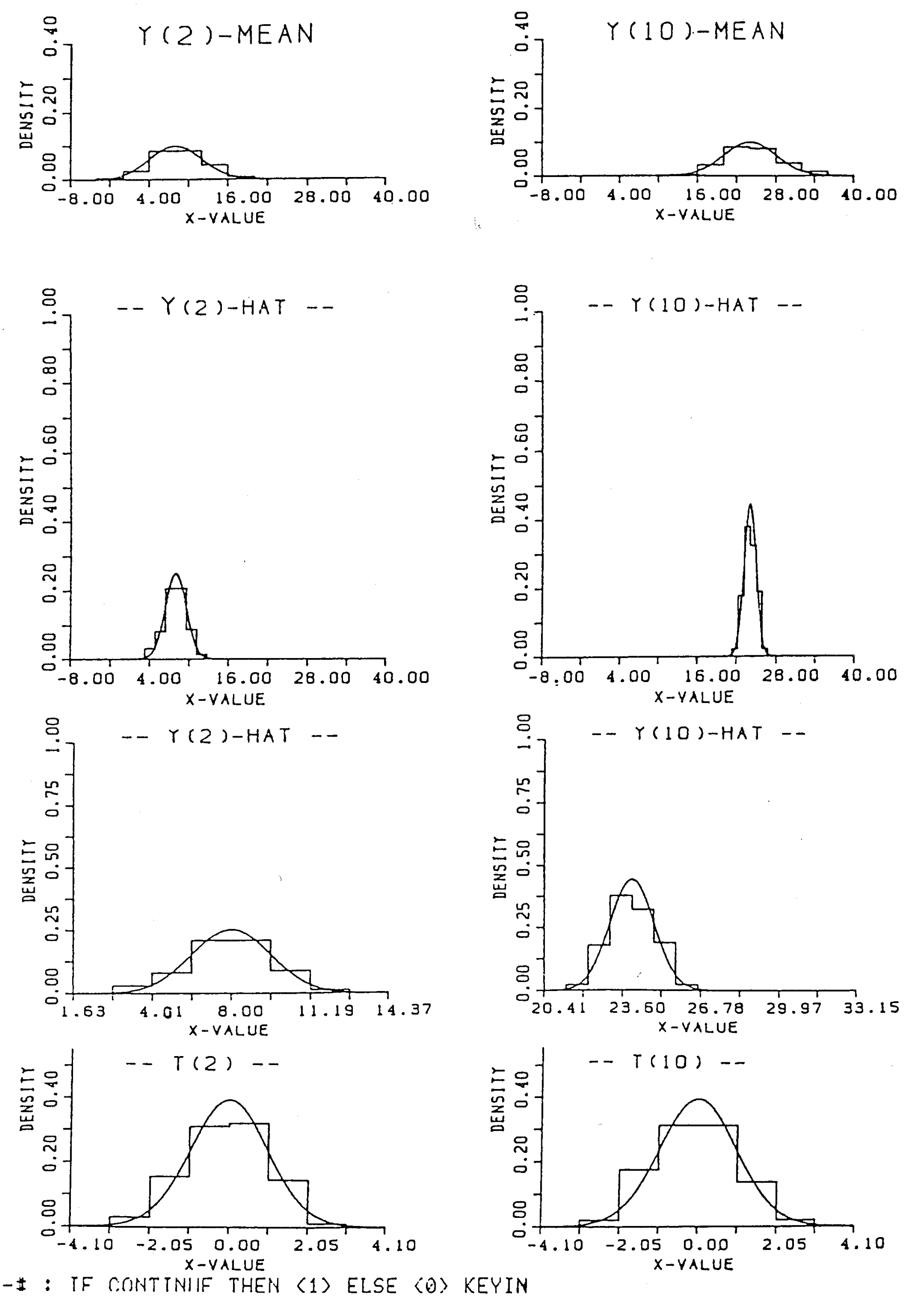


Fig.A6-3 Distributions of Y_i, $\hat{Y}_i$ and T_i (standardized value of $\hat{Y}_i$) at sample points i=2, 10.

Appendix 7. Computational Method for Variance.

Four kinds of algorithms for variance are compared. The aim is to inspire careful attention of the computation even if it is quite usual statistics such as variance. For sample variances of Normal distribution, the calculated value differs largely from its true one if it lacks the precision and is indifferent to the method. One of the authors analized the interest curve such as *Fig.A7-1, 2.* In the *Fig.A7-1,* it shows two algorithms (1) and (2) defined later with a single precision. In *Fig.A7-2,* the variance calculated by Textbook algorithm, and SINGLE-DOUBLE means Σx_i is in a single precision and Σx_i^2 in a double one. SINGLE or DOUBLE means in a same precision for both terms respectively. (1) Two-pass algorithm (Definition): $\bar{x}=\Sigma x_i/n$, $\sigma^2=\Sigma(x_i-\bar{x})^2/n$, (2) Textbook algorithm (Desk Calculator Method): $\bar{x}=\Sigma x_i/n$, $\sigma^2=\Sigma x_i^2/n-\bar{x}^2$, (3) Trial Mean algorithm: $d_i=x_i-x_1$, $\bar{x}=\Sigma d_i/n+x_1=\bar{d}+x_1$, $\sigma^2=\Sigma d_i^2/n-\bar{d}^2$, (4) West's algorithm: $\bar{x}_j=\bar{x}_{j-1}+(\bar{x}_j-\bar{x}_{j-1})/j$, $T_j=T_{j-1}+(1-1/j)(\bar{x}_j-\bar{x}_{j-1})^2$, $\bar{x}=\bar{x}_n$, $\sigma^2=T_n$.

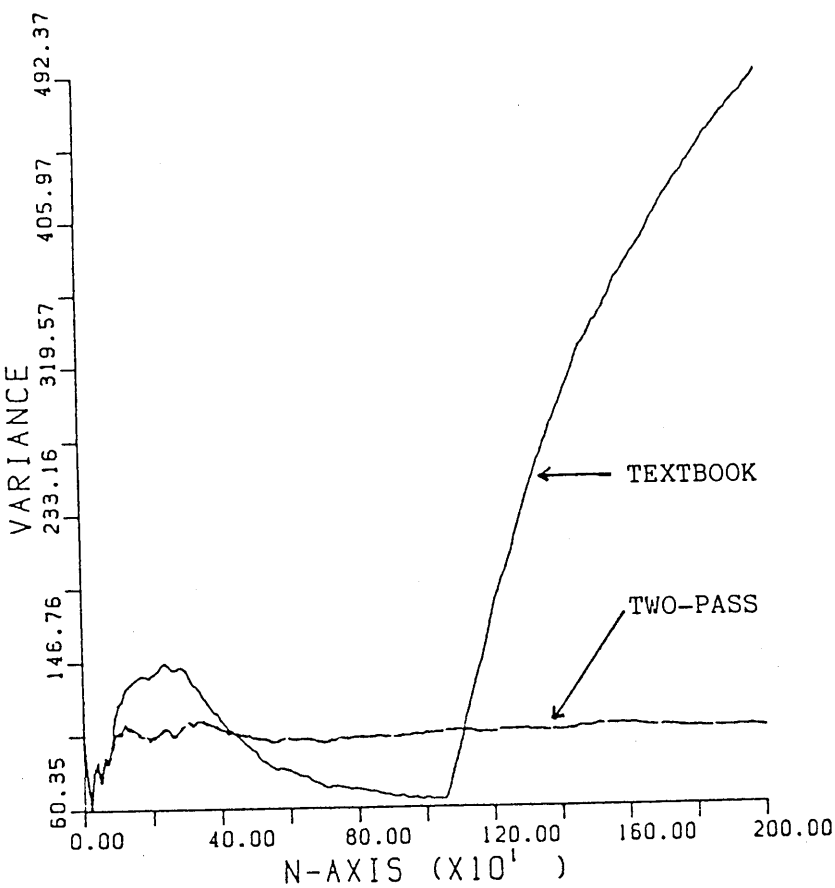


Fig.A7-1 Algorithms for variance (Textbook and Two-pass method in a single precision from N (1000, 100) with sample size=2000).

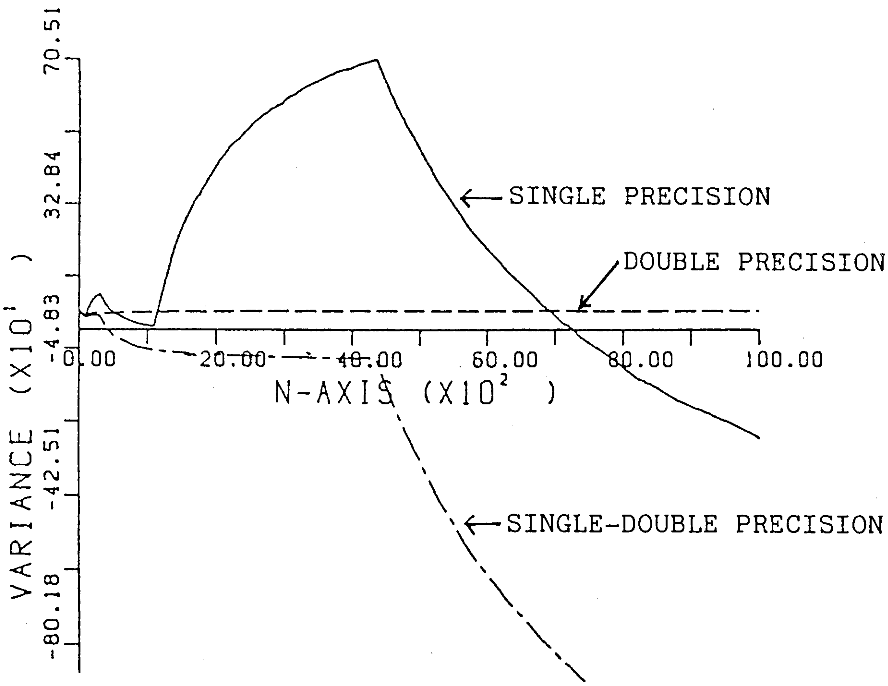


Fig.A7-2 Algorithms for variance
(Textbook method in several precisions from N (1000, 50) with sample size = 10000).