浮動小数点演算における相対丸め誤差の平均と分散について†

小沢 一文††

この論文では，仮数分布は逆数分布であるという仮定を用い，浮動小数点演算において生ずる相対丸め誤差の分布を求めている．また，適合度の検定を行うことによって，その分布の妥当性を確認している．その分布を用いて，相対丸め誤差の平均，分散を与える公式も導いている．この公式は，パラメータ $d(=b^{-t}, b$ は基底で，t は仮数の語長）のベキ級数展開の形をとっていて，その第一項は Tsao 等によって得られた公式そのものである．実験結果は，有効桁が少ないとき（d が大きいとき）は本論文の公式の方が秀れていることを示している．

1. はじめに

電子計算機を用いて数値計算を行う場合，計算機が表現できる数値は常に有限桁のものであるため，一般に丸め誤差の発生を防ぐことはできない．

浮動小数点演算においては，相対丸め誤差の上限は使用する計算機の語長に依存するが，取り扱うデータに無関係になるため，丸め誤差を絶対誤差の形よりも相対誤差の形で表現する方が，あらゆる誤差を統一的に取り扱えるという点で便利である．

したがって，これまでに浮動小数点演算の丸め誤差解析の研究においては，主に相対丸め誤差が取り扱われ，多くの成果もあげられている[1)~7)]．

しかし，相対丸め誤差の分布に関する研究には，多少の問題点が残されている．

これまでの浮動小数点演算の誤差解析に関する多くの研究では，相対丸め誤差の分布は一様分布であると仮定されてきた[3),4)]．しかし，この仮定に基いて導出された結果は，実験結果とはあまり良く一致していない[3)]．

これに対して，Tsao[5)]，Kaneko 等[6)]は，浮動小数点数の仮数分布が逆数分布[8)]であるという仮定より相対丸め誤差の分布を導出している．

しかし，これらの研究においては相対誤差の密度関数は Sterbenz[9)]の導出した誤差の上限を越えたところにおいても値をもっている．したがって，誤差の分布を説明するモデルとしては完全ではないと考えられる．また，そこで導出された平均，分散を与える公式は実験結果との比較検討がほとんどなされていないため，その公式は実用上差し支えないのかどうか不明である．

本論文では，文献 5)，6) と同じように仮数の分布が逆数分布であるような入力データに対して，相対丸め誤差の分布を導出しその適合度の検定を行う．さらに，その分布を用いて平均，分散を与える公式も導出し，実験結果との比較を行う．

2. 浮動小数点演算と相対丸め誤差

任意の実数 x を，整数 $b(b\geq 2)$ を底とする浮動小数点方式を用いて表わすと，

$$x=sgn(x)\cdot b^{e}\cdot\alpha \tag{1}$$

となる．ここで，$sgn(x)$ は符号を表わすものとする．すなわち，

$$sgn(x)=\begin{cases}1, & x>0\\ 0, & x=0\\ -1, & x<0\end{cases} \tag{2}$$

なる関数とする．また，指数 e は仮数 α が，

$$b^{-1}\leq\alpha<1 \tag{3}$$

を満足するような整数を選ぶものとする．このとき，α を b 進法で表現すると，

$$\alpha=m_1b^{-1}+m_2b^{-2}+m_3b^{-3}+\cdots+m_ib^{-i}+\cdots \tag{4}$$

となる．ここで，各 m_i は条件

$$1\leq m_1\leq b-1,\ 0\leq m_i\leq b-1,\ i=2,\ 3,\ \cdots \tag{5}$$

を満足する整数である．以下（4）の α を $\alpha=(0.m_1m_2\cdots m_i\cdots)_b$ と略記する．

† On the Mean and the Variance of Relative Round-Off Errors in Floating-Point Arithmetic by KAZUFUMI OZAWA (Sendai Radio Technical College).

†† 仙台電波工業高等専門学校

電子計算機を用いて実数 x を表現する場合，符号，指数，それに仮数をコード化して表現するわけであるが，計算機の語長は有限であるため，これらすべてを十分正確に表現できるとは限らない．

ここでは，符号と指数については常に正確に表現できるものとし，さらに仮数部に割り当てられた桁が b 進 $t(t\geq 2)$ 桁であるとし，それ以下の桁は表現できないものとする．

この仮定より，α を表現する際に α の $t+1$ 桁以下に適当な処理を施さなければならない．ここでは，$t+1$ 桁以下を無条件に切り捨てる切り捨て方式 (chopping) と，$t+1$ 桁以下で表わされる値が $b^{-t}/2$ より大きいとき $t+1$ 桁を切り上げ他は切り捨てる四捨五入方式 (rounding) を適用した場合について考察を行う．

仮数 α の b 進 t 桁近似を $\bar{\alpha}$ とすると，切り捨て方式を適用したときは，

$$\bar{\alpha}=(0, m_1m_2\cdots m_t)_b \tag{6}$$

となり，四捨五入方式を適用したときは，

$$\bar{\alpha}=\begin{cases}(0.m_1m_2\cdots m_t)_b, & 0\leq\beta<\dfrac{d}{2}\\ (0.m_1m_2\cdots(m_t+1))_b, & \dfrac{d}{2}\leq\beta<d\end{cases} \tag{7}$$

となる．ここで，$d=b^{-t}$ とし，$\beta=(0.0\cdots0m_{t+1}m_{t+2}\cdots)_b$ とする．

これより，

$$|\bar{\alpha}-\alpha|\leq\begin{cases}d, & \text{切り捨て方式,}\\ d/2, & \text{四捨五入方式,}\end{cases} \tag{8}$$

を得る．

つぎに，α の代わりに $\bar{\alpha}$ を用いた x の有限桁近似を $\bar{x}$ とすれば，仮定より指数と符号は正確に記憶されているから，相対誤差 ε は $x\neq 0$ とすると，

$$\varepsilon=(\bar{x}-x)/x=(\bar{\alpha}-\alpha)/\alpha \tag{9}$$

となる．したがって，式(3) より α の最小値が b^{-1} であり，しかも式 (8) が成り立つから

$$|\varepsilon|\leq\begin{cases}bd, & \text{切り捨て方式,}\\ bd/2, & \text{四捨五入方式,}\end{cases} \tag{10}$$

を得る[7]．これは，浮動小数点演算の誤差解析においてしばしば用いられる誤差の上限である．

しかし，α が最小値 b^{-1} であるとき，α は b 進 t 桁の仮数部によって十分正確に表現できるから，すなわち $\bar{\alpha}=\alpha$ であるから，この場合は $\varepsilon=0$ となる．したがって，式 (10) に示した上限は実際に到達不可能なものである．

これに対し，Sterbenz[9]は切り捨て方式を適用した場合に生ずる相対丸め誤差のより厳密な上限を求めた．すなわち，

$$|\varepsilon|<\frac{bd}{1+bd} \tag{11}$$

である．また，Sterbenz と同様の手法によって，四捨五入方式についてもより厳密な上限を求めると

$$|\varepsilon|\leq\frac{bd}{2(1+bd/2)} \tag{12}$$

を得る．これ等の上限は式 (10) の上限より多少小さい．また，式 (10) の上限との差は b が大きいかまたは d が大きい（t が小さい）ときは，無視できないものになる．

これまでは，ε の分布として式 (10) で与えられる範囲での一様分布が用いられてきた．また，Tsao[5]，Kaneko 等[6]も式 (11)，(12) の上限とは矛盾した ε の密度関数を導出している．

次に，これまでの議論をもとに相対丸め誤差の分布を求める．

3. 相対丸め誤差の分布

ここで，実数 x は確率変数であるとする．このとき，仮数 α も確率変数となり分布をもつが，仮数分布については以下に示すような重要な報告がある．

まず，独立な二つの数の間に四則演算を施すと，その計算結果の仮数はもとの数の仮数より逆数分布に近づき，またどちらか一方の因子の仮数が逆数分布であるならば，演算結果の仮数も逆数分布になるというものである[8]．また，もう一つは実数 x の密度関数が，例えば正規分布のように解析関数であるならば，x の仮数の分布は逆数分布に非常に近いものになるというものである[10]．

したがって，本論文ではこれ等の報告より x の仮数 α の分布は逆数分布であると仮定する．すなわち，α の確率密度関数を $P_\alpha(\alpha)$ としたとき，

$$P_\alpha(\alpha)=\frac{1}{\alpha\ln b},\quad b^{-1}\leq\alpha<1 \tag{13}$$

であるとする．

つぎに，この仮定の下で α の上位 t 桁を γ，下位桁を β としたとき，すなわち，

$$\gamma=(0.m_1m_2\cdots m_t)_b,\quad \beta=(0.00\cdots0m_{t+1}m_{t+2}\cdots)_b$$

としたとき，各々の分布について考察する．

γ は条件 (5) より全部で $(b-1)\cdot b^{(t-1)}$ 通りの値

をとるから，それらの各々に小さい順に番号を付ける．

$$\left.\begin{aligned}&\gamma_i=b^{-1}+(i-1)d,\ i=1,\ 2,\ \cdots,\ I\\&I=(b-1)b^{(t-1)}\end{aligned}\right\} \tag{14}$$

ここで，$\gamma=\gamma_i$ である確率を q_i として表わせば，

$$q_i \triangleq P_r\{\gamma=\gamma_i\}=\int_{\gamma_i}^{\gamma_{i+1}} P_\alpha(\alpha)d\alpha=(ln\gamma_{i+1}-ln\gamma_i)/lnb \tag{15}$$

となる．

つぎに，β の分布を考察する．Feldstein[11]は，仮数 α が逆数分布しているならば，t が大きくなるにつれ β の分布は限りなく一様分布に近づき，またその収束のオーダが b^{-t} であることを証明している．したがって，ここでは β の確率密度関数 $P_\beta(\beta)$ を

$$P_\beta(\beta)=d^{-1},\ 0\leq\beta<d \tag{16}$$

とする．

さらに，上位桁 γ は β に対して統計的に独立であるとする*．

これより，ε の密度関数 $P(\varepsilon)$ は

$$P(\varepsilon)=\sum_{i=1}^{I} q_i\left\{\int_0^d P_{\varepsilon|\gamma,\beta}(\varepsilon|\gamma_i,\beta)P_\beta(\beta)d\beta\right\} \tag{17}$$

として表わせるから，式 (15), (16) を用いると

$$P(\varepsilon)=\sum_{i=1}^{I} q_i d^{-1}\int_0^d P_{\varepsilon|\gamma,\beta}(\varepsilon|\gamma_i,\beta)d\beta \tag{18}$$

を得る．ここで，$P_{\varepsilon|\gamma,\beta}(\varepsilon|\gamma_i,\beta)$ は ε の条件付き密度関数であり，切り捨て方式を用いるかあるいは四捨五入方式を用いるかによって異なる．以下，切り捨て方式と四捨五入方式について (18) より $P(\varepsilon)$ を導出する．

3.1 切り捨て方式

切り捨て方式を適用したとき，式 (6) より $\bar{\alpha}=\gamma$ であるから，

$$\varepsilon=(\bar{\alpha}-\alpha)/\alpha=-\beta/(\gamma+\beta) \tag{19}$$

を得る．したがって，

$$P_{\varepsilon|\gamma,\beta}(\varepsilon|\gamma_i,\beta)=\delta\left(\varepsilon+\frac{\beta}{\gamma_i+\beta}\right) \tag{20}$$

となる．ここで，δ はディラックのデルタ関数とする．式 (20) を式 (18) に代入する．ここで，$P(\varepsilon)$ を四捨五入方式のそれと区別するため $P^c(\varepsilon)$ で表わすと，

$$P^c(\varepsilon)=\sum_{i=1}^{I} q_i d^{-1}\int_0^d \delta\left(\varepsilon+\frac{\beta}{\gamma_i+\beta}\right)d\beta \tag{21}$$

となる．ここで，$y=\varepsilon+\beta/(\gamma_i+\beta)$ なる変数変換を行うと，式 (21) は

$$P^c(\varepsilon)=\sum_{i=1}^{I} q_i d^{-1}\int_\varepsilon^{\varepsilon+d/\gamma_{i+1}} \delta(y)\gamma_i\frac{1}{(y-\varepsilon-1)^2}dy \tag{21}'$$

と変形される．この積分は，下限 ε が負で上限が正であるとき（この逆は式 (19) よりありえない）値をもつ．すなわち，

$$P^c(\varepsilon)=\sum_{i=1}^{I} P_i{}^c(\varepsilon) \tag{22}$$

と表わせば，

$$P_i{}^c(\varepsilon)=\begin{cases}\dfrac{\gamma_i(ln\gamma_{i+1}-ln\gamma_i)}{dlnb(\varepsilon+1)^2}, & -\dfrac{d}{\gamma_{i+1}}<\varepsilon<0\\ 0 & \text{その他}\end{cases} \tag{23}$$

$i=1,\ 2,\ \cdots I.$

を得る．

ここで得た $P^c(\varepsilon)$ は，ε が式 (11) の Sterbenz の上限より大きいところでは $P^c(\varepsilon)=0$ となる．

これに対して，Tsao[5], Kaneko[6]の導出した密度関数は，この上限の外においても値をもっている．**図 1** に分布の一例を示す．

3.2 四捨五入方式

四捨五入方式を適用したときの密度関数 $P^r(\varepsilon)$ は，

$$P^r(\varepsilon)=\sum_{i=1}^{I}(P_i{}^+(\varepsilon)+P_i{}^-(\varepsilon)) \tag{24}$$

である．ここで，

$$P_i{}^+(\varepsilon)=\begin{cases}\dfrac{\gamma_{i+1}(ln\gamma_{i+1}-ln\gamma_i)}{dlnb(\varepsilon+1)^2}, & 0<\varepsilon<\dfrac{d}{2\gamma_i+d}\\ 0 & \text{その他}\end{cases} \tag{25}$$

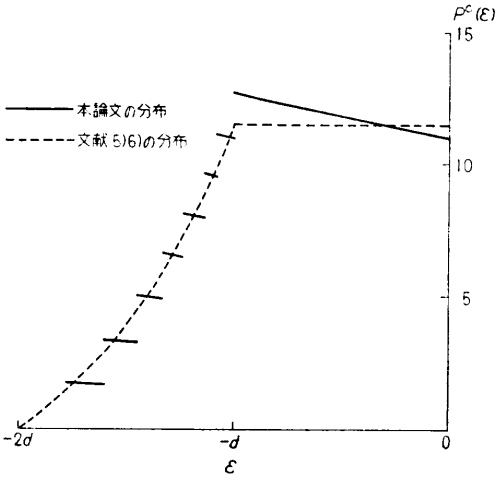

図 1 相対誤差分布の一例 ($b=2,\ t=4$)

Fig. 1. An example of relative round-off error distributions ($b=2,\ t=4$).

* これは近似的に成立する仮定である（付録参照）．なお，文献 5), 6) では α と β が互いに独立であるとしているが，証明は与えていない．

$$P_i^{-}(\varepsilon)=\begin{cases}\dfrac{\gamma_i(ln\gamma_{i+1}-ln\gamma_i)}{dlnb(\varepsilon+1)^2}, & -\dfrac{d}{2\gamma_i+d}<\varepsilon<0\\ 0 & \text{その他}\end{cases}$$

$$i=1,\ 2,\ \cdots,\ I \qquad (26)$$

である.

ここで得られた $P^r(\varepsilon)$ は，切り捨て方式の $P^c(\varepsilon)$ と同様に式(12)の上限より小さいところで値をもち，それ以外では0になる．また，その形状は原点に関して非対称である.

つぎに，ここで得られた密度関数が実際に実験結果にどの程度適合しているかを確認する．ここでは，$P^c(\varepsilon)$ の検定だけを行う.

まず，ε の分布は $P^c(\varepsilon)$ で表わされるという仮説を立て，$b=2$ とし t は3から 16 まで変化させ χ^2-検定を行う.

実験は，2進 27 桁の仮数を有する計算機において仮数の $t+1$ 桁から先を切り捨てるプログラムを作成し，これに真値 x を入力することによって行われる．真値 x は，仮数が逆数分布に非常に近い分布をもつと考えられる[10]正規乱数*を用いる．検定の手順は以下に示すとおりである.

i) 区間 $(-2d, 0]$ を 16 等分し，その各々を I_i $(i=1, 2, \cdots, 16)$ とする.

$$I_i=(-2d+d(i-1)/8,\ -2d+di/8]$$

ii) 実数 x の系列を前述のプログラムに入力し，その有限桁近似 $\bar{x}$ を求め，$\bar{x}$ と x より相対誤差 ε の標本を求める．ここで，標本の大きさ N は 10,000 とする.

iii) 得られた ε が上記のどの区間に入るかを調べ，その各区間 $I_i(i=1, 2, \cdots, 16)$ における度数 f_i を求める.

iv) つぎに，ε の分布が $P^c(\varepsilon)$ で与えられるとしたとき，ε が各区間に入ると期待される度数

$F_i=N\int_{I_i}P^c(\varepsilon)d\varepsilon$ を求め，χ^2 を次式より求める.

$$\chi^2=\sum_{F_i\neq 0}(f_i-F_i)^2/f_i$$

v) ε の分布として，文献 5)，6) で得られたものを仮定し，同じ手順で χ^2 の値を求める.

図2に t に対する χ^2 の値の変化と，$\chi^2_{0.95}$，$\chi^2_{0.99}$ を示し**，**表1，2**には $t=4$，16 における f_i と F_i

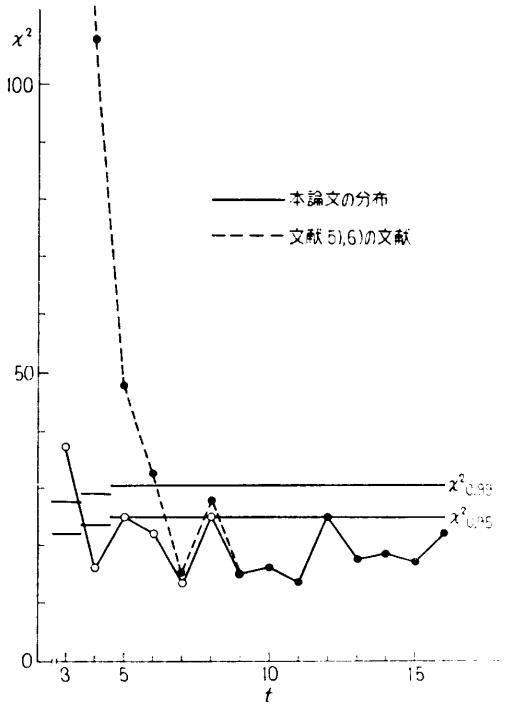

図 2 仮数の桁数 t と分布の適合度 χ^2 の関係

Fig. 2. The relation between the number of digits t and the goodness of fit χ^2 of the distributions.

を示す.

図2より，本報告のモデルでは $t=3$ の場合を除いてすべて $\chi^2<\chi^2_{0.95}$ となり，仮説が 5% の有意水準をもって受容されることが判明した．これに対し，文

表 1 各区間における度数と期待度数 $(b=2,\ t=4)$

Table 1 The frequency f_i and the expected frequencies F_i in each interval $(b=2,\ t=4)$.

区間	度数 f_i	本報告のモデルによる期待度数 F_i	文献 5)，6) のモデルによる期待度数 F_i
I_1	0	0	29
I_2	31	31	94
I_3	124	133	167
I_4	225	235	253
I_5	316	339	353
I_6	441	465	473
I_7	615	625	618
I_8	802	807	797
I_9	913	974	902
I_{10}	969	958	902
I_{11}	938	942	902
I_{12}	930	927	902
I_{13}	991	912	902
I_{14}	919	898	902
I_{15}	895	884	902
I_{16}	891	870	902
Total	10,000	10,000	10,000

* この正規乱数は，χ^2一検定を行った結果「仮数分布は逆数分布である」という仮説が受容されたものである(付録参照).

** 本報告のモデルでは，$t=3$ のとき $F_1=F_2=0$，$t=4$ のとき $F_1=0$ となり自由度が減るため $\chi^2_{0.95}$，$\chi^2_{0.99}$ の値が変化する．文献 5)，6) のモデルでは変化しない.

表 2 各区間における度数と期待度数 ($b=2$, $t=16$)
Table 2 The frequency f_i and the expected frequencies F_i in each in terval ($b=2$, $t=16$).

区間	度数 f_i	本報告のモデルによる期待度数 F_i	文献 5), 6) のモデルによる期待度数 F_i
I_1	25	29	29
I_2	81	94	94
I_3	164	167	167
I_4	222	253	253
I_5	330	353	353
I_6	476	473	473
I_7	620	618	618
I_8	834	797	797
I_9	926	902	902
I_{10}	871	902	902
I_{11}	909	902	902
I_{12}	928	902	902
I_{13}	894	902	902
I_{14}	853	902	902
I_{15}	884	902	902
I_{16}	983	902	902
Total	10,000	10,000	10,000

献 5), 6) のモデルでは t が 6 以下では $\chi^2>\chi_{0.99}^2$ となり有意水準が 1% でも仮説が棄却されている.

したがって, $t=3$ の場合を除き t が比較的小さいときは本報告のモデルの方が実験結果をより忠実に表現しているといえる.

また, 表 1, 2, 図 2 からわかるように t が大きくなるにしたがって, これら 2 つのモデルは同じものになることが予想され, どちらも検定に十分合格するモデルであるといえる.

4. 相対丸め誤差の平均, 分散

上で求めた ε の密度関数を用いて, ε の平均, 分散を与える公式を導く. ここでは, d に関する展開式の形で求める. ただし, 実用上十分な精度を得るためには, d が小さいときは平均は d, 分散は d^2 のオーダまでの展開で十分であるが, 比較的大きい場合も考慮して平均は d^2, 分散は d^3 のオーダまで展開し以下は打ち切る.

4.1 切り捨て方式

平均を $\bar{\varepsilon}_c$ で表わせば, 式 (22), (23) より

$$\bar{\varepsilon}_c=\sum_{i=1}^{I}\int_{-d/\gamma_{i+1}}^{0}\varepsilon P_i^c(\varepsilon)d\varepsilon$$
$$=\sum_{i=1}^{I}\frac{\gamma_i(ln\gamma_{i+1}-ln\gamma_i)}{dlnb}\int_{-d/\gamma_{i+1}}^{0}\frac{\varepsilon}{(\varepsilon+1)^2}d\varepsilon$$
$$=-1+\frac{1}{dlnb}\sum_{i=1}^{I}\gamma_i\left\{ln\left(1+\frac{d}{\gamma_i}\right)\right\}^2 \qquad (27)$$

を得る.

ここで, $b^{-1}\le\gamma_i$ でしかも $3\le t$ であるから, すべての i について $0<d/\gamma_i<1$ が成り立つ. したがって,

$$ln\left(1+\frac{d}{\gamma_i}\right)=\frac{d}{\gamma_i}-\frac{1}{2}\left(\frac{d}{\gamma_i}\right)^2+\frac{1}{3}\left(\frac{d}{\gamma_i}\right)^3-\cdots \qquad (28)$$

なる展開が可能である.

この展開を用いると,

$$\bar{\varepsilon}_c=-1+\frac{d}{lnb}\sum_{i=1}^{I}\left(\frac{1}{\gamma_i}-\frac{d}{\gamma_i^2}+\frac{11}{12}\frac{d^2}{\gamma_i^3}-\frac{5}{6}\frac{d^3}{\gamma_i^4}+\cdots\right) \qquad (29)$$

を得る. ここで, 上式の和を簡単化するため Euler-Maclaurin の公式[12]を用いる. それをここに示す.

$$d\sum_{i=1}^{I}f(\gamma_i)=F(1)-F(b^{-1})-d(f(1)-f(b^{-1}))/2+d^2(f'(1)-f'(b^{-1}))/12+O(d^4) \qquad (30)$$

ここで, $F(t)=\int f(t)dt$ とし, $f(t)$ は区間 $[b^{-1},1)$ で 4 回微分可能とする.

式 (29) の右辺第二項に式 (30) を適用すると,

$$\bar{\varepsilon}_c=-\frac{(b-1)}{2lnb}d+\frac{(b^2-1)}{24lnb}d^2+O(d^3) \qquad (31)$$

となる.

Tsao[5], Kaneko[6]の導出した結果は,

$$\bar{\varepsilon}=-\frac{(b-1)}{2lnb}d \qquad (32)$$

である*.

4.1.2 分 散

分散を求めるまえに 2 乗平均を求める. 2 乗平均を $\overline{\varepsilon_c^2}$ で表わせば,

$$\overline{\varepsilon_c^2}=\sum_{i=1}^{I}\frac{\gamma_i(ln\gamma_{i+1}-ln\gamma_i)}{dlnb}\int_{-d/\gamma_{i+1}}^{0}\frac{\varepsilon^2}{(\varepsilon+1)^2}d\varepsilon$$
$$=-1-2\bar{\varepsilon}_c+\frac{1}{lnb}\sum_{i=1}^{I}\frac{1}{1+(d/\gamma_i)}ln(1+(d/\gamma_i)) \qquad (33)$$

となる. ここで, $0<(d/\gamma_i)<1$ であるから前と同様に $1/(1+(d/\gamma_i))$ と $ln(1+(d/\gamma_i))$ に級数展開を施すと, 式 (33) の右辺第三項は,

$$\frac{1}{lnb}\sum_{i=1}^{I}\frac{1}{1+(d/\gamma_i)}ln(1+(d/\gamma_i))$$
$$=\frac{1}{lnb}\sum_{i=1}^{I}\left\{\frac{1}{\gamma_i}-\frac{3}{2}\left(\frac{1}{\gamma_i}\right)^2d+\frac{11}{6}\left(\frac{1}{\gamma_i}\right)^3d^2-\cdots\right\} \qquad (34)$$

---

* 文献 5), 6) では, 誤差=真値−近似値として定義しているので, これを本論文の定義にあわせて書き直した.

となる．つぎに，上式に Euler-Maclaurin の公式(30)を適用すると，$\overline{\varepsilon_c^2}$ は

$$\overline{\varepsilon_c^2}=\frac{b^2-1}{6lnb}d^2-\frac{b^3-1}{18lnb}d^3+O(d^4) \tag{35}$$

となる．

これより，分散 σ_c^2 は

$$\sigma_c^2=\overline{\varepsilon_c^2}-(\varepsilon_c)^2=\frac{b-1}{12(lnb)^2}\{2(b+1)lnb-3(b-1)\}d^2$$
$$+\frac{b-1}{72(lnb)^2}\{3(b^2-1)-4(b^2+b+1)lnb\}d^3$$
$$+O(d^4) \tag{36}$$

となる．

文献 5)，6) では

$$\sigma_c^2=\frac{b-1}{12(lnb)^2}\{2(b+1)lnb-3(b-1)\}d^2 \tag{37}$$

である．

4.2 四捨五入方式

ここでは，結果のみに留める．

4.2.1 平　均

平均を $\overline{\varepsilon_r}$ で表わせば，

$$\overline{\varepsilon_r}=-\frac{b^2-1}{48lnb}d^2+O(d^4) \tag{38}$$

である．また，文献 5)，6) では

$$\overline{\varepsilon_r}=0 \tag{39}$$

である．

4.2.2 分　散

分散を σ_r^2 で表わせば

$$\sigma_r^2=\frac{b^2-1}{24lnb}d^2+O(d^4) \tag{40}$$

である．これは，$O(d^4)$ の項を除けば文献 5)，6) と同じ結果である．

5. 実験結果との比較

つぎに，ここで得られた公式がどの程度実験結果に適合しているかを確認するため，実験結果と式(31)，(36)，(38)，(40)によって得られる理論値の比較を行う．

この実験は，前に誤差の分布を検証するのに用いた正規乱数列を用いこれを真値として採用し，その t 桁近似から相対丸め誤差の系列を求めて行う．この系列を $\varepsilon_n(n=1,2,\cdots,N)$ とし，標本平均と標本分散を

$$\bar{\varepsilon}=\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n,\quad \sigma^2=\frac{1}{N}\sum_{n=1}^{N}(\varepsilon_n-\bar{\varepsilon})^2 \tag{41}$$

より計算する．これらは，標本数 N を 10,000 という大きな数としたので，真の平均，分散を十分よく近似しているものと考える．また，桁数 t は前に分布の適合度の検定で仮説が受容された $4\leq t\leq 16$ の範囲とする．

表 3 に実験値と理論値の比較を示す．この結果より，t が大きくなるにつれ理論値と実験結果とがよく一致する傾向にあるが，t が小さいときでも文献 5)，6) の結果よりは本論文の結果の方が実験結果に近いことが明らかになった．これは，平均では d^2，分散では d^3 の項の影響であると考えられる．

本論文の公式は，t が小さいときにも有効であるため，減算によって有効桁が極端に減少するいわゆる桁落ち[13]が起きたときの誤差解析にも役立つものと思われる．

6. おわりに

浮動小数点演算における相対丸め誤差の分布を求め，それをもとに平均，分散を与える公式を求めた．

表 3 相対丸め誤差の平均と分散

Table 3 The mean and the variance of relative round-off error.

t	平均			分散		
	平均 $\bar{\varepsilon}_c$	文献 5)，6) の理論値	実験値 $\bar{\varepsilon}$	分散 σ_c^2	文献 5)，6) の理論値	実験値 σ^2
4	−0.444E−1	−0.451E−1	−0.436E−1	0.714E−3	0.786E−3	0.709E−3
5	−0.224E−1	−0.226E−1	−0.222E−1	0.187E−3	0.196E−3	0.187E−3
6	−0.112E−1	−0.113E−1	−0.112E−1	0.479E−4	0.491E−4	0.481E−4
7	−0.562E−2	−0.564E−2	−0.559E−2	0.121E−4	0.123E−4	0.120E−4
8	−0.282E−2	−0.282E−2	−0.282E−2	0.305E−5	0.307E−5	0.305E−5
9	−0.141E−2	−0.141E−2	−0.140E−2	0.765E−6	0.767E−6	0.764E−6
10	−0.704E−3	−0.702E−3	−0.705E−3	0.191E−6	0.191E−6	0.192E−6
11	−0.352E−3	−0.352E−3	−0.353E−3	0.479E−7	0.479E−7	0.482E−7
12	−0.176E−3	−0.176E−3	−0.176E−3	0.120E−7	0.120E−7	0.120E−7
13	−0.881E−4	−0.881E−4	−0.877E−4	0.300E−8	0.300E−8	0.298E−8
14	−0.440E−4	−0.440E−4	−0.440E−4	0.745E−9	0.745E−9	0.742E−9
15	−0.220E−4	−0.220E−4	−0.220E−4	0.187E−9	0.187E−9	0.187E−9
16	−0.110E−4	−0.110E−4	−0.110E−4	0.468E−10	0.468E−10	0.465E−10

さらに，分布の適合性の検定，公式によって得られた推定値と実験結果との比較も行った．

その結果，仮数の桁数 t が小さいときは同様の仮定から導出された類似のモデル[5),6)]より，実験結果に適合していることが判明した．また，t が大きいときには本報告によって与えられる誤差の平均，分散の推定値と文献 5)，6) のそれとは等しくなり，どちらも真の値の十分良い推定値になっていることが判明した．

この公式を導出するにあたり，仮数の分布は逆数分布であると仮定したが，Hamming[9)]が証明しているように四則演算を繰り返し行った場合，その演算結果の仮数は逆数分布に十分近い分布をもつから，この公式は浮動小数点演算の誤差解析において十分有効である．

なお，ここで用いた計算機は東北大学大型計算センター ACOS 77 システム 700 である．

最後に，本論文の不備な点を根気強くご指摘下さった査読者に謝意を表わします．

参 考 文 献

1) Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations, John Wiley & Sons, Inc. New York (1962).
2) Wilkinson, W. H.: Algebraic Eigenvalue Problem, Oxford University Press, Oxford (1965).
3) Liu, B. and Kaneko, T.: Error Analysis of Digital Filters Realyzed with Floating-Point Arithmetic, Proc. IEEE, Vol. 57, No. 10, pp. 1735-1747 (1969).
4) Kan, E. P. and Aggarwal, J. K.: Error Analysis of Digital Filter Employing Floating Point Arithmetic, IEEE, Vol. CT-18, No. 6, pp. 678-686 (1971).
5) Tsao, N. K.: On the Distribution of Significant Digits and Round-off Errors, Commun. ACM. Vol. 17, No. 5, pp. 269-271 (1974).
6) Kaneko, T. and Liu, B.: On Local Roundoff Error in Floating-Point Arithemetic, J. ACM, Vol. 20, No. 3, pp. 391-398 (1973).
7) Wilkinson, J. H.: Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N. J. (1963).
8) Hamming, R. W.: On the Distribution of Numbers, Bell Syst. Tech. J., Vol. 49, No. 8, 1609-1625 (1970).
9) Sterbenz, P. H.: Floating-Point Computation, pp. 74, Prentice-Hall Inc., Englewood Cliffs, N. J. (1974).
10) 小沢一文：浮動小数点数の仮数分布とその逆数分布からの偏差について，電子通信学会技術研究報告. CST-78-81, pp. 51-58 (1978).
11) Feldstein, A.: Convergence Estimates for the Distribution of Trailing Digits, J. ACM. Vol. 23, No. 2, pp. 287-296 (1976).
12) 森口繁一ほか：数学公式集Ⅱ，pp. 34，岩波書店，東京 (1976).
13) 一松 信：電子計算機と数値計算，pp. 42，朝倉書店，東京 (1973).

表 4 仮数分布の適合度の検定
Table 4. The test of goodness of fit of mantissa distribution.

区 間	度 数 f_i	期待度数 F_i
I_1	1,382	1,375
I_2	1,234	1,255
I_3	1,183	1,155
I_4	1,115	1,069
I_5	991	995
I_6	899	931
I_7	819	875
I_8	844	825
I_9	813	780
I_{10}	720	740
Total	10,000	10,000

$\chi^2=\sum_{i=1}^{10}(f_i-F_i)^2/F_i=10.12$

$\left(F_i=\int_{I_i}\frac{d\alpha}{\alpha\, ln\, 2}\times 10,000\right)$

付録 1 α の上位桁 γ と下位桁 β の独立性について．

任意の $\eta(0<\eta<d)$ について，$\gamma=\gamma_i$，$0\leq\beta<\eta$ なる確率を求めると，

$$P_r\{\gamma=\gamma_i, 0\leq\beta<\eta\}=P_r\{\gamma_i\leq\alpha<\gamma_i+\eta\}$$
$$=ln(1+\eta/\gamma_i)/lnb$$

を得る．上式において $(\eta/\gamma_i)<1$ が成り立っているから，

$$P_r\{\gamma=\gamma_i, 0\leq\beta<\eta\}=\left(\frac{\eta}{\gamma_i}-\frac{1}{2}\left(\frac{\eta}{\gamma_i}\right)^2+\cdots\right)\Big/lnb$$

なる展開が可能である．

一方，β の分布は一様分布としたから

$$P_r\{\gamma=\gamma_i\}P_r\{0\leq\beta<\eta\}=\eta ln(1+d/\gamma_i)/(dlnb)$$
$$=\frac{1}{lnb}\left(\frac{\eta}{\gamma_i}-\frac{1}{2}\left(\frac{d\eta}{\gamma_i{}^2}\right)+\cdots\right),\quad i=1,\ 2,\ \cdots,\ I$$

となる．したがって，

$$P_r\{\gamma=\gamma_i\}P_r\{0\leq\beta<\eta\}=P_r\{\gamma=\gamma_i, 0\leq\beta<\eta\}$$
$$+O(d^2)$$

が成立する．これより，γ と β はほぼ独立となる．特に d が小さくなるにつれてこの傾向は強まる．

付録 2 正規乱数の検定

ここでは，仮説「正規乱数列xの仮数の分布は逆数分布である」を立て，これを検定する．

正規乱数として区間 (0, 1) の一様乱数を12回加えて作り出されたものを用いる．すなわち，区間 (0, 1) の一様乱数の系列 $u_{n,k}(n=1, 2, \cdots, N, k=1, 2, \cdots, 12)$ より

$$x_n=\left(\sum_{k=1}^{12} u_{n,k}\right)-6,\quad n=1,\ 2,\ \cdots,\ N$$

を求め，この x_n を平均0，分散1の正規乱数列として採用する．ここで標本の大きさ N は 10,000 とする．

検定の手順は前に $P^c(\varepsilon)$ の適合度の検定を行ったときと同じである．ただし，ここでは区間 [1/2, 1) を10 等分する．各区間 I_i における度数 f_i と期待度数 F_i を**表4**に示す．

これより，χ^2 の値を計算すると $\chi^2=10.12$ となり，この仮説は有意水準を 10% としても十分に検定に合格する．

なお，ここで用いた一様乱数は ACOS-FORTRAN の基本外部関数 RAND である．

(昭和52年 9 月 26 日受付)
(昭和54年10月 25 日採録)