THE ACCURACY OF COMPUTATION WITH APPROXIMATE NUMBERS

By

Helen M. Walker
Teachers College, Columbia University

and

Vera Sanford
Oneonta State Normal School.

1. *General Considerations.*

The number of figures necessarily free from error in the result of a piece of computation may be determined by studying the relation between the number of digits in the result and the number of digits in the maximum error of the computation. It is the purpose of this essay to derive rules for the determination of the number of digits which are certain to be correct in computations based on measurement, but it must be understood that these rules state the minimum number of correct digits so that the result of a specific piece of computation may be accurate for more places than the rules indicate.

2. *Notation.*

Since the location of the decimal point has no connection with significant figures in a given number, it is assumed that the decimal point follows the last significant figure in each of the original numbers, the argument being somewhat simplified by this assumption. Accordingly the greatest error in the statement of the original numbers is ± 0.5. Let A and B be the true values of two numbers such that $A = a \cdot 10^m$ and $B = b \cdot 10^n$, where m and n are positive integers and where $0.1 \leqslant a < 1.0$ and where $0.1 \leqslant b < 1.0$. Then by the convention adopted above, the number of significant figures in A and B are m and n respectively, and the observed values are not less than $A - 0.5$ and $B - 0.5$ and not more than

$A + 0.5$ and $B + 0.5$. Let ϵ represent the maximum error in the computation and let ϵ' be the value of the largest term in the expansion of ϵ.

3. *Products.*

The greatest error in the product of A and B will occur when each is in excess by 0.5, the value of this error being

$$\epsilon = (A+.5)(B+.5) - AB = \tfrac{1}{2} a \cdot 10^{m} + \tfrac{1}{2} b \cdot 10^{n} + .25$$

For $ab > 0.1$, AB has $m+n$ digits to the left of the decimal point. For $ab < 0.1$, AB has $m+n-1$ digits to the left of the decimal point. The cases to be considered are

$$\begin{array}{ll} (\text{ I}) & m = n = 1 \\ (\text{ II}) & m = n > 1 \\ (\text{III}) & m - n = 1 \\ (\text{IV}) & m - n > 1 \end{array}$$

(I) *Let* $m = n = 1$. In this extreme case each factor consists of a single digit and the product consists of one or two digits. In this case, the figure in the unit's place is always affected by the maximum error and the figure in the ten's place, when present, is generally so affected.

(II) *Let* $m = n > 1$. Here $AB = ab \cdot 10^{2n}$ and

$$\epsilon = \tfrac{1}{2}(a+b) \cdot 10^{n} + .25 .$$

But $0.1 \leq \frac{a+b}{2} < 10$

and therefore $10^{n-1} + \frac{1}{4} \leq \epsilon < 10^{n} + \frac{1}{4}$.

The following conditions are possible:

Either (1) $ab \geq 0.1$ and AB has $2n$ digits to the left of the decimal point,

or (2) $ab < 0.1$ and AB has $2n-1$ digits to the left of the decimal point.

Either (3) ϵ has n digits to the left of the point,

or (4) ϵ has $n+1$ digits to the left of the point, the first one being 1 and all the others to the left of the decimal point being zero. This can occur only when ϵ is very near its maximum value. For example, when $n=4$, the value of ϵ must be less than 10000.25.

Then if conditions (1) and (3) are fulfilled, the result has n more digits to the left of the decimal point than has the error. A subsequent proof will show that this means that at least $n-1$ places in the result are free from error. Under conditions (1) and (4), the difference in the number of digits is $n-1$, and at least $n-2$ are not affected by the error. Similarly under conditions (2) and (3), $n-2$ digits are not affected by the error. Conditions (2) and (4) cannot occur simultaneously.

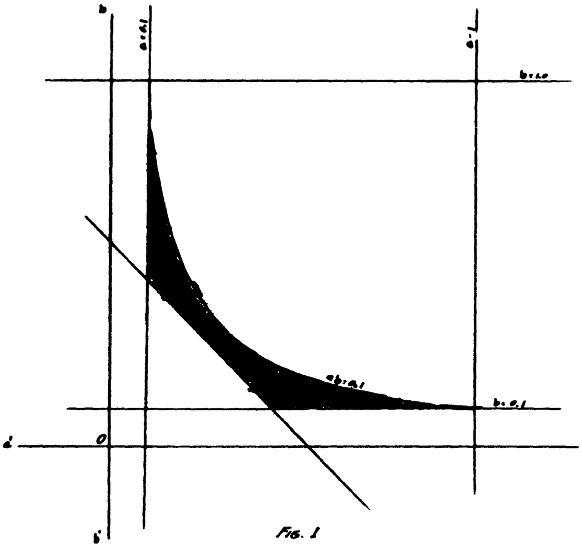

Fig. 1

The proof that conditions (2) and (4) are incompatible with the conditions that $0.1 \leqslant a < 1.0$ and $0.1 \leqslant b < 1.0$ may be obtained from fig. 1. The area within which these limits hold for a and for b is the area of the square bounded by $a = 1.0$, $a = 0.1$, $b = 1.0$, $b = 0.1$, the numerical value of this area being 0.81. The region within which $ab < 0.1$ is the area below the hyperbola $ab = 0.1$. The region within which $\frac{a+b}{2}$ is larger than a specified value is the region above the line $a+b=k$. Therefore the probability that all of these conditions shall be simultaneously fulfilled is the ratio of the shaded region in fig. 1 to the total area of the square, or to 0.81. When $\frac{a+b}{2} > 5$, this probability is only 0.000,014,8 and when $\frac{a+b}{2} > 0.55$, the probability vanishes altogether.

(III) *Let* $m-n=1$. Then $\epsilon = \frac{1}{2}(10a+b)\cdot 10^{n} + .25$.

But $1.1 \leqslant 10a+b < 11$.

Now either $n=1$ or $n>1$.

Let $n=1$. Then $5.75 \leqslant \epsilon < 55.25$.

Either (1) $ab \geqslant 0.1$ and AB has 3 digits to the left of the decimal point,

or (2) $ab < 0.1$ and AB has 2 digits to the left of the decimal point.

Either (3) $10a+b < 1.95$ and $5.75 \leqslant \epsilon < 10$,

or (4) $10a+b \geqslant 1.95$ and $10 \leqslant \epsilon < 55.25$.

If conditions (1) and (3) are met, the product has 2 more digits to the left of the decimal point than has the error. Thus one or two places will in general be free from error. Under conditions (1) and (4) or conditions (2) and (3) the number of such places free from error is 0 or 1. By an analysis similar to that given under (II) it appears that there is about one chance in ten that conditions (2) and (4) should be simultaneously met, in which case no place would be free from error.

Let $n=2$. Then $55.25 < \epsilon < 550.25$.

Either (1) $ab \geqslant 0.1$ and AB has 5 digits to the left of the decimal point,

or (2) $ab < 0.1$ and AB has 4 digits to the left of the decimal point.

Either (3) $10a+b < 1.995$ and $55.25 \leqslant \epsilon < 100$.

or (4) $10a+b \geqslant 1.995$ and $100 \leqslant \epsilon < 550.25$.

If conditions (1) and (3) are met, the product has either 2 or 3 places free from error. Under conditions (2) and (3) or conditions (1) and (4), the product has 1 or 2 places free from error. Simultaneous fulfilment of conditions (2) and (4) will be rare but not impossible. However in this case, the first digit of the error cannot be larger than 5, hence, as shown later, the number of digits free from error in the result will usually be the number in the product minus the number in the error, rather than one less than that.

For $n > 2$, the constant 0.25 forms a still smaller proportion of the error. Hence for larger values of n, if $m-n=1$, the product may be expected to have n or $n-1$ places free from error.

(IV) *Let* $m-n>1$. Here $m \geq 3$ and therefore the terms $\frac{1}{2}b\cdot 10^{n}$ and 0.25 are negligible in comparison with $\frac{1}{2}a\ 10^{m}$ and may be disregarded, since neither of them can affect the first place in the error. Then $\epsilon' = \frac{1}{2}a\cdot 10^{m}$, and therefore

$$0.5\ (10^{m-1}) < \epsilon' < 0.5\ (10^{m}).$$

Either (1) $ab \geq 0.1$ and AB has $m+n$ places to the left of the decimal point,

or (2) $ab < 0.1$ and AB has $m+n-1$ places to the left of the decimal point.

Either (3) $a < .2$ and $\epsilon' < 10^{m-1}$ so that ϵ has $m-1$ places to the left of the decimal point,

or (4) $a \geq .2$ and $10^{m-1} < \epsilon' < 0.5\ (10^{m})$, so that ϵ has m places to the left of the decimal point.

Conditions (1) and (3) would leave either $n+1$ or n places free from error in the product. Conditions (2) and (3) or conditions (1) and (4) would leave either n or $n-1$ places free from error. If conditions (2) and (4) are met, there would be $n-1$ places in the product to the left of the first digit in the error, and since this first digit is not more than 5, the error is not likely to affect the preceding digit. See section 6.

In general, therefore, if there are n significant figures in the less accurate of two approximations, the product of the two approximations will have n or $n-1$ digits free from error. The product of

two such numbers should be rounded off until it contains only as many significant figures as the less accurate of the two numbers. The last digit in the product may then contain some error.

4. *Quotients.*

The greatest error which can occur in a quotient arises when the dividend is in excess by 0.5 and the divisor in defect by the same amount.

Then $$\epsilon = \frac{A+.5}{B-.5} - \frac{A}{B} = \frac{b + a \cdot 10^{m-n}}{b(2b \cdot 10^{n} - 1)}$$

We must consider separately the cases (I) $m = n$

(II) $m > n$

(III) $m < n$

(I) *Let* $m = n$. Then $$\epsilon = \frac{b+a}{b(2b \cdot 10^{n} - 1)}$$

(a) If $m = n = 1$ there are 81 possible quotients of one-place numbers, and an examination of these shows that in only thirty-one of these cases the first digit is free from error.

(b) The case for $m = n > 1$ should be studied for specific values of n . In general, however, the error is large as $a \to 1$ and $b \to 0.1$, and is small as $a \to 0.1$ and $b \to 1.0$. Also as n increases, the influence of the constant term in the denominator becomes less. Therefore in general

$$\frac{0.55}{10} < \frac{1.1}{2 \cdot 10^{n} - 1} < \epsilon < \frac{1.1}{0.1[0.2(10^{n}) - 1]} = \frac{55}{10^{n} - 5} < \frac{0.61}{10^{n-2}}$$

If $a \to 1.0$, $b \to 0.1$ and n is large, then $\epsilon < \frac{0.61}{10^{n-2}}$ and the error has at least $n - 2$ zeros to the left of the first digit. In this case, $\frac{A}{B}$ has one digit to the left of the decimal point, so that the quotient will have at least $n-1$ digits to the left of the first digit in the error.

If $a \to 0.1$, $b \to 1.0$, and n is large, then $\epsilon > \frac{0.55}{10^{n}}$, and the error has n zeros to the left of the first digit. In this case $\frac{A}{B}$ has no digits to the left of the decimal, so that the quotient will again have $n-1$ digits to the left of the first digit in the error.

Furthermore, $\epsilon = (\frac{a}{b}+1)(2b\cdot 10^{n}-1)$, OR

$\epsilon = \frac{1}{2b\cdot 10^{n}} + \frac{a}{2b^{2}\cdot 10^{n}} +$ higher powers of 10^{-n}.
Then if $a \geqslant b$, $\frac{1}{b\cdot 10^{n}} \leqslant \epsilon' < \frac{55}{b\cdot 10^{n}}$. We then have 1 digit to the left of the decimal point in the quotient, and either $n-1$ or $n-2$ zeros preceding the first digit in the error.

If $a < b$, $\frac{0.55}{b\cdot 10^{n}} < \epsilon' < \frac{1}{b\cdot 10^{n}}$

since $0.1 > \frac{a}{b} > 10$.

In this case we have no digits to the left of the decimal point in the quotient, and either n or $n-1$ zeros preceding the first digit in the quotient.

(II) *Let* $m > n$.

(a) Let $m-n=1$.

Then $\epsilon = \frac{b+10a}{b(2b\cdot 10^{n}-1)} = (1+\frac{10a}{b})\{(2b\cdot 10^{n})^{-1}+(2b\cdot 10^{n})^{-2}+\cdots\}$.

$\therefore\ \epsilon = (1+\frac{10a}{b})(2b\cdot 10^{n})^{-1}$, the higher powers of $(2b\cdot 10^{n})^{-1}$ having no effect upon the first digit in the error.

If $a > b$, there are 2 digits to the left of the decimal point in the quotient and either $n-2$ or $n-3$ zeros in the error.

If $a < b$, there is 1 digit to the left of the point in the quotient and either $n-1$ or $n-2$ zeros in the error.

Only in rare cases will there be as few as $n-3$ zeros in front of the first digit in the error. To secure this ϵ must be greater than 10^{2-n}. This probability will differ for different values of n. For example, if $n=4$, we have as bounding conditions,

$$a > 20b^{2} - 0.101b$$
$$b < 1.0 \quad and \quad 0.1 < a < 1.0.$$

The ratio of the area bounded by $a = 20b^{2}-.101b$, $b=.1$, and $a=1.0$ to the area of the square bounded by $a=.1$, $a=1.0$, $b=.1$, and $b=1.0$, is

$$P = \frac{1}{81}\int_{b=0.1}^{b=\frac{.101+\sqrt{80.010201}}{40}} (20b^{2}-.101b)\,db = c.00084$$

which is the probability that there would be only $n-3$ zeros following the decimal point in the error.

(b) *Let* $m-n>1$. Then $\epsilon = \frac{b+a\cdot 10^{m-n}}{b(2b\cdot 10^{n}-1)}$

This situation should be studied for specific values of n. However an approximation may be obtained by letting

$$\epsilon' = \frac{b+a\cdot 10^{m-n}}{2b^2\cdot 10^{n}}$$

since subsequent terms in the expansion do not affect the first digit. If $a>b$, then $\epsilon' < \frac{1+10^{m-n+1}}{2b\cdot 10^{n}} = \frac{10^{m-2n+1}}{2b}$ + other terms which do not affect the first digit.

Also $\epsilon' > \frac{1+10^{m-n}}{2b\cdot 10^{n}} > \frac{10^{m-2n}}{2b} > 0.5(10^{m-2n})$

In this case the quotient has $m-n+1$ digits to the left of the decimal point, while the error has either $m-2n$, $m-2n+1$ or $m-2n+2$. Consequently there are either $n-1$, n, or $n+1$ digits free from error in the result.

If $a<b$, then $\frac{1+10^{m-n-1}}{2b\cdot 10^{n}} < \epsilon' < \frac{1+10^{m-n}}{2b\cdot 10^{n}}$

In this case the quotient has $m-n$ digits to the left of the decimal point, while the error has either $m-2n-1$, $m-2n$, or $m-2n+1$. Again there are either $n-1$, n, or $n+1$ digits free from error in the result.

(III) *Let* $m<n$.

Suppose $m+1=n$.

If $a>b$, the first digit of the quotient is immediately to the right of the decimal point, while there are from $m-1$ to $m+1$ zeros between the point and the first digit in the error.

If $a<b$, there is one zero between the decimal point and the first digit of the quotient, while in the error there are either m or $m+1$ zeros.

In general, therefore, if there are n digits in the less reliable of two approximations, there will be either n, $n-1$, or $n+1$ digits *free from error in their quotient.* In a few rare cases, a fortuitous combination of digits, discussed later, may throw the error back into the $n - 2$ place. *In general the quotient should be rounded off to contain only as many places as there are in the less accurate of the two numbers.*

5. *Square Root.*

When a number is in excess by 0.5, the error in its square root is $(A + 1/2)^{1/2} - A^{1/2}$

$$= 1/4 A^{-1/2} - 1/32 A^{-3/2} + 1/128 A^{-5/2} - \cdots + (-1)^{k-1} \frac{1 \cdot 3 \cdot 5 \cdots (2k-3)}{2^{2k} \cdot k!} A^{-\frac{2k-1}{2}} + \cdots .$$

When a number is in defect by 0.5, the error in its square root is $(A - 1/2)^{1/2} - A^{1/2}$

$$= 1/4 A^{-1/2} - 1/32 A^{-3/2} - 1/128 A^{-5/2} - \cdots - \frac{1 \cdot 3 \cdot 5 \cdots (2k-3)}{2^{2k} \cdot k!} A^{-\frac{2k-1}{2}} - \cdots .$$

Obviously the greater error occurs when the number is in defect by 0.5, but in either case we may neglect all terms after the first. Each term can readily be shown to be larger than the term following it, and the ratio of the first term to the second is so large that the second term cannot affect the first digit in the error.

We must consider in turn the case in which m is even and that in which m is odd.

(I) *Let* $m = 2r$.

Then $A = a \cdot 10^{2r}$ and has $2r$ digits to the left of the decimal point. $A^{1/2} = a^{1/2} \cdot 10^{r}$ and has r digits to the left of the decimal.

Now $|\epsilon'| = 1/4\, a^{-1/2} \cdot 10^{-r}$. But $1/4 < \frac{1}{4a^{1/2}} \leqslant \frac{1}{4\sqrt{0.1}} < .791$.

Since $0.25\,(10^{-r}) < |\epsilon'| \leqslant 0.791\,(10^{-r})$, the error has r zeros between its first digit and the decimal point.

Therefore when m is even, the root contains as many significant figures as the number.

(II) *Let* $m = 2r - 1$.

Then $A = a \cdot 10^{2n-1} = 10a \cdot 10^{2n-2}$

Then $A^{1/2} = (10a)^{1/2} \cdot 10^{n-1}$ and has n digits to the left of the decimal.

$$|\epsilon'| = 1/4\, A^{-1/2} = \frac{10^{1-n}}{4(10a)^{1/2}}$$

$$0.079 < \frac{1}{4\sqrt{10}} < \frac{1}{4(10a)^{1/2}} < \frac{1}{4} = 0.25$$

$$(.079)\ 10^{1-n} < |\epsilon'| < 0.25\,(10^{1-n})$$

The error then will affect the n^{th} place to the right of the decimal point, and the number of digits free from error will be $n+n-1=2n-1$ which was the number of places in the square.

(III) There is also the case where the decimal point is so placed that the second digit in the last period is not known, as in $\sqrt{32.4}$ or $\sqrt{0.46825}$.

Here $|\epsilon'| = \frac{5}{2} A^{-1/2}$ In this case also the number of digits free from error in the root is the number of digits in the original number.

In general, then, the number of digits free from error in the square root of a number is the number of digits in the number.

6. *Effect of the Error.*

The following table will illustrate how an error of n places may affect either n or $n+1$ places in the computation:

	ERROR IN EXCESS			ERROR IN DEFECT		
Result obtained by computation	6247	5986	7253	6247	5986	7253
Error	33	53	12	33	53	12
True value	6214	5933	7241	6280	6039	7265
Computed value, rounded	6200	6000	7300	6200	6000	7300
True value, rounded	6200	5900	7200	6300	6000	7300

We will now show that the chances are approximately 3 out of 4 that an error of n digits affects n and not $n+1$ places in the result. For convenience we may place the decimal point to the left of the first digit in the error, the position of the decimal point being entirely independent of the number of significant figures in the computation.

Let ϵ = error.

d = portion of the number to the right of the decimal point.

c = portion of the number to the left of the decimal point.

A = the true value of the number.

Then $c + d$ = result of computation.

$A = c + d - \epsilon$ = true value.

We will consider ϵ to be positive when the observed value is in excess and negative when it is in defect.

We will consider separately the case where the computed value is in excess and the case where it is in defect.

Suppose the result of computation to be in excess

1. (a) Then if $d > .5$ and $\epsilon > d - .5$
   (b) or $d < .5$ and $\epsilon > d + .5$
   the error will affect $n+1$ *places in the result.*

2. (a) If $d > .5$ and $\epsilon < d - .5$
   (b) or $d < .5$ and $\epsilon < d + .5$
   the error will affect only n *places in the result.*

3. (a) If $d > .5$ and $\epsilon = d - .5$
   (b) or $d < .5$ and $\epsilon = d + .5$
   (c) or $d = .5$ and $\epsilon > o$
   the error will affect either n *or* $n+1$ *places depending on whether the last digit of* c *is odd or even. This is on the assumption of the usual rule, that in rounding off the digit* 5 *the previous digit is made even.*

Since the number of digits in ϵ is finite, the values of d and of ϵ form discrete series, so that we shall have to think of $d = .5$ not as an infinitesimal but as a finite portion of the scale, ranging from $d = .495$ to $d = .505$ when $n = 2$, from $d = .4995$ to $d = .5005$ when $n = 3$, etc. If we map the region bounded by $d = 0$, $d = 1$, $\epsilon = 0$, $\epsilon = 1$ the proportions of area representing conditions (1), (2) and (3) represent the respective probabilities of these three sets of conditions. As n increases, the width of the strip $d = .5$ becomes smaller, the probability of (3) becomes smaller, and the probability of (2) approaches $3/4$.

When $n = 2$ these areas are respectively

1 (a) and 1 (b)	.245025
2 (a) and 2 (b)	.735075
3 (a), 3 (b), and 3 (c)	.0199
	1.0000000

We may assume that the last digit in c is as likely to be even as to be odd, we may say that the probability that the error will affect $n+1$ places in the result is slightly more than $1/4$ when there are two digits in ϵ. This ratio will approach $1/4$ if the number of digits in ϵ increases.

A similar argument holds when the result of computation is in defect.

7. *Summary of Rules.*

On the assumption that an error of n places affects only n places in the result we have the following rules:

If the less accurate of two approximate numbers contains n significant digits, their product and their quotient each contain n or $n-1$ significant digits.

The square root of a number contains as many significant figures as the number.

About once in four times, the error will affect one more place than these rules state, for the reasons given in section 6.