

TABLE 3. THE RANK OF  $Z'X$  FOR PARTIALLY BALANCED GROUP DIVISIBLE INCOMPLETE BLOCK DESIGNS WITH TWO ASSOCIATE CLASSES

Design	Rank
Singular	$d$
Semi-regular	$t-d+1$
Regular	$t$

## REFERENCES

- [1] Blackwell, D., "Conditional expectation and unbiased sequential estimation," *Annals of Mathematical Statistics*, 18 (1947), 105-10.
- [2] Bose, R. C., Clatworthy, W. H., and Shrikhande, S. S., "Tables on Partially Balanced Designs with two Associate Classes," *Technical Bulletin No. 107*, North Carolina Agricultural Experiment Station, Raleigh N. C., August, 1954.
- [3] Graybill, F. A., *An Introduction to Linear Statistical Models, Volume 1*. New York: McGraw-Hill, 1961.
- [4] Graybill, F. A., and Hultquist, R. A., "Some theorems concerning Eisenhart's model II," *Annals of Mathematical Statistics*, 32 (1961), 261-9.
- [5] Graybill, F. A., and Marsaglia, G., "Idempotent matrices and quadratic forms in the general linear hypothesis," *Annals of Mathematical Statistics*, 28 (1957), 678-86.
- [6] Graybill, F. A., and Weeks, D. L., "Combining inter-block and intra-block information in balanced incomplete blocks," *Annals of Mathematical Statistics*, 30 (1959), 799-805.
- [7] Lehmann, E. L., and Scheffé, H., "Completeness, similar regions, and unbiased estimation, Part I," *Sankhyā*, 10 (1950), 305-40.
- [8] Perlis, S., *Theory of Matrices*. Reading, Massachusetts: Addison and Wesley, 1952.
- [9] Thompson, W. A., Jr., "On the ratio of variances in the mixed incomplete block model," *Annals of Mathematical Statistics*, 26 (1955), 721-33.
- [10] Thompson, W. A., Jr., "The ratio of variances in a variance components model," *Annals of Mathematical Statistics*, 26 (1955), 325-9.
- [11] Weeks, D. L. and Graybill, F. A., "A minimal sufficient statistic for a general class of designs," *Sankhyā, A*, 24 (1962), 339-53.

## RATIOS OF NORMAL VARIABLES AND RATIOS OF SUMS OF UNIFORM VARIABLES

GEORGE MARSAGLIA

Boeing Scientific Research Laboratories

The principal part of this paper is devoted to the study of the distribution and density functions of the ratio of two normal random variables. It gives several representations of the distribution function in terms of the bivariate normal distribution and Nicholson's  $V$  function, both of which have been extensively studied, and for which tables and computational procedures are readily available. One of these representations leads to an easy derivation of the density function in terms of the Cauchy density and the normal density and integral. A number of graphs of the possible shapes of the density are given, together with an indication of when the density is unimodal or bimodal.

The last part of the paper discusses the distribution of the ratio  $(u_1 + \dots + u_n)/(v_1 + \dots + v_m)$  where the  $u$ 's and  $v$ 's are independent, uniform variables. The exact distribution for all  $n$  and  $m$  is given, and some approximations discussed.

### 1. INTRODUCTION

THE first part of this paper will discuss the distribution of the ratio of normal random variables; the second part, the distribution of the ratio of sums of uniform random variables. There does not seem to be much in the literature concerning the ratio of normal variables—there are some comments by Curtiss in his paper, [2], on the ratios of arbitrary variates, and papers by Fieller [4], and Geary [5], all of which are quite old. It might be thought that the subject is so simple that it was considered long ago, then dropped, but this is not quite the case. Unless the means are zero, where one easily gets the Cauchy distribution, the distribution of the ratio of normal variables does not respond readily to the devices that work so well for other important quotients in statistics, e.g., those of  $t$ ,  $z$ , or  $F$ . Curtiss remarks that it is apparently impossible to evaluate the density in closed form, a rather vague statement. We will derive the exact density of the ratio of two arbitrary normal variates by what might be called modern methods—not in the sense of using powerful new techniques, but merely by using properties of distributions that have been extensively studied in the intervening years. The density may be expressed as the product of a Cauchy density and a factor involving the normal density and integral, which might be considered a closed expression (equation (5) of Section 2). At any rate, there are now available a number of methods for handling the functions associated with the distribution and density of the ratio, and with the aid of a computer, we may study them in detail.

Aside from its frequent occurrence in problems involving the ratio of measured quantities with a random, presumably normal, error, the problem of the ratio of normal variates is of importance in regression theory. In fitting a line to points  $(x_1, y_1), \dots, (x_n, y_n)$ , the  $x$ 's assumed constant and the  $y$ 's independent normal with  $E(y_i) = \alpha + \beta x_i$ , one gets  $\hat{\alpha}$  and  $\hat{\beta}$  as estimates of  $\alpha$  and  $\beta$  by least squares. It is not unusual to find that the ratio of the two estimates

line in the form  $-\alpha/\beta$ , and thus the problem of the ratio of normal variates arises.

The following example of this problem occurs in medicine: in order to estimate the life span of the circulating red blood cells of a subject, a number of his red cells are labelled and then, by some means or other, the number of labelled cells still in the circulation is sampled, say, every 5 days for 50 days. This gives a sequence of points which are plotted and fitted with a straight line; the point where the line intercepts the time axis is used as the estimate of the red cell life span. It is important to know the distribution of this estimate about its true value—the normal red cell life span is about 120 days and shortened life spans are associated with various hematological disorders, most of them severe.

We will discuss the distribution and density of the ratio of two normal random variables in Section 2. In Section 4 we will discuss the distribution of ratios of the form  $(u_1 + \dots + u_n)/(v_1 + \dots + v_m)$  where the  $u$ 's and  $v$ 's are independent uniform variables; a recent paper, Locker and Perry [8], on this distribution for  $n=m=2$  led to its being considered here. We will find the exact distribution for all  $n$  and  $m$ , and examine the closeness of the normal approximation. On the way to finding the distribution of  $(u_1 + \dots + u_n)/(v_1 + \dots + v_m)$  we will need the distribution of a linear combination of uniform variates; some comments on this distribution and its history are in Section 3.

2. RATIOS OF NORMAL VARIABLES

We are concerned with the distribution of the ratio of two normal random variables. The problem has been discussed in the past, [2, 4, 5]. We will bring the problem up to date in this Section—give an explicit representation of the distribution in terms of what are now familiar functions, and discuss in more detail some of the properties of the distribution.

Let

$$w = \frac{a + x}{b + y} \tag{1}$$

where  $a, b$  are non-negative constants and  $x, y$  are independent standard normal random variables. It is easy to see that if  $w' = x_1/y_1$  is the ratio of two arbitrary normal random variables, correlated or not, then there are constants  $c_1$  and  $c_2$  such that  $c_1 + c_2 w'$  has the same distribution as  $w$ . It thus suffices to study the distribution of (1); translations and changes of scale will provide the distributions of the general ratio  $x_1/y_1$ .

The set of points  $(x, y)$  for which

$$\frac{a + x}{b + y} < t$$

is a region bounded by straight lines, and the normal probability measures of such regions have been extensively studied in the past few years. We should be able to express the distribution of  $w$  in terms of functions associated

with those measures, particularly the bivariate normal distribution function

$$L(h, k, \rho) = P[\xi > h, \eta > k]$$

where  $\xi$  and  $\eta$  are standard normal with covariance  $\rho$ , and the  $V$  function of Nicholson [11]:

$$V(h, q) = \int_0^h \int_0^{qx/h} \phi(x)\phi(y)dydx,$$

where  $\phi$  is the standard normal density. We have

$$\begin{aligned} P[w < t] &= P[a + x < t(b + y), b + y > 0] + P[a + x > t(b + y), b + y < 0] \\ &= P[-x + ty > a - bt, y > -b] + P[x - ty > -a + bt, y > b] \\ &= L\left(\frac{a - bt}{\sqrt{1 + t^2}}, -b, \frac{t}{\sqrt{1 + t^2}}\right) + L\left(\frac{-a + bt}{\sqrt{1 + t^2}}, b, \frac{t}{\sqrt{1 + t^2}}\right) \end{aligned}$$

Then using the elementary properties of the  $L$  and  $V$  functions (see, for example, the NBS table [10], p. vii),

$$\begin{aligned} L(-h, -k, \rho) &= L(h, k, \rho) + \int_0^h \phi(x)dx + \int_0^k \phi(x)dx \\ L(-h, -k, \rho) + L(h, k, \rho) &= 2V\left(h, \frac{k - \rho h}{\sqrt{1 - \rho^2}}\right) + 2V\left(k, \frac{h - \rho k}{\sqrt{1 - \rho^2}}\right) \\ &\quad + \frac{1}{2} + \frac{\sin^{-1} \rho}{\pi}, \end{aligned}$$

we have several representations of

$$F(t) = P\left[\frac{a + x}{b + y} < t\right]:$$

$$F(t) = L\left(\frac{a - bt}{\sqrt{1 + t^2}}, -b, \frac{t}{\sqrt{1 + t^2}}\right) + L\left(\frac{-a + bt}{\sqrt{1 + t^2}}, b, \frac{t}{\sqrt{1 + t^2}}\right), \tag{2}$$

$$F(t) = \int_0^{(bt-a)/\sqrt{1+t^2}} \phi(x)dx + \int_0^b \phi(x)dx + 2L\left(\frac{bt-a}{\sqrt{1+t^2}}, b, \frac{t}{\sqrt{1+t^2}}\right), \tag{3}$$

$$F(t) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} t + 2V\left(\frac{bt-a}{\sqrt{1+t^2}}, \frac{b+at}{\sqrt{1+t^2}}\right) - 2V(b, a). \tag{4}$$

Representation (4) appears best for numerical purposes, unless  $b$  is large, say  $b > 3$ , since we have good methods for providing values of  $V$  and  $1/\pi \tan^{-1} t$ , [9], [10], and [13]. This last reference, by D. B. Owen, also gives tables and formulas for the function

$$T(h, \lambda) = (2\pi)^{-1} \tan^{-1} \lambda - V(h, \lambda h),$$

which for some purposes is more convenient than the  $V$  function.

When  $b$  is large, the second and third terms of (3) may be replaced by .5 and

0, so that

$$P\left[\frac{a+x}{b+y} < t\right] \cong \frac{1}{2} + \int_0^{(bt-a)/\sqrt{1+t^2}} \phi(x)dx = \int_{-\infty}^{(bt-a)/\sqrt{1+t^2}} \phi(x)dx$$

provides very good numerical approximations to  $F(t)$ , plus the additional information that

$$(bw - a)/\sqrt{1+w^2}$$

is approximately normally distributed.

There is no need to go through complicated arguments involving the  $L$  and  $V$  functions in order to derive this approximation, however; it follows directly from the assumption that  $b+y > 0$ . In many applications one is more sure that the denominator of the ratio is positive than one is of the normality of  $x$  and  $y$ , so that the exact distribution of the ratio of normal variates may not approximate the practical ratio as well as the approximation given by

$$P\left[\frac{a+x}{b+y} < t\right] \cong P[a+x < bt+yt] = \int_{-\infty}^{(bt-a)/\sqrt{1+t^2}} \phi(x)dx.$$

Now we turn to the density of  $(a+x)/(b+y)$ . Let

$$h = \frac{bt-a}{\sqrt{1+t^2}}, \quad q = \frac{b+at}{\sqrt{1+t^2}}, \quad \lambda = \frac{q}{h} = \frac{b+at}{b-at}.$$

Using primes to indicate differentiation with respect to  $t$ , so that  $h' = q/(1+t^2)$ ,  $\lambda' = -(a^2+b^2)/(bt-a)^2$ , we differentiate (4) to get

$$f(t) = F' = \frac{1}{\pi(1+t^2)} + 2h'\phi(h) \int_0^q \phi(y)dy + 2\lambda' \int_0^h x\phi(x)\phi(\lambda x)dx.$$

Integrating the last term and simplifying, we get this form for  $f(t)$ , the density function of the ratio

$$f(t) = \frac{e^{-.5(a^2+b^2)}}{\pi(1+t^2)} \left[ 1 + \frac{q}{\phi(q)} \int_0^q \phi(y)dy \right], \quad q = \frac{b+at}{\sqrt{1+t^2}}. \quad (5)$$

Figure 1 shows  $f(t)$ , the density of  $(a+x)/(b+y)$ , for various values of  $a$  and  $b$ . The curves in Figure 1 were drawn by a computer; it also drew the identification for each density in the form

$$\frac{a+x}{b+y},$$

where  $a$  is a multiple of  $\frac{1}{3}$  and  $b$  a multiple of  $\frac{1}{8}$ . The values of  $a$  and  $b$  were chosen so as to give a rough indication of the possible shapes of the densities

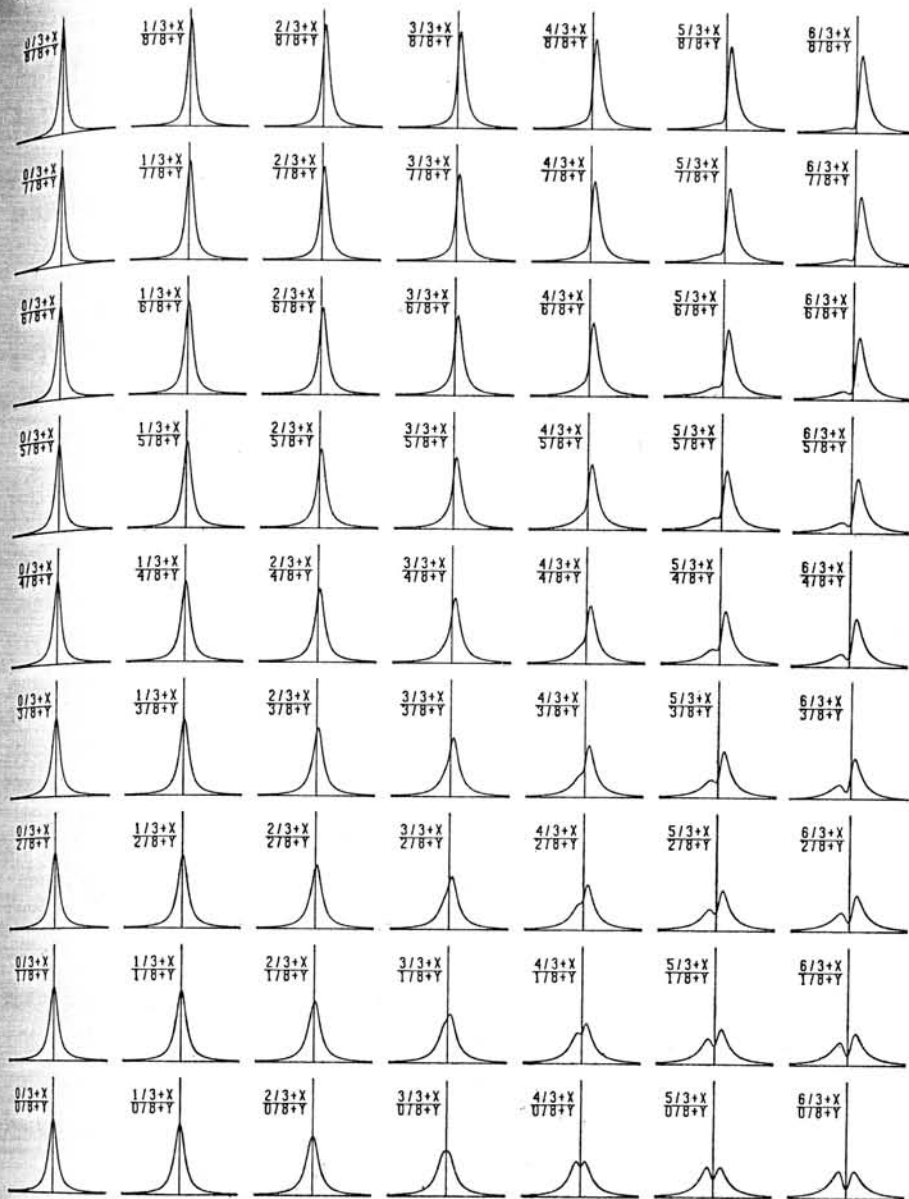


FIG. 1. Graphs of the density of  $(a+x)/(b+y)$ , where  $a > 0$ ,  $b > 0$  and  $x, y$  are independent, standard normal random variables. The formula for the density is in equation (5). Values  $a=0/3, 1/3, \dots, 6/3$  and  $b=0/8, 1/8, \dots, 8/8$  were chosen so as to represent the possible shapes of the density function.

The positive  $a, b$  quadrant may be divided into two regions according to whether the density of  $(a+x)/(b+y)$  is unimodal or bimodal, as in Figure 2. The curve that determines the two regions is asymptotic to  $a \cong 2.257$ . Thus when  $a > 2.257$ , the density of  $(a+x)/(b+y)$  is bimodal, even though it may not appear so. For example, the density of  $(10^6+x)/(10^6+y)$ ,  $x$  and  $y$  inde-

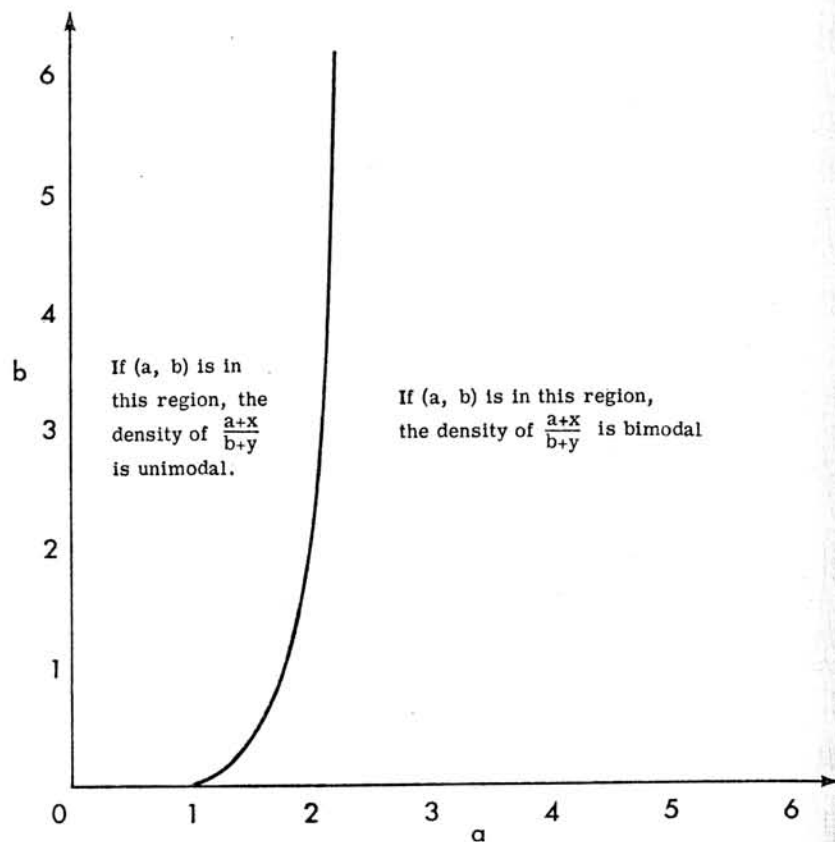


FIG. 2. The density of  $(a+x)/(b+y)$  is unimodal or bimodal according to the region of the positive quadrant in which the point  $(a, b)$  falls.

pendent standard normal, would appear to be a single spike at  $t=1$ , but in fact it has another mode somewhere in the vicinity of  $t=-10^{12}$ .

We conclude this Section with a summary.

*Summary of the Properties of the Ratio  $w = (a+x)/(b+y)$ , where  $x$  and  $y$  are Independent Standard Normal and  $a \geq 0, b \geq 0$ .*

$$w = \frac{a+x}{b+y}, \quad \text{where } a > 0, b > 0,$$

1. If  $w' = x_1/y_1$  is the ratio of any two jointly normal variables, then there are constants  $c_1$  and  $c_2$  so that  $c_1 + c_2 w'$  has the same distribution as  $w$ .

2. The distribution of  $w$ , say

$$F(t) = P\left[\frac{a+x}{b+y} < t\right],$$

may be expressed in terms of the bivariate normal distribution, or Nicholson's  $V$  function in several ways—formulas (2), (3), and (4) above.

3. When  $b$  is large, say  $b > 3$ , then

$$(bw - a)/\sqrt{1+w^2}$$

is approximately normally distributed, and

$$P[w \leq t] = P\left[\frac{a+x}{b+y} \leq t\right] \cong \int_{-\infty}^{(bt-a)/\sqrt{1+t^2}} \phi(u) du.$$

4. The density of

$$\frac{a+x}{b+y}$$

is given by formula (5). This density is plotted for various  $a$  and  $b$  in Figure 1.

5. The density of

$$\frac{a+x}{b+y}$$

is unimodal or bimodal according to the region of Figure 2 in which  $(a, b)$  lies. When  $a > 2.257$ , the density is bimodal, although one of the modes may be insignificant.

### 3. THE DISTRIBUTION OF $c_1 u_1 + \dots + c_n u_n$

Let  $u_1, \dots, u_n$  be independent random variables, each uniformly distributed over the interval  $(0, 1)$ . In the next Section we will need the distribution of a linear combination of the  $u$ 's,

$$c_1 u_1 + c_2 u_2 + \dots + c_n u_n \quad (6)$$

with the  $c$ 's positive. The general linear form in the  $u$ 's can readily be reduced to (6), for example

$$3u_1 - 2u_2 + 5u_3$$

has the same distribution as

$$3u_1 - 2(1 - u_2) + 5u_3 = 3u_1 + 2u_2 + 5u_3 - 2,$$

since  $1 - u_2$  has the same distribution as  $u_2$ .

There have been a number of discussions of the distribution of (6) in the literature—the problem (for equal  $c$ 's) dates back to Laplace [7], who solved it as a limiting form of the discrete case,<sup>1</sup> and, again with equal  $c$ 's, the result is in standard textbooks, e.g., Uspensky [17], who inverted the characteristic

<sup>1</sup> The discrete case of the problem, which may be viewed as the problem of finding the sum on  $n$  "dice," each one having a certain number of faces, has an even more curious history. In 1710, Montmort solved the problem for equal dice, as did DeMoivre in 1711, Simpson in 1740, LaGrange around 1770, and LaPlace in 1774. Montmort attempted, but did not solve, the problem of unequal dice. See Todhunter's *History* [16], Articles 148, 149, 364, 888, 915, 987.

function, and Cramér [1], proof by successive convolution. For unequal  $c$ 's the result was given by Olds [12], and the distribution appeared as a problem on volumes, [3], with subsequent remarks on its proof—particularly a development of Schoenberg [15], using recursive relations for spline curves. More recently, Roach [14], offered a geometric argument.

Thus the problem is now well known, and it is not particularly difficult, although notational difficulties, plus the fact that the problem may be viewed as one of probability, geometry, or spline functions, have led to a variety of proofs.

Roughly, the distribution of  $c_1u_1 + \dots + c_nu_n$  may be described as follows: Let  $S$  be the set of all  $2^n$  numbers which can be formed as a sum of different  $c$ 's:

$$S = \{0, c_1, \dots, c_n, c_1 + c_2, \dots, c_1 + \dots + c_n\}.$$

Then

$$P[c_1u_1 + \dots + c_nu_n < a] = \frac{1}{n!c_1c_2 \dots c_n} \sum_{s \in S, s < a} \pm (a - s)^n,$$

the  $+$  or  $-$  being according to whether there are an even or odd number of  $c$ 's used to form  $s$ . For example,

$$P[2u_1 + 3u_2 + 8u_3 < 7] = \frac{1}{3!(48)} [7^3 - (7-2)^3 - (7-3)^3 + (7-5)^3]$$

and

$$P[2u_1 + 3u_2 + 8u_3 < 12] = \frac{1}{3!(48)} [12^3 - (12-2)^3 - (12-3)^3 - (12-8)^3 + (12-5)^3 + (12-10)^3 + (12-11)^3]. \quad (7)$$

Note also that the distribution of  $2u_1 + 3u_2 + 8u_3$  is symmetric (any linear combination of independent symmetric random variables is symmetric), and that, rather than compute expression (7), one might consider

$$\begin{aligned} P[2u_1 + 3u_2 + 8u_3 < 12] &= P[2(1 - u_1) + 3(1 - u_2) + 8(1 - u_3) < 12] \\ &= P[2u_1 + 3u_2 + 8u_3 > 1] = 1 - \frac{1}{3!(48)}. \end{aligned}$$

We may formally describe the distribution of  $c_1u_1 + \dots + c_nu_n$  as follows:

*Theorem 1.* Let  $u_1, u_2, u_3, \dots, u_n$  be independent random variables, each uniformly distributed over the interval  $(0, 1)$ , and let  $c_1, c_2, \dots, c_n$  be positive constants. Let

$$F_n(a) = \text{Prob}[c_1u_1 + \dots + c_nu_n \leq a]$$

and let

$$g_n(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{x^n}{n!c_1c_2 \dots c_n} & \text{if } 0 < x. \end{cases}$$

Then, for  $0 < a < c_1 + \dots + c_n$ ,

$$\begin{aligned} F_n(a) &= g_n(a) - \sum_i g_n(a - c_i) + \sum_{i < j} g_n(a - c_i - c_j) \\ &\quad - \sum_{i < j < k} g_n(a - c_i - c_j - c_k) + \dots \end{aligned}$$

The theorem may be easily proved by induction, using the elementary results:

$$F_{n+1}(a) = \frac{1}{c_{n+1}} \int_0^{c_{n+1}} F_n(a - x) dx$$

and

$$\frac{1}{c_{n+1}} \int_0^{c_{n+1}} g_n(b - x) dx = g_{n+1}(b) - g_{n+1}(b - c_{n+1}).$$

When the  $c$ 's are all equal to 1, the result takes the following form:

$$P[u_1 + \dots + u_n < a] = \frac{1}{n!} \left[ a^n - \binom{n}{1} (a-1)^n + \binom{n}{2} (a-2)^n \dots \right],$$

where the terms are taken as long as  $a, a-1, a-2, \dots$ , are positive. More formally, for  $0 \leq a \leq n$ , and with the greatest integer notation,

$$P[u_1 + \dots + u_n < a] = \frac{1}{n!} \sum_{i=0}^{[a]} (-1)^i (a-i)^n.$$

#### 4. THE DISTRIBUTION OF $\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m}$

Let  $u_1, u_2, \dots, u_n, v_1, \dots, v_m$  be independent random variables, each uniform over  $(0, 1)$ . We want the distribution of

$$\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m}. \quad (8)$$

The distribution of (8) is of interest in studying round-off error propagation in numerical analysis, see references [6], [18]. The particular case  $m=n=2$  was worked out in detail in reference [8]. We will find the distribution of (8) for all  $n$  and  $m$ , by applying the results of the previous Section, and will, in addition, discuss approximations to the distribution.

Since  $1-v_i$  is distributed as  $v_i$ , we have

$$\begin{aligned} P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} < a\right] &= P\left[\frac{u_1 + \dots + u_n}{(1-v_1) + \dots + (1-v_m)} < a\right] \\ &= P[u_1 + \dots + u_n + av_1 + av_2 + \dots + av_m < ma] \end{aligned}$$

and hence a direct application of Theorem 1 gives (after a little thought about

how the terms combine):

$$P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} < a\right] = \frac{1}{(n+m)!a^m} \sum_{i=0}^{[ma]} \sum_{j=0}^{[(ma-i)/a]} (-1)^{i+j} \binom{n}{i} \binom{m}{j} [(m-j)a - i]^{n+m}$$

For example,

$$P\left[\frac{u_1 + \dots + u_7}{v_1 + \dots + v_5} < .9\right] = \frac{1}{12!(.9)^5} \left[ \begin{aligned} & \binom{7}{0} \left[ \binom{5}{0} (4.5)^{12} - \binom{5}{1} (3.6)^{12} + \binom{5}{2} (2.7)^{12} - \binom{5}{3} (1.8)^{12} + \binom{5}{4} (.9)^{12} \right] \\ & - \binom{7}{1} \left[ \binom{5}{0} (3.5)^{12} - \binom{5}{1} (2.6)^{12} + \binom{5}{2} (1.7)^{12} + \binom{5}{3} (.8)^{12} \right] \\ & + \binom{7}{2} \left[ \binom{5}{0} (2.5)^{12} - \binom{5}{2} (1.6)^{12} + \binom{5}{3} (.7)^{12} \right] \\ & - \binom{7}{3} \left[ \binom{5}{0} (1.5)^{12} - \binom{5}{2} (.6)^{12} \right] \\ & + \binom{7}{4} \left[ \binom{5}{0} (.5)^{12} \right] \end{aligned} \right]$$

The variate  $(u_1 + \dots + u_n)/(v_1 + \dots + v_m)$  is approximately a ratio of independent normal variables, and the discussion of Section 2 should apply. We may derive a good normal approximation directly, however, writing

$$P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} < a\right] = P[u_1 + \dots + u_n + av_1 + \dots + av_m < ma].$$

Since the sum on the right is approximately normal with mean  $.5[n+ma]$  and variance  $(a^2m+n)/12$ , we have

$$P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} < a\right] \cong \phi\left[\frac{\sqrt{3}(am-n)}{\sqrt{a^2m+n}}\right].$$

Figure 3 gives some indication of the merits of this approximation. The function plotted is

$$\text{error}(x) = P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} < x\right] - \int_{-\infty}^{\sqrt{3}(xm-n)/\sqrt{x^2m+n}} \phi(t) dt$$

for  $m=n=3, 4, 5, 6, 8, 10$ . The shapes of these error curves are similar, being stretched and flattened as  $n$  and  $m$  get large; their shapes resemble those of the normal derivatives, suggesting that the first few terms of a Gram-Charlier expansion would give even better approximations.

In case it is necessary to get the tail of the distribution with great precision,

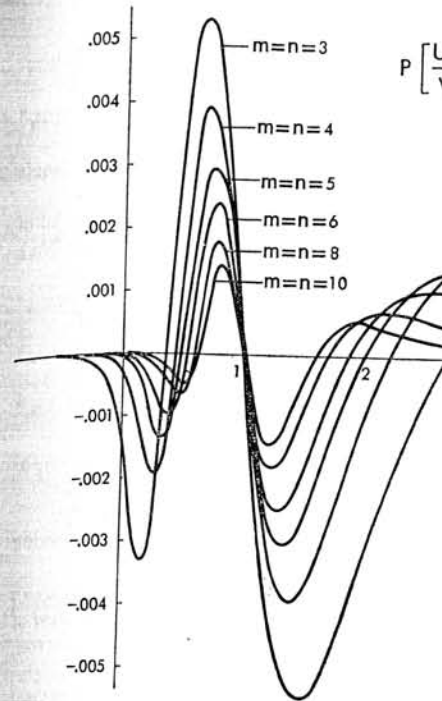


FIG 3

it is not too difficult to calculate the

$$P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} < a\right] = \frac{a^n}{(n+m)!} \left[ m^{n+m} - \binom{m}{1} (m-1)^{n+m} + \dots \right]$$

and for  $b > n$ ,

$$P\left[\frac{u_1 + \dots + u_n}{v_1 + \dots + v_m} > b\right] = \frac{b^{-m}}{(n+m)!} \left[ n^{n+m} - \binom{n}{1} (n-1)^{n+m} + \dots \right]$$

Some supplementary references: The useful references on the ratio of normal which deals with the ratio for a fiducial equation (5) of this article, and a survey by Eisenhart and Zelen, [2s], pages 1-151 the distribution of the ratio of normal v (Equation 12.157 of that reference. The proper relation  $P[Y/X < Z] = P[Y < ZX]$

## REFERENCES

- [1] Cramér, Harald, *Mathematical Methods of Statistics*. Princeton, New Jersey: Princeton University Press, 1946. pp. 244-6.
- [1s] Creasy, Monica A., "Symposium on interval estimation: limits for the ratio of means," *Journal of the Royal Statistical Society, Series B*, (1954), 186-94.
- [2] Curtiss, J. H., "On the distribution of the quotient of two chance variables," *Annals of Mathematical Statistics*, 12 (1943), 409-21.
- [2s] Eisenhart, Churchill, and Zelen, Marvin, "Elements of probability," Chapter 12 of *Handbook of Physics*, Condon and Odishaw, Editors, New York: McGraw-Hill Book Company, 1958.
- [3] Eisenstein, Maurice, and Klamkin, M. S., "Problem 59-2,  $N$ -dimensional volume," *SIAM Review*, 1 (1959), 69.
- [4] Fieller, E. C., "The distribution of the index in a normal bivariate population," *Biometrika*, 24 (1932), 428-40.
- [5] Geary, R. C., "The frequency distribution of the quotient of two normal variates," *Journal of the Royal Statistical Society*, 93 (1930), 442-6.
- [6] Inman, S., "The probability of a given error being exceeded in approximate computations," *The Mathematics Gazette*, 34 (1950), 99-113.
- [7] Laplace, P., *Théorie Analytique des Probabilités*. Paris, France: Courcier, 1812.
- [8] Locker, John, and Perry, N. C., "Probability functions for computations involving more than one operation," *Mathematics Magazine*, 35 (1962), 87-9.
- [9] Marsaglia, G., "Tables of  $1/2\pi \tan^{-1}(\lambda)$  and  $\tan^{-1}(\lambda)$  for  $\lambda = .0001, .0002, \dots, .9999$ , with some remarks on their use in finding the normal probability measure of polygonal regions," Boeing Scientific Research Laboratories, Seattle, Washington, Document D1-82-0078, (1960).
- [10] National Bureau of Standards, *Tables of the Bivariate Normal Distribution and Related Functions*, Government Printing Office, Applied Mathematics Series 50, Washington, D. C., 1959.
- [11] Nicholson, C., "The probability integral for two variables," *Biometrika*, 33 (1943), 59-72.
- [12] Olds, E. G., "A note on the convolutions of uniform distributions," *Annals of Mathematical Statistics*, 23 (1952), 282-5.
- [13] Owen, D. B., "Tables for computing bivariate normal probabilities," *Annals of Mathematical Statistics*, 27 (1956), 1075-90.
- [14] Roach, S. A., "The frequency distribution of the sample mean when each member of the sample is drawn from a different rectangular distribution," *Biometrika*, 50 (1963), 508-13.
- [15] Schoenberg, I. J., "Solution to problem 59-2,  $N$ -dimensional volume," *SIAM Review*, 2 (1960), 41-5.
- [16] Todhunter, I., *A History of the Mathematical Theory of Probability*. London: Macmillan, 1865. New York: Chelsea Reprinted Edition, 1949.
- [17] Uspensky, J. V., *Introduction to Mathematical Probability*. New York: McGraw-Hill Book Company, 1937. Pp. 277-8.
- [18] Woodward, R. S., *Probability and Theory of Errors*. New York: John Wiley, 1906.

## DESIGN FOR OPTIMAL PREDICTION IN SIMPLE LINEAR REGRESSION\*

D. W. GAYLOR AND H. C. SWEENEY

*Research Triangle Institute*

Allocation of experimental data points is considered in order to provide efficient prediction of the dependent variable in simple linear regression. The region of experimental points is taken such that it need not coincide with the region for prediction where the regression is linear. The allocation which minimizes the maximum variance for a predicted value of the dependent variable is obtained. The allocation of experimental data points which minimizes the average variance of predicted values, occurring according to a density function in the region of prediction, is derived. The relative efficiency of balanced allocation (one-half of the data points at each end of the experimental region) to minimax or minimum average variance allocation is about 90 percent for prediction near the ends of the experimental region with small samples.

### 1. INTRODUCTION

THE work in this paper concerns the selection of values of the independent variable  $x$  which should be run in order to estimate a simple linear regression function in some optimal manner. It will be assumed that the purpose of estimating the regression equation is to allow values of the dependent variable  $y$  to be predicted at values of the independent variable which may not have been run; this prediction may involve either interpolation or extrapolation. Values of the independent variable used in establishing the regression function will be denoted by  $x$ ; values of the independent variable used in prediction will be denoted by  $z$ . It is assumed that the values of the independent variable  $x$  which may be used in establishing the regression function are bounded from above and below; without loss of generality, these bounds may be taken to be  $0 \leq x \leq 1$ . In similar fashion, it will be assumed that the values of the independent variable  $z$  for which predictions are to be made are constrained to lie in some interval  $z_0 \leq z \leq z_1$ .

Optimal designs have been considered by Smith [7], de la Garza [3], Guest [4], David and Arens [2] and Kiefer and Wolfowitz [6]; in these papers, the experimental range of  $x$  exactly coincides with the range of interest of predicted values  $z$ :  $z_0=0$ ,  $z_1=1$ . Daniel and Heerema<sup>1</sup> [1] have considered the problem of optimal spacing for linear extrapolation to a single value  $z$  outside the range  $[0, 1]$ . The work reported herein generalizes the problem of optimum allocation for prediction in any range of interest  $[z_0, z_1]$  which may be completely within, partially within, or wholly outside the possible experimental

\* Research work done under contract for the Research and Development Department, Union Carbide Chemicals Division, South Charleston, West Virginia.

<sup>1</sup> Hoel and Levine [5] have recently considered prediction in regions of the type  $0 \leq z \leq t$  where  $t > 1$  and  $t_1$  is a value satisfying a certain equation.