T. R. No. 47

01c.

Technical Report

SOME RESULTS ON THE USE OF RANDOM NUMBERS IN SAMPLING FROM FINITE AND INFINITE POPULATIONS

by

Y.P.Sabharwal

विद्या विनियोगाद्विकासः

IIM

AHMEDABAD

INDIAN INSTITUTE OF MANAGEMENT

AHMEDABAD

SOME RESULTS ON THE USE OF RANDOM NUMBERS IN SAMPLING FROM FINITE AND INFINITE POPULATIONS

by
Y.P.Sabharwal

T.R. No. 47

August 1974

Indian Institute of Management
Ahmedabad

No. [illegible]

To

Chairman (Research)
IIMA

Technical Report

Title of the report: Some [illegible] on the [illegible] of Random Numbers [illegible] in Sampling from Finite and Infinite Populations

Name of the Author: [illegible] (F. B. A.)

Under which area do you like to be classified? P & QM

ABSTRACT (within 250 words)

Conceptual framework is provided for some of the well known ideas relevant to the use of random numbers for sampling. Results on the efficient choice of d (the number of digits in the column of random numbers) are envisaged to be of use to the practitioner in designing a simulation exercise. The exact and approximate methods of sampling ~~from~~ in ~~numerous~~ different situations are also given. Most of the work is limited to univariate populations, but briefly methods are indicated for dealing with bivariate or multivariate cases.

Please indicate restrictions if any that the author wishes to place upon this note: NIL

[illegible]

Signature of the Author

Date Aug. 21, 1974.

ACKNOWLEDGEMENTS

An earlier version of this note was prepared for a meeting (held in May 1969) of the Actuarial Society of India, Delhi Branch, on the advice of Mr.D.R.Iyer of L.I.C. Author received encouragement from Prof.Mohan Kaul to revise it in the present form. Thanks are also due to Mr.Rajagopalan(who was in Chair at the said meeting), Mr.Padmanabhan and other members of the Society who participated in the discussion.

CONTENTS

1. Introduction

Sampling is an important tool for drawing quick conclusions about a mass of data. Resort to sampling is often necessary for reasons other than saving in time too. An interesting science based on the concept of scientific sampling is that of Simulation. It is proposed in this expository note to discuss at length the use of random numbers in scientific sampling from finite and infite populations.

Although the basic idea is old, the subject matter presented here is relatively new. Readable text books are only just beginning to appear and the subject is not a tidy one. Further, the authors of these books assume the consideration of such results, that are presented here, to be the responsibility of the authors of text books on probability theory, who have not in fact looked into these results from this point of view.

2. Finite Populations :

By a population we mean a collection of units. A finite population comprising of N units may be specified either by its Frame(Table 1) or by the Frequency Distribution (Tables 2(a) & 2(b) of certain characteristic Y possessed by the units.

Table 1 : Frame of a Finite Population.

S.No.	Unit Identity	Location	Characteristic Y	Some Auxiliary Informations A
1	U_1	L_1	Y_1	A_1
2	U_2	L_2	Y_2	A_2
⋮	⋮	⋮	⋮	⋮
N	U_N	L_N	Y_N	A_N

Table 2(a): Frequency Distribution of the characteristic Y

Characteristic Y	Frequency f	Cumulative Frequency C^1
y_1	f_1	$C_1^1 = f_1$
y_2	f_2	$C_2^1 = f_1+f_2$
$\vdots$	$\vdots$	$\vdots$
y_k	f_k	$C_k^1 = f_1+f_2+\ldots+f_k = N$

Table 2(b) : Frequency Distribution of the characteristic Y.

Characteristic Y	Frequency f	Cumulative Frequency C^1
$l_1 - u_1$	f_1	$C_1^1 = f_1$
$l_2 - u_2$	f_2	$C_2^1 = f_1+f_2$
$\vdots$	$\vdots$	$\vdots$
$l_k - u_k$	f_k	$C_k^1 = f_1+f_2+\ldots+f_k = N$

In fact, if the populsation is specified by a frequency distribution, we can imagine a frame so that units number 1 to $f_1 = C_1^1$ possess the characteristic value y_1 or l_1-u_1, as the case may be, units number f_1+1 to $f_1+f_2 = C_2^1$ possess the characteristic value y_2 or l_2-u_2, as the case may be, and so on.

3. Infinite Populations

An infinite population* is specified by the probability distribution $\{Y_j\ p(Y)\}$ or $\{Y_j\ f(Y)\}$, where $p(Y)$ and $f(Y)$ are respectively the probability function and the probability density function, i.e.

$$p(y_j) = P_r(Y = y_j)$$

$$\text{and}\quad dP(t) = P_r(t-\tfrac{1}{2}dt \leq Y \leq t + \tfrac{1}{2}\,dt) \simeq f(t)\,dt.$$

In the latter case,

$$P_r\ (l_i \leq Y \leq u_i) = \int_{l_i}^{u_i} f(t)\,dt.$$

The physical implication of such populations is a collection of units with characteristic value Y having relative frequency distribution as given above.

4. Sample and Simple Random Sampling

A sample is a "fraction" of the population. Thus a sample of size n drawn from the population will comprise of n units (not necessarily all distinct - which calls for the quotes on fraction) $u_1, u_2 \ldots, u_n$, say. Each u_i is one of the units of the population. Further, we shall denote by x_i the characteristic value of u_i.

Sampling is the process of selecting units from the population. We shall write :

(i) for finite populations

$u_i = U_j$ to imply that the jth unit of the population is included in the sample at the ith selection, and

(ii) For infinite populations

$x_i = y$ to imply the inclusion of a unit in the sample at the ith selection which possesses the characteristic value y.

* Some authors prefer to call it a hypothetical population.

Simple random sampling (s.r.s.) is basic to all the sampling procedures. The probability rules for s.r.s. may be expressed as follows :

Simple Random Sampling with Replacement

(i) Finite Population.

$$P_r\,(u_i \equiv U_j) = \frac{1}{N} \qquad \text{for all } i \text{ and } j,$$

$$P_r\,(u_{i_1} \equiv U_{j_1},\ u_{i_2} \equiv U_{j_2}) = \begin{cases} o & i_1 = i_2,\ j_1 \neq j_2 \\ \frac{1}{N} & i_1 = i_2,\ j_1 = j_2 \\ \frac{1}{N2} & i_1 \neq i_2 \end{cases}$$

etc.

(ii) Infinite Population.

$$P_r\,(x_i = y_j) = p(y_j) \qquad \text{for all } i \text{ and } j,$$

$$P_r\,(x_{i_1} = y_{i_1},\ x_{i_2} = y_{j_2}) = \begin{cases} o & i_1 = i_2,\ j_1 \neq j_2 \\ p(y_{j_1}) & i_1 = i_2,\ j_1 = j_2 \\ p(y_{j_1}) * p(y_{j_2}) & i_2 \neq i_2 \end{cases}$$

etc.

OR $dP(x_i) = f(x_i)\,dx_i$ for all i

$$dP(x_{i_1}, x_{i_2}) = \begin{cases} f(x_{i_1}) * f(x_{i_2})\,dx_{i_1}\,dx_{i_2} & i_1 \neq i_2 \\ f(x_{i_1}) * dx_{i_1} & i_1 = i_2 \end{cases}$$

etc.,

According as Y is discrete or continuous.

Simple Random Sampling Without Replacement

(i) Finite Population

$$P_r\,(u_i \equiv U_j) = \frac{1}{N}$$

$$P_r\,(u_{i_1} \equiv U_{j_1},\ u_{i_2} \equiv U_{j_2}) = \begin{cases} o & i_1 = i_2, j_1 \neq j_2 \ \&\ i_1 \neq i_2, j_1 = j_2 \\ \frac{1}{N} & i_1 = i_2, j_1 = j_2 \\ \frac{1}{N(N-1)} & i_1 \neq i_2,\ j_1 \neq j_2 \end{cases}$$

etc.

(ii) Infinite Population:

For an infinite population the concept of sampling without replacement has little practical significance.

5. Random Numbers

A one-digit column of random numbers comprises of digits 0,1,2,3,4, 5,6,7,8 and 9, with the provision that if the t^{th} entry in the column is x_t, then

$$P_r\,(x_t = j) = \frac{1}{10} \quad \text{for all } t \text{ and } j = o(1)9,$$

$$P_r\,(x_{t_1} = j_1,\ x_{t_2} = j_2) = \begin{cases} 0 & t_1 = t_2,\ j_1 \neq j_2 \\ \frac{1}{10} & t_1 = t_2,\ j_1 = j_2 \\ \frac{1}{100} & t_1 \neq t_2. \end{cases} \qquad .....(1)$$

A number taken from this column is thus a random selection from the 10 numbers o(1)9. A two-digit column of random numbers comprises of numbers 00,01, ..., 09, 10, 11,....,99, with the provision that if the t^{th} entry is $x_t y_t$, so that the number is $x_t * 10^1 + y_t * 10^0 = 10\,x_t + y_t$, then x_t and y_t individually satisfy (1) and $P_r(x_t = i, y_t = j) = P_r(x_t = i) * P_r(y_t = j)$ for all i and j.

A number taken from this column is thus a random selection from the 100 numbers 00(01)99. Evidently a two-digit column of random numbers can be formed by placing two one-digit columns of random numbers parallel to each other, the location of the matching being determined arbitrarily. This affords an easy extension to the case of d-digit column of random numbers. A number taken from this column is a random selection from the 10^d numbers $0....0(0....01)10^d-1$. Of course, d is a positive integer.

A bibliography, with notes, of the important published series of random numbers appears in "The Advanced Theory of Statistics Vol.1" by M.G.Kendall & A.Stuart, Charles Griffin & Co.Ltd., London.

6. Sampling from finite populations :

As pointed out in § 4, s.r.s. is basic to all the sampling procedures. It is sufficient, therefore, to describe below the use of random numbers for s.r.s. only.

Consider first the case of population being specified by its frame.

Case 1 : $N = 10^d$

This case is simplest in that we can associate the number i, $(i=0,1,\ldots,10^d-1)$, of the d-digit column of random numbers to the $(i+1)$th unit of the population; make a random choice (without being prejudice as to the placement of numbers in the column) for the starting point of the column of random numbers*. Following results follows for ~~the~~ s.r.s. with and without replacement :

(i) Simple Random Sampling with Replacement

(1) Since $P_r(u_1 \equiv U_{j_1}+1) = \frac{1}{10^d} = \frac{1}{N}$

$$P_r(u_2 \equiv U_{j_2}+1 \quad u_1 \equiv U_{j_1}+1)$$

$$= P_r(u_2 \equiv U_{j_2}+1)$$

$$= \frac{1}{10^d} = \frac{1}{N} \ ,$$

with $j_1, j_2, \ldots, j_n$ as the first n entries in the column, the n units included in the sample are $U_{j_1}+1, U_{j_2}+1, \ldots, U_{j_n}+1$.

(2) Probability that any particular unit is included in the sample exactly r times is

$$\binom{n}{r}\left(\frac{1}{N}\right)^r \left(1-\frac{1}{N}\right)^{n-r} ,$$

with the mean n/N and variance $n(N-1)/N^2$.

* When a device for generating random numbers, e.g. a computer routine, is used in lieu of the column, this point is automatically taken care of.

(3) Probability $p(r,n)$ that a sample of size n contains exactly r distinct units is given by the difference equation :

$$p(r,n) = p(r-1,n-1)\ \frac{N-r+1}{N} + p(r,n-1)\ \frac{r}{N}\,,$$

$$n > 1 \text{ and } r = 2(1)n,$$

with $p(0,n) \equiv 0$,
$p(1,n) = (\frac{1}{N})^{n-1}$
$p(r,n) \equiv 0, \qquad r > n.$

The average number of distinct units in a sample of siz n is $N\left[1-(\frac{N-1}{N})^n\right]$.

(4) Probability $p_n(r)$ that a sample of exactly n units will be required to include exactly r distinct elements is given by the difference equation :

$$p_n(r) = p_{n-1}(r)\ \frac{r-1}{N} + p_{n-1}(r-1)\ \frac{N-r+1}{N}$$

The average value of n is

$$N \sum_{k=0}^{r-1} \frac{1}{N-k} \simeq N \log_e \frac{N}{N-r+1}, \text{ for large N.}$$

(5) Also, $p_n(r)$ and $p(r,n)$ are related by the relation :

$$p(r,n) = \sum_{k=r}^{n} p_k(r)\ (\frac{r}{N})^{n-k} \quad \text{and}$$

$p_n(r)$ is the coefficient of Z^n in

$$Z^{r} \prod_{k=1}^{r-1} \left(\frac{N-k}{N-kZ}\right).$$

(6) Probability $p_n^o(r)$ that a sample of exactly n units will be required to include the r preassigned units instead of r arbitrary ones is the coefficient of Z^n in

$$Z^r \prod_{k=1}^{r} \left(\frac{k}{N-(N-k)Z}\right).$$

(ii) **Simple Random Sampling Without Replacement**

(1) $u_1 \equiv U_{j_1}+1$ if the opening number in the column is j_1. Of course,

$P_r(u_1 \equiv U_{j_1}+1) = \frac{1}{N}$

Let the next number in the column be j_2^o. If $j_2^o = j_1$, then pass on to the next number. Among the following numbers, let j_2 be the first such number which is different from j_1. Then $u_2 = U_{j_2}+1$. We verify that

$$P_r(u_2 \equiv U_{j2}+1 \mid u_1 \equiv U_{j_1}+1)$$
$$= P_r(j_2 \neq j_1)$$
$$= \frac{1}{10^d-1} = \frac{1}{N-1}.$$

Next, $u_3 \equiv U_{j_3}+1$ if j_3 is the first next number other than j_1 and j_2, and so on. Thus, the n units in the sample will be $U_{j_1}+1, U_{j_2}+1, \ldots,$ $U_{j_n}+1$, where $j_1, j_2, \ldots, j_n$ are the first n distinct numbers in the column.

(2) In this case the time required to complete sampling (measured in terms of the number of random numbers required for the purpose) is a random variable. Probability that in-between j_1 and j_2 there are exactly (r-1) numbers (each equal to j_1) is $\frac{N-1}{N}(\frac{1}{N})^{r-1}$. Probability that in-between j_2 and j_3 there are exactly (r-1) numbers (each equal to j_1 or j_2) is $\frac{N-2}{N}(\frac{2}{N})^{r-1}$.

Thus the average time required for a sample of size 2 is

$$2 + \sum_{r=1}^{\infty} (r-1)\left(\frac{N-1}{N}\right)\left(\frac{1}{N}\right)^{r-1} = 2 + \frac{1}{N-1}$$

In general the average time required for a sample of size n would be

$$n + \sum_{k=1}^{n-1} \frac{k}{N-k} \simeq N \log_e N/(N-n+1), \text{ for large } N.$$

Case 2: $10^{d-1} < N < 10^d$:

In this case we associate the $(r+1)$ numbers j, $j+N$, $j+2N$, ..., $j+2\,N$, $j = 0(1)\,N-1$ and r is the largest positive integer with $N-1 + Nr \quad 10^d-1$, of the d-digit column of random numbers to the $(j+1)^{th}$ unit U_j+1 of the population. To this set of integers we shall denote by A_j+1 and by A the set $\bigcup_{j=0}^{N-1} A_j+1$. The remaining $(10^d-1)-(N-1+Nr) = 10^d - N(r+1)$ numbers of the column are not associated with any of the N units of the population. We denote the set of these integers by I. As in the previous case, make a random choice for the starting point of the column.

(1) Simple Random Sampling With Replacement

(1) $u_1 \equiv U_{j_1}+1$ if the number in the column which belongs to A is one of the $(r+1)$ numbers $j_1(N)j_1+Nr$. Of course

$$P_r(u_1 \equiv U_{j_1}+1) = P_r(\text{number} \in A_{j_1}+1 \mid \text{number} \in A)$$
$$= \frac{(r+1)/10^d}{N(r+1)/10^d} = \frac{1}{N}$$

Next : $u_2 \equiv U_{j_2}+1$ if the first next number in the column that belongs to A is one of the $(r+1)$ numbers which belong to $A_{j_2}+1$.

$P_r(u_2 \equiv U_{j_2}+1 \mid u_1 \equiv U_{j_1}+1) = P_r(u_2 \equiv U_{j_2}+1) = \frac{1}{N}$. Thus, the n units included in the sample are $U_{j_1}+1, U_{j_2}+1, \ldots, U_{j_n}+1$, where $j_1, j_2, \ldots, j_n$ are the first n numbers in the column which belong to A.

(2) The average time required to complete sampling in this case is

$$n + n \; \frac{10^d - N(r+1)}{N(r+1)} = n \, \frac{10^d}{N(r+1)} .$$

This is minimum for min. d and max. r and it conforms our way of choosing r above.

(ii) Simple Random Sampling Without Replacement

(1) As in (i),

$u_1 \equiv U_{j_1} + 1$ if the first number in the column which belongs to A is one of the $(r+1)$ numbers which belong to $A_{j_1} + 1$. Also $P_r(u_1 \equiv U_{j_1} + 1) = \frac{1}{N}$. $u_2 \equiv U_{j_2} + 1$ if the first next number which does not belong to $I \cup A_{j_1} + 1$ belongs to $A_{j_2} + 1$. $P(u_2 \equiv U_{j_2} + 1 \mid u_1 \equiv U_{j_1} + 1) =$

$$\frac{(r+1)(r+1)/N(r+1)(N-1)(r+1)}{(r+1)/N(r+1)} = \frac{1}{N-1} .$$

Next : $u_3 \equiv U_{j_3} + 1$ if the first next number which does not belong to $I \cup A_{j_1+1} \cup A_{j_2} + 1$ belongs to $A_{j_3} + 1$, and so on.

(2) The average time required to complete sampling in this case is

$$n + \sum_{k=0}^{n-1} \frac{10^d - (N-k)(r+1)}{(N-k)(r+1)} = \frac{10^d}{(r+1)} \sum_{k=0}^{n-1} \frac{1}{N-k}$$

$$\simeq \frac{10^d}{(r+1)} \log_e \{ N / (N-n+1) \} \text{ for large } N.$$

This completes the discussion of the case when the population is specified by its frame.

In fact, all that we have discussed above applies to the situations where the population is specified by the frequency distribution of the characteristic Y. At the end of § 2 we had indicated how a frame could be associated to a population specified by the frequency distribution of Y.

When the population is specified by the frequency distribution as in Table 2(a),

$$x_i = y_j \quad \text{if } u_i \equiv U_R \text{ with } \quad c^1_{j-1} < R \leq c^1_j .$$

In case the specification is as in Table 2(b), the characteristic value x_i of the unit included in the sample at the ith selection, corresponding to the case $u_i \equiv U_R$, with $c^1_{j-1} < R \leq c^1_j$, can be estimated in the following two ways :

(1) $x_i \simeq l_j + \dfrac{R - c^1_{j-1}}{f_j} \; (u_j - l_j)$, on the assumption that the frequency f_j is uniformly distributed over the class internal $(l_j - u_j)$.

(2) $x_i \simeq \frac{1}{2} \; (l_j + u_j)$, on the assumption that the frequency f_j is concentrated at the mid point of the class internal $(l_j - u_j)$.

7. <u>Sampling from Infinite Populations</u>

(i) <u>Characteristic Y is discrete</u>

<u>Case 1: k (no.of distinct characteristic values) is finite</u>.

Choose a positive integer N such that each of $N\,p(y_j)$ is integral. Let $N\,p(y_j) = f_j$, say, be the number of units possessing the characteristic value y_j in a population comprising of N units. Now proceed to draw a sample of size n from this population as in the corresponding situation considered in § 6.

Evidently, it will be best (in respect of the average time required for completing sampling) to choose min d but max. N with $10^{d-1} < N \leq 10^{d}$.

Case 2: k is infinite.

Since $p(y_j) \geqslant 0$ and $\sum_j p(y_j) = 1$, it will be possible, for all practical purposes to consider only a finite number of distinct value (or groups of values), e.g., $y_1, y_2, \ldots, y_{k-1}, \geqslant y_k$ in case of uni-model distribution with the infinite tail on the right hand side.

We can now proceed as in case 1 with a possible use of the methods outlined in § 6 estimating x_i.

While the method outlined above is general enough to cover all situations, special artifries are available in the case of standard distributions. Following are illustrations of these :

(a) Binomial Distribution

$$p(y) = \binom{n^*}{y} p^y q^{n^*-Y}, \quad p + q = 1$$
$$Y = 0(1)n^*$$

A random observation Y on this distribution would be given by

$$Y = X_1 + X_2 + \ldots + X_{n^*}$$

where X_i are independent random observations on the "true" binomial distribution.

Y	p(Y)
0	q
1	p

(b) Geometric Distribution

$$p(Y) = q^{Y-1}p, \quad p + q = 1$$
$$Y = 1(1)\infty$$

Here

$$Y = y \text{ if } \quad X_i = \begin{cases} 0 & i < y \\ 1 & i = y. \end{cases}$$

(c) Negative Binomial Distribution

$$p(Y) = \binom{Y-1}{r-1} p^r q^{Y-r} \qquad \begin{matrix} p+q = 1 \\ Y=r(1)\infty \end{matrix}$$

Here $Y = Y_1+Y_2+\ldots+Y_r$, where Y_i are independent random observations on the geometric distribution considered in (b) above.

(d) Poisson Distribution

$$p(Y) = e^{-\lambda} \lambda^Y / Y! \qquad Y = 0(1)\infty$$

Here $Y = y$ if $\sum_{i=1}^{y-0} X_i \leq 1$ and $\sum_{i=1}^{y+1} X_i = 1$,

where X_i are indeoendent random observations on the exponential distribution with parameter λ^{-1}, considered in (ii) below :

(e) Hypergeometric Distribution

$$p(Y) = \frac{\binom{n_1}{Y}\binom{n_2}{n^*-Y}}{\binom{n_1+n_2}{n^*}}, \qquad Y = \min(0, n^*-n_2)\,(1)\,\max(n_1, n^*)$$

Here

$$Y = Y_1 + Y_2 + \ldots + Y_{n^*}$$

where Y_i are independent random observations on the "true" binomial distributions.

Y_i	$p_i(Y_i)$	
0	q_i	$p_i + q_i = 1$
1	p_i,	$p_i = \dfrac{n_1-Y_1-Y_2-\ldots-Y_i-1}{(n_1+n_2)-(i-1)}$

(ii) Characteristic Y is continuous.

The three approaches that are available in this case are described in turn below :

(1) General Approach

Following result from the theory of probability distributions is fundamental to this approach :

Irrespective of the probability distribution $\{Y; f(Y)\}$ of the characteristic Y, the probability distribution of the signle-valued function

$$Z = F(Y) = \int_{-\infty}^{Y} f(t)\ dt$$

is uniform with

$$dP(Z) = dZ, \qquad 0 \leq Z \leq 1.$$

Also: in general, there is one to one correspondence between Y and Z*.

The problem of sampling from any continuous population is, therefore, reducible to that of sampling from uniform population with probability density function $f(Y) = 1$, $0 \leq Y \leq 1$. Method of sampling from this population is, therefore, described first.

Place a decimal point in front of each number in the d-digit column of random numbers. A number taken from the resulting column is a random selection from the fractions $0(10^{-d})\ 1-10^{-d}$. A random selection from the continuous interval (0,1) will thus be had by letting $d \rightarrow \infty$. The limiting case is, however, neither practicable nor necessar in general. A choice for the value of d can be made to suit the particul situation at hand, and governed by the degree of accuracy desired.

*This part of the result is true over the effective range of Y provided that f(Y) is a continuous function over this range. A special consideration will be required for the 'mixed' case.

In the general case, let R_i denote the ith sample from the uniform distribution. Then x_i would be given by the equation

$$R_i = \int_{-\infty}^{x_i} f(Y)\, dY.$$

Right hand side is a function of x_i; and success in solving the resulting equation in x_i will depend on the form of this function.

Following particular cases are of practical importance :

(a) Rectangular Population.

$$f(Y) = \frac{1}{b-a} \quad a \leq Y \leq b$$

In this case

$$R_i = \int_a^{x_i} \frac{1}{b-a}\, dY$$

$$= \frac{x_i - a}{b-a}$$

and, therefore,

$$x_i = a + (b-a) R_i$$

(b) Triangular Population I

$$f(Y) = \frac{2}{(b-a)^2} (b-Y) \quad a \leq Y \leq b$$

so that

$$R_i = \int_a^{x_i} \frac{2}{(b-a)^2} (b-Y)\, dY$$

$$= 1 - \frac{1}{(b-a)^2} (b-x_i)^2$$

and, therefore,

$$x_i = b - (b-a) \sqrt{(1 - R_i)}$$

(c) <u>(Double)-Triangular Population II</u>.

$$f(Y) = \begin{cases} m_1 \quad (Y-a) & a \leq Y \leq m \\ m_2 \quad (Y-b) & m \leq Y \leq b \end{cases}$$

with $\frac{1}{2}\, m_1\, (m-a)^2 - \frac{1}{2}\, m_2\, (b-m)^2 = 1$.

Here

$$R_i = \begin{cases} \frac{1}{2}\, m_1\, (x_i - a)^2 & R_i \leq R \\ \frac{1}{2}\, m_1\, (m - a)^2 + \frac{1}{2}\, m_2\, (x_i - b)^2 - \frac{1}{2}\, m_2\, (m-b)^2 & R_i \geq R, \end{cases}$$

where $R = \int_a^m f(Y)\, dY$

$$= \frac{1}{2}\, m_1\, (m - a)^2.$$

Hence

$$x_i = \begin{cases} a + \sqrt{2\, R_i/m_1} & R_i \leq R \\ b - \sqrt{2\, (1-R_i)/(-m_2)} & R_i \geq R \end{cases}$$

(d) <u>Exponential Population</u>

$$f(Y) = k\, e^{-k(Y - a)} \qquad k > 0, \quad Y \geq a$$

In this case

$$R_i = \int_a^{x_i} k\, e^{-k(Y - a)}\, dY$$

$$= 1 - e^{-k\, (x_i - a)}$$

Therefore,

$$x_i = a - \frac{1}{k} \log_e (1 - R_i)$$

(e) <u>Cauchy Population</u>.

$$f(Y) = \frac{1}{\pi} \quad \frac{1}{1 + (Y-a)^2}, \quad -\infty \leq Y \leq \infty$$

$$R_i = \int_{\infty}^{x_i} \frac{1}{\pi} \frac{1}{1+(Y-a)^2} dY$$

$$= \frac{1}{\pi} \left[A\ \mathrm{Tan}\ (x_i - a) - \frac{3\pi}{2} \right]$$

so that

$$x_i = a + \mathrm{Tan} \left[\pi (R_i + \tfrac{3}{2}) \right] .$$

In each of the populations considered above, the function $\int_{-\infty}^{x_i} f(Y)\ dY$ was good enough to provide a straight forward formula connecting x_i to R_i. None of the following cases affords such formulae and resort, to either numerical integration/or use of some numerical methods in individual cases or a reference to the standard tables of area under the probability curve will be necessary. There are alternative approximate methods available for such situations; and these are discussed later in this section. Here we present the general formulae.

(f) Normal Population.

$$f(Y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp \left[-\tfrac{1}{2} (Y-\mu)/\sigma \right]$$

Here $R_i = \int_{\infty}^{x_i} f(Y)\ dY$

$$= \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} \exp \left[-\tfrac{1}{2} Z^2 \right] dZ,$$

where $Z_i = (x_i - \mu)/\sigma$ is the standard normal deviate, and, therefore,

$$x_i = \mu + \sigma Z_i .$$

A reference to the table of area under the normal curve will give Z_i.

(g) The Gamma, Beta (Type I & II), Chi-square, F and Students' t Populations also belong to this category. However, there exist interrelationships between (among) these distribution; and the normal distribution; and fundamental to all these is the Gamma distribution.

For Gamme Population

$$f(Y) = \frac{1}{\Gamma(l)} \quad Y^{l-1} \quad e^{-Y} \qquad Y \geqslant 0 \quad l > 0$$

$$R_i = \int_0^{x_i} \frac{1}{\Gamma(l)} \quad Y^{l-1} \quad e^{-Y} \quad dY$$

The integral on the R.H.S. has been extensively tabulated. If l is a positive integer then Y is expressiable as the sum of l independent exponential variables with k = 1 and a = 0.

(h) <u>Bivariate Normal Population</u>

Bivariate populations arise when two characteristics Y_1, Y_2, say are associated with the various units of the population. In this case

$$f(Y_1,Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} * \left\{\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right\}\right], \quad \sigma_1, \sigma_2 > 0, -\infty \leq Y_1, Y_2 \leq \infty$$

Since the conditional distribution of Y_2, given Y_1 is normal with mean $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(Y_1-\mu_1)$, variance $\sigma_2^2(1-\rho^2)$,

the characteristic values x_{1i} and x_{2i} can be obtained from the relations

$$x_{1i} = \mu_1 + \sigma_1 \; Z_{1i}\,(R_{1i})$$

$$x_{2i} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}\;(x_{1i} - \mu_1)$$
$$+\, \sigma_2 \sqrt{1-\rho^2} \quad Z_{2i}\,(R_{2i})$$

Where Z_{1i} and Z_{2i} are the standard normal deviates corresponding to the random observations R_{1i}, R_{2i} on the uniform distribution.

An alternative approach, particularly useful for the multivariate case, would be to use ~~such~~ a transformation which replaces the variables by independent variables.

(2) <u>Discretization (Approximation I)</u>

For appropriately chosen non-overlapping intervals $(l_i - u_i)$, determine the grouped probability distribution

Y	Probability
l_1-u_1	p_1
l_2-u_2	p_2
$\vdots$	$\vdots$
l_k-u_k	p_k

$$\sum_{i=1}^{k} p_i = 1, \quad p_i = \int_{l_i}^{u_i} f(Y)\,dY.$$

Now proceed as in the corresponding case considered in § 6.

(3) Indirect Methods (Approximation II)

These methods seek to exploit the inherent characteristics of the sampling distributions one such example is the use of Central Limit Theorem in sampling from normal population, viz. If R_{ij} j= 1,2,...,K are independent random observations on the uniform distribution then

$$Z_i = \frac{1}{K} \sum_{j=1}^{K} (R_{ij} - \tfrac{1}{2}) \sqrt{\frac{1}{12K}}$$

is approximately normally distributed with mean 0 and variance 1. This fact is used to generate standard normal deviates to finally obtain

$$x_i = \mu + \sigma Z_i$$

In practice, it is sufficient to take K = 12 which in fact simplifies the formula for Z_i to

$$Z_i = \left(\sum_{j=1}^{12} R_{ij} - 6 \right).$$

BIBLIOGRAPHY

1. Feller, W. (1960) -
An Introduction to the Theory of Probability & its Applications
John Willey & Sons, Inc.

2. Naylar, T.H. Et.Al. (1966) -
Computer simulation techniques
John Willey & Sone, Inc.