

# *Summer Course (advanced) for Statisticians*

*May-June 1961*

**LECTURE NOTES No.—5**

---

## **SOME RECENT ADVANCES IN SAMPLING THEORY**

*by* M. N. Murthy, Indian Statistical Institute

AND

## **OPERATIONS REASERCH**

*by* C. S. Ramakrishnan, Indian Statistical Institute

RESEARCH AND TRAINING SCHOOL  
INDIAN STATISTICAL INSTITUTE  
203 BARRACKPORE TRUNK ROAD  
CALCUTTA-35

C4721  
15.4.64

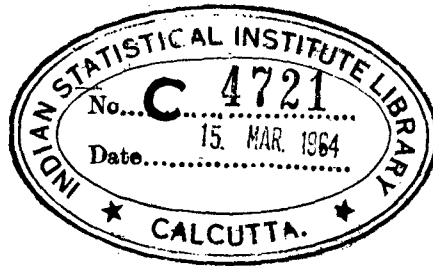
SOME RECENT ADVANCES IN SAMPLING THEORY

by

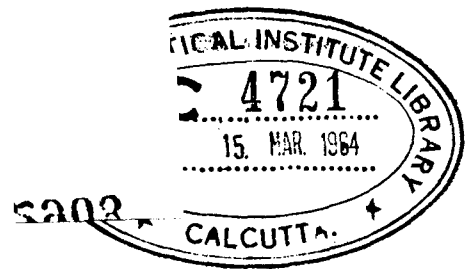
M. N. Murthy

C O N T E N T S

	<u>Page</u>
1. Introduction	1
2. Improving of Estimators	2
3. Varying Probability Sampling	4
4. Method of Ratio Estimation	9
5. Optimum Stratification	15
6. Non-Sampling Error	17
References	



Research and Training School  
Indian Statistical Institute  
203, Barrackpore Trunk Road,  
Calcutta-35.



E R R A T A

<u>Page</u>	<u>Line</u>	<u>Read</u>	<u>For</u>
3	18	(12)	(i j)
6	16	(N - i + 1)	(n - i + 1)
7	1	$\frac{y_1}{p_1} (1 - p_2)$	$\frac{y_1}{p_1} (1 - p_1)$
8	4	(3.13)	(3.11)
13	5	(4.15)	(4.14)
14	4	have	has
15	7	though .... have	through.... have
21	12	p	P
21	20	$\hat{V}(p)$	$\hat{V}(P)$
28	8	HARTLEY	HARLEY

# SOME RECENT ADVANCES IN SAMPLING THEORY

by

M.N.Murthy

## 1. INTRODUCTION

1.1. From times immemorial the concept of generalizing from a 'part' of the population to the 'whole' has been used more or less subjectively in daily life. But not until the later half of the nineteenth century, objective methods of generalizing from a part to the whole seem to have received much attention. In this case two questions arise - (i) how to select the 'part' from the 'whole' and (ii) how to generalize from the selected part to the whole. The problem is one of finding that combination of selection and estimation procedures which would minimize the risk involved in generalizing from the part to the whole per unit of cost. Alternatively, the problem may be viewed as that of finding that combination of selection and estimation procedures which would minimize the cost, ensuring at the same time a specified precision for the inference from a part to the whole.

1.2. The earlier developments in this field relate to the second question posed above and the result has been a fairly well developed theory of estimation and statistical inference based on the simplest of selection procedures, namely, equal probability sampling with replacement. Improvements in selection procedures may be considered to have been initiated by Bowley (1926) who used stratified simple random sampling with proportional allocation. Neyman (1934) considered the question of optimum allocation in stratified sampling.

1.3. During the decade 1940-50, considerable developments in sampling theory have taken place and the works of Mahalanobis (1940, 1944 and 1946), Cochran (1942), Hansen and Hurwitz (1943, 1946) and Madow and Madow (1944) need special mention. Cochran considered the question of utilizing supplementary information at the estimation stage by using ratio and regression estimators. Hansen and Hurwitz suggested the utilization of the supplementary information for

selecting the units with probability proportional to a suitable measure of size and they also considered the problem of non-response in surveys. Mahalanobis realized the importance of assessment of non-sampling errors as early as 1938 and developed the technique of interpenetrating sub-samples to assess the response errors in surveys. Madow and Madow developed the technique of sampling units systematically. Cochran (1953), Deming (1950), Hansen, Hurwitz and Madow (1953), Sukhatme (1953) and Yates (1953) have covered in detail the earlier developments in the theory of sampling in their respective books.

1.4. In these notes an attempt is made to <sup>present</sup> ~~give~~ some important advances in sampling theory during the last decade 1950-60. The recent developments are more in the nature of refinements of the earlier methods than in the nature of fundamental developments. Perceptible progress has been made in the fields of stratification, varying probability sampling, ratio estimation and assessment and control of non-sampling errors. It may be mentioned that only some of those developments in the theory of sampling which have been or are likely to be found useful in survey practice have been considered here and hence the coverage is not exhaustive.

## 2. IMPROVING OF ESTIMATORS

2.1. Three techniques of improving upon of certain types of estimators have been suggested. Goodman (1953) has shown that if  $t$  is an unbiased estimator of  $\theta$  and  $V(t) = K \theta^2$ , where  $K$  is known, then a more efficient estimator from the point of the risk function

$$\lambda(\theta)(t - \theta)^2, \lambda(\theta) > 0 \quad (2.1)$$

where  $\lambda(\theta)$  is a function of  $\theta$ , is given by

$$t' = \frac{1}{K+1} t. \quad (2.2)$$

An example of such a situation is provided by the estimator

$$s^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2 \quad (2.3)$$

of  $\sigma^2$  in case of a sample from a 'normal' population. Since

$$V(s^2) = \frac{2}{n-1} \sigma^4,$$

a better estimator of  $\sigma^2$  is given by

$$s'^2 = \frac{1}{n+1} \sum_i^n (y_i - \bar{y})^2. \quad (2.4)$$

2.2. In case of sampling without replacement, some estimators have been given by Das (1951) and Des Raj (1956), which depend on the order in which the units are selected. Murthy (1957) has shown that corresponding to any estimator based on the order in which the units are selected, there exists a more efficient estimator which does not take into account the order of selection of the units in the sample. The former may be termed '~~ordered~~ ordered estimator' and the latter 'unordered estimator'. The technique involved in improving the ordered estimator consists in taking the conditional expected value of the ordered estimator over all possible orders for a given unordered sample of units as the unordered estimator. For instance in case of selecting 2 units with varying probability without replacement, one of the ordered estimators given by Des Raj is

$$\hat{Y}_o = \frac{1}{2} \left[ \frac{y_1}{p_1} (1 + p_1) + \frac{y_2}{p_2} (1 - p_1) \right] \quad (2.5)$$

where the order of selection of the units is (ij). The corresponding unordered estimator which is more efficient than the above estimator is

$$\hat{Y}_u = \frac{1}{2-p_1-p_2} \left[ \frac{y_1}{p_1} (1-p_2) + \frac{y_2}{p_2} (1-p_1) \right]. \quad (2.6)$$

2.3. In case of sampling with replacement Basu (1958) has shown that corresponding to any estimator which takes account of the number of repetitions of the units in the sample, there exists a more efficient estimator which is based only on the distinct units in the sample without taking into consideration the number of repetitions. The procedure of improving the estimator is exactly

the same as that explained earlier and consists in taking the conditional expected value of the estimator over all possible repetitions of the given set of distinct units. For instance, in the case of simple random sampling with replacement the sample mean based only on the distinct units is more efficient than the mean of the units in the sample including their repetitions. Similarly in case of sampling  $n$  units with varying probabilities with replacement, if  $n-1$  units turn out to be distinct, then the improved estimator is given by

$$Y_d = \frac{1}{n} \left[ \frac{\bar{y}'}{\bar{p}'} + \left( \sum \frac{y_i'}{p_i'} \right) \right] \quad (2.7)$$

which is more efficient than the usual estimator  $\frac{1}{n} \left( \sum_{i=1}^n \frac{y_i}{p_i} \right)$ ;  $\bar{y}'$  and  $\bar{p}'$  being the sample means based on distinct units.

2.4. The improved estimators obtained by using the above techniques are in general difficult to calculate except for certain particular cases and the expressions for their variance estimators are rather complicated.

2.5. Attempts have been made to give generalised estimators in sampling from finite populations [Midzumo, (1950); Godambe, (1955); Nanjamma, Murthy and Sethi, (1959)], Godambe, (1955) and Roy and Chakravorthy, (1960)) have shown that there does not exist a minimum variance unbiased estimator in the non-parametric sense in sampling from a finite population. The above techniques only help in improving upon certain class of estimators and not in obtaining minimum variance estimators.

### 3. VARYING PROBABILITY SAMPLING

3.1. Hansen and Hurwitz (1943) suggested selection of units with probability proportional to a given measure of size, as this provides efficient estimators if the measure of size is related to the characteristic under consideration. One procedure of selecting one unit with probability proportional to size (pps) is to cumulate the sizes of units ( $C_i = C_{i-1} + X_i$ ;  $i = 1, 2, \dots, N$ )

and then to select the  $i$ th unit if  $C_{i-1} < R \leq C_i$  where  $R$  is a number chosen at random between 1 and  $C_N$ . This method becomes time consuming if the population size is large.

3.2. Lahiri (1951) suggested a simple method of selection of a unit with pps which avoids cumulation of the sizes. His method consists of

- (i) selection of two numbers at random, one from 1 to  $N$  (say  $i$ ) and the other from 1 to  $M$  (say  $R$ ),  $M$  being the maximum of the sizes of the units;
- (ii) selection of the  $i$ th unit if  $R \leq X_i$ , the measure of size of that unit
- (iii) rejection of the  $i$ th unit and repetition of the above operations if  $R > X_i$ .

It may be easily shown that this procedure leads to the required probabilities of selection of the units. But there is possibility of rejection of certain draws and the probability of rejecting a draw is  $1 - \bar{X}/M$ , where  $\bar{X}$  is the population mean of the sizes. The probability of rejection can however be reduced by considering the units with very large sizes as made up of two or more split units.

3.3. Suppose  $C_1$  is the cost per unit for cumulation,  $C_2$  is the cost per unit for writing down the cumulative totals and  $C_3$  is cost of drawing a random number. Then the cost involved in selecting  $n$  units with pps with replacement will be

$$C' = N(C_1 + C_2) + nC_3 \quad (3.1)$$

for the cumulative total method and

$$C'' = NC_1 + 2n \frac{M}{\bar{X}} C_3 \quad (3.2)$$

for Lahiri's method, since even in the case of the latter it is necessary to get the total of the sizes for estimation purposes. Hence Lahiri's method is to be preferred to cumulative total method from cost point of view if

$$\frac{M}{\bar{X}} < \frac{1}{2} \left( 1 + \frac{N}{n} \frac{C_2}{C_3} \right) \quad (3.3)$$



If, however, the total of the sizes is already available, then Lahiri's method is to be chosen if

$$\frac{M}{X} < \frac{1}{2} \left( 1 + \frac{N}{n} \frac{C_1 + C_2}{C_3} \right). \quad (3.4)$$

3.4. In sampling  $n$  units with pps with replacement, an unbiased estimator of the population total is given by

$$\hat{Y} = \frac{1}{n} \sum_i^n \frac{y_i}{p_i} \quad (3.5)$$

and its variance estimator is given by

$$\hat{V}(\hat{Y}) = \frac{1}{n(n-1)} \sum_i^n \left( \frac{y_i}{p_i} - \hat{Y} \right)^2. \quad (3.6)$$

The above estimator which takes into account the repetitions of the units in the sample can be improved by using the technique mentioned in section 2.

3.5. Das (1951) considered the question of varying probability sampling without replacement. In case of sampling  $n$  units with pps without replacement, he suggested the ordered estimator

$$\bar{d} = \frac{1}{n} \sum_i^n d_i \quad (3.7)$$

where

$$d_i = \frac{(1-p_1)(1-p_1-p_2)\dots(1-p_1-p_2-\dots-p_{i-1})}{(N-1)(N-2)\dots(n-i+1)} \frac{y_i}{p_1 p_2 \dots p_{i-1} p_i}$$

and an unbiased estimator of the variance of this estimator is given by

$$\hat{V}(\bar{d}) = \bar{d}^2 - \left[ \frac{1}{n} \sum_{j=1}^n d_j y_j + \frac{N-1}{\binom{n}{2}} \sum_j \sum_{j' > j} d_j y_{j'} \right]. \quad (3.8)$$

The estimator given in (3.7) can be improved upon by using the technique explained in section 2. For instance in case of  $n = 2$ , the improved unordered estimator which corresponds to this ordered estimator in case of sampling with pps without replacement is given by

$$d_u = \frac{1}{2} \left[ \frac{\frac{y_1}{p_1} (1-p_1) + \frac{y_2}{p_2} (1-p_1)}{(2-p_1-p_2)} + \frac{1}{N-1} \frac{y_1 + y_2}{p(12)} \right] \quad (3.9)$$

where  $P(12) = \frac{p_1 p_2 (2-p_1-p_2)}{(1-p_1)(1-p_2)}$ .

3.6. In case of sampling with pps without replacement Horvitz and Thompson (1952) suggested an estimator which does not take into account the order of selection of the units. The estimator suggested by them is

$$\hat{Y}_{HT} = \sum_i^n \frac{y_i}{\pi_i} \quad (3.10)$$

where  $\pi_i$  is the probability of inclusion of the  $i$ th unit in the sample. An unbiased variance estimator in this case is

$$\hat{V}_{HT}(\hat{Y}_{HT}) = \sum_{i=1}^n (1-\pi_i) \left( \frac{y_i}{\pi_i} \right)^2 + 2 \sum_{i=1}^n \sum_{j>i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \quad (3.11)$$

Since this variance estimator takes negative values often, Yates and Grundy (1953) have suggested the following variance estimator which takes negative values less often than (3.11).

$$\hat{V}_{YG}(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (3.12)$$

3.7. Des Raj (1956) considered a set of ordered estimators in sampling with pps without replacement and one such estimator is

$$\bar{t} = \frac{1}{n} \sum_i^n t_i \quad (3.13)$$

where

$$t_i = y_1 + y_2 + \dots + y_{i-1} + \frac{y_i}{p_i} (1-p_1-p_2-\dots-p_{i-1}).$$

The estimators  $(t_i, i = 1, 2, \dots, n)$  are uncorrelated and hence an unbiased variance estimator is given by

$$\hat{V}(\bar{t}) = \frac{1}{n(n-1)} \sum_{i=1}^n (t_i - \bar{t})^2 \quad (3.14)$$

This estimator can also be improved using the technique mentioned in section 2. The unordered estimator corresponding to the ordered estimator in this case for  $n = 2$  has been given in (2.6). In the general case, the unordered estimator corresponding to the ordered estimator given in (3.11) can be written as

$$Y_u = \frac{\sum_{i=1}^n y_i P(s|i)}{P(s)} \quad (3.15)$$

where  $P(s)$  is the probability of getting the unordered sample of  $n$  units and  $P(s|i)$  is the conditional probability of getting the unordered sample 's' given that the  $i$ th unit in the sample has occurred first.

3.8. Hanumantha Rao (1960) has given selection procedures which give rise to a given set of probabilities of inclusion of the units in the sample. Murthy (1960) has pointed out that a given set of probabilities of inclusion of units ( $\pi_i$ 's) can easily be achieved by selecting the units with probability proportional to  $\pi_i$ 's systematically. The procedure consists in first obtaining the cumulative totals of the  $\pi_i$ 's ( $C_i = C_{i-1} + \pi_i ; i = 1, 2, \dots, N$ ) and then selecting the units systematically with a random start from 0 to 1 and with 1 as the sampling interval. Considerable gain in efficiency can be achieved by arranging the units in some suitable order before selection. Hartley and Rao (1959) have considered this procedure of sampling when the units are assumed to be in random order.

3.9. The efficiencies of the different estimators considered in this section have been studied for a sample of 2 units selected from a small population of 4 units given below.

Unit	$U_1$	$U_2$	$U_3$	$U_4$
y	0.5	1.2	2.1	3.5
x	1	2	3	4
p	0.1	0.2	0.3	0.4

Table 1: Showing the efficiencies of the different pps selection and estimation procedures in case of the above mentioned population for a sample of size 2.

sl. no.	selection procedure	estimator	variance	efficiency (%)
1	pps with replacement	$\frac{1}{2}[(y_1/p_1)+(y_2/p_2)]$	0.50	60
2	pps without replacement	$Y_{HT}$	0.82	37
3	"	$d_u$	0.36	83
4	"	$Y_u$	0.31	97
5	pps systematic	$\frac{1}{2}[(y_1/p_1)+(y_2/p_2)]$	0.30	100

#### 4. METHOD OF RATIO ESTIMATION

4.1. As the relationship between two characteristics is of much interest, the question of estimation of ratios of certain population parameters has become quite important in many surveys. The method of ratio estimation is also being used to estimate population totals, since a ratio estimator is more efficient than the conventional unbiased estimator under certain circumstances not uncommon in actual practice. The usual method of estimating a population ratio has been to take the ratio of unbiased estimators of the numerator and denominator. If  $y$  and  $x$  are unbiased estimators of  $Y$  and  $X$  respectively, then an estimator of  $R(= Y / X)$  is given by

$$\hat{R} = y/x \quad (4.1)$$

and of  $Y$  is given by

$$\hat{Y}_r = \frac{y}{x} X \quad (4.2)$$

where  $x$  is a suitable supplementary variate and the value of  $X$  is known.

4.2. The above estimators are biased. The bias and mean square error of the estimator  $\hat{R}$  correct to the second degree of approximation are given by

$$B(y/x) = R(C_x^2 - C_{xy}) \quad (4.3)$$

$$M(y/x) = R^2(C_y^2 - 2RC_{xy} + C_x^2) \quad (4.4)$$

where  $C_x$  and  $C_y$  are the coefficients of variation of  $x$  and  $y$  respectively and  $\rho$  is the correlation coefficient between  $x$  and  $y$ . Expressions for the bias and mean square error of  $Y_r$  can be obtained by multiplying (4.3) by  $X$  and (4.4) by  $X^2$  respectively. The above expressions have been obtained under the assumptions (i)  $|\frac{x-X}{X}| < 1$  and (ii) relative moments of degree greater than 2 in  $x$  and  $y$  can be neglected. These assumptions are likely to be valid if the sample size is sufficiently large. In recent years attempts have been made to <sup>evolve</sup> ~~give~~ selection and estimation procedures which provide unbiased ratio estimators.

4.3. It is of interest to note that the relative bias in  $\hat{R}$  can be written as

$$|\frac{B(\hat{R})}{\hat{R}}| = \rho_{(\hat{R},x)} C_{\hat{R}} C_x \quad (4.5)$$

where  $C_{\hat{R}}$  and  $C_x$  are the coefficients of variation of the estimators  $\hat{R}$  and  $x$  respectively and  $\rho_{(\hat{R},x)}$  is the correlation coefficient between  $\hat{R}$  and  $x$ . This shows that the relative bias of the ratio estimator would be small if the sample size is large since in that case the coefficients of variation of  $\hat{R}$  and  $x$  are likely to be small. Further  $\rho_{(\hat{R},x)}$  is likely to be nearly zero if  $y$  and  $x$  are approximately proportional. For instance if  $C_{\hat{R}} = 0.01$ ,  $C_x = 0.05$  and  $\rho_{(\hat{R},x)} = 0.05$ , then the relative bias is 0.000025 which is negligible.

4.4. In case of ratio method of estimation for estimating a population total using the data on a suitable supplementary variate, Hartley and Ross(1954) have obtained a ratio-type of estimator which is unbiased in case of simple random sampling. Suppose  $(y_i, x_i)$  are the values of the variates  $y$  and  $x$  for the  $i$ th unit in the sample selected with equal probability without replacement ( $i = 1, 2, \dots, n$ ). The bias of the estimator

$$\hat{Y} = \bar{r} X, \quad (\bar{r} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}) \quad (4.6)$$

can be expressed as

$$B(\hat{Y}) = -N \text{cov}(\frac{y}{x}, x). \quad (4.7)$$

Since  $\text{cov} \left( \frac{Y}{X}, x \right)$  can be unbiasedly estimated, the estimator given in (4.6) can be corrected for its bias and the corrected estimator is given by

$$Y' = \bar{r} X + \frac{N-1}{n-1} n (\bar{y} - \bar{r} \bar{x}). \quad (4.8)$$

If the sampling is with replacement the factor  $(N-1)$  is to be replaced by  $N$ .

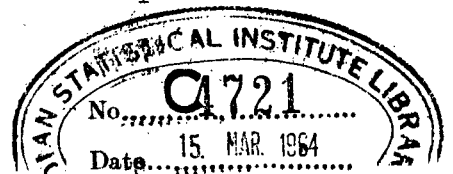
4.5. Goodman and Hartley (1958) have shown that for large samples, this unbiased ratio-type estimator is more efficient than the usual combined ratio estimator  $\left( \frac{\bar{y}}{\bar{x}} X \right)$  if the slope of the population regression line of  $y$  on  $x$  is closer to  $\frac{1}{N} \sum_{i=1}^N \frac{Y_i}{X_i}$  than to  $\frac{Y}{X}$ . However, this condition is not in general satisfied because one would ordinarily use a ratio estimator only when the regression coefficient is expected to be near  $Y/X$ . Hence the above condition is rather restrictive and the proposed unbiased ratio estimator may be less efficient than the usual combined ratio estimator in large samples.

4.6. It may be mentioned that this technique of getting an unbiased ratio estimator is applicable only if the value of the denominator of the population ratio is known, which is the case when one is using the ratio method of estimation for estimating the population mean or total using a suitable supplementary variate. But in case of estimating a population ratio where usually the value of the denominator is not known, it is not possible to use this technique.

4.7. The above method can easily be generalized to the case where  $y_i$  and  $x_i$  ( $i = 1, 2, \dots, m$ ) are unbiased estimators of the population totals  $Y$  and  $X$  based on  $m$  interpenetrating sub-samples of the same size selected according to any specified sampling design. In this case an unbiased ratio-type estimator of  $Y$  is given by

$$Y' = \bar{r} X + \frac{m}{m-1} (\bar{y} - \bar{r} \bar{x}), \quad (4.9)$$

where  $\bar{r} = \frac{1}{m} \sum_{i=1}^m \frac{y_i}{x_i}$  and  $\bar{y}$  and  $\bar{x}$  are means of the sub-sample estimates of  $Y$  and  $X$  respectively.



4.8. Murthy and Nanjamma (1959) have developed a technique of estimating the bias of an ordinary ratio estimator to any specified degree of approximation based on the interpenetrating sub-sample estimates. This estimator of bias can be used to correct the biased ratio estimator for its bias thereby obtaining an almost unbiased ratio estimator. Suppose  $(y_i, x_i)$  are unbiased estimates of  $Y$  and  $X$  based on the  $i$ th interpenetrating sub-sample ( $i=1,2,\dots,m$ ) If  $B_m$  is the bias of the estimator

$$R_m = \frac{1}{m} \sum_{i=1}^m \frac{y_i}{x_i} \quad (4.10)$$

and  $B_1$  is the bias of the estimator

$$R_1 = \frac{\bar{y}}{\bar{x}}, \quad (4.11)$$

then

$$B_m = m B_1 \quad (4.12)$$

correct to the second degree of approximation. Hence an estimator of the bias  $B_1$  which is unbiased to the second degree approximation is given by

$$\hat{B}_1 = \frac{R_m - R_1}{m - 1}. \quad (4.13)$$

This estimator of the bias can be used to correct  $R_1$  for its bias and we get an almost unbiased ratio estimator

$$R_c = R_1 - \hat{B}_1 = \frac{mR_1 - R_m}{m-1}. \quad (4.14)$$

It has been shown that under conditions not uncommon in practice, the estimator  $R_c$  is more efficient than  $R_1$ . It may be mentioned <sup>that</sup> this technique unlike that of Hartley and Ross can be applied to both the situation of estimating a population ratio and of estimating the population total using the ratio method of estimation.

4.9. Pascual (1961) has suggested the correction of the bias of the estimator  $R_1$  using result (4.12) and an unbiased estimator of  $B_m$  suggested by Hartley and Ross. This gives rise to the estimator

$$Y'' = \frac{\bar{y}}{\bar{x}} X - \frac{N-1}{n-1} (\bar{y} - \bar{r} \bar{x}) \quad (4.15)$$

in case of simple random sampling without replacement. This estimator is only almost unbiased and not completely unbiased as in the case of Hartley and Ross estimator, but this estimator is likely to be more efficient than  $Y'$  in large samples. In case of with replacement sampling,  $(N-1)$  in (4.14) is to be replaced by  $N$ . This result can also be utilized when estimates based on interpenetrating sub-samples of the same size are available.

4.10. If  $y_s$  and  $x_s$  are unbiased estimators of  $Y$  and  $X$  based on the  $s$ th sample selected according to any given sample design, the ratio estimator

$$R_s = \frac{y_s}{x_s} \quad (4.16)$$

will be unbiased for the ratio  $R (= Y/X)$  if the sample design is changed such that  $P'_s$ , the probability of selecting the  $s$ th sample, is proportional to  $x_s P_s$  where  $P_s$  is the probability of selecting the  $s$ th sample according to the original sample design, that is, if

$$P'_s = \frac{x_s P_s}{X} \quad (4.17)$$

for,

$$E(R) = \sum_s \frac{y_s}{x_s} P'_s = \frac{\sum_s y_s P_s}{X} = R.$$

4.11. Murthy, Nanjamma and Sethi (1959) have given simple modifications of many of the selection procedures commonly adopted in practice, namely, equal probability sampling, varying probability sampling, stratified sampling and multi-stage sampling, which, while retaining the form of the usual ratio estimators, make them unbiased. For many of the situations commonly met with in practice, this modification of a given sampling scheme consists essentially in first selecting one ultimate sampling unit with probability proportional to its value of the characteristic occurring in the denominator of the ratio and then selecting the remaining sample according to the original scheme of



sampling. The authors have also considered the variance and the variance estimators of the unbiased ratio estimator for the modified sampling schemes.

4.12. For instance in case of simple random sampling, Midzuno (1952) and Sen (1952) has suggested selecting one unit with pps and then the remaining units in the sample with equal probability without replacement. This makes the ratio estimator  $(\bar{y}/\bar{x})X$  unbiased for Y. Similarly in case of stratified simple random sampling without replacement, the modification in the selection procedure for getting an unbiased ratio estimator consists of selecting one unit (say the jth unit in the ith stratum) with pps from the whole population,  $(n_i-1)$  units from the remaining  $N_i-1$  units in the ith stratum and  $n_j$  units from  $N_j$  units of the jth stratum ( $j \neq i$ ) with equal probability without replacement. For this modified selection procedure the estimator

$$\hat{Y} = \frac{\sum_{i=1}^k N_i \bar{y}_i}{\sum_{i=1}^k N_i \bar{x}_i} X \quad (4.13)$$

is unbiased for the population total Y.

4.13. The efficiencies of different selection and estimation procedures considered here have been studied for sampling 2 units from the small population of 4 units given in section 3. The results of this study are presented in Table 2.

Table 2. Showing the efficiencies of different selection and estimation procedures.

sl. no.	selection procedure	estimator	bias	mean sq. error	efficiency
1	pps and srs of the remaining	$(\bar{y}/\bar{x}) X$	-	0.34	88
2	equal probability without replacement	$\bar{Y}$	-	5.44	6
3	"	$(\bar{y}/\bar{x}) X$	-0.15	0.39	77
4	"	Y'	-	0.60	50
5	"	Y''	0.10	0.55	55
6	pps systematic	$\frac{1}{2} \sum \frac{y_i}{x_i}$	-	0.30	100

## 5. OPTIMUM STRATIFICATION

5.1. The problems involved in the use of stratified sampling are (i) choice of a stratification variable, (ii) choice of the number of strata, (iii) determination of the points of demarcation of strata and (iv) allocation of total sample size to the strata. The earlier objective developments in this field were mainly confined to the consideration of problem (iv) mentioned above, through considerable attention have also been paid to the problems (i) and (ii) Dalenius (1950, 1953, 1957) studied in detail problems (ii) and (iii) and evolved techniques of determining optimum points of demarcation of strata.

5.2. Suppose the population under consideration can be represented by the frequency function  $f(x)$ ,  $a \leq x \leq b$  and it is divided into  $L$  strata by the points  $x_i$  ( $i = 1, 2, \dots, L-1$ ). Let  $W_i$  and  $\mu_i$  be the proportion of units and the mean value for the  $i$ th stratum,

$$W_i = \int_{x_{i-1}}^{x_i} f(x) dx \quad (5.1)$$

$$\mu_i = \int_{x_{i-1}}^{x_i} xf(x) dx. \quad (5.2)$$

If  $n_i$  units are selected from the  $i$ th stratum with equal probability, then an unbiased estimator of  $\mu$  is

$$\bar{x} = \sum_{i=1}^L W_i \bar{x}_i, \quad (5.3)$$

and its variance is

$$V(\bar{x}) = \sum_{i=1}^L W_i^2 \frac{\sigma_i^2}{n_i} \quad (5.4)$$

5.3. If the units are allocated / to the strata in proportion to  $W_i$ 's, then the variance becomes

$$V_p(\bar{x}) = \frac{1}{n} \sum_{i=1}^L W_i \sigma_i^2 \quad (5.5)$$

The optimum points of demarcation of strata in case of proportional allocation can be obtained by minimizing (5.5) with respect to  $x_i$  ( $i = 1, 2, \dots, L-1$ )

and they are given by

$$x_i = \frac{\mu_i + \mu_{i+1}}{2}, \quad (i = 1, 2, \dots, L-1) . \quad (5.6)$$

It may be noted that the above condition is necessary but not sufficient. In case of multi-modal functions there may be more than one set of optimum demarcation points and that set which gives rise to 'minimum minimorum' is to be selected.

5.4. In case of optimum allocation, the variance of the estimator (5.3) is given by,

$$V(\bar{x}) = \frac{1}{n} \left( \sum_{i=1}^L W_i \sigma_i \right)^2. \quad (5.7)$$

The optimum points of stratification in case of optimum allocation can be obtained by minimizing (5.7) with respect to  $x_i$  ( $i = 1, 2, \dots, L-1$ ). The optimum points of stratification are given by

$$\frac{\sigma_i^2 + (x_i - \mu_i)^2}{\sigma_i} = \frac{\sigma_{i+1}^2 + (x_{i+1} - \mu_{i+1})^2}{\sigma_{i+1}} \quad (5.8)$$

( $i = 1, 2, \dots, L-1$ ).

5.5. Since the solution of (5.8) is in general difficult, attempts have been made to evolve practical procedures which would achieve approximately optimum stratification. Dalenius and Hodges (1957) have shown that equalization of  $W_i \sigma_i$  in forming strata leads to optimum stratification which result had previously been conjectured by Dalenius and Gurney (1951).

Mahalanobis (1952) and Hansen, Hurwitz and Meadow (1953) suggested the equalization of  $W_i \mu_i$  in forming strata. It may be mentioned that the later procedure will be approximately optimum if the coefficients of variation in the different strata are roughly equal.

5.6. Dalenius has shown that for the population  $f(x) = e^{-x}$  the design where the stratum having larger units is completely enumerated is less efficient than stratified optimum allocation design if  $n/N$  is small. In both

is  
the cases, it is assumed that the stratification is done in an optimum fashion.  
It is surmised that this result is likely to be of general applicability.

5.7. Dalenius (1953) has studied the problem of determining the optimum number of strata and has conjectured the following relationship

$$V(\bar{x}, L) = \left(\frac{L-1}{L}\right)^2 V(\bar{x}, L-1). \quad (5.9)$$

Since  $V(\bar{x})$  and cost function depend on the sample size  $n$  and the number of strata  $L$ , the optimum values of  $n$  and  $L$  which minimize the variance for a given cost can be obtained empirically.

## 6. NON-SAMPLING ERRORS

6.1. The question of assessment and control of non-sampling errors has been receiving considerable attention and suitable techniques are being developed for this purpose. Mahalanobis (1940, 1944, 1946), Mahalanobis and Lahiri (1960) and Lahiri (1957) have suggested many important techniques for assessing and controlling errors in censuses and surveys. Hansen and others (1946, 1951, 1960) and Sukhatme and Seth (1952) have considered the question of non-sampling errors and have developed a suitable mathematical model. Post-enumeration checks and re-interview surveys are being made part of some of the nation-wide censuses and surveys so as to enable assessment of the non-sampling errors.

6.2. The broad sources of non-sampling errors, which are present in both in complete enumerations and sample surveys, though possibly to varying degrees, are incomplete coverage of the population or sample (including non-response) faulty definitions, defective methods of data collection and tabulation errors. In case of sample surveys the errors may also arise from defective sampling frame and selection procedures. More specifically, the non-sampling errors may arise due to omission or duplication of units, inaccurate and inappropriate methods of measurement, inappropriate arrangement or wording of questions, inadequate and ambiguous instructions, non-response, deliberate

or unconscious mis-reporting of data by respondents, carelessness on the part of investigators and computers, lack of proper supervision and defective methods of scrutiny and tabulation of data.

6.3. The conceptual set-up involved in developing a theory of <sup>sampling</sup> errors in sample surveys can be explained as follows. Suppose a sample has been chosen to be canvassed under reasonable conditions of survey and there are two populations, one of investigators and the other of computers qualified for doing the field and the processing work of the survey. If we repeatedly carry out this survey on the selected units with different samples of investigators and computers chosen with some suitable sampling designs, we may get different results because of the various possible sources of error present in <sup>the data</sup> under the usual operational conditions. Here there are three stages of randomization: selection of units, investigators and computers. The difference between the expected value of the estimator taken over all the three stages of randomization and the true value may be termed as 'total bias'. This consists of both 'sampling bias' and 'non-sampling bias'. The variance of the estimator taken over all the three stages is a measure of the divergence of the estimator from its expected value and is composed of sampling variance, variance between investigators and variance between computers and some interactions between the three sources of error. For instance the data collected by one investigator may be affected by his misunderstanding of the instructions, his pre-conceived notions about the survey, the earlier units canvassed by him etc. Thus we see that the total error consists of sampling <sup>and</sup> non-sampling biases, sampling and non-sampling variance and some interactions between the sample and the sources of non-sampling errors.

6.4. To fix the ideas let us consider the case where a simple random sample of  $n$  units drawn with replacement from a population of  $N$  units is divided at random into  $k$  equal sub-samples of  $m$  units each and these sub-samples are surveyed by  $k$  investigators selected with equal probability from a large population of  $K$  investigators qualified for this work. Let  $Y_{i,j}$  and  $Y_j$

be the value reported for the  $i$ th unit by the  $j$ th investigator and its true value respectively. Suppose  $y_{ij}$  is the value reported for the  $i$ th unit in the sample by the  $j$ th selected investigator. Here it is assumed that the response for a unit is not affected by the responses of the other units in the sample. An estimator of the population mean  $\bar{Y}$  is given by

$$\bar{y} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^m y_{ij}, \quad (n = km). \quad (6.1)$$

6.5. The expected value of the estimator (6.1) taken over the two stages of randomization is

$$E(\bar{y}) = \frac{1}{N} \sum_{i=1}^N Y'_i \quad (Y'_i = \frac{1}{K} \sum_{j=1}^K Y_{ij}), \quad (6.2)$$

and the total bias, which in this case consists wholly of response bias, is

$$B(\bar{y}) = \frac{1}{N} \sum_{i=1}^N (Y'_i - Y_i). \quad (6.3)$$

The variance of this estimator taken over the two stages of randomization can be shown to be

$$V(\bar{y}) = \frac{\sigma_s^2}{n} + \frac{\sigma_d^2}{n} [1 + (m-1)\rho] \quad (6.4)$$

where

$$\sigma_s^2 = \frac{1}{N} \sum_{i=1}^N (Y'_i - \bar{Y})^2, \quad \bar{Y}' = \frac{1}{N} \sum_{i=1}^N Y'_i \quad (6.5)$$

is the sampling variance,

$$\sigma_d^2 = \frac{1}{KN} \sum_{i=1}^N \sum_{j=1}^K (Y_{ij} - Y_{i.})^2 \quad (6.6)$$

is termed 'simple' or 'uncorrelated' response variance, and

$$\rho \sigma_d^2 = \frac{1}{KN(N-1)} \sum_{j=1}^k \sum_{i=1}^N \sum_{i' \neq i} (Y_{ij} - Y_{i.})(Y_{i'j} - Y_{i'.}) \quad (6.7)$$

where  $\rho$  is the intra-class correlation among the response deviations in a sample canvassed by one investigator.

6.6. The result in (6.4) shows the contribution to the total variance from response variation and it also brings out the impact of the intra-class correlation among the responses in a sample surveyed by one investigator (intra-investigator correlation) on the response variance. The intra class correlation will be positive if the response deviations for the different units have a consistent tendency to be in one direction for one investigator and in another direction for another investigator. Even when this correlation is small, the relative contribution to the response variation may be considerable if  $m$ , the number of units surveyed by each investigator, is large. For instance if  $\rho = 0.01$  and  $m = 1000$ , then the response variation becomes about 10 times more than that in case of  $\rho = 0$ .

6.7. An unbiased estimator of the variance of  $\bar{y}$  given in (6.4) is given by

$$\hat{V}(\bar{y}) = \frac{1}{k(k-1)} \sum_{j=1}^k (\bar{y}_{.j} - \bar{y})^2, \quad (\bar{y}_{.j} = \frac{1}{m} \sum_{i=1}^m y_{ij}) \quad (6.8)$$

6.8. Suppose the cost function is

$$c = k c_1 + n c_2 \quad (6.9)$$

where  $c_1$  is the cost of recruiting and training one investigator,  $c_2$  is the cost of surveying one unit and  $n = mk$ . The total variance given in (6.4) can be written as

$$V(\bar{y}) = \frac{\sigma^2 - \sigma_r^2}{n} + \frac{\sigma_r^2}{k} \quad (6.10)$$

where  $\sigma_r^2 = \rho \sigma_d^2$  and  $\sigma^2 = \sigma_s^2 + \sigma_d^2$ . Minimizing the variance in (6.10) with respect to  $n$  and  $k$  subject to the cost restriction (6.9), we get the optimum values of  $k$  and  $m$  as,

$$k = \frac{c}{\sqrt{c_1 \sigma_r^2} + \sqrt{c_2 (\sigma^2 - \sigma_r^2)}} \sqrt{\frac{\sigma_r^2}{c_1}} \quad (6.11)$$

and

$$m = \sqrt{\frac{c_1}{c_2}} \sqrt{\frac{\sigma^2 - \sigma_r^2}{\sigma_r^2}} \quad (6.12)$$

6.9. A number of empirical studies have been conducted in recent years to assess the magnitude of the intra-investigator correlation coefficient for different types of characteristics (Kish and Slater, 1960). It is found that for factual items, the intra-investigator correlation coefficient  $\rho$  ranges from 0 to 0.02 whereas for subjective items it is about 0.04 to 0.08 and for morbidity items it is as high as 0.11 to 0.15.

6.10. In case of estimating a population proportion from a sample of  $n$  units selected with equal probability with replacement and surveyed by a sample of  $k$  investigators at the rate of  $m$  units each ( $n = km$ ), the total variance of the sample proportion  $P$  becomes

$$V(p) = \frac{P'Q'}{n}, \quad (P' = E(p) \text{ and } Q' = 1 - P') \quad (6.13)$$

if the intra-investigator correlation coefficient is assumed to be zero. This result is interesting because it shows that the expression which is normally used as the sampling variance of a sample proportion includes not only the sampling variance but also the uncorrelated response variance (Hansen, Hurwitz and Bershada, 1960). An unbiased estimator of the variance (6.13) is given by

$$\hat{V}(P) = \frac{p q}{n-1} \quad (q = 1 - p) \quad (6.14)$$

6.11. Originally Mahalanobis made use of the technique of interpenetrating sub-samples in crop surveys to find out the differential investigator bias. For this purpose, linked pairs of grids (square parcel of land) were located at random on the maps in the form of dumb-bell shaped figures, one end of each figure representing the grid belonging to sub-sample 1 and the other end representing the grid belonging to sub-sample 2. One sub-sample was investigated by one set of investigators and the other sub-sample by an entirely different set of investigators independently.



Under certain assumptions the student's t-test may be applied to the difference between the estimates based on the two-sub-samples to test the hypothesis that there is no differential investigator bias at any given level of significance. If the difference turns out to be statistically significant, it means that the direction and magnitude of investigator bias are not of the same order for all the investigators. It may be noted that if the difference turns out to be statistically insignificant, it does not mean that the investigator bias is zero. For this result may be due to the fact that the biases are all of the same order and in the same direction.

6.12. The above method can well be applied to bring out the differential effect of different tabulation procedures, methods of data collection etc., and to bring out the variation over time. Suppose one is interested in finding out whether intensive training of the investigators for a given survey is essential or not. For this purpose, one sub-sample may be assigned to intensively trained investigators and the other sub-sample to investigators who have got only superficial training. If the difference in the results obtained from these two sub-samples turns out to be significant, there is strong case for adopting the method of intensive training in future surveys of a similar nature. On the other hand if the difference was not significant, it would mean that for this type of survey intensive training is perhaps not essential.

6.13. The technique of interpenetrating sub-samples may be used as a check on the different operations involved in large scale sample survey. Suppose one wishes to have a check on the calculations at the time of estimation. For this purpose, the sample may be divided into  $k$  suitably linked samples assigned to  $k$  different groups of computers at random and the estimates may be obtained from each of these sub-samples independently. If there is good agreement between these estimates, for all practical purposes it may be assumed with certain amount of confidence that the calculations have been done correctly. If one of these estimates differs from the others (assuming  $k$  is more than 2) and if there is good agreement between the remaining  $k-1$  estimates, one naturally

suspects the calculations done on that sub-sample and gets that estimate recalculated. Thus it is seen that suitable action can be taken on the basis of the sub-sample estimates, thereby increasing the accuracy and utility of the final results.

6.14. The technique of interpenetrating sub-samples is of help in calculating the total variation especially in large scale sample surveys where a number of characteristics are under consideration. If there are  $k$  independent interpenetrating sub-samples subjected to  $k$  different operations each providing a valid estimate of the population parameter under consideration, then an unbiased estimator of the variance of the estimator (mean of the sub-sample estimates) is given by

$$\frac{1}{k(k-1)} \sum_{i=1}^k (y_i - \bar{y})^2, \quad (\bar{y} = \frac{1}{k} \sum_{i=1}^k y_i) \quad (6.15)$$

where  $y_i$  is the estimate based on the  $i$ th sub-sample. It may be noted that this procedure gives a simple method of getting an estimator of the variance of

a ratio estimator. If  $r_i (= \frac{y_i}{x_i})$ , ( $i = 1, 2, \dots, k$ ) is an estimate of the population ratio  $R (= \frac{Y}{X})$  based on the  $i$ th sub-sample, then an unbiased estimator of the variance of

$$R' = \frac{1}{k} \sum_{i=1}^k r_i \quad (6.16)$$

is given by

$$V(R') = \frac{1}{k(k-1)} \sum_{i=1}^k (r_i - R')^2. \quad (6.17)$$

Since <sup>be</sup> it can be shown that the variance of  $R'$  and that of the combined ratio estimator

$$R'' = \frac{\sum_{i=1}^k y_i}{\sum_{i=1}^k x_i} \quad (6.18)$$

are approximately the same, (6.17) can be taken as an estimator of the variance of  $R''$ .

6.15. Suppose there are two agencies and two parties of investigators within each agency to conduct the survey. Then  $8k$  ( $k$  being an integer) independent interpenetrating sub-samples may be selected and each party of investigators in each agency may be assigned  $2k$  sub-samples at random for being surveyed. With this arrangement the total variation of the estimator may be analysed as given below.

source of variation	degrees of freedom
between agencies	1
between parties	2
within error	$8k-4$
total	$8k-1$

This analysis will help in locating the stages of operations where there is much of discrepancy. For instance if the between agency difference turned out to be statistically significant, this would mean that the survey has not been carried out according to the specifications by one or both of these agencies. Similarly a significant result for the parties will help in locating that party which is not functioning according to the specifications.

6.16. An illuminating example of a situation, where the sample survey estimate turned out to be nearer the true value than the complete enumeration figure, is provided by the Jute Survey in Bengal (India and Pakistan) during the years 1944-45 and 1945-46. (Mahalanobis and Lahiri, 1960). Jute being a cash crop of international importance, accurate figures for production become available subsequently. The official forecast in these years were based on complete enumeration of all plots. Sample surveys were also conducted by the method of actual physical observation of randomly selected plots. The results of the enquiry are given below.

Table (1): Comparison of official (complete enumeration) and sample survey estimates with very reliable trade figures, Bengal 1944-45, 1945-46.

sr. no.	item	quantity (thousand bales)	
		1944-45	1945-46
(0)	(1)	(2)	(3)
1.	trade figure	6728	7562
2.	complete enumeration	4895	6304
3.	sample survey	6480	7540
4.	discrepancy between (2) and (1)	- 27.2 $\phi$	- 16.6 $\phi$
5.	discrepancy between (3) and (1)	- 3.6 $\phi$	- 0.3 $\phi$

(Source: Mahalanobis, P.C. and Lahiri, D.B. (1960); Analysis of errors in censuses and surveys with special reference to experience in India, 32nd Session of the International Statistical Institute, Tokyo).

This interesting example shows that the sample survey provided a more accurate figure than the census because of the reduction in non-sampling errors made possible by confining the survey to a sample.

6.17. Another interesting case of response bias is provided by a study conducted in central Iowa, U.S.A. by the Iowa State College (Hendricks, 1956). In this survey the figures for volume of corn in a sample of 50 cribs arrived at on the basis of farmers' judgement estimates were compared with the objective measurements got after the harvest. The result of this study showed that the judgement estimate was about 15  $\phi$  below the objective figure.

Table (2): Objective corn yield estimates compared with estimates from reported data.

area sampled	bushels/acre			
	objective estimate		reported estimate	
	corn in field	adjusted for loss in harvest	unadjusted	adjusted to net acreage and 15 % understatement
(1)	(2)	(3)	(4)	(5)
Alabama (1948)	26	23.4	21.0 <sup>x</sup>	24.8
North Carolina (1949)	41	36.9	31.5 <sup>x</sup>	37.2
Virginia (1949)	55	49.5	42.0 <sup>x</sup>	49.6
10 Southern States (1954)	21.8	19.6	16.4 <sup>x</sup>	19.4
Central Iowa (1953)	79.3	71.4	58.3 <sup>y</sup>	68.8
Central Iowa (1954)	74.0	66.6	55.7 <sup>y</sup>	65.7

x: official estimates; y : reported data.

(Source: Hendricks, W.A. (1956): Non-sampling errors in agricultural surveys, Improving the Quality of Statistical Surveys, 31-39).

6.18. The post-enumeration survey in the 1950 Census of the U.S.A. showed an under-enumeration of 1.4 %, which may be taken as the non-sampling bias in that census. Similarly the post-enumeration survey in the 1951 census of India showed an under-enumeration of 1.1 % and the non-sampling bias in 1956 Livestock Census of India was assessed to be about 15 % (including processing errors) for large heads by a post-census check survey.

6.19. The technique of statistical quality control (SQC) may be applied to census and survey work to assess the quality of work and to improve the out-going quality with suitable corrective action. For this purpose it is desirable to use those SQC techniques which have built-in devices for initiating corrective action. More attention is to be paid to control of errors through SQC techniques than to acceptance plans for finished work. For a particular situation, the best plan is defined as that which ensures the highest out-going quality for a given cost or the lowest cost for a specified out-going quality. There is considerable scope to apply SQC techniques for control of errors in case of large scale surveys because of the large amount

REFERENCES

1. BASU, D.(1958): On sampling with and without replacement; Sankhya, 20, 287-294.
2. BIRNBAUM, Z.W. and SIRKIN, M.G.(1950): Bias due to non-availability in sampling surveys; Jour. Amer. Stat. Assn., 45, 98-111.
3. BOWLEY, A.L. (1926): Measurement of precision attained in sampling; Bull. Inter. Stat. Inst., 22, 1-62.
4. COCHRAN, W.G. (1942): Sampling theory when the sampling units are of unequal sizes; Jour. Amer. Stat. Assn., 37, 199-212.
5. " (1953) : Sampling Techniques; John Wiley & Sons, New York.
6. DALENIUS, T.(1950): The problem of optimum stratification; Skand. Akt., 33, 203-213.
7. " (1952): The problem of optimum stratification in a special type of design; Skand. Akt., 35, 61-70.
8. " (1953): The economics of one-stage stratified sampling; Sankhya, 12, 351-356.
9. " (1957): Sampling in Sweden, Almqvist & Wicksell, Stockholm.
10. DALENIUS, T and GURNEY, M. (1951): The problem of optimum stratification II; Skand. Akt., 34; 133-148.
11. DALENIUS, T. and HODGES, J.L. (Jr.) (1957): The choice of stratification points; Skand. Akt., 40, 198-203.
12. DAS, A.C.(1951): On two phase sampling and sampling with varying probabilities; Bull. Inter. Stat. Inst., 33(2), 165-112.
13. DEMING, W.E. (1950): Some theory of sampling; John Wiley & Sons, New York.
14. DESAI, J. (1956): Some estimators in sampling with varying probabilities without replacement; Jour. Amer. Stat. Assn., 51, 269-284.

15. DURBIN, J.(1960): A note on the application of Quenouille's method of bias reduction to the estimation of ratios; *Biometrika*, 47, 477-480.
16. GODAMBE, V.P. (1955): A unified theory of sampling from finite population; *Jour. Roy. Stat. Soc.*, 17(B), 269 - 278
17. GOODMAN, L.A. (1953): A simple method for improving some estimators; *Ann. Math. Stat.*, 24, 114 - 117.
18. GOODMAN, L.A. and HARLEY, H.O. (1958): The precision of unbiased ratio type estimators; *Jour. Amer. Stat. Assn.*, 53, 491 - 508.
19. HANSEN, M.H. and HURWITZ, W.N. (1943): On the theory of sampling from finite populations; *Ann. Math. Stat.*, 14, 333 - 362.
20. " " (1946) : The problem of non-response in sample surveys; *Jour. Amer. Stat. Assn.*, 41, 517 - 529.
21. HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953): *Sample Survey Methods and Theory*, Vol. I & II; John Wiley & Sons. New York.
22. HANSEN, M.H., HURWITZ, W.N. and BERSHAD, M.A. (1960) : Measurements of errors in censuses and surveys, presented at the meeting of the Inter. Stat. Inst. at Tokyo.
23. HANSEN, M.H., HURWITZ, W.N., MARKS, E.S., and MAULDIN, W.P. (1951) : Response errors in surveys; *Jour. Amer. Stat. Assn.*, 46, 146 - 190.
24. HANUMANTHA RAO (1960): Some sampling mechanism (an abstract), *Proceedings of the 47th session of the Indian Science Congress*, Part III, 35.
25. HARTLEY, H.O. and RAO, J.N.K. (1959): Sampling with unequal probabilities and without replacement, *proceedings of the Social Statistics Section of the American Statistical Association*. 56 - 59.
26. HARTLEY, H.O. and ROSS, A (1954): Unbiased ratio estimators; *Nature*, 174, 220 - 271.

27. HENDRICKS, W.A. (1956): Non-sampling errors in agricultural surveys, improving of the quality of statistical surveys; Amer. Stat. Assn. publication, 31 - 39.
28. HORVITZ, D.G. and THOMPSON, D.J.(1952): A generalisation of sampling without replacement from a finite universe; Jour. Amer. Stat. Assn., 47, 663 - 685.
29. KISH, L. and SLATER, C.W. (1960): Two studies of interview variance of socio-psychological variables, presented at the Annual Meeting of the Amer. Stat. Assn.
30. LAHIRI, D.B.(1951): A method of sample selection providing unbiased ratio estimates; Bull. Inter. Stat. Inst., 33 (2), 133 - 140.
31. " (1954): Technical paper on some aspects of the development of sample design; Sankhya, 14, 264 - 316.
32. " (1957): Recent developments in the use of techniques for assessment of errors in national surveys in India; Bull. Inter. Stat. Inst., 36 (2), 71 - 93.
33. " (1957): Observations on the use of interpenetrating samples in India; Bull. Inter. Stat. Inst., 36 (3), 144 - 152.
34. MADOW, W.G. and MADOW, L.H. (1944): On the theory of systematic sampling I; Ann. Math. Stat., 15, 1 - 24.
35. MADOW, W.G. (1949): On the theory of systematic sampling II; Ann. Math. Stat., 20, 333 - 354.
36. MAHALANOBIS, P.C. (1940): A sample survey of acreage under jute in Bengal; Sankhya, 4, 511 - 530.
37. " (1944) : On large scale sample survey; Phil. Trans. Roy. Soc., 231 B, 329 - 451.
38. " (1946): Recent experiments in statistical sampling in the Indian Statistical Institute; Jour. Roy. Stat. Soc., 109, 325 - 370.



39. MAHALANOBIS, P.C. (1952): Some aspects of the design of sample surveys; Sankhya, 12, 1-7.
40. MAHALANOBIS, P.C. and LAHIRI, D.B. (1960): Analysis of errors in censuses and surveys with special reference to experience in India, presented at the Annual Meeting of the Inter. Stat. Inst. at Tokyo.
41. MIDZUNO, H. (1950): An outline of the theory of sampling systems; Amer. Inst. Stat. Math., 1, 149 - 156.
42. MIDZUNO, H. (1952): On the sampling system with probability proportional to sum of the sizes; Amer. Inst. Stat. Math., 3, 99 - 107.
43. MURTHY, M.N. (1957): Ordered and unordered estimators in sampling without replacement; Sankhya, 18, 379 - 390.
44. " (1960): A note on probability sampling; Proceedings of the 47th session of the Indian Science Congress, Part III, 34,
45. MURTHY, M.N. and NANJAMMA, N.S. (1959): Almost unbiased ratio estimates based on interpenetrating sub-sample estimates; Sankhya, 21, 381 - 392.
46. NANJAMMA, N.S., MURTHY, M.N. and SETHI, V.K. (1959): Some sampling systems providing unbiased ratio estimators; Sankhya, 21, 299 - 314.
47. NEYMAN, J. (1934) : On the two different aspects of the representative method of stratified sampling and the method of purposive selection; Jour. Amer. Stat. Assn.; Jour. Roy. Stat. Soc., 109, 558 - 625.
48. PASCUAL, J.N. (1961): Unbiased ratio estimators in stratified sampling; Jour. Amer. Stat. Assn., 56, 70 - 87.
49. POLITZ, A.N. and SIMONS, W.R. (1949): An attempt to get the not-at-homes into the sample without call-backs; Jour. Amer. Stat. Assn., 44, 9-31.

50. ROY, J. and CHAKRAVARTI, I.M. (1960): Estimating the mean of a finite population; Ann. Math. Stat., 31, 392 - 398.
51. SEN, A.R. (1952): Present status of probability sampling and its use in estimation of characteristics (an abstract), Econometrica, 20, 103.
52. SETH, G.R. (1961): On some aspects of sampling theory and practice, Presidential Address to Statistics Section of 48th session of the Indian Science Congress.
53. SUKHATME, P.V. (1935): Contributions to the theory of representative method; Jour. Roy. Stat. Soc., Suppl., 253 - 268.
54. SUKHATME, P.V. (1953): Sampling theory of surveys with applications, Iowa State College Press.
55. SUKHATME, P.V. and SETH, G.R. (1952): Non-sampling errors in surveys; Jour. Ind. Soc. Agr. Stat., 4, 5 - 41.
56. YATES, F. (1953): Sampling methods for censuses and surveys; (Second edition); Charles Griffi & Co., London.
57. YATES, F. and GRUNDY, P.M. (1953): Selection without replacement from within strata with probability proportional to size; Jour. Roy. Stat. Soc., B, 15, 253 - 261.

# OPERATIONS RESEARCH

by

C. S. Ramakrishnan

## CONTENTS

	<u>Page</u>
1. Introduction, Scope and literature	1
2. Linear programming	
A. The problem of L.P., Classical examples and Simplex Method.	5
B. Numerical examples in L.P., including transportation and production scheduling.	15
C. Phase 1. of Simplex Method, Duality theorems	22
3. Inventory Control	
A. Case of random demand - one stage models	25
B. Case of deterministic demand - Economic Lotsize formulae.	27
4. Dynamic Programming - applied to the problem of optimum distribution of effort and the ware house problem.	28
5. Waiting Lines or Queues.	30
6. Replacement problems	34
7. Problems involving Monte Carlo methods	37

## Notes on Operations Research (CSR).

### 1. Introduction

In the last decade, we have seen a rapid use of mathematics and statistics in industry, economic planning, military strategy, a movement called by various names such as 'Operations Research', 'Management Science', 'Scientific Programming'. There has been a steady interplay between a more careful formulation of business and logistics problems and development of new mathematical disciplines and tools such as linear and non-linear programming, Queuing and Inventory Control theory etc. All these are concerned with the same fundamental task of scientifically analysing various courses of action and determining the 'best'.

A decision problem has four parts (1) a model (expressing a set of assumed empirical relations among a set of variables) which is a suitable abstraction of reality. (2) a subset of decision variables, whose values are to be chosen by the firm or decision making entity (3) an objective function of the variables, which is to be maximized and (4) procedures for analysing the effect on the objective function of alternative values of the decision variables.

Decision problems can be classified according to mainly three different criteria (1) static and dynamic (2) deterministic and stochastic and (3) the space of available strategies finite or infinite dimensional. Though pure mathematicians are themselves usually pessimistic of the use of infinity in this finite world, the modern operations researcher looks forward to the golden age of electronic computer and function spaces. Common sense has limitations which can be overcome only by deep and complex mathematics.

The leading journal in the field is 'Operations Research' (U.S.A.). Other journals are operations Research Quarterly (U.K.), Management Sciences (U.S.A.). The Naval Logistics Quarterly, Econometrica, and Applied Statistics (U.K.). The subject seems to be of equal interest to Mathematicians, Statisticians and Econometricians, and may help in bridging the gap between the three allied fields.

The first half of the course will deal with methods of O.R. which can be illustrated with numerical examples, and is based on the following books.

1. M. Sasieni, A. Yaspan, L. Fridman - Operations Research (methods and problems)
2. S. Yazsonyi - Scientific programming in business and industry
3. S. Gass - Linear programming - (methods and applications)
4. Churchman, Ackoff and Arnoff - Introduction to O.R.

The second half will deal with mathematical theory of linear and non-linear programming and relation to economic analysis and will be generally based on the following books.

1. S. Karlin - Mathematical methods and theory in Games, programming and Economics (Vol. 1).
2. Gale - Theory of Linear Economic models.

The following books are used for reference.

1. Samuelson, Dorfman and Solow - Linear programming and Economic analysis.
2. Koopmans (Editor) - Activity Analysis.
3. R.G.D. Allen - Mathematical Economics.

Time and personal limitations does not permit the author to talk on advanced methods of Inventory Control, Dynamic programming and Queuing theory. The following references on these may be consulted.

1. Arrow, Karlin and Scarf - Studies in theory of Inventory and Production.
2. R. Bellman - Dynamic programming
3. P.M. Morse - Queues, Inventories etc.
4. Bharucha Reid - Elements of Markov processes and applications.

The book of Bartlett on stochastic processes and the series of methuen monographs on applied probability and statistics edited by him are valuable for any O.R. mathematician, as are Blackwell and Girshick's Games and statistical Decisions and Feller's book on probability.

An O.R. Mathematician will be naturally attracted to the allied field of computational mathematics. The leading journals here are Journal of Association for computing machinery (U.S.A.) and Journal of Society for Industrial and Applied mathematics (U.S.A).

The following books are selected references.

1. Grabbe (Edited). Handbook of Automation, computation and control (3 volumes)
2. Bodewig - Matrix Calculus
3. Householder - Principles of Numerical Analysis.

Two other topics interesting to O.R. mathematicians are (1) the minimisation of switching functions with the help of Boolean Algebra and (2) the problem of coding in information theory. For topic (1), the readers may consult the books of Phister and Caldwell and for (2) the papers of R.C.Bose and Bell System Technical Journal, especially years 1955-56 contains valuable articles on these topics by Slepian and Mc Closkey.

Good Case Histories of O.R. are very scarce, probably because the firms are reluctant to give out the information. The leading O.R. practitioner in industry seems to be A.W.Swan (U.K.) and he has written two survey articles including a large number of case histories, one in the latest 'Productivity - measurement Review', and another in the section on O.R. in the book 'Handbook of Industrial Management' edited by Grant and Ireson. The O.R. practitioner has very little time for mathematical theory and research. He has to rely on common sense and simple statistical techniques. Knowledge of Industrial Engineering, Time and Motion Study and Costing seems to be of more use to him than mathematical methods of decision processes. It may take a long time before the gap

between academic and professional interests can be filled out. Irwin and Co. (U.S.A) are bringing out a useful series of books on scientific management, two of which are

1. Bowman and Fetter - Production analysis for management
2. " " - Analysis of Industrial operation

For history of O.R. as well as case histories, reference may be made to

- 1 and 2. Mc Closkey and Trefethon - O.R. for management (2 vols.)
3. Flagle (Edited) - O.R. and System Engineering.

The first book on O.R. is by Morse and Kimball (1951). It only use calculus and elementary probability in the problem of hunting for a submarine. There O.R. is defined as 'a scientific method of providing executive departments with a quantitative basis for taking decisions regarding operations under their control'. P.M.Morse is a physicist, head of O.R. department at M.I.T., first president of O.R. Society of U.S.A. and we close with a quotation from him.

'What characterizes Research, or distinguishes it from other activities like engineering, or administration, is that Research is aimed at understanding rather than immediate practical use. Centuries of human experience have shown that such understanding is necessary before results of a practical nature are realised'.

Some more useful references:

1. Bibliography of O.R. (ORSA)
2. Saaty - Math. methods of O.R.
3. Morgenstern (Ed) - Economic Activity analysis
4. Goode and Machol - System Engineering.

-----

2. Linear programming - introduction

From the economic (or industrial) point of view, programming deals with the problem of allocation of existing resources among different activities to be performed, so as to maximize the over all effectiveness subject to limitations on either the amounts of resources or the way in which they can be spent. For computational tractability, we first take up linear programming which is concerned mathematically with the problem of maximisation or minimisation of a linear function of several variables, subject to several linear equations and/or inequalities (including restrictions of non-negativity on some of the variables). Any inequality can be changed into an equation by the introduction of a slack (new) variable, and an unrestrained variable can be expressed (replaced) as the difference of two new variables which are restrained to be non-negative. Thus the general L.P. problem can be transformed into the standard canonical (or simplex) form viz. Maximize  $c'x$  subject to  $x \geq 0$ ,  $Ax = b$   $n$  non-negative variables subject to  $m$  ( $<n$ ) equations.

The earliest example of L.P. is the Hitchcock - Koopmans transportation problem, formulated in 1941 and solved in 1947. Given  $m$  origins or factories with stocks or supplies of a commodity  $a_1 a_2 \dots a_m$  respectively and  $n$  destinations or retail stores with demands  $b_1 b_2 \dots b_n$  with total demand not exceeding total supply, and also given the cost matrix  $c_{ij}$ , cost of shipping a unit amount from  $i$ th origin to  $j$ th destination. To determine the quantities  $x_{ij}$  to be shipped from  $i$ th origin to  $j$ th destination, so as to minimise the total cost  $\sum \sum c_{ij} x_{ij}$ . Restraints are  $x_{ij} \geq 0$   $\sum_j x_{ij} = a_i$ ;  $\sum_i x_{ij} = b_j$ .

The next example was the diet problem of Stigler (1948). Given  $n$  foods and  $m$  nutrients (vitamins, proteins etc.),  $a_{ij}$  the amount of  $i$ th nutrient in a unit amount of  $j$ th food,  $b_i$  = the minimum amount of  $i$ th nutrient;  $c_j$  = the cost per unit amount of  $j$ th food. To determine the quantities  $x_1 \dots x_n$  of the  $n$  foods to be purchased so to minimise the total cost of the diet  $\sum c_j x_j$  and so as to satisfy the minimum requirements  $x_j \geq 0$   $\sum_j a_{ij} x_j \geq b_i$ .



The problem of Assignment (origin not clear probably Brogden 1946) of  $n$  individuals to  $n$  jobs so as to maximize total effectiveness, given the matrix  $c_{ij}$  of effectiveness of  $i$ th individual in  $j$ th job, was later on discovered to be a particular case of transportation problem, with additional constraints  $x_{ij}=0$  or  $1$ . However, the problem is highly degenerate (in a technical sense to be made precise later on) and the standard transportation technique is not efficient for solving this problem. A highly ingenious method using entirely new ideas was found out by Kuhn (1955) using results of graph theory. This is called Reduced Matrix method. This and allied work on net work flow problems by Ford, Fulkerson and Gale, form the main part of what is known as Integral programming for which no general theory is yet available.

Another interesting and early L.P. problem is the Activity-analysis problem (Koopman 1951) or Linear programming analysis of the firm. A manufacturer or firm has fixed amounts  $b_1 \dots b_m$  of  $m$  resources (raw material, labor, equipment, capital). These can be used to indulge in several different activities or to manufacture several different products ( $n$  in number). Given the input matrix  $a_{ij}$  the amount of resource  $j$  required to produce one unit of commodity  $j$  and  $c_j$  the profit per unit of commodity  $j$  produced. The problem is determine  $x_1 \dots x_n$ , the levels of the activities or the amounts of the products so as to maximize the profit  $\sum c_j x_j$ . Restraints  $x_j \geq 0$   $\sum_j a_{ij} x_j \leq b_i$  i.e. the total amount of  $i$ th resource used must not exceed the given  $b_i$ . The importance of this problem is due to the economic interpretation of its so called dual.

We shall later on study the more general activity problem where there is no distinction between inputs and outputs and also the non-linear programming version

Other famous problems of L.P. are (1) Caterer's problem Jacobs (1954); (2) Production Scheduling problem (Magee); (3) The ware house problem (Cahn); (4) The travelling salesman problem; (5) Transshipment problem (Orden); (6) Generalised transportation problem (Dantzig).

Special computing techniques are available for these special programming problems. Hosts of examples of L.P. in industry (Chemical, petroleum, Coal paper)

and farm management are available in literature. These are problems of general type and the most satisfactory method (general) is the simplex method of Dantzig, inspired by conversation with Von-Neumann. All the leading mathematicians in O.R. are students, friends and /or admirers of the two famous men A. Wald and J.V. Neumann.

### 3. Nature of solution of L.P.problem and Simplex method

**Consider** the canonical L.P. problem  $\max c \cdot x$  subject to  $Ax = b$  and  $x \geq 0$ . The set of all  $x \geq 0$  satisfying  $Ax = b$  is called the set of Feasible solutions (K) and is a convex set i.e.  $x, y \in K \rightarrow \lambda x + (1-\lambda)y \in K$   $0 \leq \lambda \leq 1$ . An extreme point (or corner) of K is a point which cannot be expressed as a proper convex combination of two other points  $\in K$ . Now there are 2 possibilities; (1) K is void i.e. the problem is not feasible; (2) K is not empty, i.e. the problem is feasible. Now there is a theorem saying that if K is feasible, then K has at least one extreme point. The simplex method has two phases. Phase I is aimed at finding out an extreme point of K. It is a constructive method which in finite number of steps enables us to find out whether K is empty and if it is not, to arrive at an extreme point of K, thus going a constructive proof of the above theorem. On many occasions one may be able to get an extreme point of K without going through phase I. Having got an extreme point of K, we start phase II. This first tests whether this extreme point is an optimal solution and if it is not, either (1) goes to an adjacent extreme point having a higher value for the preference function, or (2) discovers whether the preference function is unbounded over K. As K has only a finite number of extreme points, phase II will converge in finite number of steps. But the number of steps (or iterations) required cannot be known before hand and thus it is a finite iterative method, and not a direct or analytic method which will either give an explicit formula for the solution in terms of original data (or coefficients) or enable us to compute the solution by performing a predetermined number of operations on the coefficients.

Let us assume at first that the m equations in  $Ax = b$  or  $\sum x_i P_i = b$  independent. Selecting any set of m independent columns (variables) and putting the

remaining  $n-m$  variables equal zero we get a basic solution which will be feasible if the solution is non-negative. A non-degenerate basic feasible solution is a feasible basic solution with exactly  $m$  positive  $x_i$ . For simplifying proof and procedures, we shall first make the assumption that all basic feasible solutions are non-degenerate. Every basic feasible solution corresponds to an extreme point of  $K$ .

Theorem 1.  $(x_1 \dots x_n)$  is an extreme point of  $K$  if and only if the positive  $x_j$  are coefficients of linearly independent  $P_j$  in  $\sum_1^n x_j P_j = b$ .

Thus there are at most  $\binom{n}{m}$  bases or extreme points for  $K$  and if  $K$  be bounded then it will be the convex polyhedron of these extreme points.

Theorem 2. A linear function over a convex polyhedron attains its maximum at an extreme point. If it attains its maximum at more than one extreme point, then it takes the same maximum value over the sub-convex polyhedron generated by these extreme points.

Phase II of simplex method. Let the first  $m$  col. vectors of  $A$  (without loss of generality) be independent and give a non-degenerate basic feasible solution  $(x_1, x_2, \dots, x_m, 0 \dots 0)$   $x_1 P_1 + x_2 P_2 + \dots + x_m P_m = b$   $x_i > 0$  and let

$Z_0 = x_1 c_1 + \dots + x_m c_m$  the corresponding value of the objective function.

Express the  $P_j$ 's in terms of this basis.  $P_j = x_{1j} P_1 + x_{2j} P_2 + \dots + x_{mj} P_m$   $j = 1, 2, \dots, n$  and define the quantities  $z_j = x_{1j} c_1 + \dots + x_{mj} c_m$ . Then we do the following simplex optimality test.

Theorem If  $z_j \geq c_j$  for all  $j$ , then the given solution is optimal.

Proof: Let  $y_1 P_1 + \dots + y_n P_n = b$  be any other solution and let

$y_1 c_1 + \dots + y_n c_n = Z^*$ . Then  $z_0 \geq Z^*$  as we shall show. Since  $c_j \leq z_j$

$Z^* \leq y_1 z_1 + \dots + y_n z_n$ . We have  $\sum_{j=1}^n y_j P_j = b$  and  $P_j = \sum_{i=1}^m x_{ij} P_i$ . Hence,

$\sum_j y_j \sum_i x_{ij} P_i = b$  or  $\sum_i (\sum_j y_j x_{ij}) P_i = b$  and since  $b = \sum_i x_i P_i$  and the expres-

sion in terms of the basis  $P_i$  is unique,  $\sum_j y_j x_{ij} = x_i$ .

Now  $Z^* \leq \sum y_j z_j$  and  $z_j = \sum_i x_{ij} c_i$ . Hence

$$Z^* \leq \sum_j y_j \sum_i x_{ij} c_i = \sum_i c_i \left( \sum_j y_j x_{ij} \right) = \sum_i c_i x_i = Z^*.$$

If this optimality test does not hold, we must have at least one  $C_j > Z_j$ .

Theorem If  $C_j > Z_j$  for some  $j$ , then we have two cases.

Case 1. If for any of those  $j$  for which  $C_j > Z_j$ , we have  $x_{ij} \leq 0$  then we can have a set of solutions tending to infinite value of the objective function.

Case 2. If for all  $j$  for which  $C_j > Z_j$ , we have at least one  $i$  for which  $x_{ij} > 0$ , then by including any such  $P_j$ , and dropping that  $P_i$  for which  $\frac{x_i}{x_{ij}}$  is minimum and  $> 0$ , we can get a new (adjacent) basic feasible solution having a higher value for  $z$ .

Proof.  $\sum x_i P_i + \theta P_j - \theta P_j = b$  and  $P_j = \sum x_{ij} P_j$ . Hence

$$\sum (x_i - \theta x_{ij}) P_i + \theta P_j = b \text{ and the corresponding } z = \sum (x_i - \theta x_{ij}) c_i + \theta c_j = Z_0 + \theta(C_j - Z_j).$$

If all  $x_{ij} \leq 0$ , then  $\theta$  can have any positive value and hence  $z \rightarrow \infty$ . If at least one  $x_{ij} > 0$ , then the maximum value of  $\theta$  which will still give a feasible solution is  $\theta_0 = \min_i \frac{x_i}{x_{ij}}$  over those  $i$  for which  $x_{ij} > 0$  and for this  $\theta_0$ ,

upon the non-degeneracy assumption,  $\theta_0 > 0$  and the minimum will be attained at a unique  $i$ . It can be shown that the new set of  $P_1 \dots P_m, P_j$  after dropping  $P_i$  will be a basis. In general, we can include any  $P_j$  having  $C_j > Z_j$ , but it is found empirically the most convenient and effective rule is to choose that  $j$  for which  $(C_j - Z_j)$  is maximum. Let  $\max_j (C_j - Z_j) = Z_k - C_k > 0$ .  $P_k$  is to

be introduced. Compute  $\theta_0 = \min_i \frac{x_i}{x_{ik}}$  for  $x_{ik} > 0 = \frac{x_r}{x_{rk}}$ .  $P_r$  has to be

dropped. The new solution is  $P_0 = y_1 P_1 + \dots + y_k P_k + \dots + y_m P_m$  where

$y_k = \frac{x_r}{x_{rk}}$  and  $y_i = x_i - y_k x_{ik}$  for  $i \neq r$ . Similarly to express  $P_j$ 's in terms

of the new basis,

$$P_j = y_{1j} P_1 + \dots + y_{kj} P_k + \dots + y_{mj} P_m$$

where  $y_{kj} = \frac{x_{rj}}{x_{rk}}$  and  $y_{ij} = x_{ij} - y_{kj} x_{ik}$ . The new  $c_j - z_j = (c_j - z_j) - y_{kj}(c_k - z_k)$  and the new  $-z = -z_0 - y_k(c_k - z_k)$ .

The whole computation can be arranged in a tabular form and succeeding simplex tableaus can be got by sweep out process which can be carried out even by a computer with just SSLC education. Having chosen k and r, divide rth row by  $x_{rk}$  to make the element in the (r, k) cell unity, and with the help of this new row sweep out the other elements of kth column to zero.

If the problem were in activity form  $Ax \leq b$  ( $b > 0$ ), we can immediately write down the initial simplex tableau without going thro' phase 1. In the general case  $Ax = b$ , if ~~one can~~ by general knowledge of the variables of the problem, and/or by trial and error one can get a basis  $B(m \times m)$  of columns of  $A$  such that  $x = B^{-1} b \geq 0$ , we can write down the initial tableau with the help of  $B^{-1}$ . In general, one may guess a few of the important variables which will be in the optimal solution but no fool-proof method for getting a basis with these variables is available. In general, we do not know even whether the m equations are independent, let alone a basic feasible solution. Two ways out of this have been suggested. A phase I, which is to apply simplex method to the maximize  $-\sum y_i$  subject to  $Ax + Iy = b$  and non-negativity (without loss of generality  $b > 0$ ). If this results in  $y = 0$ , we have a b.f.s. for original problem. Otherwise it is not feasible. Another way is the M-Method,  $\max c'x - My$  sub.  $Ax + Iy = b$  and non-negativity, where M is a large no, (not necessary to specify it).

A small example: Max.  $z = 3x_1 + 5x_2 + 4x_3$  subject to non-negativity and  $2x_1 + 3x_2 \leq 8$ ,  $2x_2 + 5x_3 \leq 10$ ,  $3x_1 + 2x_2 + 4x_3 \leq 15$ .

Introducing slacks  $x_4, x_5, x_6$  we have the first Simplex tableau

	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$	b
$P_4$	2	3	0	1	0	0	8
$P_5$	0	2	5	0	1	0	10
$P_6$	3	2	4	0	0	1	15
$C_j - Z_j$	3	5	4	0	0	0	$0 \rightarrow (-Z)$

$P_2$  enters and  $P_4$  drops out. The second tableau is

	$P_1$	$P_2$	$P_3$	$P_4$	$P_5$	$P_6$	b
$P_2$	$2/3$	1	0	$1/3$	0	0	$8/3$
$P_5$	$-4/3$	0	5	$-2/3$	1	0	$14/3$
$P_6$	$5/3$	0	4	$-2/3$	0	1	$29/3$
$C_j - Z_j$	$-1/3$	0	4	$-5/3$	0	0	$-40/3 \rightarrow (-Z)$

The second solution is thus  $x_2 = 8/3, x_5 = 14/3, x_6 = 29/3, x_1 = x_3 = x_4 = 0$  and the new increased value of  $Z = 40/3$ . Two more rounds will give the answer  $x_2 = 50/41, x_3 = 62/41, x_1 = 89/41, x_4 = x_5 = x_6 = 0, z = 765/41$ . See P.233 (Friedman).

The following economic (common-sense) significance of simplex optimality test is from Samuelson, Solow and Dorfman p.164.

A basic feasible program is optimal if and only if the included activities are such that no excluded activity is more profitable than its equivalent combination in terms of the included activities.

Here are two very good practical examples of L.P. in Agriculture one of (maximising) Activity analysis variety and the other of (minimising) diet variety due to Waugh (Econometrica 1955 and Journal of Farm Economics 1951).

1. Allocation of Land, Capital and Labor for a Farming Enterprise

(Table showing the quantity of resources required for each product, resources available and the net profit on each product).

Resources	Products						available resources
	Potato	Corn	Beans	Beef	Cabbage	Lettice	
Spring land	1.0	1.0	1.0	1.0	-	-	60
Fall land	-	1.0	1.0	1.0	1.0	1.0	60
Production capital	99.4	37.8	19.8	27.2	74.8	53.0	2000
Jan.-Feb. Labour	2.4	1.5	-	-	-	-	351
Mar.-April Labour	2.0	2.0	-	-	-	-	448
May-June Labour	1.8	3.3	5.3	-	-	-	479
July-Aug. Labour	-	-	2.1	-	-	-	388
Sept.-Oct. Labour	-	-	.4	-	19.1	12.4	424
Nov.-Dec. Labour	-	3.0	.4	-	9.1	26.7	359
Net profit	83.4	72.4	27.3	36.0	20.7	45.5	

2. Minimum-cost dairy-feed diet

(Table showing unit costs and nutrient contents of 10 different feeds and minimum requirements).

Contents	Feeds						
	Corn	Oats	Milo	Bran	Flour	Linseed	Cotton
Nutrient	78.6	70.1	80.1	67.2	78.9	77.0	70.6
Protein	6.5	9.4	8.8	13.7	16.1	30.4	32.8
Calcium	.02	.09	.03	.14	.09	.41	.20
Phosphorous	.27	.34	.30	1.29	.71	.86	1.22
Unit cost	2.4	2.52	2.18	2.14	2.44	3.82	3.55
	Bean	Gluten	Homing	Requirements			
	78.5	76.3	84.5	74.2			
	37.1	21.3	8.0	14.7			
	.26	.48	.22	.14			
	.59	.82	.71	.55			
	3.70	2.60	2.54				

3. An application of L.P. in planning (T.P.Chowdhri)

All employable persons are classified into six categories and the number of employments (in 1000's) created in 5 industries for an investment of Rs.100 crores is given below (included in the first Table). Prepare an investment plan which will minimise total investment and will create at least 10, 12, 15, 8, 20 and 10 thousands respectively in the above categories.

Let  $x_1 \dots x_5$  be the amount to be invested in the five industries and  $x_6 \dots x_{11}$  be slack variables (excess employment in each category). The problem is to determine  $x_1 \dots x_5 \geq 0$ , so as to minimize  $\sum_1^5 x_i$  and to satisfy the restrictions given in the following table.

Employment category	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$	$x_{11}$	Require-ment
1	.49	.64	.39	.18	.62	-1						10
2	.59	.76	.72	.30	.13		-1					12
3	.35	.54	.45	.53	.54			-1				15
4	.16	.10	.15	.32	.32				-1			8
5	3.08	2.18	1.22	2.32	.20					-1		20
6	.10	.10	.10	10.52	.10						-1	10

The problem is of the diet type. Because restrictions are all  $\geq$ , we do not get unit matrix, so as to enable us to write down a first feasible solution. To include six artificial vectors and go thro' first phase is too long. Gauss suggests a method by which only one artificial variable is included. The following method is offered by the author to get a first feasible basic solution easily. The idea is to chose one of the industries and try to find the minimal



investment in that industry necessary to provide all the minimum requirements.

We have inequalities of the type  $a_{i1} x_1 \geq b_i$   $x_1 \geq \frac{b_i}{a_{i1}}$  and the smallest value of  $x_1$  is  $\max. \frac{b_i}{a_{i1}}$ . In this case, this happens for  $i = 6$ , and  $x_1 = 10 / .10 = 100$ .

An investment of 100 crores in the first industry will meet the requirements and give a basic feasible solution with  $x_{11}$  = the sixth slack variable = 0, also  $x_2 = x_3 = x_4 = x_5 = 0$ , and  $x_6, \dots, x_{10}$  other slacks positive. To translate this idea into a first simplex tableau having a unit matrix, we have to divide the last row by its first element .10 to make it unity, and then make other elements in that column zero by subtracting every other equation from a suitable multiple of the new last equation. Here is the simplex tableau.

	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>	P <sub>9</sub>	P <sub>10</sub>	P <sub>11</sub>	P <sub>0</sub>
P <sub>6</sub>	0	-.15	.15	51.31	-.13	1					- 4.9	39
P <sub>7</sub>	0	-.17	-.13	61.77	.46		1				- 5.9	47
P <sub>8</sub>	0	-.19	-.10	36.29	-.19			1			- 3.5	20
P <sub>9</sub>	0	.06	.01	16.51	-.16				1		- 1.6	8
P <sub>10</sub>	0	.90	1.86	32.70	2.88					1	-30.8	288
P <sub>11</sub>	1	1	1	105.2	1						-10	100
Z <sub>j</sub> -C <sub>j</sub>	X	0	0	104.2	0	x	x	x	x	x	-10	100->(Z)

P<sub>4</sub> is the next vector to go into the basis. After 3 or 4 rounds we get the answer  $x_1 = 0$ ,  $x_2 = 9.79$ ,  $x_3 = 0$ ,  $x_4 = 13.54$ ,  $x_5 = 6.80$ . Total investment = 304 crores. We might have guessed that  $x_4$  is most likely to be in the optimal solution and might have got a quicker answer by introducing  $x_5$  at first instead of  $x_1$ .

1. Nutmix problem

A manufacturer wishes to determine an optimal program for mixing three grades of nuts consisting of cashew-nuts, hazels, and peanuts according to the specifications and selling price given in Table 1 and capacity limit on the inputs and buying prices given in Table 2.

Table 1

Mixture	Specifications	Selling price As / lb.
1	at least 50 ¢ cashew-nut, at most 25 ¢ peanuts	50
2	at least 25 ¢ cashews, at most 50 ¢ peanuts	35
3	nil	25

Table 2

Inputs	Capacity lb / day	Price As/lb
Cashew	100	65
peanut	100	25
Hazel	60	35

(Charnes Cooper Henderson -  
Introduction to L.P.)

2. Production Scheduling

Find the minimum cost production schedule for a single product for the first six months of the year where it is planned to meet the following sales forecast

	Jan.	Feb.	Mar.	Apr.	May	June
$S_i$	500	700	1050	1230	1140	900

R - unit inventory cost = Rs 5 per month per unit (includes storage and interest charges on goods tied up).

P - unit overtime cost = Rs 9 per unit (includes cost due to drop inefficiency)

m - Regular time capacity = 900 units per months

n - Overtime capacity = 200 units per month

$x_i$  - Amount of regular production in ith month = ?

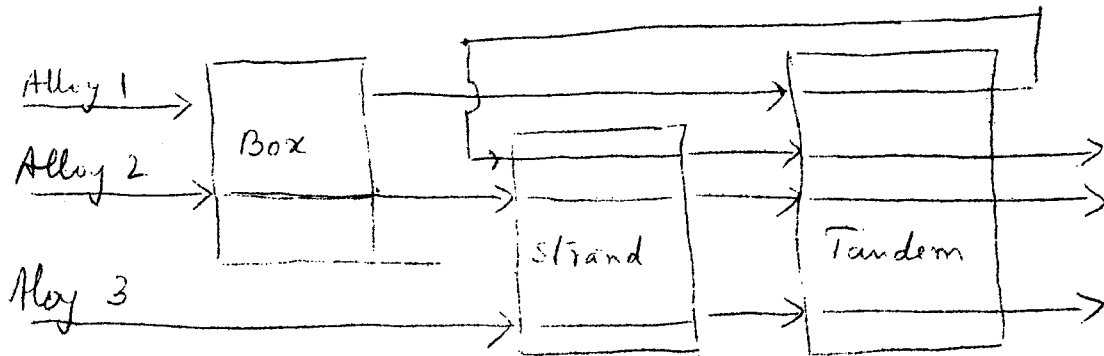
$y_i$  - Amount of overtime production in ith month = ?

Any normal costs such as regular labor, overhead, and raw material costs, which must be borne at some time if the schedule is to be met, can be disregarded as they are unaffected by the schedule. The costs affected by the schedule are essentially only added production costs such as overtime, and inventory costs resulting from the time of production.

3. Solve the following transportation problem

	store					supplies		
warehouse	1	9	12	9	6	9	10	5
	2	7	3	7	7	5	5	6
	3	6	5	9	11	3	11	2
	4	6	8	11	2	2	10	9
Requirements		4	4	6	2	4	2	

4. A steel mill produces three types of coils, each made of a different alloy. The process flow chart looks like Fig. 1. The problem is to determine the amounts of each alloy to produce, within the limitations of sales and machine capacities, so as to maximize profits.



Data on capacities and profits are given in Tables 1 and 2.

Table 1

Machine	No. of machines	8-hour Shifts per week	Down time, $\phi$
Box anneal	4	21	5
Strand anneal	1	20	10
Tandem roll	1	12	0

contd.

Table 2

Alloy	Operation	Machine rate	Sales potential	Profit per ton
1	Box anneal	28 hr per 10 tons	1250 tons per month	\$ 25
	Tandem roll (1)	50 ft per min		
	Strand anneal	20 ft per min		
	Tandem roll (2)	25 ft per min		
2	Box anneal	35 hr per 10 tons	250 tons per month	\$ 35
	Strand anneal	20 ft per min		
	Tandem roll	25 ft per min		
3	Strand anneal	16 ft per min	1500 tons per month	\$ 40
	Tandem roll	20 ft per min		

Coils for each alloy are 400 feet long and weigh 4 tons. Set up the objective function and the restrictions, from which a simplex solution to the manufacturer's problem might be obtained.

5. The strategic bomber command receives instructions to interrupt the enemy's tank production. The enemy has four key plants located in separate cities, and destruction of any one plant will effectively halt the production of tanks. There is an acute shortage of fuel, which limits the supply to 48,000 gallons for this particular mission. Any bomber sent to any particular city must have at least enough fuel for the round trip plus a reserve of 100 gallons.

The number of bombers available to the commander and their descriptions are listed in the following table.

Bomber type	Description	Miles per gallon	Number available
1	Heavy	2	48
2	Medium	3	32

Information about the location of the plants and their vulnerability to attack by a medium bomber and a heavy bomber is given below.

Plant	Distance from base, miles	Probability of destruction by	
		a heavy bomber	a medium bomber
1	450	0.10	0.08
2	480	0.20	0.16
3	540	0.15	0.12
4	600	0.25	0.20

How many of each type of bomber should be dispatched, and how should they be allocated among the four targets, in order to maximize the probability of success ? (Assume that no damage is inflicted on a plant by a bomber that fails to destroy it.)

6. The sales forecasts for a certain product, by months, are given below.

January	3,000	July	10,000
February	3,000	August	6,000
March	4,000	September	4,000
April	6,000	October	3,000
May	8,000	November	2,000
June	10,000	December	2,000

It costs \$ 1.00 per unit to increase production from one month to the next, and \$ 0.50 per unit to decrease production. Production scheduled for the month of December in the current year is 2000 units, and it is estimated that the inventory level on January 1 will be 1000 units. Storage capacity is limited to 5000 units at any one time.

Show how to obtain a production schedule for the coming year that will minimize the cost of changing production rates, while at the same time insuring that sufficient stock will be available to meet the sales forecast at all times. (Assume that production scheduled during a month becomes available for shipment just in time to meet the current month's sales demand)

7. A ship has three cargo holds, forward, aft, and center. The capacity limits are:

Forward	2000 tons	1000,000	cubic feet
Center	3000 tons	135,000	cubic feet
Aft	1500 tons	30,000	cubic feet

The following cargoes are offered; the shipowners may accept all or any part of each commodity:

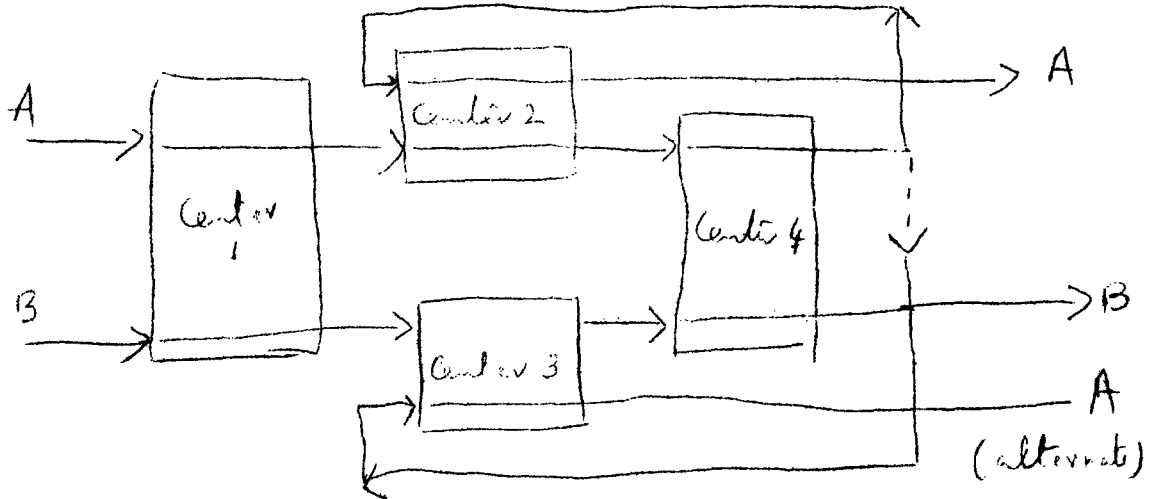
<u>Commodity</u>	<u>Amount, tons</u>	<u>Volume per ton cu ft</u>	<u>Profit per ton, \$</u>
A	6,000	60	6
B	4,000	50	8
C	2,000	25	5

In order to preserve the trim of the ship, the weight in each hold must be proportional to the capacity in tons. How should the cargo be distributed so as to maximize profit? (Note: Since the restrictions will contain an equality, the approach suggested in the previous problem may be needed in order to obtain a numerical answer.)

8. A trucking company with \$ 400,000 to spend on new equipment is contemplating three types of vehicle. Vehicle A has a 10-ton payload and is expected to average 35 miles per hour. It costs \$ 8000. Vehicle B has a 20-ton payload and is expected to average 30 miles per hour. It costs \$ 13,000. Vehicle C is a modified form of B; it carries sleeping quarters for one driver, and this reduces its capacity to 18 tons and raises the cost to \$ 15,000.

Vehicle A requires a crew of one man, and , if driven on three shifts per day, could be run for an average of 18 hours per day. Vehicles B and C require a crew of two men each, but, whereas B would be driven 18 hours per day with three shifts, C could average 21 hours per day. The company has 150 drivers available each day and would find it very difficult to obtain further crews. Maintenance facilities are such that the total number of vehicles must not exceed 30. How many vehicles of each type should be purchased if the company wishes to maximize its capacity in ton-miles per day ?

9. A plant makes two products, A and B, which are routed through four processing centres, 1, 2, 3, 4, as shown by the solid lines in Fig.



If there is spare capacity in center 3, it is possible to route product A through 3 instead of going through 2 twice, but this is more expensive.

Given the information below, how should production be scheduled so as to maximize profits? [By a 'production schedule' is meant the specification of the following three amounts: (1) the daily amount of raw material used for A, regular route, (2) the daily amount of raw material used for A, optional route, (3) the daily amount of raw material used for B. Assume that sufficient storage capacity is available at no additional cost.].

contd..

Table

Product	Center	Input, gals per hr.	% recovery	Running cost cost per hr \$
A	1	300	90	150
	2 (1st pass)	450	95	200
	4	250	85	180
	2 (2nd pass)	400	80	220
	3	350	75	250
B	1	500	90	300
	3	480	85	250
	4	400	80	240

Product	Raw material cost per gal	Sales price per finished gal	Maximum daily sales, gal of finished product
A	5	20	1700
B	6	18	1500

Centers 1 and 4 run up to 16 hours a day; centers 2 and 3 run up to 12 hours a day. A final restriction is furnished by shipping facilities, which limit the daily output of A and B to a total of 2500 gallons.



Example of phase 1.     $\max x_1 + 2x_2$ , sub  $x_1, x_2 \geq 0$ ,  $-x_1 + 3x_2 \leq 10$ ,  
 $x_1 + x_2 \leq 6$ ,  $x_1 - x_2 \leq 2$ ,  $x_1 + 3x_2 \geq 6$ ,  $2x_1 + x_2 \geq 4$ .

Basis	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	Solutions
$x_3$	-1	3	1							10
$x_4$	1	1		1						6
$x_5$	1	-1			1					2
$x_6$	1	3				-1		1		6
$x_9$	2	1					-1		1	4
$w_j - d_j$	3	4				-1	-1			10
$x_3$	-2		1			1		1		4
$x_4$	2/3			1		1/3		-1/3		4
$x_5$	4/3				1	-1/3		1/3		4
$x_2$	1/3	1				-1/3		1/3		2
$x_9$	5/3					1/3	-1	-1/3	1	2
$w_j - d_j$	5/3					1/3	-1	-4/3		2
$x_3$			1			7/5		-6/5		
$x_4$				1		1/5		2/5		
$x_5$					1	-3/5		4/5		
$x_2$	1	1				-2/5		1/5		
$x_1$	1					1/5		-3/5		
$w_j - d_j$								-1	-1	0
$-Z_j + C_j$						-3/5	-1/5			22/5

Duality in Linear Programming

1. Canonical Simplex form (Unsymmetric)

Primal (P) max.  $c'x$  subject to  $Ax = b, x \geq 0$ .

Dual (D) min.  $b'y$  subject to  $y'A \geq c'$   $y$  unrestricted.

If  $x$  be feasible (P) and  $y$  be feasible (D),  $y'b = y'Ax \geq c'x$ .

Hence, P is bounded above and D is bounded below and  $\text{Min. } D \geq \text{Max. } P$ . Now it can be easily shown that if  $x, y$  be feasible for P and D and  $c'x = b'y$ , then  $x$  and  $y$  are optimal. Also, it follows  $\text{Max. } P$  can be infinite only if D is not feasible and  $\text{min. } D$  is minus infinity only if P is infeasible. It may happen that both P and D are infeasible. The duality theorem states that necessary and sufficient condition for P(D) to have a finite solution is that both are feasible in which case the value of the dual programs are equal. The dual (D) can be placed in the primal form and then its dual will be the original primal (P) itself.

With this observation, all that remains to be proved is:- Let P be feasible and has finite solution. Then D is feasible and has the same value. The proof that follows is a constructive one based on Simplex method and is by Dantzig and Orden. Other existence proof (very elegant ones) based on theory of Linear Inequalities have been given by Kuhn, Tucher and Gale.

Let B be the optimal basis and  $x$  be the solution given by the final simplex tablean obtained by solving P. Let  $C'_B$  be the condensation of  $c'$  to the variables in the basis. We shall show that  $y' = C'_B B^{-1}$  is feasible (D) and  $b'y = c'x$ . Let  $A = (B|P)$ .  $B^{-1}A = (I | X)$ .  $B^{-1}b = x_B$  (nonzero part of  $x$ ). The vector  $z$  of indirect profits =  $(C'_B, C'_B X)$  is  $c'$  because B is optimal. Hence  $y'A = C'_B B^{-1}A = c'_B(I | X) = z' \geq c'$  and hence  $y$  is feasible (D). Also  $y'b = C'_B B^{-1}b = C'_B x_B = c'x$ .

2. Standard form (Symmetric).

Primal P max  $c'x$  subj. to  $Ax \leq b, x \geq 0$

Dual D min  $b'y$  subj. to  $y'A \geq c'$   $y' \geq 0$

**Prof.** for the duality theorem in this form can be deduced from the previous one, by placing the above standard primal in the canonical form  $(A|I)x = b$   $y'I \geq 0$  is same as  $y' \geq 0$ . When  $(A|I)$  is transformed by simplex method into  $(I | X)$  except for a permutation of columns,  $I$  is automatically transformed into  $B^{-1}$ , as  $B$  becomes  $I$ .  $y'(A|I) = z'$  as before, and hence  $y'$  is nothing but the elements of  $z$  in the last  $m$  columns. Thus, the final simplex tableau in the case symmetric problem, automatically yields the solution of the dual also and another check on optimality can be got by checking  $y'b = c$ . In the case of the unsymmetric primal, however we don't get  $B^{-1}$  and  $y' = C'_B B^{-1}$  as by products but have to compute them separately after getting the final basis  $B$ . However, instead of ordinary simplex method, we have the Revised Simplex or Inverse Matrix method, which gets  $B^{-1}$  at each stage.

3. General Linear program and its dual. Let  $(1, 2, \dots, m) \in S$  and  $(1, 2, \dots, n) = T \cup T'$

Primal.  $\max c'x$  sub.  $(Ax)_i \leq b_i$  for  $i \in S$  and  $(Ax)_i = b_i$  for  $i \in S'$   
 $x_j \geq 0$  for  $j \in T$  and  $x_j$  unrestricted for  $j \in T'$ .

Dual.  $\min b'y$  subj.  $(y'A)_j \geq c_j$  for  $j \in T$  and  $(y'A)_j = c_j$  for  $j \in T'$   
 $y_i \geq 0$  for  $i \in S$  and  $y_i$  unrestricted for  $i \in S'$ .

The duality theorem for the general case, can be deduced from the canonical case, by placing the general in the standard form.

#### 4. Equilibrium theorem

Standard form  $x, y$  feasible for  $P$  and  $D$  are optimal if and only if, for those  $i$  having  $(Ax)_i < b_i$ ,  $y_i = 0$  and for those  $j$  having  $(y'A)_j > c_j$ ,  $x_j = 0$ .

and only if

Canonical form  $x, y$  feasible are optimal if/ , for those  $j$  for which  $(y'A)_j > c_j$ ,  $x_j = 0$ .

General form (Left as an exercise).

5. While discussing phase I of simplex method, we saw that the problem of solving Linear Inequalities (i.e. obtaining feasible solutions) can be converted into a linear programming problem. From the duality theorem, solving any

linear program, can be converted into a problem of solving simultaneous linear inequalities. In fact any  $x+y$  satisfying the feasibility conditions, and the additional one  $b'y \leq c'x$  automatically solves both the programs.

6. (Num. Example of page 11)  $y' = \frac{45}{41}, \frac{24}{41}, \frac{11}{41}$   $y'b = (8 \times 45 + 10 \times 24 + 11 \times 15)/41 = 765/41 = c'x.$

Inventory control

Lemma on (differentiation of an integral containing a parameter in the limit).

$$G(q) = \int_{a(q)}^{b(q)} g(q,x) dx$$

$$\frac{dG}{dq} = \int_{a(q)}^{b(q)} \frac{\partial g}{\partial q} dx + g(b,q) \frac{\partial b}{\partial q} - g(a,q) \frac{\partial a}{\partial q}$$

Demand for accomodity in a given period is a random variable with cumulative distribution  $F(x)$ . The problem for the business firm is to determine the optimum quantity  $q$  to be stocked so as to minimise expected profit. Given that  $p$  = profit per item sold and  $L$  = Loss per item not sold but stocked.

(a) If  $F(x)$  be absolutely continuous, show that the optimum quantity is given by the equation  $F(q) = P/(P + L)$ .

(b) If  $F(x)$  is a step function, (i.e. discrete distribution of demand), show that the optimum  $q$  is given by  $F(q-1) \leq P/(P + L) \leq F(q)$ .

Examples

1. The Metal Products Company has vacant space in their plant which they are considering leasing. The shortest term lease that they can get a renter to accept is one year. The amount of space the company will need next year is uncertain but can probably be approximated by the following:

Probability of needing  $x$  or more square feet =  $\frac{-x}{150,000}$  . The space

will rent for \$ 0.30 a square foot. If Metal Products needs more space than is available, the loss in contribution to profit and overhead per year is \$ 0.80 a square foot. If the vacant space in question is 200,000 square feet, how much should be rented ?

2. Let demand 'x' be stochastic, continuous, with d.f.  $F(x)$  and p.d.f  $f(x)$ . The problem is to determine the optimum stock  $q$ , when the profit function is piecewise linear, i.e.  $G(q, x)$

$$\begin{aligned} G(q, x) &= a_1 q + b_1 x + c_1 && (x \leq q) \\ &= a_2 q + b_2 x + c_2 && (x \geq q) \end{aligned}$$

Derive the resulting equation for optimum  $q$  and show that it can be easily solved with the help of tables of  $F(q)$  and  $f(q)$ .

(b) Consider the problem of a bread shopkeeper - How much to stock ? Let  $A$  be the cost of purchasing bread,  $B$  sale price,  $C$  the refund on stale bread, and finally  $D$  the penalty for loss of goodwill when there is a dissatisfied customer (all costs are unit costs). Observe that the profit function is of the form described in (a), and deduce as a particular case from the results you have obtained in (a) that the optimum quantity  $q$  is given by  $F(q) = (B+D-A)/(B+D-C)$ .

(c) If  $A = 8$ ,  $B = 20$ ,  $C = 2$ ,  $D = 5$ , and the demand rectangular between 1000 and 2000, show that the best order quantity  $q$  is 1740 loaves of bread.

Economic lot size formulae in the case of deterministic demand

Model. 1. Classic and Simplest case. Harris (1915).

$D_y$  = Total demand in a fixed period (usually year) at a uniform rate.

$C_3$  = Cost of set up per production run or cost of ordering.

$C_{1y}$  = inventory carrying cost, per unit, per year.

Shortages not permitted. To determine  $n$ , the number of runs (or the interval between orders), and  $x$  the quantity to be ordered every time  $nx = D_y$ . Total cost per year = ordering cost + inventory carrying cost (the first one a decreasing function of  $n$ , and the 2nd an increasing function of  $n$ ) =  $C_3 \cdot n + C_{1y} \frac{x}{2} = \frac{C_3 D_y}{x} + \frac{C_{1y} x}{2}$ . This is minimum, by differentiation, when  $\frac{C_3 D_y}{x} = \frac{C_{1y} x}{2}$  or  $x = \sqrt{\frac{2C_3 D_y}{C_{1y}}}$

Alternative way of formulating the same model.

$R$  = demand/unit time (uniform)  $C_3$  = cost per run or order

$C_1$  = inventory carrying cost/unit time/unit of the article.

Total cost/unit time =  $\frac{1}{2} C_1 R t + \frac{C_3}{t} = \frac{1}{2} C_1 q + \frac{C_3 R}{q}$  min. when

$$q = \sqrt{\frac{2C_3 R}{C_1}} \quad q=Rt = \text{size of order, } t = \text{interval between order.}$$

Model. 2. In the previous case delivery was immediate. Now we assume a finite production rate  $k$ /unit time ( $k > R$ ).

Cost/unit time =  $C_3 R/q + \frac{1}{2} C_1 q(1 - R/k)$ , min. when

$$q = \sqrt{\frac{2C_3 R}{C_1(1-R/k)}} \quad \text{when } k = \infty, \text{ we get model 1.}$$

Model. 3. Shortages are allowed. Immediate delivery. Interval between orders fixed by other considering =  $T$  (month, week etc.) .  $C_2$  = ~~penalty~~ cost (per unit time) of failing to deliver one unit of article on schedule. Problem is to determine the inventory level  $z$  at the beginning of a period.

Total cost per period.  $F(z) = C_1 \frac{z^2}{2R} + C_2 \frac{(TR-z)^2}{2R}$  minimum when  $Z=TR \frac{C_2}{C_1+C_2}$ .

Model.4. Generalisation of 3, interval between runs  $t$ , not fixed. Now  $F(t,z) =$

$$\text{Total cost/unit time} = \frac{1}{t} \left[ \frac{C_1 z^2}{2R} + \frac{C_2 (tR-z)^2}{2R} \right] + \frac{C_3}{t} \text{ minimum, when}$$

$$z = tR \frac{C_2}{C_1+C_2} \text{ and } t = \sqrt{\frac{2C_3(C_1+C_2)}{C_1 R C_2}} ; q = Rt = \sqrt{\frac{2C_3 R}{C_1} \frac{C_1+C_2}{C_2}}$$

5: A Model with stochastic demand, with withdrawal from stock continuous interval between runs given. (Say week) and  $c_1$  = unit inventory carrying cost/week and  $c_2$  = unit shortage cost/week. But the demand  $x$  is random with  $p(x)$   $x = 0, 1, 2, \dots$ , to find the quantity  $q$  to be stocked to minimise expected cost.

$$\text{Cost} = c(q, x) = \begin{cases} \frac{c_1 q^2}{2x} + \frac{c_2 (x-q)^2}{2x} & \text{if } x > q \\ c_1 (q - \frac{x}{2}) & \text{if } x \leq q \end{cases}$$

Expected cost  $E(q)$  is min., for the smallest  $q$  satisfying

$$\sum_0^{\infty} p(x) + (q + \frac{1}{2}) \sum_{j+1}^{\infty} p(\frac{x}{2}) > c_1/c_1+c_2$$

(For Numerical Examples: See Ackoff et al or Sasieni et al).

Dynamic Programming (Simple Examples)

D.P. is a new technique by R Bellman to solve multistage allocation problems, by reducing a problem in  $n$  stages into  $n$  problems in single stages using the principle of optimality: 'whatever be the state of first decision, the remaining decisions must be optimal wrt the state resulting from the first decision.'

1. The problem of optimum distribution of effect: To allocate a given amount  $A$  of resource or capital amidst  $n$  sectors or regions so as to maximize the total return  $\sum_1^n f_i(x_i)$   $f_i(x_i)$  - return by allocating  $x_i \geq 0$  to the  $i$ th sector  $x_1 + x_2 + \dots + x_n = A$ .

Numerical Example of distribution of 12 sales men among 3 areas

$x_1$	0	1	2	3	4	5	6	7	8	9	10	11	12
$f_1(x_1)$	38	41	48	60	66	72	83	96	102	100	95	89	82
$f_2(x_2)$	40	42	50	58	66	75	82	88	95	99	100	100	100
$f_3(x_3)$	60	64	68	78	90	102	109	119	124	125	125	125	125

Answer  $x_1 = 7, x_2 = 0, x_3 = 5$ . Return = 238

Compute successively

$$F_2(X) = \max_{0 \leq x_1 \leq A} [f_1(x_1) + f_2(X-x_1)] \quad X = 0, 1, 2, A$$

$$F_3(X) = \max_{0 \leq x_3 \leq A} [f_3(x_3) + F_2(X-x_3)] \quad \dots$$

$$F_n(X) = \max_{0 \leq x_n \leq A} [f_n(x_n) + F_{n-1}(X-x_n)]$$

2. Ware house problem.

A man is engaged in buying and selling identical items, each of which requires considerable storage space. He operates a ware house of capacity 500 items. He can order on the 15th day of each month, at prices shown below for delivery on the first of following month. During a month, he can sell any amount up to total stock on hand in the begining of the month at market prices given below. If he starts the year with 200 items in stock, how much should be plan to purchase and sell each month to maximize his profit.

Month	i	J	F	M	A	M	J	J	A	S	O	N	D
Cost price	$c_i$	150	155	165	160	160	160	155	150	155	150	150	150
Selling price	$p_i$	165	165	165	175	170	155	155	155	160	170	175	170

- Let  $x_i$  = amount to be sold during the ith month
- $y_i$  = " " ordered on 15th of ith month
- $s_i$  = stock at the beginning of ith month
- H = ware house capacity.

Total profit =  $\sum (p_i x_i - c_i y_i)$  max. subject to

$$0 \leq x_i \leq S_i \quad \text{and} \quad S_i \leq H \quad \text{where} \quad S_i = S_{i-1} + y_{i-1} - x_{i-1}$$

This is a problem in L.P. involving 34 original variable and 24 stocks which can be solved more easily by Dynamic programmy.

Let  $f_n(S)$  = max. profit when stock level is s and n months to go

$$f_n(S) = \max_{\substack{0 \leq x \leq S \\ 0 \leq y \leq H+x}} [p_{12-n} x - c_{12-n} y + f_{n-1}(S+y-x)]$$



The corners of the admissible region are  $(0,0)$ ,  $(S,0)$ ,  $(S,H)$ ,  $(0,H-S)$ .

We get  $f_1(S_{12}) = 170 S_{12}$ .  $f_2(S_{11}) = 175 S_{11} + 20 H$ ,  $f_3(S_{10}) = 170 S_{10} + 40 H$ .  
 ...  $f_{12}(S_1) = 165 S_1 + 130 H = \underline{98,000}$ . The solution is

x	300	500	500	500	500	0	0	y	500	500	500
y	500	500	500	500	0	0	$y \leq 500$	500	500	500	0

The optimal policy is of the form: Do nothing for  $k$  periods ( $k = 0$  here). Then oscillate between a full and empty warehouse.

Waiting Lines or Queues

(1) input-mechanism (2) service mechanism (3) queue discipline. These are the three structural features which specify a queue.

M/M/1. Poisson input with mean arrival rate  $\lambda$ . Exponential service mean service rate  $\mu$ . One counter. Probability that are  $n$  people in system at 't' =  $P_n(t)$

$$P_n(t+dt) = P_{n+1}(t)\mu dt + P_{n-1}(t)\lambda dt + P_n(t)(1-\lambda dt-\mu dt) + o(dt)$$

$$\frac{dP_n(t)}{dt} = \mu P_{n+1}(t) + \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t).$$

Steady state probabilities  $p_n = \lim_{t \rightarrow \infty} P_n(t)$  exist iff  $\lambda < \mu$  and are given by

$$p_{n+1} = (1-\rho)p_n - \rho p_{n-1}, \quad p_1 = \rho p_0, \quad \rho = \frac{\lambda}{\mu}. \quad p_n = \rho^n \cdot p_0. \quad \sum p_n = 1 = p_0(1-\rho).$$

$E(n) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$ . This is the expected number of units in the system

(either in service or waiting). The queue length  $m$  is zero if  $n = 0, 1$  and  $= n - 1$  if  $n > 1$ .  $E(m) = \sum_{n=2}^{\infty} (n-1)p_n = \sum_{n=1}^{\infty} np_n - p_1 - \sum_{n=2}^{\infty} p_n = E(n) - (1-p_0) = \bar{E}(n) - \rho$ .

Distribution of waiting time 'w', of a newly arrived unit.

$$P(w = 0) = P(n=0) = 1-\rho. \quad \text{For } w > 0. \quad p(\text{waiting time} > w) = \frac{\lambda}{\mu} e^{-(\mu-\lambda)w}.$$

Total time spent in the system 'v' = waiting time and service time =  $w+u$ .

This is an exponential distribution ( $P(\text{total time} > v) = e^{-(\mu-\lambda)v}$ ).

$$E(v) = E(w) + \frac{1}{\mu} = \frac{1}{\mu-\lambda} \quad E(w) = \frac{\lambda}{\mu(\mu-\lambda)}.$$

## Problems in Queues

1. Arrivals at a telephone booth are considered to be Poisson, with an average time of 10 minutes between one arrival and the next. The length of a phone call is assumed to be distributed exponentially, with mean 3 minutes. (a) What is the average length of the queues that form from time to time? (b) What is the probability that a person arriving at the booth will have to wait? (c) What is the probability that an arrival will have to wait more than 10 minutes before the phone is free? (d) The telephone company will install a second booth when convinced that an arrival would expect to have to wait at least 3 minutes for the phone. By how much should the flow of arrivals increase in order to justify a second booth? (e) Estimate the fraction of a day that the phone will be in use.
2. At what average rate must a clerk at a supermarket work in order to insure a probability of 0.90 that the customer will not have to wait longer than 12 minutes? It is assumed that there is only one counter, to which customers arrive in a Poisson fashion at an average rate of 15 per hour. The length of service by the clerk has an exponential distribution.
3. A repairman is to be hired to repair machines which break down at an average rate of 3 per hour. Breakdowns are distributed in time in a manner that may be regarded as Poisson. Non-productive time on any one machine is considered to cost the company \$5 per hour. The company has narrowed the choice down to 2 repairmen, one slow but cheap, the other fast but expensive. The slow cheap repairman asks \$3 per hour; in return, he will service broken-down machines exponentially at an average rate of 4 per hour. The fast expensive repairman demands \$5 per hour, and will repair machines exponentially at an average rate of 6 per hour. Which repairman should be hired?

4. A supermarket has two girls ringing up sales at the counters. If the service time for each customer is exponential with mean 4 minutes, and if people arrive in a Poisson fashion at the counter at the rate of 10 an hour. (a) What is the probability of having to wait for service ? (b) What is the expected percentage of idle time for each girl?
5. An insurance company has 3 claims adjusters in its branch office. People with claims against the company are found to arrive in a Poisson fashion at an average rate of 20 per 8 hour day. The amount of time than an adjuster spends with a claimant is found to have an exponential distribution with mean service time 40 minutes. Claimants are processed in the order of their appearance. (a) How many hours a week can an adjuster expect to spend with claimants. (b) How much time, on the average, does a claimant spend in the branch office ?
6. At a certain airport it takes exactly 5 minutes to land an aeroplane, once it is given the signal to land. Although incoming planes have scheduled arrival times, the wide variability in arrival times produces an effect which makes the incoming planes appear to arrive in a Poisson fashion at an average rate of 6 per hour. This produces occasional stack-ups at the airport which can be dangerous and costly. Under these circumstances, how much time will a pilot expect to spend circling the field waiting to land ?
7. A hospital clinic has a doctor examining every patient brought in for a general checkup. The doctor averages 4 minutes on each phase of the checkup although the distribution of time spent on each phase is approximately exponential. If each patient goes through four phases in the checkup and if the arrivals of the patients to the doctor's office are approximately Poisson at an average rate of 3 per hour, what is the average time spent by a patient waiting in the doctor's

office ? What is the average time spent in the examination ? What is the most probable time spent in the examination ?

8. A warehouse in a small state receives orders for a certain item and sends them by truck as soon as possible to the customer. The orders arrive in a Poisson fashion at a mean rate of 0.9 per day. Only one item at a time can be shipped by truck from the warehouse, which is located in the central part of the state. Because the customers are located in various places in the state, the distribution of service time in days has a distribution with probability distribution  $4te^{-2t}$ . What is the expected delay between the arrival of an order and the arrival of the item to the customer ? Service time is defined here as the time the truck takes to load, get to the customer, unload, and return to the warehouse. Loading and unloading times are small compared with travel time.
9. The arrival distribution of ships in a harbour is Poisson with mean rate  $\lambda$  ships per week, and service-time distribution is exponential with mean rate of  $\mu$  unloadings per week. The queue discipline is : First come. First served. The ships have to wait till the dock is free. If the harbour is to accomodate the ships routed to it, it must be able to handle them at least as fast as they arrive, i.e., the utilization factor  $\rho = \lambda/\mu$  must be less than unity. The problem is to decide how much less than unity  $\rho$  should be made in order to balance the cost occasioned by the delay of shipping with the cost of running the dock. We assume that there is only one unloading dock in the harbour, and the cost per unloading operation is proportional to the speed of service i.e.  $= D/\mu$  where D is the cost per operation per service rate of operating the dock. Also suppose that the average cost of having one ship idle in the port, (the cost of crew, overhead etc.)

is proportional to the mean waiting time in port, i.e.,  $= CW$  where  $C$  is the cost of ship delay per week,  $W$  is the mean waiting time including service, expressed in weeks.

Given  $\lambda, C, D$ , the problem is to determine the value of  $\mu$ , (or  $\rho$  as  $\lambda$  is given) so that the total average cost per ship unloaded (which is the sum of the above two costs) is minimum.

You will get a simple expression for  $\mu$  in terms of  $\lambda, C, D$ . From this formula, show that we may conclude as follows.

(1) If the crew cost  $C$  is small compared to the dock cost  $D$ , the optimum service rate need not be much larger than arrival rate i.e., we should keep dock utilization high ( $\rho$  near unity) at the expense of ship delay.

(2) But if the crew cost  $C$  is large compared to the dock cost  $D$ , the optimum value of  $\mu$  would be considerably larger than  $\lambda$  i.e., we could afford to have our efficient dock idle most of the time in order to reduce costly ship delays.

### Replacement

#### Items that deteriorate

$A$  is the initial cost of machine,  $C_1, C_2, \dots, C_n$  costs incurred during the 1st, 2nd, ...,  $n^{\text{th}}$  periods. Interest rate 100  $i\%$  per period

$v = 1/1+i$ . Find the value of  $n$ , which minimises total costs.

The discounted value  $K_n$  of all future costs associated with a policy of replacing after each  $n$  periods =  $[A + \sum C_i v^{i-1}] [1 + v^n + v^{2n} + \dots]$

$$K_n < K_{n-1} \text{ if } C_n < K_{n-1}(1-v) = \frac{(A + C_1) + C_2 v + \dots + C_{n-1} v^{n-2}}{1 + v + v^2 + \dots + v^{n-2}}$$

This gives the following rule of determining  $n$ . Do not replace if the next period is less than the weighted average of previous costs. Replace if the next period is greater than the weighted average of previous costs.

Items that fail

Given the life table of an item, we start with items & replace the items as they fail, at the end of  $t$  periods we adopt a group replacement policy. To determine the optimum value of  $t$ .  $C_1$  cost of individual replacement of  $C_2$  - unit cost of group replacement  $f(x)$  = no. of failures in the  $x^{\text{th}}$  period.

Total cost per period =  $K(t)/t$  where  $K(t) = N C_1 + C_2 \sum_{i=1}^{t-1} f(x)$ .

We have to prepare a table of  $K(t)/t$  for  $t = 1, 2, \dots$  and find its minimum

$t$	1	2	3	...	$K$
probability of failure	$p_1$	$p_2$	$p_3$	...	$p_r$
number of replacements	$f(1)$	$f(2)$	...		$f(K+1)$

$$f(0) = N. \quad f(1) = N p_1 \quad f(2) = N p_2 + f(1) p_1$$

$$f(3) = N p_3 + f(1) p_2 + f(2) p_1 \quad \& \text{ so on.}$$

$f(x)$  usually increase, then decrease & again increase & decrease and it will oscillate until it settles down to a stable value  $N/E(\xi)$  where  $E(\xi) = \text{Expected life} = \sum i p_i$

Replacement Problems

1.(a) A truck owner finds from his past records that the costs per year of running a truck whose purchase price is \$ 6000 are as given below:

Year	1	2	3	4	5	6	7	8
Running costs	1000	1200	1400	1800	2300	2800	3400	4000
Resale price	3000	1500	750	375	200	200	200	200

At what age is a replacement due ?

- (b) The same truck owner has 3 trucks, 2 of which are two year old and the third one year old. He is considering a new type of truck with 50% more capacity than one of the old ones at a unit price of \$ 8000. He estimates that the running costs and resale price for the new truck will be as follows :

Year	1	2	3	4	5	6	7	8
Running costs	1200	1500	1800	2400	3100	4000	5000	6100
Resale price	4000	2000	1000	500	300	300	300	300

Assuming that the loss of flexibility due to fewer trucks is of no importance, and that he will continue to have sufficient work for three of the old trucks, what should he do ?

2. A manufacturer is offered two machines A and B. A is priced at \$ 5000, and running costs are estimated at \$ 800 for each of the first 5 years, increasing by \$ 200 per year in the sixth and subsequent years. Machine B, which has the same capacity as A, costs \$ 2500 but will have running costs of \$ 1200 per year for 6 years, increasing by \$ 200 per year thereafter. If money is worth 10% every year, which machine should be purchased ? (Assume that the machines will eventually be sold for scrap at a negligible price).
3. The following failure rates have been observed for a certain type of light bulb.

<u>end of week</u>	<u>prob. of failure to date</u>
1	0.05
2	0.13
3	0.25
4	0.43
5	0.68
6	0.88
7	0.96
8	1.00

The cost of replacing an individual failed bulb is \$ 1.25. The decision is made to replace all bulbs simultaneously at fixed intervals, and also to replace individual bulbs as they fail in service. If the cost of group replacement is 30 cents per bulb, what is the best interval between group replacements ? At what group replacement price per bulb would a policy of strictly individual replacement become preferable to the adopted policy ?

### Monte Carlo Methods

1. A bakery delivers fresh bread to one of its retail stores every day. The number of loaves delivered each day is not constant, but has the following distribution :

Loaves per day :	10	11	12	13	14	15	16
Probability :	0.05	0.10	0.20	0.30	0.20	0.10	0.05

The number of customers desiring bread each day has the distribution :

No. of customers:	5	6	7	8	9	10
Probability :	0.10	0.15	0.20	0.40	0.10	0.05

Finally, the probability that a customer in need of bread wants 1, 2 or loaves is described by

Loaves to a customer :	1	2	3
Probability :	0.40	0.40	0.20

Estimate by Monte Carlo methods the average number of loaves of bread left over per day, and the average number of sales lost per day owing to lack of bread. Assume that left-over bread is given away at the end of each day.

2. Buses are scheduled to pass a certain corner every 15 minutes but actually the arrival of a bus varies normally about its scheduled arrival time, with a standard deviation of 3 minutes.



Passengers arrive in a Poisson fashion with mean arrival rate of 4 persons per hour, and the number of empty seats on the bus has a Poisson distribution with mean  $3/2$ . No standing is permitted. Use Monte Carlo methods to find the average waiting time of an arrival.

3. In a queueing system customers arrive in a Poisson fashion every 20 minutes on the average. Each customer proceeds in turn through 3 servicing stations, in a prescribed order. Service times in 3 stages are distributed in the following way:

Stage I : Normal, mean 10 minutes, S.D. 5 minutes.

Stage II : Exponential, with  $\lambda = 1/15$  service per minute.

Stage III : Service time constant, at 15 minutes.

Using Monte Carlo methods, find the expected time in the system and the expected time spent waiting, for a typical arrival.

000SA.5 SCS 61.5



C4721

