AD-754 577

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS

Larry H. Crow

Army Materiel Systems Analysis Agency
Aberdeen Proving Ground, Maryland

April 1972

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

AD 754577

AD

AMSAA

TECHNICAL REPORT NO. 56

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS

LARRY H. CROW

APRIL 1972

Approved for public release;
distribution unlimited.

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY
Aberdeen Proving Ground, Maryland

43

DISPOSITION

Destroy this report when no longer needed. Do not return it to the originator.

DISCLAIMER

The findings in this report are not to be construed as an official Department of the Army position.

WARNING

Information and data contained in this document are based on the input available at the time of preparation. The results may be subject to change and should not be construed as representing the AMC position unless so specified.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) U.S. Army Materiel Systems Analysis Agency Aberdeen Proving Ground, Maryland		2a. REPORT SECURITY CLASSIFICATION Unclassified
		2b. GROUP
3. REPORT TITLE ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS		
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)		
5. AUTHOR(S) (First name, middle initial, last name) Larry H. Crow		
6. REPORT DATE April 1972	7a. TOTAL NO. OF PAGES 46	7b. NO. OF REFS 3
8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report No. 56	
b. PROJECT NO. AMCMS Code 2270.1031		
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
d.		
10. DISTRIBUTION STATEMENT		
11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY U.S. Army Materiel Command Washington, D.C.	

13. ABSTRACT

This report discusses three sampling procedures available for estimating convex combinations of proportions and illustrates the feasibility and practical usefulness of these procedures. The first procedure is an analogue to simple random sampling from a single population. The remaining two procedures are infinite population analogues to optimum and proportional sampling from a single finite population. The practical aspects of selecting a procedure are discussed and numerical examples are given.

DD FORM 1 NOV 65 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

Unclassified
Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Estimation Convex combinations Proportions Allocation Missile testing						

II

Unclassified

Security Classification

TECHNICAL REPORT NO. 56

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS

Larry H. Crow

April 1972

AMCMS Code 2270.1051

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY
ABERDEEN PROVING GROUND, MARYLAND

/

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY

TECHNICAL REPORT NO. 56

LHCrow/dga
Aberdeen Proving Ground, Md.
April 1972

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS

ABSTRACT

This report discusses three sampling procedures available for estimating convex combinations of proportions and illustrates the feasibility and practical usefulness of these procedures. The first procedure is an analogue to simple random sampling from a single population. The remaining two procedures are infinite population analogues to optimum and proportional sampling from a single finite population. The practical aspects of selecting a procedure are discussed and numerical examples are given.

Next page is blank.

CONTENTS

		Page
	ABSTRACT	3
1.	INTRODUCTION AND SUMMARY	7
2.	NOTATION AND NOMENCLATURE	7
3.	RANDOM ALLOCATION	8
4.	OPTIMUM AND PROPORTIONAL ALLOCATION	17
5.	SELECTION OF A SAMPLING PROCEDURE	24
6.	NUMERICAL EXAMPLES	25
	REFERENCES	43
	DISTRIBUTION LIST	44

Next page is blank.

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS

1. INTRODUCTION AND SUMMARY

Interest in the subject of this report arose recently from an investigation of appropriate testing plans for the Improved Hawk and SAM-D missiles. The testing objectives for the two missile systems are similar and involve hypothesis testing on the true value of a certain convex combination of proportions. Aside from missile testing, however, parameters of this type are of general practical interest since they occur naturally in a variety of statistical problems. The purpose of this report is to discuss sampling techniques available for estimating convex combinations of proportions and to illustrate the feasibility and practical usefulness of these techniques.

Specifically, three sampling procedures are discussed in Sections 3 and 4. The first procedure, it will be shown, may be considered as an analogue to simple random sampling from a single population. The remaining two procedures are infinite population analogues to optimum and proportional sampling from a single finite population. (See Reference 1, Sections 5.3 and 5.5.)

In Section 5 the practical aspects of selecting a particular sampling procedure will be discussed and in Section 6 numerical examples of these procedures will be given.

2. NOTATION AND NOMENCLATURE

A *dichotomus population* is defined to be a population (finite or infinite) in which a proportion p', say, of the items have a certain characteristic and a proportion $1-p'$ of the items do not have it. Throughout this report it will be assumed that there are k infinite dichotomus populations denoted symbolically by $P_1,\ldots,P_K$. Also, if an item drawn at random from one of these populations has the characteristic of interest then it will be said that a *success* has been realized.

Let E and VAR denote expectation and variance, respectively, and let X_i be the random variable

$$X_i = \begin{cases} 1 & \text{if an item drawn at random from } P_i \text{ results in a success} \\ 0 & \text{otherwise} \end{cases}$$

$i=1,\dots,k$. It is assumed that $\text{Prob}(X_i=1)=p_i$, where $0 \le p_i \le 1$, $i=i,\dots,k$.

Observe, also, that $E(X_i)=p_i$, and $VAR(X_i)=p_i(1-p_i)$, $i=1,\dots,k$.

If f_i, $i=1,\dots,k$, are numbers such that $0 \le f_i$ and

$$1= \sum_{i=1}^{k} f_i,$$

then

$$p= \sum_{i=1}^{k} f_i p_i$$

is a *convex combination* of $p_1,\dots,p_k$. The results derived in the remainder of this paper are based on the assumption that the f_i's are known constants, the p_i's are unknown parameters and that p is the parameter to be estimated.

3. RANDOM ALLOCATION

In considering an estimate of a parameter the experimenter, of course, is mainly interested in how much confidence he has that the estimate will be close to the true value of the parameter. It would seem, therefore, that to have a high degree of confidence in an estimate of p would require that samples be taken from each of the k populations. If this is true, then to obtain a "good" estimate of p, when k is large, may be a very formidable task, indeed. Fortunately, however, this is not the case.

A procedure shall be presented in this section that will enable an experimenter to estimate p without necessarily having to sample from each of the populations. This procedure is not detrimental to the interest of the experimenter since the precision and all other statistical properties of the proposed estimator depend only on the sample size and not on the number of populations. Other desirable properties of this estimator will, also, be presented in the sequel.

Exact distribution theory is derived in this section, too, that will allow one to construct operating characteristic curves to test hypothesis on the true value of p. Also, the analogy of this procedure to simple random sampling from a single dichotomus population shall be discussed.

The procedure is defined next.

Definition 3.1

Let n be the total sample size. For the r^{th} observation, $r=1,\ldots,n$, choose one of the populations $P_1,\ldots,P_k$, with the probability of selecting P_i equal to f_i, $i=1,\ldots,k$. Let P_{i_r} denote the population selected. Next choose an item at random from P_{i_r} and observe whether or not a success was obtained. This procedure is called *random allocation*.

The analogy of this procedure to simple random sampling is of particular interest. Let P be a dichotomus population and suppose that the proportion of items in P that have the characteristic of interest varies. Specifically, assume that at any given instant the probability is f_i that the proportion of items in P with the characteristic of interest is p_i, $i=1,\ldots,k$. Hence, if one draws an item at random from P then the probability is

$$p=\sum_{i=1}^{k} f_i p_i$$

that a success is observed. One may, therefore, simply assume that P is a dichotomus population with the proportion of items with the characteristic of interest equal to p. Thus, random allocation, from this point of view, merely describes the process of taking a random sample of size n from P.

A statistical analysis of this procedure will now be developed.

Let $\bar{X}_r$ be the random vector

$$\bar{X}_r=(X_{1r},\ldots,X_{kr}),$$

where

$$X_{ir}=\begin{cases}1 & \text{if } r^{th} \text{ item is drawn from } P_i \text{ and a success is obtained}\\ 0 & \text{otherwise}\end{cases}$$

$i=1,\ldots,k, r=1,\ldots,n.$

Since selecting a population is statistically independent from drawing an item it follows that

$$\text{Prob}(X_{ir}=1)=f_i p_i,$$

$i=1,\ldots,k, r=1,\ldots,n.$ Now, let

$$Y_r=\sum_{i=1}^{k} X_{ir},$$

$r=1,\ldots,n$, and note that the events that determine $\bar{X}_r$ are disjoint. It follows then that

$$\text{Prob}(Y_r=1)=\sum_{i=1}^{k} f_i p_i,$$

$r=1,\ldots,n.$ That is, $Y_1,\ldots,Y_n$ are independent Bernoulli random variables with mean p. Observe, also, that

$$Y_r = \begin{cases} 1 & \text{if } r^{th} \text{ observation results} \\ & \text{in a success} \\ 0 & \text{otherwise} \end{cases}$$

$r=1,\ldots,n$.

The maximum likelihood estimate (MLE) of p will be given in the next theorem. It will be shown later that the MLE is, in fact, the "best" estimate of p when data arise from random allocation.

Theorem 3.2

For a total sample of size n the MLE of p from random allocation is

$$p_{ran} = \sum_{r=1}^{n} y_r/n. \tag{3.1}$$

i.e., the MLE of p is the mean number of successes in the sample.

Proof

It is straightforward to show that the likelihood function for the experiment is

$$L(\bar{X}_1,\ldots,\bar{X}_n) = \prod_{r=1}^{n} \prod_{i=1}^{k} \left(f_i p_i\right)^{X_{ir}} \left(1-p\right)^{1-\sum_{i=1}^{k} X_{ir}}. \tag{3.2}$$

Now, note that $f_i p_i$ may be written as $c_i p$, $i=1,\ldots,k$, where

$$\sum_{i=1}^{k} c_i = 1.$$

Hence, (3.2) becomes

$$L(\bar{X}_1,\ldots,\bar{X}_n)=\prod_{r=1}^{n}\prod_{i=1}^{k}\left(c_i p\right)^{X_{ir}}\left(1-p\right)^{1-\sum_{i=1}^{k} X_{ir}} . \tag{3.3}$$

Thus,

$$\log L(\bar{X}_1,\ldots,\bar{X}_n)=\sum_{r=1}^{n}\sum_{i=1}^{k} X_{ir}\log c_i p + n\log(1-p)$$

$$-\sum_{r=1}^{n}\sum_{i=1}^{k} X_{ir}\log(1-p).$$

Setting the derivative of $\log L(\bar{X}_1,\ldots,\bar{X}_n)$ equal to 0, it follows that

$$\sum_{r=1}^{n}\sum_{i=1}^{k} X_{ir}/p - \left(n - \sum_{r=1}^{n}\sum_{i=1}^{k} X_{ir}\right)\Big/\left(1-p\right) = 0.$$

Therefore,

$$p_{ran}=\sum_{r=1}^{n}\sum_{i=1}^{k} X_{ir}/n.$$

Since

$$Y_r=\sum_{i=1}^{k} X_{ir},$$

$r=1,\ldots,n$, the theorem is proved.

A result concerning the distribution of np_{ran} will be given next. Notice that this result does *not* depend on k.

Theorem 3.3

The distribution of np_{ran} is Binomial with parameters (n,p).

$$\text{i.e., } \mathrm{Prob}(p_{ran}=m)=\binom{n}{m}p^m(1-p)^{n-m},$$

$m=0,1,\ldots,n.$

Proof

Since $Y_1,\ldots,Y_n$ are independent Bernoulli random variables with mean p then

$$\sum_{r=1}^{n} Y_r$$

is a Binomial random variable with parameters (n,p). The theorem is proved because

$$np_{ran} = \sum_{r=1}^{n} Y_r.$$

Some optimum properties of p_{ran} are given in the following theorem.

Theorem 3.4

The MLE p_{ran} of p has the following optimum properties:

(i) sufficient (i.e., uses all the information available from random allocation in estimating p).

(ii) unbiased (i.e., the expectation of p_{ran} is p).

(iii) efficient (i.e., has the smallest variance among all unbiased estimates of p obtained from random allocation).

(iv) strongly consistent (i.e., p_{ran} will always converge to p as n becomes large).

(v) converges in probability (i.e., $\text{Prob}(|p_{ran}-p|<\varepsilon) \rightarrow 1$ as $n \rightarrow \infty$, for any $\varepsilon>0$).

Proof

(i) By (3.3) the likelihood function may be written as

$$L(\bar{X}_1,\ldots,\bar{X}_n)=A\cdot B$$

where

$$A=\prod_{i=1}^{k} c_i^{np_{ran}}$$

and

$$B=p^{np_{ran}}(1-p)^{n-np_{ran}}.$$

This factorization is a necessary and sufficient condition to insure that p_{ran} be sufficient. (See Reference 2, Section 5.5.1.)

(ii) This follows directly from Theorem 3.3.

(iii) One can show that the likelihood function, given by (3.3), implies that the information, $I(p)$, in the sample (Reference 2, 6.1.3) equals

$$n[p(1-p)]^{-1}.$$

The Cramer-Rao lower bound for an unbiased estimator of p (Reference 2, 6.1.3) is

$$[I(p)]^{-1} = n^{-1}p(1-p).$$

But by Theorem 3.3

$$VAR(p_{ran})=n^{-1}p(1-p).$$

That is, VAR(p_{ran}) equals the lower bound. This gives the desired result.

(iv) Since $Y_1,\ldots,Y_n$ are independent, identically distributed random variables the strong law of large numbers (Reference 3, 2c.3) implies that as n gets large,

$$\sum_{r=1}^{n} Y_r/n$$

converges to $E(Y_1)$. But

$$E(Y_1)=p.$$

(v) Strong convergence, given by (iv), implies convergence in probability (Reference 3, 2c.2).

The fact that p_{ran} converges in probability does not, in itself, reflect the precision the experimenter may expect in this estimator. That is, it does not tell him for a sample of size n what degree of confidence he has that

$$|p_{ran}-p|$$

is small. The next theorem attempts to answer this question. It is important to note that this result is independent of k, the number of populations.

Theorem 3.5

For any $\epsilon>0$ and for any $n\geq 1$

$$\text{Prob}(|p_{ran}-p|<\epsilon) \geq 1-1/4n\epsilon^2.$$

Proof

By the Chebyshev Inequality (Reference 3, 2c.2)

$$\text{Prob}(|p_{ran}-p|\geq\epsilon) \leq \text{VAR}(p_{ran})/\epsilon^2.$$

But

$$\mathrm{VAR}(p_{ran}) = p(1-p)/n \leq 1/4n,$$

since $0 \leq p \leq 1$. Thus,

$$\mathrm{Prob}(|p_{ran}-p| \geq \varepsilon) \leq 1/4n\varepsilon^2$$

or equivalently,

$$\mathrm{Prob}(|p_{ran}-p| < \varepsilon) \geq 1-1/4n\varepsilon^2.$$

Another result, independent of k, which is of practical significance for large sample size is the following.

Theorem 3.6

Let n be the total sample size and let $q=1-p$. Then the asymptotic distribution of

$$\frac{p_{ran}-p}{\left(\frac{pq}{n}\right)^{1/2}}$$

is Normal (0,1), as $n \to \infty$.

Proof

This is a direct result of the Central Limit Theorem (Reference 3, 2c.5).

Remark

Whereas Definition 3.1 specifies that an item be drawn immediately after the population is selected this is not, in practice, necessary. In practice random sample sizes $n_1,\ldots,n_k$, say, may be determined first, where n_i is the number of items to be drawn from P_i, $i=1,\ldots,k$, and

$$n = \sum_{i=1}^{k} n_i.$$

After these sample sizes are determined, samples are drawn from the respective populations and p_{ran} is obtained.

4. OPTIMUM AND PROPORTIONAL ALLOCATION

If independent random samples are taken from each of the k populations it seems natural to estimate p_i by the mean number of successes in the sample from P_i, $i=1,\dots,k$. Moreover, the convex combination of these estimates would, also, seem the natural function to use as an estimate of p since the weights $f_1,\dots,f_k$ are assumed known. This approach to estimating p is considered in this section.

For a fixed total sample size n the question arises as to how one should assign the sample sizes $n_1,\dots,n_k$, where n_i is the size of the sample to be drawn from P_i, $i=1,\dots,k$. This section discusses two procedures for determining these sample sizes so as to obtain estimators of p which have certain desirable properties. It is to be remarked that these procedures are analogus to optimum and proportional sampling methods from a finite population.

Let $\hat{p}_i$ be the mean number of successes in a sample of size n_i taken from P_i, $i=1,\dots,k$. The corresponding estimator for p is then

$$\hat{p}=\sum_{i=1}^{k} f_i\hat{p}_i.$$

Now, $\hat{p}_i$ is the MLE of p_i, $i=1,\dots,k$. Further, $\hat{p}_i$ is an unbiased estimator of p_i, $i=1,\dots,k$, and, also, enjoys having the smallest variance of all such estimators. In general, *unbiasness* and *minimum variance* are very desirable properties for an estimator to possess. By virtue of the fact that $\hat{p}_i$ is an unbiased estimator it follows that $\hat{p}$ is, also, an unbiased estimator since

$$E(\hat{p})=\sum_{i=1}^{k} f_i E(\hat{p}_i)=p.$$

Observe, also, that

$$\mathrm{VAR}(\hat{p}) = \sum_{i=1}^{k} f_i^2\,\mathrm{VAR}(\hat{p}_i) = \sum_{i=1}^{k} f_i^2 p_i(1-p_i)/n_i. \qquad (4.1)$$

From the above, it is seen that the variance of $\hat{p}$ depends directly on the sample sizes $n_i, i=1,\ldots,k$, which are chosen by the experimenter. The problem of allocating the n_i to insure that $\hat{p}$ will have the smallest possible variance will be considered next.

Theorem 4.1

For a total sample size n, $VAR(\hat{p})$ is minimized if

$$n_i = w_i n,$$

where

$$w_i = \frac{f_i[p_i(1-p_i)]^{1/2}}{\sum_{j=1}^{k} f_j[p_j(1-p_j)]^{1/2}}, \tag{4.2}$$

$i=1,\ldots,k.$

Proof

Minimizing $VAR(\hat{p})$, given by (4.1), subject to $n_1+\ldots+n_k=n$ is equivalent to minimizing

$$\sum_{i=1}^{k} f_i^2 p_i(1-p_i)/n_i + \lambda(n_1+n_2+\cdots+n_k-n) \tag{4.3}$$

where λ is a Lagrange multiplier. Taking the derivative of (4.3) with respect to n_i and setting this equal to 0 yields

$$-f_i^2 p_i(1-p_i)/n_i^2 + \lambda = 0. \tag{4.4}$$

Thus,

$$n_i = f_i[p_i(1-p_i)]^{1/2}/\lambda^{1/2},\ i=1,\ldots,k. \tag{4.5}$$

Since

$$n = \sum_{j=1}^{k} n_j$$

Equation 4.5 implies that

$$\lambda^{1/2} = \sum_{j=1}^{k} f_j [p_j(1-p_j)]^{1/2}/n.$$

Hence, w_i, given by (4.2), is immediate for $i=1,\ldots,k$.

Definition 4.2

Let n be the total sample size. When $n_i=w_in$, w_i given by (4.2), $i=1,\ldots,k$, the procedure is called *optimum allocation* and $\hat{p}$ is denoted by p_{opt}.

The weights w_i may or may not be known in practice. If they are not known then, of course, optimum allocation cannot be employed. Another allocation method which is always feasible in practice is defined next.

Definition 4.3

Let n be the total sample size. When $n_i=f_in$, $i=1,\ldots,k$, the procedure is called *proportional allocation* and $\hat{p}$ is denoted by p_{prop}.

In the next section a comparison of the three allocation procedures presented in this report will be made. This comparison will be based on a few important factors which, it is felt, should be taken into account when choosing among these procedures. One point of interest will, of course, be the relative precision of the corresponding estimators.

Now, by the way optimum allocation was defined it is known that

$$\mathrm{VAR}(p_{opt}) \leq \mathrm{VAR}(p_{prop}).$$

The relationship between the variances of p_{prop} and p_{ran} will be considered next.

Theorem 4.4

Let n be the total sample size. Then

$$VAR(p_{prop}) = \sum_{i=1}^{k} f_i p_i (1-p_i)/n. \tag{4.6}$$

Proof

This result follows easily since

$$VAR(p_{prop}) = \sum_{i=1}^{k} f_i^2 p_i (1-p_i)/n_i,$$

and

$$n_i = f_i n, i=1,\ldots,k.$$

Theorem 4.5

$$VAR(p_{prop}) \leq VAR(p_{ran})$$

Proof

Let $n \geq k$ be the total sample size. By (4.6)

$$VAR(p_{prop}) = \sum_{i=1}^{k} f_i p_i (1-p_i)/n$$

and by Theorem 3.3

$$VAR(p_{ran}) = p(1-p)/n,$$

where, of course,

$$p=\sum_{i=1}^{k} f_i p_i .$$

Observe, now, that

$$\mathrm{VAR}(p_{prop})=\sum_{i=1}^{k} f_i p_i (1-p+p-p_i)/n$$

$$=\sum_{i=1}^{k} f_i p_i (1-p)/n+\sum_{i=1}^{k} f_i p_i (p-p_i)/n$$

$$=\mathrm{VAR}(p_{ran})+\sum_{i=1}^{k} f_i p_i (p-p_i)/n.$$

It will next be shown that

$$\sum_{i=1}^{k} f_i p_i (p-p_i)\leq 0. \tag{4.7}$$

To show this, consider a random variable Z which assumes the values $p_1,\ldots,p_k$, and suppose that

$$\mathrm{Prob}(Z=p_i)=f_i,$$

$i=1,\ldots,k$. Recall here that

$$\sum_{i=1}^{k} f_i=1.$$

Observe, now, that

$$\sum_{i=1}^{k} f_i p_i (p-p_i) = \left(\sum_{i=1}^{k} f_i p_i\right)^2 - \sum_{i=1}^{k} f_i p_i^2$$

$$= - \mathrm{VAR}(Z).$$

Since VAR(Z) is nonnegative (4.7) is established and, hence, the theorem is proved.

The next theorem gives two asymptotic properties of p_{opt} and p_{prop} as the total sample size increases.

Theorem 4.6

The estimators p_{opt} and p_{prop} of p have the following properties:

(i) strongly consistent (i.e., $p_{opt}(p_{prop})$ will always converge to p as n becomes large).

(ii) converges in probability (i.e., $\mathrm{Prob}(|\hat{p}-p|<\varepsilon) \rightarrow 1$ as $n \rightarrow \infty$, for any $\varepsilon>0$, $\hat{p}=p_{opt},p_{prop}$).

Proof

The proof is similar to the one given for Theorem 3.4 (iv) and (v) and is, therefore, omitted.

In Theorem 3.5 a result was given that, in practice, would aid an experimenter to assess the confidence he may place in p_{ran} as an estimator of p for a certain sample size n. The key to the result was the fact that

$$\mathrm{VAR}(p_{ran}) = p(1-p)/n \leq 1/4n.$$

This inequality is, in fact, sharp; i.e., it cannot be improved on. In this section it has been shown that

$$\mathrm{VAR}(p_{opt}) \leq \mathrm{VAR}(p_{prop}) \leq \mathrm{VAR}(p_{ran}),$$

with equality if $p_1 = p_2 = \ldots = p_k$.

Consequently, 1/4n is an upper bound on the variances of p_{opt} and p_{prop} and the bound is, also, sharp. The significance of the following theorem is, therefore, established.

Theorem 4.7

For any $\varepsilon>0$ and for any $n\geq k$

$$Prob(|\hat{p}-p|<\varepsilon) \geq 1-1/4n\varepsilon^2,$$

$$\hat{p}=p_{opt}, p_{prop}.$$

Proof

The proof is omitted because of its similarity to the proof of Theorem 3.5.

Another result which may be useful when the sample size n is large will be given next.

Theorem 4.8

Let n be the total sample size. Then the asymptotice distribution of

$$\frac{\hat{p}-p}{[VAR(\hat{p})]^{1/2}}$$

is Normal(0,1), as $n\rightarrow\infty$, where $\hat{p} = p_{opt}, p_{prop}$.

Proof

This result follows immediately from the Central Limit Theorem (Reference 3, 2c.5).

5. SELECTION OF A SAMPLING PROCEDURE

In this section some of the practical aspects of choosing among the three sampling procedures presented in this report shall be discussed. This discussion may, also, serve as a brief review of some of the basic properties of these procedures.

It is difficult, if not impossible, to consider all possible situations an experimenter may face when selecting a sampling procedure. It may, also, be difficult to ascertain which procedure is the "best" to use in practice for a particular situation. It may happen in practice that a procedure is "best" only because of its computational simplicity or possibly because the theoretical aspects are easy to explain to ones employer. The following discussion, however, is general in nature and is intended to serve only as a guide when selecting a procedure.

Regardless of any other considerations, the experimenter must usually make at least two basic decisions before he selects a sampling procedure. Firstly, he must decide whether the chief purpose of the experiment is to estimate p or to test a hypothesis on the true value of p. Secondly, he must decide whether or not the sample size n will be larger than k, the number of populations. The choice of a sampling procedure for these four basic situations shall be discussed.

Suppose first that it is decided that $k>n$. Immaterial of whether the experimental goal is estimation or hypothesis testing, the only appropriate procedure to use is random allocation. Hypothesis tests are based on the statistic np_{ran}, which by Theorem 3.3, has a Binomial distribution.

Consider, now, the case when $k \leq n$ and the purpose of the experiment is to estimate p. For most problems encountered in practice it happens that

$$VAR(p_{opt}) < VAR(p_{prop}) < VAR(p_{ran}).$$

Since the estimator with the smallest variance is generally considered to be best it follows that optimum allocation should be used if possible. The weights $w_i, i=1,\ldots,k$, given by (4.2) may not, of course, be known in practice. If this is the case then proportional allocation is the appropriate sampling procedure.

Finally, assume that $k \leq n$ and the experimenter wishes to test a hypothesis on the true value of p. In this case a statistic with a known distribution that depends on p must be available. Distributions, in a closed form, of statistics obtained from optimum and proportional allocation are not known to the author. However, by Theorem 3.3, the distribution of np_{ran} is Binomial with parameters (n,p). It appears then that random allocation is generally the appropriate procedure to use under these circumstances. It is, also, remarked that if n is very large relative to k then the distributions of p_{opt} and p_{prop}, properly normalized, are approximately normal with mean zero and variance one. Unfortunately, the normalizing factors are functions of the individual $p_i, i=1,\dots,k$, and not explicitly a function of p. It is further noted that if n is very large relative to k, then proportional allocation and random allocation are approximately the same procedure. That is, the sample sizes to be drawn from the individual populations are approximately equal.

6. NUMERICAL EXAMPLES

In this section several numerical examples shall be given which will demonstrate the practical significance of the three sampling procedures presented in this paper. Before these examples are given, it should be pointed out that if one is sampling from a dichotomus population and the probability of a success is equal to a fixed constant for all trials, then the population is infinite. This observation will be useful in the following examples.

Example 1

At the end of the development period of a certain surface-to-air missile system, the Army wishes to conduct tests that will demonstrate, at an appropriate significance level, a particular performance requirement of the system. This requirement is not concerned with the detection of threatening aircraft but only with the performance of the missile from launch to detonation of the warhead.

This missile is designed to destroy air-breathing aircraft and several factors such as target speed and range, use of electronic counter measures and others are assumed to affect this capability. These factors, taken together in various combinations yield a total of 52 possible states that a threat may assume. In the event of an enemy attack the probability that a missile will encounter an aircraft presenting a particular threat is known. However, it is not known what the probability is that a missile will destroy an aircraft given that the aircraft is presenting a certain threat.

The unconditional probability that a missile will destroy an aircraft is called the *expected kill probability* (EKP) and it is required that testing demonstrate that this probability is sufficiently high before the production phase of the missile system is begun.

Suppose that the 52 possible states a threatening aircraft can assume are numbered and let f_i be the known probability of state i, i=1,...,52. Further, let p_i be the unknown probability that a missile will destroy an aircraft imposing a threat in the i^{th} state, i=1,...,52. From this notation it follows that the EKP is

$$p=\sum_{i=1}^{52} f_i p_i .$$

Because of cost only 40 missiles can be allocated for the testing. Also, since 40, the sample size, is less than 52, the number of possible threats, and hypothesis testing is the goal, then the random allocation procedure is appropriate. In general terms the null hypothesis, H_N, and the alternate hypothesis, H_A, are defined as:

H_N: EKP is acceptable

H_A: EKP is not acceptable.

The decision to accept or reject H_N is to be based on the statistic $40p_{ran}$, which, by Theorem 3.3, has a Binomial distribution with parameters (40,p). It is decided that if $40p_{ran}$ is at least 25 then H_N is accepted while if $40p_{ran}$ is less than 25 then H_N is rejected in favor of H_A. The corresponding operating characteristic (OC) curve, which is based on Binomial probabilities, is given in Figure 6.1.

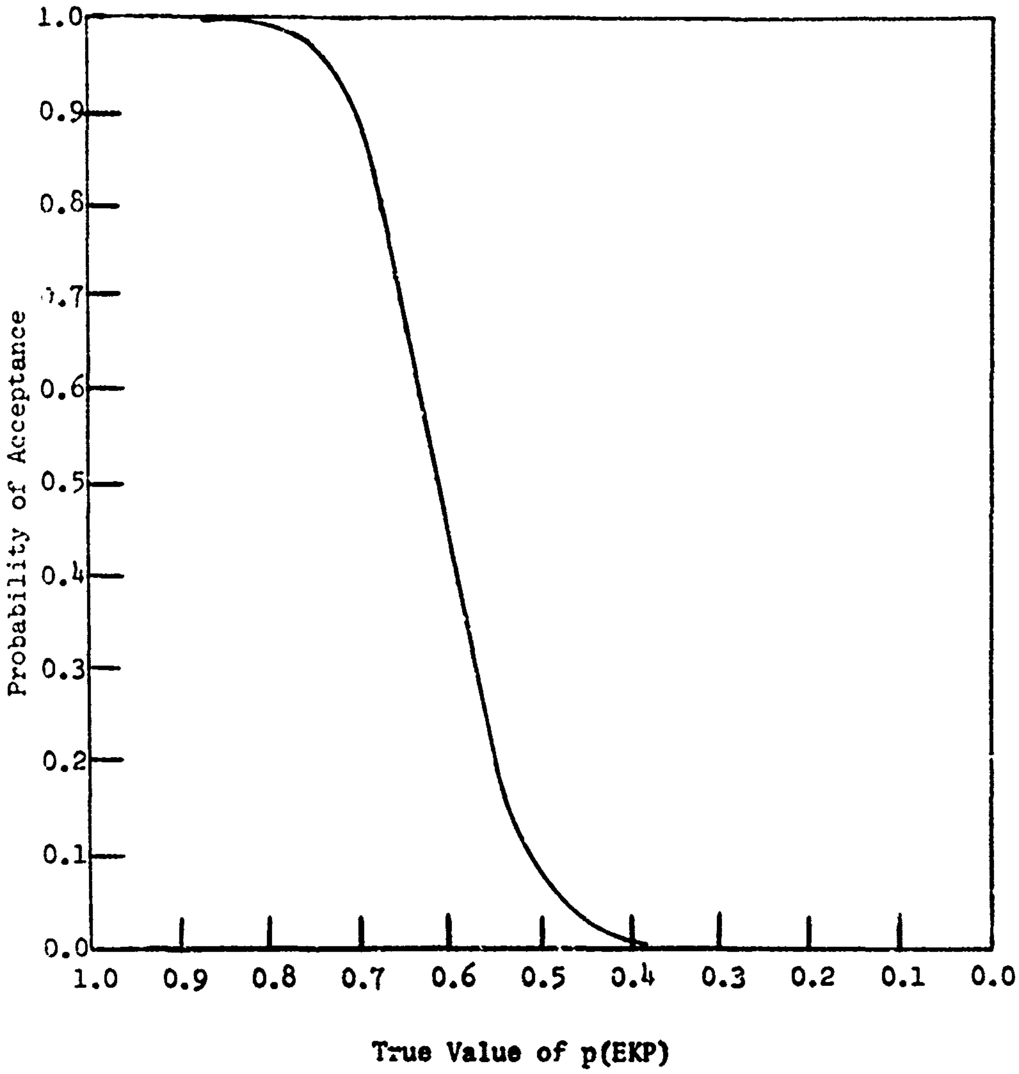

Figure 6.1 **Operating Characteristic Curve for a Sample Size Equal to 40 and an Allowable Number of Failures Equal to 15**

For a particular value of p on the abscissa of this curve the resulting value on the ordinate is

$$\mathrm{Prob}(np_{ran} \geq 25 | p) = \sum_{j=25}^{40} \binom{40}{j} p^j q^{40-j},$$

where $q = 1-p$.

For purposes of a numerical example the missile tests were simulated on a computer using the values of the f_i's and p_i's given in Table 6.1 (Note that in practice the f_i's are known but the p_i's are unknown). The simulation produced first the number of missiles to be fired at aircraft presenting the different threats. These numbers, denoted by n_i, $i=1,\ldots,52$, are given in Table 6.1. Next the missile firings were simulated and S_i, $i=1,\ldots,52$, given in Table 6.1, are the number of successful firings. This simulation yielded a value of 30 for $40p_{ran}$. Therefore, the null hypothesis, H_N, should be accepted implying that the EKP is acceptable.

TABLE 6.1 INFORMATION AND SIMULATED DATA PERTAINING TO HYPOTHESIS TEST ON THE EKP

Threat i	Frequency f_i	Destroy Probability p_i	Sample Size n_i	Successful Firings S_i
1	.0576	.72	3	3
2	.0144	.60	1	1
3	.0384	.60	1	0
4	.0384	.72	2	1
5	.0096	.60	0	0
6	.0256	.60	2	2
7	.0096	.55	2	2
8	.0064	.55	0	0
9	.01728	.72	2	1
10	.00432	.60	0	0
11	.06912	.60	2	0
12	.01152	.72	1	1
13	.00288	.60	0	0
14	.04608	.60	2	2
15	.01728	.55	1	0
16	.01152	.55	0	0
17	.0192	.72	0	0
18	.0048	.60	0	0
19	.0128	.72	0	0
20	.0032	.60	0	0
21	.0480	.72	3	3
22	.0120	.60	0	0
23	.0120	.60	0	0
24	.1440	.72	4	4
25	.0360	.60	2	2
26	.0360	.60	1	1
27	.0030	.55	0	0
28	.0090	.55	0	0

TABLE 6.1 (CONTINUED)

Threat i	Frequency f_i	Destroy Probability p_i	Sample Size n_i	Successful Firings S_i
29	.0396	.72	2	2
30	.0099	.60	0	0
31	.0044	.60	0	0
32	.1188	.55	6	2
33	.0297	.55	1	1
34	.0132	.72	1	1
35	.0011	.60	0	0
36	.0033	.72	0	0
37	.0020	.72	0	0
38	.0005	.55	0	0
39	.0060	.72	0	0
40	.0015	.55	0	0
41	.0016	.72	0	0
42	.0004	.69	0	0
43	.0144	.72	0	0
44	.0036	.60	0	0
45	.0016	.72	0	0
46	.0004	.60	0	0
47	.0144	.72	1	1
48	.0036	.60	0	0
49	.0008	.72	0	0
50	.0002	.55	0	0
51	.0072	.72	0	0
52	.0018	.55	0	0
	$EKP = \sum_{i=1}^{52} f_i p_i = 0.643883$		$n = \sum_{i=1}^{52} n_i = 40$	$np_{ran} = \sum_{i=1}^{52} S_i = 30$

Example 2

Consider again the missile testing problem discussed in Example 1. Suppose now that instead of hypothesis testing the Army wishes to estimate the EKP from data obtained from random allocation. How large should the sample size n be in order for the Army to have an a priori probability of at least 0.95 that the estimator p_{ran} will differ from the true EKP, p, by less than 0.1? That is, what value of n will insure that

$$\text{Prob}(|p_{ran} - p| < 0.1) \geq 0.95?$$

It is known from Theorem 3.3 that np_{ran} has a Binomial distribution with parameters (n,p). Hence,

$$\text{Prob}(|p_{ran} - p| \leq 0.1) = \text{Prob}(|np_{ran} - np| \leq 0.1n)$$

$$= \sum_{|x-np| \leq 0.1n} \binom{n}{x} p^x q^{n-x}$$

where

$q = 1-p$ and the summation is extended over the values of x for which $|x - np| \leq 0.1n$. One can see the difficulty in attempting to determine directly an n so that no matter what p is the inequality

$$\sum_{|x - np| \leq 0.1n} \binom{n}{x} p^x q^{n-x} \geq 0.95$$

holds. By using the results of Theorem 3.5 one may, however, easily find such an n. Now, by this theorem

$$\text{Prob}(|p_{ran} - p| < 0.1) \geq 1-1/(0.1)^2\, 4n.$$

Therefore, if

$$1-1/(0.1)^2\, 4n \geq 0.95$$

then it follows that

$$n \geq 25/0.05 = 500.$$

Consequently, if 500 or more missiles are fired then the a priori probability is at least 0.95 that p_{ran} will differ from the (unknown) EKP, p, by less than 0.1.

Inasmuch as the above result is correct, it will next be shown that a much smaller sample size n will, also, suffice. The derivation of the smaller sample size will be based on the fact, given in Theorem 3.6, that p_{ran} is approximately normally distributed if n is sufficiently large. From practice, it has been observed that a sample size of 30 or more is generally sufficiently large to insure that this approximation is adequate. Hence, if $n \geq 30$ then

$$\text{Prob}(|p_{ran} - p| < 0.1)$$

$$= \text{Prob}\left(\frac{-0.1}{\sigma} < \frac{p_{ran}-p}{\sigma} < \frac{0.1}{\sigma}\right)$$

$$\doteq \frac{1}{\sqrt{2\pi}} \int_{-0.1/\sigma}^{0.1/\sigma} e^{-t^2/2}\, dt$$

where

$$\sigma = \left(pq/n\right)^{1/2}$$

and "$\doteq$" denotes "approximately equal to."

From a table of cumulative standard normal probabilities one finds that if

$$0.1/\sigma \geq 1.96$$

then

$$\frac{1}{\sqrt{2\Pi}} \int_{-0.1/\sigma}^{0.1/\sigma} e^{-t^2/2}\, dt \geq 0.95.$$

Hence, if

$$n \geq (19.6)^2/4 \doteq 96$$

then

$$\text{Prob}(|p_{ran} - p| < 0.1) \geq 0.95$$

regardless of the unknown value of p. That is, a sample size of 96, instead of 500, is large enough to obtain the required precision.

This estimation problem was simulated on a computer using the values for the f_i's and p_i's given in Table 6.1 and a total sample size of $n = 96$. In Table 6.2 n_i is the number of missiles to be fired at aircraft presenting the i^{th} threat and S_i is the number of successful firings, $i=1,\ldots,52$. From Table 6.1 the true value of the EKP is

$$p = 0.643883$$

and from Table 6.2 the simulated estimate of p is

$$p_{ran} = 0.614583.$$

TABLE 6.2 SIMULATED DATA PERTAINING TO THE ESTIMATION OF THE EKP FROM RANDOM ALLOCATION

Threat i	Sample Size n_i	Successful Firings S_i	Threat i	Sample Size n_i	Successful Firings S_i
1	5	4	27	1	0
2	2	1	28	1	0
3	6	5	29	3	2
4	2	1	30	2	2
5	0	0	31	0	0
6	4	3	32	9	7
7	2	1	33	4	2
8	0	0	34	1	0
9	0	0	35	0	0
10	2	0	36	0	0
11	4	3	37	0	0
12	1	1	38	0	0
13	0	0	39	0	0
14	4	2	40	0	0
15	2	1	41	0	0
16	2	0	42	0	0
17	4	3	43	1	1
18	0	0	44	1	0
19	1	1	45	0	0
20	0	0	46	0	0
21	2	2	47	2	1
22	1	0	48	0	0
23	0	0	49	0	0
24	15	7	50	0	0
25	4	2	51	1	1
26	6	5	52	1	1
				$n = \sum_{i=1}^{52} n_i$	$P_{ran} = \sum_{i=1}^{52} S_i / n$
				$= 96$	$= 0.614583$

It has been noted before, and is well worth mentioning again, that the statistical properties of p_{ran} are independent of k, the number of dichotomus populations. This means that a sample size of 40 (96) would still yield the same OC curve (precision) in Example 1 (2) even if the number of possible threats was not 52 but *any* arbitrarily large number. Observe, also, that the number of threat states actually sampled from was only 20 in Example 1 and only 31 in Example 2.

Example 3

Suppose a small, inexpensive item is produced by a certain factory for the Army and the Army wishes to estimate the fraction of defective items produced from a total sample size of 1000. Three assembly lines contribute to the factory's production of the item and because of differences in machinery and personnel the fractions of defective items produced by the assembly lines differ. However, after production the items are stored together and are not distinguished as to the particular assembly line which produced them.

Ordinarily, a random sample of 1000 items would be taken from the storage area and the fraction of defective items produced by the factory would be estimated by the mean number of defective items in the sample. If the proportions f_1, f_2, f_3, say, of the items produced by Assembly Lines 1, 2 and 3, respectively, are known then the random allocation procedure may, also, be used to estimate the fraction of defective items produced. Any additional effort or cost it may take to implement the random allocation procedure would probably not be justified since, in Section 3, it was noted that this procedure and the random sampling procedure are statistically equivalent. Thus, it is immaterial which of the two procedures is used because the respective estimators afford the same precision.

Assuming that the f_i's are known the Army may, also, use the proportional allocation procedure since 1000, the sample size, is greater than 3, the number of assembly lines. This procedure, it has been pointed out, generally yields a more precise estimator than does random allocation. Because the f_i's are easily obtainable from the manufacturer the Army decided to use the proportional allocation procedure and take advantage of this extra precision in estimating the fraction of defective items produced.

This estimation problem was simulated on a computer when Assembly Line 1 produces items 5 percent defective and accounts for 25 percent of the production; Assembly Line 2 produces items 4 percent defective and accounts for 35 percent of the production, while Assembly Line 3 produces items 3 percent defective and accounts for 40 percent of the production. Of course, the fraction of defective items produced by each assembly line would be unknown in practice.

The simulation results are given in Table 6.3, where n_i is the number of items to be drawn from Assembly Line i and S_i is the number of defective items in the sample, i=1, 2, 3. The actual fraction, p, say, of defective items produced by the factory is

$$p = 0.0385.$$

The estimate of p, resulting from the simulation, is

$$p_{prop} = 0.0330.$$

TABLE 6.3 INFORMATION AND SIMULATED DATA PERTAINING TO THE ESTIMATION OF THE FRACTION DEFECTIVE FROM PROPORTIONAL ALLOCATION

Assembly Line i	Proportion of Production f_i	Fraction of Items Defective p_i	Sample Size n_i	Number of Defectives S_i
1	0.250	0.050	250	7
2	0.350	0.040	350	16
3	0.400	0.030	400	10
	$p = \sum_{i=1}^{3} f_i p_i = 0.0385$		$n = \sum_{i=1}^{3} n_i = 1000$	$p_{prop} = \sum_{i=1}^{3} f_i S_i / n_i = 0.0330$

Example 4

Consider again the estimation problem discussed in Example 3. Suppose the manufacturer informs the Army that from past experience with the manufacturing of similar type items the production pro esses of the three assembly lines usually exhibit a certain relationship. Particularly, the variation of the production process for Assembly Line 2 is generally about 20 percent less than the variation of the production process for Assembly Line 1. Also, the variation of the production process for Assembly Line 3 has been observed to usually be about 40 percent less than the corresponding variation for Assembly Line 1. The differences in the age and design of the machinery in the three assembly lines were given as the primary reason for the differences in the production processes. With this information the Army may use the optimum allocation procedure to estimate the fraction of defective items produced by the factory and, thereby, obtain the best possible precision.

Again let f_i be the proportion of the production contributed to Assembly Line i and, also, let p_i be the fraction of the items produced by Assembly Line i which are defective, i=1, 2, 3. For a total sample size n it follows from Theorem 4.1 that the sample size n_i to be taken from Assembly Line i is

$$n_i = w_i n, \tag{6.1}$$

where

$$w_i = \frac{f_i[p_i(1-p_i)]^{1/2}}{\sum_{j=1}^{3} f_j[p_j(1-p_j)]^{1/2}}$$

i = 1, 2, 3.

From remarks made in Section 2 it is known that if

$$U_i = \begin{cases} 1 & \text{if an item drawn at random from Assembly Line } i \text{ is defective} \\ 0 & \text{otherwise} \end{cases}$$

then $E(U_i) = p_i$ and $VAR(U_i) = p_i(1-p_i)$, $i=1, 2, 3$.

Hence,

$$w_i = \frac{f_i \ [VAR(U_i)]^{1/2}}{\sum_{j=1}^{3} f_j \ [VAR(U_j)]^{1/2}}$$

$i = 1, 2, 3$.

Also, from the information provided by the manufacturer it is known that $VAR(U_2)$ is approximately equal to 0.8 $VAR(U_1)$ and $VAR(U_3)$ is approximately equal to 0.6 $VAR(U_1)$. Regarding these as actual equalities one has that

$$n_2 = \frac{f_2}{f_1} (0.8)^{1/2} n_1 \tag{6.2}$$

and

$$n_3 = \frac{f_3}{f_1} (0.6)^{1/2} n_1 . \tag{6.3}$$

Therefore, for a total sample size of n it follows that

$$n = [1 + \frac{f_2}{f_1} (0.8)^{1/2} + \frac{f_3}{f_1} (0.6)^{1/2}] n_1 .$$

Consequently,

$$n_1 = n/[1 + \frac{f_2}{f_1} (0.8)^{1/2} + \frac{f_3}{f_1} (0.6)^{1/2}]. \tag{6.4}$$

If, in fact

$$VAR(U_2) = 0.8\ VAR(U_1) \tag{6.5}$$

and

$$VAR(U_3) = 0.6\ VAR(U_1) \tag{6.6}$$

then the sample sizes given by (6.2) - (6.4) are optimum for a fixed total sample size n. If the variances of U_2 and U_3 are given only approximately by the right-hand sides of (6.5) and (6.6), respectively, then the sample sizes n_1, n_2 and n_3 are, of course, only *nearly* optimum.

For a total sample size of n = 1000 this estimation problem was simulated on a computer using the same f_i's and p_i's given in Example 3. The exact relationships between the variances of U_1, U_2 and U_3 are determined from p_1 p_2 and p_3 and are:

$$VAR(U_2) = 0.808\ VAR(U_1) \tag{6.7}$$

and

$$VAR(U_3) = 0.613\ VAR(U_1). \tag{6.8}$$

For the simulation the sample sizes n_1, n_2 and n_3 were determined assuming that the relationships (6.5) and (6.6) hold. The resulting sample sizes are:

$$n_1 = 286; \tag{6.9}$$

$$n_2 = 359; \tag{6.10}$$

$$n_3 = 355. \tag{6.11}$$

These sample sizes are only *nearly* optimum. The *true* optimum sample sizes, based on Equation (6.1), are:

$$n_1 = 285;$$
$$n_2 = 358;$$
$$n_3 = 357.$$

Clearly, then, the sample sizes given by (6.9) - (6.11) may, for all practical purposes, be considered optimum.

The simulation results are given in Table 6.4, where n_i is the number of items to be drawn from Assembly Line i and S_i is the number of defective items in the sample, i=1, 2, 3. From Table 6.3 the true fraction of defective items produced by the factory is

$$p = 0.0385.$$

The estimate of p, obtained from this simulation, is

$$p_{opt} = 0.0335.$$

TABLE 6.4 SIMULATED DATA PERTAINING TO THE ESTIMATION OF THE FRACTION DEFECTIVE FROM OPTIMUM ALLOCATION

Assembly Line i	Sample Size n_i	Number of Defectives S_i
1	286	10
2	359	15
3	355	9
	$n = \sum_{i=1}^{3} n_i$	$p_{opt} = \sum_{i=1}^{3} f_i \, S_i/n_i$
	$= 1000$	$= 0.0335$

Next page is blank.

REFERENCES

1. Cochran, W. G. (1953). Sampling Techniques, Wiley, New York.

2. Lindgren, B. W. (1962). Statistical Theory, The MacMillan Company, New York.

3. Rao, C. R. (1968). Linear Statistical Inference and its Applications, Wiley, New York.

Next page is blank.

AD ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS

U.S. Army Materiel Systems Analysis Agency (AMSAA), Aberdeen Proving Ground, Maryland 21005, AMSAA Technical Report No. 56, Report UNCLASSIFIED

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS, Larry H. Crow, AMSAA; AMCMS Code No. 2270.1031, April 1972

Trace Index Terms

Estimates of convex combinations
Random allocation estimator
Optimum allocation estimator
Proportional allocation estimator

1. Estimates of convex combinations 2. Random allocation estimator 3. Optimum allocation estimator 4. Proportional allocation estimator

---

AD Crow, Larry H.

U.S. Army Materiel Systems Analysis Agency (AMSAA), Aberdeen Proving Ground, Maryland 21005, AMSAA Technical Report No. 56, Report UNCLASSIFIED

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS, Larry H. Crow, AMSAA; AMCMS Code No. 2270.1031, April 1972

Trace Index Terms

Estimates of convex combinations
Random allocation estimator
Optimum allocation estimator
Proportional allocation estimator

1. Estimates of convex combinations 2. Random allocation estimator 3. Optimum allocation estimator 4. Proportional allocation estimator

---

AD Estimates of convex combinations

U.S. Army Materiel Systems Analysis Agency (AMSAA), Aberdeen Proving Ground, Maryland 21005, AMSAA Technical Report No. 56, Report UNCLASSIFIED

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS, Larry H. Crow, AMSAA; AMCMS Code No. 2270.1031, April 1972

Trace Index Terms

Estimates of convex combinations
Random allocation estimator
Optimum allocation estimator
Proportional allocation estimator

1. Estimates of convex combinations 2. Random allocation estimator 3. Optimum allocation estimator 4. Proportional allocation estimator

---

AD Random allocation estimator

U.S. Army Materiel Systems Analysis Agency (AMSAA), Aberdeen Proving Ground, Maryland 21005, AMSAA Technical Report No. 56, Report UNCLASSIFIED

ESTIMATION OF A CONVEX COMBINATION OF PROPORTIONS, Larry H. Crow, AMSAA; AMCMS Code No. 2270.1031, April 1972

Trace Index Terms

Estimates of convex combinations
Random allocation estimator
Optimum allocation estimator
Proportional allocation estimator

1. Estimates of convex combinations 2. Random allocation estimator 3. Optimum allocation estimator 4. Proportional allocation estimator