The Design and Analysis of Pattern Recognition Experiments

By W. H. HIGHLEYMAN

(Manuscript received March 2, 1961)

A popular procedure for testing a pattern recognition machine is to present the machine with a set of patterns taken from the real world. The proportion of these patterns which are misrecognized or rejected is taken as the estimate of the error probability or rejection probability for the machine. In Part I, this testing procedure is discussed for the cases of unknown and known a priori probabilities of occurrence of the pattern classes. The differences between the tests that should be made in the two cases are noted, and confidence intervals for the test results are indicated. These concepts are applied to various published pattern recognition results by determining the appropriate confidence interval for each result.

In Part II, the problem of the optimum partitioning of a sample of fixed size between the design and test phases of a pattern recognition machine is discussed. One important nonparametric result is that the proportion of the total sample used for testing the machine should never be less than that proportion used for designing the machine, and in some cases should be a good deal more.

PART I — ON ANALYSIS

INTRODUCTION

There are two distinct and consecutive processes usually involved in the feasibility study of a pattern recognition method or machine. The first process is the actual design of the machine. This might be based upon a set of sample patterns which the experimenter has gathered, from which he estimates the parameters of the machine. Alternatively, the experimenter may base his design on some *a priori* knowledge concerning the pertinent characteristics of the pattern classes under study. The second process is then the testing of this machine either in its hardware form or by its simulation on a general purpose computer. A differ-

ent set of sample patterns from that used in the design is used in this stage.

The popular procedure for interpreting the test results is to take the proportion of patterns in the test data which have been misrecognized or rejected by the machine as the estimates of the error probability and rejection probability, respectively, for the machine. There are several questions which might be raised concerning this testing procedure, such as:

1. Are these estimates the best estimates?
2. If so, how good are these estimates?
3. How does the estimate improve as the sample size is increased?

Questions such as these are discussed in Part I of this paper. Two cases are considered; one is the case in which the *a priori* probabilities of class occurrence are unknown, and the other case assumes full knowledge of the *a priori* probabilities.

Case 1. Unknown a priori Probabilities — Random Sampling

Let the number of allowable pattern classes be c. It will be assumed that, for each allowable class i, there exists an *a priori* probability of occurrence ω_i, a probability of error e_i, and a probability of rejection r_i. (For the rest of this paper, the term "error" will refer to an undetected error; all detected errors will be assumed to be rejected.) These probabilities are unknown to the experimenter, who is interested in estimating the overall probability of error for the machine.

$$e = \sum_{i=1}^{c} \omega_i e_i , \tag{1}$$

and the over-all probability of rejection,

$$r = \sum_{i=1}^{c} \omega_i r_i . \tag{2}$$

Let him perform the following experiment, which will be called random sampling. Consider the patterns to be randomly generated by a "pattern source" according to the *a priori* probabilities of occurrence. He takes a pattern from the source, identifies it, and then lets his pattern recognition machine attempt identification. He notes which of the three possible outcomes occurs: correct recognition, misrecognition, or rejection. This experiment is repeated n times, resulting in m_e samples which have been misrecognized and m_r samples which have been rejected.

Since these outcomes are mutually exclusive, and each experiment independent, then the resulting random variables, m_e and m_r, clearly

are distributed according to the multinomial probability distribution. That is, the joint probability distribution of m_e and m_r, $P(m_e, m_r)$, is given by

$$P(m_e, m_r) = \binom{n}{m_e m_r} e^{m_e} r^{m_r} (1 - e - r)^{n - m_e - m_r}. \tag{3}$$

The maximum-likelihood estimates for e and r, denoted by $\hat{e}$ and $\hat{r}$, are then[1]

$$\hat{e} = \frac{m_e}{n}, \qquad \hat{r} = \frac{m_r}{n}, \tag{4}$$

which are the estimates in common use. Further, each of these estimates is proportional to a single random variable having a binomial distribution; therefore, $n\hat{e}$ and $n\hat{r}$ are themselves binomially distributed. The mean value of each estimate is the parameter for which it is an estimate; the variance of each is[1]

$$\sigma_{\hat{e}}^2 = \frac{1}{n^2} \sigma_{m_e}^2 = \frac{e(1 - e)}{n} \tag{6}$$

$$\sigma_{\hat{r}}^2 = \frac{r(1 - r)}{n}. \tag{7}$$

Because it is known that $n\hat{e}$ and $n\hat{r}$ are binomially distributed, confidence intervals can be applied to these estimates.* These confidence intervals require rather involved computations, but fortunately have been plotted for several values of n by various people.[3,4] In Fig. 1 is shown such a plot of intervals for a 95 per cent confidence level computed by C. S. Clopper and E. S. Pearson. The use of this graph is fairly simple. A vertical line extended upward from the observed value of the estimate given on the abscissa will intersect the pair of curves pertaining to the particular sample size used. Projecting these two intersections horizontally onto the ordinate axis gives an interval for the parameter being estimated. The probability is 0.95 that the interval drawn in this manner includes the parameter. For instance, if a sample size of $n = 250$ yielded 50 errors, then the estimate of the probability of error is 0.20. Using Fig. 1 it can be stated that, with probability 0.95, the true probability of error is included in the interval from 0.15 to 0.27.

* Mattson[2] has used a similar argument for determining convergence of an adaptive system. However, he used Tchebycheff's inequality to obtain confidence intervals which are necessarily larger than if he had used such intervals pertaining to the binomial distribution.

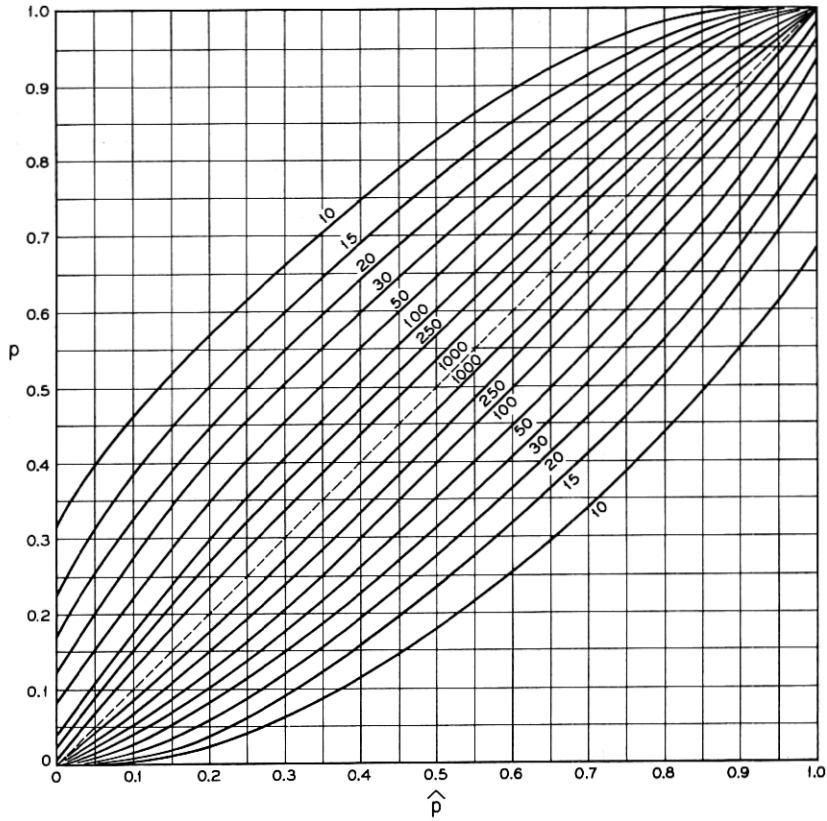

Fig. 1 — 95 per cent confidence intervals for a binomially distributed variable.

Case 2. Known a priori Probabilities — Selective Sampling

It is now assumed that the *a priori* probability of occurrence for each class, ω_i, is known. To take advantage of this knowledge, the experimenter takes n_i samples from each class i such that

$$\frac{n_i}{n} = \omega_i, \tag{8}$$

where n is the total number of samples. This process will be referred to as selective sampling.* (It will be assumed that the ω_i are such that (8) can be fulfilled with the desired sample size, n.)

* This sort of sampling dichotomy has been previously noted by others. For instance, Bowley[5] and Neyman[6] have referred to these two methods as "unrestricted" and "stratified" sampling.

The machine is again allowed to attempt recognition of these patterns, resulting in m_{e_i} samples from class i being misrecognized, and m_{r_i} samples from class i being rejected.

For any class i, the joint probability distribution for m_{e_i} and m_{r_i} again is multinomial:

$$P(m_{e_i}, m_{r_i}) = \left(m_{e_i}{}^{n_i} m_{r_i}\right) e_i^{m_{e_i}} r_i^{m_{r_i}} (1 - e_i - r_i)^{n_i - m_{e_i} - m_{r_i}}. \quad (9)$$

Since each of these distributions is independent of the others in this experiment, then the joint probability of the outcome for all c classes is the product of the individual probabilities (9):

$$P(m_{e_1}, \cdots, m_{e_c}, m_{r_1}, \cdots, m_{r_c}) = \prod_{i=1}^{c} \left(m_{e_i}{}^{n_i} m_{r_i}\right) e_i^{m_{e_i}} r_i^{m_{r_i}} (1 - e_i - r_i)^{n_i - m_{e_i} - m_{r_i}}. \quad (10)$$

This is no longer a multinomial probability distribution. However, since the maximum-likelihood estimate of a sum of independent variables is the sum of the maximum-likelihood estimates, then these estimates for e and r are

$$\hat{e} = \frac{\sum_{i=1}^{c} m_{e_i}}{n}, \quad (11)$$

$$\hat{r} = \frac{\sum_{i=1}^{c} m_{r_i}}{n}, \quad (12)$$

which again agree with the popular practice of using the proportions as estimates. The random variables of which $n\hat{e}$ and $n\hat{r}$ are values are not now binomially distributed, since a sum of binomially distributed variables is not itself a binomial distribution in general.

The mean of each estimate is again the particular parameter being estimated. The variance of each of these estimates can be computed:

$$\sigma_{\hat{e}}'^2 = \frac{1}{n^2} \sum_{i=1}^{c} \sigma_{m_{e_i}}'^2 = \frac{1}{n^2} \sum_{i=1}^{c} n_i e_i (1 - e_i) = \frac{1}{n} \sum_{i=1}^{c} \omega_i e_i (1 - e_i), \quad (13)$$

in which use of (8) is made, and the prime distinguishes this variance from that for random sampling. Similarly,

$$\sigma_{\hat{r}}'^2 = \frac{1}{n} \sum_{i=1}^{c} \omega_i r_i (1 - r_i). \quad (14)$$

It is of interest to compare these variances for selective sampling with those obtained for the case of random sampling. Since the variance

for $\hat{r}$ has the same form as $\hat{e}$ in both cases, it is necessary to consider only one of them, say $\hat{e}$. First note that $\sigma_{\hat{e}}^2$ can be written, using (1) and (6), as

$$\sigma_{\hat{e}}^2 = \frac{1}{n}\left(\sum_{i=1}^{c} \omega_i e_i\right)\left(1 - \sum_{k=1}^{c} \omega_k e_k\right). \tag{15}$$

From (13),

$$\sigma_{\hat{e}}^2 - \sigma_{\hat{e}}'^2 = \frac{1}{n}\sum_{i=1}^{c} \omega_i e_i^2 - \frac{1}{n}\left(\sum_{i=1}^{c} \omega_i e_i\right)^2. \tag{16}$$

Noting that $\sum_{i=1}^{c} \omega_i = 1$, (16) can be written as

$$\sigma_{\hat{e}}^2 - \sigma_{\hat{e}}'^2 = \frac{1}{n}\sum_{i=1}^{c} \omega_i \left(e_i - \sum_{k=1}^{c} \omega_k e_k\right)^2 = \frac{1}{n}\sum_{i=1}^{c} \omega_i (e_i - e)^2 = \sigma_e^2 \geqq 0. \tag{17}$$

Hence, the variance in the case of random sampling is greater than the variance in the case of selective sampling, the difference being what might be interpreted as the variance of the class errors. That is, if e_i is treated as a random variable with probability distribution ω_i, then σ_e^2 is the variance of e_i. (A similar derivation holds for the variance of the rejection probability estimates.) That the selective sampling variance should be smaller than the random sampling variance might be expected, since in selective sampling more information is used, namely the *a priori* probabilities.

Although statements have been made concerning the mean and variance of the estimates in the selective sampling case, nothing has been said yet concerning confidence intervals. This is a much more complicated problem than that in the case of random sampling, since the estimates do not have a simple distribution function. In fact, the confidence intervals will in general depend on the particular set of e_i's (or r_i's) pertaining to the machine, and not simply on e (or r).

However, for small probabilities, the binomial distribution is quite closely approximated by the Poisson distribution, the fit becoming perfect as the probability approaches zero. For any reasonable recognition machine, one would expect the probabilities of error and rejection to be small; consequently, the marginal form of (9) for m_{e_i} or m_{r_i} may be approximated by a Poisson distribution. The estimates given by (11) and (12) are now sums of random variables with Poisson distributions (approximately) which are then themselves Poisson distributed. If the over-all error is also small, as is usually the case, the binomial-Poisson approximation can now be used in reverse, and one may state that, for small error rates, the error and rejection estimates

(11) and (12) are approximately binomially distributed. Consequently, one can use Fig. 1 to obtain 95 per cent confidence intervals for the error and rejection probabilities. Further, from (17), we would expect this confidence interval to be on the safe side, that is, the actual 95 per cent confidence interval should be slightly smaller than this.

APPLICATION TO PUBLISHED RESULTS

To illustrate the ease of determining these confidence intervals, some published results in pattern recognition are listed in Table 1 along with the 95 per cent confidence intervals as determined from Fig. 1. It should be emphasized that Table I is not meant to compare one method against another, since the methods obviously treat problems of various complexities. Rather, the table is meant to compare the accuracies of the various evaluating experiments.

Three points of caution should be noted concerning the validity of the confidence intervals in this table. First, the author is not positive that the test data is different from the design data in every case. Second, to the best of the author's knowledge, in every case the number of samples taken from each allowable pattern class was predetermined. This is selective sampling; therefore, it is assumed that the proportion of samples taken from each class represents its *a priori* probability of occurrence. The third assumption is that the patterns used to test the machine are a reasonable sampling from the real-life world of patterns, and are not biased toward either well-formed or poorly-formed (noisy) patterns.

CONCLUSION

Two important cases concerning the testing of pattern recognition methods or machines have been considered: Random sampling for the case of unknown *a priori* probabilities of class occurrence, and selective sampling for the case of known *a priori* probabilities. The most predominant form of testing in the present day art is to assume that the pattern classes have equal *a priori* probabilities of occurrence, and consequently to use equal sample sizes for each class; this is a special case of selective sampling.

It has been shown that, for both cases, the maximum-likelihood estimate for the error probability or rejection probability is simply the proportion of samples misrecognized or rejected. In the case of random sampling, the estimates are binomially distributed, and accurate confidence intervals can be obtained. In the case of selective sampling, tighter estimates are obtained which are approximately binomially distributed

TABLE I — 95 PER CENT CONFIDENCE INTERVALS FOR SOME PUBLISHED RESULTS

Author	Pattern Classes	Measured Characteristics	Recognition Criteria	Sample Size	Error	95% Confidence Interva
Baran, Estrin[7]	Machine Printed Numbers	Presence of ink in elements of 30 x 32 matrix	Maximize *a posteriori* probability (Bayes' Equation)	480	9%	7%–12%
Bledsoe, Browning[8]	Hand-Printed Alpha-Numerics	Presence of mark in elements of 10 x 15 matrix	Matching 2-tuples of matrix elements against table	180	21.6%	13%–29%
Bomba[9]	Hand-Printed Alpha-Numerics	Topological features (orientation of straight lines, intersections, etc.)	Decision tree	112	0%	0%–4%
Doyle[10]	Hand-Printed AEILMNORST	Simply measured topological features	Maximize *a posteriori* probability (Bayes' Equation)	~450	12.5%	10%–16%
Frishkopf[11]	Handwritten words	Extremes, and interconnections between extremes	Cross-correlation against dictionary	160	68%	57%–77%
Harmon[11]	Unsegmented Handwritten Letters	Topological features (cusps, closures, special marks, etc.)	Decision tree	412	41.1%	37%–46%
Mathews, Denes[12]	Spoken digits	Frequency vs time spectra	Cross-correlation against previous averaged spectra from same speaker	99	6%	2%–12%
Marill, Green[13]	Handwritten A,B,C (done as example only)	Distance of character from field edge along eight different line segments	Likelihood function assuming normal distribution of measures	90	3%	1%–10%
Sebestyen[14]	Spoken digits	Frequency vs time spectra	Minimization of non-Euclidean distance measure to average spectra	20	0%	0%–18%

for small error rates. Conservative confidence limits may then be obtained for these estimates.

Using these notions, the experimenter can now determine the sample size required to obtain results which he deems significant. Alternatively, if he has a limited sample size, he can determine the significance of his results. Note that in both cases considered, the variance is inversely proportional to the sample size. This does not mean that the confidence interval is inversely proportional to the square root of the sample size, however, since a binomial rather than a normal distribution pertains. However, perusal of Fig. 1 seems to indicate that this is a good rule of thumb. Note also that the total number of samples required to obtain a certain confidence in the results seems to be independent of the number of allowable pattern classes. This is an interesting philosophical point to ponder.

PART II — ON DESIGN

INTRODUCTION

Part I of this paper was concerned with the estimation of the performance of a given pattern recognition machine. There it was shown how confidence intervals could be found for these estimates. These results are nonparametric in that they hold for any categorization machine (or procedure) regardless of its structure.

We now consider the following problem. An experimenter desires to solve a particular pattern recognition problem. He has at his disposal a set of different methods for solving this problem, but it is not clear to him which is the best to use. Consequently, he desires to estimate the performance of each method when applied to this problem, and choose the best. Let us assume that each method is characterized by certain key parameters which, when known, completely determine the recognition machine. To evaluate any particular recognition method, the experimenter plans to design the corresponding machine by estimating its parameters on the basis of one sampling from the real world of patterns, and then to test this machine based on another sampling (either by constructing the machine or by simulating it).

However, in many practical applications, the total sample size available to the experimenter for design and test purposes is limited. For instance, he may be interested in building a machine to read hand-printed numbers, but he may not have an automatic scanner available to him. Since simulating a scanner by hand is very tedious, he may not be willing to scan more than a certain number of samples.

Or, he may be interested in distinguishing between radar returns caused by missiles and those caused by decoys. Since it is expensive to actually run the sort of experiment required to gather data for this problem, budget limitations will certainly place a limit on the number of available samples.

Another example is in the field of automatic diagnosis of diseases. The experimenter may, for instance, be interested in building a machine which would determine the presence of cancer based on a list of symptoms. However, records have been maintained for only a certain number of people who have contracted this disease, and the sample size is thus definitely limited.

The following problem then arises. If the total sample size is fixed, what is the optimum partitioning of this sample between the design and test phases? This is a rather loose, but concise, statement of the problem. A more accurate one follows.

Assume that the experimenter is concerned with the study of a particular pattern recognition method as applied to some particular problem. The optimum pattern recognition machine based upon this method would have an error probability e_o. The experimenter is interested in estimating e_o so that he can decide whether the particular method under study is adequate for the solution of his problem, or alternately whether it is better than another method. To do this, he takes a sample of a certain size t from the real-life world of patterns. He desires to use part of this sample to design a machine according to the particular method under study. The machine which he thus designs will have an actual error probability $e \geqq e_o$ (both quantities are unknown to the experimenter). He then uses the remaining part of his original sample to test the machine (according to the procedures of Part I). He thus obtains an estimate of e, which will be denoted by $\hat{e}$. It will be shown that $\hat{e}$ is a biased estimate of e_o, and that the bias can be computed. Consequently $\hat{e}$ can be adjusted so that it gives an unbiased estimate, $\hat{e}_o$, of e_o. The optimum partitioning of the total sample will be defined as that partitioning which minimizes the variance of $\hat{e}_o$. Thus, if the experimenter follows this procedure, he will obtain an unbiased minimum variance estimate of e_o, the optimum error probability. Of course, if he finally decides that a particular method is applicable, he can then redesign the corresponding machine with the entire sample size.

OPTIMUM SAMPLE PARTITIONING

We are interested, then, in minimizing the quantity

$$\sigma_{\hat{e}_o}^2 = E[\hat{e}_o - e_o)^2] = E[\hat{e}_o^2] - e_o^2, \tag{18}$$

where $E[x]$ and σ_x^2 denote the expected value and variance of x, respectively.

Let us first digress and consider the biased estimate $\hat{e}$. Since $\hat{e}$ is discrete (it is the proportion of test samples misrecognized), its expected value can be written

$$E[\hat{e}] = \sum \hat{e}p(\hat{e}),$$

where the summation is over all values of $\hat{e}$, and $p(x)$ denotes the probability of x. But

$$p(\hat{e}) = \int p(\hat{e} \mid e)p(e)\, de,$$

where $p(\hat{e} \mid e)$ is the probability of $\hat{e}$ given e, and the integral is over all (continuous) values of e (by definition $e_o \leqq e \leqq 1$). Hence

$$E[\hat{e}] = \sum \hat{e} \int p(\hat{e} \mid e)p(e)de = \int [\sum \hat{e}p(\hat{e} \mid e)]\, p(e)de.$$

Let us henceforth consider only the case of random sampling. Then $\hat{e}$ is proportional to a binomially distributed variable ($n\hat{e}$) with parameter e. Therefore the term in brackets, which is the expected value of $\hat{e}$ given the parameter e, is just e. Then

$$E[\hat{e}] = \int ep(e)de = E[e]. \tag{19}$$

$E[e]$ is a function only of the parameters of the problem and the design sample size; it is not a random variable.

We next determine $E[\hat{e}^2]$. By going through a process analogous to the above, and by making use of (19), we obtain

$$\sigma_{\hat{e}}^2 = E[(\hat{e} - E[e])^2] = E[\hat{e}^2] - (E[e])^2 = \frac{E[e(1 - e)]}{n},$$

where n is the size of the test sample. Hence

$$E[\hat{e}^2] = \frac{E[e(1 - e)]}{n} + (E[e])^2. \tag{20}$$

We now determine $E[e]$. Let the optimum machine be described by c different parameters δ_{oi}, $1 \leqq i \leqq c$. The design of the machine consists of estimating the parameters δ_{oi} by making measurements on a set of sample patterns (the design sample). Let the estimates of these parameters be denoted δ_i, $1 \leqq i \leqq c$. Then the error probability e

of the resulting machine is a function of the estimates of the true parameters:

$$e = e(\delta_1, \delta_2, \cdots, \delta_c).$$

One can now expand e in a Taylor series expansion about its minimum point, e_o. Since this is a minimum point, all the coefficients of the linear terms will be zero. If the error deviation $(e - e_o)$ is small, terms above the second order term may be neglected:

$$e \approx e_o + \tfrac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \Bigg|_{\delta_o} (\delta_i - \delta_{oi})(\delta_j - \delta_{oj}).$$

The expected value of the error for the resulting machine is then

$$E[e] = e_o + \tfrac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \Bigg|_{\delta_o} E[(\delta_i \, \delta_{oi})(\delta_j - \delta_{oj})].$$

If it is assumed that the estimates are unbiased, i.e., $E(\delta_i) = \delta_{oi}$, then the above equation may be written as

$$E[e] = e_o + \tfrac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} a_{ij}\sigma_{ij} \tag{21}$$

where

$$a_{ij} = a_{ji} = \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \Bigg|_{\delta_o},$$

σ_{ij} is the covariance of the estimates for δ_{oi} and δ_{oj}, and $\sigma_{ii} = \sigma_i^2$ is the variance of the estimate for δ_{oi}. (21) is valid for small values of the quantity $(e - e_o)$.

It may be worth-while to digress here to a simple example which may help to clarify the definitions of the above terms. Zachary Oglethorpe is not only a crafty fisherman, but is also a good gadgeteer. He has decided to try to build equipment which will determine each day whether he should use a surface bait or a deep water bait in order to catch the maximum number of fish. He has means available to measure the water temperature, the magnitude of surface ripple, and the atmospheric pressure, and therefore decides to use these as his measurements. He denotes values of these measurements by m_1, m_2, and m_3 respectively.

Mr. Oglethorpe has been recording values of these measurements every day for the past six months, and has noted on each day whether he was more successful with surface or deep water bait. He thus has a

total sample size of roughly 180 samples, some from one pattern class (surface bait), and some from the other pattern class (deep water bait). Since each sample was taken without *a priori* knowledge of the class to which it belonged, then this constitutes random sampling; that is, the proportion of samples in each class is an estimate of the *a priori* probability of occurrence of that class.

Our crafty fisherman decides to build a decision making, or pattern recognition, machine by building a correlator for each of the two possible decisions (or pattern classes). That is, the machine will make the following two calculations:

$$\text{Surface bait} = \delta_1 m_1 + \delta_2 m_2 + \delta_3 m_3 ,$$

$$\text{Deep water bait} = \delta_4 m_1 + \delta_5 m_2 + \delta_6 m_3 .$$

The class achieving the highest value represents the desired decision. Let us assume that, according to some theory, the optimum values of the δ_i are the means for each measurement within the appropriate pattern class, normalized so that the sum of the squares of the coefficients of each linear form is unity. That is, δ_1 is proportional to the mean water temperature when surface bait should be used, and so forth, and is normalized with δ_2 and δ_3 so that $\delta_1^2 + \delta_2^2 + \delta_3^2 = 1$.

Thus the parameters δ_i completely characterize this pattern recognition machine in that, given values for each δ_i, $1 \leqq i \leqq 6$, the machine may be built. The optimum values for each δ_i are the appropriate normalized means, which are the δ_{oi} of the previous equations. Mr. Oglethorpe obtains estimates of these optimum parameters by taking normalized averages over a portion of the appropriate data. These estimates are the δ_i of the previous equations, and are the actual numbers on which he would base the construction of his machine. Note that, in this case, these estimates are unbiased and efficient, and may very well be independent of each other (e.g., the probability distribution of the water temperature when surface bait should be used may be independent of the values of surface ripple magnitude and atmospheric pressure).

Having thus designed his fisherman's aid with a portion of his data, he now tests it with the remainder of the data to determine its accuracy. He does not want to use it if there is a good probability that it is less accurate than he has found his own intuition to be. This then leads us to the basic problem being studied: How should Zachary Oglethorpe split his total sample between the design and the testing of his machine to obtain the best estimate of the accuracy of the machine? Again, if

the estimated accuracy of his machine were sufficient, he would then be wise to redesign it, basing the new design on the entire sample.

We now return to the study of this sample partitioning. Let each parameter be estimated with m samples.* If each of these estimates is an efficient and unbiased estimate, and if the estimates are independent (either because the estimates are statistically independent, or because different samples are used to estimate each), then all $\sigma_{ij} = 0$, $i \neq j$, and all σ_i^2 will be proportional to $1/m$. Hence one can rewrite (21) as

$$E[e] = e_o + \frac{b}{m}, \tag{22}$$

where b is some constant calculated from (21). (Often, $E[e]$ is in the form (22) even if the estimates are not independent.)

Let t be the total sample size, and p be the number of sets of m samples used to design the machine. p is chosen to be the smallest number which insures that $E[e]$ is of the form (22). It is often simply the number of allowable pattern classes, since, of course, parameters of different classes must be estimated with different samples. If n is the test sample size, then

$$t = n + pm. \tag{23}$$

From (19) and (22),

$$E[\hat{e}] = E[e] = e_o + \frac{b}{m}. \tag{24}$$

Consequently, $\hat{e}$ is a biased estimate of e_o. The adjusted estimate $\hat{e}_o$, given by

$$\hat{e}_o = \hat{e} - \frac{b}{m}, \tag{25}$$

is an unbiased estimate of e_o, with variance given by (18). This variance can now be rewritten using (25):

$$\sigma_{\hat{e}_o}^2 = E[\hat{e}_o^2] - e_o^2 = E\left[\left(\hat{e} - \frac{b}{m}\right)^2\right] - e_o^2$$

$$= E[\hat{e}^2] - 2\frac{b}{m}E[\hat{e}] + \left(\frac{b}{m}\right)^2 - e_o^2.$$

* This is not always desirable, since some parameters may be easier to estimate than others, or there may be more data available for some parameters than others. However, this condition is assumed here for simplicity, as are the following assumptions of efficiency, unbiasedness, and independence.

From (20) and (24),

$$\sigma_{\hat{e}_o}^2 = \frac{E[e(1-e)]}{n} + (E[e])^2 - 2\frac{b}{m}e_o - \left(\frac{b}{m}\right)^2 - e_o^2.$$

$$= \frac{E[e(1-e)]}{n} + (E[e])^2 - \left(e_o + \frac{b}{m}\right)^2.$$

Thus, from (24),

$$\sigma_{\hat{e}_o}^2 = \frac{E[e(1-e)]}{n}. \tag{26}$$

If $b/m \ll 1$ (which will certainly be true for any reasonable design), then

$$\sigma_{\hat{e}_o}^2 \approx \frac{E[e(1-e_o)]}{n} = (1-e_o)\frac{e_o + \frac{b}{m}}{n} = (1-e_o)\frac{e_o + \frac{pb}{t-n}}{n}, \tag{27}$$

where the relation (23) was used.

We wish to choose n such that (27) is minimized. Differentiating (27) and equating to zero, one obtains

$$\frac{e_o t}{pb} = \frac{2\frac{n_o}{t} - 1}{\left(1 - \frac{n_o}{t}\right)^2}, \tag{28}$$

where n_o is that value of n satisfying (28); it is the optimum test sample size in the sense previously discussed. n_o/t is of course the proportion of the total sample used for the test. One interesting result is immediately obvious: n_o/t must be greater than 0.5 for all cases. The equation (28) is plotted in Fig. 2, from which the following general statements can be made.

1. The proportion of the total sample that should be used to test the machine should never be less than 50 per cent.
2. If $e_o t/pb < 0.1$, then the proportion used for design should be about 50 per cent.
3. The proportion of the total sample that should be used to test the machine becomes larger as:
   a. The total sample size increases,
   b. the error of the optimum machine increases,
   c. the effectiveness of the design increases (pb decreases).

Here $1/pb$ is taken as a measure of the effectiveness of the design,

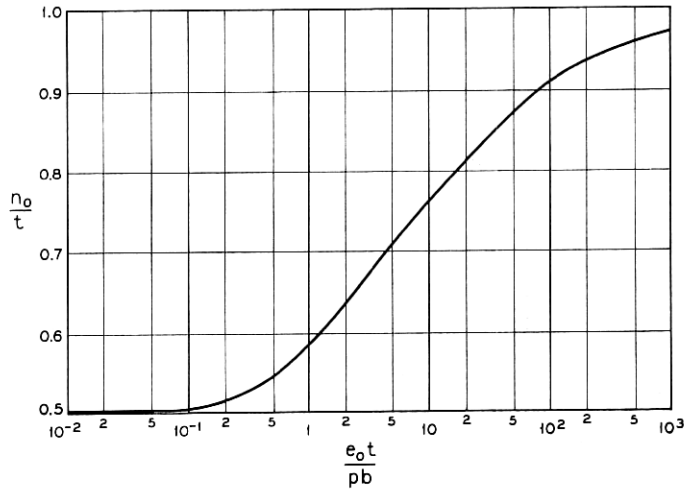

Fig. 2 — Optimum sample partitioning.

since pb is the product of the expected deviation from optimum, $E[e - e_o]$, and the design sample size, pm.

These results indicate just how a sample should be split between the design and test stages of a feasibility study of a pattern recognition method. If the experimenter follows this procedure, he will obtain an estimate $\hat{e}_o$ of e_o which is unbiased and has minimum variance.

The value of this minimum variance can be expressed as

$$\sigma^2_{\hat{e}_{o\min}} = \frac{e_o(1 - e_o)}{n}\left(1 + \frac{1 - \frac{n_o}{t}}{2\frac{n_o}{t} - 1}\right),$$

which was obtained by eliminating pb between (27) and (28). Note that this is the variance that would have been obtained if the optimum machine were tested with n samples, increased by a factor which accounts for the design error.

AN EXAMPLE OF OPTIMUM SAMPLE PARTITIONING

As an illustration of these ideas, consider the following example (perhaps the simplest of the n-dimensional problems). A pattern recognition machine is to be designed using the optimum decision function[15,16] which will distinguish between q classes. The occurrence of each class is equally probable *a priori*, and all costs of misrecognition are the same. The receptor makes a set of k measurements m_j, $1 \leqq j \leqq k$, on each

input pattern. It is known that each measurement is normally distributed with variance σ, and that all measurements are independent. Further, it is known that the distances between the mean vectors in measurement space* are all equal. (Consequently, there can be no more than $k + 1$ pattern classes. The tips of the mean vectors are the vertices of a regular polytope.)

It can then be shown that the optimum decision function partitions the measurement space into polytopes which are bounded by those hyperplanes which are the perpendicular bisectors of the line segments joining all pairs of means. The hyperplane separating two classes, say classes 1 and 2, is the set of all points $(x_1, \cdots, x_k)$, represented by the vector $\bar{X}$, which satisfy

$$\bar{x} \cdot (\bar{\mu}_1 - \bar{\mu}_2) = \tfrac{1}{2}(\bar{\mu}_1 \cdot \bar{\mu}_1 - \bar{\mu}_2 \cdot \bar{\mu}_2), \tag{29}$$

where $\bar{\mu}_i$ is the mean vector of class i.[13]

The design procedure consists of estimating each mean vector from a sampling; denote the estimated mean vector for class i by $\bar{x}_i$. The distribution of the estimate of a mean vector from a normal distribution with covariance matrix $[V]$ is also normal with covariance matrix $1/m\ [V]$, where m is the sample size used in the estimate.[17] Since the measurements are independent in this case, then so will be the estimates of the means of the various measurements. Furthermore, each estimate will have a variance of σ^2/m. Consequently, only one set of samples of size m from each pattern class is required to insure that the form (22) is valid, and p is hence equal to the number of allowable pattern classes, q.

It is shown in the Appendix that b is given by

$$b = \frac{q(q-1)}{4} \frac{\Delta\mu}{2\sigma} N\left(\frac{\Delta\mu}{2\sigma}\right),$$

where $\Delta\mu$ is the distance between any pair of mean vectors, and $N(\Delta\mu/2\sigma)$ is the value of the standard normal density function for the variable $\Delta\mu/2\sigma$. The equation (28) then becomes

$$\frac{4e_o t}{q^2(q-1)\dfrac{\Delta\mu}{2\sigma} N\left(\dfrac{\Delta\mu}{2\sigma}\right)} = \frac{2\dfrac{n_o}{t} - 1}{\left(1 - \dfrac{n_o}{t}\right)^2}. \tag{30}$$

* A geometric interpretation of categorization problems is often useful. By measurement space, we mean a k-dimensional space in which each coordinate represents one of the k receptor measurements. Thus any set of measurements which have been made on an input pattern may be represented as a point in measurement space. The decision function may be thought of as partitioning the measurement space into regions corresponding to the different allowable pattern classes and into rejection regions.

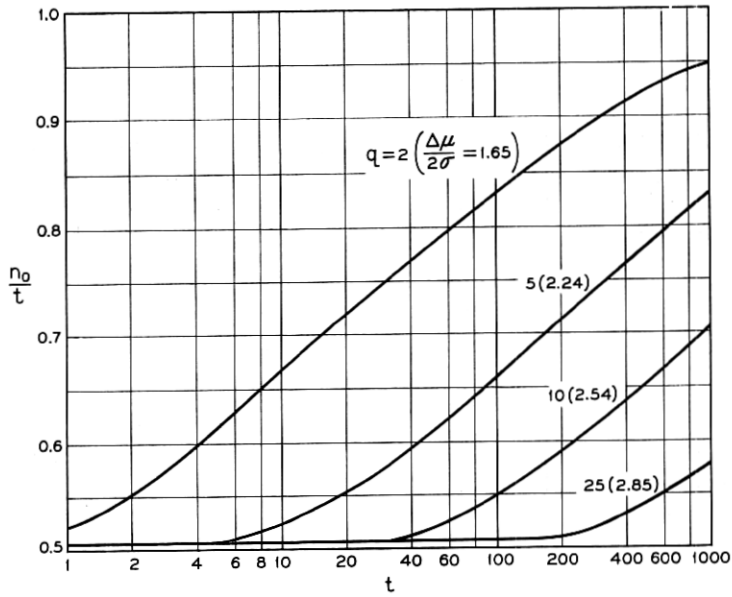

Fig. 3 — Optimum sample partitioning for symmetric Gaussian case.

Some curves representing (30) are plotted in Fig. 3 in which the proportion of the total sample to be used in the test, n_o/t, is shown as a function of t, the total sample size, with the number of allowable pattern classes, q, as a parameter. e_o was held constant at 0.05 (which involves the choosing of the proper value of $\Delta\mu/2\sigma$ for each q).

From Fig. 3 it is seen that, for many cases, the sample should be split evenly between design and test, as one might intuitively suspect. However, there are some drastic deviations from this. For instance, if the categorizer is to separate only two classes, and 1000 samples are available, then only 50 of these should be used to design the machine, and 950 should be used to test it. Consequently, it is seen that intuition may go wrong in some cases.

CONCLUSION

This paper has begun an analysis of some of the problems which arise in the design and analysis of pattern recognition experiments. In Part II, the problem of optimum sample partitioning between the design and test phases of a pattern recognition machine was investigated for the

case of a fixed total sample size and no overlap between the design and test samples. The general relation between the optimum partitioning and the total sample size, optimum error rate, and design efficiency was derived. From this, it was apparent that the test sample size should never be smaller than the design sample size. These results are nonparametric in the sense that they do not depend on the detailed structure of the recognition machine. It is only necessary that the deviation of the designed machine from the optimum machine be small, and that the design of the machine be done in such a way that (22) holds.

However, the actual computation of the optimum sample partitioning does depend strongly on the detailed structure of the machine through the quantity b. Since this computation is quite difficult even in the simplest of cases, the interesting question arises as to the possibility of estimating b from the sample. Another interesting phase of this problem which has not been attacked here concerns the case when the design sample and test sample overlap — that is, some of the sample patterns from the design sample are also used in the test sample. In the limit, this reduces to using the total sample for both design and test purposes. In this case, the results of the test are usually not very reliable. Consequently, there may be some sample partitioning with overlap which is better (in the sense discussed in this paper) than for either the case of no overlap or the case of total overlap.

ACKNOWLEDGMENT

I would like to thank Prof. A. E. Laemnel of the Polytechnic Institute of Brooklyn for suggesting this problem, and Messrs. E. Wolman, W. H. Williams, and E. N. Pinson for their very helpful discussions concerning these topics.

APPENDIX

We determine here the coefficient b in (22) for the example discussed in this paper. If the mean vectors are more than about 3σ apart, then only a small error is made if the total error is approximated by adding the errors of each hyperplane taken alone. That is, the integrals on the wrong side of the hyperplane that are counted more than once will be quite small compared to the integrals counted only once.

Due to the symmetry of the problem, the error associated with each hyperplane for the optimum decision function is identical, and the derivatives of (21) will also be identical for each hyperplane. Since there are

$q(q-1)/2$ hyperplanes, b may be expressed (from (21) and (22)) as

$$\frac{b}{m} = \frac{q(q-1)}{2} \frac{1}{2} \sum_{i=1}^{k} \left[\frac{\partial^2 e_{12}}{\partial \bar{x}_{i1}^2} \bigg|_{\mu_1,\mu_2} + \frac{\partial^2 e_{12}}{\partial \bar{x}_{i2}^2} \bigg|_{\mu_1,\mu_2} \right] \frac{\sigma^2}{m}, \tag{31}$$

where the hyperplane separating classes 1 and 2 is taken as typical, and the independence of the estimates is used. e_{12} is the error associated with this hyperplane, μ_1 and μ_2 are the mean vectors of these classes, and $\bar{x}_1$ and $\bar{x}_2$ are the estimates of the mean vectors.

There is no loss in generality if μ_1 is taken as zero, and all the components of μ_2 (μ_{12}, $\cdots$, μ_{k2}) are taken as zero except for μ_{12}. That is,

$$\mu_1 = (0,0, \cdots ,0)$$
$$\mu_2 = (\mu,0, \cdots ,0),$$

where μ_{12} is denoted μ, $\mu > 0$. Consequently, the optimum boundary is given by

$$x_1 = \mu/2.$$

A sampling of size m is taken from each class, and the mean vectors are estimated, giving

$$\bar{x}_1 = (\bar{x}_{11}, \bar{x}_{21}, \cdots, \bar{x}_{k1})$$
$$\bar{x}_2 = (\bar{x}_{12}, \bar{x}_{22}, \cdots, \bar{x}_{k2}).$$

A boundary given by (29) is computed based on the above estimates, and this, together with the other estimated boundaries, determines the structure of the machine.

The error e_1 associated with this particular boundary for class 1 is

$$e_1 = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \cdot \int_{\xi_1(x_2,\ldots x_{1k})}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_1}{\sigma}\right)^2 dx_1,$$

where $\xi_1(x_2, \cdots, x_k)$ is the value of x_1 on the boundary, and is given by (from (29))

$$\xi_1(x_2, \cdots, x_n) = -\sum_{i=2}^{k} \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\bar{x}_{11} - \bar{x}_{12}} x_i + \frac{1}{2} \sum_{i=1}^{k} \frac{(\bar{x}_{i1}^2 - \bar{x}_{i2}^2)}{\bar{x}_{11} - \bar{x}_{12}}$$
$$= \frac{\bar{x}_{11} + \bar{x}_{12}}{2} - \frac{1}{2} \sum_{i=2}^{k} \frac{2(\bar{x}_{i1} - \bar{x}_{i2})x_i - (\bar{x}_{i1}^2 - \bar{x}_{i2}^2)}{x_{11} - x_{12}}.$$

Then

$$\frac{\partial e_1}{\partial \bar{x}_{i1}} = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{\xi_1}{\sigma}\right)^2\right)\left(\frac{x_i - \bar{x}_{i1}}{\bar{x}_{11} - \bar{x}_{12}}\right), \quad 2 \leqq i \leqq n.$$

$$\frac{\partial^2 e_1}{\partial \bar{x}_{i1}^2} = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left(\frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{\xi_1}{\sigma}\right)^2\right) \cdot \left[\frac{\xi_1}{\sigma^2}\left(\frac{x_i - \bar{x}_{i1}}{\bar{x}_{11} - \bar{x}_{12}}\right)^2 - \left(\frac{1}{\bar{x}_{11} - \bar{x}_{12}}\right)\right], \quad 2 \leqq i \leqq n.$$

$$\left.\frac{\partial^2 e_1}{\partial \bar{x}_{i1}^2}\right|_{\mu_1,\mu_2} = \left(\frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{\mu/2}{\sigma}\right)^2\right) \cdot \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left[-\frac{x_i}{2\sigma^2} + \frac{1}{\mu}\right]$$

$$= \frac{1}{\sigma} N\left(\frac{\mu}{2\sigma}\right)\left[-\frac{1}{2\sigma^2} E[x_i] + \frac{1}{\mu}\right]$$

$$= \frac{1}{\mu\sigma} N\left(\frac{\mu}{2\sigma}\right), \qquad 2 \leqq i \leqq n,$$

where $N(\mu/2\sigma)$ is the value of the standard normal density function for the variate $\mu/2\sigma$. In a like manner,

$$\left.\frac{\partial^2 e_2}{\partial \bar{x}_{i1}^2}\right|_{\mu_1,\mu_2} = -\frac{1}{\mu\sigma} N\left(-\frac{\mu}{2\sigma}\right) = -\frac{1}{\mu\sigma} N\left(\frac{\mu}{2\sigma}\right), \qquad 2 \leqq i \leqq n,$$

where e_2 is the error associated with this boundary for class 2. Since the total error for this boundary is $e_{12} = e_1 + e_2$, then

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{i1}^2}\right|_{\mu_1,\mu_2} = \left.\frac{\partial^2 e_1}{\partial \bar{x}_{i1}^2}\right|_{\mu_1,\mu_2} + \left.\frac{\partial^2 e_2}{\partial \bar{x}_{i1}^2}\right|_{\mu_1,\mu_2} = 0, \qquad 2 \leqq i \leqq n.$$

A like result holds for $\dfrac{\partial^2 e_{12}}{\partial \bar{x}_{i2}^2}$, $2 \leqq i \leqq n$. Going through this same procedure for $\bar{x}_{11}$,

$$\frac{\partial e_1}{\partial \bar{x}_{11}} = -\prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left[\frac{1}{2\sigma} N\left(\frac{\xi_1}{\sigma}\right)\right].$$

$$\frac{\partial^2 e_1}{\partial \bar{x}_{11}^2} = -\prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left[-\frac{1}{4\sigma}\left(\frac{\xi_1}{\sigma^2}\right) N\left(\frac{\xi_1}{\sigma}\right)\right].$$

$$\left.\frac{\partial^2 e_1}{\partial \bar{x}_{11}^2}\right|_{\mu_1,\mu_2} = \frac{1}{8}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

Similarly,

$$\left.\frac{\partial^2 e_2}{\partial \bar{x}_{11}{}^2}\right|_{\mu_1,\mu_2} = \frac{1}{8}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

Hence

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{11}{}^2}\right|_{\mu_1,\mu_2} = \frac{1}{4}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

It would also be found that

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{12}{}^2}\right|_{\mu_1,\mu_2} = \frac{1}{4}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

This analysis is perfectly general for arbitrary mean vectors, providing that μ is merely interpreted as the distance between a pair of mean vectors (all such distances being assumed identical). This distance will henceforth be written $\Delta\mu$ to indicate that it is a difference of means. Therefore, from (31), we find that

$$b = \frac{q(q-1)}{4}\frac{\Delta\mu}{2\sigma} N\left(\frac{\Delta\mu}{2\sigma}\right).$$

REFERENCES

1. Fraser, D. A. S., *Statistics: An Introduction*, John Wiley and Sons, Inc., New York, 1960.
2. Mattson, R. L., Master's Thesis, E. E. Dept., M.I.T., May, 1959.
3. Clopper, C. S., and Pearson, E. S., Biometrika, **26,** 1934, p. 404.
4. Pearson, E. J., and Hartley, H. O., *Biometrika Tables for Statisticians*, The University Press, Cambridge, 1954, p. 204.
5. Bowley, A. L., Bull. Int. Stat. Inst., **22,** 1926, p. 1.
6. Neyman, J., J. Royal Stat. Soc., **97,** Pt. 4, 1934, p. 558.
7. Baran, P., and Estrin, G., I.R.E. Wescon Record, Pt. 4, 1960, p. 29.
8. Bledsoe, W. W., and Browning, I., Proc. E.J.C.C., Dec., 1959, p. 225.
9. Bomba, J. S., Proc. E.J.C.C., Dec., 1959, p. 218.
10. Doyle, W., Proc. W.J.C.C., 1960, p. 133.
11. Frishkopf, L. S., and Harmon, L. D., Proc. 4th London Symposium on Information Theory, 1960.
12. Mathews, M. V., and Denes, P., J. Acous. Soc. Amer., **32,** Nov., 1960, p. 1450.
13. Marill, T., and Green, D. M., I.R.E. Trans. on Elec. Comp., **EC-9,** Dec., 1960, p. 472.
14. Sebestyen, G. S., I.R.E. Trans. on Information Theory, **IT-7,** Jan., 1961, p. 44.
15. Middleton, D., and Van Meter, D., J. Soc. Ind. and App. Math., **3,** Sept., 1955, p. 192; **4,** June, 1956, p. 86.
16. Chow, C. K., I.R.E. Trans. on Elec. Comp., **EC-6,** Dec., 1957, p. 247.
17. Anderson, T. W., *An Introduction to Multivariate Statistics*, John Wiley and Sons, New York, 1958.