ARPA ORDER NO. 189-61

MEMORANDUM
RM-4405-ARPA
FEBRUARY 1965

SUFFICIENCY AND INFORMATION RATE OF MULTI-STAGE STATISTICAL TESTS

Julian J. Bussgang and Michael B. Marcus

PREPARED FOR:
ADVANCED RESEARCH PROJECTS AGENCY

The RAND Corporation
SANTA MONICA · CALIFORNIA

ARPA ORDER NO. 189-61

MEMORANDUM
RM-4405-ARPA
FEBRUARY 1965

SUFFICIENCY AND INFORMATION RATE OF MULTI-STAGE STATISTICAL TESTS

Julian J. Bussgang and Michael B. Marcus

This research is supported by the Advanced Research Projects Agency under Contract No. SD-79. Any views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of ARPA.

The RAND Corporation
1700 MAIN ST • SANTA MONICA • CALIFORNIA • 90406

PREFACE

In the design of phased-array radars, processing equipment and/or radar power can be saved if sequential detection (multi-stage statistical test) criteria are used. This Memorandum demonstrates theoretically in what sense Wald's sequential testing is optimal. The study is novel in that it shows that sequential testing is optimal in an information theoretic sense.

The work was undertaken as basic research in technology applicable to the design of electronically scanned radars of potential use in ballistic missile defenses. It is part of a continuing study for ARPA on low-altitude defense against ballistic missiles.

Dr. Julian J. Bussgang, co-author of this Memorandum, is President of SIGNATRON, Inc., Lexington, Massachusetts, and is a Consultant to The RAND Corporation.

BLANK PAGE

SUMMARY

In this Memorandum some fundamental aspects of multi-stage tests of alternate statistical hypotheses are discussed. Section II is devoted to the formulation of the problem and the definition of the quantities of interest. Section III demonstrates certain fundamental equalities of the conditional distributions of the sample size which occur in Wald's sequential probability ratio test. These equalities, which to the authors' knowledge have not been noted before, imply that the terminal decision is a sufficient statistic for the estimation of the true hypothesis regardless of the terminal stage. In Section IV a further consequence of these equalities is demonstrated. Using information theoretic concepts, the rate of transmission of a statistical test is defined and a test procedure, constructed to satisfy these equalities, is shown to minimize this rate. The information theoretic view of an alternate decision problem has been suggested before, but only for a fixed sample test.(1) The results in the Memorandum provide an alternate approach to the study of the optimality of multi-stage tests of alternate statistical hypotheses and suggest a criterion for designing such tests based on the conditional distributions of the sample size rather than on the average risk.

CONTENTS

I. DEFINITION OF AN ALTERNATE HYPOTHESIS STATISTICAL TEST

As a general framework for alternate hypothesis statistical tests involving a discrete sample we consider that there exists a real valued Borel probability measure defined on $\overset{\infty}{\underset{i=1}{X}} R_i$ where $R_i = R$, R being the real line. This measure is known up to a parameter θ that can have one of two values.* These values for θ form two hypotheses about the measure which are characteristically denoted by H_0 (the null hypothesis) and H_1 (the alternate hypothesis). Similarly, we shall denote the corresponding measures by μ_0 and μ_1. We assume also that there exist a priori probabilities π and $1-\pi$ that the measures are μ_0 and μ_1, respectively. Each possible measure generates a stochastic process, Ω_0 or Ω_1, with elements $x \in \overset{\infty}{\underset{i=1}{X}} R_i$ called paths, which are sequences of real numbers.

In an actual statistical test there is some mechanism for obtaining numbers called observations. We assume that they can be obtained one at a time; obtaining the nth observation will be called the nth stage of the test. When the observations are written in order $x_1, \ldots, x_n$ (we call the sequence of observations a sample) they represent the first n values of a particular realization or path of either the stochastic process Ω_0 or the stochastic process Ω_1. A multi-stage alternate hypothesis test is a decision procedure that uses

*We consider that the measure underlying the sample is one of $\mu_\theta (\theta = 0,1)$ in order to cast our problem as one of parameter estimation. However, by parameter estimation we mean more than estimating a parameter that appears in a distribution function that might generate the measure, like the mean of a Gaussian distribution. We view the parameter θ as an index for the two possible values of the measure.

the sample to determine, subject to certain pre-established probabilities of error, to which of these processes the path belongs. That is, it determines whether the underlying measure is μ_0 or μ_1.

At each stage of the alternate hypothesis test one of three decisions can be made: that the hypothesis H_0 is true, that the hypothesis H_1 is true or that another observation should be made. Trying to achieve maximum generality we impose on these tests only two conditions: that at each stage these three decisions are mutually exclusive, and that with probability one the test eventually leads to the acceptance of one of the two hypotheses.

Tests will be said to have the same power if they have the same error probabilities. The error probabilities are denoted as follows:

$\alpha \equiv$ probability that H_1 is accepted when H_0 is true

$\beta \equiv$ probability that H_0 is accepted when H_1 is true

For any alternate hypothesis test for which a decision is made at each stage, the collection of paths that lead to the acceptance of $H_0(H_1)$ at the $n^{\underline{th}}$ stage is a cylinder set in $\Omega=\Omega_0 \cup \Omega_1$. This cylinder set will be denoted by $C_n^* \left(C_n^{**}\right)$.

We define

$$\mu_\theta^*(n) \left(\mu_\theta^{**}(n)\right) \equiv \text{measure of } C_n^* \left(C_n^{**}\right) \qquad \theta=0,1.$$

The conditions imposed upon the tests insure that these measures are well defined. For all tests of the same power (i.e., to which there corresponds a specific pair (α, β)), we see that under the two conditions imposed above

$$\sum_{n=1}^{\infty} \mu_0^*(n)=1-\alpha, \sum_{n=1}^{\infty} \mu_1^*(n)=\beta, \sum_{n=1}^{\infty} \mu_0^{**}(n)=\alpha \text{ and } \sum_{n=1}^{\infty} \mu_1^{**}(n)=1-\beta.$$

Thus we can define four conditional probability density functions $p_\theta^*(n)$ and $p_\theta^{**}(n)$, $\theta=0,1$. As an example we have

$$p_0^{**}(n) = \frac{\mu_0^{**}(n)}{\sum_{n=1}^{\infty} \mu_0^{**}(n)} = \frac{\mu_0^{**}(n)}{\alpha} \equiv \text{probability that } H_1 \text{ is accepted at the } n^{\underline{th}} \text{ stage given that } H_0 \text{ is the true hypothesis}$$

Lastly, we term the acceptance of H_0 decision zero, D_0, and the acceptance of H_1 decision one, D_1.

II. IMPLICATIONS OF THE EQUALITIES $p_0^*(n) = p_1^*(n)$ and $p_0^{**}(n) = p_1^{**}(n)$

We consider tests of the same power characterized by an (α, β) and are concerned with the completed tests and the decision to which they lead. For a completed test we have knowledge about two random variables, the stage at which the test terminates and whether it terminates in D_0 or D_1. This is the case regardless of what functions of the sample are used to arrive at the decisions. By $P(H_i|D_j, n)$ i,j=0,1 we mean the probability that H_i is the true hypothesis given that the test ended at the $n^{\underline{th}}$ stage with the acceptance of hypothesis j. The probability $P(H_\theta|D_j, n)$ is the a <u>posteriori</u> probability of the true hypothesis H_θ in a multi-stage alternate hypothesis test. There are four functions of this kind; an example is

$$P(H_0|D_1, n) = \frac{\pi\alpha p_0^{**}(n)}{\pi\alpha p_0^{**}(n) + (1-\pi)\ (1-\beta)\ p_1^{**}(n)} \tag{1}$$

So far we have attempted to portray alternate hypothesis tests in their greatest generality; in practice it is common to use a fixed sample or Wald sequential probability ratio test. In the first case, $p_\theta^*(N) = p_\theta^{**}(N) \approx 1$, θ=0,1 where N is the pre-assigned stage at which the test terminates. In the latter case the functions $p_\theta^*(n)$ and $p_\theta^{**}(n)$ (i=0,1) n = 1,2..., are generally difficult to calculate. Experience indicates that unless the $p_\theta^*(n)$, $p_\theta^{**}(n)$ can be obtained trivially, as in a fixed sample test, their calculation is a major and frequently unsolvable problem. It is useful to consider when the conditional probabilities that the correct decision was made would be independent of the stage at which the test terminates. From the form of expression (1) it is clear that this is the case if and only if

$$p_0^*(n) = p_1^*(n),\ p_0^{**}(n) = p_1^{**}(n) \quad (n = 1,2,\ldots,) \qquad (2)$$

We have the following theorem:

Theorem 3.1 The *a posteriori* probability of satisfying either hypothesis in a multi-stage test of alternate hypotheses is independent of the stage at which the test ended if and only if (2) is satisfied.

Proof:

Sufficiency is obvious from expression (1); for necessity we notice that

$$\frac{p_1^{**}(n)}{p_0^{**}(n)} = \text{const.}$$

Since both $p_1^{**}(n)$ and $p_0^{**}(n)$ are probability measures, the constant must be one.

We employ Theorem 3.1 to demonstrate the statistical sufficiency of the terminal decision of an alternate hypothesis test, when the test procedure is such that (2) is satisfied and the test is used to estimate θ. The outcome of a specific test is a random variable Γ which takes on the values $\left\{D_j, n\right\}$ $j = 0,1$; $n = 1,2\ldots$. Let T be a function of Γ such that $T\left(\Gamma = \left\{D_j, n\right\}\right) = D_j$. Statistical sufficiency can be defined by the following statement: (Ref. 2) "If the conditional distribution of θ given $X = x$ depends only on

$T(x)$ then T is a sufficient statistic for θ." Thus in the problem of observing the random variable Γ and estimating θ, by Theorem 3.1, the statistic $T\left(\left\{D_j, n\right\}\right) = D_j$ is a sufficient statistic if and only if (2) is satisfied.

It is common to speak of a statistic as being sufficient for the estimation of a parameter of a stochastic process when the statistic is a function of the observations of the process. Here the statistic is a function of both the observations and the test procedure which is chosen. It is clear that when an alternate hypothesis test procedure is chosen. and the outcome of this test procedure is used as an estimate for the parameter, that considerable information about the parameter contained in the observations might be lost. The point of view that we take in this Memorandum is that we are studying the outcome of a multi-stage alternate hypothesis test, and not the composition of the sample. The only utilizable information that these tests convey is the decision that they lead to. Thus it is important to know when the terminal decision is a sufficient statistic with respect to the true hypothesis. The fact that (2) implies sufficiency of the test statistic establishes the significance of the equalities (2).

We now show that (2) is satisfied by the Wald test. (This is also true for a fixed sample size Neyman-Pearson test which satisfies (2) as a trivial case.) Since the Wald test employs the likelihood ratio, it is necessary to introduce additional assumptions on μ_0 and μ_1 to insure that this ratio exists. The likelihood ratio at stage n is a function of the first n observations of a particular sample path

$$\frac{d\mu_1(x_1, x_2, \ldots, x_n)}{d\mu_0(x_1, x_2, \ldots, x_n)}$$

This function exists as a Radon-Nikodym derivative as long as μ_1 is absolutely continuous with respect to μ_0.

The Wald sequential probability ratio test is a multi-stage test of alte.nate hypotheses that continues as long as

$$B < \frac{d\mu_1(x_1, \ldots, x_n)}{d\mu_0(x_1, \ldots, x_n)} < A \qquad (A>1,\ B<1 \text{ are constints}) \qquad (3)$$

and ceases with the acceptance of H_0 if the left inequality is violated and with the acceptance of H_1 if the right inequality is violated.

There is a fundamental approximation used in connection with Wald tests that is frequently referred to as "neglecting the excess over the boundary". This approximation consists of assuming that when the sequential test terminates there is equality at either the left side or right side of (3). The approximation becomes exact when the sample paths are continuous with independent increments and when the probability density function for the value of each increment is continuous.

It is well known that with this approximation B is taken to be $\frac{\beta}{1-\alpha}$ and A is taken to be $\frac{1-\beta}{\alpha}$.

In the terminology of this Memorandum, Wald's approximation consists of saying that for those paths which lead to $D_0(D_1)$ at the

$n^{\underline{th}}$ stage

$$\frac{d\mu_1(x_1,\ldots, x_n)}{d\mu_0(x_1,\ldots, x_n)} = B \qquad \left(\frac{d\mu_1(x_1,\ldots, x_n)}{d\mu_0(x_1,\ldots, x_n)} = A\right) \tag{4}$$

and that this is true for all n. We will assume that for all paths in C_n^* (C_n^{**}) the likelihood ratio at the $n^{\underline{th}}$ stage is constant, but that the constant can be different for each n. Our assumption is meaningful whenever Wald's assumption is. Of course since we consider more general measures there are cases where the assumption will not agree with reality. The following theorem shows how important the assumption is and displays some of the special properties of the Wald test.

<u>Theorem 3.2</u> Assume that

$$\frac{d\mu_1(x_1,\ldots, x_n)}{d\mu_0(x_1,\ldots, x_n)}$$

is constant for all paths in C_n^* and C_n^{**}, and consider alternate hypothesis tests in which the function of the observations at the $m^{\underline{th}}$ stage that is used to perform the estimation is

$$\frac{d\mu_1(x_1,\ldots, x_m)}{d\mu_0(x_1,\ldots, x_m)}$$

Then (2) is satisfied if and only if the test is a Wald test.

Proof:

Suppose we have a Wald test and $(x_1,\ldots, x_n,\ldots) \in C_n^{**}$, then

$$\frac{d\mu_1(x_1,\ldots, x_n)}{d\mu_0(x_1,\ldots, x_n)} = \frac{1-\beta}{\alpha}$$

Integrating over C_n^{**} we have

$$\int_{C_n^{**}} \frac{d\mu_1(x)}{d\mu_0(x)} \; d\mu_0(x) = \frac{1-\beta}{\alpha} \int_{C_n^{**}} d\mu_0(x)$$

so $p_1^{**}(n) = p_0^{**}(n)$. A similar result holds for paths in C_n^{*}.

Now let $p_1^{**}(\neg) = p_0^{**}(n)$. Then $\mu_1^{**}(n) = \frac{1-\beta}{\alpha}\,\mu_0^{**}(n)$.

We write this as

$$\int_{C_n^{**}} \frac{d\mu_1(x)}{d\mu_0(x)} \; d\mu_0(x) = \frac{1-\beta}{\alpha} \int_{C_n^{**}} d\mu_0(x) \qquad (5)$$

where by x we mean the cylinder set represented by $(x_1,\ldots, x_n,\ldots)$.

Since we assume that

$$\frac{d\mu_1(x)}{d\mu_0(x)}$$

is a constant for $x \in C_n^{**}$, it follows that

$$\frac{d\mu_1(x)}{d\mu_0(x)} = \frac{1-\beta}{\alpha}$$

for all tests that lead to D_1 at the $n^{\underline{th}}$ stage. A similar result holds if $p_1^{*}(n) = p_0^{*}(n)$. Thus the test is a Wald test and the theorem is proven.

It is obvious that for any alternate hypothesis test of fixed sample size (2) is satisfied. This shows the importance of the assumption of the constancy of the likelihood ratio for proving the converse of Theorem 3.2.

III. INFORMATION THEORETIC APPROACH

By regarding a statistical test of alternate hypotheses as a problem of transmitting messages over a noisy channel and by defining the information rate per decision we are able to provide additional insight into the nature of these tests. In particular, we are able to interpret the optimality of the Wald test from the information theoretic point of view. We first restate the basic formalism of information theory.(3)

Let X, Y and Z be three discrete random variables which occur together and let $\{x_i\}$ $i=1,\ldots,\ell$ be the set of the ℓ possible different values of X $\{y_j\}$ $j=1,\ldots,m$ be the set of the m possible different values of Y and $\{z_k\}$ $k=1,\ldots,n$ be the set of the n possible different values of Z. Denote the probability of the joint occurence of x_i for X, y_j for Y and z_k for Z by $P\left[X=x_i,\ Y=y_j,\ Z=z_k\right]$. The joint entropy of X, Y and Z is then defined by

$$H(X,Y,Z) = - \sum_{i,j,k=1}^{l,m,n} P\left[X=x_i,\ Y=y_j,\ Z=z_k\right] \log P\left[X=x_i,\ Y=y_j,\ Z=z_k\right] \quad (6)$$

The logarithmic base in this expression and in those that follow is the same but is otherwise arbitrary; the choice of the base corresponds to the choice of a unit for measuring entropy and is usually base 2. A change in the logarithmic base introduces only a mutliplicative scale factor which is of no consequence in this work. The joint entropy of X and Y and the entropy of X alone are defined by

$$H(X,Y) = -\sum_{i,j=1}^{\ell,m} P\left[X=x_i,\ Y=y_j\right]\log P\left[X=x_i,\ Y=y_j\right] \tag{7}$$

and

$$H(X) = -\sum_{i=1}^{\ell} P\left[X=x_i\right]\log P\left[X=x_i\right] \tag{8}$$

in which $P\left[X=x_i,\ Y=y_j\right]$ is the joint probability that $X=x_i$ and $Y=y_j$ and $P\left[X=x_i\right]$ is the probability that $X=x_i$.

Suppose that an information source can, by some random mechanism, generate one of two messages which are indexed 0 and 1. The index of the message actually generated at a particular time is taken as the random variable X. The entropy H(X) is said to measure the amount of information contained in a message generated by that source. If the information source feeds a noisy channel, the receiver at the output of the channel receives the message corrupted by noise. The receiver decodes the message, i.e., estimates whether X was 0 or 1. The estimate of X made by the receiver is the random variable Y which has the same two possible values as X. The entropy H(Y) is said to measure the amount of information generated by the receiver. The rate of transmission R(X;Y) is defined as the sum of the amount of information generated by the source and the amount of information generated by the receiver minus the amount of information common to both the transmitter and the receiver

$$R(X;Y) = H(X) + H(Y) - H(X,Y) \tag{9}$$

If the channel is so noisy that the variables X and Y are independent, $H(X,Y)= H(X)+H(Y)$ and the rate is zero. If the channel is not noisy at all, X is the same as Y, $H(X,Y)= H(X)= H(Y)$ and $R(X;Y)$ is equal to $H(X)$. In effect $R(X;Y)$ measures that part of information generated by the source that must reach the receiver in order that the receiver generate an amount of information $H(Y)$. The quantity $H(X)-R(X;Y)$ is called the "equivocation" of X given Y and measures the amount of unwanted information reaching the receiver that is generated by channel noise.

Next, these basic concepts of the information theory are applied to statistical tests of alternate hypotheses. We have two distinct hypotheses H_0 and H_1 which occur with a priori probabilities π and $1-\pi$. The random variable X is the index of the true hypothesis so that $P\left[X=0\right]=\pi$ and $P\left[X=1\right]=1-\pi$. The statistical test can terminate with the acceptance of the hypothesis H_0 which is the decision D_0, or with the acceptance of the hypothesis H_1 which is the decision D_1. The random variable Y is the index of the accepted decision. If α and β are the specified probabilities of errors, we have $\alpha = P\left[Y=1 \mid X=0\right]$ and $\beta = P\left[Y=0 \mid X=1\right]$. The particular stage N at which the test can terminate is also a random variable depending on the particular sample and on the test procedure. The relevant conditional probabilities that the test will terminate at a particular stage n are denoted by

$$\begin{aligned}
p_0^*(n) &= P\left[N=n \mid X=0,\ Y=0\right] \\
p_1^*(n) &= P\left[N=n \mid X=1,\ Y=0\right] \\
p_0^{**}(n) &= P\left[N=n \mid X=0,\ Y=1\right] \\
p_1^{**}(n) &= P\left[N=n \mid X=1,\ Y=1\right]
\end{aligned} \qquad (10)$$

where n is a positive integer.

The rate R(X;Y,N) of information per decision, i.e., the amount of information which must reach the receiver in order that the estimate Y of X can be achieved with error probabilities α and β at the $N^{\underline{th}}$ stage is, by an extension of (9)

$$R(X;Y,N) = H(X) + H(Y,N) - H(X,Y,N) \qquad (11)$$

The equivocation of X given Y and N, which measures the amount of information required to estimate Y at the $N^{\underline{th}}$ stage, is

$$H(X) - R(X;Y,N) = H(X,Y,N) - H(Y,N)$$

Substituting in (11) from (6), (7), (8) and (10), we find that R(X;Y,N) can be written in the form:

$$R(X;Y,N) = Q(\pi) - \Omega Q\left(\frac{(1-\alpha)\pi}{\Omega}\right) - (1-\Omega)Q\left(\frac{\alpha\pi}{1-\Omega}\right) + G\left[\pi, 1-\alpha, \beta, p_0^*(n), p_1^*(n)\right]$$
$$+ G\left[\pi, \alpha, 1-\beta, p_C^{**}(n), p_1^{**}(n)\right] \qquad (12)$$

in which $Q(t) = -t \log t - (1-t) \log (1-t)$, $\Omega = (1-\alpha)\pi + \beta(1-\pi)$ and $G(\cdot)$ is a function expressing the dependence of R(X;Y,N) on the terminal stage,

$$G\left[\pi, 1-\alpha, \beta, p_0^*(n), p_1^*(n)\right] =$$
$$= \Omega \sum_{n=1}^{\infty} \left\{ - \left[\frac{(1-\alpha)\pi}{\Omega} p_0^*(n) + \frac{\beta(1-\pi)}{\Omega} p_1^*(n) \right] \log \left[\frac{(1-\alpha)\pi}{\Omega} p_0^*(n) + \frac{\beta(1-\pi)}{\Omega} p_1^*(n) \right] \right.$$
$$\left. + \frac{(1-\alpha)\pi}{\Omega} p_0^*(n) \log p_0^*(n) + \frac{\beta(1-\pi)}{\Omega} p_1^*(n) \log p_1^*(n) \right\} \qquad (13)$$

The sum of the last two terms in (12) is in effect R(N;X|Y).

Each term of the summation in (13) is of the form $-\phi(g_1t_1+g_2t_2)+g_1\phi(t_1)+g_2\phi(t_2)$ where $g_1>0, g_2>0, g_1+g_2=1$ and $\phi(t)=t\log t$ is a continuous convex function. We assume that $\alpha<\frac{1}{2}$ and $\beta<\frac{1}{2}$ so that $g_1t_1+g_2t_2$ lies between t_1 and t_2. It follows that each such term is strictly positive and is zero if and only if $p_0^*(n) = p_1^*(n)$ and $p_0^{**}(n) = p_1^{**}(n)$ for all n. When a test procedure is used such that both of these conditions are satisfied, the minimum value of R(X;Y,N) over all possible tests of power (α, β) is achieved and we have

$$\begin{aligned}\text{Min } R(X;Y,N) &= Q(\pi)-\Omega\, Q\left(\frac{1-\alpha}{\Omega}\pi\right) - (1-\Omega)Q\left(\frac{\alpha}{1-\Omega}\pi\right) \qquad (14)\\ &= H(X)-H(X|Y).\end{aligned}$$

An alternate form (14), obtained by rearranging the different terms, is

$$\begin{aligned}\text{Min } R(X;Y,N) &= Q(\Omega)-\pi\, Q(1-\alpha)-(1-\pi)Q(\beta)\\ &= H(Y)-H(Y|X).\end{aligned}$$

These results are expressed in the following theorem:

Theorem 4.1 Among all the procedures for conducting a statistical test of alternate hypotheses, the procedure which is designed to satisfy the conditions $p_0^*(n) = p_1^*(n)$ and $p_0^{**}(n) = p_1^{**}(n)$ for all n requires the minimum rate of information to attain the desired probabilities of error α and β for any a priori probability π and $1-\pi$. This minimum rate is given by (14).

Corollary 4.1 For a sample consisting of independent and identically distributed variables, the Wald test requires the least rate of information to attain the desired probabilities of error α and β for any given *a priori* probabilities π and $1-\pi$.

This Corollary follows from the fact that by Theorem 3.1 the Wald test satisfies the conditions of Theorem 4.1.

An interesting qualitative argument can be based on Theorem 4.1. It is plausible to suppose that the amount of information in a sample is a monotonically increasing function of the average sample size. This assumption together with Theorem 4.1 implies that the test procedure designed to satisfy the conditions of Theorem 4.1 requires, on the average, the smallest average sample size to provide a statistical test with the power (α, β).

The result (14) also implies this additional Theorem:

Theorem 4.2 When the test procedure is designed to satisfy the conditions $p_0^*(n) = p_1^*(n)$ and $p_0^{**}(n) = p_1^{**}(n)$ for all n, the rate of transmission $R(X;Y,N)$ does not depend on the terminal stage N.

Proof:

We observe by writing out $R(X;Y)$ in terms α and β that

$$\text{Min } R(X;Y,N) = R(X;Y).$$

This theorem is a complementary result of the notion of sufficiency discussed in the previous section. Another result which is less obvious can be stated in the form of the following Theorem:

Theorem 4.3 Consider two different test procedures which have probabilities of error less than 0.5. If these test procedures require the same rate of information per decision [but only one procedure is designed to satisfy (2)] the procedure that satisfies (2) cannot have probabilities of error (α, β) both larger than the corresponding probabilities of error of the other test. This holds for any a priori probabilities π and $1-\pi$.

Proof:

Let α' and β' be the probabilities of error of the first and second kind of the test that satisfies (2) and α and β the corresponding probability of the other test. The information rate per decision of the test that satisfies (2) is by (14)

$$R' = Q(\pi) - \Omega' Q\left(\frac{1-\alpha'}{\Omega'}\,\pi\right) - (1-\Omega')Q\left(\frac{\alpha'}{1-\Omega'}\,\pi\right)$$

in which

$$\Omega' = (1-\alpha')\pi + \beta'(1-\pi)$$

The information rate of the other test is given by (12). Since both rates are assumed to be equal and the G(·) functions are positive, we must have

$$Q(\pi) - \Omega\, Q\left(\frac{1-\alpha}{\Omega}\,\pi\right) - (1-\Omega)Q\left(\frac{\alpha}{1-\Omega}\,\pi\right) < Q(\pi) - \Omega' Q\left(\frac{1-\alpha'}{\Omega'}\,\pi\right) - (1-\Omega')Q\left(\frac{\alpha'}{1-\Omega'}\,\pi\right) \tag{15}$$

Suppose we assume that

$$\begin{aligned} 0.5 &> \alpha' \geq \alpha \\ 0.5 &> \beta' \geq \beta \end{aligned} \tag{16}$$

This assumption implies the following inequalities (see Fig. 1)

$$0 < \frac{(1-\alpha)}{\Omega}\pi \le \frac{(1-\alpha')}{\Omega'}\pi < \pi < \frac{\alpha'}{1-\Omega'}\pi \le \frac{\alpha}{1-\Omega}\pi < 1$$

for any $0 < \pi < 1$. Let P_1, P_1', P, P_2', P_2 be the points on the curve $y = Q(t)$ corresponding to $(1-\alpha)\pi/\Omega$, $(1-\alpha')\pi/\Omega'$, π, $\alpha'\pi/(1-\Omega')$, $\alpha\pi/1-\Omega$. Since $Q(t)$ is continuous, concave and non-linear, the chord $P_1' P_2'$ lies above the chord P_1P_2 except when both equality signs in (16) hold and the chords coincide. Suppose R and R' are the points on the chords P_1P_2 and $P_1' P_2'$ corresponding to $t=\pi$. Then P lies above R' and R' lies above R except when $\alpha=\alpha'$ and $\beta=\beta'$, in which case R and R' coincide. Let $\overline{PR}$ be the distance from P to R. The inequality (16) therefore implies $\overline{PR} \ge \overline{PR'}$. But the left-hand side of (15) is the distance $\overline{PR}$ and the right-hand side of (15) is the distance $\overline{PR'}$. Thus (15) represents the inequality $\overline{PR} < \overline{PR'}$ which therefore cannot be achieved under condition (16). Conversely the inequality (16) contradicts (15).

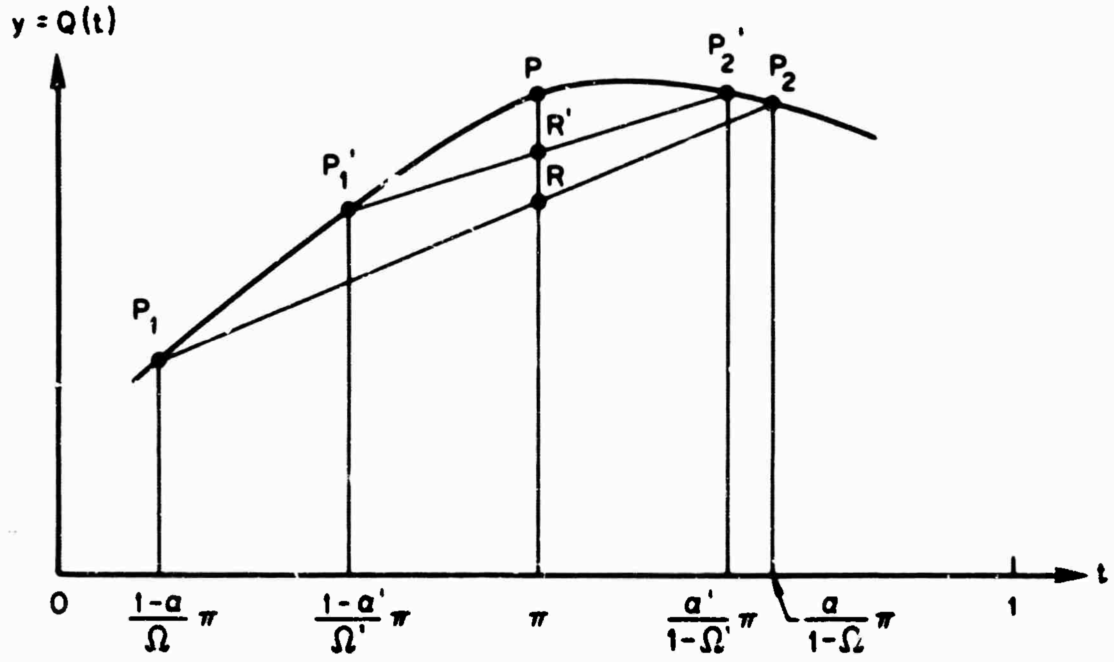

Fig. 1--Geometrical Relationships for the Proof of Theorem 4.2

IV. CONCLUSIONS

It is important to notice the difference between the usual communication problem and the decision problem. In the communication problem the channel is specified, and one desires to maximize the rate of transmission. This is achieved through the coding of messages which is said to match the information source to the channel. The maximum rate of transmission (with respect to all the admissible sources) that can be achieved for a particular channel is known as the capacity of the channel. By contrast, in the decision problem the experimenter assumes a priori the hypotheses H_0 and H_1, but not the test procedure; thus the information source rather than the channel is specified. The test procedure (i.e., the test statistic and the decision regions) that plays the part of the channel can be chosen by the experimenter. The probabilities of error determine the amount of information which must be generated by the receiver. The relevant design problem is now to select that test procedure that requires least information to complete the test, i.e., that minimizes the rate of transmission. We might consider this as the problem of matching the channel to the source.

We find that the Wald test not only minimizes the average risk but also minimizes the rate of transmission independently of the a priori probabilities. The proof of the optimality of the Wald test in the sense of minimum average risk applies only to the alternate hypotheses tests on identically distributed, independent samples.(4) It is suggestive to apply the Theorem 4.1 to the design of multi-stage statistical tests of alternate hypotheses even in the case of correlated and non-identically distributed observations by requiring that the test procedure be constructed to satisfy (2). This rule of construction would determine the boundaries of the proper decision regions which need not be parallel lines. Another extension of Theorem 4.1 applies to the design of multi-stage statistical tests of multiple hypotheses where by analogy to the case of two hypotheses, the optimum decision rule would be specified by the relevant equalities among conditional probabilities at each stage.

REFERENCES

1. Middleton, D., and D. Van Meter, "Detection and Extraction of Signals in Noise from the Point of View of Statistical Decision Theory II," J. SIAM, Vol. 4, No. 2, June 1956, pp. 86-119.

2. Lehmann, E. Testing Statistical Hypothesis, John Wiley & Sons, Inc., New York, 1959, p. 20.

3. Shannon, C. E., and W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, Ill, 1949.

4. Wald, A., and J. Wolfowitz, "Optimum Character of the Sequential Probability Ratio Test," Ann. Math. Stat., Vol. 19, No. 3, September 1948, p. 326.