

INFORMATION THEORY
AND DECISION MAKING

MINORU SAKAGUCHI

Visiting Professor of Statistics, The George Washington University
National Science Foundation Fellow, Senior Foreign Scientist
Assistant Professor Mathematics, University of Electro-Communications,
Tokyo, Japan.

5/21/64
AKS

STATISTICS DEPARTMENT
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON D.C. 20006
June 1964

Copy 11

INFORMATION THEORY
AND DECISION MAKING

MINORU SAKAGUCHI

Visiting Professor of Statistics, The George Washington University
National Science Foundation Fellow, Senior Foreign Scientist
Assistant Professor Mathematics, University of Electro-Communications,
Tokyo, Japan.

STATISTICS DEPARTMENT
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON D.C. 20006

June 1964

FOREWORD

This volume is a revised edition of a series of Seminar lecture notes on Information Theory and Decision Making presented by Dr. Sakaguchi in the Statistics Department of The George Washington University during the academic year 1963-1964.

Support for Dr. Sakaguchi as Visiting Professor of Statistics, his research program, and Seminar was generously provided by The National Science Foundation under its program of Fellowships For Senior Foreign Scientists. We take this opportunity to express our appreciation on behalf of The George Washington University, The Statistics Department, and Dr. Sakaguchi.

Support for this publication was provided by a grant from the Research Council of The George Washington University. The support of Dean A. E. Burns of the Graduate Council and Professor H. F. Bright, Chairman of the Department of Statistics is gratefully acknowledged.

The generous support by so many people within and outside The George Washington University which made the program with Dr. Sakaguchi possible was most encouraging and has contributed immensely to the research and teaching program of the Statistics Department. It was a real pleasure for me personally to participate in Dr. Sakaguchi's Seminar and work with him in preparing the present publication.



S. Kullback
Professor of Statistics

TABLE OF CONTENTS

Introduction	1
Section 1 - Fisher's amount of information and statistical estimation.	2
Section 2 - Kullback-Leibler's information and testing simple hypotheses.	9
Section 3 - Kullback-Leibler information and testing composite hypotheses.	36
Section 4 - Kullback-Leibler information and sequential tests of hypotheses.	46
Section 5 - Statistical experiments and information provided by them	59
Section 6 - Comparison of experiments.	79
Section 7 - Sequential design of experiments.	107
Section 8 - Capacity of statistical experiments.	142
Appendix A	160
Appendix B	165
Appendix C	175
Appendix D	187
References	198

INTRODUCTION

I believe that the concept of information is very important in almost all branches of research in natural science, just as is the concept of potential or energy. Generally speaking, every time when we make an observation or perform an experiment, we draw and use some information provided by it. How much can we find on the basis of the observed data or obtained experimental results concerning the problem about which we are seeking? Of course these concepts are highly abstract in nature, but various approaches have been made to provide a quantitative analysis of information in the last 30 years. Especially, the rapid progress of modern mathematical statistics has provided a useful and powerful tool for the development of this analysis.

1. Fisher's amount of information and statistical estimation.

The most classic problem in mathematical statistics is that of estimating statistical parameters. R. A. Fisher proposed that a desirable statistic used as an estimator of a statistical parameter must possess the following properties:

- (a) unbiasedness
- (b) consistency
- (c) asymptotic efficiency, i.e., yielding minimum variance σ^2 among all statistics which are, when properly normalized, asymptotically distributed as $N(\theta, \sigma^2)$,

and he introduced the concept of sufficient statistics to derive the estimator satisfying these properties. Let the likelihood of the sample be $L(x; \theta)$, the p.d.f. of the statistic T be $\Phi(t; \theta)$. We call the statistic $T(x)$ the sufficient statistic if and only if we have

$$(1.1) \quad L(x; \theta) = \Phi(t; \theta)M(x),$$

where $M(x)$ is a function of x only and independent of θ , $\Phi(t; \theta)$ is a function of θ and the realized value t of T depending on x only through $t = T(x)$. Let $T_1(x)$ be a sufficient statistic and $T_2(x)$ be another statistic which is not a function of $t_1 = T_1(x)$. After making a change of variables

$y_1 = T_1(x_1, \dots, x_n)$, $y_2 = T_2(x_1, \dots, x_n)$, $y_3 = x_3, \dots, y_n = x_n$ in the identity

$$L(x; \theta) dx = \Phi_1(t_1; \theta)M_1(x)dx$$

and integrating both sides with respect to y_3, \dots, y_n we get the following equation as the simultaneous probability element of T_1, T_2 ,

$$(1.2) \quad \Phi(t_1, t_2; \theta) dt_1 dt_2 = \Phi_1(t_1; \theta) M(t_1, t_2) dt_1, dt_2,$$

which means that the conditional probability distribution of T_2 given $T_1 = t_1$ is independent of the unknown true value of parameter θ .

And we can say:

The realized value t_1 of the sufficient statistic T_1 gives all the information which the sample x can provide about the unknown parameter θ , and no information can be added even if the value of any other statistic T_2 (of course not a function of T_1) be known.

Since a sufficient statistic exhausts in the above sense the information which the sample can give, it plays a fundamental role in every statistical procedure, not only in statistical estimation. Then what is the advantage of using a sufficient statistic as an estimator of a statistical parameter? We shall consider the problem of point estimator only. The following theorems will answer the above question.

Let the p.d.f. of the population distribution be $f(x, \theta)$. We assume a set of appropriate regularity conditions which guarantee validity of interchanging the order of integration with respect to x and differentiation with respect to θ about some quantities such as $f(x, \theta)$, $\frac{\partial}{\partial \theta} f(x, \theta)$ and so on. We call this set of conditions the regularity assumption (R). Then we have

Theorem 1.1 Under the regularity assumption (R), the variance $V_\theta(T)$ of an unbiased estimator $T(x_1, \dots, x_n)$ of the parameter θ satisfies the inequality

$$(1.3) \quad V_{\theta}(T) \geq I(\theta)^{-1},$$

where

$$(1.4) \quad I(\theta) \equiv E_{\theta} \left[- \frac{\partial^2}{\partial \theta^2} \log L(X_1, \dots, X_n; \theta) \right].$$

If T is a function of a sufficient statistic, the equality holds in

(1.3) identically. (Cramér-Rao inequality)

Theorem 1.2 We assume the existence of an unbiased estimator and a sufficient statistic. Let T be a sufficient statistic. If the efficient estimator, i.e., the estimator with uniformly minimum variance among those which are unbiased and have finite variances, exist, then it is a function of T. (Fundamental theorem of estimation)

Since $L(x; \theta) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta)$ we have by (1.4)

$$(1.5) \quad I(\theta) = ni(\theta)$$

where

$$(1.6) \quad i(\theta) \equiv E_{\theta} \left[- \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) \right].$$

The function $i(\theta)$ is a function proper to the set $\{f(x, \theta) | \theta \in \Theta\}$ of the given family of p.d.f.'s which characterizes the estimation problem together with the criterion measure of error in estimation; Theorems 1.1 and 1.2 state that we can measure the best-possible accuracy of our estimation problem by use of $i(\theta)$. Fisher thus coined for this the name "intrinsic accuracy". When $i(\theta)$ is large we can estimate accurately in the sense of theorem 1.1. If the assumption in theorem 1.2 be satisfied, and if we adopt an unbiased sufficient estimator T_1 , then it exhausts the information which the sample can give, and the uniform inequality in (1.3) holds. In this case we have by (1.1), (1.3), and

(1.4)

$$V_{\theta}(T) \geq I(\theta)^{-1} = (E_{\theta}[-\frac{\partial^2}{\partial \theta^2} \log \phi(T_1, \theta)])^{-1} = V_{\theta}(T_1).$$

Here we regard the quantity $E_{\theta}[-\frac{\partial^2}{\partial \theta^2} \log \phi(T_1, \theta)]$ as expressing how efficient the unbiased sufficient statistic T_1 is in the point estimation of θ . Let us define for any estimator T

$$(1.7) \quad I(T) \equiv E_{\theta}[-\frac{\partial^2}{\partial \theta^2} \log \phi(T, \theta)]$$

where ϕ is the p.d.f. of T . Then we can see that if we equate the efficiency of an estimator to a large amount of information which the estimator provides concerning the point estimation of the parameter, the introduction of this quantity (1.7) is quite reasonable.

That is, we have

Theorem 1.3 Under the regularity assumption (R), we have

- (i) $I(T) \geq 0$, and the equality holds if and only if the distribution of T is independent of θ ,
- (ii) $I(T_1) \leq I(T_1, T_2)$, and the equality holds if and only if the conditional distribution of T_2 given the realized value of T_1 is independent of θ ,
- (iii) $I(T_1, T_2) = I(T_1) + I(T_2)$, if T_1 and T_2 are mutually independent,
- (iv) $I(T) \leq I(X_1, \dots, X_n)$, and the equality holds if and only if T is a sufficient statistic.

Now that we have been making clear the important role of sufficient statistics, several questions will arise, as a practical matter, of how

to derive these sufficient statistics, how to choose the "best" sufficient statistic among many of those etc. It is one of the most interesting facts in mathematical statistics that we can derive a reasonable estimator automatically by use of the maximum-likelihood method. The following theorems will show some connections of the maximum-likelihood principle with the concept of information.

We shall say that a non-trivial solution of the likelihood equation $\frac{\partial \log L}{\partial \theta} = 0$ is a quasi-maximum-likelihood estimator. Then under (R) we have

Theorem 1.4 If a sufficient statistic T exists, every q.m.l. estimator is a function of T.

Theorem 1.5 Some q.m.l. estimator has the consistency property (b):

For any $\delta > 0$ we have

$$\lim_{n \rightarrow \infty} P_{\theta_0} \left\{ \frac{\partial \log L}{\partial \theta} = 0 \text{ has a root in } (\theta_0 - \delta, \theta_0 + \delta) \right\} = 1$$

(Duguè, 1937)

Theorem 1.6 The q.m.l. estimator, stated in the above theorem, is asymptotically distributed as $N(\theta_0, I(\theta_0)^{-1})$.

By theorem 1.1 and theorem 1.5, theorem 1.6 means that some q.m.l. estimator has the asymptotic efficiency (c). Moreover we have

Theorem 1.7 Let us assume that (R) be satisfied and that a sufficient statistic exists. Let $\hat{\theta}(x)$ be a q.m.l. estimator, and let ρ be the radius of curvature at the point $\theta = \hat{\theta}(x)$ of the curve $y = \log L(x; \theta)$.

Then we have

$$(1.8) \quad \rho^{-1} = E_{\theta} \left[- \frac{\partial^2}{\partial \theta^2} \log L \right] \Big|_{\theta = \hat{\theta}} = I(\hat{\theta}) \quad (\text{Huzurbazar, 1949})$$

Proof. Since (R) is satisfied and a sufficient statistic exists we have by the well-known Koopman's theorem

$$\log f(x; \theta) = K(\theta)u(x) + a(x) + b(\theta)$$

and hence

$$\log L(x; \theta) = K(\theta)t(x) + A(x) + B(\theta),$$

where $t(x) = \sum_1^n u(x_i)$ is a sufficient statistic, $K(\theta)$ and $B(\theta)$ are some functions of θ , and $A(x)$ is a function of x . If we fix x , we get

$$0 = \frac{\partial \log L(x, \theta)}{\partial \theta} \Big|_{\theta = \hat{\theta}} = K'(\hat{\theta})t(x) + B'(\hat{\theta})$$

$$\frac{\partial^2 \log L(x, \theta)}{\partial \theta^2} \Big|_{\theta = \hat{\theta}} = K''(\hat{\theta})t(x) + B''(\hat{\theta})$$

And we have

$$0 = E_{\theta} \left[\frac{\partial \log L(x, \theta)}{\partial \theta} \right] = K'(\theta)E_{\theta}(t(X)) + B'(\theta)$$

$$-I(\theta) = E_{\theta} \left[\frac{\partial^2 \log L(x, \theta)}{\partial \theta^2} \right] = K''(\theta)E_{\theta}(t(X)) + B''(\theta)$$

Eliminating $t(x)$ and $E_{\theta}(t(X))$ from the above four equations we easily obtain the equality

$$- \frac{\partial^2 \log L(x, \theta)}{\partial \theta^2} \Big|_{\theta = \hat{\theta}} = E_{\theta} \left[- \frac{\partial^2 \log L(x, \theta)}{\partial \theta^2} \right] \Big|_{\theta = \hat{\theta}}$$

We thus have

$$\rho^{-1} = \left[\frac{-\partial^2 \log L / \partial \theta^2}{\{1 + (\partial \log L / \partial \theta)^2\}^{3/2}} \right]_{\theta = \hat{\theta}} = - \frac{\partial^2 \log L}{\partial \theta^2} \Big|_{\theta = \hat{\theta}}$$

$$= E_{\theta} \left[- \frac{\partial^2 \log L}{\partial \theta^2} \right] \Big|_{\theta = \hat{\theta}} = I(\hat{\theta})$$

This completes the proof.

Corollary to Theorem 1.7. Under the assumptions above, the solution of the likelihood equation is unique and maximizes the likelihood. (It is no longer "quasi"-maximal).

Proof. Every solution of the likelihood equations gives a stationary maximum. Hence, if there were two or more distinct solutions all would be stationary maxima, and between two stationary maxima we should have a stationary minimum, under (R). But there is no solution which gives a minimum. Several examples of intrinsic accuracy $i(\theta)$ are listed in the following.

	$f(x, \theta)$	$i(\theta)$
binomial:	$\binom{N}{x} \theta^x (1-\theta)^{N-x}$	$N/\{\theta(1-\theta)\}$
Poisson:	$e^{-\theta} \theta^x / (x!)$	θ^{-1}
normal:	$(2\pi)^{-1/2} e^{-(x-\theta)^2/2}$	1
normal:	$(2\pi)^{-1/2} e^{-x^2/(2\theta)}$	$2\theta^{-1}$
exponential:	$\theta^{-1} e^{-x/\theta}$	θ^{-2}
gamma:	$x^{\theta-1} e^{-x} / \Gamma(\theta)$	$\partial^2 \log \Gamma(\theta) / \partial \theta^2$
beta:	$\{x(1-x)\}^{\theta-1} / B(\theta, \theta)$	$\partial^2 \log B(\theta, \theta) / \partial \theta^2$

where we have set $\Gamma(\theta) \equiv \int_0^{\infty} x^{\theta-1} e^{-x} dx$ and $B(\theta, \theta) \equiv \int_0^1 x^{\theta-1} (1-x)^{\theta-1} dx$.

2. Kullback-Leibler's information and testing simple hypotheses.

Consider the probability spaces $(\mathcal{X}, \mathcal{B}, \mu_i)$ $i=0,1$, where μ_i is a probability measure defined over the measurable space $(\mathcal{X}, \mathcal{B})$. We assume that $\mu_0 \equiv \mu_1$ (absolutely continuous with respect to each other). Let λ be a probability measure such that $\mu_i \equiv \lambda$ ($i=0,1$) and $f_i(x)$, $i=0,1$ are generalized probability densities with respect to λ measure:

$$\mu_i(E) = \int_E f_i(x) d\lambda, \quad \text{for all } E \in \mathcal{B}, \quad (i=0,1)$$

The easiest problem, probably, in statistical inference is that of testing a simple hypothesis against a simple alternative. Suppose that the hypothesis H_0 specifies that n independently distributed observations X_1, \dots, X_n have density $f_0(x)$ whereas the alternative H_1 specifies the density $f_1(x)$.

In testing statistical hypotheses we generally have two types of errors, that is,

- (a) Type I error: the error of accepting H_1 when H_0 is true,
- (b) Type II error: the error of accepting H_0 when H_1 is true.

Since we cannot derive a test which simultaneously minimizes the probabilities of both types of errors, we try to derive

- (A) a test which minimizes the probability of one type of error among all other tests with fixed probability of another type of error,
- (B) a test which minimizes some weighted averages of probabilities of the two types of errors,

or

(C) a test which minimizes the maximum of the probabilities of the two types of errors.

A leading principle in this test problem is the so-called likelihood-ratio principle. This principle insists that when we have a sample $X = (X_1, \dots, X_n)$, the sample should be considered to support the hypothesis H_1 if the likelihood-ratio

$$L_1(X)/L_0(X) = \prod_{i=1}^n \{f_1(X_i)/f_0(X_i)\}$$

is large, and to support the hypothesis H_0 if the likelihood-ratio is small. It is well known that the class of best tests are "likelihood-ratio tests" characterized by critical regions (i.e. the rejection region of the null hypothesis H_0) which are given by

$$\{x | L_1(x)/L_0(x) \geq a\}$$

for some positive constants a .

Kullback-Leibler defined "the mean information numbers for discriminating two probability densities" as

$$(2.1) \quad \begin{cases} I(0:1) \equiv E_0 \left[\log \frac{f_0(X)}{f_1(X)} \right] = \int f_0(x) \log \frac{f_0(x)}{f_1(x)} d\lambda \\ I(1:0) \equiv E_1 \left[\log \frac{f_1(X)}{f_0(X)} \right] = \int f_1(x) \log \frac{f_1(x)}{f_0(x)} d\lambda \end{cases}$$

It should be mentioned here that, if ζ and $1-\zeta$ are the a priori probabilities that H_0 and H_1 are the true hypotheses respectively, we have

$$\log \frac{f_1(x)}{f_0(x)} = -\log \frac{1-\zeta}{\zeta} - \left\{ -\log \left(\frac{(1-\zeta)f_1(x)}{\zeta f_0(x) + (1-\zeta)f_1(x)} \bigg/ \frac{\zeta f_0(x)}{\zeta f_0(x) + (1-\zeta)f_1(x)} \right) \right\}$$

in which the first and second terms in the right-hand side express logarithms of the prior probability ratio and the posterior probability ratio, respectively.

The numbers defined by (2.1) are really "information numbers" since we have

Theorem 2.1 (i) $I(0:1) \geq 0$ with equality if and only if $f_0(x) = f_1(x)[\lambda]$.

A similar result holds true for $I(1:0)$.

(ii) For independent random variables we have

$$\int f_i(x,y) \log \frac{f_1(x,y)}{f_0(x,y)} d\lambda(x,y) = \int f_i^{(1)}(x) \log \frac{f_1^{(1)}(x)}{f_0^{(1)}(x)} d\lambda^{(1)}(x) + \int f_i^{(2)}(y) \log \frac{f_1^{(2)}(y)}{f_0^{(2)}(y)} d\lambda^{(2)}(y) \quad (i=0,1)$$

where $f_i(x,y) = f_i^{(1)}(x) f_i^{(2)}(y)$ ($i=0,1$), $\lambda^{(1)}(x) = \int f_1^{(2)}(y) d\lambda(x,y)$,

and $\lambda^{(2)}(y) = \int f_1^{(1)}(x) d\lambda(x,y)$.

(iii) Diminishing property under transformations: if $g_i(y)$ is the p.d.f. of a statistic $y = T(x)$ under each hypothesis H_i , then

$$(2.2) \quad \int f_1(x) \log \frac{f_1(x)}{f_0(x)} d\lambda \geq \int g_1(y) \log \frac{g_1(y)}{g_0(y)} d\lambda T^{-1}(y)$$

with equality if and only if T is sufficient, i. e.

$$(2.3) \quad f_1(x)/f_0(x) = g_1 T(x)/g_0 T(x) \quad [\lambda]$$

Proof (i) follows from the inequality $Z \log(Z/Z') \geq Z - Z'$ for any $Z, Z' \geq 0$.

(ii) is evident.

In (2.2) we have

(the left-hand side) - (the right-hand side)

$$= \int f_1(x) \log(f_1(x)/f_0(x)) d\lambda - \int f_1(x) \log(g_1 T(x)/g_0 T(x)) d\lambda$$

$$= \int f_1 \log \frac{f_1}{f_0 \cdot g_1 T / g_0 T} d\lambda \geq 0$$

since

$$\int f_0(x) \cdot (g_1 T(x)/g_0 T(x)) d\lambda = \int g_1(y) d\lambda T^{-1}(y) = 1.$$

Corollary to theorem 2.1. For any set $E \in \mathfrak{B}$ with $\lambda(E) > 0$, we have

$$\int_E f_0(x) \log \frac{f_0(x)}{f_1(x)} d\lambda \geq \mu_0(E) \log \frac{\mu_0(E)}{\mu_1(E)}$$

with equality if and only if $\frac{f_0(x)}{f_1(x)} = \text{const.}$ $[\lambda]$ in E .

Proof. Use Theorem 2.1(i) for $g_i(x) \equiv f_i(x)/\mu_i(E)$ ($i=0,1$).

An immediate consequence is that if $\{E_j\}$ is a finite or infinite partition of \mathfrak{X} into pairwise disjoint sets,

$$I(0:1) \geq \sum_j \mu_0(E_j) \log \frac{\mu_0(E_j)}{\mu_1(E_j)}$$

with equality if and only if $\frac{f_0(x)}{f_1(x)} = \frac{\mu_0(E_j)}{\mu_1(E_j)}$ $[\lambda]$ in E_j ($j=1,2,\dots$).

In other words the grouping of observations generally causes a loss of information and the information is not diminished by the grouping if and only if the conditional density of x given E_j is the same under both hypotheses.

It is to be noted that K-L information numbers are not irrelevant to Fisher's intrinsic accuracy since we have

$$(2.4) \quad I(\theta:\theta+\Delta\theta) = \int f(x,\theta) \log \frac{f(x,\theta)}{f(x,\theta+\Delta\theta)} d\lambda = \frac{1}{2} i(\theta) \Delta\theta^2 + o(\Delta\theta^3).$$

Now let us see in the following in what sense the K-L information numbers measure the information for discriminating between two statistical hypotheses. Let us treat H_0 as the null hypothesis, and W as the critical region. The type I and type II errors are

$$\alpha = \Pr\{x \in W | H_0\} \quad \text{and} \quad \beta = \Pr\{x \in W^c | H_1\}, \text{ respectively.}$$

We now have

Theorem 2.2

$$(2.5) \quad \begin{cases} (i) & I(0:1) \geq \left(\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta} \right) / n \\ & I(1:0) \geq \left(\beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha} \right) / n \end{cases}$$

(ii) Let $\alpha_n^*(\beta_n^*)$ be the minimum possible value of $\alpha(\beta)$ for a fixed value of $\beta(\alpha)$. Then we have

$$(2.6) \quad \begin{cases} \lim_{n \rightarrow \infty} \left(-\frac{1}{n} \log \alpha_n^* \right) = I(1:0) \\ \lim_{n \rightarrow \infty} \left(-\frac{1}{n} \log \beta_n^* \right) = I(0:1) \end{cases}$$

This theorem expresses, in the second part, that the most powerful test of H_0 against the alternative H_1 with the critical level α has the consistency property (i.e. the probability of the type II error tends to 0 as $n \rightarrow \infty$) and that the order of consistency will be determined by the K-L information number $I(0:1)$.

Proof. (i) Denoting the likelihood function of the sample by $L_i(x) = L_i(x_1, \dots, x_n)$ ($i=0,1$) we have

$$\begin{aligned} \int L_i(x) \log \frac{L_i(x)}{L_{1-i}(x)} d\lambda^n &= n \int f_i(x) \log \frac{f_i(x)}{f_{1-i}(x)} d\lambda \\ &\geq n \left(\mu_i(W) \log \frac{\mu_i(W)}{\mu_{1-i}(W)} + \mu_i(W^c) \log \frac{\mu_i(W^c)}{\mu_{1-i}(W^c)} \right) \end{aligned}$$

by the corollary to Theorem 2.1.

(ii) We shall prove the first half of (2.6) only. If $I(1:0) < \infty$, we have by the weak law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f_1(X_i)}{f_0(X_i)} \xrightarrow{(n \rightarrow \infty)} I(1:0) \text{ in probability under } H_1,$$

that is, for any $\epsilon, \delta > 0$ and for sufficiently large n

$$\Pr \left\{ \frac{L_1(X)}{L_0(X)} \geq e^{n(I(1:0) - \epsilon)} \mid H_1 \right\} \geq 1 - \beta_0$$

$$\Pr \left\{ \frac{L_1(X)}{L_0(X)} \leq e^{n(I(1:0) + \epsilon)} \mid H_1 \right\} \geq 1 - \delta.$$

With $W_1 \equiv \{(x_1, \dots, x_n) \mid \frac{L_1(x)}{L_0(x)} \geq e^{n(I(1:0) - \epsilon)}\}$ we have

$$1 \geq \Pr\{W_1 \mid H_1\} \geq e^{n(I(1:0) - \epsilon)} \Pr\{W_1 \mid H_0\} \geq e^{n(I(1:0) - \epsilon)} \alpha_n^*.$$

$$(*) \quad \therefore \lim_{n \rightarrow \infty} \frac{1}{n} \log \alpha_n^* \geq I(1:0).$$

With $W_2 \equiv \{(x_1, \dots, x_n) \mid \frac{L_1(x)}{L_0(x)} \leq e^{n(I(1:0) + \epsilon)}\}$ we have

$$\begin{aligned} \Pr\{W_1 \cap W_2 \mid H_1\} &= 1 - \Pr\{W_1^c \cup W_2^c \mid H_1\} \geq 1 - \Pr\{W_1^c \mid H_1\} - \Pr\{W_2^c \mid H_1\} \\ &= \Pr\{W_1 \mid H_1\} - \Pr\{W_2^c \mid H_1\} \geq 1 - \beta_0 - \delta \end{aligned}$$

and

$$\Pr\{W_1 \cap W_2 | H_1\} \leq e^{n(I(1:0)+\epsilon)} \Pr\{W_1 \cap W_2 | H_0\},$$

so that

$$\overline{\lim}_{n \rightarrow \infty} -\frac{1}{n} \log \alpha_n^* \leq I(1:0)$$

Combining this with (*) we obtain the first half of (2.6).

We shall next state an important theorem based on the criterion (B) mentioned earlier.

Theorem 2.3 Let ζ be the a priori probability that H_0 is true. For the Bayes test with the rejection region

$$W_\zeta \equiv \{x | \zeta L_0(x) < (1-\zeta)L_1(x)\}$$

of H_0 , we have

$$(2.7) \quad r(W_\zeta) \equiv \zeta \int_{W_\zeta} L_0(x) d\lambda^n + (1-\zeta) \int_{W_\zeta^c} L_1(x) d\lambda^n \leq \rho^n$$

uniformly in ζ , where

$$(2.8) \quad \rho \equiv \inf_{0 < t < 1} \int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda$$

(Joshi, 1957)

Proof. We have

$$\begin{aligned} \rho^n &= \left(\inf_{0 < t < 1} \int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda \right)^n \\ &= \inf_{0 < t < 1} \left(\int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda \right)^n \\ &= \inf_{0 < t < 1} \int [L_1(x)]^t [L_0(x)]^{1-t} d\lambda^n \equiv \inf_{0 < t < 1} h(t), \end{aligned}$$

say, and

$$\begin{aligned}
h(t) &= \int [L_1(x)]^t [L_0(x)]^{1-t} d\lambda^n \\
&= \int_{W_\zeta} \left[\frac{L_1(x)}{L_0(x)} \right]^t L_0(x) d\lambda^n + \int_{W_\zeta^c} \left[\frac{L_0(x)}{L_1(x)} \right]^{1-t} L_1(x) d\lambda^n \\
&\geq \left(\frac{\zeta}{1-\zeta} \right)^t \int_{W_\zeta} L_0(x) d\lambda^n + \left(\frac{1-\zeta}{\zeta} \right)^{1-t} \int_{W_\zeta^c} L_1(x) d\lambda^n \\
&\geq r(W_\zeta) \quad (\because 0 < (1-\zeta)^t \zeta^{1-t} < 1, \text{ if } 0 < t < 1).
\end{aligned}$$

By the well-known Hölder's inequality

$$\left| \int f(x)g(x) d\lambda \right| \leq \left(\int |f(x)|^p d\lambda \right)^{1/p} \left(\int |g(x)|^q d\lambda \right)^{1/q}$$

($p^{-1} + q^{-1} = 1$; $p, q > 1$), we have $\int f_1^t f_0^{1-t} d\lambda \leq \left(\int f_1 d\lambda \right)^t \cdot \left(\int f_0 d\lambda \right)^{1-t} = 1$,

and hence, by (2.8), $0 \leq \rho \leq 1$ with the first equality if and only if $f_0(x) = f_1(x)$ [λ]. Theorem 2.3 shows that the Bayes test with respect to any prior distribution is consistent (i.e. the weighted average $r(W_\zeta)$ of the two error probabilities tends to 0 as $n \rightarrow \infty$) and that the order of consistency will be determined by the number ρ defined by (2.8) or equivalently

$$(2.8') \quad -\log \rho = -\log \left(\inf_{0 < t < 1} \int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda \right).$$

The above number was first defined by H. Chernoff (1952), and so we shall hereafter call this the Chernoff information number. The Chernoff information number is symmetric, that is, if we denote (2.8') by $I(0,1)$ then $I(0,1) = I(1,0)$. And we have

Theorem 2.4

(i) $-\log \rho \geq 0$, with equality if and only if $f_0(x) = f_1(x)[\lambda]$.

(ii) Additive for independent and identically distributed random

variables: if $-\log \rho(X_1, X_2)$ is the Chernoff information number

corresponding to $L_i(x_1, x_2) = f_i(x_1)f_i(x_2)$ ($i=0,1$), then $-\log \rho(X_1, X_2) = -2\log \rho$.

(iii) If $y = T(x)$ is a sufficient statistic, i.e., its p.d.f. $g_i(y)$

($i=0,1$) under each hypothesis H_i satisfies $f_1(x)/f_0(x) = g_1 T(x)/g_0 T(x)[\lambda]$,

then $-\log \rho_T = -\log \rho$, where $-\log \rho_T$ is the Chernoff information number
corresponding to $g_i(y)$'s.

(Chernoff, 1952)

It should be remarked here that although the Chernoff information number is additive for independent and identically distributed observations, it is not additive for independent but non-identically distributed observations. Let (X, Y) represent an observation consisting of independent but differently distributed random variables X and Y . Hence the densities of (X, Y) under H_i ($i=0,1$) has the form $f_i(x)g_i(y)$ and it is easy to see from (2.8) that

$$-\log \rho(X, Y) \leq -\log \rho_X - \log \rho_Y.$$

On the other hand K-L information numbers $I(0:1)$ and $I(1:0)$ yield equality in the above relation, i.e.

$$I(i:1-i|X, Y) = I(i:1-i|X) + I(i:1-i|Y) \quad (i=0,1).$$

It is of interest to notice the connection between the two information numbers of K-L and Chernoff. Let Z be a random variable with the assumption

$M(t) \equiv E(e^{tZ}) < \infty$, for t in some neighborhood of 0.

Then we easily have

$$M(0) = 1, \quad M'(0) = Ez$$

and $M(t)$ is convex (Fig. 2.1).

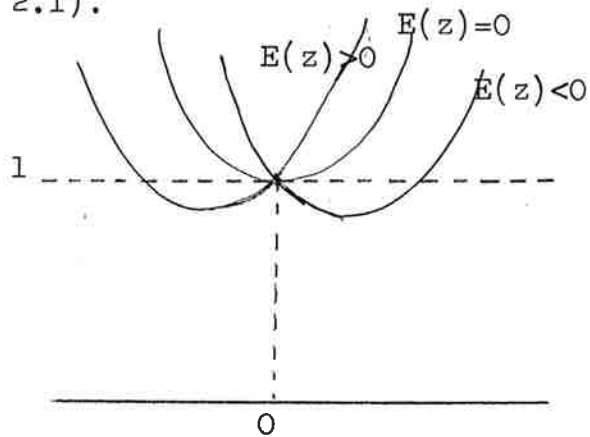


Figure 2.1 Graphs of $M(t)$

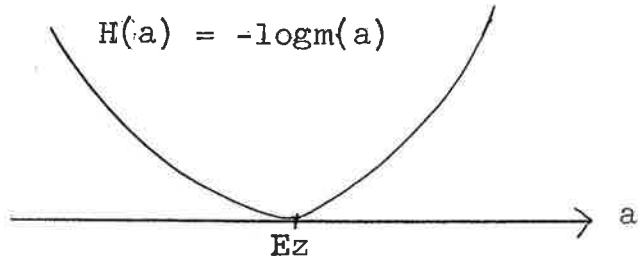


Figure 2.2

Theorem 2.5 Let

$$m(a) \equiv \inf_t e^{-ta} M(t) = \inf_t E(e^{t(Z-a)})$$

$$H(a) \equiv -\log m(a) = -\log \left[\inf_t E(e^{t(Z-a)}) \right]$$

Then we have (fig. 2.2)

(i)

$$(2.9) \quad H(a) = at(a) - \log M(t(a)) \quad \text{where } t(a) \text{ is determined by the equation}$$

$$(2.10) \quad a = \frac{M'(t(a))}{M(t(a))}$$

(ii) $H(a)$ is convex

(iii) $H(Ez) = 0$, $H'(Ez) = 0$, $H''(Ez) = 1/\text{Var}z$

Proof. Differentiating $e^{-ta}M(t)$ with respect to t and setting the derivative equal to 0 gives (i).

(ii) By (2.9) and (2.10)

$$H'(a) = \frac{d}{da} (t(a)a - \log M(t(a))) = t(a),$$

$$H''(a) = t'(a) = \left[\frac{M(t)^2}{M'(t)M(t) - M'(t)^2} \right]_{t=t(a)} > 0$$

by (2.10) and Schwarz' inequality,

Since $t(a)$ is monotone increasing, we have by (2.10)

$$t(a) = 0 \iff a = Ez.$$

If we take

$$Z = \text{Log}(f_1(x)/f_0(x))$$

then we have that

$$M_0(t) \equiv E_0(e^{tz}) = E_0 \left\{ \left[\frac{f_1(X)}{f_0(X)} \right]^t \right\} = \int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda$$

$$M_1(t) \equiv E_1(e^{tz}) = E_1 \left\{ \left[\frac{f_1(X)}{f_0(X)} \right]^t \right\} = M_0(1+t)$$

are both convex (Fig. 2.3) and

$$M_0(0) = M_0(1) = 1$$

$$M_0'(0) = E_0 \left(\log \frac{f_1(X)}{f_0(X)} \right) = -I(0:1)$$

$$M_0'(1) = E_1 \left(\log \frac{f_1(X)}{f_0(X)} \right) = I(1:0).$$

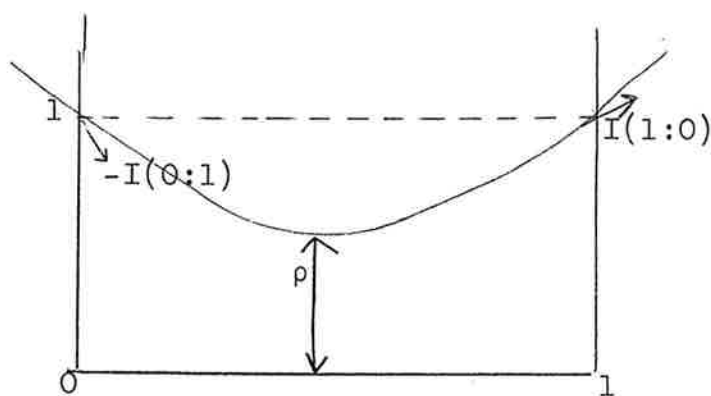


Figure 2.3 Graph of $M_0(t)$

Following the previous theorem we introduce the functions ($i=0,1$)

$$m_i(a) \equiv \inf_t e^{-ta} M_i(t),$$

$$H_i(a) \equiv -\log m_i(a).$$

Then we have

Theorem 2.6 The Chernoff information number defined by (2.8') is equal to

$$\begin{aligned} -\log \rho &= -\log \left(\inf_{0 < t < 1} M_0(t) \right) \\ &= H_0(0) = \max_{-I(0:1) \leq a \leq I(1:0)} \min(H_0(a), H_1(a)). \end{aligned}$$

Proof. Note that

$$\rho \equiv \inf_t M_0(t) = \inf_{0 < t < 1} M_0(t),$$

$$H_0(0) = -\log m_0(0) = -\log \rho.$$

If we define $t_i(a)$ ($i=0,1$) as in the proof of the previous theorem, then we have (Fig. 2.4)

$$t_0(a) = t_1(a) + 1, \quad m_1(a) = e^a m_0(a), \quad H_1(a) = H_0(a) - a.$$

Since $t_0(a)$ is increasing, when a varies in $E_0 z = -I(0:1) \leq a \leq I(1:0) = E_1 z$, $t_0(a)$ varies in

$$0 = t_0(E_0 z) \leq t_0(a) \leq t_0(E_1 z) = t_1(E_1 z) + 1 = 1.$$

It follows thus

$$H_0(0) = \max_{-I(0:1) \leq a \leq I(1:0)} \min(H_0(a), H_1(a)).$$

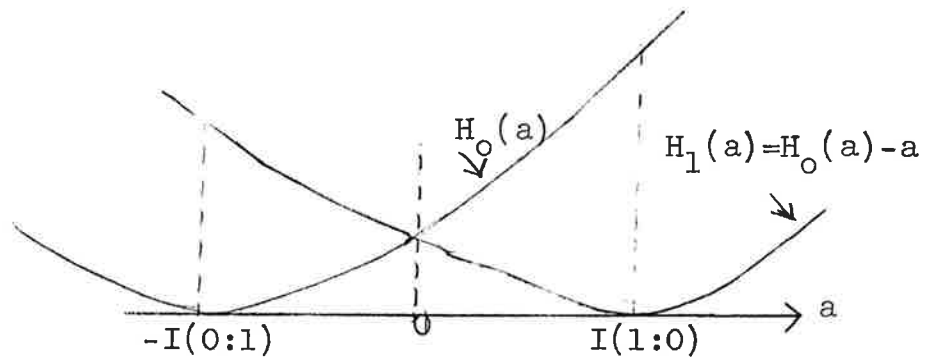


Figure 2.4

The following two corollaries of theorem 2.7 express some interesting connections between the two kinds of information numbers and the function $H(a)$. First we shall state

Theorem 2.7 Let

$$G_\theta \equiv \{g(x) \mid g(x) \geq 0[\lambda], \int g(x) d\lambda = 1, \int T(x)g(x) d\lambda = \theta\},$$

where $T(x)$ is any given statistic, and let $f(x)$ be a given p.d.f. Then
we have

$$\min_{g(x) \in G_\theta} I(g:f) = -\log m(\theta) = \theta t(\theta) - \log M(t(\theta))$$

where

$$M(t) \equiv \int e^{tT(x)} f(x) d\lambda, \quad m(\theta) \equiv \inf_t \{e^{-t\theta} M(t)\}$$

and $t(\theta)$ is determined by

$$\theta = \frac{M'(t(\theta))}{M(t(\theta))} .$$

The minimizing p.d.f. is given by

$$g(x) = g_*(x) \equiv e^{t(\theta)T(x)} f(x) / M(t(\theta))$$

and moreover we have

$$J(g_*, f) \equiv I(g_* : f) + I(f : g_*) = \left(\theta - \int T(x) f(x) d\lambda \right) t(\theta) .$$

(Kullback, 1954)

Proof. An application of the calculus of variations to

$$\int g(x) \left(\log \frac{g(x)}{f(x)} - \mu - tT(x) \right) d\lambda \rightarrow \min_g$$

with Lagrange multipliers μ and t yields the minimizing $g = f e^{tT} / \int f e^{tT} d\lambda$,

where t is connected with θ by $\theta = \int T e^{tT} f d\lambda / \int e^{tT} f d\lambda$. For a more rigorous proof we use the convex property of the function $y \log y$:

$$\begin{aligned} \int g(x) \left(\log \frac{g(x)}{f(x)} - \mu - tT(x) \right) d\lambda &= \int f(x) h(x) \log \frac{h(x)}{e^{\mu+tT(x)}} d\lambda \quad \left(h(x) \equiv \frac{g(x)}{f(x)} \right) \\ &\geq \left(\int f(x) h(x) d\lambda \right) \log \frac{\int f(x) h(x) d\lambda}{\int f(x) e^{\mu+tT(x)} d\lambda} = -\mu - \log M(t) \end{aligned}$$

with equality if and only if $h(x) = e^{\mu+tT(x)}$.

The Cramer-Rao inequality which is fundamental in the theory of statistical estimation will be derived from this theorem as follows.

Let $f_0(x) = f(x, \theta)$ and $f_1(x) = f(x, \theta + \Delta\theta)$ be two densities in a parametric family $\{f(x, \theta) | \theta \in \Omega\}$ of p.d.f.'s and let $\mu_i T^{-1}$ ($i=0,1$) be induced measures by the transform $T(x)$ which is an unbiased estimator of θ , and let

$\{\mu_0 T^{-1}, \mu_1 T^{-1}\} \ll \lambda T^{-1}$, $d\mu_i T^{-1} = g_i d\lambda T^{-1}$ ($i=0,1$). Then

$$\begin{aligned} E_1 \left\{ \log \frac{g_1(Y)}{g_0(Y)} \right\} &\leq E_1 \left\{ \log \frac{f_1(X)}{f_0(X)} \right\} && \text{(by (iii) of theorem 2.1)} \\ &= \frac{(\Delta\theta)^2}{2} \left(E_\theta \left\{ - \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) \right\} + o(1) \right) . && (\Delta \rightarrow 0) \end{aligned}$$

If $H(\theta)$ is the function calculated from the p.d.f. $g_0(y)$ we have from the above theorem

$$\begin{aligned} E_1 \left\{ \log \frac{g_1(Y)}{g_0(Y)} \right\} &\geq H(\theta + \Delta\theta) = H(\theta) + \Delta\theta H'(\theta) + \frac{(\Delta\theta)^2}{2} H''(\theta_1) \\ & && (\theta \leq \theta_1 \leq \theta + \Delta\theta) \end{aligned}$$

and by letting $\Delta\theta \rightarrow 0$

$$\frac{1}{\text{Var}_\theta(Y) \cdot E_\theta \left\{ - \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) \right\}} \leq 1$$

since $H(\theta) = H'(\theta) = 0$ and $H''(\theta) = 1/\text{Var}_\theta Y$ by theorem 2.5 (iii).

Another interesting example of an application of theorem 2.7 is as follows: Let W be a given set and let $T(x) = \chi_W(x) =$ indicator of W .

Then we get, if we set $\mu(W) = \int_W f d\lambda$,

$$\min_{g \in G_\theta} I(g:f) = \theta \log \left(\frac{\theta}{\mu(W)} \right) + (1-\theta) \log \left(\frac{(1-\theta)}{(1-\mu(W))} \right),$$

and

$$g_*(x) = \begin{cases} \theta f(x) / \mu(W) , & x \in W \\ (1-\theta) f(x) / (1-\mu(W)) , & x \in W^c . \end{cases}$$

Corollary 1 to theorem 2.7 Let

$$G_a \equiv \{g(x) \mid g(x) \geq 0, \int g(x) d\lambda = 1, \int xg(x) d\lambda = a\}$$

and let $f(x)$ be a given p.d.f. Then we have

$$\min_{g(x) \in G_a} I(g:f) = H(a) \\ = -\log m(a) = at(a) - \log M(t(a)),$$

where

$$M(t) \equiv \int e^{tx} f(x) d\lambda, \quad m(a) \equiv \inf_t \{e^{-ta} M(t)\},$$

and $t(a)$ is determined by

$$(2.10) \quad a = M'(t(a))/M(t(a)).$$

The minimizing p.d.f. is given by

$$g(x) = g_*(x) \equiv e^{t(a)x} f(x) / M(t(a)),$$

and moreover we have

$$J(g_*, f) \equiv I(g_*:f) + I(f:g_*) = \left(a - \int x f(x) d\lambda \right) t(a).$$

Proof. Let $T(x) = x$ in theorem 2.7.

Corollary 2 to theorem 2.7. Let $f_i(x)$ ($i=0,1$) be given p.d.f.'s and let

$$G \equiv \{g(x) \mid g(x) \geq 0[\lambda], \int g(x) d\lambda = 1, I(g:f_0) = I(g:f_1)\}.$$

Then we have

$$\min_{g(x) \in G} I(g:f_0) = -\log \rho,$$

the Chernoff information number for deciding between two densities f_0

and f_1 , i.e.,

$$= -\log M_0(t^*)$$

where

$$M_0(t) \equiv \int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda$$

and t^* is determined by the equation

$$M'_0(t^*) = \int [f_1(x)]^{t^*} [f_0(x)]^{1-t^*} \log \frac{f_1(x)}{f_0(x)} d\lambda = 0.$$

The minimizing p.d.f. is given by

$$g(x) = g_*(x) = [f_1(x)]^{t^*} [f_0(x)]^{1-t^*} / M_0(t^*).$$

and moreover we have

$$J(g_*, f_0) \equiv I(g_*:f_0) + I(f_0:g_*) = t^* I(f_0:f_1),$$

$$J(g_*, f_1) \equiv I(g_*:f_1) + I(f_1:g_*) = (1-t^*) I(f_1:f_0).$$

Proof. Since $I(g:f_0) = I(g:f_1)$ means $\int g \log (f_1/f_0) d\lambda = 0$, we can take $T(x) = \log (f_1(x)/f_0(x))$ and $\theta = 0$ in theorem 2.7.

Some theorems connected with the two corollaries stated above will be presented later in section 1.6.

The next theorem is based on the criterion (C) mentioned earlier.

Theorem 2.8 Let T_n^0 be a test of a null hypothesis H_0 with the critical region

$$(2.11) \quad Z_n \equiv \frac{1}{n} \sum_{i=1}^n \log \frac{f_1(x_i)}{f_0(x_i)} > \frac{\sigma_0 I_1 - \sigma_1 I_0}{\sigma_0 + \sigma_1}$$

where

$$I_i \equiv I(i:1-i), \quad \sigma_i^2 \equiv \int \left(\log \frac{f_1}{f_0} \right)^2 f_i d\lambda - I_i^2 \quad (i=0,1).$$

Let $r(i, T_n)$ be the error probability of a test T_n when the hypothesis H_i is true. Then the test T_n^0 is asymptotically minimax in the sense that

$$\frac{\inf_{T_n} \max_{i=0,1} r(i, T_n)}{\max_{i=0,1} r(i, T_n^0)} \xrightarrow{(n \rightarrow \infty)} 1$$

and its asymptotically minimax risk is given by

$$\Phi((I_0 + I_1)\sqrt{n}/(\sigma_0 + \sigma_1))$$

where $\Phi(x) \equiv \int_x^{\infty} \varphi(t) dt$ and $\varphi(t) \equiv (2\pi)^{-1/2} e^{-t^2/2}$ (Sakaguchi, 1955)

Proof. Since Z_n is the sum of independent random variables, it follows from the central limit theorem that asymptotically we have

$$Z_n \sim \begin{cases} N(-I_0, \sigma_0^2/n), & \text{under } H_0 \\ N(I_1, \sigma_1^2/n), & \text{under } H_1. \end{cases}$$

Hence when n is large we have

$$r(0, T_n^0) = \Pr(Z_n > a | H_0) = \Phi((a + I_0)\sqrt{n}/\sigma_0) (1 + o(1))$$

$$r(1, T_n^0) = \Pr(Z_n \leq a | H_0) = \{1 - \Phi((a - I_1)\sqrt{n}/\sigma_1)\} (1 + o(1)),$$

where we have set $a \equiv (\sigma_0 I_1 - \sigma_1 I_0) / (\sigma_0 + \sigma_1)$. Since we have $\frac{a + I_0}{\sigma_0} + \frac{a - I_1}{\sigma_1} = 0$, we get

$$(*) \quad \lim_{n \rightarrow \infty} \frac{\max_{i=0,1} r(i, T_n^0)}{\Phi((I_0 + I_1)\sqrt{n}/(\sigma_0 + \sigma_1))} = 1.$$

It is well known that the probability ratio test of the type (2.11) is a Bayes decision function in the case of simple loss functions (in fact, the test with the critical region $Z_n > A$ is a Bayes decision function with respect to the a priori probabilities $e^{nA}/(1 + e^{nA})$ for H_0 and $(1 + e^{nA})^{-1}$ for H_1). Since a Bayes solution with equal risks is a minimax solution and the minimax risk is equal to the Bayes risk, we have

$$\lim_{n \rightarrow \infty} \frac{\inf_{T_n} \max_{i=0,1} r(i, T_n)}{\Phi((I_0 + I_1)\sqrt{n}/(\sigma_0 + \sigma_1))} = 1$$

which, with (*), completes the proof.

For a simple illustration, let

$$f_i(x) = e^{-\mu_i} \mu_i^x / x! \quad (i=0,1 : x=0,1,2,\dots).$$

Then we have

$$\sigma_i^2 = \mu_i \left(\log(\mu_i / \mu_0) \right)^2 \quad (i=0,1),$$

$$\frac{\sigma_0 I_1 - \sigma_1 I_0}{\sigma_0 + \sigma_1} = \sqrt{\mu_0 \mu_1} \log \frac{\mu_1}{\mu_0} - (\mu_1 - \mu_0) \quad (\text{if } \mu_1 > \mu_0),$$

$$\frac{I_0 + I_1}{\sigma_0 + \sigma_1} = \frac{\mu_1 - \mu_0}{\sqrt{\mu_0} + \sqrt{\mu_1}} \quad (\text{if } \mu_1 > \mu_0).$$

Before concluding this section we shall state an important theorem which shows the usefulness of some "information-statistics".

Consider a parametric family of p.d.f.'s $\{f_\theta(x)\}_\theta$. Let

$$H_j : \theta = \theta_j \quad (j=1,\dots,m)$$

be m simple hypotheses, and let the prior probabilities of these hypotheses be denoted by $\alpha_1, \dots, \alpha_m$ respectively, where $\alpha_j > 0$ ($j=1, \dots, m$)

and $\sum_{j=1}^m \alpha_j = 1$.

We now define the Kullback-Leibler information-statistic for a random sample of n independent observations as

$$(2.12) \quad \hat{I}(* : H_i | O_n) \equiv [nI(f_\theta : f_{\theta_i})]_{\theta = \hat{\theta}}$$

where $\hat{\theta}$ is a quasi-maximum likelihood estimator of θ . This is a "directed divergence" from the sample to the simple hypothesis H_i .

Theorem 2.9 If $\{f_{\theta}(x)\}_{\theta}$ is an exponential family of distributions:

$$f_{\theta}(x) = e^{\theta^T(x)} h(x) / M(\theta)$$

where $M(\theta) = \int e^{\theta^T(x)} h(x) d\lambda$, then we have

$$\hat{I}(* : H_i | O_n) - \hat{I}(* : H_j | O_n) = \log \frac{P(O_n | H_j)}{P(O_n | H_i)}. \quad (\text{Kupperman, 1958})$$

Proof. Since for the exponential family of distributions stated in the theorem we have

$$I(\theta : \theta_i) = (\theta - \theta_i) \frac{\partial}{\partial \theta} \log M(\theta) - \log \frac{M(\theta)}{M(\theta_i)},$$

$$\left. \frac{\partial}{\partial \theta} \log M(\theta) \right|_{\theta = \hat{\theta}} = \frac{1}{n} \sum_{k=1}^n T(x_k),$$

and

$$\log \frac{P(O_n | H_j)}{P(O_n | H_i)} = (\theta_j - \theta_i) \sum_{k=1}^n T(x_k) + n \log \frac{M(\theta_i)}{M(\theta_j)},$$

we obtain

$$\begin{aligned} \hat{I}(* : \theta_i | O_n) &= n(\hat{\theta} - \theta_i) \left[\frac{\partial}{\partial \theta} \log M(\theta) \right]_{\theta = \hat{\theta}} - n \log \frac{M(\hat{\theta})}{M(\theta_i)} \\ &= (\hat{\theta} - \theta_i) \sum_{k=1}^n T(x_k) - n \log \frac{M(\hat{\theta})}{M(\theta_i)} \end{aligned}$$

Hence we have

$$\begin{aligned} \hat{I}(* : \theta_i | O_n) - \hat{I}(* : \theta_j | O_n) &= (\theta_j - \theta_i) \sum_{k=1}^n T(x_k) + n \log \frac{M(\theta_i)}{M(\theta_j)} \\ &= \log \frac{P(O_n | H_j)}{P(O_n | H_i)}. \end{aligned}$$

Since we have by Bayes' theorem

$$\log \frac{P(H_j | O_n)}{P(H_i | O_n)} = \log \frac{P(O_n | H_j)}{P(O_n | H_i)}$$

the above theorem can now be restated as follows:

Corollary. For the exponential family of distributions we have

$$\hat{I}(* : H_i | O_n) \leq \hat{I}(* : H_j | O_n),$$

if and only if

$$\log \frac{P(H_i | O_n)}{P(H_i)} \geq \log \frac{P(H_j | O_n)}{P(H_j)} .$$

We shall finally give some short tables of $I(f_0 : f_1)$, $J(f_0, f_1)$, $-\log p$ and $H(a)$ in the following tables:

Table 2.1

$f_i(x)$ ($i=0,1$)	$I(0:1)$	$J(0,1) \equiv I(0:1) + I(1:0)$
Binomial: $\binom{N}{x} p_1^x q_1^{N-x}$	$N \left(p_0 \log \frac{p_0}{p_1} + q_0 \log \frac{q_0}{q_1} \right)$	$N \left\{ (p_0 - p_1) \log \frac{p_0}{p_1} + (q_0 - q_1) \log \frac{q_0}{q_1} \right\}$
Poisson: $e^{-\mu_i} \mu_i^x / (x!)$	$\mu_0 \log \frac{\mu_0}{\mu_1} - (\mu_0 - \mu_1)$	$(\mu_0 - \mu_1) \log \frac{\mu_0}{\mu_1}$
Normal: $(2\pi\sigma_i^2)^{-1} \exp \frac{-(x-\mu_i)^2}{2\sigma_i^2}$	$\frac{1}{2} \left\{ \frac{\sigma_0^2}{\sigma_1^2} - 1 + \left(\frac{\mu_0 - \mu_1}{\sigma_1} \right)^2 \right\} - \log \frac{\sigma_0}{\sigma_1}$	$\frac{1}{2} \left\{ \left(\frac{\sigma_0}{\sigma_1} - \frac{\sigma_1}{\sigma_0} \right)^2 + \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2} \right) (\mu_0 - \mu_1)^2 \right\}$
Exponential: $\beta_i^{-1} e^{-x/\beta_i}$	$-\log \frac{\beta_0}{\beta_1} - \left(1 - \frac{\beta_0}{\beta_1} \right)$	$\left(\sqrt{\frac{\beta_0}{\beta_1}} - \sqrt{\frac{\beta_1}{\beta_0}} \right)^2$
Gamma: $x^{\alpha_i-1} e^{-x} / \Gamma(\alpha_i)$	$\log \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_0)} + (\alpha_0 - \alpha_1) \frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)}$	$(\alpha_0 - \alpha_1) \left(\frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)} - \frac{\Gamma'(\alpha_1)}{\Gamma(\alpha_1)} \right)$
Beta: $\{x(1-x)\}^{s_i-1} / B(s_i, s_i)$	$(s_0 - s_1) G(s_0)$ where $G(s) = \frac{d}{ds} \log B(s, s)$	$(s_0 - s_1) (G(s_0) - G(s_1))$

Table 2.2

$f_i(x)$ ($i=0,1$)	$g_*(x) = [f_1(x)]^{t^*} [f_0(x)]^{1-t^*} / M_0(t^*)$	$-\log \rho$
Binomial:	$(p^*)^x (q^*)^{1-x}$	$p^* \log \frac{p^*}{p_1} +$
$p_1^x q_1^{1-x}$	where	$q^* \log \frac{q^*}{q_1}$
	$p^* = \left(\log \frac{q_1}{q_0} \right) / \left(\log \frac{q_1}{q_0} + \log \frac{p_0}{p_1} \right)$	
Poisson:	$e^{-\mu^*} (\mu^*)^x / (x!)$	$\mu^* \log \frac{\mu^*}{\mu_1} -$
$e^{-\mu_1} \mu_1^x / (x!)$	where	$(\mu^* - \mu_1)$
	$\mu^* = \frac{\mu_1 - \mu_0}{\log(\mu_1 / \mu_0)}$	
Normal:	$(2\pi)^{-1/2} e^{-(x-\mu^*)^2/2}$	$(\mu_1 - \mu_0)^2 / 8$
$(2\pi)^{-1/2} e^{-(x-\mu_1)^2/2}$	where	
	$\mu^* = (\mu_1 + \mu_0) / 2$	
Exponential:	$(\beta^*)^{-1} e^{-x/\beta^*}$	$-\log \frac{\beta^*}{\beta_1} -$
$\beta_1^{-1} e^{-x/\beta_1}$	where	$\left(1 - \frac{\beta^*}{\beta_1} \right)$
	$\beta^* = \frac{1}{\beta_1^{-1} - \beta_0^{-1}} \log \left(\frac{\beta_1^{-1}}{\beta_0^{-1}} \right)$	

Table 2.3

$f(x)$	$g_*(x) = e^{xt(a)} f(x) / M(t(a))$	$H(a) = -\log m(a) = I(g_*; f)$
Binomial: $\binom{N}{x} p^x q^{N-x}$	$\binom{N}{x} \left(\frac{a}{N}\right)^x \left(\frac{1-a}{N}\right)^{N-x}$	$a \log \frac{a}{Np} + (N-a) \log \frac{N-a}{Nq}$
Poisson: $e^{-\mu} \mu^x / (x!)$	$e^{-a} a^x / (x!)$	$a \log \frac{a}{\mu} - (a-\mu)$
Normal: $(2\pi)^{-1/2} e^{-(x-\mu)^2/2}$	$(2\pi)^{-1/2} e^{-(x-a)^2/2}$	$(a-\mu)^2/2$
Gamma: $x^{\alpha-1} e^{-x/\beta} / \Gamma(\alpha) \beta^\alpha$	$\frac{x^{\alpha-1} e^{-\alpha x/a}}{\Gamma(\alpha) (a/\alpha)^\alpha}$	$a \log \frac{\alpha}{a/\beta} - \left(\alpha - \frac{a}{\beta}\right)$

PROBLEMS

(1) Show, by using Theorem 2.1, that

$$\sum_{i=1}^n x_i \log \frac{x_i}{y_i} \geq (x_1 + \dots + x_n) \log \frac{x_1 + \dots + x_n}{y_1 + \dots + y_n},$$

for all $x_i, y_i > 0$.

(2) Using the Schwarz inequality and the inequalities

$$\frac{x-y}{x} \leq \log \frac{x}{y} \leq \frac{x-y}{y} \quad \text{if } \frac{x}{y} > 0,$$

show that

$$\frac{1}{4} \left(\int |f_0(x) - f_1(x)| d\lambda \right)^2 \leq 2 \left(1 - \int \sqrt{f_0(x)f_1(x)} d\lambda \right) \leq I(0:1) \leq \int \left| \frac{f_0(x)}{f_1(x)} - 1 \right| f_0(x) d\lambda$$

(3) Let $f(x)$ be a probability density function with $\int f(x) \log f(x) d\lambda < \infty$. Define $H[f] = - \int f(x) \log f(x) d\lambda$. Prove the following statements:

(a) Let $\mathcal{X} = (-\infty, \infty)$, and λ be the Lebesgue measure. If $f(x)$ has mean μ and variance σ^2 , then $H[f]$ is maximized by the normal density

$$f^*(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-\mu)^2/2\sigma^2} \text{ and } H[f^*] = \frac{1}{2} + \log(\sqrt{2\pi} \sigma).$$

(b) Let $\mathcal{X} = (0, \infty)$, and λ be the Lebesgue measure. If $f(x)$ has mean $a(>0)$, then $H[f]$ is maximized by the exponential density $f^*(x) = a^{-1} e^{-x/a}$ and $H[f^*] = 1 + \log a$.

(c) Let \mathcal{X} be a bounded subset with Lebesgue measure V in the finite dimensional Euclidean space, and λ be Lebesgue measure. Then $H[f]$ is maximized by the uniform density $f^*(x) = 1/V$ and $H[f^*] = \log V$.

(d) Let $\mathcal{X} = \{0, 1, 2, \dots\}$ and λ be counting measure. If $f(x)$ has mean $a(>0)$, then $H[f]$ is maximized by the geometrical density $f^*(x) = \frac{1}{1+a} \left(\frac{a}{1+a} \right)^x$ ($x=0, 1, 2, \dots$) and $H[f^*] = \frac{1}{p} \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right)$, where $p = \frac{1}{1+a}$.

(e) Let $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ and λ be counting measure. If $f(x)$ has mean μ with $\min_{1 \leq i \leq n} x_i \leq \mu \leq \max_{1 \leq i \leq n} x_i$, then $H[f]$ is maximized by the

exponential density $f^*(x) = e^{\beta x} / \sum_{j=1}^n e^{\beta x_j}$ ($x=x_1, \dots, x_n$), where β is

defined by $\frac{\sum_{j=1}^n x_j e^{\beta x_j}}{\sum_{j=1}^n e^{\beta x_j}} = \mu$. And $H[f^*] = \log \left(\sum_{j=1}^n e^{\beta x_j} \right) - \beta \mu$.

(4) Let

$$I(p_0:p_1) \equiv p_0 \log \frac{p_0}{p_1} + (1-p_0) \log \frac{1-p_0}{1-p_1}$$

where $0 \leq p_0, p_1 \leq 1$. Show that

$$I(p_0:p_1) \begin{cases} \geq \\ \leq \end{cases} I(p_1:p_0), \text{ if and only if } p_0(1-p_0) \begin{cases} \geq \\ \leq \end{cases} p_1(1-p_1).$$

(Private communication from Kullback who owes this to I. J. Good).

[Hint: Compute $(I(p_0:p_1) - I(p_1:p_0))/(p_0 - p_1)$ and use the convex decreasing property of the function $x \log \frac{x+1}{x-1}$ for $x > 1$.]

(5) Let $\{E_j\}$ be any measurable partition of \mathcal{X} into pairwise disjoint sets. Prove that

$$I(0:1) = \sup_{\{E_j\}} \sum_{j=1}^{\infty} \mu_0(E_j) \log \frac{\mu_0(E_j)}{\mu_1(E_j)}.$$

(6) Let $\{E_j\}$ be any measurable partition of \mathcal{X} into pairwise disjoint sets. Prove that for any $0 < t < 1$

$$\int [f_1(x)]^t [f_0(x)]^{1-t} d\lambda \leq \sum_j [\mu_1(E_j)]^t [\mu_0(E_j)]^{1-t}.$$

(7) Let $f(x)$ be a given p.d.f. and let $g_*(x)$ be an exponential density generated by $f(x)$:

$$g_*(x) \equiv e^{tT(x)} f(x) / M(t)$$

where $T(x)$ is any given statistic and

$$M(t) \equiv \int e^{tT(x)} f(x) d\lambda.$$

(a) Prove that for any $E \in \mathcal{B}$

$$\left\{ \begin{array}{l} t \min_{x \in E} T(x) \\ t \max_{x \in E} T(x) \end{array} \right\} \leq \log \frac{\int_E g_*(x) d\lambda}{\int_E f(x) d\lambda} + \log M(t) \leq \left\{ \begin{array}{l} t \max_{x \in E} T(x) \\ t \min_{x \in E} T(x) \end{array} \right\}.$$

(b) Let $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $T(x) = x$. Use the result of (a) to show

that

$$\log \frac{\Phi(a-t)}{\Phi(a)} \geq at - \frac{t^2}{2} \quad \text{for all } a, t > 0,$$

where

$$\Phi(a) \equiv \int_a^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

(8) Consider the measurable transformations $T^{(N)}(x)$ of the probability spaces $(\mathcal{X}, \mathcal{B}, \mu_i)$, $i=0,1$ onto the probability spaces $(\mathcal{Y}, \mathcal{C}, \nu_i^{(N)})$, where $\nu_i^{(N)}(G) = \mu_i(\{x | T^{(N)}(x) \in G\})$ for $G \in \mathcal{C}$; that is, $\{T^{(N)}(x)\}_{N=1}^{\infty}$ is a

sequence of statistics and N may be the sample size. Show that if ν_i ($i=0,1$) is a probability measure on the measurable space $(\mathcal{Y}, \mathcal{C})$ such that

$$\lim_{N \rightarrow \infty} \nu_i^{(N)}(G) = \nu_i(G) \quad \text{for all } G \in \mathcal{C} \quad (i=0,1)$$

then

$$I(\nu_0 : \nu_1) \leq \lim_{N \rightarrow \infty} I(\nu_0^{(N)} : \nu_1^{(N)}) \leq \overline{\lim}_{N \rightarrow \infty} I(\nu_0^{(N)} : \nu_1^{(N)}) \leq I(\mu_0 : \mu_1).$$

3. Kullback-Leibler information and testing composite hypotheses.

We have shown in the previous section that Kullback-Leibler and Chernoff information numbers play a remarkable role in the asymptotic theory of statistical tests of two simple hypotheses. We have shown that the amount of information contained in the dichotomous experiment measures in some sense how difficult it is to discriminate between two probability densities with the best test. We shall show in this section some applications of these information numbers to the statistical theory of composite hypotheses testing.

Let $\theta \in \Omega$ be a 1-1 index on a class of distributions on a probability space with elements x and let ω_0, ω_1 be disjoint subsets of Ω . Let $\{X_n\}$ be a sequence of independent random variables with a common distribution indexed by $\theta \in \Omega$. A test $\varphi = \{\varphi_k\}$ with φ_k depending only on X_1, \dots, X_k will be described by the probabilities $\varphi_k(X_1, \dots, X_k)$ assigned to the rejection of the hypothesis $H_0: \theta \in \omega_0$.

Let the risk function when adopting the test φ be given by

$$(3.1) \quad r(\theta, \varphi) = w_0(\theta) E_{\theta}(1 - \varphi(X)) + w_1(\theta) E_{\theta}(\varphi(X))$$

where $w_i(\theta)$ ($i=0,1$) represents the loss of accepting the hypothesis $H_i: \theta \in \omega_i$, when θ is the true parameter value. It is assumed that both $w_i(\theta)$'s are non-negative, bounded and $w_i(\theta) = 0$ for $\theta \in \omega_i$.

We have the following result.

Theorem 3.1 Let $\theta_0 \in \omega_0$ and $\theta_1 \in \omega_1$ be fixed and define

$$\rho_{\theta} = \inf_t E_{\theta} \left\{ \left[\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \right]^t \right\}$$
$$\rho = \inf_{0 < t < 1} E_{\theta_0} \left\{ \left[\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \right]^t \right\}$$

$$I(\theta:\theta_i) = \int f_\theta(x) \log(f_\theta(x)/f_{\theta_i}(x)) d\lambda \quad (i=0,1).$$

If there exist $\theta_0 \in \omega_0$ and $\theta_1 \in \omega_1$ such that

$$(a) \quad I(\theta:\theta_i) < I(\theta:\theta_{1-i}) \quad \text{for all } \theta \in \omega_i \quad (i=0,1)$$

$$(b) \quad \sup_{\theta \in \omega_0} \rho_\theta = \sup_{\theta \in \omega_1} \rho_\theta = \rho,$$

then the test

$$\varphi_n^*(X_1, \dots, X_n) = \begin{cases} 1, & \text{if } \sum_{k=1}^n \log(f_{\theta_1}(X_k)/f_{\theta_0}(X_k)) > 0 \\ 0, & \text{otherwise} \end{cases}$$

is asymptotically efficient in the minimax sense, that is

$$(3.2) \quad \frac{\min \left\{ n \mid \inf_{\varphi_n} \sup_{\theta \in \Omega} r(\theta, \varphi_n) \leq \alpha \right\}}{\min \left\{ n \mid \sup_{\theta \in \Omega} r(\theta, \varphi_n^*) \leq \alpha \right\}} \xrightarrow{(\alpha \rightarrow 0)} 1$$

and we have

$$(3.3) \quad \text{the denominator in the left-hand side of (3.2)} \sim (-\log \alpha) / (-\log \rho).$$

(MacKay, 1959)

(Lemma) Let $\{X_n\}$ be a sequence of independent random variables with common distribution, and assume that $\rho \equiv \inf_t E(e^{tX_1})$ exists. If $EX_1 < 0$

then given any positive $\epsilon < \rho$, we have for sufficiently large n

$$(\rho - \epsilon)^n \leq \Pr \left\{ \sum_{1}^n X_i \geq 0 \right\} \leq \rho^n$$

and

$$(\rho - \epsilon)^n = o \left(\Pr \left\{ \sum_{1}^n X_i \geq 0 \right\} \right) \quad (\text{Chernoff, 1952})$$

Proof.
$$E \left[\exp \left\{ \sum_{1}^n t X_i \right\} \right] \geq \Pr \left\{ \sum_{1}^n X_i \geq 0 \right\} E \left[\exp \left\{ \sum_{1}^n t X_i \right\} \mid \sum_{1}^n X_i \geq 0 \right]$$

$$\geq \Pr \left\{ \sum_{1}^n X_i \geq 0 \right\}, \quad \text{if } t > 0$$

since $EX_1 < 0$ we have

$$\begin{aligned} \rho^n &= \left(\inf_{t>0} E e^{tX_1} \right)^n = \left(\inf_{t>0} E e^{tX_1} \right)^n \\ &= \inf_{t>0} E e^{t \sum_{i=1}^n X_i} \geq \Pr \left\{ \sum_{i=1}^n X_i \geq 0 \right\}. \end{aligned}$$

Thus the right-half of the inequalities in the lemma is proved. To prove the second half, or equivalently, the order relation is not simple. We shall not present the proof here, and readers are suggested to refer to the original paper (Chernoff, 1952).

Proof of theorem 3.1. If ζ^* is a prior probability distribution on Ω concentrating on θ_0 and θ_1 and assigning to each θ_i probability $w_i(\theta_{1-i})/(w_0(\theta_1) + w_1(\theta_0))$ ($i=0,1$), then φ_n^* is Bayes with respect to ζ^* . For we have for every test φ_n

$$\begin{aligned} r(\zeta^*, \varphi_n) &= w \left(E_{\theta_0} [\varphi_n(X)] + E_{\theta_1} [1 - \varphi_n(X)] \right) \\ &= w \left\{ 1 + \int (L_{\theta_0}(x) - L_{\theta_1}(x)) \varphi_n(x) d\lambda^n \right\} \end{aligned}$$

where $w = w_0(\theta_1)w_1(\theta_0)/(w_0(\theta_1) + w_1(\theta_0))$.

Since $E_{\theta_0} \left\{ \log(f_{\theta_1}(X_k)/f_{\theta_0}(X_k)) \right\} < 0$, $E_{\theta_1} \left\{ \log(f_{\theta_1}(X_k)/f_{\theta_0}(X_k)) \right\} > 0$

we have by the Lemma

$$\begin{aligned} P_{\theta_0} \left\{ \sum_{k=1}^n \log \frac{f_{\theta_1}(X_k)}{f_{\theta_0}(X_k)} \geq 0 \right\} &\geq \left(\inf_{t>0} E_{\theta_0} \left\{ \exp \left(t \log \frac{f_{\theta_1}(X_k)}{f_{\theta_0}(X_k)} \right) \right\} - \epsilon \right)^n = (\rho - \epsilon)^n \\ P_{\theta_1} \left\{ \sum_{k=1}^n \log \frac{f_{\theta_1}(X_k)}{f_{\theta_0}(X_k)} \leq 0 \right\} &\geq \left(\inf_{t>0} E_{\theta_1} \left\{ \exp \left(-t \log \frac{f_{\theta_1}(X_k)}{f_{\theta_0}(X_k)} \right) \right\} - \epsilon \right)^n = (\rho - \epsilon)^n, \end{aligned}$$

for sufficiently large n , more precisely for $n \geq n(\epsilon)$, say. Hence

$$\begin{aligned} r(\zeta^*, \varphi_n^*) &= w \left\{ E_{\theta_0}(\varphi_n^*(X)) + E_{\theta_1}(1 - \varphi_n^*(X)) \right\} \\ &= w \left(P_{\theta_0} \left\{ \sum_{k=1}^n \log \frac{f_{\theta_1}(X_k)}{f_{\theta_0}(X_k)} > 0 \right\} + P_{\theta_1} \left\{ \sum_{k=1}^n \log \frac{f_{\theta_1}(X_k)}{f_{\theta_0}(X_k)} \leq 0 \right\} \right) \\ &\equiv 2w(\rho - \epsilon)^n \end{aligned}$$

and for $\alpha < 2w(\rho - \epsilon)^{n(\epsilon)}$ we have

$$(*) \quad N[\zeta^*] = \min \left\{ n \mid \inf_{\varphi_n} r(\zeta^*, \varphi_n) \leq \alpha \right\} \geq \frac{\log \alpha - \log(2w)}{\log(\rho - \epsilon)}.$$

Similarly we have from (a), (b) and the Lemma

$$\begin{aligned} E_{\theta} \left\{ \log(f_{\theta_1}(X_k)/f_{\theta_0}(X_k)) \right\} &= I(\theta : \theta_0) - I(\theta : \theta_1) \begin{cases} < 0, & \theta \in \omega_0 \\ > 0, & \theta \in \omega_1 \end{cases} \\ \sup_{\theta \in \Omega} r(\theta, \varphi_n^*) &= \max \left\{ \sup_{\theta \in \omega_0} w_1(\theta) E_{\theta}(\varphi_n^*(X)), \sup_{\theta \in \omega_1} w_0(\theta) E_{\theta}(1 - \varphi_n^*(X)) \right\} \\ &\leq W \max \left\{ \sup_{\theta \in \omega_0} \rho_{\theta}^n, \sup_{\theta \in \omega_1} \rho_{\theta}^n \right\} = W \rho^n \end{aligned}$$

$$\text{where } W \equiv \max \left\{ \sup_{\theta \in \omega_0} w_1(\theta), \sup_{\theta \in \omega_1} w_0(\theta) \right\}.$$

If we take

$$n = N_{\varphi^*} - 1 \equiv \min \left\{ n \mid \sup_{\theta \in \Omega} r(\theta, \varphi_n^*) \leq \alpha \right\} - 1$$

then

$$(**) \quad N_{\varphi^*} \leq \frac{\log \alpha - \log W}{\log \rho} + 1.$$

Since we have

$$N[\zeta^*] \leq \min \left\{ n \mid \inf_{\varphi_n} \sup_{\theta \in \Omega} r(\theta, \varphi_n) \leq \alpha \right\} \leq N_{\varphi^*}$$

we obtain by (*) and (**) the relations (3.2) and (3.3) stated in the Theorem.

The condition (a) of theorem 3.1 states that the two composite hypotheses ω_0 and ω_1 must be properly "separated". For example these two hypotheses must be separated as in Fig. 3.1(a) and not as in Fig. 3.1 (b) and (c).

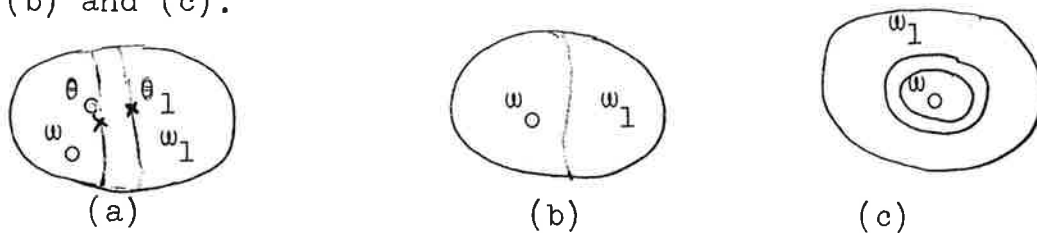


Figure 3.1

And the condition (b) of theorem 3.1 requires θ_0 and θ_1 to be the "representatives", in the sense stated, of ω_0 and ω_1 respectively.

Consider for example the case of the exponential family of distributions. Let

$$f_{\theta}(x) = e^{\theta T(x)} h(x)/M(\theta), \quad M(\theta) = \int e^{\theta T(x)} h(x) d\lambda.$$

Take $a < b$, and let $\omega_0 = \{\theta | \theta \leq a\}$ and $\omega_1 = \{\theta | \theta \geq b\}$ be the two subsets. Then it can be shown that a and b are the "representatives" of ω_0 and ω_1 respectively, that is, we have

Theorem 3.2 $\theta_0 = a$ and $\theta_1 = b$, in this case, satisfy MacKay's conditions (a) and (b) in Theorem 3.1.

Proof. Since we have by the strict convexity of $w(\theta)$

$$I(\theta:\theta') = w(\theta') - w(\theta) - (\theta' - \theta)w'(\theta), \quad (w(\theta) \equiv \log M(\theta))$$

$$I(\theta:b) - I(\theta:a) = w(b) - w(a) - (b-a)w'(\theta)$$

$$\begin{cases} > 0, & \text{if } \theta \leq a \\ < 0, & \text{if } \theta \geq b \end{cases}$$

the validity of the condition (a) is evident. In order to see that condition (b) holds true we have to calculate ρ_{θ} . It thus follows that

$$\begin{aligned}
\rho_\theta &= \inf_t E_\theta \left\{ \left[\frac{f_b(X)}{f_a(X)} \right]^t \right\} = \inf_t E_\theta \left\{ \left[\frac{M(a)}{M(b)} \right]^t e^{t(b-a)T(x)} \right\} \\
&= \inf_t E_\theta \left[e^{t(b-a)} \left\{ T(x) - (\omega(b) - \omega(a)) / (b-a) \right\} \right] \\
&= \inf_t E_\theta \left[e^{t(T(x) - \omega'(\theta_{ab}^*))} \right]
\end{aligned}$$

where θ_{ab}^* is the unique root of the equation

$$\omega'(\theta_{ab}^*) = (\omega(b) - \omega(a)) / (b-a).$$

By theorem 2.5(i) we have

$$\begin{aligned}
-\log \rho_\theta &= -\log \left(\inf_t E_\theta \left[e^{t(T(x) - \omega'(\theta_{ab}^*))} \right] \right) \\
&= (\theta_{ab}^* - \theta) \omega'(\theta_{ab}^*) - \log \frac{M(\theta_{ab}^*)}{M(\theta)} \\
&= \omega(\theta) - \omega(\theta_{ab}^*) - (\theta - \theta_{ab}^*) \omega'(\theta_{ab}^*)
\end{aligned}$$

since

$$E_\theta (e^{tT(x)}) = \int e^{(t+\theta)T(x)} h(x) d\lambda / M(\theta) = M(t+\theta) / M(\theta).$$

It follows that

$$\frac{1}{\rho_\theta} \frac{\partial \rho_\theta}{\partial \theta} = \omega'(\theta_{ab}^*) - \omega'(\theta) \begin{cases} > 0, & \text{if } \theta \leq a \\ < 0, & \text{if } \theta \geq b. \end{cases}$$

It is easily shown that the three values $\sup_{\theta \leq a} \rho_\theta = \rho_a$, $\sup_{\theta \geq b} \rho_\theta = \rho_b$

and ρ are equal and the common value is

$$\begin{aligned}
&\exp \left[- \left\{ \omega(a) - \omega(\theta_{ab}^*) - (a - \theta_{ab}^*) \omega'(\theta_{ab}^*) \right\} \right] \\
&= \left[\frac{1}{M(a)} \right]^{(b - \theta_{ab}^*) / (b-a)} \left[\frac{1}{M(b)} \right]^{(\theta_{ab}^* - a) / (b-a)} \bigg/ \frac{1}{M(\theta_{ab}^*)}.
\end{aligned}$$

Our second example of MacKay's conditions is as follows: Let us consider the test of the difference of two binomial parameters. We have $\Omega = \{(\xi, \eta) \mid 0 \leq \xi, \eta \leq 1\}$, $\omega_0 = \{\xi - \eta \geq \delta\}$ and $\omega_1 = \{\xi - \eta \leq -\delta\}$, where δ is a given positive number. Let $\hat{\xi}$ and $\hat{\eta}$ be the maximum likelihood estimators of ξ and η respectively. Then we can show that the test

$$\varphi_{m,n}^* = \begin{cases} 1, & \text{if } \lambda \left(\hat{\xi} - \frac{1}{2} \right) \leq \hat{\eta} - \frac{1}{2} \\ 0, & \text{otherwise,} \end{cases}$$

is asymptotically efficient in the minimax sense, where

$$\lambda = \frac{m}{n} \frac{\log((1-\xi_1)/\xi_1)}{\log((\xi_1+\delta)/(1-\xi_1-\delta))}, \quad (\xi_1 < \frac{1}{2} < \xi_1 + \delta)$$

m and n are the sizes of the samples from the two populations, and ξ_1 is the unique root of the equation

$$m(\xi^{-1} - (1-\xi)^{-1}) + n((\xi+\delta)^{-1} - (1-\xi-\delta)^{-1}) = 0.$$

The two "representatives" (ξ_0, η_0) and (ξ_1, η_1) are on the boundary line $\xi - \eta = \delta$ and $\xi - \eta = -\delta$ respectively and symmetrically situated about the centre of the unit square (Fig. 3.2).

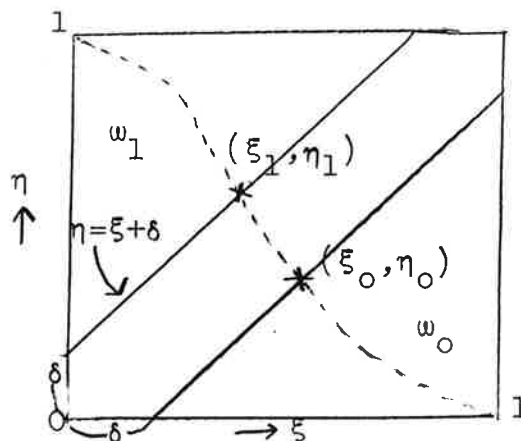


Figure 3.2

The next theorem is a generalization of theorem 2.3 to the composite hypotheses case.

Theorem 3.3 Assume that there exist $\theta_0 \in \omega_0$ and $\theta_1 \in \omega_1$ satisfying MacKay's conditions (a) and (b) in theorem 3.1. Then for any prior probability distribution ζ on Ω the average risk of the Bayes test ω_ζ with respect to ζ satisfies the inequality

$$r(\zeta) \equiv \int_{\Omega} r(\theta, \omega_\zeta) d\zeta \leq W \rho^n$$

for every n and independently of ζ , where $W = \max \left(\sup_{\theta \in \omega_0} w_1(\theta), \sup_{\theta \in \omega_1} w_0(\theta) \right)$

(Sakaguchi, 1961)

Proof. Let $\rho_\theta(t) \equiv E_\theta \left\{ \left[f_{\theta_1} / f_{\theta_0} \right]^t \right\}$ then we have by (a)

$$\rho'_\theta(0) = E_\theta \left\{ \log(f_{\theta_1} / f_{\theta_0}) \right\} = I(\theta : \theta_0) - I(\theta : \theta_1) \begin{cases} < 0, & \theta \in \omega_0 \\ > 0, & \theta \in \omega_1 \end{cases} .$$

Hence we get

$$\rho_\theta = \inf_t \rho_\theta(t) = \begin{cases} \inf_{t>0} \rho_\theta(t), & \theta \in \omega_0 \\ \inf_{t<0} \rho_\theta(t), & \theta \in \omega_1 \end{cases}$$

from which we obtain

$$(3.4) \quad \rho_\theta^n = \begin{cases} \inf_{t>0} \int \left[L_{\theta_1}(x) / L_{\theta_0}(x) \right]^t L_\theta(x) d\lambda^n, & \theta \in \omega_0 \\ \inf_{t<0} \int \left[L_{\theta_1}(x) / L_{\theta_0}(x) \right]^t L_\theta(x) d\lambda^n, & \theta \in \omega_1 \end{cases}$$

where $L_\theta(x)$ represents the likelihood $\prod_{i=1}^n f_\theta(x_i)$. Let ζ^* be a prior

probability distribution on Ω concentrating on θ_0 and θ_1 and assigning

to θ_0 probability p , the value of which is to be chosen later. The Bayes test with respect to ζ^* is the non-randomized test

$$\varphi_{\zeta^*}(x) = \begin{cases} 1, & \text{if } pw_1(\theta_0)L_{\theta_0}(x) < (1-p)w_0(\theta_1)L_{\theta_1}(x), \\ 0, & \text{otherwise.} \end{cases}$$

Let $R \equiv \left\{ x \mid pw_1(\theta_0)L_{\theta_0}(x) < (1-p)w_0(\theta_1)L_{\theta_1}(x) \right\}$ then we have from (3.1)

$$\begin{aligned} r(\theta, \varphi_{\zeta^*}) &= w_0(\theta)E_{\theta}(1-\varphi_{\zeta^*}(X)) + w_1(\theta)E_{\theta}(\varphi_{\zeta^*}(X)) \\ &= w_0(\theta)\int_{R^c} L_{\theta}(x)d\lambda^n + w_1(\theta)\int_R L_{\theta}(x)d\lambda^n \end{aligned}$$

We thus get

$$(3.5) \quad \sup_{\theta \in \Omega} r(\theta, \varphi_{\zeta^*}) \leq W \max \left(\sup_{\theta \in \omega_1} \int_{R^c} L_{\theta}(x)d\lambda^n, \sup_{\theta \in \omega_0} \int_R L_{\theta}(x)d\lambda^n \right)$$

where $W = \max \left(\sup_{\theta \in \omega_0} w_1(\theta), \sup_{\theta \in \omega_1} w_0(\theta) \right)$. If we set

$$h(t, \theta) \equiv \int (L_{\theta_1}(x)/L_{\theta_0}(x))^t L_{\theta}(x)d\lambda^n = \left(\int_R + \int_{R^c} \right) (L_{\theta_1}/L_{\theta_0})^t L_{\theta} d\lambda^n$$

in (3.4) it follows that

$$h(t, \theta) \geq \begin{cases} \int_R \left[\frac{L_{\theta_1}}{L_{\theta_0}} \right]^t L_{\theta} d\lambda^n \geq \left[\frac{pw_1(\theta_0)}{(1-p)w_0(\theta_1)} \right]^t \int_R L_{\theta} d\lambda^n, & \text{if } t > 0 \\ \int_{R^c} \left[\frac{L_{\theta_0}}{L_{\theta_1}} \right]^{-t} L_{\theta} d\lambda^n \geq \left[\frac{(1-p)w_0(\theta_1)}{pw_1(\theta_0)} \right]^{-t} \int_{R^c} L_{\theta} d\lambda^n, & \text{if } t < 0 \end{cases}$$

and hence that

$$\begin{aligned} \int_R L_{\theta} d\lambda^n &\leq h(t, \theta) \left[\frac{(1-p)w_0(\theta_1)}{pw_1(\theta_0)} \right]^t, & \text{if } t > 0 \\ \int_{R^c} L_{\theta} d\lambda^n &\leq h(t, \theta) \left[\frac{pw_1(\theta_0)}{(1-p)w_0(\theta_1)} \right]^{-t}, & \text{if } t < 0. \end{aligned}$$

Now we choose $p = w_0(\theta_1)/(w_0(\theta_1) + w_1(\theta_0))$ and denote ζ^* and R with this choice of p by ζ^0 and R^0 respectively. Then we have by (3.4)

$$(3.6) \quad \left\{ \begin{array}{l} \int_{R^0} L_\theta d\lambda^n \leq \inf_{t>0} h(t, \theta) = \rho_\theta^n, \quad \text{if } \theta \in \omega_0 \\ \int_{R^{0c}} L_\theta d\lambda^n \leq \inf_{t<0} h(t, \theta) = \rho_\theta^n, \quad \text{if } \theta \in \omega_1 \end{array} \right\} .$$

From (3.5), (3.6) and the condition (b) of theorem 3.1 we get

$$\sup_{\theta \in \Omega} r(\theta, \varphi_{\zeta^0}) \leq W \rho^n.$$

Hence for any prior probability distribution ζ on Ω the average risk of the Bayes test φ_ζ with respect to ζ satisfies the inequality

$$\begin{aligned} \int_{\Omega} r(\theta, \varphi_\zeta) d\zeta &\leq \int_{\Omega} r(\theta, \varphi_{\zeta^0}) d\zeta && \text{(by the Bayes property of } \varphi_\zeta) \\ &\leq \sup_{\theta \in \Omega} r(\theta, \varphi_{\zeta^0}) \leq W \rho^n \end{aligned}$$

and the proof of our theorem 2.3 is completed.

4. Kullback-Leibler information and sequential tests of hypotheses.

We shall begin this section by an asymptotic study of the problem of sequentially testing a simple hypothesis against a simple alternative. Suppose as usual $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ are two simple hypotheses, and the experiment yields a variable x whose density is $f_i(x)$ under H_i , $i = 0,1$. The Bayes strategies are the Wald sequential probability-ratio tests. These are characterized by two numbers A and B with $B < 0 < A$, (Fig. 4.1), and consist of reacting to the first m observations x_1, \dots, x_m by

$$(4.1) \quad \begin{cases} \text{accepting } H_1 \text{ if } Z_m \geq A \\ \text{accepting } H_0 \text{ if } Z_m \leq B \\ \text{continuing sampling as long as } B < Z_m < A, \text{ where} \\ Z_m = \sum_{i=1}^m \log(f_1(x_i)/f_0(x_i)). \end{cases}$$

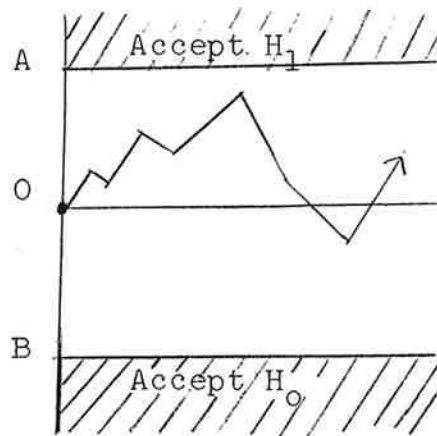


Figure 4.1

The appropriate numbers A and B are determined by the a priori probability ζ of H_0 and the costs. These are the cost c per observation and the loss $w_i (i=0,1)$ due to rejecting H_i when it is true. The

risks corresponding to a sequential strategy are given by

$$(4.2) \quad \begin{cases} R_0 = w_0 \alpha + c E_0(n) \\ R_1 = w_1 \beta + c E_1(n) \end{cases}$$

where α and β are the two probabilities of error, by that strategy, and n is the random sample size. Of course A and B are determined so as to minimize the average risk $\zeta R_0 + (1-\zeta)R_1$.

The following fundamental theorem in sequential test theory is a generalization of the non-sequential result stated in Theorem 2.2(i).

Theorem 4.1 For any closed sequential test S with strength (α, β) , we have

$$(4.3) \quad \begin{cases} E_0(n|S) \geq \left\{ (1-\alpha) \log \frac{1-\alpha}{\beta} + \alpha \log \frac{\alpha}{1-\beta} \right\} / I(0:1) \\ E_1(n|S) \geq \left\{ \beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha} \right\} / I(1:0) \end{cases}$$

or equivalently by setting $L(\theta) = P_\theta \{ S \text{ accepts } H_0 \}$ ($\theta=0,1$),

$$(4.3') \quad E_\theta(n|S) \geq \left\{ L(\theta) \log \frac{L(\theta)}{L(\theta')} + (1-L(\theta)) \log \frac{1-L(\theta)}{1-L(\theta')} \right\} / I(\theta:\theta')$$

($\theta=0,1; \theta \neq \theta'$)

(Wald, 1945)

Proof. Let $Z = \log(f_\theta(x)/f_{\theta'}(x))$. By the Wald-Blackwell theorem we have for any closed sequential test

$$(*) \quad E_\theta(Z_1 + \dots + Z_n) = E_\theta(n) E_\theta(Z).$$

The left-hand side is

$$E_\theta(Z_1 + \dots + Z_n) = E_\theta \left\{ \log \left(\frac{P_{\theta n}}{P_{\theta' n}} \right) \right\} \quad (P_{\theta n} \equiv \prod_{i=1}^n f_\theta(x_i))$$

$$(*) = \sum_{N=1}^{\infty} E_\theta(n=N) \left(\int_{Q_{0N}} + \int_{Q_{1N}} \right) \frac{P_{\theta N}}{P_{\theta' N}} \log \frac{P_{\theta N}}{P_{\theta' N}} d \lambda^N(x)$$

where $Q_{iN} = \left\{ (x_1, \dots, x_n) \mid S \text{ accepts } H_i \text{ with sample size } N \right\}$ ($i=0,1$).

Since

$$L(\theta) = P_{\theta} \left\{ S \text{ accepts } H_0 \right\} = \sum_{N=1}^{\infty} P_{\theta}(n=N) P_{\theta}(Q_{0N})$$

$$1 - L(\theta) = P_{\theta} \left\{ S \text{ accepts } H_1 \right\} = \sum_{N=1}^{\infty} P_{\theta}(n=N) P_{\theta}(Q_{1N})$$

and by using the convex property of K-L informations (cor. to theorem 2.1), we have

$$\begin{aligned} (*) &\geq \sum_{N=1}^{\infty} P_{\theta}(n=N) \left\{ P_{\theta}(Q_{0N}) \log \frac{P_{\theta}(Q_{0N})}{P_{\theta'}(Q_{0N})} + P_{\theta}(Q_{1N}) \log \frac{P_{\theta}(Q_{1N})}{P_{\theta'}(Q_{1N})} \right\} \\ &\geq L(\theta) \log \frac{L(\theta)}{L(\theta')} + (1-L(\theta)) \log \frac{1-L(\theta)}{1-L(\theta')} . \end{aligned}$$

Combining the last inequality and (*) and (*) we have the desired result (4.3').

Wald approximations: If the means and variances of the random variable $Z = \log(f_1(x)/f_0(x))$ are all small enough (in absolute value)

$$(4.4) \quad \left\{ \begin{array}{l} \alpha \approx e^{-A} \\ \beta \approx e^B \\ E_0(n) \approx \frac{(1-\alpha) \log \frac{1-\alpha}{\beta} + \alpha \log \frac{\alpha}{1-\beta}}{I(0:1)} \approx \frac{-B}{I(0:1)} \\ E_1(n) \approx \frac{\beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha}}{I(1:0)} \approx \frac{A}{I(1:0)} \end{array} \right.$$

for the Wald sequential probability-ratio test.

Proof. Set $E_{\theta}^*(Z_1 + \dots + Z_n) \approx -B$, $E_{\theta}^{**}(Z_1 + \dots + Z_n) \approx A$ in (*) and (*)

of the proof of Theorem 4.1.

$$E_{\theta}(n) = \frac{E_{\theta}(Z_1 + \dots + Z_n)}{E_{\theta}(Z)} \approx \frac{L(\theta)(-B) + (1-L(\theta))A}{I(\theta:\theta')}$$

Moreover we have

$$\begin{aligned} -B &\approx -\log\beta \approx \log\frac{1-\alpha}{\beta} \\ A &\approx -\log\alpha \approx \log\frac{1-\beta}{\alpha} \end{aligned}$$

Consider the given family $\{f(x,\theta)\}_{\theta}$ of p.d.f's and let

$$\begin{aligned} H_0: \theta &= \theta_0 \\ H_1: \theta &= \theta_1 = \theta_0 + \Delta \end{aligned}$$

The following theorem shows that when θ_1 is close to θ_0 , the percentage saving of the Wald sequential procedure compared with the best non-sequential test, is independent of the particular function $f_{\theta}(x)$ and the particular values θ_1 and θ_0 , provided $f_{\theta}(x)$ satisfies some weak conditions.

Theorem 4.2 Assume that the family $\{f(x,\theta)\}_{\theta}$ satisfies the regularity assumption (R) mentioned in Section 1.1. Let $E_i(n)$ ($i=0,1$) denote the expected sample size under the hypothesis H_i required by the Wald sequential test with strength (α,β) . Let N be the size of the sample to achieve the same strength (α,β) by the most powerful non-sequential test. Then we have

$$(4.5) \quad \begin{cases} \lim_{\Delta \rightarrow 0} \frac{E_0(n)}{N} = \frac{(1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta}}{\frac{1}{2}(U_{\alpha} + U_{\beta})^2} \\ \lim_{\Delta \rightarrow 0} \frac{E_1(n)}{N} = \frac{\beta\log\frac{\beta}{1-\alpha} + (1-\beta)\log\frac{1-\beta}{\alpha}}{\frac{1}{2}(U_{\alpha} + U_{\beta})^2} \end{cases}$$

where U_α and U_β are defined by the relations

$$\alpha = \int_{U_\alpha}^{\infty} (2\pi)^{-1/2} e^{-t^2/2} dt, \quad \beta = \int_{U_\beta}^{\infty} (2\pi)^{-1/2} e^{-t^2/2} dt$$

(Paulson, 1947)

Proof. For the Wald sequential test with strength (α, β) we have

$$(*) \quad \begin{cases} E_0(n) = \left\{ (1-\alpha) \log \frac{1-\alpha}{\beta} + \alpha \log \frac{\alpha}{1-\beta} \right\} / I(0:1) + o(1) \\ E_1(n) = \left\{ \beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha} \right\} / I(1:0) + o(1) \end{cases}$$

when $\Delta \rightarrow 0$. On the other hand the most powerful non-sequential test has the critical region

$$Z_N \equiv \frac{1}{N} \sum_{i=1}^N \log \frac{f(x_i, \theta_0 + \Delta)}{f(x_i, \theta_0)} \geq k.$$

When $\Delta \rightarrow 0$, $N \rightarrow \infty$, Z_N has the asymptotic distribution

$$Z_N \sim \begin{cases} \mathbb{N}(-I(0:1), \sigma_0^2/N), & \text{under } H_0 \\ \mathbb{N}(I(1:0), \sigma_1^2/N), & \text{under } H_1 \end{cases},$$

where σ_i^2 ($i=0,1$) is the variance of $\log(f(X, \theta_1)/f(X, \theta_0))$ under H_1 .

Hence we find the N required for a test with strength (α, β) by solving for N from the relations

$$\begin{cases} \alpha = \Pr \left\{ Z_N \geq k | H_0 \right\} = \int_{(k+I(0:1))/(\sigma_0/\sqrt{N})}^{\infty} (2\pi)^{-1/2} e^{-t^2/2} dt \\ \beta = \Pr \left\{ Z_N < k | H_1 \right\} = \int_{-\infty}^{(k-I(1:0))/\sigma_1/\sqrt{N}} (2\pi)^{-1/2} e^{-t^2/2} dt \end{cases}$$

so that

$$\begin{cases} U_\alpha = \sqrt{N} (k + I(0:1)) / \sigma_0 \\ U_{1-\beta} = -U_\beta = \sqrt{N} (k - I(1:0)) / \sigma_1 \end{cases}$$

and we get

$$(*) \quad N = \left(\frac{\sigma_0 U_\alpha + \sigma_1 U_\beta}{J(0,1)} \right)^2.$$

Now when $\Delta \rightarrow 0$ we have

$$I(0:1), I(1:0) \sim \frac{\Delta^2}{2} i(\theta_0) + o(\Delta^3) \quad \left(i(\theta) \equiv E_{\theta} \left\{ - \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) \right\} \right)$$

$$\begin{aligned} \sigma_0^2 &= E_0 \left[\left\{ \log \frac{f(X, \theta_1)}{f(X, \theta_0)} \right\}^2 \right] - I(0:1)^2 \\ &= E_0 \left[\left\{ \Delta \frac{\partial}{\partial \theta} \log f(X, \theta) + \frac{\Delta^2}{2} \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) + o(\Delta^3) \right\}^2 \right] - \\ &\quad \left(\frac{\Delta^2}{2} i(\theta_0) + o(\Delta^3) \right)^2 \end{aligned}$$

$$= \Delta^2 i(\theta_0) + o(\Delta^3)$$

$$\sigma_1^2 = E_1 \left[\left\{ \log \frac{f(X, \theta_1)}{f(X, \theta_0)} \right\}^2 \right] - I(1:0)^2 = \Delta^2 i(\theta_0) + o(\Delta^3)$$

from which we have by (*)

$$N \sim \frac{(U_{\alpha} + U_{\beta})^2}{\Delta^2 i(\theta_0)}, \quad (\Delta \rightarrow 0).$$

Combining this and (*) we obtain the desired result.

Since the Wald sequential test is optimum, as is well-known, it is suggested that

$$\frac{\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}}{\frac{1}{2} (U_{\alpha} + U_{\beta})^2} \leq 1, \quad \text{for all } 0 < \alpha, \beta < 1,$$

or equivalently

$$(4.6) \quad \Phi(x) \log \frac{\Phi(x)}{1-\Phi(y)} + (1-\Phi(x)) \log \frac{1-\Phi(x)}{\Phi(y)} \leq \frac{1}{2} (x+y)^2, \quad (-\infty < x, y < \infty),$$

where

$$\Phi(x) = \int_x^{\infty} \varphi(t) dt, \quad \varphi(t) = (2\pi)^{-1/2} e^{-t^2/2}.$$

We shall give here a direct analytic proof of this inequality

(Sakaguchi, 1955). By Jensen's inequality for convex functions, we have

$$f\left(\frac{\int_a^b tg(t)dt}{\int_a^b g(t)dt}\right) \leq \frac{\int_a^b f(t)g(t)dt}{\int_a^b g(t)dt},$$

where $f(t)$ is convex and $g(t)$ is non-negative in (a,b) . For $a = x$, $b = \infty$, $f(t) = e^{-kt}$ ($-\infty < k < \infty$) and $g(t) = \varphi(t) \equiv (2\pi)^{-1/2} e^{-t^2/2}$, this reduces to

$$\exp\left\{-k\varphi(x)/\Phi(x)\right\} \leq e^{k^2/2} \Phi(x+k)/\Phi(x)$$

so that

$$(4.6^*) \quad \log \frac{\Phi(x)}{\Phi(x+k)} \leq \frac{k^2}{2} + k \frac{\varphi(x)}{\Phi(x)}, \quad (-\infty < k, x < \infty),$$

since

$$\int_x^\infty t\varphi(t)dt = \varphi(x), \quad \int_x^\infty e^{-kt}\varphi(t)dt = e^{k^2/2}\Phi(x+k), \quad (-\infty < k, x < \infty).$$

From (4.6*) and the relation $\Phi(x) + \Phi(-x) = 1$, we obtain the desired inequality (4.6) as follows: If we set $k = -x-y$ in (4.6*) and multiply both sides by $\Phi(x)$, then

$$\Phi(x) \log \frac{\Phi(x)}{1-\Phi(y)} \leq \frac{(x+y)^2}{2} \Phi(x) - (x+y)\varphi(x).$$

If we change x to $-x$ in (4.6*), set $k = x+y$ and then multiply both sides by $1-\Phi(x)$

$$(1-\Phi(x)) \log \frac{1-\Phi(x)}{\Phi(y)} \leq \frac{(x+y)^2}{2} (1-\Phi(x)) + (x+y)\varphi(x).$$

Adding these two inequalities we get (4.6).

It is to be noted that we have

$$\lim_{\beta \rightarrow 1-\alpha-0} \frac{\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}}{\frac{1}{2} (U_\alpha + U_\beta)^2} = \frac{\varphi(U_\alpha)^2}{\alpha(1-\alpha)}, \quad \left(\frac{d}{d\alpha} U_\alpha = -1/\varphi(U_\alpha) \right)$$

which is symmetric about $\alpha = \frac{1}{2}$, and equals $\frac{2}{\pi} \doteq 0.637$ at $\alpha = \frac{1}{2}$.

Another fact to be noted is that by expanding $\log \Phi(x+k)$ about $k = 0$ we have from (4.6*)

$$\varphi^2 - x\varphi\Phi \leq \Phi$$

from which we can obtain an inequality for the Mill's ratio Φ/φ

$$\Phi/\varphi \geq \frac{2}{\sqrt{x^2+4} + x} = \frac{1}{2} \left(\sqrt{x^2+4} - x \right).$$

Let us now consider the case of composite hypotheses.

Theorem 4.3 Let S be any closed sequential test for deciding between
 $H_0: \theta \in \omega_0$ and $H_1: \theta \in \omega_1$ such that

$$\begin{aligned} P_\theta \left\{ S \text{ accepts } H_1 \right\} &\leq \alpha, & \text{if } \theta \in \omega_0 \\ P_\theta \left\{ S \text{ accepts } H_0 \right\} &\leq \beta, & \text{if } \theta \in \omega_1 \end{aligned}$$

where $\alpha + \beta \leq 1$. Then we have

$$(4.7) \quad \begin{cases} E_\theta(n|S) \geq \left\{ (1-\alpha) \log \frac{1-\alpha}{\beta} + \alpha \log \frac{\alpha}{1-\beta} \right\} / \inf_{\theta' \in \omega_1} I(\theta: \theta'), & \theta \in \omega_0 \\ E_\theta(n|S) \geq \left\{ \beta \log \frac{\beta}{1-\alpha} + (1-\beta) \log \frac{1-\beta}{\alpha} \right\} / \inf_{\theta' \in \omega_0} I(\theta: \theta'), & \theta \in \omega_1 \end{cases} \quad (\text{Hoeffding, 1953})$$

Proof. For any closed sequential test with the operating characteristic function $L(\theta) \equiv P_\theta \left\{ S \text{ accepts } H_0 \right\}$ we have

$$E_\theta(n) I(\theta: \theta') \geq L(\theta) \log \frac{L(\theta)}{L(\theta')} + (1-L(\theta)) \log \frac{1-L(\theta)}{1-L(\theta')}, \quad \theta, \theta' \in \Omega$$

by (4.3') of Theorem 4.1. It follows that

$$\begin{aligned}
 & E_{\theta}(n) (\zeta I(\theta:\theta_0) + (1-\zeta)I(\theta:\theta_1)) \\
 \geq & \zeta L(\theta) \log \frac{L(\theta)}{L(\theta_0)} + (1-L(\theta)) \log \frac{1-L(\theta)}{1-L(\theta_0)} + (1-\zeta) L(\theta) \log \frac{L(\theta)}{L(\theta_1)} \\
 & \qquad \qquad \qquad + (1-L(\theta)) \log \frac{1-L(\theta)}{1-L(\theta_1)} \\
 = & L(\theta) \log \frac{L(\theta)}{[L(\theta_0)]^{\zeta} [L(\theta_1)]^{1-\zeta}} + (1-L(\theta)) \log \frac{1-L(\theta)}{[1-L(\theta_0)]^{\zeta} [1-L(\theta_1)]^{1-\zeta}} \\
 \geq & -\log \left([L(\theta_0)]^{\zeta} [L(\theta_1)]^{1-\zeta} + [1-L(\theta_0)]^{\zeta} [1-L(\theta_1)]^{1-\zeta} \right).
 \end{aligned}$$

Now the function

$$K(x,y) = (1-x)^{\zeta} y^{1-\zeta} + x^{\zeta} (1-y)^{1-\zeta}$$

is increasing in x and y if $x + y < 1$. Since $1-L(\theta_0) \leq \alpha$ and $L(\theta_1) \leq \beta$ we have $K(1-L(\theta_0), L(\theta_1)) \leq K(\alpha, \beta)$ if $\alpha + \beta \leq 1$. Thus we get

$$(4.8) \quad E_{\theta}(n) \geq \frac{-\log(\alpha^{\zeta}(1-\beta)^{1-\zeta} + (1-\alpha)^{\zeta}\beta^{1-\zeta})}{\zeta I(\theta:\theta_0) + (1-\zeta)I(\theta:\theta_1)}$$

for every

$$0 < \zeta < 1, \quad \theta_0 \in \omega_0, \quad \theta_1 \in \omega_1.$$

The best inequality we can obtain from (4.8) is

$$(4.9) \quad E_{\theta}(n) \geq \sup_{0 < \zeta < 1} \frac{-\log(\alpha^{\zeta}(1-\beta)^{1-\zeta} + (1-\alpha)^{\zeta}\beta^{1-\zeta})}{\zeta \inf_{\theta_0 \in \omega_0} I(\theta:\theta_0) + (1-\zeta) \inf_{\theta_1 \in \omega_1} I(\theta:\theta_1)}$$

If $\theta \in \omega_0$, the above ratio becomes

$$\frac{1}{1-\zeta} \log \left\{ (1-\beta) \left(\frac{\alpha}{1-\beta} \right)^{\zeta} + \beta \left(\frac{1-\alpha}{\beta} \right)^{\zeta} \right\} / \inf_{\theta_1 \in \omega_1} I(\theta:\theta_1)$$

which is continuous in ζ . Hence letting $\zeta \rightarrow 1$ we obtain the first inequality of (4.7). The second inequality will be obtained similarly.

Theorem 4.3 is clearly a generalization of Theorem 4.1 which is the simple-hypotheses case. The lower bound of (4.7) for the expected sample size is attained by Wald sequential tests with specific choice of $f_{\theta}(x)$, $\theta_0 \in \omega_0$, and $\theta_1 \in \omega_1$ except for trivial cases. Exceptions for trivial cases are due to the fact that there may exist, when $\alpha + \beta = 1$, a trivial test satisfying (4.7) with both sides equal to 0 which rejects H_0 with probability α without sampling any observations.

Thus far, we have not discussed sequential analysis from a large-sample point of view. At first glance, it may seem as though the very nature of sequential analysis is such as to rule out large-sample theory. That it is not so becomes clear when one considers that reducing the cost of sampling should increase the expected sample size. In fact, let us suppose that the cost per observation is c . Consider the Bayes procedure corresponding to a fixed a priori probability ζ that H_0 is true. The expected risk

$$\zeta R_0 + (1-\zeta)R_1 = \zeta(w_0\alpha + cE_0(n)) + (1-\zeta)(w_1\beta + cE_1(n))$$

is minimized by a Wald sequential probability-ratio test. As $c \rightarrow 0$, $E_0(n)$ and $E_1(n) \rightarrow \infty$, but $\zeta R_0 + (1-\zeta)R_1 \rightarrow 0$. Minimizing the Wald approximation (4.4)

$$\zeta R_0 + (1-\zeta)R_1 \approx \zeta \left(w_0 e^{-A} - \frac{cB}{I_0} \right) + (1-\zeta) \left(w_1 e^B + \frac{cA}{I_1} \right)$$

for the expected risk, where $I_i = I(i:1-i)$, $i = 0,1$, we find that

$$A \approx -\log c + \log(\zeta w_0 I_1 / (1-\zeta)) \approx -\log c$$

$$B \approx \log c - \log((1-\zeta) w_1 I_0 / \zeta) \approx \log c$$

$$\alpha \approx \frac{\zeta}{1-\zeta} \cdot \frac{c}{w_0 I_1}$$

$$\beta \approx \frac{1-\zeta}{\zeta} \cdot \frac{c}{w_1 I_0}$$

(4.10)

$$E_0(n) \approx -\log c / I_0$$

$$E_1(n) \approx -\log c / I_1$$

$$R_0 = w_0 \alpha + c E_0(n) \approx -c \log c / I_0$$

$$R_1 = w_1 \beta + c E_1(n) \approx -c \log c / I_1 .$$

The risk corresponding to the optimum strategy is mainly the cost of experimentation. The optimum strategy and its risk depend mainly on c , I_0 and I_1 and are relatively insensitive to the loss w_0 and w_1 of making the wrong decision and to the a priori probability ζ .

PROBLEMS

(1) Prove the inequality (4.6) by using the fact that

$$I(f:g) = \frac{1}{2} (x + y)^2,$$

where $f(t) \equiv \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t-x)^2}{2}\right\}$ and $g(t) \equiv \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t+y)^2}{2}\right\}$, and the convex property of $I(f:g)$ (see the corollary to theorem 2.1)

(2) Sketch the function

$$F(\alpha, \beta) = \frac{\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}}{\frac{1}{2} (U_\alpha + U_\beta)^2}$$

in the unit square $0 \leq \alpha, \beta \leq 1$.

(3) Let $f(x, \theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\}$, $w_0 = (-\infty, \delta]$, $w_1 = [\delta, \infty)$, where

δ is a given positive number. Consider the case $\alpha = \beta$ in theorem 4.3.

(a) Show that when $\theta=0$ the lower bound of the expected sample number of any closed sequential test (that is, the right-hand side of (4.9)) is

$$-\frac{1}{\delta^2} \log \left\{ 4\alpha(1-\alpha) \right\} \quad (\equiv M, \text{ say}).$$

(b) Show that the non-sequential uniformly most powerful test with sample size m and size α , that is,

$$\sup_{\theta \in w_0} P_\theta \left\{ \text{test accepts } H_1 \right\} = \alpha,$$

is the test with the critical region

$$\bar{x} + \delta \geq U_\alpha / \sqrt{m}, \quad \text{where } \frac{1}{\sqrt{2\pi}} \int_{U_\alpha}^{\infty} e^{-t^2/2} dt = \alpha.$$

(c) If for the above non-sequential test

$$\sup_{\theta \in \omega_1} P_{\theta} \left\{ \text{test accepts } H_0 \right\} \leq \alpha$$

then show that the least sample size N must be the smallest integer $\geq \left(\frac{U_{\alpha}}{\delta} \right)^2$.

(d) If $0 \leq \alpha \leq \frac{1}{2}$ and δ is taken such that $\frac{U_{\alpha}}{\delta}$ is an integer, then show

that

$$\frac{1}{2} \leq \frac{M}{N} = \frac{-\log \{4\alpha(1-\alpha)\}}{U_{\alpha}^2} \leq \frac{2}{\pi}.$$

5. Statistical experiments and information provided by them

Experiment. An observation of a univariate random variable X is said to be a performance of an experiment with the random variable X . Hence the experiment will result in an observation x , belonging to a space \mathcal{X} . The space \mathcal{X} has a σ -field B of subsets. We shall consider a dominated parameteric set of probability measures, each defined on the measurable space (\mathcal{X}, B) . We shall describe it by $\{p(x|\theta) | \theta \in \Theta\}$, where $p(x|\theta)$ denotes the generalized p.d.f. with respect to a common dominating measure, and Θ is any parameter space. Then the couple

$$(5.1) \quad E = \left[(\mathcal{X}, B), \{p(x|\theta) | \theta \in \Theta\} \right]$$

characterizes an experiment E .

With this definition, the notion of the experiment corresponds to the following communication system with noise (Fig. 5.1). It consists of essentially two parts:

(i) The input space is the set Θ of symbols θ . These symbols are transmitted one by one by some discrete stochastic process, in which each choice of θ is made with probability $p(\theta)$, successive choices being independent.

(ii) The noisy channel is such that the output space is a set \mathcal{X} . We assume that successive symbols are independently perturbed by the noise. The channel, therefore, is described by the set of transition probabilities $p(x|\theta)$, $\theta \in \Theta$, the probability of the transmitted symbol θ being received as $x \in \mathcal{X}$.

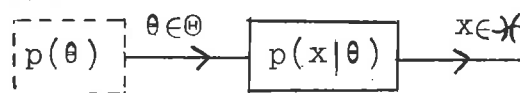


Figure 5.1

The statistical decision problem of deciding which θ is the "true" transmitted symbol will be represented by Fig. 5.2.

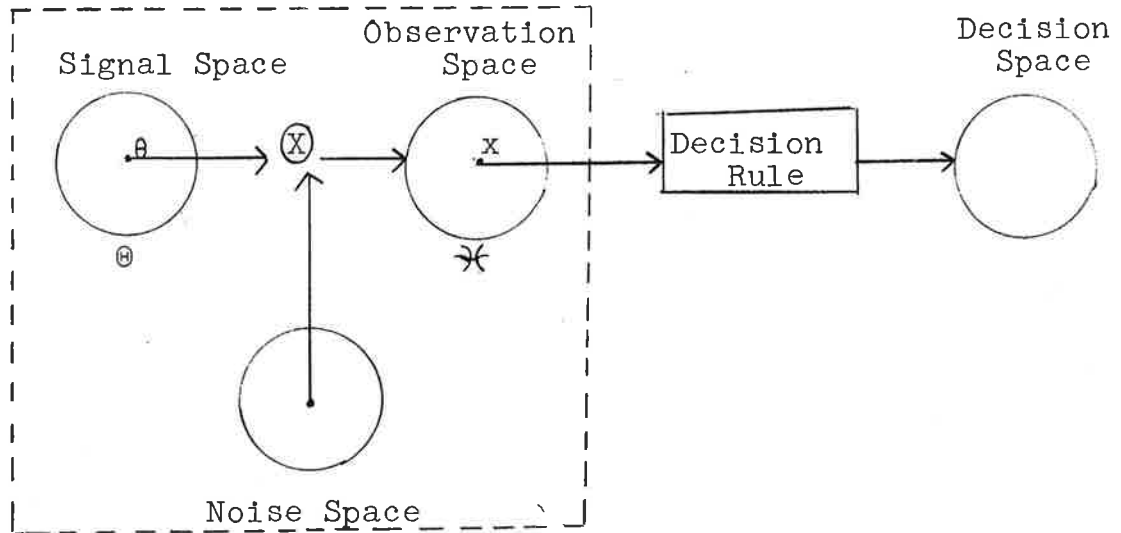


Figure 5.2

The part in the frame of the broken line in Fig. 5.2 is the "communication channel with noise" and this part has no relation with statistical decision theory. Fig. 5.1 is equivalent to this part. In order to

clarify the input and output spaces we sometimes write (5.1) as

$$(5.1') \quad \mathcal{E} = \left[\Theta, \left\{ p(x|\theta) \mid \theta \in \Theta \right\}, \mathcal{X} \right].$$

The transmission rate of the noisy channel is defined by Shannon (1948) as the difference of input entropy and the received conditional entropy.

Hence, when the a priori p.d.f. of input symbols θ is $p(\theta)$, the transmission rate is given by

$$\begin{aligned} T(\theta; x) &= H(\theta) - H_x(\theta) \\ &= \iint p(\theta)p(x|\theta) \log \frac{p(x|\theta)}{p(x)} d\theta dx \\ &= \int p(\theta) I \left[P(x|\theta) : p(x) \mid p(\theta) \right] d\theta, \end{aligned}$$

where $p(x) = \int p(\theta)p(x|\theta)d\theta$. Hereafter, for simplicity in notation, we

shall not distinguish between random variables and the values assumed by them, nor shall we attempt to be specific in describing the density functions. Thus $p(\theta)$ and $p(x)$ will denote the density functions of the random variables θ and x , respectively, without any suggestion that they have the same density. Moreover we shall denote integration with respect to the dominating measures on \mathcal{X} and Θ by dx and $d\theta$ respectively, again for simplicity of notation. We shall, following Lindley (1956), define the amount of information provided by the experiment (5.1) with the prior knowledge $p(\theta)$ by

$$(5.2) \quad I(\mathcal{E}, p(\theta)) = T(\theta; x) = \iint p(\theta) p(x|\theta) \log \frac{p(x|\theta)}{p(x)} d\theta dx.$$

We give an example. Let \mathcal{E} be a dichotomous experiment:

$$(5.3) \quad \mathcal{E} = [(\mathcal{X}, B), \{f_1(x), f_2(x)\}].$$

Then if the prior probability is ζ for $\theta=1$ we have (fig. 5.3)

(5.4)

$$I(\mathcal{E}, \zeta) = \zeta \int f_1(x) \log \frac{f_1(x)}{\zeta f_1(x) + (1-\zeta) f_2(x)} dx + (1-\zeta) \int f_2(x) \log \frac{f_2(x)}{\zeta f_1(x) + (1-\zeta) f_2(x)} dx$$

$$= \zeta I(f_1; f_\zeta) + (1-\zeta) I(f_2; f_\zeta),$$

$$\frac{d}{d\zeta} I(\mathcal{E}, \zeta) = I(f_1; f_\zeta) - I(f_2; f_\zeta),$$

$$\frac{d}{d\zeta} I(\mathcal{E}, \zeta) |_{\zeta=0} = I(f_1; f_2),$$

$$\frac{d}{d\zeta} I(\mathcal{E}, \zeta) |_{\zeta=1} = -I(f_2; f_1),$$

$$\frac{d^2}{d\zeta^2} I(\mathcal{E}, \zeta) = -\int (f_1 - f_2)^2 / f_\zeta dx,$$

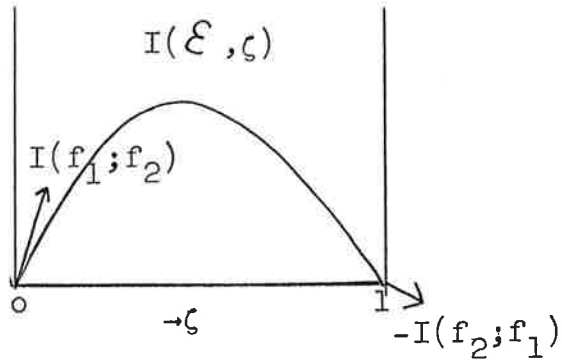


Figure 5.3

where $f_{\zeta}(x) = \zeta f_1(x) + (1-\zeta)f_2(x)$.

For example, if the $f_i(x)$'s are binomial densities:

$$f_i(x) = \theta_i^x (1-\theta_i)^{1-x} \quad (i = 1, 2; \quad x = 0, 1),$$

then we have

(5.5)

$$I(\mathcal{E}(\theta_1, \theta_2), \zeta) = \zeta I(\theta_1; \theta_{\zeta}) + (1-\zeta) I(\theta_2; \theta_{\zeta}) = S(\theta_{\zeta}) - \zeta S(\theta_1) - (1-\zeta) S(\theta_2),$$

where $\theta_{\zeta} = \zeta \theta_1 + (1-\zeta) \theta_2$ and

$$I(\theta_i, \theta_{\zeta}) = \theta_i \log \frac{\theta_i}{\theta_{\zeta}} + (1-\theta_i) \log \frac{1-\theta_i}{1-\theta_{\zeta}}, \quad (i = 1, 2),$$

$$S(\theta) \equiv -\theta \log \theta - (1-\theta) \log(1-\theta), \quad (0 \leq \theta \leq 1).$$

Example 5.1 Normal experiment.

Let $\mathcal{E}(\sigma) = [(-\infty, \infty), \{p(x|\theta) \mid -\infty < \theta < \infty\}, (-\infty, \infty)]$, where

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}. \quad \text{If we take } p(\theta) = \frac{1}{\sqrt{2\pi}v} \exp\left\{-\frac{(\theta-m)^2}{2v^2}\right\}$$

then

$$p(x) = \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi}v} \exp\left\{-\frac{(\theta-m)^2}{2v^2}\right\} d\theta = \frac{1}{\sqrt{2\pi(\sigma^2+v^2)}} \exp\left\{-\frac{(x-m)^2}{2(\sigma^2+v^2)}\right\}$$

$$\begin{aligned}
I(p(x|\theta): p(x)|p(\theta)) &= (1/2) \left(\frac{\sigma^2}{\sigma^2+v^2} - 1 + \frac{(\theta-m)^2}{\sigma^2+v^2} - \log \frac{\sigma^2}{\sigma^2+v^2} \right) \\
&= (1/2) \left(\log \left(1 + \frac{v^2}{\sigma^2} \right) - \frac{v^2}{\sigma^2+v^2} + \frac{(\theta-m)^2}{\sigma^2+v^2} \right),
\end{aligned}$$

and

$$\begin{aligned}
I(\xi(\sigma), p(\theta)) &= \int p(\theta) I(p(x|\theta): p(x)|p(\theta)) d\theta \\
&= \int \frac{1}{\sqrt{2\pi} v} \exp\left\{-\frac{(\theta-m)^2}{2v^2}\right\} \cdot (1/2) \left(\log \left(1 + \frac{v^2}{\sigma^2} \right) - \frac{v^2}{\sigma^2+v^2} + \frac{(\theta-m)^2}{\sigma^2+v^2} \right) d\theta \\
&= (1/2) \log \left(1 + \frac{v^2}{\sigma^2} \right).
\end{aligned}$$

Example 5.2 Binomial experiment.

Let $\xi = \left[[0,1], \{p(x|\theta) | 0 \leq \theta \leq 1\}, [0,1] \right]$, where $p(x|\theta) = \theta^x (1-\theta)^{1-x}$, ($x = 0,1; 0 \leq \theta \leq 1$). If we take $p(\theta) = \theta^{a-1} (1-\theta)^{b-1} / B(a,b)$ ($0 \leq \theta \leq 1; a, b > 0$) then

$$\begin{aligned}
p(x) &= \int_0^1 p(\theta) p(x|\theta) d\theta = \int_0^1 \frac{\theta^{a-1} (1-\theta)^{b-1}}{B(a,b)} \theta^x (1-\theta)^{1-x} d\theta \\
&= \frac{B(a+x, b+1-x)}{B(a,b)} = \begin{cases} \frac{a}{a+b}, & \text{if } x = 1 \\ \frac{b}{a+b}, & \text{if } x = 0, \end{cases}
\end{aligned}$$

$$I(p(x|\theta): p(x)|p(\theta)) = \theta \log \frac{\theta}{a/(a+b)} + (1-\theta) \log \frac{1-\theta}{b/(a+b)},$$

and

$$\begin{aligned}
I(\xi, p(\theta)) &= \int_0^1 p(\theta) I(p(x|\theta): p(x)|p(\theta)) d\theta \\
&= \int_0^1 \frac{\theta^{a-1} (1-\theta)^{b-1}}{B(a,b)} \left(\theta \log \frac{\theta}{a/(a+b)} + (1-\theta) \log \frac{1-\theta}{b/(a+b)} \right) d\theta.
\end{aligned}$$

Note that if y is a $B(r,s)$ -distributed random variable, then

$$E(\log y) = \psi(r) - \psi(r+s), \quad E(\log(1-y)) = \psi(s) - \psi(r+s),$$

where $\psi(r) = \frac{d}{dr} \log \Gamma(r)$. Using this we have

$$\begin{aligned}
I(\mathcal{E}, p(\theta)) &= S\left(\frac{a}{a+b}\right) + \frac{a}{a+b} \psi(a+1) + \frac{b}{a+b} \psi(b+1) - \psi(a+b+1) \\
&= S\left(\frac{a}{a+b}\right) + \frac{a\psi(a) + b\psi(b)}{a+b} - \psi(a+b) + \frac{1}{a+b}
\end{aligned}$$

where $\psi(r+1) = \psi(r) + 1/r$

and $S(\theta) = -\theta \log \theta - (1-\theta) \log(1-\theta)$, $(0 \leq \theta \leq 1)$.

Sum, average and mixture of two experiments.

Given the two experiments

$$\mathcal{E}_i = \left[\Theta, \{p(x_i|\theta) | \theta \in \Theta\}, \mathcal{X}_i \right] \quad (i = 1, 2)$$

with a common input space, we define the sum of the two experiments as

$$(5.6) \quad (\mathcal{E}_1, \mathcal{E}_2) \equiv \left[\Theta, \{p(x_1, x_2|\theta) | \theta \in \Theta\}, \mathcal{X}_1 \times \mathcal{X}_2 \right]$$

where $p(x_1, x_2|\theta)$ is any p.d.f. with the marginal densities

$$\int p(x_1, x_2|\theta) dx_i = p(x_{3-i}|\theta), \quad (i = 1, 2; \theta \in \Theta).$$

We have, by the additive law of information transmission

$$(5.7) \quad I((\mathcal{E}_1, \mathcal{E}_2)) = T(\theta; x_1, x_2) = T(\theta; x_1) + T_{x_1}(\theta; x_2),$$

where the last term is the expected value w.r.t. $p(x_1)$ of the transmitted information based on the conditional p.d.f. $p(\theta, x_2|x_1)$.

Or more precisely

$$\begin{aligned}
I((\mathcal{E}_1, \mathcal{E}_2), p(\theta)) &= \iiint p(\theta) p(x_1, x_2|\theta) \log \frac{p(x_1, x_2|\theta)}{p(x_1, x_2)} d\theta dx_1 dx_2 \\
&= \iiint p(\theta) p(x_1, x_2|\theta) \left(\log \frac{p(x_1|\theta)}{p(x_1)} + \log \frac{p(x_2|\theta, x_1)}{p(x_2|x_1)} \right) d\theta dx_1 dx_2 \\
&= \iint p(\theta) p(x_1|\theta) \log \frac{p(x_1|\theta)}{p(x_1)} d\theta dx_1 \\
&\quad + \int p(x_1) dx_1 \iint p(\theta|x_1) p(x_2|\theta, x_1) \log \frac{p(x_2|\theta, x_1)}{p(x_2|x_1)} d\theta dx_2
\end{aligned}$$

which is, say,

$$(5.8) \quad I((\mathcal{E}_1, \mathcal{E}_2), p(\theta)) = I(\mathcal{E}_1, p(\theta)) + E_{x_1} \left\{ I(\mathcal{E}_2(x_1), p(\theta|x_1)) \right\}$$

The experiment $\mathcal{E}_2(x_1)$ in the above expression may be represented by Fig. 5.4

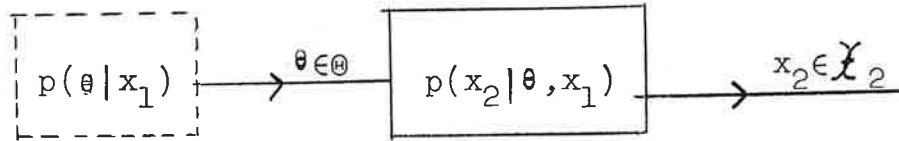


Figure 5.4

Let us call \mathcal{E}_1 and \mathcal{E}_2 (mutually) independent (in the sum $\mathcal{E}_1 + \mathcal{E}_2$) if

$$p(x_1, x_2 | \theta) = p(x_1 | \theta) p(x_2 | \theta), \text{ for all } \theta \in \Theta.$$

If \mathcal{E}_1 and \mathcal{E}_2 are independent $\mathcal{E}_2(x_1)$ is equivalent to \mathcal{E}_2 .

Theorem 5.1 If the two experiments \mathcal{E}_1 and \mathcal{E}_2 in the sum (5.6) are independent we have

$$(i) \quad I(\mathcal{E}_2) - E_{x_1} \left\{ I(\mathcal{E}_2(x_1), p(\theta|x_1)) \right\} = T(x_1; x_2)$$

(ii) $I((\mathcal{E}_1, \mathcal{E}_2)) \leq I(\mathcal{E}_1) + I(\mathcal{E}_2)$, with equality if and only if X_1 and X_2 are statistically independent.

Proof. The left-hand side of (i) = $I(\mathcal{E}_2) + I(\mathcal{E}_1) - I((\mathcal{E}_1, \mathcal{E}_2))$
 $= T(\theta; x_2) + T(\theta; x_1) - T(\theta; x_1, x_2)$
 $= (H(x_2) - H_\theta(x_2)) + (H(x_1) - H_\theta(x_1)) - (H(x_1, x_2) - H_\theta(x_1, x_2))$
 $= H(x_2) + H(x_1) - H(x_1, x_2) = T(x_1; x_2),$

since we have $H_\theta(x_1, x_2) = H_\theta(x_1) + H_\theta(x_2)$ from the independence of \mathcal{E}_1 and \mathcal{E}_2 . (ii) follows from (5.8) and (i).

It is worthwhile to note the following relations: If we denote, for simplicity, the mean information of the conditional experiment in the last term of (5.8) by $I(\mathcal{E}_2 | \mathcal{E}_1)$, we have

$$(5.8') \quad I((\mathcal{E}_1, \mathcal{E}_2)) = I(\mathcal{E}_1) + I(\mathcal{E}_2 | \mathcal{E}_1),$$

and if \mathcal{E}_1 and \mathcal{E}_2 are independent

$$(5.9) \quad I((\mathcal{E}_1, \mathcal{E}_2)) = I(\mathcal{E}_1) + I(\mathcal{E}_2) - T(x_1; x_2),$$

$$(5.10) \quad I(\mathcal{E}_2) \geq I(\mathcal{E}_2 | \mathcal{E}_1).$$

Corollary 5.1 For repetitive performances $\mathcal{E}^{(n)} = (\mathcal{E}, \dots, \mathcal{E})$ of the independent and common experiments, the information $I(\mathcal{E}^{(n)})$ is a concave, increasing function of n .

Proof. It suffices to show that

$$0 \leq j_{n+1} - j_n \leq j_n - j_{n-1}$$

where $j_n = I(\mathcal{E}^{(n)})$. We have by (5.8') and (5.10)

$$j_{n+1} - j_n = I(\mathcal{E}_{n+1} | \mathcal{E}^{(n)}) \geq 0,$$

$$j_n - j_{n-1} = I(\mathcal{E}_n | \mathcal{E}^{(n-1)}) = I(\mathcal{E}_{n+1} | \mathcal{E}^{(n-1)}) \geq I(\mathcal{E}_{n+1} | \mathcal{E}^{(n)}).$$

For other important complex experiments we introduce in the following the weighted average and the mixture. As before, let

$$\mathcal{E}_i = \left[\Theta, \{p(x_i | \theta) | \theta \in \Theta\}, \mathcal{X}_i \right] \quad (i = 1, 2)$$

be two experiments with the common input space Θ . If $0 \leq \zeta \leq 1$ and

$\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ we call the experiment

$$(5.11) \quad \zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2 = \left[\Theta, \{p(x | \theta) | \theta \in \Theta\}, \mathcal{X}_1 \cup \mathcal{X}_2 \right]$$

where

$$p(x | \theta) = \begin{cases} \zeta p(x_1 | \theta), & \text{if } x = x_1 \in \mathcal{X}_1 \\ (1-\zeta) p(x_2 | \theta), & \text{if } x = x_2 \in \mathcal{X}_2, \end{cases}$$

a weighted average of \mathcal{E}_1 and \mathcal{E}_2 (with weight ζ on \mathcal{E}_1).

Now let

$$\mathcal{E}_i = \left[\Theta, \{p_i(x | \theta) | \theta \in \Theta\}, \mathcal{X}_i \right] \quad (i = 1, 2)$$

be two experiments with the common input space Θ and the same output space \mathcal{X} . We call an experiment

$$(5.12) \zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2 \equiv \left[\Theta, \left\{ \zeta p_1(x|\theta) + (1-\zeta)p_2(x|\theta) \mid \theta \in \Theta \right\}, \mathcal{X} \right]$$

a mixture of \mathcal{E}_1 and \mathcal{E}_2 (with weight ζ on \mathcal{E}_1).

The weighted average and the mixture of two experiments will be represented by diagrams as in Fig. 5.5 (a) and (b) respectively.

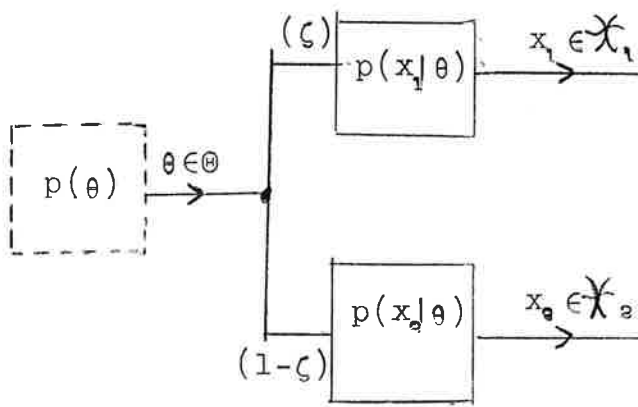


Figure 5.5 (a)

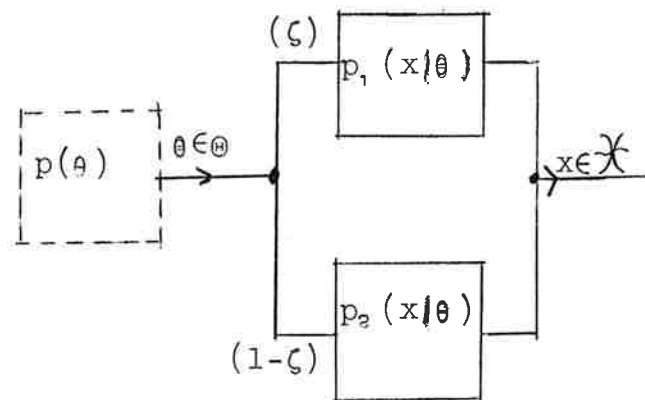


Figure 5.5 (b)

The weighted average $\zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2$, defined by (5.11) and represented by Fig. 5.5 (a), can be thought of as being performed as follows: With probability ζ , a value x_1 is obtained according to the density $p(x_1) = \int p(\theta)p(x_1|\theta)d\theta$; with probability $1-\zeta$, x_2 is obtained according to $p(x_2) = \int p(\theta)p(x_2|\theta)d\theta$. The experimenter is informed not only of a value x_1 or x_2 , but also which event of probability ζ or $1-\zeta$ took place. On the other hand, the mixture $\zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2$, defined by (5.12) and represented by Fig. 5.5 (b), can be thought of as follows: A value x is obtained according to $p_1(x|\theta)$ and $p_2(x|\theta)$ with probabilities ζ and $1-\zeta$ respectively. The experimenter, in this case, is informed only of x , and not of

which event of probability ζ or $1-\zeta$, took place.

As may be intuitively expected we have

Theorem 5.2 (i) For the weighted average (5.11)

$$I(\zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2) = \zeta I(\mathcal{E}_1) + (1-\zeta) I(\mathcal{E}_2)$$

(ii) For the mixture (5.12)

$$I(\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2) \leq \zeta I(\mathcal{E}_1) + (1-\zeta) I(\mathcal{E}_2).$$

Proof. (i) is evident. Let \mathcal{E}^* denote an experiment which informs which events of probability ζ or $1-\zeta$ took place. Then

$$(\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2, \mathcal{E}^*) = \zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2.$$

Hence we have

$$\begin{aligned} I(\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2) &\leq I((\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2, \mathcal{E}^*)) \\ &= I(\zeta \mathcal{E}_1 + (1-\zeta) \mathcal{E}_2) = \zeta I(\mathcal{E}_1) + (1-\zeta) I(\mathcal{E}_2). \end{aligned}$$

by (5.8') and the first part of this theorem.

More precisely in the second part of the above theorem, we have

$$\begin{aligned} &\zeta I(\mathcal{E}_1) + (1-\zeta) I(\mathcal{E}_2) - I(\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2) \\ &= \zeta \iint p(\theta) p_1(x|\theta) \log \frac{p_1(x|\theta)}{p_1(x)} d\theta dx + (1-\zeta) \iint p(\theta) p_2(x|\theta) \log \frac{p_2(x|\theta)}{p_2(x)} d\theta dx \\ &\quad - \iint p(\theta) (\zeta p_1(x|\theta) + (1-\zeta) p_2(x|\theta)) \log \frac{\zeta p_1(x|\theta) + (1-\zeta) p_2(x|\theta)}{\zeta p_1(x) + (1-\zeta) p_2(x)} d\theta dx \\ &= \int p(\theta) d\theta \left[\left\{ \zeta I(p_1(x|\theta); p_\zeta(x|\theta)) + (1-\zeta) I(p_2(x|\theta); p_\zeta(x|\theta)) \right\} \right. \\ &\quad \left. - \left\{ \zeta I(p_1(x); p_\zeta(x)) + (1-\zeta) I(p_2(x); p_\zeta(x)) \right\} \right], \end{aligned}$$

where $p_\zeta(x|\theta) = \zeta p_1(x|\theta) + (1-\zeta) p_2(x|\theta)$,

$$p_\zeta(x) = \zeta p_1(x) + (1-\zeta) p_2(x).$$

The amount of information $I(\mathcal{E}, p(\theta))$ provided by an experiment is

convex in the mixture of experiments as the above theorem shows, and it is concave in the prior knowledge: that is,

Theorem 5.3 $I(\mathcal{E}, p(\theta))$ is concave in the prior knowledge. That is,
if $p_1(\theta)$ and $p_2(\theta)$ are two prior knowledges, then

$I(\mathcal{E}, tp_1(\theta) + (1-t)p_2(\theta)) \geq tI(\mathcal{E}, p_1(\theta)) + (1-t)I(\mathcal{E}, p_2(\theta)),$
 for all $0 \leq t \leq 1$.

Proof. It is readily shown that

$$(5.14) \quad I(\mathcal{E}, tp_1(\theta) + (1-t)p_2(\theta)) = I(\mathcal{E}^*, t) + tI(\mathcal{E}, p_1(\theta)) + (1-t)I(\mathcal{E}, p_2(\theta)),$$

where

$$\mathcal{E}^* = \left[(\mathcal{X}, B), \{p_1(x), p_2(x)\} \right],$$

$$p_i(x) = \int p_i(\theta) p(x|\theta) d\theta \quad (i = 1, 2),$$

and

$$I(\mathcal{E}^*, t) = t \int p_1(x) \log \frac{p_1(x)}{tp_1(x) + (1-t)p_2(x)} dx \\
 + (1-t) \int p_2(x) \log \frac{p_2(x)}{tp_1(x) + (1-t)p_2(x)} dx.$$

The above theorem will have several applications in later sections.

Uncertainty and information functions.

When the input space Θ is finite, we say that the experiment is finite. For finite experiments non-negativity which is the most fundamental property of the information provided by the experiment is not particular to Shannon's uncertainty measure. In fact the next theorem 5.4 shows that for any concave uncertainty function we can well define the amount of information provided by the experiment by the amount of decrease of the uncertainty after the performance of the experiment.

Let

$$(5.15) \quad \mathcal{E} = \left[(\mathcal{X}, B), \left\{ f_1(x), \dots, f_k(x) \right\} \right]$$

be a finite experiment. Let \mathfrak{H} denote the space of all probability- k vectors $\xi = (\xi_1, \dots, \xi_k)$, i.e., $\xi_i \geq 0$ ($i = 1, \dots, k$) and $\sum_1^k \xi_i = 1$.

An uncertainty function U is a non-negative measurable function defined on \mathfrak{H} . Intuitively, the value $U(\xi)$ is meant to represent the uncertainty of an experimenter about the true value of θ when his prior knowledge over Θ is ξ . The information $I[\mathcal{E}, \xi; U]$ in a finite experiment (5.15) when the prior knowledge is ξ , relative to the uncertainty function U , is defined as

$$(5.16) \quad I[\mathcal{E}, \xi; U] = U(\xi) - E[U(\xi(X)) | \xi],$$

where $\xi(x) = (\xi_1(x), \dots, \xi_k(x))$, with

$$\xi_i(x) = \xi_i f_i(x) / \sum_{j=1}^k \xi_j f_j(x) \quad (i = 1, \dots, k),$$

and the expectation $E[\cdot | \xi]$ means $\sum_{i=1}^k \xi_i E_i[\cdot]$.

Theorem 5.4 Let U be a given uncertainty function defined on \mathfrak{H} . Then $I[\mathcal{E}, \xi; U] \geq 0$, for all experiments (5.15) and all $\xi \in \mathfrak{H}$, if and only if U is concave.

Proof. If $U(\xi)$ is concave, then by the familiar Jensen's inequality $E[U(\xi(X)) | \xi] \leq U(E[\xi(X) | \xi]) = U(\xi)$, for all \mathcal{E} and $\xi \in \mathfrak{H}$.

It follows that $I[\mathcal{E}, \xi; U] \geq 0$.

Conversely suppose that $I[\mathcal{E}, \xi; U] \geq 0$ for all \mathcal{E} and ξ . Let ξ and ν be any two vectors in \mathfrak{H} and let $0 < t < 1$. Consider the experiment (5.15) in which $f_j(x)$'s are binomial densities with the parameters

$$t\xi_j / (t\xi_j + (1-t)\nu_j), \quad (j = 1, \dots, k).$$

Let $\pi = t\xi + (1-t)\nu$. If the prior knowledge is π , then the posterior probabilities after observing X are

$$\pi(1) = \left(\frac{\pi_1 t \xi_1 / (t \xi_1 + (1-t) \nu_1)}{\sum_{j=1}^k \pi_j t \xi_j / (t \xi_j + (1-t) \nu_j)}, \dots \right) = \xi,$$

$$\pi(0) = \left(\frac{\pi_1 (1-t) \nu_1 / (t \xi_1 + (1-t) \nu_1)}{\sum_{j=1}^k \pi_j (1-t) \nu_j / (t \xi_j + (1-t) \nu_j)}, \dots \right) = \nu.$$

Hence, since, by assumption, $I[\xi, \xi; U] \geq 0$ it follows that

$$U(t\xi + (1-t)\nu) = U(\pi) \geq E[U(\pi(X)) | \pi] = tU(\pi(1)) + (1-t)U(\pi(0))$$

$$= tU(\xi) + (1-t)U(\nu).$$

The above theorem indicates that it might be reasonable to consider some concave uncertainty functions, other than the famous Shannon entropy function $U(\xi) = - \sum_{j=1}^k \xi_j \log \xi_j$. An example of such a function is

$$(5.17) \quad U(\xi) = \min(\xi_1, \dots, \xi_k).$$

Using this uncertainty measure the information in an experiment becomes

$$(5.18) \quad I[\xi, \xi; U] = \min_{1 \leq i \leq k} \xi_i - \int \min_{1 \leq i \leq k} (\xi_i f_i(x)) dx.$$

An important class of concave uncertainty functions can be derived from standard statistical decision problems. In a statistical decision problem there is given a decision space A and a loss function $L(\theta, a)$, assumed to be non-negative and bounded on $\Theta \times A$. Let

$$(5.19) \quad U(\xi) \equiv \inf_{a \in A} \sum_{j=1}^k \xi_j L(\theta_j, a).$$

This is the risk from the optimal decision. Since it is known that the

above $U(\xi)$ is continuous and concave on Ξ , $E [U(\xi(X)) | \xi]$ is the risk resulting from the Bayes' decision procedure using the observation X , and the information $I [\mathcal{E}, \xi; U]$ is the reduction in risk that can be attained by performing the experiment \mathcal{E} .

If we take, for example, $A = \{1, \dots, k\}$ and $L(\theta, a) = \begin{cases} 0, & \text{if } j = a \\ 1, & \text{if } j \neq a \end{cases}$ then (5.19) yields (5.17) and (5.18). In particular if $k = 2$, (5.18) is the reduction of the risk of Bayes decision rule deciding between two densities f_1 and f_2 with usual zero-one loss:

$$I [\mathcal{E}, \xi; U] = \min(\xi_1, \xi_2) - \left[\begin{array}{cc} \xi_1 \int_{\frac{f_2 > \xi_1}{f_1 \xi_2}} f_1(x) dx + \xi_2 \int_{\frac{f_2 \leq \xi_1}{f_1 \xi_2}} f_2(x) dx \end{array} \right].$$

Sequential sampling rules

Let \mathcal{E} be an experiment with observation X that can be replicated independently and indefinitely. Then a random sequential sample of observations X_1, X_2, \dots , each X_i having the same distribution as X , can be obtained.

Consider the sequential sampling rule whereby observations are taken as long as $U(\xi(X_1, \dots, X_n)) > \delta$ for some given $\delta > 0$, and sampling stops as soon as $U(\xi(X_1, \dots, X_n)) \leq \delta$ for some value of n . This sampling rule yields, in some cases, reasonable statistical procedures.

Example 5.3 Suppose that Θ contains only two points, and \mathcal{E} is a dichotomous experiment (5.3). Suppose that U is a continuous, concave function of ξ_1 with $U(0) = U(1) = 0$. Then, for any a priori probability ξ_1 for $\theta = 1$, sampling as long as $U(\xi_1(X_1, \dots, X_n)) > \delta$ is equivalent to sampling as long as $\delta_1 < \xi_1(X_1, \dots, X_n) < \delta_2$ for some δ_1 and δ_2 ,

which in turn is equivalent to sampling as long as $A < \prod_{i=1}^n \frac{f_2(X_i)}{f_1(X_i)} < B$.

Thus the sampling rule is a Wald sequential probability ratio test.

Example 5.4 Suppose that $\Theta = (-\infty, \infty)$ and $\mathcal{E}(\sigma)$ is the normal experiment defined in Example 5.1. Suppose that the uncertainty function $U(p(\theta))$ for the prior distribution $p(\theta)$ over Θ is taken as $U(p(\theta)) = -\int p(\theta) \log p(\theta) d\theta$. If we consider $p(\theta) = \frac{1}{\sqrt{2\pi} v} \exp\left\{-\frac{(\theta-m)^2}{2v^2}\right\}$

$\equiv \frac{1}{v} \phi\left(\frac{\theta-m}{v}\right)$, say

$$p(\theta | x_1, \dots, x_n) = p(\theta) \frac{\prod_{i=1}^n p(x_i | \theta)}{\int p(\theta) \prod_{i=1}^n p(x_i | \theta) d\theta}$$

$$= \left(\frac{1}{v^2} + \frac{1}{\sigma^2/n}\right)^{1/2} \phi\left(\frac{\theta - \left(\frac{m}{v^2} + \frac{\bar{x}}{\sigma^2/n}\right)}{\left(\frac{1}{v^2} + \frac{1}{\sigma^2/n}\right)^{-1/2}}\right),$$

then $U(p(\theta | x_1, \dots, x_n)) = \frac{1}{2} + \log \left\{ \sqrt{2\pi} \left(\frac{1}{v^2} + \frac{1}{\sigma^2/n}\right) \right\}$.

Thus, for a given prior distribution $p(\theta)$, sampling as long as $U(p(\theta | X_1, \dots, X_n)) > \delta$ is equivalent to taking a sample of fixed size.

Example 5.5 Suppose that $\Theta = [0, 1]$ and \mathcal{E} is the binomial experiment defined in Example 5.2. Suppose that as in the above example, $U(p(\theta)) = -\int p(\theta) \log p(\theta) d\theta$. If we take $p(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$, then $p(\theta | x_1, \dots, x_n) = \theta^{a+\sum_{i=1}^n x_i - 1} (1-\theta)^{b+n-\sum_{i=1}^n x_i - 1} / B(a+\sum_{i=1}^n x_i, b+n-\sum_{i=1}^n x_i)$. It is easily shown

that if y is a $B(r,s)$ -distributed random variable, we have as its entropy

$$H(y) \equiv -\int p(y) \log p(y) dy$$

$$= \log B(r,s) + (r+s-2)\psi(r+s) - (r-1)\psi(r) - (s-1)\psi(s),$$

where $\psi(r)$ is defined in Example 5.2, and

$$H(y) \simeq \frac{1}{2} + \log \sqrt{2\pi} + \frac{1}{2} \log \frac{rs}{(r+s)^3}, \text{ for sufficiently large } r,s.$$

Hence, using this result

$$\begin{aligned} U(p(\theta | x_1, \dots, x_n)) &= \log B\left(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i\right) + (a+b-n-2)\psi(a+b-n) \\ &- \left(a + \sum_{i=1}^n x_i - 1\right)\psi\left(a + \sum_{i=1}^n x_i\right) - \left(b + n - \sum_{i=1}^n x_i - 1\right)\psi\left(b + n - \sum_{i=1}^n x_i\right) \\ &\simeq \frac{1}{2} + \log \sqrt{2\pi} + \frac{1}{2} \log \frac{\left(a + \sum_{i=1}^n x_i\right)\left(b + n - \sum_{i=1}^n x_i\right)}{(a+b+n)^3}. \end{aligned}$$

The sampling rule whereby observations are taken as long as $U(p(\theta | x_1, \dots, x_n)) > \delta$ can be described graphically as follows. Suppose the prior distribution over θ is $B(a_0, b_0)$. Then, in the ab -plane, start at the point (a_0, b_0) and after each observation move one unit up or to the right according as the observed value is 1 or 0. Stop sampling as soon as the curve $ab = (a+b)^3 \delta'$, for some appropriate $\delta' > 0$, is crossed (Fig. 5.6).

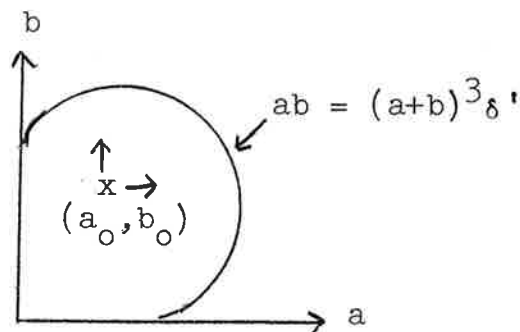


Figure 5.6

PROBLEMS

(1) For binomial dichotomous experiments, we call the quantity $C(\theta_1, \theta_2) = \max_{0 \leq \zeta \leq 1} I(\mathcal{E}(\theta_1, \theta_2), \zeta)$, the capacity of the binomial

dichotomy, and the maximizing $(\zeta^*, 1-\zeta^*)$ the matching input probabilities. Show that ζ^* satisfies the equation

$$\frac{\log \frac{1-\theta_1 \zeta^*}{\theta_1 \zeta^*}}{\theta_1 \zeta^*} = \frac{S(\theta_2) - S(\theta_1)}{\theta_2 - \theta_1},$$

that is,

$$\zeta^* = \left\{ \theta_2 - \left(1 + \exp \left\{ \frac{S(\theta_2) - S(\theta_1)}{\theta_2 - \theta_1} \right\} \right)^{-1} \right\} / (\theta_2 - \theta_1).$$

(2) Let $\mathcal{E}_X = [(X, B), \{f_1(x), f_2(x)\}]$ and $\mathcal{E}_Y = [(Y, C), \{g_1(y), g_2(y)\}]$

be two dichotomous experiments. If

$$I_X(f_1 : f_{3-i}) \geq I_Y(g_1 : g_{3-i}), \quad i = 1, 2,$$

then is it true that

$$I(\mathcal{E}_X, \zeta) \geq I(\mathcal{E}_Y, \zeta), \quad \text{for all } 0 \leq \zeta \leq 1 \quad ?$$

If not construct an example.

(3) In the above problem, let $f_i(x)$ and $g_i(x)$ be binomial densities with parameters a_i, b_i ($i = 1, 2$) with $a_1 < a_2$. Determine in the unit square the following sets of points:

$$(a) \quad I(a_i : a_{3-i}) \geq I(b_i : b_{3-i}), \quad i = 1, 2,$$

$$(b) \quad I(a_i : a_\zeta) \geq I(b_i : b_\zeta), \quad i = 1, 2$$

where $a_\zeta = \zeta a_1 + (1-\zeta)a_2$ and $b_\zeta = \zeta b_1 + (1-\zeta)b_2$.

$$(c) \quad I(\mathcal{E}_X, \zeta) \geq I(\mathcal{E}_Y, \zeta), \quad \text{for all } 0 \leq \zeta \leq 1.$$

(4) Let the p.d.f. of the input symbol be

$$p(\theta) = ae^{-a\theta}, \quad 0 \leq \theta < \infty$$

where a is a given positive constant.

(a) If the channel is given by

$$p(x|\theta) = \begin{cases} 1/\epsilon, & \text{if } |x-\theta| < \epsilon/2 \\ 0, & \text{otherwise,} \end{cases}$$

show that the information provided by this experiment is

$$\frac{1}{a\epsilon} \left(\frac{\pi^2}{6} - \sum_{j=1}^{\infty} \frac{e^{-ja\epsilon}}{j^2} \right).$$

(b) If the channel is given by

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\theta)^2}{2\sigma^2} \right\}, \quad -\infty < x < \infty,$$

show that the information provided by this experiment is

$$-\frac{1}{2} (1 + \gamma^2 + \log(2\pi\gamma^2)) + \exp \left\{ -\frac{\gamma^2}{2} \right\} \sum_{j=2}^{\infty} \frac{\rho_j}{j(j-1)},$$

where $\gamma = a\sigma$ and

$$\rho_j = \int_0^1 e^{x d \Phi^j(x/\sigma)}, \quad \Phi(x) = \int_x^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt.$$

(Rapoport and Horvath, 1960)

(5) Let $\mathcal{E}^{(n)}$ be n independent replications of a common experiment \mathcal{E} .

(a) For $\mathcal{E} = \mathcal{E}(\sigma)$ in Example 5.1 and $p(\theta) = \frac{1}{\sqrt{2\pi}\nu} \exp \left\{ -\frac{(\theta-m)^2}{2\nu^2} \right\}$, show

that
$$I(\mathcal{E}^{(n)}, p(\theta)) = \frac{1}{2} \log \left(1 + \frac{\nu^2}{\sigma^2/n} \right).$$

(b) For $\mathcal{E} =$ binomial experiment with $\Theta = [0,1]$ in Example 5.2 and

$p(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$, calculate the value of $I(\mathcal{E}^{(n)}, p(\theta))$.

(c) In (b), let $U(p(\theta)) = \inf_{0 \leq a \leq 1} \int p(\theta)(a-\theta)^2 d\theta$, i.e., the risk of

Bayes point-estimator of θ . Derive the sequential sampling rule which is equivalent to continue sampling as long as $U(p(\theta|X_1, \dots, X_n)) > \delta$.

(6) Let $\mathcal{E} = [(\mathcal{X}, B), \{f_1(x), f_2(x)\}]$ be a dichotomous experiment. Two densities $f_1(x)$ and $f_2(x)$ generate a parametric family of densities

$$F = \left\{ f_\zeta(x) = \zeta f_1(x) + (1-\zeta)f_2(x) \mid 0 \leq \zeta \leq 1 \right\}.$$

(a) Show that the Fisher information (i.e., intrinsic accuracy) for F is

$$\begin{aligned} I(\zeta) &= \int \left\{ -\frac{\partial^2}{\partial \zeta^2} \log f_\zeta(x) \right\} f_\zeta(x) dx \\ &= \int (f_1 - f_2)^2 / f_\zeta dx = \frac{1}{\zeta(1-\zeta)} \left(1 - \int f_1 f_2 / f_\zeta dx \right). \end{aligned}$$

(b) Let $S(\zeta) = \int f_1 f_2 / f_\zeta dx$. For

$$f_i(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left\{ -\frac{(x-\mu_i)^2}{2\sigma^2} \right\} \quad (i=1,2; \mu_1 < \mu_2)$$

show that

$$\begin{aligned} S(\zeta) &= \sum_{m=0}^{\infty} (-1)^m \exp \left\{ \frac{m(m+1)d^2}{2} \right\} \left\{ \frac{1}{\zeta} \left(\frac{1-\zeta}{\zeta} \right)^m \Phi \left(\frac{(2m+1)d}{2} + \frac{1}{d} \log \frac{\zeta}{1-\zeta} \right) \right. \\ &\quad \left. + \frac{1}{1-\zeta} \left(\frac{\zeta}{1-\zeta} \right)^m \Phi \left(\frac{(2m+1)d}{2} - \frac{1}{d} \log \frac{\zeta}{1-\zeta} \right) \right\}, \end{aligned}$$

where $\Phi(y) \equiv \int_y^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$ and $d = \frac{\mu_2 - \mu_1}{2}$. Show that for given d ,

$S(\zeta)$ is symmetric about, and has a unique minimum at $\zeta = \frac{1}{2}$.

(Hill, 1963)

(c) Show that if we take $U(\zeta) = \zeta(1-\zeta)$ in (5.16), then

$$I[\mathcal{E}, \zeta; U] = \zeta(1-\zeta) \left(1 - \int f_1 f_2 / f_\zeta dx \right) = \zeta^2(1-\zeta)^2 I(\zeta),$$

where $I(\zeta)$ is the Fisher information defined in (a).

(7) For any concave uncertainty function $U(\xi)$, prove the following statements.

(a) $E[U(\xi(X))|\xi]$ is concave in ξ .

(b) Let $e^{(n)}$ be n independent replications of a common finite experiment e . Then

$$I[e^{(n)}, \xi; U] \equiv U(\xi) - E[U(\xi(X_1, \dots, X_n))|\xi]$$

is an increasing function of $n = 1, 2, \dots$.

6. Comparison of experiments.

Let $\mathcal{E}_X = [(\mathcal{X}, B), \{p(x|\theta) | \theta \in \Theta\}]$ and $\mathcal{E}_Y = [(\mathcal{Y}, C), \{p(y|\theta) | \theta \in \Theta\}]$ be two given experiments with a common input space Θ . If

$$I(\mathcal{E}_X, p(\theta)) \geq I(\mathcal{E}_Y, p(\theta)) \quad \text{for all } p(\theta),$$

then we say, following Lindley (1956), that \mathcal{E}_X is not less informative than \mathcal{E}_Y and write this symbolically as $\mathcal{E}_X \geq \mathcal{E}_Y$. The relation $\mathcal{E}_X \geq \mathcal{E}_Y$ defines a partial order; that is, two experiments will not generally be comparable, but the relation is transitive. The more-informative-than relation can be defined in the familiar way and will be denoted by

$$\mathcal{E}_X > \mathcal{E}_Y.$$

Example 6.1 Let $0 < \zeta < 1$ and k, m and $\zeta k + (1-\zeta)m$ be three positive integers. Then by the corollary to theorem 5.1

$$\mathcal{E}^{(\zeta k + (1-\zeta)m)} \geq \zeta \mathcal{E}^{(k)} + (1-\zeta) \mathcal{E}^{(m)}.$$

Thus the weighted average of the two experiments each consisting of a fixed number of independent repetitions is not more informative than one experiment which consists of an average number of independent repetitions of the two experiments.

Example 6.2 Blackwell's 2 x 2 Table (Blackwell and Girshik, 1954).

Consider a large population in which each individual has or has not each of two characteristics H, S and in which the proportions h, s of individuals with characteristics H, S are known. What is not known is the proportion w of individuals having both characteristics; we take $\theta = w/h$ as an unknown parameter. We assume without loss of generality, that $0 \leq h \leq s \leq 1-s \leq 1-h \leq 1$. The statistician might consider four

experiments $\mathcal{E}(H)$, $\mathcal{E}(\bar{H})$, $\mathcal{E}(S)$, and $\mathcal{E}(\bar{S})$. The performance of the experiment $\mathcal{E}(H)$, for example, is to observe 1000 individuals with characteristic H and count the number of individuals possessing characteristic S . All the four experiments are binomial experiments with $\Theta = [0,1]$ and $p(x|\theta)$ is a binomial density with parameters $\zeta a + (1-\zeta)\theta$, where $0 \leq \zeta \leq 1$ and a is a constant proportion independent of θ but varies for each experiment.

	S	\bar{S}	Total
H	w		h
\bar{H}			1-h
Total	s	1-s	

Let

$$p_1(x|\theta) = a^x(1-a)^{1-x}$$

$$p_2(x|\theta) = \theta^x(1-\theta)^{1-x}$$

and let

$$\mathcal{E}_i = [(\mathcal{X}, B), \{p_i(x|\theta) \mid 0 \leq \theta \leq 1\}] \quad (i = 1, 2).$$

If $\mathcal{E} = [(\mathcal{X}, B), \{p(x|\theta) \mid 0 \leq \theta \leq 1\}]$ is a binomial experiment with parameters $\zeta a + (1-\zeta)\theta$ then $\mathcal{E} = \zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2$ and by theorem 5.2

$$\begin{aligned} I(\mathcal{E}) &= I(\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2) \leq \zeta I(\mathcal{E}_1) + (1-\zeta) I(\mathcal{E}_2) \\ &= (1-\zeta) I(\mathcal{E}_2) \leq I(\mathcal{E}_2) \end{aligned}$$

because $I(\mathcal{E}_1) = 0$,

since $\zeta = 0$ for $\mathcal{E}(H)$, $I(\mathcal{E}(H))$ is greatest among those of the other three experiments. Thus we find that the experiment associated with the rarest characteristic is most informative.

As stated in the previous section the amount of information provided by an experiment has been defined by the Shannon information transmission through the channel:

$$T(\theta; x) = H(\theta) - H_x(\theta) = \int p(\theta) I(p(x|\theta):p(x)|p(\theta)) d\theta.$$

Thus we might be able to introduce the concept of strongly more informativeness as follows: Let $\mathcal{E}_i = [(\mathcal{X}_i, B_i), \{p_i(x_i|\theta) | \theta \in \Theta\}]$ ($i=1,2$) be two given experiments with a common input space Θ . \mathcal{E}_1 is said to be strongly not-less informative than \mathcal{E}_2 when

$$\int p_1(x|\theta_0) \log \frac{p_1(x|\theta_0)}{\int p(\theta) p_1(x|\theta) d\theta} dx \equiv I(p_1(x|\theta_0):p_1(x)|p(\theta))$$

$$\geq I(p_2(x|\theta_0):p_2(x)|p(\theta)) \equiv \int p_2(x|\theta_0) \log \frac{p_2(x|\theta_0)}{\int p(\theta) p_2(x|\theta) d\theta} dx,$$

for all $\theta_0 \in \Theta$ and $p(\theta)$.

This relation will be written symbolically as $\mathcal{E}_1 \stackrel{(s)}{\geq} \mathcal{E}_2$.

Sufficiency of experiments.

Following Blackwell (1953), \mathcal{E}_1 is said to be sufficient for \mathcal{E}_2 with respect to Θ and denoted symbolically by $\mathcal{E}_1 \stackrel{(s)}{\geq} \mathcal{E}_2(\Theta)$ (or simply by $\mathcal{E}_1 \stackrel{(s)}{\geq} \mathcal{E}_2$) when there exists a stochastic transformation of X_1 (given by a set of distribution functions $\{G(z|x_1) | -\infty < x_1 < \infty\}$) to a random variable Z such that, for each $\theta \in \Theta$, $Z(X_1)$ and X_2 have identical distributions. Roughly speaking, that \mathcal{E}_1 is sufficient for \mathcal{E}_2 says that a random variable with the same distributions as X_2 can be generated from X_1 and an auxiliary randomization.

Let $\mathcal{E}_i = [\Theta, \{p_i(x_i|\theta) | \theta \in \Theta\}, \mathcal{X}_i]$ ($i=1,2$) be two given experiments. By our simple notation used in the previous section we can state:

\mathcal{E}_1 is sufficient for \mathcal{E}_2 when there exists

$p(x_2|x_1) = \text{indep. of } \theta$; $p_2(x_2|\theta) \equiv \int_{\mathcal{X}_1} p(x_2|x_1)p_1(x_1|\theta)dx_1$, for all θ and x_2 .

Theorem 6.1 If $\mathcal{E}_1 \subset \mathcal{E}_2$ then $\mathcal{E}_1 \stackrel{(s)}{\geq} \mathcal{E}_2$.

Proof. For any $\theta_0 \in \Theta$ and $p(\theta)$ we have

$$\begin{aligned} I(p_2(x|\theta_0):p_2(x)|p(\theta)) &= \int_{\mathcal{X}_2} p_2(x_2|\theta_0) \log \frac{p_2(x_2|\theta_0)}{\int p(\theta)p_2(x_2|\theta)d\theta} dx_2 \\ &= \int dx_2 \left[\left(\int p(x_2|x_1)p_1(x_1|\theta_0)dx_1 \right) \log \frac{\int p(x_2|x_1)p_1(x_1|\theta_0)dx_1}{\int p(x_2|x_1)p_1(x_1)dx_1} \right] \\ &\leq \int dx_2 \left[\int p(x_2|x_1)dx_1 \left\{ p_1(x_1|\theta_0) \log \frac{p_1(x_1|\theta_0)}{p_1(x_1)} \right\} \right] \\ &= \int p_1(x_1|\theta_0) \log \frac{p_1(x_1|\theta_0)}{p_1(x_1)} dx_1 \\ &= I(p_1(x_1|\theta_0):p_1(x)|p(\theta)). \end{aligned}$$

When the two experiments are both finite experiments with a common input space, we have a similar theorem to the above for any concave uncertainty functions. Let

$$\mathcal{E}_X = [(\mathcal{X}, B), \{f_1(x), \dots, f_k(x)\}]$$

$$\mathcal{E}_Y = [(\mathcal{Y}, C), \{g_1(y), \dots, g_k(y)\}]$$

be two finite experiments with a common input space $\Theta = \{1, \dots, k\}$.

Theorem 6.2 If $\mathcal{E}_X \subset \mathcal{E}_Y$, then for any concave uncertainty function

$U(\xi)$ defined on Ξ

$$I[\mathcal{E}_{X,\xi;U}] \geq I[\mathcal{E}_{Y,\xi;U}], \quad \text{for all } \xi \in \Xi.$$

(DeGroot, 1962)

In order to prove this theorem we shall prove the following lemma. Let Ξ be as before the set of all probability- k vectors.

Lemma For a fixed $\xi \in \Xi$ we define, on the set

$A \equiv \left\{ (a_1, \dots, a_k) \mid a_1, \dots, a_k \geq 0 \right\}$, a non-negative real-valued function

$$W(a) = (a \cdot \xi) U(a \otimes \xi),$$

where $a \cdot \xi$ is the scalar product of the two vectors, and $a \otimes \xi$ is defined by

$$a \otimes \xi = \begin{cases} \left(\frac{a_1 \xi_1}{a \cdot \xi}, \dots, \frac{a_k \xi_k}{a \cdot \xi} \right), & \text{if } a \cdot \xi \neq 0 \\ (1, 0, \dots, 0), & \text{if } a \cdot \xi = 0. \end{cases}$$

Then if $U(\xi)$ is concave on Ξ , then $W(a)$ is concave on A .

Proof. Take any $a, b \in A$ and any constants α and β with $0 < \alpha = 1 - \beta < 1$.

Then since

$$\begin{aligned} \xi \otimes (\alpha a + \beta b) &= \left(\frac{\alpha a_1 \xi_1 + \beta b_1 \xi_1}{\alpha(a \cdot \xi) + \beta(b \cdot \xi)}, \dots \right) \quad (\text{if denominator } \neq 0) \\ &= \frac{\alpha(a \cdot \xi)}{\alpha(a \cdot \xi) + \beta(b \cdot \xi)} a \otimes \xi + \frac{\beta(b \cdot \xi)}{\alpha(a \cdot \xi) + \beta(b \cdot \xi)} b \otimes \xi \end{aligned}$$

it follows that from the concavity of U

$$\begin{aligned} W(\alpha a + \beta b) &= (\alpha(a \cdot \xi) + \beta(b \cdot \xi)) U(\xi \otimes (\alpha a + \beta b)) \\ &\geq \alpha(a \cdot \xi) U(a \otimes \xi) + \beta(b \cdot \xi) U(b \otimes \xi) \\ &= \alpha W(a) + \beta W(b). \end{aligned}$$

Proof of Theorem 6.2 It suffices to show that $E[U(\xi(X))|\xi] \leq E[U(\xi(Y))|\xi]$ for all $\xi \in \Xi$. Since $\mathcal{E}_X \prec \mathcal{E}_Y$, there exists a p.d.f. $p(y|x)$ independent of i , such that

$$g_i(y) = \int p(y|x) f_i(x) d\mu(x), \quad \text{a.e.}(v), \quad (i=1, \dots, k).$$

Thus we have for any $\xi \in \Xi$

$$\begin{aligned} E[U(\xi(Y))|\xi] &= \int (\xi \cdot \vec{g}(y)) U\left(\frac{\xi_1 g_1(y)}{\xi \cdot \vec{g}(y)}, \dots, \frac{\xi_k g_k(y)}{\xi \cdot \vec{g}(y)}\right) d\nu(y) \\ &= \int W(\vec{g}(y)) d\nu(y) \\ &= \int W(\int p(y|x) \vec{f}(x) d\mu(x)) d\nu(y) \\ &= \int W\left(\frac{\int p(y|x) \vec{f}(x) d\mu(x)}{\int p(y|x) d\mu(x)}\right) \left(\int p(y|x) d\mu(x)\right) d\nu(y) \\ &\quad \text{(because } W(\alpha a) = \alpha W(a)) \\ &\geq \int d\nu(y) \int W(\vec{f}(x)) p(y|x) d\mu(x) \quad \text{(by the Lemma)} \\ &= \int W(\vec{f}(x)) d\mu(x) \\ &= E[U(\xi(X))|\xi]. \end{aligned}$$

Example 6.3 Let $\mathcal{E}(\sigma) = [(-\infty, \infty), \{p(x|\theta) | -\infty < \theta < \infty\}, (-\infty, \infty)]$ be a normal experiment, where $p(x|\theta)$ is a normal density with mean θ and variance σ^2 . The value of σ^2 is assumed to be known. If $\sigma_1 < \sigma_2$ then $\mathcal{E}(\sigma_1) \stackrel{(s)}{>} \mathcal{E}(\sigma_2)$. The proof is as follows: Let $x_j (j=1,2)$ be the output variable in the experiment $\mathcal{E}(\sigma_j)$. If u is a random variable independent of x_1 and has a normal density with mean 0 and variance $\sigma_2^2 - \sigma_1^2$, then $Z = x_1 + u$ has identical distribution with x_2 . Thus $\mathcal{E}(\sigma_1) \prec \mathcal{E}(\sigma_2)$. We have $\mathcal{E}(\sigma_1) \stackrel{(s)}{\geq} \mathcal{E}(\sigma_2)$ by theorem 6.1, and moreover $\mathcal{E}(\sigma_1) \stackrel{(s)}{>} \mathcal{E}(\sigma_2)$ from example 5.1.

Example 6.4 If in an experiment the input space and output space are both finite, the experiment will be characterized by a stochastic matrix

$$P = (p_{ij} | i=1, \dots, k; j=1, \dots, N_1) ; \text{ all } p_{ij} \geq 0, \sum_{j=1}^{N_1} p_{ij} \equiv 1.$$

Let

$$Q = (q_{ij} | i=1, \dots, k; j=1, \dots, N_2) ; \text{ all } q_{ij} \geq 0, \sum_{j=1}^{N_2} q_{ij} \equiv 1,$$

be another stochastic matrix with the same k , then the experiment with P is sufficient for the experiment with Q (simply written by $P \xi Q$), if and only if there exists

$$(6.1) \text{ stoch. } M \begin{matrix} N_1 \\ \times \\ N_2 \end{matrix} ; \quad PM = Q.$$

Now let the experiments be binomial dichotomies:

$$P = \begin{pmatrix} a_1 & 1-a_1 \\ a_2 & 1-a_2 \end{pmatrix}, \quad Q = \begin{pmatrix} b_1 & 1-b_1 \\ b_2 & 1-b_2 \end{pmatrix}$$

with $a_1 < a_2$. Then P^{-1} exists and

$$P^{-1}Q = \frac{1}{a_1 - a_2} \begin{pmatrix} 1-a_2 & -(1-a_1) \\ -a_2 & a_1 \end{pmatrix} \begin{pmatrix} b_1 & 1-b_1 \\ b_2 & 1-b_2 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{b_2(1-a_1) - b_1(1-a_2)}{a_2 - a_1} & \frac{(1-a_1)(1-b_2) - (1-a_2)(1-b_1)}{a_2 - a_1} \\ \frac{a_2 b_1 - a_1 b_2}{a_2 - a_1} & \frac{a_2(1-b_1) - a_1(1-b_2)}{a_2 - a_1} \end{pmatrix}$$

has row-sums 1. Non-negativity of the four elements of this matrix yields

$$(6.2) \quad \frac{1-a_2}{1-a_1} \leq \left\{ \begin{array}{l} \frac{1-b_2}{1-b_1} \leq \frac{b_2}{b_1} \\ \frac{b_2}{b_1} \leq \frac{1-b_2}{1-b_1} \end{array} \right\} \leq \frac{a_2}{a_1},$$

which is a necessary and sufficient condition that $P \subseteq Q$ in this case. For fixed parameters $a_1 < a_2$, the shaded region and the dotted region in Fig. 6.1 show the sets $\{(b_1, b_2) | P \subseteq Q\}$ and $\{(b_1, b_2) | P \rightarrow Q\}$, respectively. The experiment corresponding to (c_1, c_2) , in the figure, for example, is not comparable with the experiment with P .

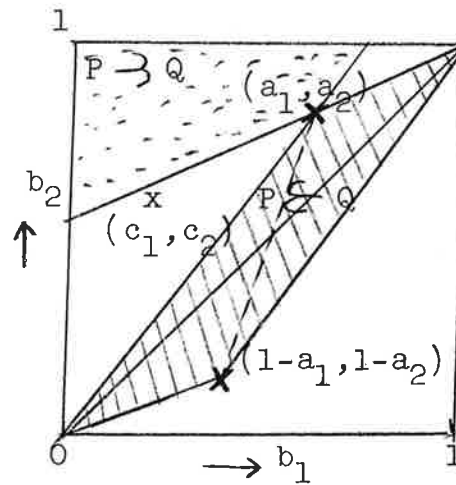


Figure 6.1

Comparison of experiments characterized by stochastic matrices

We shall discuss a simpler case where the input space Θ and the output space \mathfrak{X} are both finite, so that the experiment \mathcal{E} is characterized by a stochastic matrix

$$P = (p_{ij} | i=1, \dots, k; j=1, \dots, N).$$

Let A be a bounded, closed and convex subset of Euclidean k space.

A fixed sample-size decision problem with finite states of nature and finite sample space is specified by a pair (P, A) : a decision function (d.f) is represented by a $N \times k$ matrix

$$D = \begin{bmatrix} l_1(1), \dots, & l_k(1) \\ l_1(2), \dots, & l_k(2) \\ & \dots \\ l_1(N), \dots, & l_k(N) \end{bmatrix},$$

where each row-vector $(l_1(j), \dots, l_k(j))$ is a point of A , and represents the loss vector when the sample point j is observed and the d.f. D used. Since $(PD)_{ii} = \sum_{j=1}^N p_{ij} l_i(j)$ is the expected loss of D when the true state of nature is i , the risk vector of D is

$$(6.3) \quad \text{diag. } (PD) \equiv ((PD)_{11}, \dots, (PD)_{kk}).$$

Let

$$B(P,A) = \left\{ \text{risk vector of } D \mid \text{all possible } D \right\}.$$

It may be shown that in any decision problem (P,A) , the set $B(P,A)$ is a bounded closed and convex set containing A .

Let P, Q be any stochastic matrices with the same k . We say that P is more informative than Q , written $P \supset Q$, if

$$B(P,A) \supset B(Q,A), \text{ for all bounded, closed and convex } A.$$

Theorem 6.3 Let

$$P = (p_{ij} \mid i=1, \dots, k; j=1, \dots, N_1),$$

$$Q = (q_{ij} \mid i=1, \dots, k; j=1, \dots, N_2)$$

be two stochastic matrices with the same k . Each of the following five conditions is equivalent to $P \supset Q$.

$$(1) \quad \begin{pmatrix} \forall & C \\ N_2 & \times k \end{pmatrix} \exists \text{ stoch. } \begin{matrix} M \\ N_1 & \times & N_2 \end{matrix} ; \quad \text{diag. } (PMC) = \text{diag. } (QC).$$

$$(2) \exists \text{ stoch. } \begin{matrix} M \\ N_1 \times N_2 \end{matrix} ; \quad PM = Q.$$

$$(3) (\forall C) \begin{matrix} N_2 \times k \\ N_1 \end{matrix} \exists \text{ stoch. } \begin{matrix} M \\ N_1 \times N_2 \end{matrix} ; \quad \text{Trace (PMC)} \leq \text{Trace (QC)}.$$

$$(4) \sum_{j=1}^{N_2} (\sum_i p_{ij}) \varphi \left(\frac{p_{1j}}{\sum_i p_{ij}}, \dots, \frac{p_{kj}}{\sum_i p_{ij}} \right) \geq \sum_{j=1}^{N_2} (\sum_i q_{ij}) \varphi \left(\frac{q_{1j}}{\sum_i q_{ij}}, \dots, \frac{q_{kj}}{\sum_i q_{ij}} \right),$$

for all continuous, convex $\varphi(\xi)$ on Ξ .

$$(5) \exists \text{ stoch. } \begin{matrix} T \\ N_2 \times N_1 \end{matrix} ;$$

$$\begin{matrix} T \\ N_2 \times N_1 \end{matrix} \cdot \begin{bmatrix} p_{11} & \dots & p_{k1} \\ \frac{p_{11}}{\sum_i p_{i1}} & \dots & \frac{p_{k1}}{\sum_i p_{i1}} \\ \dots & \dots & \dots \\ p_{1N_1} & \dots & p_{kN_1} \\ \frac{p_{1N_1}}{\sum_i p_{iN_1}} & \dots & \frac{p_{kN_1}}{\sum_i p_{iN_1}} \end{bmatrix} = \begin{bmatrix} q_{11} & \dots & q_{k1} \\ \frac{q_{11}}{\sum_i q_{i1}} & \dots & \frac{q_{k1}}{\sum_i q_{i1}} \\ \dots & \dots & \dots \\ q_{1N_2} & \dots & q_{kN_2} \\ \frac{q_{1N_2}}{\sum_i q_{iN_2}} & \dots & \frac{q_{kN_2}}{\sum_i q_{iN_2}} \end{bmatrix},$$

and

$$\left(\sum_i q_{i1}, \dots, \sum_i q_{iN_2} \right) \cdot \begin{matrix} T \\ N_2 \times N_1 \end{matrix} = \left(\sum_i p_{i1}, \dots, \sum_i p_{iN_1} \right).$$

We note the following three facts.

(a) Condition (2) says that $P \prec Q$. Hence the theorem asserts that $P \succ Q$ if and only if $P \prec Q$.

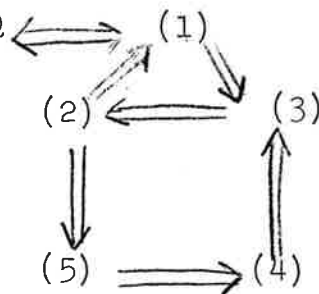
(b) Condition (3) also has a simple interpretation; for trace (QC) is n times the expected loss of C when the states of nature are equiprobable in decision problem (Q, A) . Thus $P \succ Q$ if and only if for any A the Bayes risk when the states of nature are equiprobable in (P, A) is less than or equal to the corresponding Bayes risk in (Q, A) .

(c) If we write $\xi^0 = (\frac{1}{k}, \dots, \frac{1}{k})$, $U = -\varphi$, and define $I [P, \xi^0; U] \equiv U(\xi^0) - E [U(\xi^0(X)) | \xi^0]$ as in (5.16), then condition (4) says that $P \supset Q$ if and only if

$I [P, \xi^0; U] \geq I [Q, \xi^0; U]$, for all continuous, concave uncertainty functions U .

Proof of theorem 6.3.

We proceed as in the diagram. $P \supset Q$



$(2) \Rightarrow (1) \Rightarrow (3)$ is evident. $P \supset Q \Rightarrow (1)$:

Let A be the convex hull determined by the rows of C . Then C is a possible d.f. in (Q, A) . Since any d.f. D in (P, A) is an $N_1 \times k$ matrix, each row of which is a convex linear combination of the rows of C ,

$$\exists \text{ stoch. } \begin{matrix} M \\ N_1 \times N_2 \end{matrix} ; \begin{matrix} D \\ N_1 \times k \end{matrix} = \begin{matrix} M \\ N_2 \times k \end{matrix} C$$

Choosing D with $\text{diag}(PD) = \text{diag}(QC)$ yields on M satisfying (1).

$(1) \rightarrow P \supset Q$:

Let A be any bounded convex and closed set in k space, and let C be any d.f. in (Q, A) . MC is a d.f. in (P, A) and from (1)

$$\exists \text{ stoch. } \begin{matrix} M \\ N_1 \times N_2 \end{matrix} ; (\text{risk vector of } C \text{ in } (Q, A)) = (\text{risk vector of } MC \text{ in } (P, A)) \in B(P, A).$$

Hence we have $B(Q, A) \subset B(P, A)$.

(2) \Rightarrow (5):

From (2) we have

$$\sum_{j=1}^{N_1} p_{ij} m_{j\ell} = q_{i\ell} \quad (i=1, \dots, k; \ell=1, \dots, N_2).$$

$$\therefore \frac{q_{i\ell}}{\sum_{i=1}^k q_{i\ell}} = \frac{\sum_j p_{ij} m_{j\ell}}{\sum_i q_{i\ell}} = \sum_{j=1}^{N_1} \frac{p_{ij}}{\sum_i p_{ij}} \cdot \left\{ \frac{\sum_i p_{ij} m_{j\ell}}{\sum_i p_{i\ell}} \right\}.$$

Let $\left\{ \right\} = t'_{j\ell} = t_{\ell j}$, then $(t_{\ell j})$ is stochastic, and satisfies (5).

For,

$$\begin{pmatrix} \frac{p_{11}}{\sum_i p_{i1}} & \frac{p_{12}}{\sum_i p_{i2}} & \dots & \frac{p_{1N_1}}{\sum_i p_{iN_1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{k1}}{\sum_i p_{i1}} & \frac{p_{k2}}{\sum_i p_{i2}} & \dots & \frac{p_{kN_1}}{\sum_i p_{iN_1}} \end{pmatrix} \cdot \begin{matrix} T' \\ N_1 \times N_2 \end{matrix} = \begin{pmatrix} \frac{q_{11}}{\sum_i q_{i1}} & \frac{q_{12}}{\sum_i q_{i2}} & \dots & \frac{q_{1N_2}}{\sum_i q_{iN_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{q_{k1}}{\sum_i q_{i1}} & \frac{q_{k2}}{\sum_i q_{i2}} & \dots & \frac{q_{kN_2}}{\sum_i q_{iN_2}} \end{pmatrix},$$

$$\sum_{\ell=1}^{N_2} (\sum_i q_{i\ell}) t_{\ell j} = \sum_{\ell=1}^{N_2} (\sum_i q_{i\ell}) \left\{ \frac{\sum_j p_{ij} m_{j\ell}}{\sum_i q_{i\ell}} \right\} = \sum_i p_{ij} \quad (j=1, \dots, N_1).$$

(5) \Rightarrow (4) :

$$\begin{aligned} & \sum_{j=1}^{N_1} (\sum_i p_{ij}) \varphi \left(\frac{p_{1j}}{\sum_i p_{ij}}, \dots, \frac{p_{kj}}{\sum_i p_{ij}} \right) = \sum_{\ell=1}^{N_2} (\sum_i q_{i\ell}) \sum_{j=1}^{N_1} t_{1j} \varphi \left(\frac{p_{1j}}{\sum_i p_{ij}}, \dots \right) \\ & \geq \sum_{\ell=1}^{N_2} (\sum_i q_{i\ell}) \varphi \left(\sum_j \left(\frac{p_{1j}}{\sum_i p_{ij}} \right) t_{\ell j}, \dots \right) \\ & = \sum_{\ell=1}^{N_2} (\sum_i q_{i\ell}) \varphi \left(\frac{q_{1\ell}}{\sum_i q_{i\ell}}, \dots, \frac{q_{k\ell}}{\sum_i q_{i\ell}} \right). \end{aligned}$$

(4) \Rightarrow (3):

Let C be any $N_2 \times k$ matrix, and let $\varphi(\xi) = - \min_{1 \leq \ell \leq N_2} \sum_{i=1}^k c_{\ell i} \xi_i$.

Then we can show that

$$\min_{\text{stoch. } V} \text{trace (QVC)}_{N_2 \times N_1} = - \sum_{j=1}^{N_2} (\sum_i q_{ij}) \varphi \left(\frac{q_{1j}}{\sum_i q_{ij}}, \dots, \frac{q_{kj}}{\sum_i q_{ij}} \right).$$

In fact, for any stochastic V ,
 $N_2 \times N_1$

$$\begin{aligned} \text{trace (QVC)} &= \sum_{i=1}^k \sum_{j=1}^{N_2} \sum_{\ell=1}^{N_2} q_{ij} v_{j\ell} c_{\ell i} = \sum_{j=1}^{N_2} (\sum_i q_{ij}) \sum_{\ell=1}^{N_2} v_{j\ell} \sum_{i=1}^k c_{\ell i} \left(\frac{q_{ij}}{\sum_i q_{ij}} \right) \\ &\geq - \sum_{j=1}^{N_2} (\sum_i q_{ij}) \sum_{\ell=1}^{N_2} v_{j\ell} \varphi \left(\frac{q_{1j}}{\sum_i q_{ij}}, \dots, \frac{q_{kj}}{\sum_i q_{ij}} \right) \\ &= - \sum_{j=1}^{N_2} (\sum_i q_{ij}) \varphi \left(\frac{q_{1j}}{\sum_i q_{ij}}, \dots \right) \end{aligned}$$

and it is clear the equality holds for some appropriate V .

Similarly we can show

$$\min_{\text{stoch. } M} \text{trace (PMC)}_{N_1 \times N_2} = - \sum_{j=1}^{N_1} (\sum_i p_{ij}) \varphi \left(\frac{p_{1j}}{\sum_i p_{ij}}, \dots, \frac{p_{kj}}{\sum_i p_{ij}} \right),$$

so that by (4)

$$\min_M \text{trace (PMC)} \leq \min_V \text{trace (QVC)} \leq \text{trace (QC)}.$$

(3) \Rightarrow (2):

Consider the 0-sum 2-person game in which

I chooses any matrix C with all $0 \leq c_{ij} \leq 1$,
 $N_2 \times k$

II chooses any stoch. $\begin{matrix} M \\ N_1 \times N_2 \end{matrix}$

with payoff trace $\{(PM-Q)C\}$.

Since the pure-strategy spaces in this game are both bounded, convex and closed, there exists a value v_0 and the optimal strategies C_0, M_0 ; i.e.

$$\text{trace} \{(PM_0-Q)C\} \leq v_0 \leq \text{trace} \{(PM-Q)C_0\}, \quad \text{for all } C, M.$$

Hence from (3), $v_0 \leq 0$, so that $\text{trace} \{(PM_0-Q)C\} \leq 0$ for all C with

$0 \leq c_{ij} \leq 1$. With $U = PM_0 - Q$,

$$\sum_{i=1}^k \sum_{j=1}^{N_2} u_{ij} c_{ji} \leq 0 \quad \therefore \text{all } u_{ij} \leq 0.$$

But $\sum_{j=1}^{N_2} u_{ij} \equiv 0$, hence $u_{ij} \equiv 0$ follows.

Comparison of dichotomous experiments.

Let

$$(6.4) \quad \begin{aligned} \mathcal{E}_X &= [(\mathcal{X}, B), \{f_1(x), f_2(x)\}] \\ \mathcal{E}_Y &= [(\mathcal{Y}, C), \{g_1(y), g_2(y)\}] \end{aligned}$$

be two dichotomous experiments with a common input space $\Theta = \{1, 2\}$. It is required to decide which of the two input symbols is transmitted, or equivalently which of the two hypotheses is true, on the basis of one observation, either of X or of Y , with the usual zero-or-one loss.

As an example of a situation in which this type of question may arise, consider the problem of deciding between utilizing a use test as against a specifications test for acceptance of a lot of manufactured items. A large lot of items has been produced and a decision is to be

made between, say, p_1 and p_2 as being the proportion of defectives in the lot. Let $X = 1$ or 0 according as an item selected at random is defective or not when we subject it to a use test. Let $Y = 1$ or 0 according as an item selected at random is defective (i.e., it fails to meet certain specifications) or not, when we subject it to a specifications test. If the conditional probabilities $\alpha = \Pr \{ Y = 1 | \text{non-defective} \}$ and $\beta = \Pr \{ Y = 0 | \text{defective} \}$ are known, then our decision problem is described by the diagram Fig. 6.2, where $f_i(x)$ ($i=1,2$) and $g_i(y)$ ($i=1,2$) are all binomial densities with known parameters p_i ($i=1,2$) and $p_i(1-\beta) + (1-p_i)\alpha$ ($i=1,2$), respectively,

Exp Hyp.	\mathcal{E}_X	\mathcal{E}_Y
$(\zeta) H_1$	$f_1(x)$	$g_1(y)$
$(1-\zeta) H_2$	$f_2(x)$	$g_2(y)$

Figure 6.2

Now let $R_X(\zeta)$ and $R_Y(\zeta)$ be the Bayes risks with respect to the a priori distribution $(\zeta, 1-\zeta)$ when performing experiment \mathcal{E}_X and \mathcal{E}_Y , respectively. Hence for \mathcal{E}_X

$$(6.5) \quad R_X(\zeta) = \zeta \int \frac{f_2(x)}{f_1(x)} f_1(x) d\lambda + (1-\zeta) \int f_2(x) d\lambda.$$

$$\frac{f_2(x)}{f_1(x)} > \frac{\zeta}{1-\zeta} \qquad \frac{f_2(x)}{f_1(x)} \leq \frac{\zeta}{1-\zeta}$$

It is of some interest to investigate conditions that hold true for $R_X \leq R_Y$ (i.e., $R_X(\zeta) \leq R_Y(\zeta)$ for all $0 \leq \zeta \leq 1$). Interest in the case of one observation may seem curious; but it is not so. If for one

observation $R_X \leq R_Y$, then for any number of observations, any set of actions and any loss function, consistent use of X will never yield a greater Bayes risk than Y.

Let $F_1(u)$ and $F_2(u)$ be the c.d.f.'s of $\frac{f_2(X)}{f_1(X)}$ under H_1 and H_2 , respectively, i.e., $F_1(u) = \int_0^u \frac{f_2(x)}{f_1(x)} dx$ ($i=1,2$).

Then since

$$(6.6) \begin{cases} F_2(u) = \int_0^u v dF_1(v) \\ F_1(u) = \int_0^u \frac{dF_2(v)}{v} \end{cases}$$

we can write, with $\eta = \frac{\zeta}{1-\zeta}$, that

$$(6.7) \quad \frac{R_X(\zeta) - \zeta}{1-\zeta} = \int_0^\eta (u-\eta) dF_1(u) = \int_0^\infty \min(u-\eta, 0) dF_1(u) \\ = \int_0^\eta (1 - \frac{\eta}{u}) dF_2(u) = \int_0^\infty \frac{\min(u-\eta, 0)}{u} dF_2(u).$$

With the analogous definitions of c.d.f.'s of $\frac{g_2(Y)}{g_1(Y)}$, we get the parallel expressions for the risk associated with Y.

Theorem 6.4 Let \mathcal{E}_X and \mathcal{E}_Y be two dichotomous experiments defined by (6.4). Then

(i) each of the following two conditions is equivalent to $R_X = R_Y$ (i.e., $R_X(\zeta) = R_Y(\zeta)$ for all $0 \leq \zeta \leq 1$).

(1) f_2/f_1 and g_2/g_1 have the same distributions under H_1

(2) f_2/f_1 and g_2/g_1 have the same distributions under H_2 .

(ii) Let $F_1(u)$ be the c.d.f. of $\frac{f_2(X)}{f_1(X)}$ under H_1 and let $G_1(u)$ be the c.d.f. of $\frac{g_2(Y)}{g_1(Y)}$ under H_1 , i.e.,

$$F_1(u) = \int_{\frac{f_2(x)}{f_1(x)} \leq u} f_1(x) d\lambda(x), \quad G_1(u) = \int_{\frac{g_2(y)}{g_1(y)} \leq u} g_1(y) d\mu(y).$$

Then $R_X \leq R_Y$ if and only if

$$(6.8) \quad \int_0^\eta F_1(u) du \geq \int_0^\eta G_1(u) du, \quad \text{for all } 0 < \eta < \infty.$$

(Bradt and Karlin, 1956)

Proof. (i) The sufficiency is immediate. To show the necessity, suppose $R_X = R_Y$; then by (6.7) for all $0 < \eta < \infty$,

$$(*) \quad \int_0^\infty \min(u-\eta, 0) dF_1(u) = \int_0^\infty \min(u-\eta, 0) dG_1(u).$$

This equation holds with the common integrand replaced by

$$\min(u-a, 0) - \min(u-a-\frac{1}{n}, 0), \quad (a > 0; n=1, 2, \dots).$$

Thus we have in the left-hand side of (*)

$$- \int_{a < u \leq a + \frac{1}{n}} (u-a-\frac{1}{n}) dF_1(u) + \frac{1}{n} \int_{u < a} dF_1(u) = \dots$$

$$\frac{1}{n} \left[\int_{a < u \leq a + \frac{1}{n}} (1-n(u-a)) dF_1(u) + F_1(a) \right]$$

for all n . Hence,

$$F_1(a) + \int_{a < u \leq a + \frac{1}{n}} (1-n(u-a)) dF_1(u) = G_1(a) + \int_{a < u \leq a + \frac{1}{n}} (1-n(u-a)) dG_1(u).$$

Letting $n \rightarrow \infty$, we get $F_1(a) = G_1(a)$, i.e., (1).

(ii) Integrating by parts we get

$$\begin{aligned} \frac{R_X(\zeta) - \zeta}{1-\zeta} &= \int_0^\eta (u-\eta) dF_1(u) \\ &= \left[(u-\eta)F_1(u) \right]_{u=0}^\eta - \int_0^\eta F_1(u) du = - \int_0^\eta F_1(u) du. \end{aligned}$$

Note that $1-F_1(u) = \int_{\frac{f_2(x)}{f_1(x)} > \frac{u/(1+u)}{1/(1+u)}} f_1(x) dx$ is the error probability under

H_1 of the Bayes test with respect to $\left(\frac{u}{1+u}, \frac{1}{1+u} \right)$.

The following theorem shows that the experiment with uniformly smaller Bayes risk is more informative in the Shannon-Lindley sense.

Theorem 6.5 Let \mathcal{E}_X and \mathcal{E}_Y be two dichotomous experiments defined by
 (6.4). If $R_X \leq R_Y$ then $\mathcal{E}_X \geq \mathcal{E}_Y$, i.e., for all $0 \leq \zeta \leq 1$

$$(6.9) \quad \left\{ \begin{array}{l} I_X(f_1 : f_\zeta) \geq I_Y(g_1 : g_\zeta) \\ I_X(f_2 : f_\zeta) \geq I_Y(g_2 : g_\zeta) \end{array} \right.$$

where

$$f_\zeta = \zeta f_1 + (1-\zeta) f_2, \quad g_\zeta = \zeta g_1 + (1-\zeta) g_2.$$

Proof. Let $G_1(u)$ and $G_2(u)$ be the c.d.f.'s of $\frac{g_2(Y)}{g_1(Y)}$ under H_1 and H_2 ,

respectively. Then from the analogous relation to (6.7) and the assumption of the theorem, we have

$$\int_0^{\infty} \min(u-\eta, 0) dF_1(u) \leq \int_0^{\infty} \min(u-\eta, 0) dG_1(u), \quad \text{for all } 0 \leq \eta < \infty.$$

Since any continuous concave function $\varphi(u)$ on $[0, \infty)$ is uniformly approximated by

$$l(u) + \sum_{i=1}^h a_i \min(u-\eta_i, 0),$$

where $l(u)$ is a linear function and $a_i, \eta_i > 0$ ($i=1, \dots, h$), and since $\int u dF_1(u) = \int u dG_1(u) = 1$, we have

$$\int_0^{\infty} \varphi(u) dF_1(u) \leq \int_0^{\infty} \varphi(u) dG_1(u), \quad \text{for all cont. concave } \varphi(u).$$

In particular for $\varphi(u) = \log(\zeta + (1-\zeta)u)$

$$\begin{aligned} I_X(f_1 : f_\zeta) &= E_1 \left\{ \log \frac{f_1(X)}{f_\zeta(X)} \right\} = - \int_0^{\infty} \log(\zeta + (1-\zeta)u) dF_1(u) \\ &\geq - \int_0^{\infty} \log(\zeta + (1-\zeta)u) dG_1(u) = E_1 \left\{ \log \frac{g_1(Y)}{g_\zeta(Y)} \right\} = I_Y(g_1 : g_\zeta), \end{aligned}$$

while for $\varphi(u) = -u \log \frac{u}{\zeta + (1-\zeta)u}$

$$\begin{aligned} I_X(f_2 : f_\zeta) &= E_2 \left\{ \log \frac{f_2(X)}{f_\zeta(X)} \right\} = \int_0^{\infty} u \log \frac{u}{\zeta + (1-\zeta)u} dF_1(u) \\ &\geq \int_0^{\infty} u \log \frac{u}{\zeta + (1-\zeta)u} dG_1(u) = E_2 \left\{ \log \frac{g_2(Y)}{g_\zeta(Y)} \right\} = I_Y(g_2 : g_\zeta). \end{aligned}$$

These two inequalities yield the conclusion of the theorem.

Theorem 6.6 Let \mathcal{E}_X and \mathcal{E}_Y be two dichotomous experiments defined by

(6.4). $R_X \leq R_Y$, if and only if

$I[\mathcal{E}_X, \xi^0; U] \geq I[\mathcal{E}_Y, \xi^0; U]$, for all continuous concave uncertainty functions U , where $\xi^0 = (\frac{1}{2}, \frac{1}{2})$ is the uniform prior

probability-vector (Sakaguchi, 1957).

Proof. As is seen in the proof of theorem 6.5, $R_X \leq R_Y$ is equivalent to the universal inequality (6.10). Replacing the uncertainty function U by the function $\varphi(\zeta)$ of the prior probability ζ for H_1 , we have

$$\begin{aligned} E \left[U(\xi^0(X)) | \xi^0 \right] &= \frac{1}{2} \int \varphi \left(\frac{f_1(x)}{f_1(x) + f_2(x)} \right) (f_1(x) + f_2(x)) d\lambda \\ &= \frac{1}{2} \int \varphi \left(\frac{1}{1 + f_2(x)/f_1(x)} \right) \left(1 + \frac{f_2(x)}{f_1(x)} \right) d\lambda \\ &= \frac{1}{2} \int_0^{\infty} (1+u) \varphi \left(\frac{1}{1+u} \right) dF_1(u). \end{aligned}$$

Thus, in order to prove the theorem, it suffices to show that (6.10) is equivalent to

$$(*) \int_0^{\infty} (1+u) \varphi \left(\frac{1}{1+u} \right) dF_1(u) \leq \int_0^{\infty} (1+u) \varphi \left(\frac{1}{1+u} \right) dG_1(u), \text{ for all cont. concave}$$

φ . Now, if $\varphi(u)$ is concave on $(0, \infty)$ then the same is true for $(1+u)\varphi\left(\frac{1}{1+u}\right)$. Hence, (6.10) implies (*). Suppose, conversely, that (*) is true. Then, in particular, for $\varphi(u) = \min(a-u, 0)$ ($0 < a < 1$) we have

$$(1+u)\varphi\left(\frac{1}{1+u}\right) = (1+u) \min\left(a - \frac{1}{1+u}, 0\right) = a \min\left(u - \frac{1-a}{a}, 0\right),$$

so that

$$\int_0^{\infty} \min\left(u - \frac{1-a}{a}, 0\right) dF_1(u) \leq \int_0^{\infty} \min\left(u - \frac{1-a}{a}, 0\right) dG_1(u), \text{ for all } 0 < a < 1.$$

This inequality implies (6.10), as is shown in the proof of theorem 6.5.

Example 6.5. Let \mathcal{E}_X be a binomial experiment $\mathcal{E}(a_1, a_2)$, with $0 \leq a_1 < a_2 \leq 1$. Since, under H_1 , $\frac{f_2(X)}{f_1(X)}$ assumes the values $\frac{a_2}{a_1}$ and $\frac{1-a_2}{1-a_1}$ with probabilities a_1 and $1-a_1$, respectively, we have

$$F_1(u) = \begin{cases} 0, & \text{if } 0 \leq u < \frac{1-a_2}{1-a_1}, \\ 1-a_1, & \text{if } \frac{1-a_2}{1-a_1} \leq u < \frac{a_2}{a_1}, \\ 1, & \text{if } \frac{a_2}{a_1} \leq u < \infty, \end{cases}$$

so that

$$\int_0^\eta F_1(u) du = \begin{cases} 0, & \text{if } 0 \leq \eta < \frac{1-a_2}{1-a_1}, \\ (1-a_1)\eta - (1-a_2), & \text{if } \frac{1-a_2}{1-a_1} \leq \eta < \frac{a_2}{a_1}, \\ \eta-1, & \text{if } \frac{a_2}{a_1} \leq \eta < \infty. \end{cases}$$

Thus if $\mathcal{E}(b_1, b_2)$ with observation Y is another binomial dichotomous experiment with $0 \leq b_1, b_2 \leq 1$, then $R_X \leq R_Y$ if and only if

$$\frac{1-a_2}{1-a_1} \leq \left\{ \begin{array}{l} \frac{1-b_2}{1-b_1} \leq \frac{b_2}{b_1} \\ \frac{b_2}{b_1} \leq \frac{1-b_2}{1-b_1} \end{array} \right\} \leq \frac{a_2}{a_1}$$

i.e., (6.2) again. (Fig. 6.3).

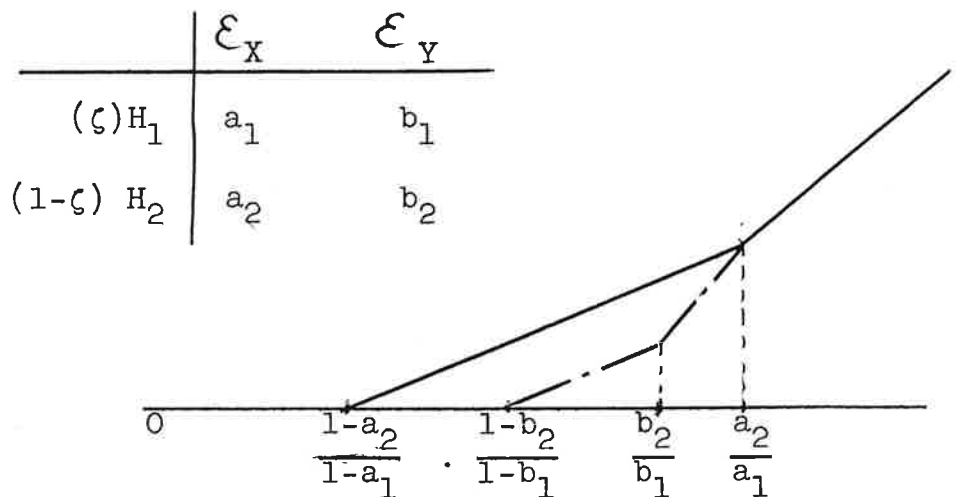


Figure 6.3

For fixed parameters $0 \leq a_1 < a_2 \leq 1$, the shaded region and the dotted region in Fig. 6.4, which is the same as the previous Fig. 6.1, show the sets $\{(b_1, b_2) | R_X \leq R_Y\}$ and $\{(b_1, b_2) | R_X \geq R_Y\}$, respectively. The experiment corresponding to (c_1, c_2) in the figure, for example, is not comparable with the experiment $\mathcal{E}(a_1, a_2)$.

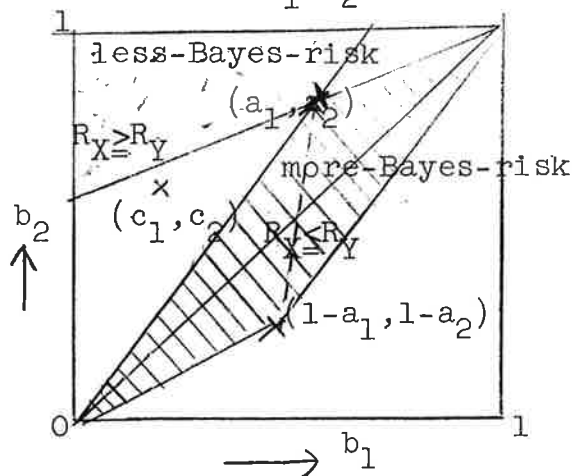


Figure 6.4

The curve connecting the points $(1,1)$ and $(0, 1 - e^{-I_X(1:2)})$, and the other curve connecting $(0,0)$ and $(e^{-I_X(2:1)}, 1)$, in Fig. 6.5, determine the two sets $\{(b_1, b_2) | I_X(i:3-i) \geq I_Y(i:3-i), i=1,2\}$ and $\{(b_1, b_2) | I_X(i:3-i) \leq I_Y(i:3-i), i=1,2\}$. These are described by the shaded region and the dotted region in Fig. 6.5, respectively.

Comparing Fig. 6.5 with Fig. 6.4, by theorem 6.5 we find the following fact: Let \mathcal{E}_X and \mathcal{E}_Y be two dichotomous experiments. If $R_X \leq R_Y$ then $I_X(i:3-i) \geq I_Y(i:3-i)$ ($i=1,2$). But the converse does not hold true.

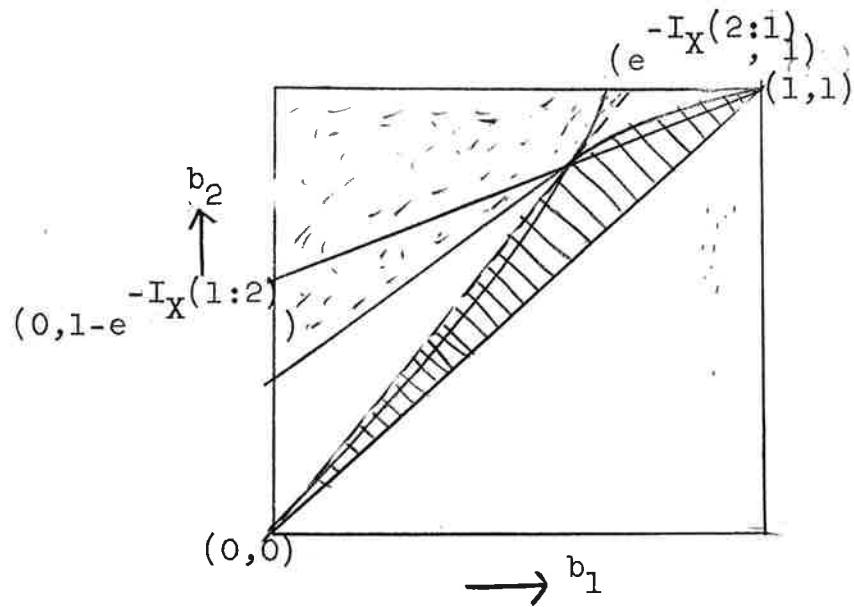


Figure 6.5

Example 6.6 (Bradt and Karlin, 1956)

We now consider the case in which both X and Y have normal distributions under each hypothesis. Let the situation be given by

		Exp.	
		\mathcal{E}_X	\mathcal{E}_Y
Hyp.	$(\zeta) H_1$	$N(0,1)$	$N(0,1)$
	$(1-\zeta) H_2$	$N(\mu, \sigma^2)$	$N(m, v)$

where $\mu, m \geq 0$ and $\sigma^2 \geq 1$.

The Kullback-Leibler informations are

$$I_X(1:2) = \frac{1}{2} \left(\log \sigma^2 - 1 + \frac{\mu^2 + 1}{\sigma^2} \right)$$

$$I_X(2:1) = \frac{1}{2} (-\log \sigma^2 - 1 + \sigma^2 + \mu^2)$$

and those for Y with the obvious substitutions. Thus we get

$$\left\{ \begin{array}{l} I_X(1:2) \geq I_Y(1:2), \text{ if and only if } m^2 \leq h_1(v) \equiv v \left(\log \sigma^2 + \frac{\mu^2 + 1}{\sigma^2} \right) - v \log v - 1, \\ I_X(2:1) \geq I_Y(2:1), \text{ if and only if } m^2 \leq h_2(v) \equiv \mu^2 + \sigma^2 - (v + \log \frac{\sigma^2}{v}). \end{array} \right.$$

We now consider the risk functions. Since $\frac{g_2(y)}{g_1(y)} > \eta$ is equivalent to

$$\left| y + \frac{m}{v-1} \right| > \frac{\sqrt{v}}{v-1} \sqrt{m^2 + (v-1) \log(v\eta^2)}, \text{ if } \eta > \frac{1}{\sqrt{v}} \exp\left\{ -m^2 / (2(v-1)) \right\} \text{ and } v > 1,$$

the Bayes risk based on Y is

$$\begin{aligned} R_Y(\zeta) &= \zeta \int_{\frac{g_2(y)}{g_1(y)} > \eta} g_1(y) d\mu + (1-\zeta) \int_{\frac{g_2(y)}{g_1(y)} \leq \eta} g_2(y) d\mu \quad (\eta = \frac{\zeta}{1-\zeta}) \\ &= \zeta \Pr \left\{ \left| Y + \frac{m}{v-1} \right| > 1 \mid H_1 \right\} + (1-\zeta) \Pr \left\{ \left| Y + \frac{m}{v-1} \right| \leq 1 \mid H_2 \right\} \end{aligned}$$

where $1 \equiv \frac{\sqrt{v}}{v-1} \sqrt{m^2 + (v-1) \log(v\eta^2)}$, and

$$\begin{aligned} \frac{R_Y(\zeta) - \zeta}{1-\zeta} &= -\eta \Pr \left\{ \left| Y + \frac{m}{v-1} \right| \leq 1 \mid H_1 \right\} + \Pr \left\{ \left| Y + \frac{m}{v-1} \right| \leq 1 \mid H_2 \right\} \\ &= \int_{-1}^1 \left\{ -\eta \Phi\left(t - \frac{m}{v-1}\right) + \frac{1}{\sqrt{v}} \Phi\left(\frac{t - \frac{mv}{v-1}}{\sqrt{v}}\right) \right\} dt. \quad (\Phi(t) \equiv \frac{e^{-t^2/2}}{\sqrt{2\pi}}). \end{aligned}$$

Comparing this expression with the obvious analogue for X it can be shown that

(i) For $v > 1$,

$$(*) \left\{ \begin{array}{l} R_X \leq R_Y, \text{ if and only if } v \leq \sigma^2 \text{ and } m^2 \leq h_3(v) \equiv (v-1) \left(\frac{\mu^2}{\sigma^2 - 1} + \log \frac{\sigma^2}{v} \right) \\ R_X \geq R_Y, \text{ if and only if } v \geq \sigma^2 \text{ and } m^2 \geq h_3(v) \end{array} \right\}$$

(ii) For $v < 1$, neither $R_X \leq R_Y$ nor $R_X \geq R_Y$ can hold.

The results obtained may be summarized in Fig. 6.6. One of the somewhat curious aspects of the results is that no matter how great $m-\mu$ may be, the X having the smaller variance under H_2 cannot yield $R_X \leq R_Y$.

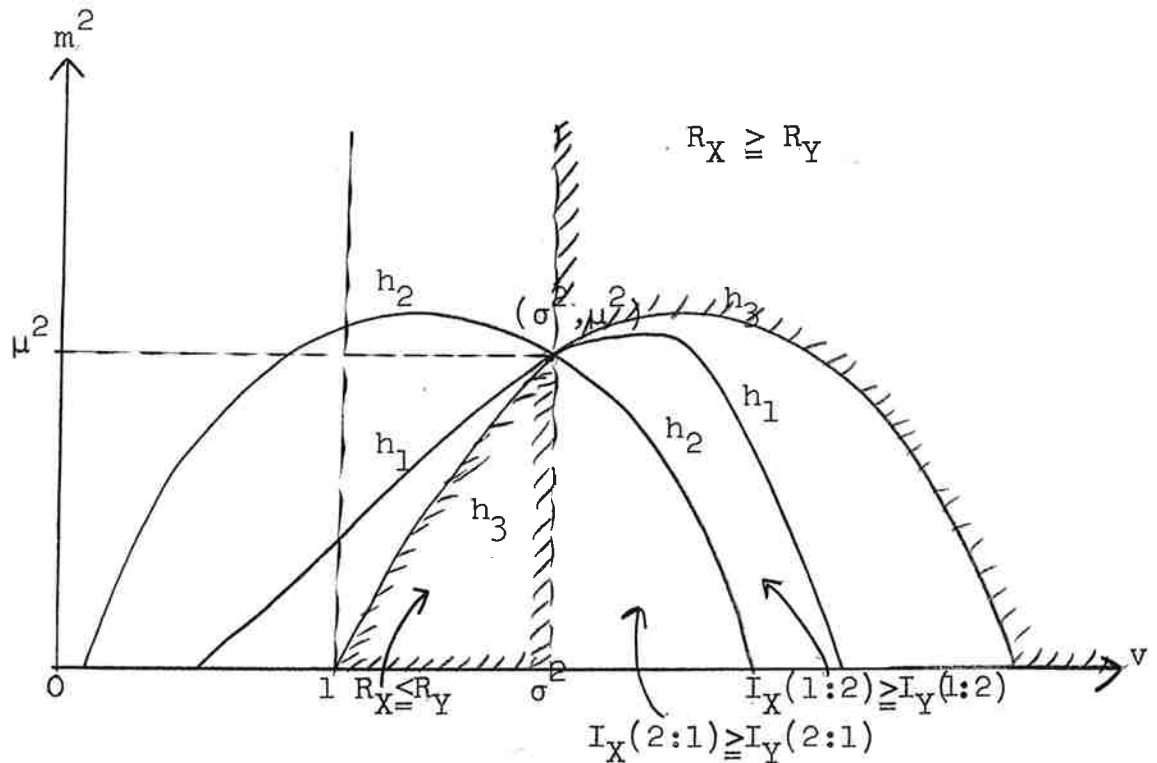


Figure 6.6

Summarizing the results of this section we present here the following diagram (Fig. 6.7) which shows the implication relations between various criteria for comparing two finite experiments.

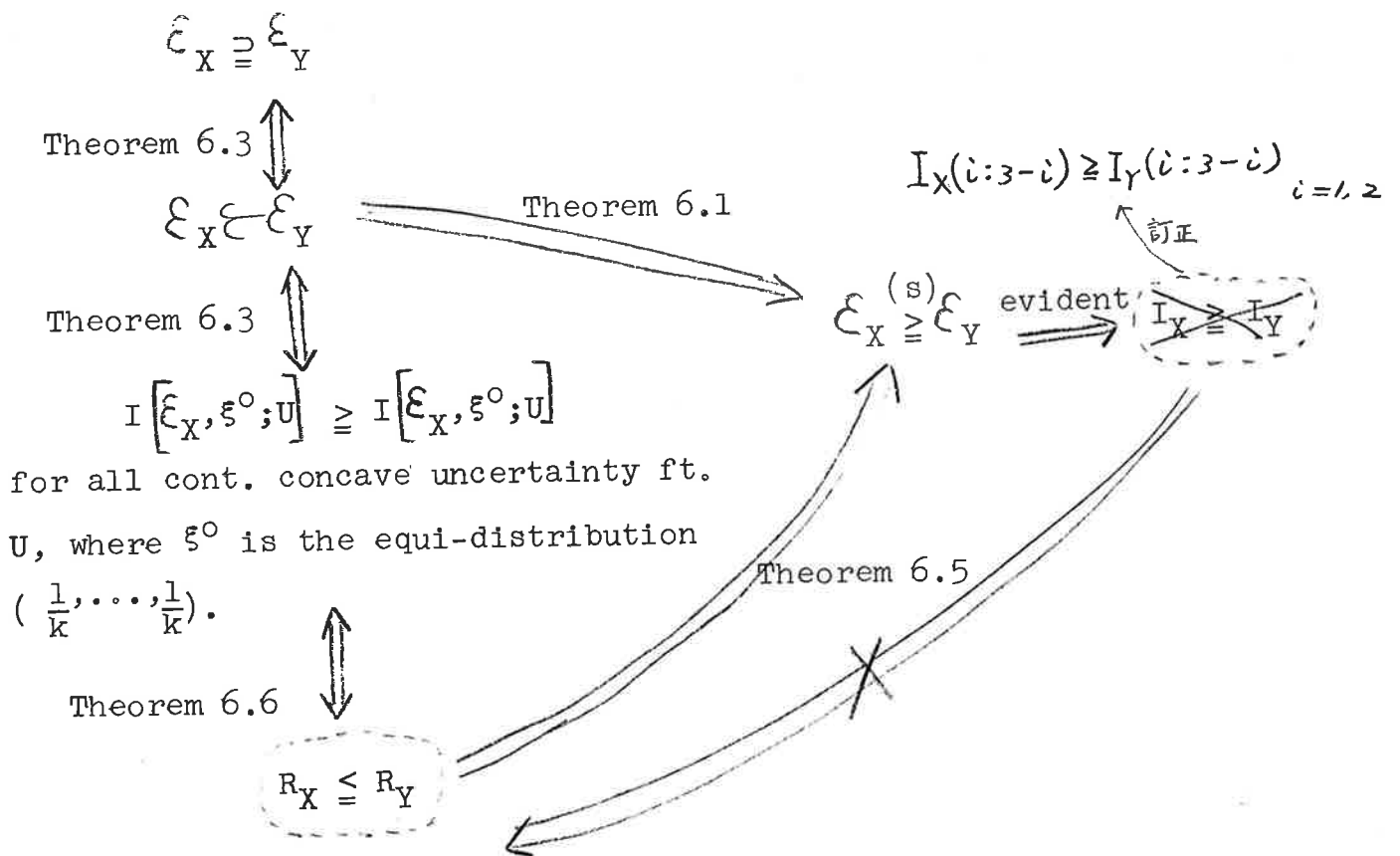


Figure 6.7

The two criteria surrounded by dotted circles can be considered only when both of the experiments are dichotomous, and the equivalence relations (upper left in the diagram) guaranteed by Theorem 6.3 hold for experiments characterized by stochastic matrices.

PROBLEMS

(1) In (*) in Example 6.6, first show that for each fixed ζ , $\frac{R_Y(\zeta) - \zeta}{1 - \zeta}$

is

(a) a non-increasing function of m for fixed $v > 1$,

(b) a non-decreasing function of $v \geq 1$ for $m^2 = h_3(v)$. Then

using these results prove the statement (*).

(2) For dichotomous Gamma experiments:

$$f_i(x) = \frac{\theta_i^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta_i x} \quad (i=1,2; \theta_1, \theta_2 > 0; x > 0)$$

$$g_i(y) = \frac{\omega_i^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\omega_i y} \quad (i=1,2; \omega_1, \omega_2 > 0; y > 0).$$

	\mathcal{E}_X	\mathcal{E}_Y
H_1	θ_1	ω_1
H_2	θ_2	ω_2

Show that $R_X \leq R_Y$ if and only if

$$\frac{\theta_2}{\theta_1} \left\{ \begin{array}{l} \leq \\ \geq \end{array} \right\} \frac{\omega_2}{\omega_1} \left\{ \begin{array}{l} \leq \\ \geq \end{array} \right\} 1.$$

(3) By a theorem in statistical decision theory the Bayes risks $R_X(\zeta)$ in dichotomous experiments are known to be continuous and concave.

(a) For the binomial experiment $\mathcal{E}(a_1, a_2)$, find and draw the function $R_X(\zeta)$ and show that the ζ which maximizes

$$\min(\zeta, 1 - \zeta) - R_X(\zeta)$$

is $\frac{1}{2}$, independently of a_1, a_2 .

(b) Do the same things for the dichotomous normal experiment:

$$H_1: N(0,1), \quad H_2: N(\mu, \sigma^2).$$

(4) Let $\mathcal{E}_i \equiv [(\mathcal{X}_i, \mathcal{B}_i), \{p_i(x|\theta) | \theta \in \Theta\}]$ ($i=1,2$) be two finite experiments with the same input space $\Theta = \{1, \dots, k\}$. Let us suppose that both $p_i(x|\theta)$ are twice-differentiable with respect to $\theta \in \Theta$ and well-behaved enough to define the Fisher information

$$I_i(\theta) \equiv E_{\theta} \left[- \frac{\partial^2}{\partial \theta^2} \log p_i(X|\theta) \right] \quad (i=1,2)$$

and to justify double differentiation with respect to θ under the integral sign with respect to x . Prove that if $\mathcal{E}_1 \geq \mathcal{E}_2$, $I_1(\theta) \geq I_2(\theta)$ for all $\theta \in \Theta$.

(Stone, 1961)

Hint: Take any $\theta_0 \in \Theta$ and $\{\theta^{(n)}\}_{n=1}^{\infty} \subset \Theta$ with $\theta^{(n)} \xrightarrow{(n \rightarrow \infty)} \theta_0$.

Let $p^{(n)}(\theta)$ be the prior distribution assigning probabilities $\frac{1}{2}$ to each of $\theta^{(n)}$ and θ_0 . Then $8I(\mathcal{E}_i, p^{(n)}) / (\theta^{(n)} - \theta_0)^2 \xrightarrow{(n \rightarrow \infty)} I_i(\theta_0)$.

.7 Sequential design of experiments.

We now turn our attention to sequential experiments such that at each stage of experimentation the experimenter can freely choose any experiment from a given class of experiments.

Specifically, let $\Theta = \{1, 2, \dots, k\}$ be the common input space. Let \mathfrak{F} be a given class of experiments: i.e., each experiment in \mathfrak{F} is represented by

$$[(X, \theta), \{f_1(x), \dots, f_k(x)\}],$$

$$[(Y, c), \{g_1(y), \dots, g_k(y)\}],$$

and so on. At each stage of sequential experiments, the experimenter is free to select any one of the experiments in \mathfrak{F} and observe its value. The selection of any one experiment to be performed can depend upon the outcomes of all experiments that have been performed at earlier stages. All observations are assumed to be independent in the sense that any experiment that is to be performed at one stage, is independent of all experiments that have been performed at earlier stages.

Truncated sequential designs.

Now let U be a given non-negative uncertainty function defined on $\Xi = \{(\xi_1, \dots, \xi_k) \mid \xi_i \geq 0 (i=1, \dots, k), \sum_{i=1}^k \xi_i = 1\}$. Consider the following problem of truncated sequential design:

A fixed number N of experiments is to be performed, and it is desired to select the experiments X_1, X_2, \dots, X_N sequentially so as to maximize

$$I[(e_1, \dots, e_N), \xi; U] \equiv U(\xi) - E[U(\xi(X_1, \dots, X_N)) \mid \xi]$$

where $\xi \in \Xi$ is the prior distribution over $\Theta = \{1, \dots, k\}$.

By the familiar dynamic programming technique (Bellman, 1957) of working backward, an explicit rule for the derivation of the optimal design can be given. Let

$U_j(\xi)$ = the expected uncertainty obtained by using the optimal design when the knowledge about θ at present is ξ and we have j stages remaining.

Then we have

$$(7.1) \quad U_{j+1}(\xi) = \min_{X \in \mathfrak{F}} E[U_j(\xi(X)) | \xi]$$

($j=0,1,\dots,N-1$; $U_0(\xi) = U(\xi)$). In the following derivation it is assumed that all minima taken over the class \mathfrak{F} are actually attained at some $X \in \mathfrak{F}$. Let $X_{n-j}^*(\xi)$ be the experiment by which the minimum in the right-hand side of (7.1) is attained. Suppose that after having observed the first j experiments, the posterior distribution over Θ is ξ^j . Then, clearly, the optimal choice for the $(j+1)$ th experiment is given by $X_{j+1}^*(\xi^j)$. The collection of optimal choices $X_1^*(\xi^0)$ (where $\xi^0 = \xi$), $X_2^*(\xi^1), \dots, X_N^*(\xi^{N-1})$ determines the optimal design.

The main trouble with this constructive method of the optimal design is that it is often very difficult to compute for moderate values of N . The following theorem gives a sufficient condition to make the situation extremely simple.

Theorem 7.1. If there exists an experiment $X^* \in \mathfrak{F}$ such that

$$E[U(\xi(X^*)) | \xi] = \min_{X \in \mathfrak{F}} E[U(\xi(X)) | \xi], \text{ for all } \xi \in \Xi,$$

then for all $\xi \in \Xi$ and fixed N , performing N replications of X^* is the optimal design (for the truncated- N problem). (DeGroot, 1962).

Proof. From the hypothesis of the theorem, if it can be shown that

$$U_2(\xi) = \min_{X \in \mathcal{F}} E[U_1(\xi(X)) | \xi] = E[U_1(\xi(X^*)) | \xi], \text{ for all } \xi \in \Xi,$$

then $X_{N-1}^*(\xi^{N-2}) = X^*$ for all ξ^{N-2} , and by induction, repeated use of X^* would be optimal.

Let $X \in \mathcal{F}$ be any experiment other than X^* . From the well-known Bayes property:

$$\xi(X, X^*) = \xi(X^*, X) = \xi(X^*)(X),$$

it follows that

$$E[U(\xi(X^*, X)) | \xi] = E[E\{U(\xi(X^*)(X)) | \xi(X^*)\} | \xi]$$

$$(*) \quad E[U(\xi(X, X^*)) | \xi] = E[E\{U(\xi(X)(X^*)) | \xi(X)\} | \xi] = E[U_1(\xi(X)) | \xi].$$

But

$$U_1(\xi) \leq E[U(\xi(X)) | \xi] \quad \text{for all } \xi \text{ and } X$$

therefore

$$U_1(\xi(X^*)) \leq E[U(\xi(X^*)(X)) | \xi(X^*)].$$

Taking $E[\quad | \xi]$ on both sides and considering (*) we get

$$E[(U_1(\xi(X^*)) | \xi] \leq E[U_1(\xi(X)) | \xi] \quad \text{for all } \xi \text{ and } X.$$

This proves the theorem.

Corollary to theorem 7.1. Let $\mathcal{E}_X = [(X, \mathcal{G}), \{f_1(x), \dots, f_k(x)\}]$ and $\mathcal{E}_Y = [(Y, \mathcal{C}), \{g_1(y), \dots, g_k(y)\}]$ be two finite experiments. Each of the following two conditions is sufficient for performing N replications of X to be the optimal design (for the truncated- N problem).

$$(1) \text{ For } U(\xi) = -\sum_1^k \xi_1 \log \xi_1; \mathcal{E}_X \geq \mathcal{E}_Y$$

$$(2) \text{ For } k = 2 \text{ and } U(\xi) = \min(\xi_1, \xi_2); R_X \leq R_Y$$

Proof. (1) implies by definition (5.16) and (6.1), that for all ξ

$$E[U(\xi(X)) | \xi] = U(\xi) - I[\mathcal{E}_X, \xi; U]$$

$$\leq U(\xi) - I[\mathcal{E}_Y, \xi; U] = E[U(\xi(Y)) | \xi]$$

$$\text{with } U(\xi) = -\sum_1^k \xi_1 \log \xi_1.$$

While (2) implies (1) by theorem 6.5. But another proof is as follows:

If we take $U(\xi) = \min(\xi_1, \xi_2)$, i.e., the Bayes risk with respect to (ξ_1, ξ_2) when using the usual zero-one loss, we have

$$\begin{aligned} E[U(\xi(X)) | \xi] &= \int U\left(\frac{\xi_1 f_1(x)}{\xi_1 f_1(x) + \xi_2 f_2(x)}, \frac{\xi_2 f_2(x)}{\xi_1 f_1(x) + \xi_2 f_2(x)}\right) (\xi_1 f_1(x) + \xi_2 f_2(x)) d\lambda \\ &= \xi_1 \int_{\frac{f_2(x)}{f_1(x)} > \frac{\xi_1}{\xi_2}} f_1(x) d\lambda + \xi_2 \int_{\frac{f_2(x)}{f_1(x)} \leq \frac{\xi_1}{\xi_2}} f_2(x) d\lambda = R_X(\xi_1) \end{aligned}$$

by (6.5). Hence condition (2) implies for all $0 \leq \xi_1 \leq 1$

$$E[U(\xi(X)) | \xi] = R_X(\xi_1) \leq R_Y(\xi_1) = E[U(\xi(Y)) | \xi].$$

Now let us consider a truncated design problem in which the available experiments are dichotomous and we are interested in reducing the Bayes risk. Let the situation be given by

$$(7.2) \quad \begin{array}{c|cc} & \epsilon_X & \epsilon_Y \\ \hline (\zeta)H_1 & f_1(x) & g_1(y) \\ (1-\zeta)H_2 & f_2(x) & g_2(y) \end{array}$$

The uncertainty function is $U(\zeta, 1-\zeta) = \min(\zeta, 1-\zeta)$. The condition (2) in the corollary to theorem 7.1 is too restrictive, for it requires the uniform inequality in $0 \leq \zeta \leq 1$. Hence let us now consider the case in which neither $R_X \leq R_Y$ nor $R_X \geq R_Y$ is guaranteed.

Defining

$R_j(\zeta)$ = Bayes risk obtained by using the optimal design when the prior probability for H_1 is ζ and we have j stages remaining,

we can write the equations (7.1) as

$$(7.3) \quad R_{j+1}(\zeta) = \min \begin{cases} X: \int R_j \left(\frac{\zeta f_1(x)}{\zeta f_1(x) + (1-\zeta) f_2(x)} \right) (\zeta f_1(x) + (1-\zeta) f_2(x)) d\lambda \\ Y: \int R_j \left(\frac{\zeta g_1(y)}{\zeta g_1(y) + (1-\zeta) g_2(y)} \right) (\zeta g_1(y) + (1-\zeta) g_2(y)) d\mu \end{cases}$$

($j=0, \dots, N-1$; $R_0(\zeta) = \min(\zeta, 1-\zeta)$).

We transform this expression into another useful expression.

Let

$$(7.4) \quad Z = \log \frac{\zeta}{1-\zeta}, \quad \text{i.e.,} \quad \zeta = \frac{e^{z/2}}{e^{z/2} + e^{-z/2}}$$

and let

$$(7.5) \quad u(x) = \log \frac{f_2(x)}{f_1(x)}, \quad v(y) = \log \frac{g_2(y)}{g_1(y)}.$$

Then after taking one observation ζ is transformed to

$$\frac{\zeta f_1(x)}{\zeta f_1(x) + (1-\zeta)f_2(x)} = \frac{e^{\frac{1}{2}(z-u(x))}}{e^{\frac{1}{2}(z-u(x))} + e^{-\frac{1}{2}(z-u(x))}}, \quad \text{if X is observed}$$

$$\frac{\zeta g_1(y)}{\zeta g_1(y) + (1-\zeta)g_2(y)} = \frac{e^{\frac{1}{2}(z-v(y))}}{e^{\frac{1}{2}(z-v(y))} + e^{-\frac{1}{2}(z-v(y))}}, \quad \text{if Y is observed}$$

and we have

$$\zeta f_1(x) + (1-\zeta)f_2(x) = \sqrt{f_1(x)f_2(x)} \frac{e^{\frac{1}{2}(z-u(x))} + e^{-\frac{1}{2}(z-u(x))}}{e^{z/2} + e^{-z/2}},$$

$$\zeta g_1(y) + (1-\zeta)g_2(y) = \sqrt{g_1(y)g_2(y)} \frac{e^{\frac{1}{2}(z-v(y))} + e^{-\frac{1}{2}(z-v(y))}}{e^{z/2} + e^{-z/2}}.$$

Hence if we set

$$h_j(z) = (e^{z/2} + e^{-z/2}) R_j \left(\frac{e^{z/2}}{e^{z/2} + e^{-z/2}} \right) \quad (j=0,1,\dots,N)$$

we obtain

$$(7.6) \quad h_{j+1}(z) = \min \begin{cases} X: \rho \int h_j(z-u(x)) f(x) d\lambda \\ Y: k \int h_j(z-v(y)) g(y) d\mu \end{cases}$$

$(-\infty < z < \infty; j = 0, 1, \dots, N-1; h_0(z) = \min(e^{z/2}, e^{-z/2}))$, where

$$(7.7) \quad \begin{cases} f(x) = \sqrt{f_1(x)f_2(x)} / \rho, & \rho = \int \sqrt{f_1(x)f_2(x)} d\lambda \\ g(y) = \sqrt{g_1(y)g_2(y)} / k, & k = \int \sqrt{g_1(y)g_2(y)} d\mu. \end{cases}$$

If the situation is given by $\lambda = \mu$, $g_1 = f_2$, $g_2 = f_1$:

$$(7.8) \quad \begin{array}{c|cc} & \mathcal{E}_X & \mathcal{E}_Y \\ \hline (\zeta)H_1 & f_1(x) & f_2(y) \\ (1-\zeta)H_2 & f_2(x) & f_1(y) \end{array}$$

that is, a two-armed bandit-problem type, then (7.3) and (7.6) become

$$(7.9) \quad R_{j+1}(\xi) = \min \begin{cases} X: \int R_j \left(\frac{\zeta f_1(x)}{\zeta f_1(x) + (1-\zeta) f_2(x)} \right) (\zeta f_1(x) + (1-\zeta) f_2(x)) d\lambda \\ Y: \int R_j \left(\frac{\zeta f_2(x)}{\zeta f_2(x) + (1-\zeta) f_1(x)} \right) (\zeta f_2(x) + (1-\zeta) f_1(x)) d\lambda \end{cases}$$

($j=0,1,\dots,N-1$; $R_0(\zeta) = \min(\zeta, 1-\zeta)$), and

$$(7.10) \quad h_{j+1}(z) = \min \begin{cases} X: \rho \int h_j(z-u(x)) f(x) d\lambda \\ Y: \rho \int h_j(z+u(y)) f(y) d\lambda \end{cases}$$

($-\infty < z < \infty$; $j=0,1,\dots,N-1$; $h_0(z) = \min(e^{z/2}, e^{-z/2})$), respectively, where $u(x)$, $f(x)$ and ρ are defined by (7.5) and (7.7).

In spite of this rather simple form of the functional equation it is quite difficult to solve even for a specific case. For this reason it is interesting and useful to investigate some design criteria which have some justification though not optimal. We shall first show in the following the optimal designs for small N , in the case of binomial distributions, to illustrate the highly complicated structure of the optimal designs.

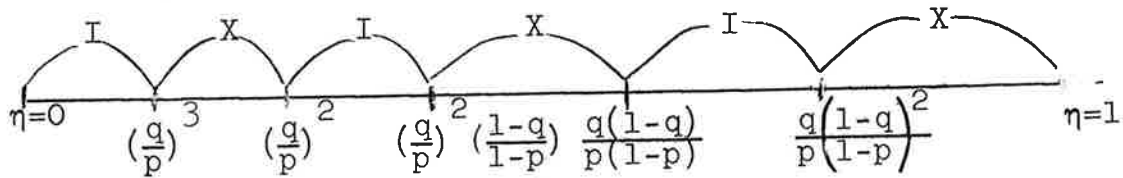
Suppose that, in (7.8) $f_1(x)$ and $f_2(x)$ are binomial densities with parameters p and q , respectively. Then we have from (7.10)

$$(7.11) \quad h_{j+1}(z) = \min \begin{cases} X: \sqrt{pq} h_j(z - \log \frac{q}{p}) + \sqrt{(1-p)(1-q)} h_j(z - \log \frac{1-q}{1-p}) \\ Y: \sqrt{pq} h_j(z + \log \frac{q}{p}) + \sqrt{(1-p)(1-q)} h_j(z + \log \frac{1-q}{1-p}) \end{cases}$$

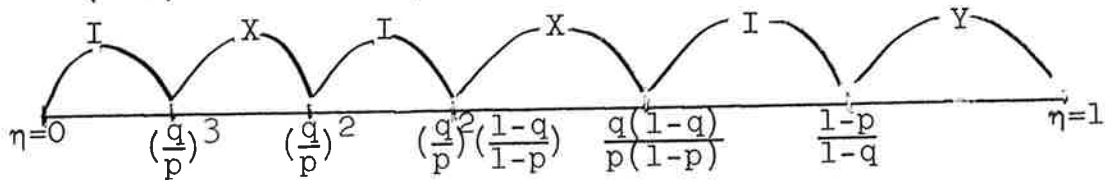
($-\infty < z < \infty$; $j=0,1,\dots,N-1$; $h_0(z) = \min(e^{z/2}, e^{-z/2})$)

We shall now present here the optimal choices at the first step in the optimal design for the case $N = 3$. (The description of the optimal designs for the remaining steps will be omitted.) The solution is expressed in $0 \leq \eta \leq 1$ corresponding to $0 \leq \zeta \leq \frac{1}{2}$, which by symmetry may be extended to $0 \leq \zeta \leq 1$. (I denotes indifference.)

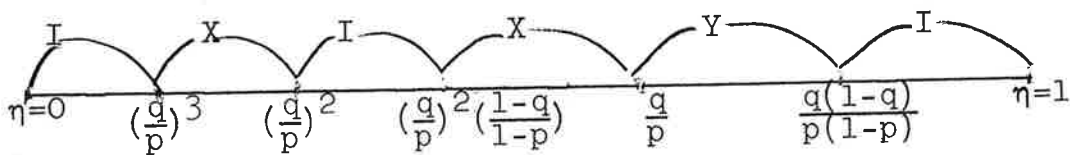
(a) For $\frac{q}{p} < \left(\frac{1-p}{1-q}\right)^3 < \left(\frac{1-p}{1-q}\right)^2 < 1$:



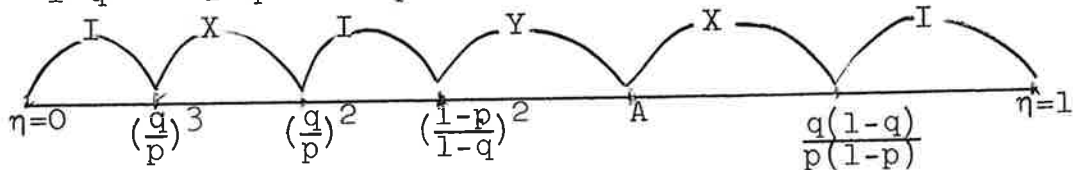
(b) For $\left(\frac{1-p}{1-q}\right)^3 < \frac{q}{p} < \left(\frac{1-p}{1-q}\right)^2 < 1$:



(c) For $\left(\frac{1-p}{1-q}\right)^2 < \frac{q}{p} < \left(\frac{1-p}{1-q}\right)^{3/2} < 1$:



(d) For $\left(\frac{1-p}{1-q}\right)^2 < \left(\frac{1-p}{1-q}\right)^{3/2} < \frac{q}{p} < 1$:



where $A = \frac{(1-p)^2(1-p-q) - q^2(1-q)}{(1-q)^2(1-p-q) - p^2(1-p)}$.

Let us now turn our attention to the problem of finding a sequential design which will maximize the sum of N independent observations, in the same statistical situation as before. Let $W_j(\zeta)$ be the expected sum of j observations when ζ is the a priori probability for H_1 and the optimal design is used. Suppose that $\mu_i = \int x f_i(x) d\lambda$ ($i=1,2$) exists and $\mu_1 > \mu_2$. Then we have

$$W_{j+1}(\zeta) = \max \begin{cases} X: \{\zeta E_{f_1} + (1-\zeta) E_{f_2}\} \left\{ X + W_j \left(\frac{\zeta f_1(X)}{\zeta f_1(X) + (1-\zeta) f_2(X)} \right) \right\} \\ Y: \{\zeta E_{f_2} + (1-\zeta) E_{f_1}\} \left\{ Y + W_j \left(\frac{\zeta f_2(Y)}{\zeta f_2(Y) + (1-\zeta) f_1(Y)} \right) \right\} \end{cases}$$

$$= \max \begin{cases} X: \mu_2 + (\mu_1 - \mu_2)\zeta + \int W_j(\zeta_X) (\zeta f_1(x) + (1-\zeta) f_2(x)) d\lambda \\ Y: \mu_2 + (\mu_1 - \mu_2)(1-\zeta) + \int W_j(\zeta_Y) (\zeta f_2(y) + (1-\zeta) f_1(y)) d\lambda \end{cases}$$

($j=0,1,\dots,N-1$; $W_0(\zeta) \equiv 0$), where $\zeta_X = \frac{\zeta f_1(x)}{\zeta f_1(x) + (1-\zeta) f_2(x)}$ and

$\zeta_Y = \frac{\zeta f_2(y)}{\zeta f_2(y) + (1-\zeta) f_1(y)}$. Without loss of generality we can take $\mu_1=1$,

$\mu_2=0$. And if we let $R_j(\zeta) = j - W_j(\zeta)$ we obtain

$$(7.12) \quad R_{j+1}(\zeta) = \min \begin{cases} X: 1-\zeta + \int R_j(\zeta_X) (\zeta f_1(x) + (1-\zeta) f_2(x)) d\lambda \\ Y: \zeta + \int R_j(\zeta_Y) (\zeta f_2(y) + (1-\zeta) f_1(y)) d\lambda \end{cases}$$

($j=0,\dots,N-1$; $R_0(\zeta) \equiv 0$).

By a stationary design we shall mean a design in which the choice at the $(m+1)$ -th step is a function only of the a posteriori probability after the m -th step, ζ^m .

Theorem 7.2 For the sequential design problem (7.12),

(i) the stationary design D^* , which chooses at $(m+1)$ -th step

$$\begin{Bmatrix} X \\ Y \end{Bmatrix} \text{ according as } \zeta^m \begin{Bmatrix} \geq \\ < \end{Bmatrix} \frac{1}{2},$$

is the optimal design.

(ii) $\lim_{N \rightarrow \infty} R_N(\zeta) = R(\zeta)$ exists for every ζ , and satisfies

$$(7.13) \quad R(\zeta) = \min \begin{cases} 1 - \zeta + \int R(\zeta_X)(\zeta f_1(x) + (1 - \zeta)f_2(x)) d\lambda \\ \zeta + \int R(\zeta_Y)(\zeta f_2(y) + (1 - \zeta)f_1(y)) d\lambda \end{cases} \quad (\text{Feldman, 1962})$$

The first part of this theorem states that the optimal design for the problem is a stationary and "one-step optimal" design, i.e., which chooses experiments each time as though there were one stage remaining. While the second part of the theorem says that the optimal design D^* is consistent, i.e., under D^* , $N^{-1}R_N(\zeta) \xrightarrow{(N \rightarrow \infty)} 0$, or in the original presentation of the problem, $N^{-1}W_n(\zeta) \xrightarrow{(N \rightarrow \infty)} \max(\mu_1, \mu_2)$. Before proving the theorem we present the following lemma.

Lemma Let the posterior probability of H_1 given ζ , be

$$\begin{cases} \zeta_X = \frac{\zeta f_1(x)}{\zeta f_1(x) + (1 - \zeta)f_2(x)} & , \quad \text{if } X \text{ is observed,} \\ \zeta_Y = \frac{\zeta f_2(y)}{\zeta f_2(y) + (1 - \zeta)f_1(y)} & , \quad \text{if } Y \text{ is observed.} \end{cases}$$

Then ζ_X and ζ_Y are both stochastically increasing functions of ζ , i.e., for every fixed w

$$\Pr\{\zeta_X \geq w | \zeta\} = \zeta \Pr\{\zeta_X \geq w | H_1\} + (1 - \zeta) \Pr\{\zeta_X \geq w | H_2\}$$

are increasing functions. (Similar definition holds for ζ_Y).

Proof. We prove only for ζ_X , the proof for ζ_Y being analogous. We have

$$(*) \Pr\{\zeta_X \geq w|\zeta\} = \zeta P_1\left\{\frac{f_1(X)}{f_2(X)} \geq \frac{1-\zeta}{\zeta} \cdot \frac{w}{1-w}\right\} + (1-\zeta)P_2\left\{\frac{f_1(X)}{f_2(X)} \geq \frac{1-\zeta}{\zeta} \cdot \frac{w}{1-w}\right\}$$

and generally we have

$$P_1\left\{\frac{f_1(X)}{f_2(X)} \geq r\right\} \begin{cases} \geq rP_2\left\{\frac{f_1(X)}{f_2(X)} \geq r\right\}, \\ = 1 - P_1\left\{\frac{f_1(X)}{f_2(X)} < r\right\} \geq 1 - rP_2\left\{\frac{f_1(X)}{f_2(X)} < r\right\}. \end{cases}$$

Using the upper inequality if $r \geq 1$, and the lower inequality if $r < 1$, we get

$$P_1\left\{\frac{f_1(X)}{f_2(X)} \geq r\right\} \geq P_2\left\{\frac{f_1(X)}{f_2(X)} \geq r\right\}, \quad \text{for all } r > 0.$$

Now from (*), $P\{\zeta_X \geq w|\zeta\}$ is a convex linear combination of two non-decreasing functions, one of which is uniformly greater than the other.

Proof of Theorem 7.2.

Writing (7.12) as

$$R_{j+1}(\zeta) = \min \begin{cases} X: R_{j+1}^X(\zeta) \\ Y: R_{j+1}^Y(\zeta) \end{cases} \quad (j=0,1,\dots,N-1)$$

it suffices to show that $\Delta_j(\zeta) \equiv R_j^Y(\zeta) - R_j^X(\zeta)$ is, for each $j=1,\dots,N$,

(a) continuous and increasing in $0 < \zeta < 1$; and

(b) $\Delta_j\left(\frac{1}{2}\right) = 0$.

We can show by induction, that for every j , $R_j(\zeta)$ is continuous and symmetric about $\zeta = \frac{1}{2}$ and that $\Delta_j\left(\frac{1}{2}\right) = 0$.

So we shall next show the increasing property of $\Delta_j(\zeta)$. This is clearly true for $\Delta_1(\zeta) = 2\zeta - 1$. Let the inductive hypothesis be

H_m : $\Delta_j(\zeta)$ is increasing in $0 < \zeta < 1$, for $j=1, \dots, m$.

Let $R_{j+1}^{XY}(\zeta)$ be the "risk" corresponding to the design which chooses X then Y for the first two steps and then follows the optimal design for the remaining $(j-1)$ steps. Let $R_{j+1}^{YX}(\zeta)$ be similarly defined. Then writing $\zeta E_{f_1} + (1-\zeta)E_{f_2}$ and $\zeta E_{f_2} + (1-\zeta)E_{f_1}$ simply as E, we have

$$R_{m+1}^{XY}(\zeta) = 1-\zeta + E[R_m^Y(\zeta_X)],$$

$$R_{m+1}^X(\zeta) = 1-\zeta + E[R_m(\zeta_X)]$$

$$= 1-\zeta + E[\pi^*(\zeta_X)R_m^X(\zeta_X) + (1-\pi^*(\zeta_X))R_m^Y(\zeta_X)], \quad (\text{by } H_m)$$

$$R_{m+1}^{YX}(\zeta) = \zeta + E[R_m^X(\zeta_Y)],$$

$$R_{m+1}^Y(\zeta) = \zeta + E[R_m(\zeta_Y)]$$

$$= \zeta + E[\pi^*(\zeta_Y)R_m^X(\zeta_Y) + (1-\pi^*(\zeta_Y))R_m^Y(\zeta_Y)], \quad (\text{by } H_m)$$

where $\pi^*(\zeta) \equiv \begin{cases} 1, & \text{if } \zeta \geq \frac{1}{2}, \text{ and} \\ 0, & \text{if } \zeta < \frac{1}{2} \end{cases}$

$$R_{m+1}^{XY}(\zeta) = R_{m+1}^{YX}(\zeta),$$

we have

$$\Delta_{m+1}(\zeta) = R_{m+1}^Y(\zeta) - R_{m+1}^X(\zeta)$$

$$= (R_{m+1}^{XY}(\zeta) - R_{m+1}^X(\zeta)) - (R_{m+1}^{YX}(\zeta) - R_{m+1}^Y(\zeta))$$

$$\begin{aligned}
&= E[\pi^*(\zeta_X)(R_m^Y(\zeta_X) - R_m^X(\zeta_X))] - E[(1 - \pi^*(\zeta_Y))(R_m^X(\zeta_Y) - R_m^Y(\zeta_Y))] \\
&= E[\pi^*(\zeta_X)\Delta_m(\zeta_X)] + E[(1 - \pi^*(\zeta_Y))\Delta_m(\zeta_Y)].
\end{aligned}$$

By H_m and the lemma, $\Delta_m(\zeta_X)$ and $\Delta_m(\zeta_Y)$ are both stochastically increasing. The last expression, therefore, is the sum of two (strictly) increasing terms except when $\Pr\{\zeta_X \leq \frac{1}{2} | \zeta\} = \Pr\{\zeta_Y < \frac{1}{2} | \zeta\} = 0$. But by the lemma such ζ does not exist.

(ii) To prove the second part of the theorem, it suffices to show convergence of the series

$$\sum_{j=1}^{\infty} \Pr\{\zeta^j < \frac{1}{2} | H_1, \zeta, D^*\}, \quad \sum_{j=1}^{\infty} \Pr\{\zeta^j \geq \frac{1}{2} | H_2, \zeta, D^*\}.$$

Because of symmetry we need only show the convergence of the first series

Let

$$Z_{j+1} = \begin{cases} z_j + u(x), & \text{if } x \text{ is observed} \\ z_j - u(x), & \text{if } y \text{ is observed.} \end{cases}$$

Then for $j = 1, 2, \dots$

$$\begin{aligned}
\Pr\{\zeta^j < \frac{1}{2} | H_1, \zeta, D^*\} &= \Pr\{z_j < 0 | H_1, \zeta, D^*\} \\
&\leq E[e^{-\frac{1}{2}z_j} | H_1, \zeta, D^*] \quad (\text{because } P\{Z \leq 0\} \leq E[e^{-\frac{1}{2}Z}]) \\
&= \Pr\{\zeta^{j-1} \geq \frac{1}{2} | H_1, \zeta, D^*\} E[e^{-\frac{1}{2}(z_{j+1} + u(x))} | \zeta^{j-1} \geq \frac{1}{2}, H_1, \zeta, D^*] \\
&\quad + \Pr\{\zeta^{j-1} < \frac{1}{2} | H_1, \zeta, D^*\} E[e^{-\frac{1}{2}(z_{j+1} - u(y))} | \zeta^{j-1} < \frac{1}{2}, H_1, \zeta, D^*] \\
&= \rho E[e^{-\frac{1}{2}z_{j-1}} | H_1, \zeta, D^*] \quad (\text{because } E_1[e^{-\frac{1}{2}u(X)}] = E_2[e^{\frac{1}{2}u(Y)}] = \rho) \\
&= \dots = \rho^j e^{-\frac{1}{2}z_0}.
\end{aligned}$$

Non-truncated sequential designs.

Now let us consider non-truncated types of sequential design problems. That is, the problem of optimally choosing the experiments to be performed sequentially from a class of available experiments is considered when the goal is to minimize over X_1, X_2, \dots

$E[\text{the first } n \text{ such that } U(\xi(X_1, \dots, X_n)) < \delta | \xi]$
 for a given $\delta > 0$ and each fixed ξ .

The following simple example illustrates interesting points involved and some peculiarities that can arise in this type of problem.

Example 7.1 (De Groot, 1962)

Let the statistical situation be given by

Exp. Hyp.	ϵ_X	ϵ_Y
$(\xi_1) H_1$	$f_1(x) \equiv 1, \text{ in } (0,1)$	$g_1(y) = \begin{cases} \epsilon, & y = 0 \\ \bar{\epsilon}, & y = 1 \end{cases}$
$(\xi_2) H_2$	$f_2(x) \equiv 1, \text{ in } (1-\alpha, 2-\alpha)$	$g_2(y) = \begin{cases} \bar{\epsilon}, & y = 0 \\ \epsilon, & y = 1 \end{cases}$

where $0 < \alpha < 1$ and $0 < \epsilon = 1 - \bar{\epsilon} < \frac{1}{2}$.

Let the uncertainty function be $U(\xi) = \min(\xi_1, \xi_2)$.

Since for any prior knowledge ξ the posterior probabilities for H_1 are

$$\xi_1(x) = \begin{cases} 1, & 0 \leq x < 1-\alpha \\ \xi_1, & 1-\alpha \leq x \leq 1 \\ 0, & 1 < x \leq 2-\alpha \end{cases}$$

$$\xi_1(y) = \begin{cases} \epsilon \xi_1 / (\epsilon \xi_1 + \bar{\epsilon} \xi_2), & y = 0 \\ \bar{\epsilon} \xi_1 / (\bar{\epsilon} \xi_1 + \epsilon \xi_2), & y = 1 \end{cases}$$

$$\xi_1(y_1, y_2) = \begin{cases} \epsilon^2 \xi_1 / (\epsilon^2 \xi_1 + \bar{\epsilon}^2 \xi_2), & \text{if } y_1 = y_2 = 0 \\ \bar{\epsilon}^2 \xi_1 / (\bar{\epsilon}^2 \xi_1 + \epsilon^2 \xi_2), & \text{if } y_1 = y_2 = 1 \\ \xi_1, & \text{if otherwise} \end{cases}$$

we have for $\xi \in \mathbb{H}_0 \equiv \{\xi \mid \epsilon < \xi_1 < \bar{\epsilon}\}$

$$E[U(\xi(X)) \mid \xi] = \alpha U(\xi) + (1-\alpha) \cdot 0 = \alpha \min(\xi_1, \xi_2)$$

$$E[U(\xi(Y)) \mid \xi] = (\epsilon \xi_1 + \bar{\epsilon} \xi_2) \min(\xi_1(0), 1 - \xi_1(0)) + (\bar{\epsilon} \xi_1 + \epsilon \xi_2) \min(\xi_1(1), 1 - \xi_1(1))$$

$$= \begin{cases} \xi_1, & 0 \leq \xi_1 \leq \epsilon \\ \alpha, & \epsilon \leq \xi_1 \leq \bar{\epsilon} \\ \xi_2, & \bar{\epsilon} \leq \xi_1 \leq 1. \end{cases}$$

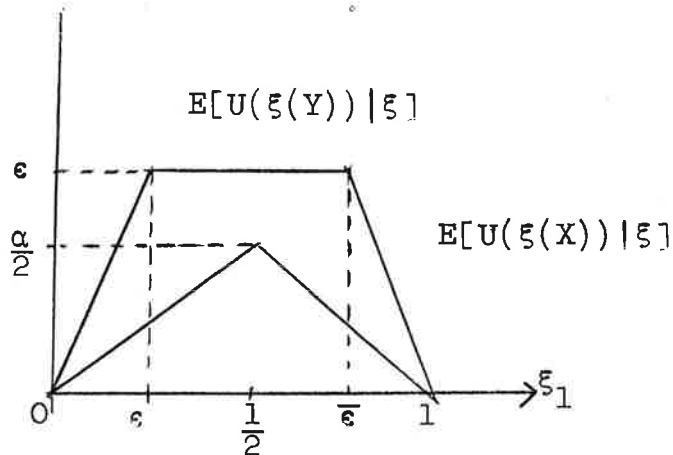


Figure 7.1

Thus we get from theorem 7.1 and figure 7.1

Proposition 1 If

$$(+)$$

$$\frac{\alpha}{2} \leq \epsilon$$

then for all $\xi \in \Xi$ and fixed n , performing n replications of X is the optimal design.

Let S_X denote the sequential design whereby the experiment \mathcal{E}_X is replicated repeatedly. S_X continues as long as $1-\alpha \leq X \leq 1$. But $\Pr\{1-\alpha \leq X \leq 1 | \xi\} = \alpha$ for all ξ . Hence

$$(*) E[n(S_X) | \xi] = \sum_{m=1}^{\infty} \Pr\{n(S_X) \geq m | \xi\} = \sum_{m=1}^{\infty} \alpha^{m-1} = \frac{1}{1-\alpha}, \quad \xi \in \Xi_0.$$

Now let S_Y denote the sequential design whereby the experiment \mathcal{E}_Y is replicated repeatedly.

(i) When $\xi_1 = \frac{1}{2}$, $\xi_1(0) = \epsilon = 1 - \xi_1(1)$, so that $E[n(S_Y) | \xi] = 1$.

(ii) suppose that $\epsilon < \xi_1 < \frac{1}{2}$. Since $\xi_1(0) < \epsilon$, $\xi_1(1,1) \geq \bar{\epsilon}$ and $\epsilon < \xi_1(1,0) < \frac{1}{2}$, S_Y continues in the pattern $1,0,1,0,1,0,\dots$ and ceases as soon as this pattern is violated. Hence

$$\begin{aligned} E[n(S_Y) | \xi] &= \sum_{m=0}^{\infty} \Pr\{n(S_Y) > m | \xi\} \\ &= \sum_{m=0}^{\infty} \Pr\{n(S_Y) > 2m | \xi\} + \sum_{m=0}^{\infty} \Pr\{n(S_Y) > 2m+1 | \xi\} \\ &= \sum_0^{\infty} (\epsilon\bar{\epsilon})^m + \sum_0^{\infty} (\epsilon\bar{\epsilon})^m (\bar{\epsilon}\xi_1 + \epsilon\xi_2) = \frac{1 + \bar{\epsilon}\xi_1 + \epsilon\xi_2}{1 - \epsilon\bar{\epsilon}}. \end{aligned}$$

(iii) By the similar argument we find that when $\frac{1}{2} < \xi_1 < \bar{\epsilon}$, S_Y continues in the pattern $0,1,0,1,0,1,\dots$ and ceases as soon as the pattern is violated, and

$$E[n(S_Y) | \xi] = (1 + \epsilon\xi_1 + \bar{\epsilon}\xi_2) / (1 - \epsilon\bar{\epsilon}).$$

Summarizing (i), (ii) and (iii) we obtain

$$(*) \sup_{\xi \in \Xi_0} E[n(S_Y) | \xi] = \frac{3}{2(1 - \epsilon\bar{\epsilon})}.$$

Thus (*) and (*) give

Proposition 2 If

$$(\ddagger) \quad \frac{3}{2(1-\epsilon\bar{\epsilon})} \leq \frac{1}{1-\alpha}$$

then for all $\xi \in \Xi$, S_Y is the optimal design

Proof. Since $E[n(S_Y) | \xi] < E[n(S_X) | \xi]$ for all $\xi \in \Xi_0$, S_Y is better than S_X . And if the optimal design observes X at some stage, then from that stage on it must repeat observations of X until it stops. Hence the optimal design never observes X.

Combining Propositions 1 and 2 we obtain

Proposition 3 If α and ϵ simultaneously satisfy both (+) and (\ddagger), (e.g., $\alpha = \frac{1}{2}$, $\epsilon = \frac{1}{4}$), then for such values it is true that for any fixed number of experiments repeated observations of X is optimal, whereas for any unfixed number of experiments repeated observations of Y is optimal.

Let us again consider the two-armed-bandit problem described by (7.8). Suppose that our problem of sequential design is that of minimizing the expected number of observations required to move the a posteriori probability for H_1 outside of an interval $(r, 1-r)$:

$$E[\text{the first } n \text{ such that } \zeta(Z_1, \dots, Z_n) \notin (r, 1-r) | \zeta] \rightarrow \min_{Z_1, Z_2, \dots}$$

where $\zeta(Z_1, \dots, Z_n)$ is the a posteriori probability for H_1 after having observed Z_1, \dots, Z_n , and each Z_i is the i -th observation either from \mathcal{E}_X or from \mathcal{E}_Y .

Let us define

$N(\zeta)$ = the expected sample size obtained by using the optimal sequential

design when the prior probability for H_1 is ζ .

Then we have

$$N(\zeta) = \min \begin{cases} X: 1 + \int N\left(\frac{\zeta f_1(x)}{\zeta f_1(x) + (1-\zeta)f_2(x)}\right) (\zeta f_1(x) + (1-\zeta)f_2(x)) d\lambda \\ Y: 1 + \int N\left(\frac{\zeta f_2(x)}{\zeta f_2(x) + (1-\zeta)f_1(x)}\right) (\zeta f_2(x) + (1-\zeta)f_1(x)) d\lambda \end{cases}$$

if $r < \zeta < 1-r$, and $N(\zeta) = 0$ if otherwise. After making the transformations (7.4), (7.5), and (7.7) we obtain

$$h(z) = e^{z/2} + e^{-z/2} + \min \begin{cases} X: \rho \int h(z-u(x)) f(x) d\lambda \\ Y: \rho \int h(z+u(y)) f(y) d\lambda \end{cases}$$

if $-A < z < A \equiv \log \frac{1-r}{r}$, and $h(z) = 0$ if otherwise.

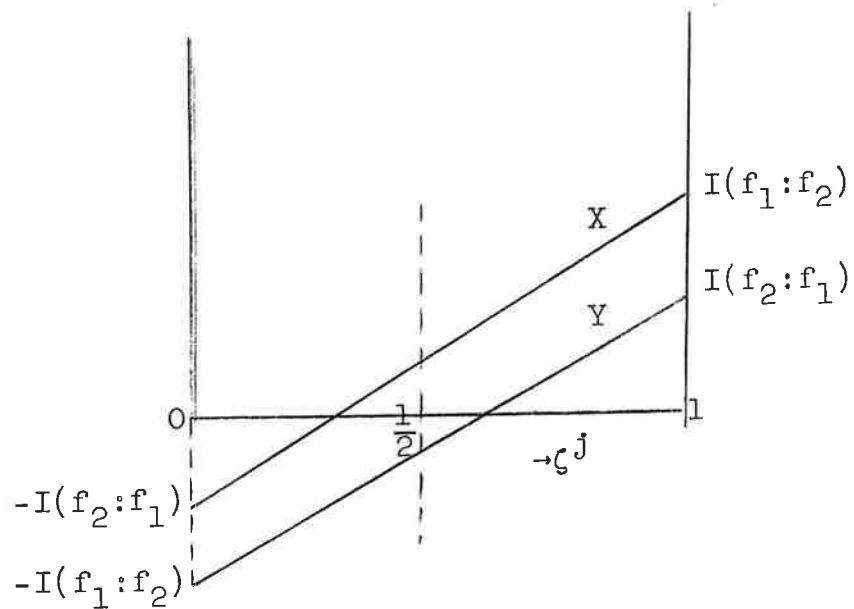
It seems very difficult to discuss the solvability and derivation of solution of this functional equation. We, therefore, consider some reasonable sequential designs though not optimal.

There are two random walks, both on the z -axis with boundaries at A and $-A$, one of which is determined by the results of observations of X and the other is determined by the results of observations of Y . Let $z_j = \log \frac{\zeta^j}{1-\zeta^j}$ be the point at which the walk has arrived after having made j observations. Then if at the $(j+1)$ -th step X is observed the expected movement in the walk is

$$\begin{aligned} E_X[z_{j+1} - z_j | \zeta^j] &= E_X \left[\log \frac{\zeta^{j+1}(X)}{1-\zeta^{j+1}(X)} - \log \frac{\zeta^j}{1-\zeta^j} \mid \zeta^j \right] \\ &= \zeta^j I(f_1:f_2) - (1-\zeta^j) I(f_2:f_1) \\ &= \zeta^j J(f_1;f_2) - I(f_2:f_1). \end{aligned}$$

Similarly if Y is observed, the expected movement is

$$E_Y[z_{j+1} - z_j | \zeta^j] = \zeta^j J(f_1;f_2) - I(f_1:f_2).$$



Thus, if $I(f_1:f_2) > I(f_2:f_1)$, the X-walk yields an expected step greater (smaller) in magnitude than the Y-walk for $\zeta^j > (<) \frac{1}{2}$. This consideration leads to the conjecture that one reasonable design may be, at the $(j+1)$ -th step,

(i) to take $\begin{Bmatrix} X \\ Y \end{Bmatrix}$ -walk, if $\zeta^j \begin{Bmatrix} > \\ < \end{Bmatrix} \frac{1}{2}$.

(If $I(f_1:f_2) < I(f_2:f_1)$, the same results will hold with X and Y interchanged.) It is easily seen that this design also coincides with

(ii) choosing the random variable corresponding to the larger of $I_X(\zeta^j) \equiv \zeta^j I(f_1:f_2) + (1-\zeta^j) I(f_2:f_1)$ and $I_Y(\zeta^j) \equiv \zeta^j I(f_2:f_1) + (1-\zeta^j) I(f_1:f_2)$

It is hardly to be expected that this stationary design would be optimal but we know that when $R_X \leq R_Y$ it will never lead to use of Y (by theorem 6.4) and it coincides with the optimal design for the truncated problems (by the corollary to theorem 7.1). This subsection concludes with a general result which shows that this design is better than either S_X (sequential design requiring X at every step) or S_Y .

Let us consider the more general case described by (7.2). Let, for any sequential design S , $E[n|S, x] \equiv E[\text{the first } n \text{ such that } \zeta(Z_1, \dots, Z_n) \notin (r, 1-r) | S, \zeta]$, where $z = \log \frac{\zeta}{1-\zeta}$.

Theorem 7.3 Let X and Y have densities f_i and g_i , respectively, under hypothesis H_i such that both $\log(f_2(X)/f_1(X))$ and $\log(g_2(Y)/g_1(Y))$ assume positive and negative values with positive probability. Let $S^{(1)}$ and $S^{(2)}$ be two stationary designs and let S be that design which requires, at the $(j+1)$ -th step,

the random variable corresponding to $\min_{i=1,2} E[n|S^{(i)}, z_j]$.

Then S is better than both of $S^{(1)}$ and $S^{(2)}$, that is,

$$E[n|S, z] \leq \min_{i=1,2} E[n|S^{(i)}, z], \quad \text{for all } -\infty < z < \infty.$$

(Bradt and Karlin, 1956)

Proof. Each sequential design $S^{(i)}$, is specified by the set

$$\Gamma^{(i)} \equiv \{z \mid |z| \leq A, S^{(i)} \text{ requires } X \text{ at } z\}.$$

We have

$$E[n|S^{(i)}, z] = \begin{cases} 1 + E_X\{E[n|S^{(i)}, z-u(X)]\}, & \text{if } z \in \Gamma^{(i)} \\ 1 + E_Y\{E[n|S^{(i)}, z-v(Y)]\}, & \text{if } z \in (-A, A) - \Gamma^{(i)} \\ 0, & \text{if } |z| > A = \log \frac{1-r}{r}. \end{cases}$$

Define the set Δ by

$$\Delta \equiv \{z \mid |z| \leq A, E[n|S^{(1)}, z] < E[n|S^{(2)}, z]\}.$$

Then we have, for $|z| \leq A$

$$\min_{i=1,2} E[n|S^{(i)}, z] = \begin{cases} E[n|S^{(1)}, z], & z \in \Delta \\ E[n|S^{(2)}, z], & z \notin \Delta \end{cases}$$

$$= \begin{cases} 1 + E_X\{E[n|S^{(1)}, z-u(X)]\}, & z \in \Delta \cap \Gamma^{(1)} \\ 1 + E_Y\{E[n|S^{(1)}, z-v(Y)]\}, & z \in \Delta - \Gamma^{(1)} \\ 1 + E_X\{E[n|S^{(2)}, z-u(X)]\}, & z \in \Gamma^{(2)} - \Delta \\ 1 + E_Y\{E[n|S^{(2)}, z-v(Y)]\}, & z \in (\Gamma^{(2)} \cup \Delta)^c \end{cases}$$

$$E[n|S, z] = \begin{cases} 1 + E_X\{E[n|S, z-u(X)]\}, & z \in (\Delta \cap \Gamma^{(1)}) \cup (\Gamma^{(2)} - \Delta) \\ 1 + E_Y\{E[n|S, z-v(Y)]\}, & z \in (\Delta - \Gamma^{(1)}) \cup (\Gamma^{(2)} \cup \Delta)^c, \end{cases}$$

so that

$$G(Z) \equiv \min_{i=1,2} E[n|S^{(i)}, z] - E[n|S, z]$$

$$= \begin{cases} E_X\{E[n|S^{(1)}, z-u(X)] - E[n|S, z-u(X)]\}, & z \in \Delta \cap \Gamma^{(1)} \\ E_Y\{E[n|S^{(1)}, z-v(Y)] - E[n|S, z-v(Y)]\}, & z \in \Delta - \Gamma^{(1)} \\ E_X\{E[n|S^{(2)}, z-u(X)] - E[n|S, z-u(X)]\}, & z \in \Gamma^{(2)} - \Delta \\ E_Y\{E[n|S^{(2)}, z-v(Y)] - E[n|S, z-v(Y)]\}, & z \in (\Gamma^{(2)} \cup \Delta)^c \end{cases}$$

$$\geq \begin{cases} E_X\{G(z-u(X))\}, & z \in (\Delta \cap \Gamma^{(1)}) \cup (\Gamma^{(2)} - \Delta) \\ E_Y\{G(z-v(Y))\}, & z \in (\Delta - \Gamma^{(1)}) \cup (\Gamma^{(2)} \cup \Delta)^c \end{cases}$$

Take a sequence of points $\{z_m\}_{m=1}^{\infty}$ in $(-A, A)$ such that $G(z_m) \xrightarrow{(m \rightarrow \infty)} \inf_z G(z)$.

Then by the above inequality there exists

$$\{z_m'\} \subset \{z_m\}; \quad \inf_z G(z) = \lim_{m' \rightarrow \infty} G(z_{m'} - u(X)), \quad \text{with probability 1}$$

or there exists

$$\{z_m''\} \subset \{z_m\}; \quad \inf_z G(z) = \lim_{m'' \rightarrow \infty} G(z_{m''} - v(Y)), \quad \text{with probability 1}$$

and since $u(X)$ and $v(Y) > 0$ with positive probability there exists

$$\lambda > 0; \quad \lim_{m' \rightarrow \infty} G(z_{m'}, -\lambda) = \inf_z G(z)$$

or

$$\lim_{m'' \rightarrow \infty} G(z_{m''}, -\lambda) = \inf_z G(z).$$

Let us denote this last sequence ($\{z_{m'}, -\lambda\}$ or $\{z_{m''}, -\lambda\}$) by $\{z_{m(1)}\}$, so that

$$\lim_{m(1) \rightarrow \infty} G(z_{m(1)}) = \inf_z G(z). \quad \text{Repeating the argument a finite number of}$$

times, we arrive at a sequence $z_{m(s)} \leq -A$ and $\inf_z g(z) = \lim_{m(s) \rightarrow \infty} G(z_{m(s)}) = 0$.

Tests of composite hypotheses

We shall present here a procedure devised by H. Chernoff (1959) for the sequential design of experiments where the problem is one of testing composite hypotheses. We assume that there are two terminal decisions and a class of available experiments. After each observation, the statistician decides whether to continue to experiment or not. If he decides to continue, he must select one of the available experiments. If he decides to stop he must select one of the two terminal decisions.

Suppose that we have to test a composite hypothesis $H_1: \theta \in \omega_1$ against the alternative $H_2: \theta \in \omega_2$. There is available a set of experiments $\{\mathcal{E}\}$, each of which may be replicated. Designate the n -th experiment by \mathcal{E}^n . Although the choice of the $(n+1)$ -st experiment may depend on the past, once it is selected \mathcal{E}^{n+1} is assumed independent of $\mathcal{E}^1, \dots, \mathcal{E}^n$. The risk of a procedure is the expected cost of sampling plus the expected loss due to the probability of making the wrong terminal decision. Let c be the cost of sampling per observation. The main reason that our studies

treat an asymptotic theory is this: while optimal designs are difficult to characterize, asymptotically optimal results (in some appropriate sense) may be available with less difficulty.

Now let us recall some fundamental results in the simplest case. The best test for sequentially deciding between two simple hypotheses

$$H_1: f(x) = f_1(x) \quad \text{and} \quad H_2: f(x) = f_2(x)$$

is the Wald sequential probability-ratio test. This test determines two numbers A and B with $B < 0 < A$ and we accept, by this test, H_2 if

$$Z_n \equiv \sum_{i=1}^n \log \frac{f_2(X_i)}{f_1(X_i)} \geq A, \quad H_1 \text{ if } Z_n \leq B, \text{ and continue sampling as long}$$

as $B < Z_n < A$. For any fixed a priori probabilities $\zeta, 1-\zeta$ for the two hypotheses, we have, asymptotically as $c \rightarrow 0$

$$\begin{cases} A \approx -\log c \\ B \approx \log c \\ \alpha \equiv \Pr\{\text{Test accepts } H_2 | H_1\} \approx c/I(f_2:f_1) \\ \beta \equiv \Pr\{\text{Test accepts } H_1 | H_2\} \approx c/I(f_1:f_2) \\ E_1(N) \approx -\log c / I(f_1:f_2) \\ E_2(N) \approx -\log c / I(f_2:f_1) \end{cases}$$

so that

$$\begin{aligned} r_1 &\equiv w_1 \alpha + c E_1(N) \approx -c \log c / I(f_1:f_2) \\ r_2 &\equiv w_2 \beta + c E_2(N) \approx -c \log c / I(f_2:f_1) \end{aligned}$$

where w_i ($i=1,2$) is the loss due to the wrong decision when H_i is the true hypothesis. The risk corresponding to the best test is mainly the cost of experimentation. The following statistical procedure is a natural and reasonable extension of the above test to the case of composite

hypotheses and multi-experiments.

(a) After n observations we compute the maximum likelihood estimator $\hat{\theta}_n$ of θ .

(b) Define a set-valued function $a(\theta)$ by

$$a(\theta) = \omega_{3-i}, \text{ if and only if } \theta \in \omega_i \quad (i=1,2),$$

Let $\tilde{\theta}_n$ be the maximum likelihood estimator of θ under the hypothesis alternative to $\hat{\theta}_n$.

(c) Compute

$$L_n \equiv \sum_{i=1}^n \log \frac{f(x_i, \hat{\theta}_n, \epsilon^i)}{f(x_i, \tilde{\theta}_n, \epsilon^i)} \quad (\geq 0),$$

where ϵ^i is the experiment selected at the i -th step and $f(x, \theta, \epsilon^i)$ is the density of the outcome x by the experiment ϵ^i .

(d) Stop sampling at the n -th observation and select the hypothesis of $\hat{\theta}_n$ if $L_n \geq -\log c$

(e) Sampling is to be continued as long as $L_n < -\log c$. If sampling is continued we select at the $(n+1)$ st step the mixed experiment η which maximizes

$$\min_{\theta' \in a(\hat{\theta}_n)} I(\hat{\theta}_n : \theta' | \eta),$$

where η is a probability measure over $\{\epsilon\}$, and if $\{\epsilon\}$ is finite and

$\eta = (\eta_1, \dots, \eta_m)$ is a probability- m vector

$$I(\theta : \theta' | \eta) = \sum_{j=1}^m \eta_j I(\theta : \theta' | \epsilon_j) = \sum_{j=1}^m \eta_j \int f(x, \theta, \epsilon_j) \log \frac{f(x, \theta, \epsilon_j)}{f(x, \theta', \epsilon_j)} d\lambda.$$

Theorem 7.4 Suppose that $\Omega = \omega_1 \cup \omega_2$ is finite and the available experiments are finite in number. Then for the above procedure

(i) the risk function $r(\theta)$ satisfies asymptotically as $c \rightarrow 0$

$$r(\theta) \approx -c \log c/I^*(\theta), \quad \text{for all } \theta,$$

where $I^*(\theta) \equiv \sup_{\eta} \inf_{\theta' \in a(\theta)} I(\theta : \theta' | \eta)$.

(ii) The procedure is asymptotically optimal in the following sense:

If an alternative procedure with risk $\bar{r}(\theta)$ satisfies

$$\lim_{c \rightarrow 0} \frac{\bar{r}(\theta_1)}{r(\theta_1)} < 1 \quad \text{for some } \theta_1$$

then

$$\lim_{c \rightarrow 0} \frac{\bar{r}(\theta_2)}{r(\theta_2)} = \infty \quad \text{for some } \theta_2 \quad (\text{Chernoff, 1959})$$

The proof will not be given here and the interested reader is referred to the original paper (Chernoff, 1959).

Example 7.2 (Chernoff, 1959)

It is desired to compare the efficacy of two drugs. The experiments \mathcal{E}_1 and \mathcal{E}_2 consist of using the first and second drugs respectively, where the outcome of these experiments has binomial densities with parameters p_1 and p_2 , respectively. The two hypotheses are

$$H_1: \theta \in \omega_1, \quad H_2: \theta \in \omega_2$$

where $\theta = (p_1, p_2)$, $\omega_1 = \{(p_1, p_2) | p_1 > p_2\}$ and $\omega_2 = \{(p_1, p_2) | p_1 \leq p_2\}$.

After n observations consisting of n_i ($i=1,2$) trials of drug i , which led to m_i successes, the maximum likelihood estimator of θ is

$$\hat{\theta}_n = (\hat{p}_{1n}, \hat{p}_{2n}) = (m_1/n_1, m_2/n_2).$$

If $m_1/n_1 > m_2/n_2$, the maximum likelihood estimator of θ under H_2 is given

by

$$\tilde{\theta}_n = (\tilde{p}_{1n}, \tilde{p}_{2n}) = \left(\frac{m_1+m_2}{n_1+n_2}, \frac{m_1+m_2}{n_1+n_2} \right).$$

Since $\frac{m_1+m_2}{n_1+n_2} = \frac{n_1}{n_1+n_2} \hat{p}_{1n} + \frac{n_2}{n_1+n_2} \hat{p}_{2n}$, $\tilde{\theta}_n$ is a weighted average of (\hat{p}_1, \hat{p}_1)

and (\hat{p}_2, \hat{p}_2) with the weights proportional to the frequencies of \mathcal{E}_1 and \mathcal{E}_2 .

We have

$$L_n \equiv \sum_{i=1}^n \log \frac{f(x_i, \hat{\theta}_n, \mathcal{E}^i)}{f(x_i, \tilde{\theta}_n, \mathcal{E}^i)} = \sum_{i=1}^n \log \left[\left(\frac{m_i}{n_i} \right)^{x_i} \left(1 - \frac{m_i}{n_i} \right)^{1-x_i} \bigg/ \left(\frac{m_1+m_2}{n_1+n_2} \right)^{x_i} \left(1 - \frac{m_1+m_2}{n_1+n_2} \right)^{1-x_i} \right]$$

$$= \sum_{j=1}^2 n_j \left[\frac{m_j}{n_j} \log \frac{m_j/n_j}{\frac{m_1+m_2}{n_1+n_2}} + \left(1 - \frac{m_j}{n_j} \right) \log \frac{1 - \frac{m_j}{n_j}}{1 - \frac{m_1+m_2}{n_1+n_2}} \right]$$

$$I((p_1, p_2) : (p'_1, p'_2) | \mathcal{E}_j) = p_j \log \frac{p_j}{p'_j} + (1-p_j) \log \frac{1-p_j}{1-p'_j} \quad (j=1,2)$$

and

$$I((p_1, p_2) : (p'_1, p'_2) | \eta) = \sum_{j=1}^2 \eta_j I((p_1, p_2) : (p'_1, p'_2) | \mathcal{E}_j)$$

$$= \zeta (p_1 \log \frac{p_1}{p'_1} + (1-p_1) \log \frac{1-p_1}{1-p'_1}) + (1-\zeta) (p_2 \log \frac{p_2}{p'_2} + (1-p_2) \log \frac{1-p_2}{1-p'_2})$$

if we set $\eta_1 = \zeta$.

Next we have to compute

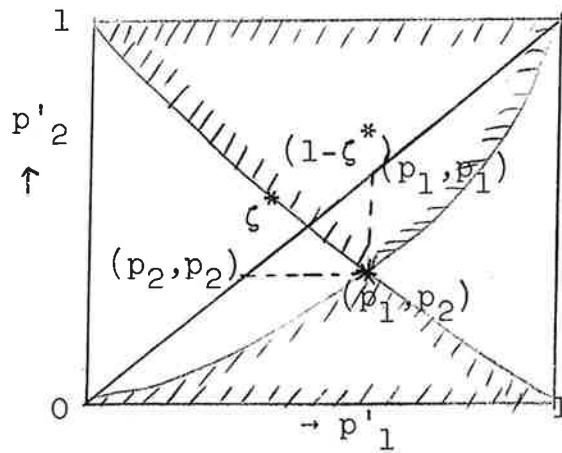
$$\max_{0 \leq \zeta \leq 1} \min_{(p'_1, p'_2) \in a((p_1, p_2))} I((p_1, p_2) : (p'_1, p'_2) | \eta).$$

In general the set $\{(p'_1, p'_2) | I((p_1, p_2) : (p'_1, p'_2) | \mathcal{E}_1) \leq I((p_1, p_2) : (p'_1, p'_2) | \mathcal{E}_2)\}$

for a fixed (p_1, p_2) is easy to characterize, and is given in Figure 7.2.

This figure also indicates the maximin strategy $\eta^* = (\zeta^*, 1-\zeta^*)$ of the statistician for the game with a payoff matrix $I((p_1, p_2) : (p'_1, p'_2) | \eta)$,

where nature is considered as the minimizing player controlling the point (p'_1, p'_2) .



The set of points (p'_1, p'_2) satisfying

$$p_1 \log \frac{p'_1}{p_1} + (1-p_1) \log \frac{1-p'_1}{1-p_1} <$$

$$p_2 \log \frac{p'_2}{p_2} + (1-p_2) \log \frac{1-p'_2}{1-p_2} .$$

Figure 7.2

In Table 7.1, we tabulate as functions of $\theta = (p_1, p_2)$.

- a) $\theta^*(\theta) = (p_1^*(\theta), p_2^*(\theta))$, the minimax strategy of nature.
- b) $\eta^*(\theta) = (\zeta^*(\theta), 1-\zeta^*(\theta))$, the maximin strategy of the statistician,
- c) $I^*(\theta) = \max_{0 \leq \zeta \leq 1} \inf_{\theta' \in a(\theta)} I(\theta : \theta' | \eta)$

$$= I(\theta : \theta^*(\theta) | \eta^*(\theta)) = I(\theta : \theta^*(\theta) | \mathcal{E}_1)$$

$$= I(\theta : \theta^*(\theta) | \mathcal{E}_2), \text{ the value of the game,}$$

and

$$d) e(\theta) \equiv \frac{\frac{1}{2} \min_{\theta' \in a(\theta)} (I(\theta : \theta' | \mathcal{E}_1) + I(\theta : \theta' | \mathcal{E}_2))}{I^*(\theta)},$$

the relative efficiency of the standard procedure which uses each drug half the time.

By symmetry we need only consider $p_1 > p_2$ and $p_1 + p_2 \leq 1$.

		p_2			
		0.05	0.10	0.20	0.40
p_1	$P_1^*(\theta)$	0.0740			
	$\zeta^*(\theta)$	0.526			
0.10	$I^*(\theta)$	0.0020			
	$e(\theta)$	0.995			
0.20		0.118	.148		
		0.547	.520		
		0.120	.0044		
		0.992	.995		
0.40		0.205	.239	.279	
		0.557	.537	.515	
		0.0429	.0277	.0105	
		0.988	.992	.999	
0.60		0.297	.333	.394	.500
		0.551	.534	.515	.500
		0.0855	.0648	.0375	.0087
		0.991	.995	.999	1.000
0.80		0.400	.438	.500	
		0.533	.517	.500	
		0.145	.120	.0837	
		0.996	.997	1.000	
0.90		0.463	.500		
		0.514	.500		
		0.187	.160		
		0.999	1.000		

Table 7.1

The results in theorem 7.4 were extended in several directions. Albert (1960) treated the case where there are infinitely many possible states of nature and Bessler (1960) attacked the problem where there are k terminal decisions. As an example we shall now reformulate a problem of the greatest mean and indicate Bessler's procedure.

Example 7.3 (Bessler, 1960)

Let $\Pi_j: N(\mu_j, 1)$ ($j=1,2,3$) be three normal populations with unknown means μ_j and common known variance 1. Let \mathcal{E}_j ($j=1,2,3$) be the experiment which consists of sampling from Π_j . It is desired to determine which population has the greatest mean.

Here the states of nature are represented by $\theta = (\mu_1, \mu_2, \dots, \mu_3)$. Let the first $n = n_1 + n_2 + n_3$ observations consist of n_j ($j=1,2,3$) observation on Π_j with sample mean \bar{X}_j . Then if $\bar{X}_1 = \max(\bar{X}_1, \bar{X}_2, \bar{X}_3)$, $a(\hat{\theta}_n) = \{\theta \mid \mu_1 \leq \mu_2 \text{ or } \mu_1 \leq \mu_3\}$ and

$$\tilde{\theta}_n = (\tilde{\mu}_{1n}, \tilde{\mu}_{2n}, \tilde{\mu}_{3n}) = \left(\frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}, \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}, \bar{X}_3 \right) \text{ or } \left(\frac{n_1 \bar{X}_1 + n_3 \bar{X}_3}{n_1 + n_3}, \bar{X}_2, \frac{n_1 \bar{X}_1 + n_3 \bar{X}_3}{n_1 + n_3} \right).$$

Thus

$$L_n = \sum_{i=1}^n \log \{ f(X_i, \hat{\theta}_n, e^i) / f(X_i, \tilde{\theta}_n, e^i) \} = \begin{cases} \frac{1}{2} \min \left\{ \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^2, \frac{n_1 n_3}{n_1 + n_3} (\bar{X}_1 - \bar{X}_3)^2 \right\}, & \text{if } \bar{X}_1 = \max(\bar{X}_1, \bar{X}_2, \bar{X}_3) \\ \dots \end{cases}$$

$$\max_{\eta} \min_{\theta' \in a(\hat{\theta}_n)} I(\hat{\theta}_n; \theta' | \eta) = \begin{cases} \frac{1}{2} \max_{\substack{\eta_j \geq 0 \\ \eta_1 + \eta_2 + \eta_3 = 1}} \min\left\{ \frac{\eta_1 \eta_2}{\eta_1 + \eta_2} (\bar{X}_1 - \bar{X}_2)^2, \frac{\eta_1 \eta_3}{\eta_1 + \eta_3} (\bar{X}_1 - \bar{X}_3)^2 \right\}, \\ \text{if } \bar{X}_1 = \max(\bar{X}_1, \bar{X}_2, \bar{X}_3) \\ \text{-----} \end{cases}$$

The maximin strategy $\eta^*(\hat{\theta}_n)$ is a function of $\gamma \equiv (\bar{X}_1 - \bar{X}_3)^2 / (\bar{X}_1 - \bar{X}_2)^2$ and we have if $\bar{X}_1 > \bar{X}_2 > \bar{X}_3$

$\eta^* \backslash \gamma$	1.0	1.5	2	10	100	∞
η_1^*	0.414	0.43	0.48	0.49	0.50	0.50
η_2^*	0.293	0.38	0.47	0.49	0.50	0.50
η_3^*	0.293	0.19	0.05	0.02	0.00	0.00

Here we see that if \bar{X}_3 is very far below \bar{X}_1 and \bar{X}_2 (i.e., $\gamma \gg 1$) all the sampling is divided between Π_1 and Π_2 which seems natural. On the otherhand if \bar{X}_2 and \bar{X}_3 are close to each other (i.e., $\gamma \doteq 1$) they share somewhat less than 60% of the sampling.

The risk of our procedure is given by

$$r(\theta) \approx -c \log c / I^*(\theta).$$

where

$$I^*(\theta) \equiv \begin{cases} \frac{1}{2} \max_{\substack{\eta_j \geq 0 \\ \eta_1 + \eta_2 + \eta_3 = 1}} \min\left\{ \frac{\eta_1 \eta_2}{\eta_1 + \eta_2} (\mu_1 - \mu_2)^2, \frac{\eta_1 \eta_3}{\eta_1 + \eta_3} (\mu_1 - \mu_3)^2 \right\}, \text{ if} \\ \text{-----} \end{cases} \quad \mu_1 = \max(\mu_1, \mu_2, \mu_3)$$

the value of which for various values of $\gamma' \equiv (\mu_1 - \mu_3)^2 / (\mu_1 - \mu_2)^2$ ($\mu_1 > \mu_2 > \mu_3$) is tabulated as follows:

γ'	1.0	1.5	2	10	100	∞
$I^*(\theta)/(\mu_1 - \mu_2)^2$	0.085	0.101	0.119	0.123	0.125	0.125

Finally this procedure can be compared with the more standard one where each population is sampled equally often. The standard procedure (with the same stopping rule as for the above one) has risk

$$\bar{r}(\theta) \approx -c \log c / \bar{I}(\theta),$$

where

$$\bar{I}(\theta) = \begin{cases} \frac{1}{2} \min\left\{ \frac{1}{6}(\mu_1 - \mu_2)^2, \frac{1}{6}(\mu_1 - \mu_3)^2 \right\} = \frac{1}{12} (\mu_1 - \mu_2)^2, & \text{if } \mu_1 > \mu_2 > \mu_3 \\ \dots \dots \dots \end{cases}$$

Hence the relative efficiency of the standard procedure is measured

by $e(\theta) \equiv \frac{r(\theta)}{\bar{r}(\theta)} = \frac{\bar{I}(\theta)}{I(\theta)}$ and is computed as a function of

$$\gamma' \equiv \frac{(\mu_1 - \mu_3)^2}{(\mu_1 - \mu_2)^2} \quad (\mu_1 > \mu_2 > \mu_3)$$

γ'	1.0	1.5	2	10	100	∞
$e(\theta)$	0.97	0.83	0.70	0.68	0.67	0.67

Finally we shall make a concluding remark. An important problem is that of modifying the procedure so that it should be good for reasonable sample sizes. From the experimentation point of view, when n is not large, $\hat{\theta}_n$ may be a relatively poor estimator of θ , although the approach described above treats $\hat{\theta}_n$ as a very good estimator of θ .

PROBLEMS 7

- (1) Prove the result stated on page 114.
- (2) Consider the sequential design problem of maximizing N independent observations, under the situation given by (7.2)
 - (a) Prove that if $f_2(X)/f_1(X)$ and $g_2(Y)/g_1(Y)$ have the same distributions under H_1 , and also under H_2 , then the stationary and "one-step optimal" design D^* : Choose $\begin{Bmatrix} X \\ Y \end{Bmatrix}$, according as $(\zeta E_{f_1} + (1-\zeta)E_{f_2})X \begin{cases} > \\ = \\ < \end{cases} (\zeta E_{g_1} + (1-\zeta)E_{g_2})Y$ is optimal.
 - (b) Using the above result find an optimal design for the situation

	\mathcal{E}_X	\mathcal{E}_Y
$(\zeta) H_1$	$N(0,1)$	$N(\mu,1)$
$(1-\zeta) H_2$	$N(\mu,1)$	$N(0,1)$

where $\mu > 0$ is a given constant.

- (3) Let the experiments \mathcal{E}_X and \mathcal{E}_Y consist in tossing a coin A and B with probabilities p and q , respectively, of obtaining heads. Assume that p and

p and q are both unknown to the experimenter. Let us consider the following two designs:

Rule D_1 : For the first toss choose \mathcal{E}_X or \mathcal{E}_Y at random. On the second stage and after, use the principle "staying on a winner", i.e., for $j = 1, 2, \dots$, if the j -th toss results in heads, stick to the same coin for the $(j+1)$ th toss, while if the j th toss results in tails, switch to the other coin for the $(j+1)$ th toss.

Rule D_0 : Choose \mathcal{E}_X or \mathcal{E}_Y , for the first toss, at random. Then stick to it for all later tosses.

For any design D let

$$L(D|p, q) \equiv \lim_{N \rightarrow \infty} (\max(p, q) - E[\frac{1}{N} \sum_{i=1}^N X_i | D; p, q])$$

be the loss by the experimenter who uses D due to ignorance of the true state of "nature". Show that D_1 is uniformly better than D_0 , or more precisely, for all $0 \leq p, q \leq 1$

$$0 \leq \delta(1-\delta/(1-\gamma)) = L(D_1 | p, q) \leq L(D_0 | p, q) = \delta,$$

where $\gamma = |p+q|/2$ and $\delta = (p-q)/2$. (Robbins, 1952).

(4) Consider the truncated- N problem for the same experiments as in the preceding problem (3) with the goal of maximizing the expected yield $E(\sum_{i=1}^N X_i | p, q)$. Assume that p and q are both unknown, but a known prior distribution $\xi(p, q)$ can be specified. Let $W_N(d\xi, D)$ denote the expected value of the sum of N observations if $\xi(p, q)$ is the prior distribution and the design D is used. Let $D^{(j)}$ be the j -th step optimal design, i.e., the design which maximizes the expected winnings over the next j plays.

(a) Show that if $N = 2$ and $\xi(p, q) = F(p)G(q)$, then the optimal design chooses \mathcal{E}_X on the first experiment, if and only if

$$\max(\mu_1 - \mu'_1, \mu_2 - \mu'_1 \mu_1) \geq \max(\mu'_1 - \mu_1, \mu'_2 - \mu'_1 \mu_1),$$

where μ_1 and μ_2 are the first two moments of F and μ'_1 and μ'_2 are those of G . Next, by using this result, show that if $dF(p) = \Phi(p)dp$, $dG(q) = \Psi(q)dq$ with continuous and positive Φ and Ψ in $(0,1)$, then there exists an N such that $D^{(1)}$ is non-optimal for the truncated- N problem.

(b) Show that there exist prior distributions $\xi(p,q)$ such that for $N = 3$, $W_3(d\xi(p,q), D^{(1)}) > W_3(d\xi(p,q), D^{(2)})$. This shows that $D^{(j)}$ is not always an improvement over $D^{(j-1)}$.

(c) "Staying on a winner" (see the problem (3)) is not always a characteristic of an optimal design: Suppose $\xi(p,q)$ concentrates probability 0.8 on $(0.1,0)$ and 0.2 on $(0.9,1)$. Then show that, for $N = 2$, the optimal design "stays on a loser". (Bradt, Johnson and Karlin, 1956)

(5) Let \mathcal{E}_X and \mathcal{E}_Y be the same experiments as in problem (3). Assume that q is known, and p is unknown. Let $d\xi$ be an a priori distribution over $0 \leq p \leq 1$, and after $s + f$ trials with \mathcal{E}_X resulting s successes, let it be modified to become

$$d\xi(s,f) = \frac{p^s(1-p)^f d\xi}{\int_0^1 p^s(1-p)^f d\xi}.$$

It is required to determine designs of sequential choice which will maximize expected number of successes during an infinite period. The discount ratio $0 < a < 1$ is introduced in order to make the sum finite and to place more emphasis on the early trials during the learning process. Now let

$W(s,f)$ = the expected discounted number of future successes, using an optimal design after s successes out of $s + f$ trials with \mathcal{E}_X .

(a) Show that we have the recurrence relation

$$W(s,f) = \max \begin{cases} X: & b(s,f)(1 + aW(s+1,f)) + (1-b(s,f))aW(s,f+1) \\ Y: & q/(1-a) \end{cases}$$

($s, f = 0, 1, 2, \dots$), where $b(s,f) \equiv \int_0^1 p d\xi(s,f)$.

(b) Prove by using successive approximation and induction that this functional equation has a unique solution, and that there exists a function $\hat{p}(f,s)$ uniquely such that an optimal design is given by:

$$\text{choose } \begin{cases} X \\ Y \end{cases}, \text{ as } \hat{p}(f,s) \begin{cases} \geq \\ < \end{cases} q. \quad (\text{Bellman, 1956}).$$

(6) Do the results in Example 7.1 still hold true if we use the uncertainty function $U(\xi) = - \sum_1^2 \xi_1 \log \xi_1$?

8. Capacity of statistical experiments.

Let $\mathcal{e} = [(\mathcal{X}, \mathcal{Q}), \{f_1(x), \dots, f_k(x)\}]$ be a finite experiment with input space $\mathcal{Q} = \{1, \dots, k\}$. In Section 1.5 we defined the amount of information provided by the experiment \mathcal{e} by

$$(8.1) \quad I(\mathcal{e}, \xi) = \sum_{i=1}^k \xi_i \int f_i(x) \log \frac{f_i(x)}{\sum_{i=1}^k \xi_i f_i(x)} d\lambda = \sum_{i=1}^k \xi_i I(f_i; \sum_{i=1}^k \xi_i f_i),$$

where $\xi \in \mathbb{E}^k$ is the a priori probability k -vector representing the prior knowledge over the input space \mathcal{Q} . $I(\mathcal{e}, \xi)$ is continuous and, by Theorem 5.3, concave in the prior knowledge. Let us define the capacity of an experiment \mathcal{e} by the maximum information for all possible prior distributions:

$$(8.2) \quad C(\mathcal{e}) = C(f_1, \dots, f_k) = \max_{\xi \in \mathbb{E}^k} I(\mathcal{e}, \xi).$$

Then we have

Theorem 8.1 If there exists a $\xi^* \in \mathbb{E}^k$ with $\xi_i^* > 0$ ($i=1, \dots, k$) such that

$$(8.3) \quad I(f_i; \sum_{i=1}^k \xi_i^* f_i) = \underline{\text{indep. of } i}$$

then ξ^* maximizes the information $I(\mathcal{e}, \xi)$ and

$$C(f_1, \dots, f_k) = I(\mathcal{e}, \xi^*) = I(f_i; \sum_{i=1}^k \xi_i^* f_i).$$

Proof. By the information identity for mixtures of prior knowledges (see (5.14)), for any $\xi, \nu \in \mathbb{E}^k$ and $0 \leq t \leq 1$ we have

$$I(\mathcal{e}, t\xi + (1-t)\nu) = I(\mathcal{e}^*, t) + tI(\mathcal{e}, \xi) + (1-t)I(\mathcal{e}, \nu),$$

where

$$\mathcal{e}^* = [(\mathcal{X}, \mathcal{B}), \{ \sum_{i=1}^k \xi_i f_i(x), \sum_{i=1}^k \nu_i f_i(x) \}],$$

$$I(\mathcal{e}^*, t) = t I \left(\sum_{i=1}^k \xi_i f_i : \sum_{i=1}^k (t \xi_i + (1-t) \nu_i) f_i \right) + (1-t) I \left(\sum_{i=1}^k \nu_i f_i : \sum_{i=1}^k (t \xi_i + (1-t) \nu_i) f_i \right).$$

Hence $I(\mathcal{e}, \xi)$ is strictly concave in ξ , so that if the Lagrange equations in the usual calculus of variations yield a solution which is not on the boundary of Ξ^k , it provides the maximum. Differentiating

$$\sum_{i=1}^k \xi_i \int f_i(x) \log(f_i(x) / \sum_{i=1}^k \xi_i f_i(x)) d\lambda - \iota \sum_{i=1}^k \xi_i,$$

where ι is the Lagrangian multiplier, with respect to $\xi_i (i=1, \dots, k)$ and equating to zero, we obtain (8.3).

Theorem 8.2 Let $p_{i.} = (p_{i1}, p_{i2}, \dots, p_{ik}) (i=1, \dots, k)$ be k probability k -vectors which are linearly independent. Suppose that they define an experiment \mathcal{e} . Then the maximum value of $I(\mathcal{e}, \xi)$ for variations of $\xi \in \Xi$ occurs for $(\xi_1^*, \dots, \xi_k^*)$ such that

$$(8.4) \quad \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = (p_{ij})^{-1} \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_k \end{bmatrix},$$

$$(8.5) \quad (\xi_1^*, \dots, \xi_k^*) (p_{ij}) = \left(\frac{e^{-b_1}}{\sum_i e^{-b_i}}, \dots, \frac{e^{-b_k}}{\sum_i e^{-b_i}} \right),$$

where $H_i = - \sum_{j=1}^k p_{ij} \log p_{ij} (i=1, \dots, k)$.

The capacity $C(\mathcal{e})$ is given by $\log \left(\sum_{i=1}^k e^{-b_i} \right)$.

Proof. $I(\theta, \xi) = \sum_i \xi_i \sum_j p_{ij} \log \frac{p_{ij}}{\sum_i \xi_i p_{ij}}$

$$= - \sum_j (\sum_i \xi_i p_{ij}) \log (\sum_i \xi_i p_{ij}) - \sum_i \xi_i H_i.$$

Since $H_i = \sum_j p_{ij} b_j$ from (8.4), we get by setting $\eta_j = \sum_i \xi_i p_{ij}$

$$I(\theta, \xi) = \sum_j \eta_j (-\log \eta_j - b_j).$$

The following lemma, combined with (8.5) yields the statement of the theorem

Lemma If b_1, \dots, b_k are given real numbers, then the maximum of

$$\sum_{j=1}^k \eta_j (-\log \eta_j - b_j)$$

for variations of $\eta \in \mathbb{R}^k$ is attained by

$$\eta_j^* = e^{-b_j} / \sum_{j=1}^k e^{-b_j} \quad (j=1, \dots, k),$$

and the maximum value is $\log (\sum_j e^{-b_j})$.

Proof. We have by Theorem 2.1

$$0 \leq \sum_{j=1}^k \eta_j \log \frac{\eta_j}{e^{-b_j} (\sum_j e^{-b_j})^{-1}} = \sum_j \eta_j (\log \eta_j + b_j) + \log (\sum_j e^{-b_j}).$$

Example 8.1 (Sakaguchi, 1959)

For binomial dichotomous experiments represented by a stochastic matrix

$$\begin{pmatrix} p_1 & 1-p_1 \\ p_2 & 1-p_2 \end{pmatrix}, \quad \text{with } p_1 \neq p_2,$$

we solve the simultaneous equations

$$\begin{cases} p_1 b_1 + (1-p_1)b_2 = H_1 (\equiv -p_1 \log p_1 - (1-p_1) \log(1-p_1)) \\ p_2 b_1 + (1-p_2)b_2 = H_2 (\equiv -p_2 \log p_2 - (1-p_2) \log(1-p_2)) \end{cases}$$

and

$$\begin{cases} p_1 \xi_1^* + p_2 \xi_2^* = e^{-b_1} / (e^{-b_1} + e^{-b_2}) \\ \xi_1^* + \xi_2^* = 1. \end{cases}$$

From these equations we obtain

$$\begin{cases} b_1 = \frac{(1-p_2)H_1 - (1-p_1)H_2}{p_1 - p_2} \\ b_2 = \frac{-p_2 H_1 + p_1 H_2}{p_1 - p_2} \end{cases}$$

and

$$\xi_1^* = 1 - \xi_2^* = \frac{1}{p_1 - p_2} \left(\frac{e^{-b_1}}{e^{-b_1} + e^{-b_2}} - p_2 \right),$$

respectively. We then have $C(\mathcal{E}) = \log(e^{-b_1} + e^{-b_2})$. The values of ξ_1^* and $C(\mathcal{E})$ in bits to five decimal places for p_1, p_2 in increments of 0.01 are tabulated by Phillips (1962).

The capacity of finite experiments measures, in some sense, how divergent k probability densities f_1, \dots, f_k are. In fact, if $f_1(x) = \dots = f_k(x)$ [λ], then we have $C(f_1, \dots, f_k) = 0$. And the capacity of an experiment is invariant under any permutations (i_1, \dots, i_k) of $(1, \dots, k)$.

Theorem 8.3 Let $Q = (q_{ij} | i=1, \dots, m; j=1, \dots, k)$ be any stochastic matrix, that is, $q_{ij} \geq 0$ and $\sum_{j=1}^k q_{ij} = 1$. Then we have

$$C\left(\sum_{j=1}^k q_{1j} f_j, \dots, \sum_{j=1}^k q_{mj} f_j\right) \leq C(f_1, \dots, f_k). \quad (\text{Sakaguchi, 1957})$$

Proof. Every row vector $q_{i\cdot}$ of Q can be considered as a prior knowledge for the experiment $\mathcal{E} = [(\mathcal{X}, \mathcal{B}), \{f_1(x), \dots, f_k(x)\}]$. Consider the mixture $\sum_{i=1}^m \alpha_i q_{i\cdot}$ ($\alpha \in \mathbb{E}^m$) of m prior knowledges $q_{1\cdot}, \dots, q_{m\cdot}$. By the information identity for mixtures of prior knowledges (see(5.14)) we have

$$(8.6) \quad I(\mathcal{E}, \sum_{i=1}^m \alpha_i q_{i\cdot}) = I(\mathcal{E}^*, \alpha) + \sum_{i=1}^m \alpha_i I(\mathcal{E}, q_{i\cdot})$$

where

$$\mathcal{E}^* = [(\mathcal{X}, \mathcal{B}), \{ \sum_{j=1}^k q_{1j} f_j(x), \dots, \sum_{j=1}^k q_{mj} f_j(x) \}],$$

$$I(\mathcal{E}^*, \alpha) = \sum_{i=1}^m \alpha_i I\left(\sum_{j=1}^k q_{1j} f_j; \sum_{i=1}^m \alpha_i \sum_{j=1}^k q_{ij} f_j\right).$$

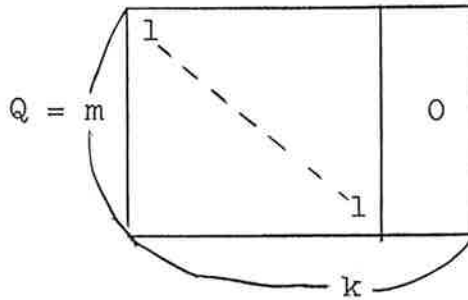
Dropping the second term of the right-hand side of the expression (8.6) and taking the supremum in $\alpha \in \mathbb{E}^m$ we get

$$\begin{aligned} C\left(\sum_{j=1}^k q_{1j} f_j, \dots, \sum_{j=1}^k q_{mj} f_j\right) &= \sup_{\alpha} I(\mathcal{E}^*, \alpha) \leq \sup_{\alpha} I(\mathcal{E}, \sum_{i=1}^m \alpha_i q_{i\cdot}) \\ &\leq \sup_{\xi} I(\mathcal{E}, \xi) = C(f_1, \dots, f_k). \end{aligned}$$

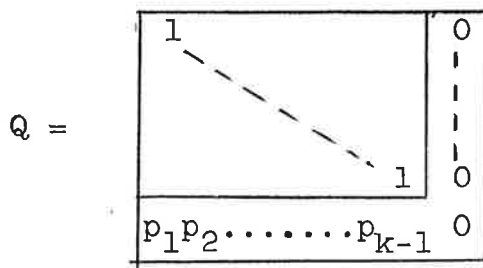
Corollary to theorem 8.3. (i) If $m \leq k$, then $C(f_1, \dots, f_m) \leq C(f_1, \dots, f_k)$.

(ii) If $p \in \mathbb{E}^{k-1}$, then $C(f_1, \dots, f_{k-1}, \sum_{i=1}^{k-1} p_i f_i) = C(f_1, \dots, f_{k-1})$.

Proof. (i) is easily seen by taking



To prove (ii) we consider the equation (8.6) with $m = k$ and



Thus we have

$$I(\epsilon; \alpha_1 + p_1 \alpha_n, \dots, \alpha_{k-1} + p_{k-1} \alpha_n, 0) = I(\epsilon^*, \alpha) + \alpha_n I(\epsilon; p_1, \dots, p_{k-1}, 0)$$

where

$$\epsilon^* = [(\mathcal{X}, \mathcal{B}), \{f_1(x), \dots, f_{k-1}(x), \sum_{i=1}^{k-1} p_i f_i(x)\}].$$

Dropping the second term in the right-hand side and taking the supremum in $\alpha \in \mathbb{E}^k$ we obtain

$$C(f_1, \dots, f_{k-1}, \sum_{i=1}^{k-1} p_i f_i) \leq C(f_1, \dots, f_{k-1}).$$

Since the reverse inequality is also true by (i) we have the desired result.

For experiments with $k = 2$ (dichotomous experiments) we have

Theorem 8.4 Let $f_1(x)$ and $f_2(x)$ be given generalized probability densities. Let

$$G \equiv \{g(x) \mid g(x) \geq 0[\lambda], \int g(x)d\lambda = 1, I(f_1:g) = I(f_2:g)\}.$$

Then we have

$$\min_{g \in G} I(f_2:g) = C(f_1;f_2),$$

and the minimizing $g(x)$ is given by

$$(8.7) \quad g(x) = \tilde{g}(x) \equiv \tilde{t}f_1(x) + (1-\tilde{t})f_2(x),$$

where $0 < \tilde{t} < 1$ is the unique root of the equation

$$(8.8) \quad I(f_1:\tilde{t}f_1 + (1-\tilde{t})f_2) = I(f_2:\tilde{t}f_1 + (1-\tilde{t})f_2).$$

Moreover we have

$$\begin{cases} J(f_2, \tilde{g}) \equiv I(f_2:\tilde{g}) + I(\tilde{g}:f_2) = \tilde{t}I(f_1:f_2) \\ J(f_1, \tilde{g}) \equiv I(f_1:\tilde{g}) + I(\tilde{g}:f_1) = (1-\tilde{t})I(f_2:f_1). \end{cases} \quad (\text{Sakaguchi, 1961})$$

Proof. It is easy to see that we have to consider the problem of maximizing the integral $\int f_2(x) \log g(x) d\lambda$ under the constraints that $\int g(x) d\lambda = 1$ and

$$-\int (f_2(x) - f_1(x)) \log g(x) d\lambda = H_2 - H_1$$

where

$$H_i \equiv -\int f_i(x) \log f_i(x) d\lambda \quad (i=1,2).$$

The usual technique of Lagrangian multipliers leads to the equations (8.7) and (8.8). Consider the function

$$\varphi(t) \equiv \int (f_1 - f_2) \log(tf_1 + (1-t)f_2) d\lambda.$$

This function is strictly increasing in t and satisfies $\varphi(0) < H_2 - H_1 < \varphi(1)$

and $H_2 - H_1 = \varphi(0) + I(f_1:f_2) = \varphi(1) - I(f_2:f_1).$

Hence the equation $\varphi(t) = H_2 - H_1$ has a unique root \tilde{t} with $0 < \tilde{t} < 1$. This completes the proof.

Another proof is as follows (A private communication to the author from Kullback): If $0 \leq t_1 = 1 - t_2 \leq 1$, then for any $g(x) \in G$ we have

$$\begin{aligned} \sum_{i=1}^2 t_i I(f_i : \sum_{i=1}^2 t_i f_i) &= \sum_{i=1}^2 t_i I(f_i : g) - I(\sum_{i=1}^2 t_i f_i : g) \\ &= I(f_2 : g) - I(\sum_{i=1}^2 t_i f_i : g). \end{aligned}$$

Since there exists a unique $0 < t_1 = \tilde{t} < 1$, such that $\varphi(\tilde{t}) = H_2 - H_1$, or equivalently, $I(f_1 : \sum_{i=1}^2 \tilde{t}_i f_i) = I(f_2 : \sum_{i=1}^2 \tilde{t}_i f_i)$, we have from Theorem 8.1

$$C(f_1, f_2) = I(f_2 : g) - I(\sum_{i=1}^2 \tilde{t}_i f_i : g)$$

$$\therefore C(f_1, f_2) \leq \min_{g \in G} I(f_2 : g)$$

But the equality is attained by $g = \tilde{t}f_1 + (1-\tilde{t})f_2 \in G$. Hence

$$C(f_1, f_2) = \min_{g \in G} I(f_2 : g).$$

Reversing the direction of the pseudo-distance $I(f:g)$, we obtain the following theorem which has already been proved in Section 2. For reference we shall state it here again.

Corollary 2 to theorem 2.7 Let $f_i(x)$ ($i=1,2$) be given p.d.f.'s and let

$$G \equiv \{g(x) \mid g(x) \geq 0[\lambda], \int g(x) d\lambda = 1, I(g:f_1) = I(g:f_2)\}$$

Then we have

$$(8.9) \quad \min_{g \in G} I(g:f_2) = -\log F(t^*)$$

(i.e., the Chernoff information number for deciding between two

densities f_1 and f_2), where

$$F(t) \equiv \int [f_1(x)]^t [f_2(x)]^{1-t} d\lambda$$

and t^* is determined by the equation

$$(8.10) \quad F'(t^*) = \int [f_1(x)]^{t^*} [f_2(x)]^{1-t^*} \log \frac{f_1(x)}{f_2(x)} d\lambda = 0.$$

The minimizing p.d.f. is given by

$$(8.11) \quad g(x) = g_*(x) = [f_1(x)]^{t^*} [f_2(x)]^{1-t^*} / F(t^*)$$

and moreover we have

$$\begin{cases} J(g_*, f_2) \equiv I(g_*:f_2) + I(f_2:g_*) = t^* I(f_2:f_1) \\ J(g_*, f_1) \equiv I(g_*:f_1) + I(f_1:g_*) = (1-t^*) I(f_1:f_2). \end{cases}$$

It is clear that the preceding two theorems can be generalized to the case of $k \geq 3$. We shall state, in the following, corresponding theorems without proof.

Theorem 8.4' Let $f_1(x), \dots, f_k(x)$ be given generalized probability densities. Let

$$G \equiv \{g(x) | g(x) \geq 0[\lambda], \int g(x) d\lambda = 1, I(f_1:g) = \dots, I(f_k:g)\}.$$

Assume that there exists a $(\tilde{t}_1, \dots, \tilde{t}_k) \in \mathbb{E}^k$ with $\tilde{t}_i > 0$ ($i=1, \dots, k$)

such that

$$I(f_1: \sum_1^k \tilde{t}_i f_i) = \dots = I(f_k: \sum_1^k \tilde{t}_i f_i).$$

Then we have

$$\min_{g \in G} I(f_k:g) = C(f_1, \dots, f_k)$$

and the minimizing $g(x)$ is given by

$$g(x) = \tilde{g}(x) = \sum_1^k \tilde{t}_i f_i(x).$$

Moreover we have

$$J(f_i, \tilde{g}) \equiv I(f_i : \tilde{g}) + I(\tilde{g} : f_i) = \sum_{j \neq i} \tilde{t}_j I(f_j : f_i), \quad (i=1, \dots, k).$$

Corollary 2' to theorem 2.7 Let $f_i(x)$ ($i=1, \dots, k$) be given p.d.f.'s and
let

$$G \equiv \{g(x) \mid g(x) \geq 0, \int g(x) d\lambda = 1, I(g : f_1) = \dots = I(g : f_k)\}.$$

Then we have

$$\min_{g \in G} I(g : f_k) = -\log F(t_1^*, \dots, t_k^*),$$

where

$$F(t_1, \dots, t_k) \equiv \int [f_1(x)]^{t_1} \dots [f_k(x)]^{t_k} d\lambda$$

and t_1^*, \dots, t_k^* are determined by

$$\left\{ \begin{array}{l} \frac{\partial F}{\partial t_j}(t_1^*, \dots, t_k^*) = \int [f_1(x)]^{t_1^*} \dots [f_k(x)]^{t_k^*} \log \frac{f_j(x)}{f_k(x)} d\lambda = 0 \quad (j=1, \dots, k-1) \\ \sum_{j=1}^k t_j^* = 1. \end{array} \right.$$

The minimizing p.d.f. is given by

$$g(x) = g_*(x) \equiv [f_1(x)]^{t_1^*} \dots [f_k(x)]^{t_k^*} / F(t_1^*, \dots, t_k^*),$$

and moreover we have

$$J(g_*, f_i) \equiv I(g_* : f_i) + I(f_i : g_*) = \sum_{j \neq i} t_j^* I(f_i : f_j), \quad (i=1, \dots, k).$$

We might call the two kinds of densities $\tilde{g}(x)$ in Theorem 8.4 and $g_*(x)$ in Corollary 2 to Theorem 2.7, as the 'average' or the 'centre of gravity' of two densities f_1 and f_2 . If f_1 and f_2 belong to the same type of the

exponential family of distributions:

$$(8.12) \quad f_i(x) = f(x|\theta_i) = e^{\theta_i x} h(x)/M(\theta_i) \quad (i=1,2),$$

where $h(x) \geq 0$ and $M(\theta) \equiv \int e^{\theta x} h(x) d\lambda$, then we can easily compute the 'average' density $g_*(x)$ by (8.10) and (8.11). This is found to be a density with the same type of the exponential family with a weighted-average parameter-value θ^* :

$$(8.13) \quad g_*(x) = f(x|\theta_*) = e^{\theta^* x} h(x)/M(\theta^*)$$

where θ^* is the unique root of the equation

$$w'(\theta^*) = \frac{w(\theta_2) - w(\theta_1)}{\theta_2 - \theta_1},$$

in which $w(\theta) \equiv \log M(\theta)$, and can be represented by

$$(8.14) \quad \theta^* = \frac{\theta^* - \theta_2}{\theta_1 - \theta_2} \theta_1 + \frac{\theta_1 - \theta^*}{\theta_1 - \theta_2} \theta_2 \equiv t^* \theta_1 + (1-t^*) \theta_2.$$

This weight t^* is the number just introduced in equations (8.10) and (8.11) and we easily see that $0 < t^* < 1$. The minimum distance, or equivalently, the Chernoff information number is, by (8.9), equal to

$$I(g_*:f_1) \equiv I(g_*:f_2) = \frac{\theta^* - \theta_2}{\theta_1 - \theta_2} w(\theta_1) + \frac{\theta_1 - \theta^*}{\theta_1 - \theta_2} w(\theta_2) - w(\theta^*).$$

On the other hand another kind of 'average' density, $\tilde{g}(x)$, which appeared in Theorem 8.4 cannot be found explicitly even in the case of densities in the exponential family. The difficulty lies mainly in the fact that even if f_1 and f_2 belong to the same type of exponential family of distributions the average density $tf_1 + (1-t)f_2$ does not generally belong to the exponential family.

If, for the two exponential densities (8.12), a unique number $0 < \bar{t} < 1$ can be found such that

$$(8.15) \quad \int f(x|\theta_1) \log \frac{f(x|\theta_1)}{f(x|\theta_2)} d\lambda = \int f(x|\theta_2) \log \frac{f(x|\theta_2)}{f(x|\theta_1)} d\lambda,$$

where $\bar{\theta} = \bar{t}\theta_1 + (1-\bar{t})\theta_2$, then the common amount of information is defined as the pseudo-capacity of the dichotomous experiment composed of $f(x|\theta_1)$ and $f(x|\theta_2)$ and is denoted by $C'(\theta_1, \theta_2)$.

Theorem 8.5 The pseudo-capacity of the dichotomous experiments for the exponential family is given by

$$(8.16) \quad C'(\theta_1, \theta_2) = \omega(\bar{\theta}) - \omega(\theta_1) - (\bar{\theta} - \theta_1)\omega'(\theta_1),$$

where $\omega(\theta) \equiv \log M(\theta)$ and

$$(8.17) \quad \begin{aligned} \bar{\theta} &= \frac{(\theta_1 \omega'(\theta_1) - \theta_2 \omega'(\theta_2)) - (\omega(\theta_1) - \omega(\theta_2))}{\omega'(\theta_1) - \omega'(\theta_2)} \\ &= \frac{I(f(x|\theta_1):f(x|\theta_2))\theta_1 + I(f(x|\theta_2):f(x|\theta_1))\theta_2}{J(f(x|\theta_1), f(x|\theta_2))} \end{aligned}$$

Moreover we have the limiting value

$$\lim_{\theta_1, \theta_2 \rightarrow \theta} \frac{C'(\theta_1, \theta_2)}{(\theta_1 - \theta_2)^2} = \frac{1}{8} \omega''(\theta) = \frac{1}{8} E_{\theta} \left[- \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right].$$

(Sakaguchi, 1959)

Proof. Since we have

$$I(f(x|\theta_1):f(x|\theta_2)) = \omega(\theta) - \omega(\theta_1) - (\theta - \theta_1)\omega'(\theta_1)$$

by straightforward calculation, we have

$$\omega(\bar{\theta}) - \omega(\theta_1) - (\bar{\theta} - \theta_1)\omega'(\theta_1) = \omega(\bar{\theta}) - \omega(\theta_2) - (\bar{\theta} - \theta_2)\omega'(\theta_2),$$

from which we get the left half of (8.17). To see that $\bar{\theta}$ can be written

$\bar{\theta} = \bar{t}\theta_1 + (1-\bar{t})\theta_2$ with $0 < \bar{t} < 1$, we use the fact that for the exponential family

$$w''(\theta) = E_{\theta} [(X - E_{\theta}X)^2] > 0, \text{ for all } \theta.$$

It follows that

$$(\bar{\theta}-\theta_1)(\bar{\theta}-\theta_2) = \frac{\{\omega(\theta_1)-\omega(\theta_2)-(\theta_1-\theta_2)\omega'(\theta_2)\}\{\omega(\theta_1)-\omega(\theta_2)-(\theta_1-\theta_2)\omega'(\theta_1)\}}{(\omega'(\theta_1) - \omega'(\theta_2))^2} < 0.$$

Since

$$\begin{aligned} J(f(x|\theta_1), f(x|\theta_2)) &= I(f(x|\theta_1):f(x|\theta_2)) + I(f(x|\theta_2):f(x|\theta_1)) \\ &= (\theta_1-\theta_2)(\omega'(\theta_1)-\omega'(\theta_2)), \end{aligned}$$

we obtain the right half of (8.17).

To yield the limiting value in the theorem we may note that

$$\frac{\bar{\theta}-\theta_2}{\theta_1-\theta_2} = \frac{\omega(\theta_2)-\omega(\theta_1)-(\theta_2-\theta_1)\omega'(\theta_1)}{(\theta_1-\theta_2)(\omega'(\theta_1)-\omega'(\theta_2))} \xrightarrow{(\theta_1, \theta_2 \rightarrow \theta)} \frac{1}{2}.$$

Thus we have

$$\begin{aligned} \frac{C'(\theta_1, \theta_2)}{(\theta_1-\theta_2)^2} &= \frac{\omega(\bar{\theta})-\omega(\theta_1)-(\bar{\theta}-\theta_1)\omega'(\theta_1)}{(\theta_1-\theta_2)^2} \\ &= \frac{1}{2} \left(\frac{\bar{\theta}-\theta_1}{\theta_1-\theta_2} \right)^2 [\omega''(\theta)]_{\theta_1 \leq \theta \leq \bar{\theta}} \xrightarrow{(\theta_1, \theta_2 \rightarrow \theta)} \frac{1}{8} \omega''(\theta). \end{aligned}$$

It should be remarked here that the parameter values θ^* in (8.13) and (8.14), and $\bar{\theta}$ in (8.16) and (8.17) are both weighted averages of θ_1 and θ_2 but they do not always coincide. The situation may be seen from Fig. 8.1.

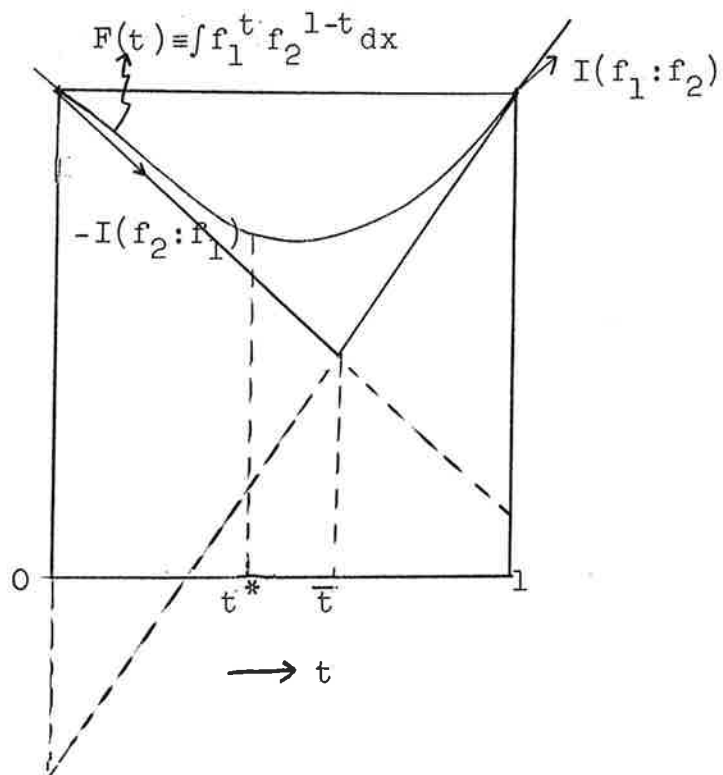


Figure 8.1

The two values t^* and \bar{t} do not always coincide,

where $\theta^* = t^* \theta_1 + (1-t^*) \theta_2$

$$\bar{\theta} = \bar{t} \theta_1 + (1-\bar{t}) \theta_2$$

in (8.14) and (8.17) respectively and

$$\bar{t} \equiv I(f_1:f_2) / (I(f_1:f_2) + I(f_2:f_1)).$$

We shall list several computed examples of the foregoing discussions in Table 8.1

Table 8.1

$f_i(x) (i=1,2)$	$g_*(x) \equiv [f_1(x)]^{t^*} [f_2(x)]^{1-t^*} / F(t^*)$	$-\log F(t^*)$
Binomial: $p_i^x q_i^{1-x}$	$(p^*)^x (q^*)^{1-x}$ where $p^* \equiv \log \frac{q_1}{q_2} / \left(\log \frac{q_1}{q_2} + \log \frac{p_1}{p_2} \right)$	$p^* \log \frac{p^*}{p_1} +$ $q^* \log \frac{q^*}{q_1}$
Poisson: $e^{-\mu_i} \mu_i^x / (x!)$	$e^{-\mu^*} (\mu^*)^x / (x!)$ where $\mu^* = \frac{\mu_1 - \mu_2}{\log \mu_1 - \log \mu_2}$	$\mu^* \log \frac{\mu^*}{\mu_1} -$ $(\mu^* - \mu_1)$
Normal: $(2\pi)^{-1/2} e^{-(x-\mu_i)^2/2}$	$(2\pi)^{-1/2} \exp\{-(x-\mu^*)^2/2\}$ where $\mu^* = (\mu_1 + \mu_2)/2$	$(\mu_1 - \mu_2)^2/8$
Exponential: $\beta_i^{-1} e^{-x/\beta_i}$	$(\beta^*)^{-1} e^{-x/\beta^*}$ where $\beta^* = \frac{1}{\beta_1^{-1} - \beta_2^{-1}} \log(\beta_1^{-1}/\beta_2^{-1})$	$-\log \frac{\beta^*}{\beta_1} -$ $\left(1 - \frac{\beta^*}{\beta_1}\right)$

(Table 8.1 continued on next page)

Table 8.1 (Continued)

θ_i	$\bar{\theta}$	Pseudo-capacity
$\log \frac{p_i}{q_i}$	$-\frac{H_1 - H_2}{p_1 - p_2}$ where $H_i = p_i \log \frac{1}{p_i} + q_i \log \frac{1}{q_i}$	$\log(e^{-X_1} + e^{-X_2})$ where $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} p_1 q_1 \\ p_2 q_2 \end{pmatrix}^{-1} \begin{pmatrix} H_1 \\ H_2 \end{pmatrix}$
$\log \mu_i$	$\frac{\mu_1 \log \mu_1 - \mu_2 \log \mu_2}{\mu_1 - \mu_2} - 1$	$\mu_1 \log \frac{\mu_1}{e^{\bar{\theta}}} - (\mu_1 - e^{\bar{\theta}})$
μ_i	$(\mu_1 + \mu_2)/2$	$(\mu_1 - \mu_2)^2/8$
$-\beta_i^{-1}$	$\frac{-1}{\beta_1 - \beta_2} \log \frac{\beta_1}{\beta_2}$	$\log \left(\frac{-1}{\bar{\theta} \beta_1} \right) - (\bar{\theta} \beta_1 + 1)$

The notion of the capacity of statistical experiments can be extended in two directions. One is the case where the experiment is not finite and the other is the case in which the information is measured by the generally-defined information $I[\mathcal{E}, \xi; U] = U(\xi) - E[U(\xi(X)) | \xi]$, using a concave uncertainty function $U(\xi)$ other than $U(\xi) = -\sum_{i=1}^k \xi_i \log \xi_i$.

We shall show two examples in these directions in the following:

Example 8.2 Consider the normal experiment

$$\mathcal{E}(\sigma) = [(\chi, \beta), \{p(x|\theta) | -\infty < \theta < \infty\}], \text{ where } p(x|\theta) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

If we take $p(\theta) = \frac{1}{\sqrt{2\pi} v} \exp\left\{-\frac{(\theta-m)^2}{2v^2}\right\}$, then we have by Example 5.1,

$$I(\mathcal{E}(\sigma), p(\theta)) = \iint p(\theta) p(x|\theta) \log \frac{p(x|\theta)}{p(x)} d\theta dx = \frac{1}{2} \log \left(1 + \frac{v^2}{\sigma^2}\right).$$

Let $P = \{ \text{p.d.f. } p(\theta) | p(\theta) = \frac{1}{\sqrt{2\pi} v} \exp\left\{-\frac{(\theta-m)^2}{2v^2}\right\}; -\infty < m < \infty; 0 < v \leq V \}$.

If we limit our attention to the prior distributions $p(\theta)$ in P only, then we have

$$\max_{p(\theta) \in P} I(\mathcal{E}(\sigma), p(\theta)) = \frac{1}{2} \log \left(1 + \frac{V^2}{\sigma^2}\right).$$

This fact is widely known and used in communication engineering in which σ^2 and V^2 are understood as noise power and signal power, respectively. The capacity of the Gaussian channel is strictly decreasing in noise power and increasing in signal power.

Example 8.3 Let $\mathcal{E} = [X, \mathcal{B}], \{f_1(x), f_2(x)\}$ be a dichotomous experiment. If we adopt the concave uncertainty function $U(\xi) = \min(\xi_1, 1 - \xi_1)$, then we have

$$\begin{aligned} I[\mathcal{E}, \xi; U] &= U(\xi) - E[(U(\xi(X))) | \xi] \\ &= \min(\xi_1, \xi_2) - \left(\begin{array}{l} \xi_1 \int f_1(x) d\lambda + \xi_2 \int f_2 d\lambda \\ \frac{f_2}{f_1} > \frac{\xi_1}{\xi_2} \qquad \frac{f_2}{f_1} \leq \frac{\xi_1}{\xi_2} \end{array} \right), \end{aligned}$$

which expresses the reduction of the risk of Bayes decision rule deciding between two densities f_1 and f_2 with usual zero-one loss. We can show that the above information attains the maximum at $\xi_1 = \xi_2 = \frac{1}{2}$ and

the capacity is equal to $\frac{1}{2} \int_{f_2 \leq f_1} (f_1 - f_2) d\lambda$.

The proof is as follows: By a well-known theorem in statistical decision theory the Bayes risk relative to the prior distribution $\xi_1, 1 - \xi_1$ is a continuous and concave function of ξ_1 . We have

$$\begin{aligned} I[\mathcal{C}, \xi; U] &= \min (\xi_1 - E[U(\xi(X)) | \xi], 1 - \xi_1 - E[U(\xi(X)) | \xi]) \\ &= \min (h_1(\xi_1), h_2(\xi_1)), \quad \text{say.} \end{aligned}$$

Since the minimum of two convex functions has the maximum at some endpoint or at the point at which the two convex functions have equal values, and since $h_1(0) = h_2(1) = 0$, we have the desired result.

Appendix A

The role of sufficient statistics in statistical decision theory.

The following theorem expresses the fundamental role of sufficient statistics in modern statistical decision theory. Let p_1, \dots, p_m be m probability distributions defined on the same measurable space. Let us assume that we have to make some statistical decision on the basis of observations of random variables x and y and a given weight function $w(i, d)$. For simplicity we shall assume that the random variables are discrete and finite. We shall denote by δ_{xy} the randomized decision functions based on x and y , and similarly we shall denote by d_{xy} the non-randomized decision functions based on x and y . We shall say that a class of d.f.'s $\{\delta_x\}$ is uniformly complete in the class of d.f.'s $\{\delta_{xy}\}$, if and only if $\{\delta_x\}$ is complete in $\{\delta_{xy}\}$ with respect to any weight function $w(i, d)$.

Theorem. $\{\delta_x\}$ is uniformly complete in $\{\delta_{xy}\}$, if and only if x is sufficient for y , that is, the conditional probabilities $p_i(y|x)$ ($i=1, \dots, m$) are independent of i .

(Elfving, 1952)

In order to prove this theorem we shall need the following three lemmas.

Lemma 1. If x is sufficient for y , then for every a priori distribution ξ , the Bayes decision function with respect to it has the form d_x .

Proof. The average risk of a randomized d.f. δ_{xy} when the a priori

distribution is ξ is written as

$$\begin{aligned} r(\xi, \delta_{xy}) &= \sum_i \xi_i \sum_{x,y} p_i(x,y) \sum_d \delta_d(x,y) w(i,d) \\ &= \sum_{x,y} p(y|x) \sum_d \delta_d(x,y) \left\{ \sum_i \xi_i p_i(x) w(i,d) \right\}. \end{aligned}$$

In order to minimize the above expression we have to assign, for each x and y , the probability 1 to that d which minimizes the expression $\{ \dots \}$.

Lemma 2 Let (a_{ij}) be a matrix game. We consider a mixed strategy η of the minimizing player and fix it. Let B be the set of all available pure strategies of the minimizing player. If a subset B^* of B exist such that

$$\min_{j \in B^*} \sum_i \xi_i (a_{ij} - \alpha_i) = \min_{j \in B} \sum_i \xi_i (a_{ij} - \alpha_i) \quad \text{for all } \xi,$$

where $\alpha_i \equiv E(i, \eta) \equiv \sum_{j \in B} a_{ij} \eta_j$, then for every minimax strategy η_0^* of the matrix game $(a_{ij} - \alpha_i)$, when the minimizing player is allowed to use only the mixed strategies which are convex combinations of pure strategies in B^* , we have

$$E(\xi, \eta_0^*) \leq E(\xi, \eta) \quad \text{for all } \xi.$$

Proof. For every ξ we have

$$\begin{aligned} E(\xi, \eta_0^*) - E(\xi, \eta) &= \sum_i \xi_i \left(\sum_{j \in B^*} a_{ij} \eta_{0j}^* - \sum_{j \in B} a_{ij} \eta_j \right) \\ &= \sum_i \xi_i \sum_{j \in B^*} (a_{ij} - \alpha_i) \eta_{0j}^* \\ &\leq \max_{\xi} \sum_i \sum_{j \in B^*} (a_{ij} - \alpha_i) \xi_i \eta_{0j}^* \end{aligned}$$

$$\begin{aligned}
&= \min_{\zeta^*} \max_{\xi} \sum_i \sum_{j \in B^*} (a_{ij} - \alpha_i) \xi_i \zeta_j^* = \max_{\xi} \min_{\zeta^*} [\dots] \\
&= \max_{\xi} \min_{\zeta} \sum_{i,j} (a_{ij} - \alpha_i) \xi_i \zeta_j \quad (\text{By Lemma 1}) \\
&\leq 0
\end{aligned}$$

Lemma 3. If $\{\delta_x\}$ is complete in $\{\delta_{xy}\}$, that is, if for every $\delta_{xy} \in \{\delta_{xy}\}$, there exists a $\delta_x \in \{\delta_x\}$ such that

$$r(\xi, \delta_x) \leq r(\xi, \delta_{xy}) \quad \text{for all } \xi,$$

then for every a priori distribution ξ , the Bayes decision function with respect to it has the form d_x .

Proof. Let δ_{xy} be the Bayes d.f. with respect to ξ . Since there exists from the assumption of the lemma, a d.f. δ_x with $r(\xi, \delta_x) \leq r(\xi, \delta_{xy})$, and the reverse inequality holds true by the definition of δ_{xy} , we have

$$(*) \quad r(\xi, \delta_x) = r(\xi, \delta_{xy}).$$

Hence we have to show that this δ_x is equivalent to some d_x . We use here the well-known fact that every randomized d.f. is equivalent to some mixed d.f. η_x . Let $\{d_j(x)\}_{j=1}^{\nu}$ be the exhaustive set of the non-randomized d.f.'s depending upon x only and lying in the set $\{d_{xy}\}$.

Then we have

$$\begin{aligned}
r(\xi, \delta_x) &= \sum_i \xi_i \sum_{j=1}^{\nu} \eta_{xj} r(i, d_j(x)) \\
&= \sum_i \xi_i \sum_j \eta_{xyj} r(i, d_j(x, y))
\end{aligned}$$

$$\text{where } \eta_{xyj} = \begin{cases} \eta_{xj}, & j = 1, \dots, \nu \\ 0, & j \geq \nu + 1. \end{cases}$$

Since we have

$$r(\xi, \delta_{xy}) \leq \sum_i \xi_i r(i, d_j(x, y)) \quad \text{for all } j,$$

we must have by (*)

$$r(\xi, \delta_x) = \sum_i \xi_i r(i, d_j(x, y)) \quad \text{for at least one } j \ (1 \leq j \leq \nu).$$

If otherwise we have $r(\xi, \delta_x) < \sum_i \xi_i \sum_j \eta_{xyj} r(i, d_j(x, y)) = r(\xi, \delta_x)$, a contradiction.

Proof of the Theorem.

(i) Assume that x is sufficient for y . Take any d.f. δ_{xy}^0 and the equivalent mixed d.f. and fix these. Since we have

$$r(\xi, \delta_{xy}) = \sum_i \xi_i \sum_j \eta_{xyj} r(i, d_j(x, y)) \equiv \sum_{i,j} \xi_i \eta_{xyj} a_{ij}, \text{ say}$$

we get

$$r(\xi, \delta_{xy}) - r(\xi, \delta_{xy}^0) = \sum_{i,j} \xi_i \eta_{xyj} (a_{ij} - \alpha_i),$$

where $\alpha_i \equiv \sum_j \eta_{xyj}^0 a_{ij}$.

Since x is sufficient for y , there exists, by Lemma 1, a subset B^* of B such that

$$\min_{j \in B^*} \sum_i (a_{ij} - \alpha_i) \xi_i = \min_{j \in B} \sum_i (a_{ij} - \alpha_i) \xi_i \quad \text{for all } \xi.$$

If η_0^* is a minimax mixed-d.f. when limited in the mixtures of non-randomized d.f. d_x in B^* , then we have by Lemma 2

$$r(\xi, \eta_0^*) = \sum_i \sum_{j \in B^*} a_{ij} \xi_i \eta_0^* \leq \sum_{i,j} a_{ij} \xi_i \eta_{xyj}^0 = r(\xi, \delta_{xy}^0).$$

Here the mixed d.f. η_0^* depends upon x only, and some randomized d.f. δ_x equivalent to η_0^* can be taken. We thus have shown that $\{\delta_x\}$ is uniformly complete in $\{\delta_{xy}\}$.

(ii) Let us assume that $\{\delta_x\}$ is uniformly complete in $\{\delta_{xy}\}$. The Bayes d.f. with respect to ξ has the form d_x and is independent of y . Hence, for any $y' \neq y''$,

$$(*) \left\{ \begin{array}{l} \text{if } d_1 \text{ and } d_2 \text{ are two non-randomized d.f. which minimize} \\ \sum_i \xi_i p_i(x, y') w(i, d) \text{ and } \sum_i \xi_i p_i(x, y'') w(i, d), \text{ respectively, then} \end{array} \right.$$

we must have $d_1 = d_2$.

If otherwise, the Bayes solution will have the form d_{xy} . Now let us assume that x is not sufficient for y . Then there exist x, y' , and y'' ($y' \neq y''$) such that

$$p_1(x), p_2(x) > 0 \text{ and}$$

$$\begin{vmatrix} p_1(x, y') & p_2(x, y') \\ p_1(x, y'') & p_2(x, y'') \end{vmatrix} = p_1(x)p_2(x) \begin{vmatrix} p_1(y' | x) & p_2(y' | x) \\ p_1(y'' | x) & p_2(y'' | x) \end{vmatrix} \neq 0.$$

We can, thus, choose $\xi_1 > 0$, $\xi_2 = 1 - \xi_1 > 0$ and $w(i, d)$ ($i=1, 2; d=d_1, d_2$) such that the simultaneous equations

$$\begin{aligned} p_1(x, y') \xi_1 w(1, d) + p_2(x, y') \xi_2 w(2, d) &= \begin{cases} 0, \\ 1, \end{cases} & d = \begin{cases} d_1 \\ d_2 \end{cases} \\ p_1(x, y'') \xi_1 w(1, d) + p_2(x, y'') \xi_2 w(2, d) &= \begin{cases} 1, \\ 0, \end{cases} & d = \begin{cases} d_1 \\ d_2 \end{cases} \end{aligned}$$

are satisfied. If we set

$$\xi_3 = \dots = \xi_m = 0; w(i, d_j) \equiv 1, (j \geq 3),$$

then this contradicts with the proposition (*).

Appendix B

Wald sequential-probability-ratio test.

The purpose of this appendix is to give a brief sketch of the theory of sequential probability ratio tests, in order to aid the comprehension of the discussions in Section 4.

Suppose the random variables X_1, X_2, \dots are distributed independently and identically, and their common density function $f(x)$ is known to be either $f_0(x)$ or $f_1(x)$. Let the possible terminal decisions be

d_0 accept the hypothesis $H_0: f(x) = f_0(x)$

d_1 accept the hypothesis $H_1: f(x) = f_1(x)$.

Let the loss of the terminal decision d_j ($j=0,1$), after observing X_1, \dots, X_m and stopping with X_m when the hypothesis H_i ($i=0,1$), is true, be given by

$$L(i, d_j | X_1, \dots, X_m) = \begin{cases} cm + w_i, & j \neq i \\ cm, & j = i \end{cases}$$

($c, l_0, l_1 > 0; i, j=0,1; m=1, 2, \dots$) and

$$L(i, d_j) = \begin{cases} w_i, & j \neq i \\ 0, & j = i, \end{cases}$$

for $m = 0$. In the above expression c is a given positive constant which represents the cost of sampling per observation. Thus the risk function of the decision rule δ is

$$r(i, \delta) = \sum_{j=0}^1 L(i, d_j) \Pr(d_j | \delta) + \sum_{m=1}^{\infty} E \left[\sum_{j=0}^1 L(i, d_j | X_1, \dots, X_m) \Pr(d_j | X_1, \dots, X_m; \delta) \middle| \delta, H_i \right]$$

Let $\delta^{[i]}$ ($i=0,1$) be the particular decision rule which chooses d_i , without any observation.

Theorem 1. There exist two values $0 \leq \underline{\zeta} \leq \bar{\zeta} \leq 1$, such that

$$\text{Bayes decision rule r.t. } \zeta, 1-\zeta \begin{cases} \text{is } \delta^{[1]}, & \text{if } 0 \leq \zeta \leq \underline{\zeta} \\ \text{observe } X_1, & \text{if } \underline{\zeta} < \zeta < \bar{\zeta} \\ \text{is } \delta^{[0]}, & \text{if } \bar{\zeta} \leq \zeta \leq 1. \end{cases}$$

Proof. We shall carry the proof in several steps.

(a) $\delta^{[0]}$ is Bayes r.t. $1,0$. If $\delta^{[0]}$ is Bayes r.t. $\zeta, 1-\zeta$, then it is Bayes r.t. $\zeta', 1-\zeta'$ for all $\zeta' \geq \zeta$. For, suppose that there exists

$\zeta < \zeta' < 1$; $\delta^{[0]}$ is Bayes r.t. $\zeta, 1-\zeta$, but non-Bayes r.t. $(\zeta', 1-\zeta')$.

Then there exists

$$\delta; \zeta' r(0, \delta) + (1-\zeta') r(1, \delta) < \zeta' r(0, \delta^{[0]}) + (1-\zeta') r(1, \delta^{[0]}) = (1-\zeta') r(1, \delta^{[0]}),$$

$$\therefore r(1, \delta) < r(1, \delta^{[0]})$$

For this δ we have

$$(1-\zeta) r(1, \delta^{[0]}) = \zeta r(0, \delta^{[0]}) + (1-\zeta) r(1, \delta^{[0]}) \leq \zeta r(0, \delta) + (1-\zeta) r(1, \delta)$$

and hence

$$\frac{1-\zeta}{\zeta} \leq \frac{r(0, \delta)}{r(1, \delta^{[0]}) - r(1, \delta)} < \frac{1-\zeta'}{\zeta'}$$

which contradicts $\zeta < \zeta'$.

(b) Thus

$$\bar{\zeta} \equiv \inf. \{ \zeta \mid \delta^{[0]} \text{ is Bayes r.t. } (\zeta', 1-\zeta'), \text{ for all } \zeta' \geq \zeta \}$$

exists. We want to show that $\bar{\zeta}$ is the minimum number. Let $\bar{\zeta} < 1$ without loss of generality. Suppose, on the contrary, that $\delta^{[0]}$ is non-Bayes r.t. $\bar{\zeta}, 1-\bar{\zeta}$.

Then there exists

$$\delta; \bar{\zeta}r(0,\delta)+(1-\bar{\zeta})r(1,\delta) < \bar{\zeta}r(0,\delta^{[0]})+(1-\bar{\zeta})r(1,\delta^{[0]}),$$

and hence for slightly larger numbers $\zeta' = \bar{\zeta} + \epsilon$ we still have the same strict inequality, which shows a contradiction.

(c) By the same argument we can prove that $\delta^{[1]}$ is Bayes r.t. 0,1, and that

$$\underline{\zeta} \equiv \max\{\zeta | \delta^{[1]} \text{ is Bayes r.t. } (\zeta', 1-\zeta'), \text{ for all } \zeta' \leq \zeta\}$$

exists

(d) Next we want to show that $\underline{\zeta} \leq \bar{\zeta}$. Suppose that $\underline{\zeta} > \bar{\zeta}$. Then both $\delta^{[0]}$ and $\delta^{[1]}$ would be Bayes r.t., $\zeta, 1-\zeta$, for $\bar{\zeta} \leq \forall \zeta \leq \underline{\zeta}$. Hence

$$\zeta r(0,\delta^{[0]})+(1-\zeta)r(1,\delta^{[0]}) = \zeta r(0,\delta^{[1]})+(1-\zeta)r(1,\delta^{[1]})$$

$$\therefore \zeta(r(0,\delta^{[1]})+r(1,\delta^{[0]})) = r(1,\delta^{[0]}), \text{ for } \bar{\zeta} \leq \text{for all } \zeta \leq \underline{\zeta}.$$

which is a contradiction.

(e) Finally we must show that, if $\underline{\zeta} < \zeta < \bar{\zeta}$ then a Bayes rule (δ^* , say) r.t. $\zeta, 1-\zeta$ certainly observes X_1 . With

$$p_i \equiv \Pr\{\delta^* \text{ chooses } d_i \text{ without any observation}\} (i=0,1),$$

we find that $p_0 + p_1 < 1$. For, if $p_0 + p_1 = 1$ then

$$\left\{ \begin{array}{l} \zeta r(0,\delta^{[0]})+(1-\zeta)r(1,\delta^{[1]}) \\ \zeta r(0,\delta^{[1]})+(1-\zeta)r(1,\delta^{[1]}) \end{array} \right\} > \zeta r(0,\delta^*)+(1-\zeta)r(1,\delta^*) \quad (*)$$

$$= \zeta \sum_{i=0}^1 p_i r(0,\delta^{[i]})+(1-\zeta) \sum_{i=0}^1 p_i r(1,\delta^{[i]})$$

$$= \sum_{i=0}^1 p_i \{\zeta r(0,\delta^{[i]})+(1-\zeta)r(1,\delta^{[i]})\}$$

which is a contradiction. Actually, we have $p_0 + p_1 = 0$. For, denoting

by δ' a decision rule which observes X_1 with probability 1 and thereafter behaves exactly the same as δ^* , we have

$$r(i, \delta^*) = \sum_{j=0}^1 p_j r(i, \delta^{[j]}) + (1-p_0-p_1)r(i, \delta'), \quad i=0,1$$

$$\begin{aligned} \therefore \zeta r(0, \delta^*) + (1-\zeta)r(1, \delta^*) &= \sum_{j=0}^1 p_j (\zeta r(0, \delta^{[j]}) + (1-\zeta)r(1, \delta^{[j]})) \\ &\quad + (1-p_0-p_1)(\zeta r(0, \delta') + (1-\zeta)r(1, \delta')). \end{aligned}$$

Hence, by (*), if $p_0 + p_1 > 0$

$$\begin{aligned} \zeta r(0, \delta^*) + (1-\zeta)r(1, \delta^*) &> (p_0+p_1)(\zeta r(0, \delta^*) + (1-\zeta)r(1, \delta^*)) \\ &\quad + (1-p_0-p_1)(\zeta r(0, \delta') + (1-\zeta)r(1, \delta')) \end{aligned}$$

$$\therefore \zeta r(0, \delta^*) + (1-\zeta)r(1, \delta^*) > \zeta r(0, \delta') + (1-\zeta)r(1, \delta')$$

which is a contradiction.

Theorem 2. Let $\underline{\zeta} < \zeta < \bar{\zeta}$. Then, by Theorem 6.1, the Bayes rule r.t.

$\zeta, 1-\zeta$ certainly observes X_1 . It is given by

$$\left\{ \begin{array}{ll} \text{Accept } H_1, & \text{if } \prod_{j=1}^m \{f_1(X_j)/f_0(X_j)\} \geq \frac{\zeta}{1-\zeta} \left(\frac{1}{\underline{\zeta}} - 1 \right) \\ \text{accept } H_0, & \text{if } \quad \leq \frac{\zeta}{1-\zeta} \left(\frac{1}{\bar{\zeta}} - 1 \right); \text{ and} \\ \text{continue sampling as long as} & \frac{\zeta}{1-\zeta} \left(\frac{1}{\underline{\zeta}} - 1 \right) < \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad < \frac{\zeta}{1-\zeta} \left(\frac{1}{\bar{\zeta}} - 1 \right). \end{array} \right.$$

Proof. For every positive integer m

$$r(i, \delta) = \Pr\{n < m | H_1, \delta\} E[\text{loss} | n < m; H_1, \delta] + \Pr\{n \geq m | H_1, \delta\} E[\text{loss} | n \geq m; H_1, \delta].$$

Let the first term in the right-hand side of the above expression be denoted by $K_1(m|\delta)$, and the second term rewritten as

$$\sum_{x_1, \dots, x_m} \dots \sum \Pr\{n \geq m | \delta; X_1 = x_1, \dots, X_{m-1} = x_{m-1}\} E[\text{loss} | n \geq m; H_i, \delta; X_1 = x_1, \dots, X_m = x_m] \prod_{j=1}^m f_i(x_j)$$

(Unless δ uses randomization in describing its sampling rule, the first factor of the summand would have the value 0 or 1.)

Let δ_{x_1, \dots, x_m} be the decision rule which behaves exactly the same as δ after observing $X_1 = x_1, \dots, X_m = x_m$. Then we have

$$E[\text{loss} | n \geq m; H_i, \delta; X_1 = x_1, \dots, X_m = x_m] = cm + r(i, \delta_{x_1, \dots, x_m})$$

and so

$$r(i, \delta) = K_i(m | \delta) + \sum_{x_1, \dots, x_m} \dots \sum g(x_1, \dots, x_{m-1} | \delta) \{cm + r(i, \delta_{x_1, \dots, x_m})\} \prod_{j=1}^m f_i(x_j),$$

where $g(x_1, \dots, x_{m-1} | \delta) \equiv \Pr\{n \geq m | \delta; X_1 = x_1, \dots, X_{m-1} = x_{m-1}\}$. Thus with

$f(x_1, \dots, x_m) = \zeta f_0(x_1) \dots f_0(x_m) + (1-\zeta) f_1(x_1) \dots f_1(x_m)$, we have

$$\zeta r(0, \delta) + (1-\zeta) r(1, \delta) = \zeta K_0(m | \delta) + (1-\zeta) K_1(m | \delta)$$

$$+ \sum_{x_1, \dots, x_m} \dots \sum g(x_1, \dots, x_{m-1} | \delta) \left\{ cm + \frac{\zeta f_0(x_1) \dots f_0(x_m)}{f(x_1, \dots, x_m)} r(0, \delta_{x_1, \dots, x_m}) \right.$$

$$\left. + \frac{(1-\zeta) f_1(x_1) \dots f_1(x_m)}{f(x_1, \dots, x_m)} r(1, \delta_{x_1, \dots, x_m}) \right\} f(x_1, \dots, x_m).$$

In order to minimize the above expression for every fixed x_1, \dots, x_m and m , δ_{x_1, \dots, x_m} must be the Bayes solution r.t. $\frac{\zeta f_0(x_1) \dots f_0(x_m)}{f(x_1, \dots, x_m)}$,

$\frac{(1-\zeta) f_1(x_1) \dots f_1(x_m)}{f(x_1, \dots, x_m)}$, which is given from Theorem 1, by

$$\left\{ \begin{array}{ll} \text{Accept } H_1, & \text{if } 0 \leq \frac{\zeta f_0(x_1) \dots f_0(x_m)}{f(x_1, \dots, x_m)} \leq \underline{\zeta}, \\ \text{observes } X_{m+1}, & \text{if } \underline{\zeta} < \quad \quad \quad < \bar{\zeta}; \text{ and} \\ \text{accept } H_0, & \text{if } \bar{\zeta} \leq \quad \quad \quad \leq 1. \end{array} \right.$$

By Theorems 1 and 2 we have shown that the Wald sequential-probability-ratio test is optimal in the sense that for given prior probabilities $\zeta, 1-\zeta$ it minimizes the average risk $\zeta r(0, \delta) + (1-\zeta)r(1, \delta)$. We shall next prove another optimal property.

Theorem 2'. Let a sequential-probability-ratio test δ with the boundaries A and B with $B < 0 < A$ have the strength (α, β) , i.e., $\Pr\{\delta \text{ accepts } H_1 | H_0\} = \alpha$ and $\Pr\{\delta \text{ accepts } H_0 | H_1\} = \beta$. Then for any other test δ' such that

$$\Pr\{\delta' \text{ accepts } H_1 | H_0\} \leq \alpha, \text{ and}$$

$$\Pr\{\delta' \text{ accepts } H_0 | H_1\} \leq \beta,$$

we have

$$E_1(n|\delta) \leq E_1(n|\delta'), \quad i=0,1.$$

Proof. For any $0 < \zeta < 1$ we have

$$\zeta(\ell_0 \alpha + cE_0(n|\delta)) + (1-\zeta)(\ell_1 \beta + cE_1(n|\delta)) = \zeta r(0, \delta) + (1-\zeta)r(1, \delta)$$

$$\leq \zeta r(0, \delta') + (1-\zeta)r(1, \delta') = \zeta(\ell_0 \Pr\{\delta' \text{ accepts } H_1 | H_0\} + cE_0(n|\delta'))$$

$$+ (1-\zeta)(\ell_1 \Pr\{\delta' \text{ accepts } H_0 | H_1\} + cE_1(n|\delta'))$$

$$\therefore \zeta E_0(n|\delta) + (1-\zeta)E_1(n|\delta) \leq \zeta E_0(n|\delta') + (1-\zeta)E_1(n|\delta').$$

This inequality must hold for all $0 < \zeta < 1$. Hence, from continuity considerations we must have $E_i(n|\delta) \leq E_i(n|\delta')$, $i=0,1$.

If we can find the values of A and B such that $r(0, \delta_\zeta) = r(1, \delta_\zeta)$, then δ_ζ with such A and B is a minimax decision rule. But, actually, we cannot find such exact values, so we shall have to be satisfied with approximations.

Before stating approximations we present the following theorem.

Theorem 3 For the Wald sequential-probability-ratio test, we have

(i) Closure; or more precisely, there exists

$$0 < p < 1; P_\theta(n > N) < \text{const } p^N, \text{ for sufficiently large } N.$$

(ii) Wald-Blackwell identity:

$$E_\theta(Z_1 + \dots + Z_n) = E_\theta(n)E_\theta(Z).$$

Proof. (i) Assume that $P_\theta(Z > 0) > 0$. The argument for the case $P_\theta(Z < 0) > 0$ is similar. Since there exists a $\Delta > 0$ such that $P_\theta(Z > \Delta) > 0$, we have if $\gamma \equiv \frac{A+(-B)}{\Delta} + 1$,

$$P_\theta\left\{\sum_{i=1}^{\gamma} z_i > A+(-B)\right\} \geq (P_\theta\{z > \frac{A+(-B)}{\gamma}\})^\gamma \geq (P_\theta\{z > \Delta\})^\gamma \equiv p^* > 0$$

and

$$\begin{aligned} P_\theta\{n > k\gamma\} &\leq P_\theta\left\{\sum_{i=m\gamma+1}^{(m+1)\gamma} z_i \leq A-B, m=0,1,\dots,k-1\right\} \\ &= (P_\theta\left\{\sum_{i=1}^{\gamma} z_i \leq A-B\right\})^k \leq (1-p^*)^k = [(1-p^*)^{1/\gamma}]^{k\gamma} \equiv p^{k\gamma}, \text{ (say).} \end{aligned}$$

Thus setting $N = k\gamma + h$ ($0 \leq h \leq \gamma - 1$), we obtain if $N \geq \gamma$

$$P_{\theta}\{n > N\} \leq P_{\theta}\{n > k\gamma\} \leq p^{k\gamma} = p^N/p^h \leq p^N/p^{\gamma-1}.$$

(ii) Let $Z_N \equiv \sum_{i=1}^N z_i$, and consider N large and fix it. We have

$$P_{\theta}(n > N)E_{\theta}(Z_N | n > N) = E_{\theta}(Z_N) - P_{\theta}(n \leq N)E_{\theta}(Z_N | n \leq N)$$

$$= NE_{\theta}(Z) - P_{\theta}(n \leq N)\{E_{\theta}(Z_N | n \leq N) + E_{\theta}(N - n | n \leq N)E_{\theta}(Z)\}$$

$$= N\{1 - P_{\theta}(n \leq N)\}E_{\theta}(Z) - P_{\theta}(n \leq N)E_{\theta}(Z_N | n \leq N) + P_{\theta}(n \leq N)E_{\theta}(n | n \leq N)E_{\theta}(Z),$$

in which the left hand side $\xrightarrow{(N \rightarrow \infty)} 0$ (because $E_{\theta}(|Z_N| | n > N) \leq \max(A, -B)$)

and in the right hand side

$$P_{\theta}(n \leq N)E_{\theta}(Z_N | n \leq N) \xrightarrow{(N \rightarrow \infty)} E_{\theta}(Z_N),$$

$$P_{\theta}(n \leq N)E_{\theta}(n | n \leq N) \xrightarrow{(N \rightarrow \infty)} E_{\theta}(n)$$

and

$$NP_{\theta}(n > N) \xrightarrow{(N \rightarrow \infty)} 0 \text{ (because } E_{\theta}(n) < \infty \text{)}.$$

Let

$S_{im} = \{x_1, \dots, x_m\}$ | If $X_1 = x_1, \dots, X_m = x_m$, then the Bayes test δ_C stops sampling and accepts H_i ($i=0,1; m=1,2,\dots$). Our approximation is to

assume that

$$\left\{ \begin{array}{ll} \sum_{j=1}^m \log(f_1(x_j)/f_0(x_j)) = B, & \text{for all } (x_1, \dots, x_m) \in S_{0m}; \text{ and} \\ & \\ & = A, & \text{for all } (x_1, \dots, x_m) \in S_{1m} \end{array} \right.$$

This approximation looks drastic, but it works fairly well if the densities $f_0(x)$ and $f_1(x)$ do not differ greatly from each other. Since

$$\left\{ \begin{array}{l} \alpha(A,B) = \sum_{m=1}^{\infty} P_0(S_{1m}) \\ \beta(A,B) = \sum_{m=1}^{\infty} P_1(S_{0m}), \end{array} \right. \quad \left\{ \begin{array}{l} 1-\alpha(A,B) = \sum_{m=1}^{\infty} P_0(S_{0m}) \\ 1-\beta(A,B) = \sum_{m=1}^{\infty} P_1(S_{0m}), \end{array} \right.$$

we have, using the above approximation

$$\left\{ \begin{array}{l} e^A \alpha(A,B) \doteq 1-\beta(A,B) \\ e^B (1-\alpha(A,B)) \doteq \beta(A,B), \end{array} \right.$$

which is equivalent to

$$\left. \begin{array}{l} \alpha(A,B) \doteq \frac{1-e^B}{e^A-e^B} \\ \beta(A,B) \doteq e^B \frac{e^A-1}{e^A-e^B} \end{array} \right\} \quad (*)$$

Moreover we have by the approximation

$$\left. \begin{array}{l} E_0(z_1+\dots+z_n) \doteq \alpha(A,B)A + (1-\alpha(A,B))B \\ E_1(z_1+\dots+z_n) \doteq \beta(A,B)B + (1-\beta(A,B))A \end{array} \right\} \quad (**)$$

From (*) and (**),

$$\left\{ \begin{array}{l} E_0(n) = \frac{E_0(z_1+\dots+z_n)}{I(0:1)} = \left(A \frac{1-e^B}{e^A-e^B} + B \frac{e^A-1}{e^A-e^B} \right) / I(0:1) \\ E_1(n) = \frac{E_1(z_1+\dots+z_n)}{I(1:0)} = \left(B e^B \frac{e^A-1}{e^A-e^B} + A e^A \frac{1-e^B}{e^A-e^B} \right) / I(1:0) \end{array} \right.$$

and by setting

$$a = e^A, \quad b = e^B \quad (\text{so that } 0 < b < 1 < a),$$

we have

$$\begin{cases} r(0, \delta) = w_0 \alpha + c E_0(n) = w_0 \frac{1-b}{a-b} + \frac{c}{I(0:1)} \left(\frac{1-b}{a-b} \log a + \frac{a-1}{a-b} \log b \right) \\ r(1, \delta) = w_1 \beta + c E_1(n) = w_1 \frac{b(a-1)}{a-b} + \frac{c}{I(1:0)} \left(\frac{1-b}{a-b} a \log a + \frac{a-1}{a-b} b \log b \right) \end{cases}$$

These are approximations for the risks of the Wald sequential-probability-ratio test.

Appendix C

Multivariate information transmission

The transmission of information requires the presence of a source of information coupled with an appropriate channel; the two together form what it is called an information system, or communication system. Here an information system is described in terms of joint probabilities of inputs and outputs, and a channel is defined by its transition probabilities.

The formulae are written as if x , y , etc. were continuous real variables; the obvious modifications must be made if they are discrete, vector-valued, etc.

Let us consider a communication channel and its input and output. Transmitted information measures the amount of association between the input and the output of the channel. If input and output are independent no information is transmitted. On the other hand, if both are perfectly correlated, all the input information is transmitted through the channel. In most cases, naturally, information transmission is found between these extremes.

We are interested in the amount of information transmitted. Suppose that we have a bivariate probability distribution with the density function $p(x,y)$. This means that if the input variable assumes a value or signal x , then noise of the channel alters it, at the output, to a value between y and $y + dy$ with probability $p(y|x)dy$ where $p(y|x) = p(x,y)/\int p(x,y)dy$, and that the rules governing the selection of

signals at the input must be constructed so that they take on values between x and $x + dx$ with probabilities $p(x)dx = dx \int p(x,y)dy$. To avoid complexity we use the notation $p(\cdot)$ to represent the density functions of the random variables, without any suggestion that they have the same density.

Under these conditions, and if successive signals are independent the amount of information transmitted per signal is defined by Shannon as

$$(1) \quad T(x;y) = H(x) + H(y) - H(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy,$$

where $H(x) \equiv -\int p(x) \log p(x) dx$, $H(y) \equiv -\int p(y) \log p(y) dy$ and $H(x,y) \equiv -\iint p(x,y) \log p(x,y) dx dy$. We know by Theorem 2.1 that $T(x,y)$ is non-negative and equals zero if and only if x and y are independent.

When we introduce the conditional entropies

$$H_x(y) \equiv -\iint p(x,y) \log p(y|x) dx dy, \quad \text{etc.}$$

we easily have the following additive formula:

$$H(x_1, x_2, \dots, x_k) = H(x_1) + H_{x_1}(x_2) + \dots + H_{x_1, \dots, x_{k-1}}(x_k), \quad (k \geq 2).$$

Thus we have another expression of $T(x;y)$ as

$$(2) \quad T(x;y) = H(x) - H_y(x) = H(y) - H_x(y).$$

Now let us consider the case where we have several sources that transmit to y . Then we take the input variable as multidimensional and we have, for instance,

$$(3) \quad T(u,v;y) = H(u,v) + H(y) - H(u,v,y) = H(u,v) - H_y(u,v) \\ = H(y) - H_{u,v}(y).$$

We can express $T(u, v; y)$ as a combination of the bivariate transmissions between u and y , and v and y . Define $T_{u=u_0}(v; y)$ as transmitted information between v and y for a particular value of u , namely u_0 . If we set

$$T_u(v; y) = \int T_{u=u_0}(v; y) p(u_0) du_0,$$

we easily see that

$$\begin{aligned} (4) \quad T_u(v; y) &= H_u(v) + H_u(y) - H_u(v, y) \\ &= H_u(v) - H_{u, y}(v) = H_u(y) - H_{u, v}(y). \end{aligned}$$

Hence we have from (2), (3), and (4)

$$(5) \quad T(u, v; y) = T(u; y) + T_u(v; y) = T(v; y) + T_v(u; y),$$

which means that the additive formula for information transmission also holds true. We have from (2) and (4)

$$\begin{aligned} T(v; y) - T_u(v; y) &= (H(y) - H_v(y)) - (H_u(y) - H_{u, v}(y)) \\ &= (H(v) - H_y(v)) + (H_u(v) - H_{u, y}(v)) \geq 0. \end{aligned}$$

These identities show the symmetry of the left hand side expression in the arguments u and v , and u and y . Since the symmetry between v and y is clear from the definitions (1) and (3), we have

$$\begin{aligned} (6) \quad A(uv; y) &\equiv T(v; y) - T_u(v; y) \\ &= T(u; y) - T_v(u; y) \\ &= T(u; v) - T_y(u; v) \geq 0. \end{aligned}$$

We shall, following McGill (1956), call $A(uv; y)$ the interaction information between the three variables. It is the gain or loss in transmitted information between any two of the variables, due to additional knowledge of the third variable.

We shall derive another important expression for $A(uvy)$ as follows: subtracting $T(v;y)$ from both sides of the first identity of (5) we have $T(u,v;y) - T(v;y) = T(u;y) - A(uvy)$. From (3) we have $T(u,v;y) = H(u) + H_u(v) + H(y) - H(u,v,y)$. Then by these two identities and the fact that $T(u;v) = H(v) - H_u(v)$, we finally get

$$(7) \quad A(uvy) = - (H(u) + H(v) + H(y) - H(u,v,y)) + (T(u;v) + T(u;y) + T(v;y)).$$

We shall show in the following several examples in which transmission of information is effectively analyzed by using the above.

Example 1

According to the definition (6) interaction information is positive (negative) when the effect of holding one of the interacting variables constant is to decrease (increase) the amount of association between the other two. And we have

Theorem 1. A necessary and sufficient condition for the three random variables u , v , and y to be independent is that we have

$$(a) \quad T(u;v) = T(v;y) = T(u;y) = 0; \text{ and}$$

$$(b) \quad A(uvy) = 0.$$

Proof. Necessity is almost clear. Since any pair of the two among the three independent variables are independent, the condition (a) is necessary and (b) is an immediate consequence of (7). To prove sufficiency we must show that independence results from $H(u,v,y) = H(u) + H(v) + H(y)$. Since

$$\iiint p(u,v,y) \log \frac{p(u,v,y)}{p(u)p(v)p(y)} \, du \, dv \, dy = 0$$

we have, by Theorem 2.1, $p(u,v,y) = p(u)p(v)p(y)$, which completes the proof.

It is well known that the condition (a) alone is not sufficient for independence between the three variables. Let $\mathcal{X} = \{a_1, a_2, a_3, a_4\}$ be a probability space with probabilities $1/4$ for each elementary event a_i ($i=1, \dots, 4$). Set

$$A = \{a_1, a_2\}, \quad B = \{a_1, a_3\}, \quad C = \{a_1, a_4\}$$

and let u, v , and y be indicator functions of the events A , B , and C , respectively. Then we easily see that the three random variables are pairwise independent but are not mutually independent. We have

$$H(uvy) \doteq H(u, v, y) - (H(u) + H(v) + H(y)) = 2 \log 2 - 3 \log 2 = -\log 2$$

Example 2

Let $\{x_n\}_{n=1}^{\infty}$ be a stationary (in the strict sense) Markov process with order s (≥ 1). Since for every n ($> s$) and m

$$\begin{aligned} H(x_{m+1}, x_{m+2}, \dots, x_{m+n}) &= H(x_{m+1}, \dots, x_{m+s}) + (n-s) H_{x_{m+1}, \dots, x_{m+s}}(x_{m+s+1}) \\ &= H(x_1, \dots, x_s) + (n-s) H_{x_1, \dots, x_s}(x_{s+1}) \end{aligned}$$

by stationarity of the process, we have

$$(8) \quad \lim_{n \rightarrow \infty} n^{-1} H(x_{m+1}, \dots, x_{m+n}) = H_{x_1, \dots, x_s}(x_{s+1}).$$

Thus the mean information per symbol contained in sufficiently long messages is equal to the conditional entropy of the process. The conditional entropy determines the redundancy of the process, i.e.

$$(9) \quad R_s \equiv \frac{H(x_{s+1}) - H_{x_1, \dots, x_s}(x_{s+1})}{H(x_{s+1})} = \frac{T(x_1, \dots, x_s; x_{s+1})}{H(x_{s+1})}$$

Evidently we have $0 < R_s \leq 1$, and $R_s = 1$ if and only if the value of

x_{s+1} is determined by the set of values (x_1, \dots, x_s) .

Redundancy measures inter-symbol correlation of the message. To analyze the redundancy of the process is equivalent to analyzing the transmitted information $T(x_1, \dots, x_s; x_{s+1})$. A natural generalization of (5) yields

$$(10) \quad T(x_1, \dots, x_s; x_{s+1}) = T(x_s; x_{s+1}) + T_{x_s}(x_{s-1}; x_{s+1}) + T_{x_{s-1}, x_s}(x_{s-2}; x_{s+1}) \\ + \dots + T_{x_2, \dots, x_s}(x_1; x_{s+1}) \\ = T(x_1; x_{s+1}) + T_{x_1}(x_2; x_{s+1}) + T_{x_1, x_2}(x_3; x_{s+1}) + \dots + T_{x_1, \dots, x_{s-1}}(x_s; x_{s+1}).$$

If we note the relation

$$(11) \quad T(x_1, \dots, x_s; x_{s+1}) = T(x_1; x_2, \dots, x_{s+1})$$

we obtain another expression

$$(12) \quad T(x_1, \dots, x_s; x_{s+1}) = T(x_1; x_2) + T_{x_2}(x_1; x_3) + T_{x_2, x_3}(x_1; x_4) + \dots + \\ T_{x_2, \dots, x_s}(x_1; x_{s+1}).$$

A proof of (11) is as follows:

$$T(x_1, \dots, x_s; x_{s+1}) = H(x_1, \dots, x_s) + H(x_{s+1}) - H(x_1, \dots, x_{s+1}) \\ = H(x_2, \dots, x_{s+1}) + H(x_1) - H(x_1, \dots, x_{s+1}) \quad (\text{by stationarity}) \\ = T(x_1; x_2, \dots, x_{s+1}).$$

Example 3 Series connection of channels. (Sakaguchi, 1957)

Hereafter we shall denote a communication system by

$$\mathcal{E} = [\mathcal{X}, \{p(y|x) | x \in \mathcal{X}\}, \mathcal{Y}]$$

where \mathcal{X} and \mathcal{Y} represent the input and output spaces, respectively. The a priori probability distribution $p(x)$ over the input space is not relevant

to the definition of the channel. If we connect the input of a channel

$\mathcal{E}_2 = [\mathcal{Y}, \{p(z|y) | y \in \mathcal{Y}\}, \mathcal{Z}]$ to the output of another channel

$\mathcal{E}_1 = [\mathcal{X}, \{p(y|x) | x \in \mathcal{X}, u\}]$, then we have a new channel

$$\mathcal{E}_1 + \mathcal{E}_2 = [\mathcal{X}, \{p(z|x) | x \in \mathcal{X}\}, \mathcal{Z}]$$

where

$$(13) \quad p(z|x) = \int p(y|x)p(z|y)dy.$$

We shall show that a positive interaction information is produced

between the variables x , y and z . Clearly we have $p(z|x,y) = p(z|y)$

and $p(x|y,z) = p(x|y)$. Hence we have $H_{x,y}(z) = H_y(z)$ and $H_{y,z}(x) = H_y(x)$.

Thus we obtain from (4)

$$T_y(x;z) = H_y(x) - H_{y,z}(x) = H_y(z) - H_{y,x}(z) = 0$$

and therefore from (6)

$$(14) \quad 0 < A(xyz) = T(z;x) = T(x;y) - T_z(x;y) = T(y;z) - T_x(y;z)$$

The information transmitted between x and z is due to a correlation with a third variable y . Holding the interacting variable constant causes the

transmitted information to disappear. The equations in (14) also show

that the series connection, or equivalently, forming the "convolution"

(13) of two channels decreases the association between the two ends of each channel.

Example 4 Parallel connection of channels.

Given two channels

$$\mathcal{E}_i = [\mathcal{X}, \{p_i(y|x) | x \in \mathcal{X}\}, \mathcal{Y}] \quad (i=1,2)$$

with common input space \mathcal{X} and common output space \mathcal{Y} , we consider the mixture

$$\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2 \equiv [\mathcal{X}, \{\zeta p_1(y|x) + (1-\zeta) p_2(y|x) | x \in \mathcal{X}\}, \mathcal{Y}]$$

where $0 \leq \zeta \leq 1$. This can be thought of as follows: with probability ζ , a value y is observed according to the density $p_1(y|x)$; with probability $1-\zeta$, y is observed according to $p_2(y|x)$. The recipient is informed only of y and not of which channel with probability ζ or $1-\zeta$, worked.

We shall discuss the effect of forming the mixture. We introduce a third variable u , which is independent of x , and informs the recipient about which channel \mathcal{E}_1 or \mathcal{E}_2 worked, but not about the value of y . Let us denote the transmitted information through $\mathcal{E}_1, \mathcal{E}_2$ and $\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2$ when the same a priori distribution $p(x)$ over \mathcal{X} is assumed, as

$T^{(1)}(x;y)$, $T^{(2)}(x;y)$ and $T(x;y)$, respectively. From (6) and the definition of the variable u , the interaction information $A(xyu)$ is non-negative:

$$0 > A(xyu) = -T_y(x;u) = T(y;u) - T_x(y;u) = T(x;y) - T_u(x;y).$$

Since $T_u(x;y) = \zeta T^{(1)}(x;y) + (1-\zeta) T^{(2)}(x;y)$ we obtain

$$(15) \quad T(x;y) \leq \zeta T^{(1)}(x;y) + (1-\zeta) T^{(2)}(x;y),$$

which is the same inequality as in (ii) of Theorem 5.2.

Theorem 2 The transmitted information $T(x;y)$ through the mixture $\zeta \mathcal{E}_1 * (1-\zeta) \mathcal{E}_2$ is convex in $0 \leq \zeta \leq 1$. For any prior distribution $p(x)$, the uniform (in ζ) equality of (15) holds if and only if

$$\frac{p_1(y|x)}{p_2(y|x)} = \frac{p_1(y)}{p_2(y)}, \quad \text{identically.}$$

Proof. We can rewrite

$$T(x;y) = \iint p(x) (\zeta p_1(y|x) + (1-\zeta)p_2(y|x)) \log \frac{\zeta p_1(y|x) + (1-\zeta)p_2(y|x)}{\zeta p_1(y) + (1-\zeta)p_2(y)} dx dy$$

with $p_i(y) \equiv \int p(x)p_i(y|x)dx$ ($i=1,2$). Direct differentiation gives

$$\frac{d}{d\zeta} T(x;y) = \iint p(x) (p_1(y|x) - p_2(y|x)) \log \frac{\zeta p_1(y|x) + (1-\zeta)p_2(y|x)}{\zeta p_1(y) + (1-\zeta)p_2(y)} dx dy,$$

$$\frac{d^2}{d\zeta^2} T(x;y) = \iint p(x) \frac{(p_1(y|x) - p_2(y|x))^2}{\zeta p_1(y|x) + (1-\zeta)p_2(y|x)} dx dy - \int \frac{(p_1(y) - p_2(y))^2}{\zeta p_1(y) + (1-\zeta)p_2(y)} dy,$$

which, by convexity of the function $\psi(u,v) \equiv \frac{(u-v)^2}{\zeta u + (1-\zeta)v}$ in the region

$\{(u,v) | u,v \geq 0\}$, is found to be non-negative. The second part of the theorem follows at once from the strict convexity of the related function $\psi(u,v)$.

Example 5 Regression and information transmission.

Let $p(x,y)$ be the density function of a bivariate probability distribution. We introduce the third random variable

$$z = y - \varphi(x)$$

where $\varphi(x)$ is the regression function of y on x , i.e.

$$\varphi(x) \equiv \int y p(y|x) dy.$$

Since $H_{x,y}(z) = H_{y,z}(x) = H_{z,x}(y) = 0$, we have from (4)

$$T_z(x;y) = H_z(x) = H_z(y),$$

$$T_y(x;z) = H_y(x) = H_y(z).$$

Thus we obtain from (6)

$$(16) \quad T(x;y) = T(x;z) + T_z(x;y) - T_y(x;z) = T(x;z) + H(y) - H(z).$$

In the right hand side of the above equalities we note that $T(x;z) \geq 0$

but $H(y) - H(z) \geq 0$. We call, following Feron and Fourgeaud (1951), these quantities $T(x;z)$ and $H(y) - H(z)$, the elastic part and the hard part, respectively, of the transmitted information $T(x;y)$.

It can be shown that $T(x;z) = 0$ if and only if

$$(17) \quad p(y|x) = A(y - \varphi(x)),$$

where $A(z)$ is some density function such that $A(z) \geq 0$, $\int A(z) dz = 1$ and $\int zA(z) dz = 0$.

Quite similarly if we define

$$w = x - \psi(y),$$

where $\psi(y)$ is the regression function of x on y , i.e.

$$\psi(y) \equiv \int xp(x|y) dx$$

then we have

$$T(x;y) = T(y;w) + H(x) - H(w)$$

in which $T(y;w) \geq 0$ but $H(x) - H(w) \geq 0$. We again have $T(y;w) = 0$ when and only when the conditional density is written in the form

$$p(x|y) = B(x - \psi(y))$$

by some density function $B(w)$ with the mean value 0.

We say that the bivariate probability distribution has hard correlations when $T(x;z) = T(y;w) = 0$. We know that bivariate normal distributions have hard correlations and linear regressions. And we have

Theorem 3 The only bivariate probability law which has hard correlations and linear regressions is the Gaussian law.

Proof. The Gaussian law has the stated properties. We shall prove the converse. Without losing generality we can, by the assumptions, write

$$(18) \quad \begin{cases} p(y|x) = A(y - \rho_1 x) \\ p(x|y) = B(x - \rho_2 y) \end{cases}$$

where ρ_1 and ρ_2 are the linear regression coefficients. By simple applications of the Schwartz inequality we obtain $\rho_1^2 \leq \text{Var. } x / \text{Var. } y$, $\rho_2^2 \leq \text{Var. } y / \text{Var. } x$ and thus we have $|\rho_1 \rho_2| < 1$ in the non-trivial case.

Differentiating the identity

$$\log f(x) + \log A(y - \rho_1 x) = \log g(y) + \log B(x - \rho_2 y)$$

($f(x)$ and $g(y)$ denoting the densities of x and y respectively) with respect to x and y , we get

$$(19) \quad \rho_1 \underline{A}''(y - \rho_1 x) = \rho_2 \underline{B}''(x - \rho_2 y)$$

where we have set $\underline{A}(z) \equiv \log A(z)$ and $\underline{B}(z) \equiv \log B(z)$.

If we set in (19) $x = 0$ and $y = 0$, we have

$$\rho_1 \underline{A}''(y) = \rho_2 \underline{B}''(-\rho_2 y), \quad \rho_1 \underline{A}''(-\rho_1 x) = \rho_2 \underline{B}''(x)$$

respectively. Hence we have by $|\rho_1 \rho_2| < 1$

$$\underline{A}''(y) = (\rho_2 / \rho_1) \underline{B}''(-\rho_2 y) = \underline{A}''(\rho_1 \rho_2 y) = \dots = \underline{A}''(\rho_1^n \rho_2^n y) = \dots = \underline{A}''(0) \equiv -1/\beta, \text{ say}$$

$$\underline{B}''(x) = (\rho_1 / \rho_2) \underline{A}''(-\rho_1 x) = \underline{B}''(\rho_1 \rho_2 x) = \dots = \underline{B}''(\rho_1^n \rho_2^n x) = \dots = \underline{B}''(0) \equiv -1/\gamma, \text{ say.}$$

Integrating we get by (18)

$$(20) \quad \begin{cases} p(y|x) = A(y - \rho_1 x) = (2\pi\beta)^{-1/2} \exp\{-(y - \rho_1 x)^2 / (2\beta)\} \\ p(x|y) = B(x - \rho_2 y) = (2\pi\gamma)^{-1/2} \exp\{-(x - \rho_2 y)^2 / (2\gamma)\} \end{cases}$$

and $\beta, \gamma > 0$. Since $p(y|x)/p(x|y) = g(y)/f(x)$, we must have

$$(21) \quad \rho_1 / \beta = \rho_2 / \gamma.$$

If we set $x = 0$ and $y = 0$ in the relation $f(x)p(y|x) = g(y)p(x|y)$, we obtain from (20) that

$$f(x) = g(0)(\gamma/\beta)^{-1/2} \exp \left\{ -\frac{x^2}{2} / \frac{\gamma}{1-\rho_1\rho_2} \right\},$$

$$g(y) = f(0)(\beta/\gamma)^{-1/2} \exp \left\{ -\frac{y^2}{2} / \frac{\beta}{1-\rho_1\rho_2} \right\}.$$

Eliminating the constants we finally get

$$p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_1^2} - 2\rho \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right\}$$

where $\sigma_1 \equiv \gamma/(1-\rho_1\rho_2)$, $\sigma_2 \equiv \beta/(1-\rho_1\rho_2)$ and $\rho \equiv (\rho_1\rho_2)^{1/2}$. We note here that since (21) and $\beta, \gamma > 0$ we have $\rho_1\rho_2 > 0$. Thus we have finished the proof.

Appendix D

Solutions of Problems

Problems Section 2

(1) By theorem 2.1 (1)

$$\sum_{i=1}^n \left\{ \left(\frac{x_i}{\sum x_i} \right) \log \left(\frac{x_i / y_i}{\sum x_i / \sum y_i} \right) \right\} \geq 0.$$

(2) We have $I(0:1) = \int f_0 \log \frac{f_0}{f_1} d\lambda = 2 \int f_0^{1/2} f_1^{1/2} \log \frac{f_0^{1/2}}{f_1^{1/2}} d\lambda$

$$\geq 2 \left(\int f_0 d\lambda \right) \log \frac{\int f_0 d\lambda}{\int \sqrt{f_0 f_1} d\lambda} = -2 \log \int \sqrt{f_0 f_1} d\lambda$$

$$\geq 2(1 - \int \sqrt{f_0 f_1} d\lambda),$$

$$\begin{aligned} \int |f_0 - f_1| d\lambda &= \int |f_0^{1/2} - f_1^{1/2}| (f_0^{1/2} + f_1^{1/2}) d\lambda \\ &\leq \left\{ \int (f_0^{1/2} - f_1^{1/2})^2 d\lambda \right\}^{1/2} \left\{ \int (f_0^{1/2} + f_1^{1/2})^2 d\lambda \right\}^{1/2} \\ &= (2 - 2 \int \sqrt{f_0 f_1} d\lambda)^{1/2} (2 + 2 \int \sqrt{f_0 f_1} d\lambda)^{1/2} \\ &\leq 2\sqrt{2} (1 - \int \sqrt{f_0 f_1} d\lambda)^{1/2}, \end{aligned}$$

and

$$\begin{aligned} I(0:1) &= \int f_0 \log \frac{f_0}{f_1} d\lambda \leq \int f_0 \cdot \frac{f_0 - f_1}{f_1} d\lambda \\ &= \int \left(\frac{f_0}{f_1} - 1 \right) \left\{ \left(\frac{f_0}{f_1} - 1 \right) f_1 + f_1 \right\} d\lambda \\ &= \int \left(\frac{f_0}{f_1} - 1 \right)^2 f_1 d\lambda. \end{aligned}$$

(4) We find that

$$A \equiv \frac{I(p_0:p_1) - I(p_1:p_0)}{p_0 - p_1} = \frac{p_0 + p_1}{p_0 - p_1} \log \frac{\frac{p_0 + p_1}{p_0 - p_1} + 1}{\frac{p_0 + p_1}{p_0 - p_1} - 1} - \frac{q_0 + q_1}{q_0 - q_1} \log \frac{\frac{q_0 + q_1}{q_0 - q_1} + 1}{\frac{q_0 + q_1}{q_0 - q_1} - 1}$$

Let $\frac{p_0+p_1}{p_0-p_1} = x$, $\frac{q_0+q_1}{q_0-q_1} = y$. The function $x \log \frac{x+1}{x-1}$ is symmetric, and

convex and decreasing for $x > 1$. Since

$$x - y = \frac{2}{p_0-p_1} \quad \text{and} \quad |x| - |y| = \frac{2(p_0+p_1-1)}{|p_0-p_1|}$$

it follows that, if $(p_0+p_1-1)(p_0-p_1) = p_1q_1 - p_0q_0 > 0$, then

$$y < -1 < 1 < x, \quad |x| > |y| \quad \text{and} \quad A < 0$$

or

$$x < -1 < 1 < y, \quad |x| < |y| \quad \text{and} \quad A > 0$$

both of which yield $I(p_0:p_1) < I(p_1:p_0)$. Similarly the case in which

$(p_0+p_1-1)(p_0-p_1) = p_1q_1 - p_0q_0 < 0$ gives $I(p_0:p_1) > I(p_1:p_0)$.

(5) By corollary to theorem 2.1 the inequality

$$I(f_0:f_1) \geq \sup_{\{E_j\}} \sum_j \mu_0(E_j) \log \frac{\mu_0(E_j)}{\mu_1(E_j)}$$

is evident. We show the reverse inequality.

From the existence of $I(f_0:f_1)$ we can find a K such that

$$(*) \quad -\frac{\epsilon}{2} \leq \mu_0\left\{\left|\log \frac{f_0(x)}{f_1(x)}\right| > K\right\} \log \frac{\mu_0\left\{\left|\log \frac{f_0(x)}{f_1(x)}\right| > K\right\}}{\mu_1\left\{\left|\log \frac{f_0(x)}{f_1(x)}\right| > K\right\}} \leq \frac{\epsilon}{2}$$

Let $\left\{\left|\log \frac{f_0(x)}{f_1(x)}\right| \leq K\right\} = \sum_{i=1}^n E_i$, $\left\{\left|\log \frac{f_0(x)}{f_1(x)}\right| > K\right\} = E_{n+1}$, with n

sufficiently large such that

$$\log \bar{h}_i - \log \underline{h}_i \leq \frac{\epsilon}{2}, \quad i = 1, \dots, n$$

where $\underline{h}_i = \inf_{x \in E_i} \frac{f_0(x)}{f_1(x)}$ and $\bar{h}_i = \sup_{x \in E_i} \frac{f_0(x)}{f_1(x)}$.

Since

$$\mu_0(E_i) \log h_i \leq \left\{ \begin{array}{l} \mu_0(E_i) \log \frac{\mu_0(E_i)}{\mu_1(E_i)} \\ \int_{E_i} \log \frac{f_0(x)}{f_1(x)} d\mu_0 \end{array} \right\} \leq \mu_0(E_i) \log \bar{h}_i,$$

we have

$$\begin{aligned} & \left| \sum_{i=1}^n \mu_0(E_i) \log \frac{\mu_0(E_i)}{\mu_1(E_i)} - \int_{f_1^{-1}} \log \frac{f_0(x)}{f_1(x)} d\mu_0 \right| \\ & \leq \sum_{i=1}^n \left| \mu_0(E_i) \log \frac{\mu_0(E_i)}{\mu_1(E_i)} - \int_{E_i} \log \frac{f_0(x)}{f_1(x)} d\mu_0 \right| \\ & \leq \sum_{i=1}^n \mu_0(E_i) (\log \bar{h}_i - \log h_i) \leq \frac{\epsilon}{2}. \end{aligned}$$

This, together with (*) yields

$$\sum_{i=1}^{n+1} \mu_0(E_i) \log \frac{\mu_0(E_i)}{\mu_1(E_i)} \geq \int_{f_1^{-1}} \log \frac{f_0(x)}{f_1(x)} d\mu_0 - \epsilon$$

from which we obtain the desired inequality.

(8) The right half of the inequalities is evident. To show the left half, we note that

$$\sum_j \nu_0^{(N)}(G_j) \log \frac{\nu_0^{(N)}(G_j)}{\nu_1^{(N)}(G_j)} \leq I(\nu_0^{(N)}; \nu_1^{(N)})$$

where the sum is taken over any set of pairwise disjoint $\{G_j\}$ such that $U = \cup_j G_j$.

Problems Section 4.

(1) Take $S_1 = \{t|t>0\}$, $S_2 = \{t|t<0\}$ in corollary to theorem 2.1

$$(3) \quad (a) \quad \inf_{\theta_0 \in \omega_0} I(\theta : \theta_0) = \begin{cases} 0 & , \quad \theta \in \omega_0 \\ (\theta + \delta)^2 / 2 & , \quad \theta \notin \omega_0 \end{cases}$$

$$\inf_{\theta_1 \in \omega_1} I(\theta : \theta_1) = \begin{cases} (\theta - \delta)^2 / 2 & , \quad \theta \in \omega_1 \\ 0 & , \quad \theta \notin \omega_1 \end{cases}$$

Thus it follows that

$$E_0(n) \geq \sup_{0 < \zeta < 1} - \frac{2}{\delta^2} \log \{ \alpha^\zeta (1-\alpha)^{1-\zeta} + (1-\alpha)^\zeta \alpha^{1-\zeta} \}$$

$$= \left[\quad \quad \quad \right]_{\zeta = \frac{1}{2}} = - \frac{1}{\delta^2} \log \{ 4\alpha(1-\alpha) \} (= M, \text{ say}).$$

(c) Since

$$\sup_{\theta \in \omega_1} P_\theta (\text{Test accepts } H_0) = 1 - P_\theta \left(\bar{X} + \delta \geq \frac{U}{\sqrt{m}} \right)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{U_\alpha - 2\sqrt{m}\delta} \exp\left\{-\frac{t^2}{2}\right\} dt,$$

then, in order for the above expression to be

$$\leq \alpha = \frac{1}{\sqrt{2\pi}} \int_{U_\alpha}^{\infty} \exp\left\{-\frac{t^2}{2}\right\} dt$$

we must have $U_\alpha - \sqrt{m}\delta \leq 0$.

Problems Section 5.

(4) (a) Since

$$p(x) = \int_0^{\infty} p(\theta)p(x|\theta)d\theta$$

$$= \begin{cases} 0, & x < -\frac{\epsilon}{2} \\ \int_0^{x+\frac{\epsilon}{2}} \frac{a}{\epsilon} e^{-a\theta} d\theta = \frac{1}{\epsilon}(1-e^{-a(x+\frac{\epsilon}{2})}), & -\frac{\epsilon}{2} < x < \frac{\epsilon}{2} \\ \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} \frac{a}{\epsilon} e^{-a\theta} d\theta = \frac{e^{-ax}}{\epsilon}(e^{a\epsilon/2} - e^{-a\epsilon/2}), & x > \frac{\epsilon}{2} \end{cases}$$

it follows that

$$T(\theta;x) = \left(\int_{-\frac{\epsilon}{2}}^{\frac{\epsilon}{2}} dx \int_0^{x+\frac{\epsilon}{2}} + \int_{\frac{\epsilon}{2}}^{\infty} dx \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} \right) p(\theta,x) \log \frac{p(\theta,x)}{p(\theta)p(x)} d\theta$$

$$= - \int_{-\frac{\epsilon}{2}}^{\frac{\epsilon}{2}} dx \int_0^{x+\frac{\epsilon}{2}} \frac{ae^{-a\theta}}{\epsilon} \log(1 - e^{-a(x+\frac{\epsilon}{2})}) d\theta$$

$$- \int_{\frac{\epsilon}{2}}^{\infty} dx \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} \frac{a}{\epsilon} e^{-a\theta} \log(e^{-ax}(e^{a\epsilon/2} - e^{-a\epsilon/2})) d\theta$$

$$= - \frac{1}{a\epsilon} \int_0^{1-e^{-a\epsilon}} \frac{z \log z}{1-z} dz - \frac{1}{a\epsilon} \int_0^{1-e^{-a\epsilon}} \log u du$$

$$(z = 1 - e^{-a(x+\frac{\epsilon}{2})}, \quad u = e^{-ax}(e^{a\epsilon/2} - e^{-a\epsilon/2}))$$

$$= - \frac{1}{a\epsilon} \int_0^{1-e^{-a\epsilon}} \frac{\log z}{1-z} dz = - \frac{1}{a\epsilon} \int_{e^{-a\epsilon}}^1 \frac{\log(1-t)}{t} dt$$

$$= \frac{1}{a\epsilon} \sum_{j=1}^{\infty} \left(\frac{1-e^{-ja\epsilon}}{j^2} \right)$$

(6) (a) The Fisher information regarding ζ is

$$\begin{aligned} I(\zeta) &= \int \left(-\frac{\partial^2}{\partial \zeta^2} \log f_\zeta(x) \right) f_\zeta(x) d\lambda = - \int \left(\frac{\partial}{\partial \zeta} \log f_\zeta(x) \right)^2 f_\zeta(x) d\lambda \\ &= \int \left(\frac{f_1(x) - f_2(x)}{f_\zeta(x)} \right)^2 f_\zeta(x) d\lambda \\ &= - \frac{1}{\zeta(1-\zeta)} \int \frac{(f_\zeta(x) - f_1(x))(f_\zeta(x) - f_2(x))}{f_\zeta(x)} d\lambda \\ &= \frac{1}{\zeta(1-\zeta)} \left(1 - \int f_1 f_2 / f_\zeta d\lambda \right) \end{aligned}$$

(b) We consider the case where f_1 and f_2 are such that $f_2(x)/f_1(x)$ is continuous, strictly decreasing and $\lim_{x \rightarrow \infty} \frac{f_2(x)}{f_1(x)} = 0$.

Let f_1 , f_2 and ζ be fixed, and let r be the unique root of the equation

$$\frac{f_2(x)}{f_1(x)} = \frac{\zeta}{1-\zeta} \quad \text{if it exists.}$$

If no root exists we define $r = -\infty$ or 0 according as the domain of the densities is the entire real line or the non-negative axis. Under this convention we have, by the bounded convergence theorem,

$$\begin{aligned} S(\zeta) &= \left(\int_{-\infty}^r + \int_r^\infty \right) \frac{f_1(x) f_2(x)}{f_\zeta(x)} d\lambda \\ &= \frac{1}{1-\zeta} \int_{-\infty}^r \frac{f_1}{1 + \frac{\zeta f_1}{(1-\zeta) f_2}} d\lambda + \frac{1}{\zeta} \int_r^\infty \frac{f_2}{1 + \frac{(1-\zeta) f_2}{\zeta f_1}} d\lambda \\ &= \frac{1}{1-\zeta} \sum_{m=0}^{\infty} (-1)^m \int_{-\infty}^r \left(\frac{\zeta f_1(x)}{(1-\zeta) f_2(x)} \right)^m f_1(x) d\lambda + \frac{1}{\zeta} \sum_{m=0}^{\infty} (-1)^m \int_r^\infty \left(\frac{(1-\zeta) f_2(x)}{\zeta f_1(x)} \right)^m d\lambda \end{aligned}$$

since

$$0 \leq \left\{ \begin{array}{l} \frac{(1-\zeta) f_2(x)}{\zeta f_1(x)} \\ \zeta f_1(x) \\ \frac{\zeta f_1(x)}{(1-\zeta) f_2(x)} \end{array} \right\} \leq 1, \quad \text{if} \quad x \left\{ \begin{array}{l} \geq \\ \leq \end{array} \right\} r.$$

Now, for the normal densities with equal variances, we find

$$r = \bar{\mu} + \frac{\sigma^2}{\mu_2 - \mu_1} \log \frac{1-\zeta}{\zeta}, \quad (\bar{\mu} \equiv (\mu_1 + \mu_2)/2)$$

Using the above expansion we obtain the desired expression.'

(7) (a) Let ξ and ν be any two distributions in \mathbb{R}^k and let $\pi = \alpha\xi + (1-\alpha)\nu$,

where $0 < \alpha < 1$. Then

$$\begin{aligned} \pi(x) &= \left(\frac{\pi_1 f_1(x)}{\sum \pi_j f_j(x)}, \dots, \frac{\pi_k f_k(x)}{\sum \pi_j f_j(x)} \right) \\ &= \left(\frac{\alpha \xi_1 f_1(x) + (1-\alpha) \nu_1 f_1(x)}{\alpha q(x) + (1-\alpha) r(x)}, \dots \right) \\ &= \frac{\alpha q(x)}{\alpha q(x) + (1-\alpha) r(x)} \left(\frac{\xi_1 f_1(x)}{q(x)}, \dots \right) + \frac{(1-\alpha) r(x)}{\alpha q(x) + (1-\alpha) r(x)} \left(\frac{\nu_1 f_1(x)}{r(x)}, \dots \right) \end{aligned}$$

where $q(x) \equiv \sum \xi_j f_j(x)$ and $r(x) \equiv \sum \nu_j f_j(x)$. It follows that from the concavity of U

$$U(\pi(x)) \geq \frac{\alpha q(x)}{\alpha q(x) + (1-\alpha) r(x)} U(\xi(x)) + \frac{(1-\alpha) r(x)}{\alpha q(x) + (1-\alpha) r(x)} U(\nu(x)).$$

Hence

$$\begin{aligned} E[U(\pi(X)) | \pi] &= \int U(\pi(x)) (\alpha q(x) + (1-\alpha) r(x)) d\lambda \\ &\geq \alpha \int U(\xi(x)) q(x) d\lambda + (1-\alpha) \int U(\nu(x)) r(x) d\lambda \\ &= \alpha E[U(\xi(X)) | \xi] + (1-\alpha) E[U(\nu(X)) | \nu]. \end{aligned}$$

(b) Let $U_n \equiv E[U(\xi(X_1, \dots, X_n)) | \xi]$. Then

$$\begin{aligned} U_n &= E\{E[U(\xi(X_1, \dots, X_{n-1})(X_n)) | \xi(X_1, \dots, X_{n-1})] | \xi\} \\ &= E\{U(\xi(X_1, \dots, X_{n-1})) - I[e, \xi(X_1, \dots, X_{n-1}); U] | \xi\} \\ &= U_{n-1} - E\{I[e, \xi(X_1, \dots, X_{n-1}); U] | \xi\} \leq U_{n-1}. \end{aligned}$$

Problems Section 6.

(3) (a) Assuming $a_1 < a_2$ we have

$$\int_0^\eta F_1(u) du = \begin{cases} 0, & 0 \leq \eta \leq \frac{1-a_2}{1-a_1} \\ (1-a_1)\eta - (1-a_2), & \frac{1-a_2}{1-a_1} < \eta \leq \frac{a_2}{a_1} \\ \eta - 1, & \frac{a_2}{a_1} < \eta < \infty, \end{cases}$$

and

$$R_X(\zeta) = \begin{cases} \zeta, & 0 \leq \zeta \leq \frac{1-a_2}{2-a_1-a_2} \\ (a_1+a_2-1)\zeta + (1-a_2), & \frac{1-a_2}{2-a_1-a_2} < \zeta \leq \frac{a_2}{a_1+a_2} \\ 1-\zeta, & \frac{a_2}{a_1+a_2} < \zeta \leq 1 \end{cases}$$

The graph of $R_X(\zeta)$ is shown by Figure a.

(b) We have

$$R_X(\zeta) = \begin{cases} \zeta, & 0 \leq \zeta \leq \zeta_0 = \frac{\sigma^{-1} e^{-\mu^2/2(\sigma^2-1)}}{1 + \sigma^{-1} e^{-\mu^2/2(\sigma^2-1)}} \\ \zeta + \int_{-\lambda}^{\zeta} \left\{ -\zeta \Phi\left(t - \frac{\mu}{\sigma^2-1}\right) + (1-\zeta) \frac{1}{\sigma} \Phi\left(\frac{t - \frac{\mu\sigma^2}{\sigma^2-1}}{\sigma}\right) \right\} dt, & \zeta_0 < \zeta \leq 1, \end{cases}$$

where $\lambda \equiv \frac{\sigma}{\sigma^2-1} \sqrt{\mu^2 + 2(\sigma^2-1)\log(\sigma\eta)}$ and $\eta = \frac{\zeta}{1-\zeta}$, if $\sigma > 1$; and

$$R_X(\zeta) = \begin{cases} 1-\zeta + \int_{-\lambda'}^{\zeta} \left\{ \zeta \Phi\left(t + \frac{\mu}{1-\sigma^2}\right) - (1-\zeta) \frac{1}{\sigma} \Phi\left(\frac{t + \frac{\mu\sigma^2}{1-\sigma^2}}{\sigma}\right) \right\} dt, & 0 \leq \zeta \leq \zeta_0 \\ 1-\zeta, & \zeta_0 < \zeta \leq 1, \end{cases}$$

where $\lambda' \equiv \frac{\sigma}{1-\sigma^2} \sqrt{\mu^2 - 2(1-\sigma^2)\log(\sigma\eta)}$ and $\eta = \frac{\zeta}{1-\zeta}$, if $\sigma < 1$.

The graph of $R_X(\zeta)$ is shown by Figure b.

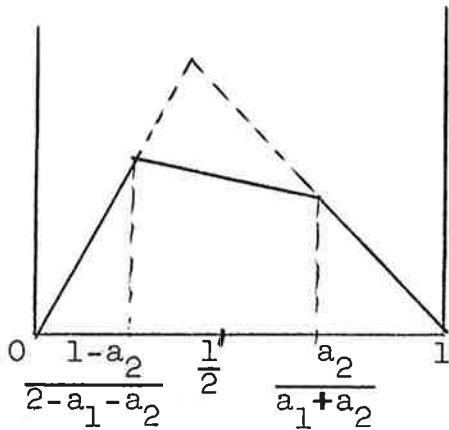


Figure a (Case for $a_1+a_2 < 1$)

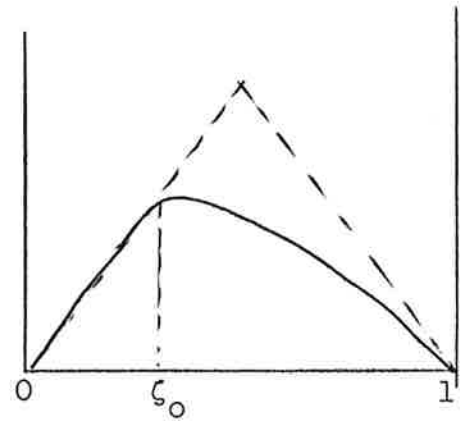


Figure b (Case for $\sigma > 1$)

(4) Take any $\theta_0 \in \Theta$ and $\{\theta^{(n)}\}_{n=1}^{\infty} \subset \Theta$ with $\theta^{(n)} \xrightarrow{(n \rightarrow \infty)} \theta_0$. Let $p^{(n)}(\theta)$ be the prior distribution assigning probability $\frac{1}{2}$ to each of $\theta^{(n)}$ and θ_0 . Then

$$\begin{aligned} I(e_1, p^{(n)}) &= \iint p_1(x|\theta) \log \frac{p_1(x|\theta)}{p_1(x)} dx dp^{(n)}(\theta) \\ &= \frac{1}{2} \int p_1(x|\theta_0) \log \frac{p_1(x|\theta_0)}{\frac{1}{2}(p_1(x|\theta_0)+p_1(x|\theta^{(n)}))} dx \\ &\quad + \frac{1}{2} \int p_1(x|\theta^{(n)}) \log \frac{p_1(x|\theta^{(n)})}{\frac{1}{2}(p_1(x|\theta_0)+p_1(x|\theta^{(n)}))} dx \end{aligned}$$

Thus we get

$$8I(e_1, p^{(n)}) / (\theta^{(n)} - \theta_0)^2 \xrightarrow{(n \rightarrow \infty)} I_1(\theta_0).$$

Taking the limit $n \rightarrow \infty$ in both sides of

$$I(e_1, p^{(n)}) \geq I(e_2, p^{(n)})$$

we obtain

$$I_1(\theta_0) \geq I_2(\theta_0).$$

Problems Section 7.

(2) (a) Let $W_j(\zeta)$ be the expected sum of j observations when ζ is the a priori probability for H_1 and the optimal design is used. Then

$$W_{j+1}(\zeta) = \max \left[\begin{array}{l} X: \{ \zeta E_{f_1} + (1-\zeta) E_{f_2} \} \left\{ X + W_j \left(\frac{\zeta f_1(X)}{\zeta f_1(X) + (1-\zeta) f_2(X)} \right) \right\} \\ Y: \{ \zeta E_{g_1} + (1-\zeta) E_{g_2} \} \left\{ Y + W_j \left(\frac{\zeta g_1(Y)}{\zeta g_1(Y) + (1-\zeta) g_2(Y)} \right) \right\} \end{array} \right]$$

($\zeta = 0, 1, \dots, N-1$; $W_0(\zeta) \equiv 0$). By the hypothetical assumption stated, the right hand side equals

$$\{ \zeta E_{f_1} + (1-\zeta) E_{f_2} \} W_j \left(\frac{\zeta f_1(X)}{\zeta f_1(X) + (1-\zeta) f_2(X)} \right) + \max \left[\begin{array}{l} X: (\zeta E_{f_1} + (1-\zeta) E_{f_2}) \cdot X \\ Y: (\zeta E_{g_1} + (1-\zeta) E_{g_2}) \cdot Y \end{array} \right].$$

(b) Since

$$\log \frac{f_2(X)}{f_1(X)} = \mu X - \frac{\mu^2}{2} \sim \begin{cases} N(-\frac{\mu^2}{2}, \mu^2), & \text{under } H_1 \\ N(\frac{\mu^2}{2}, \mu^2), & \text{under } H_2 \end{cases}$$

$$\log \frac{g_2(Y)}{g_1(Y)} = -\mu Y + \frac{\mu^2}{2} \sim \begin{cases} N(-\frac{\mu^2}{2}, \mu^2), & \text{under } H_1 \\ N(\frac{\mu^2}{2}, \mu^2), & \text{under } H_2 \end{cases}$$

the conditions stated in (a) are satisfied. The optimal design is:

$$\text{Choose } \left\{ \begin{array}{l} X \\ Y \end{array} \right\} \text{ according as } \zeta \left\{ \begin{array}{l} \leq \\ > \end{array} \right\} \frac{1}{2}.$$

(3) For the design D_1 , let α_i denote the probability of obtaining heads on the i -th toss. To avoid trivialities we shall suppose that p and q are not both 0 or both 1; then $|p+q-1| < 1$. It is easy to show that

$$\alpha_{i+1} = (p+q-1)\alpha_i + (p+q-2pq)$$

from which it follows that

$$\alpha_i = (p+q-1)^{i-1} \left(\alpha_1 - \frac{p+q-2pq}{2-p-q} \right) + \frac{p+q-2pq}{2-p-q} \xrightarrow{(i \rightarrow \infty)} \gamma + \frac{\delta^2}{1-\gamma}.$$

Hence in using the design D_1

$$\lim_{N \rightarrow \infty} E \left[\frac{1}{N} \sum_{i=1}^N X_i \mid D_1; p, q \right] = \lim_{N \rightarrow \infty} \left(\frac{\alpha_1 + \dots + \alpha_N}{N} \right) = \gamma + \frac{\delta^2}{1-\gamma},$$

so that

$$\begin{aligned} L(D_1 \mid p, q) &= \max(p, q) - \lim_{N \rightarrow \infty} E \left[\frac{1}{N} \sum_{i=1}^N X_i \mid D_1; p, q \right] \\ &= (\gamma + \delta) - \left(\gamma + \frac{\delta^2}{1-\gamma} \right) = \delta(1-\delta/(1-\gamma)). \end{aligned}$$

The corresponding quantity $L(D_0 \mid p, q)$ is easily seen to have the value $(\gamma + \delta) - \gamma = \delta$.

(4) (a) Suppose for definiteness that $\mu_1 \geq \mu_1'$. If X is used first the expected yield is

$$(*) \mu_1 + \mu_1 \left(\frac{\mu_2}{\mu_1} \right) + (1-\mu_1) \max \left(\frac{\mu_1 - \mu_2}{1-\mu_1}, \mu_1' \right)$$

Since $(*) \geq 2\mu_1 \geq \mu_1 + \mu_1' \geq 2\mu_1'$, X followed by optimal is better than Y followed always by X, which is better than Y followed always by Y.

Of the other two strategies starting with Y, the one requiring X if $Y = 1$ has expected yield $2\mu_1' + \mu_1\mu_1' - \mu_2'$ which can be shown to be less than or equal to that for the strategy requiring Y if $Y = 1$, namely

$$(+)$$

$$\mu_1' + \mu_2' + (1-\mu_1')\mu_1.$$

We have $(*) \geq (+)$, if and only if

$$\text{either } \mu_2 \geq \mu_2', \text{ or } \mu_1 + \mu_1\mu_1' \geq \mu_1' + \mu_2'.$$

Combining these results we get the first half of the statement.

(6) Proposition 2 holds true for any continuous uncertainty functions with $U(\xi) = 0$ for $\xi_1 = 0$ or 1. However if $U(\xi) = -\sum_{i=1}^2 \xi_i \log \xi_i$, Proposition 1 is no longer true.

REFERENCES

- Bellman, R., (1956), "A problem in the sequential design of experiments", Sankhya, 16, 221-229.
- Bellman, R., (1957), Dynamic Programming, Princeton University Press.
- Bessler, S. A., (1960), Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments; Part I, Stanford Univ. Tech. Report, No. 55; Part II, ibid. No. 56.
- Blackwell, D., (1953), "Equivalent comparisons of experiments", Ann. Math. Stat., 24, 265-272.
- Blackwell, D., and Girshick, M. A., (1954), Theory of Games and Statistical Decisions, John Wiley, New York.
- Bradt, R. N., Johnson, S. M., and Karlin, S., (1956), On sequential designs for maximizing the sum of n observations, Ann. Math. Stat., 27, 1060-1074.
- Bradt, R. N., and Karlin, S., (1956), "On the design and comparison of certain dichotomous experiments", Ann. Math. Stat., 27, 390-409.
- Chernoff, H., (1952), "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", Ann. Math. Stat., 23, 492-505.
- Chernoff, H., (1956), "Large-sample theory: parametric case", Ann. Math. Stat., 27, 1-22.
- Chernoff, H., (1959), "Sequential design of experiments", Ann. Math. Stat. 30, 755-770.
- Cramér, H., (1946), Mathematical Methods of Statistics, Princeton Univ. Press.

- DeGroot, M. H., (1962), "Uncertainty, information and sequential experiments", Ann. Math. Stat., 33, 404-419.
- Elfving, G., (1952), "Sufficiency and completeness in decision function theory", Ann. Acad. Sci. Fennicas, 135.
- Feldman, D., (1962), "Contributions to the two-armed bandit problem", Ann. Math. Stat., 33, 847-856.
- Féron, R. and Fouregeaud, C., (1951), "Information et régression", C. R. Acad. Sci. Paris, 232, 1636-1638.
- Hill, B. M., (1963), "Information for estimating the proportions in mixtures of exponential and normal distributions", Journ. Amer. Stat. Assoc., 58, No. 304, 918-932.
- Hoefding, H., (1953), "A lower bound for the average sample number of a sequential test", Ann. Math. Stat., 24, 127-130.
- Huzurbazar, V. S., (1949), "On a property of distribution admitting sufficient statistics", Biometrika, 36, 71-74.
- Joshi, D. D., (1957), "L'information en statistique mathématique et dans la théorie des communications", Thèse, Faculté des Sciences de l'Université de Paris, June.
- Kullback, S., (1959), Information Theory and Statistics, John Wiley, New York.
- Kupperman, M., (1958), "Probabilities of hypotheses and information-statistics in sampling from exponential-class populations", Ann. Math. Stat., 29, 571-575.
- Lindley, D. V., (1956), "On a measure of the information provided by an experiment", Ann. Math. Stat., 27, 986-1005.

- MacGill, W. J., (1954), "Multivariate information transmission", Trans. I.R.E. PGIT - 4, 93-111, Sept.
- MacKay, J. H., (1959), "Asymptotically efficient tests based on the sum of observations", Ann. Math. Stat., 30, 806-813.
- Mallows, C. L., (1959), "The information in an experiment", Jour. Roy. Stat. Soc., Ser. B, 21, 67-72.
- Mourier, E., (1946), "Etude du choix entre deux lois de probabilités", C. R. Acad. Paris, 223, 712-714.
- Paulson, E., (1947), "A note on the efficiency of the Wald sequential test", Ann. Math. Stat., 18.
- Rapaport, A. and Horvath, W. J., (1960), "The theoretical channel capacity of a single neuron as determined by various coding systems", Information and Control, 3, 335-350.
- Robbins, H., (1952), "Some aspects of the sequential design of experiments", Bull. Amer. Math. Soc., 58, 527-535.
- Sakaguchi, M., (1955), "Notes on statistical applications of information theory II", Rep. Stat. Appl. Res., JUSE., 4, 57-68.
p. 26
- Sakaguchi, M., (1957), "Notes on statistical applications of information theory III", Rep. Stat. Appl. Res., JUSE, 5, 9-16.
p. 98
- Sakaguchi, M., (1957), "Notes on information transmission in multivariate probability distributions", Rep Univ. Electro-Comm. No. 9, 25-31.
p. 146
- Sakaguchi, M., (1959), "Notes on statistical applications of information theory IV", Rep. Stat. Appl. Res., JUSE., 6, 54-57.
p. 153

Sakaguchi, M., (1961), "Notes on statistical applications of information
p.198 theory V", Rep. Stat. Appl. Res., JUSE., 8, 173-181.

Stone, M., (1961), "Non equivalent comparisons of experiments and their
use for experiments involving location parameters",
Ann. Math. Stat., 32, 326-332.

Wald, A., (1947), Sequential Analysis, John Wiley, New York.

Weiss, L., (1961), Statistical Decision Theory, McGraw-Hill Book, Co.,
New York.

Phillips R.D. (1963), *Table Useful in Statistics and Information
Theory*, Federal Systems Division,
IBM Corporation, Rockville, Maryland

