LIKELIHOOD RATIOS IN ASYMPTOTIC STATISTICAL THEORY

By

BRIAN GILBERT LEROUX

B.Sc., Carleton University, 1982

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 1985

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date March 18, 1985

ABSTRACT

This thesis deals with two topics in asymptotic statistics. A concept of asymptotic optimality for sequential tests of statistical hypotheses is introduced. Sequential Probability Ratio Tests are shown to have asymptotic optimality properties corresponding to their usual optimality properties. Secondly, the asymptotic power of Pearson's chi-square test for goodness of fit is derived in a new way.

The main tool for evaluating asymptotic performance of tests is the likelihood ratio of two hypotheses. In situations examined here the likelihood ratio based on a sample of size n has a limiting distribution as $n \to \infty$ and the limit is also a likelihood ratio. To calculate limiting values of various performance criteria of statistical tests the calculations can be made using the limiting likelihood ratio.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

The author, being one who thrives on encouragement, wishes to thank Professors Cindy Greenwood and John Petkau for the constant supply they gave.

INTRODUCTION

The motivation behind some of this work lies in a problem concerning a sequential procedure for testing the mean of a normal distribution. The following discussion of this problem follows [3]. There are observed independent identically distributed observations X_1, X_2,... assumed to be distributed as $N(\mu, \sigma^2)$ for a known σ^2. It is required to find a sequential procedure for testing whether μ is positive or negative (sequential procedures are discussed in Section 1.1). The criterion by which procedures are to be judged is the Bayes Risk. This is defined in terms of a cost function having two components, one due to reaching an incorrect conclusion and a second depending on the number of observations on which the conclusion is based. The proposed costs are $K|\mu|$ for making an error (K is a constant) and a cost of c per observation.

The average cost for a given procedure will depend on μ. To avoid problems involved with this it is assumed that μ is a random variable, also with a normal distribution. If its mean and variance are specified the average cost can be averaged further against this distribution for μ. The result is the Bayes Risk.

In the development of a sequential procedure which minimizes the Bayes Risk the partial sums of the observations are replaced by a Brownian motion. This is a reasonable approximation if the number of observations can be expected to be large, and this can be expected when the cost c is small. A procedure which is optimal (minimizes Bayes

Risk) in the continuous time setting is derived and then applied (with a small adjustment) to the discrete time setting. It is desired to have a result stating that this procedure is asymptotically optimal in some sense which can be made precise. Asymptotic here refers to c approaching zero. Results along these lines can be found in [13] where the setting is the more complicated situation of sequential medical trials in which further components of cost are considered (see [4]).

This author attempted to establish similar results using the theory of weak convergence of likelihood ratios which will be discussed in Chapter 1. Success was met only in simple hypothesis testing settings where there are only two possible states of nature. In Chapter 2 are presented discussions of asymptotically optimal sequential procedures which are based on the likelihood ratio. It is believed that the methods used there could be applied successfully in more complicated situations.

Another area for application of likelihood ratio theory lies in the calculation of asymptotic performance of other tests not necessarily based on the likelihood ratio. In Chapter 3 the asymptotic power of chi-square tests is studied via the theory of Chapter 1. It is indicated there that the chi-square test is asymptotically inefficient compared to a test based on the likelihood ratio.

CHAPTER 1

THE THEORY OF LIKELIHOOD RATIOS

1.1 Likelihood Ratios and Hypothesis Testing

We describe the general hypothesis testing problem of distinguishing two probability measures. On a set Ω let there be probability measures P_0 and P_1. A random element X of Ω is chosen and the question is asked: was X chosen based on the distribution P_0 or the distribution P_1? A decision rule for answering the question is a subset D of Ω; if X belongs to D then it is decided that P_1 is the true distribution, otherwise P_0. In common language D is a test of the simple hypothesis $H_0:P_0$ versus the simple hypothesis $H_1:P_1$. D is also called the rejection region because the occurrence of the event D leads to the rejection of the null hypothesis H_0 in favor of the alternative H_1.

Each decision rule has associated with it two error probabilities:

$\alpha(D) = P_0(D)$ = probability of rejecting H_0 when it is true, and

$\beta(D) = P_1(D^c)$ = probability of accepting H_1 when it is false,

called the type I error and type II error respectively. $\alpha(D)$ is also called the level and $1-\beta(D)$ the power of the test D.

Because it is generally impossible to minimize both types of error simultaneously, various criteria for comparing decision rules have been employed. In many cases the best rules are based on the likelihood ratio which we will now define.

Given two probability measures P_0 and P_1 such that P_1 is absolutely continuous with respect to P_0 ($P_1 \ll P_0$), their likelihood ratio is the Radon-Nikodym derivative dP_1/dP_0. This can be generalized by defining the likelihood ratio of any two probability measures P_0, P_1 on a measure space (Ω, F) by

$$Z = \frac{dP_1/d\mu}{dP_0/d\mu} \tag{1.1}$$

where μ is any measure on (Ω, F) such that $P_1 \ll \mu$ and $P_0 \ll \mu$ (such as $\mu = P_0 + P_1$). The conventions $1/0 = \infty$ and $0/0 = 0$ are used in (1.1). In order to show that Z does not depend on the particular choice of μ the following result is needed.

Lebesque Decomposition. For any $A \in F$,

$$\int_A Z \, dP_0 = P_1(A \cap \{Z < \infty\}).$$

Proof: Let $\lambda_0 = \frac{dP_0}{d\mu}$ and $\lambda_1 = \frac{dP_1}{d\mu}$. Then for any $A \in F$

$$\int_A Z \, dP_0 = \int_{A \cap \{Z<\infty\}} Z \, dP_0 = \int_{A \cap \{Z<\infty\}} \lambda_0 Z \, d\mu = \int_{A \cap \{Z<\infty\}} \lambda_1 \, d\mu = P_1(A \cap \{Z < \infty\})$$

since $\{Z = \infty\} = \{\lambda_0 = 0\}$ and $\lambda_0 Z = \lambda_1$ on $\{Z < \infty\}$.

Now because $P_0(\lambda_0 = 0) = 0$, Z is a finite random variable on the probability space (Ω, F, P_0). For any $A \subset \{Z < \infty\}$ the integral $\int_A Z \, dP_0 = P_1(A)$ is determined and so Z is uniquely determined on

(Ω, F, P_0). By symmetry $1/Z$ is uniquely determined on (Ω, F, P_1) and hence Z is also uniquely determined on (Ω, F, P_1). The notation dP_1/dP_0 is used to denote this extension of the Radon-Nikodym derivative and from here on dP_1/dP_0 will denote Z as defined in (1.1).

Example. When P_0 and P_1 are probability distributions on R with densities f_0 and f_1 respectively the ratio of densities

$$Z = f_1/f_0$$

is the likelihood ratio of P_0 and P_1. It need not be assumed that the support of f_0 is contained in the support of f_1.

Since Z is expected to be larger when H_1 is true a reasonable test of H_0 vs. H_1 uses the decision rule

$$D^* = \{Z \geq C\}$$

where C is a constant which determines the level of the test. This rule has the following nice property.

Neyman-Pearson Lemma. If D is any test of H_0 vs H_1 satisfying $\alpha(D) \leq \alpha(D^*)$ then $\beta(D) \geq \beta(D^*)$.

Proof: By the Lebesque Decomposition

$$\beta(D) = P_1(D^c) = \int_{D^c} Z \, dP_0 + P_1(D^c \cap \{Z = \infty\}).$$

Since $D^{*c} \cap \{Z = \infty\} = \emptyset$,

$$P_1(D^c \cap \{Z = \infty\}) \geq P_1(D^{*c} \cap \{Z = \infty\}).$$

Also

$$\int_{D^c} Z \, dP_0 - \int_{D^{*c}} Z \, dP_0 = \int_{D^*} (I_{D^c} - I_{D^{*c}}) \, Z \, dP_0 + \int_{D^{*c}} (I_{D^c} - I_{D^{*c}}) \, Z \, dP_0$$

$$\geq C \int_{D^*} (I_{D^c} - I_{D^{*c}}) \, dP_0 + C \int_{D^{*c}} (I_{D^c} - I_{D^{*c}}) \, dP_0$$

since $Z \geq C$ on D^* and $I_{D^c} - I_{D^{*c}} \leq 0$ on D^{*c}.

Therefore

$$\int_{D^c} Z \, dP_0 - \int_{D^{*c}} Z \, dP_0 \geq C \int (I_{D^c} - I_{D^{*c}}) \, dP_0 = C[P_0(D^c) - P_0(D)^{*c})]$$

$$= C[P_0(D^*) - P_0(D)] \geq 0.$$

This result says that among all rules having type I error at most $P_0(D^*)$, D^* has the smallest type II error. Equivalently, among all rules with type II error at most $P_1(D^{*c})$, D^* has the smallest type I error. A simple and symmetric formulation is:

> no rule can simultaneously have both a smaller type I error and a smaller type II error than D^*.

If $P_0(D^*) = \alpha$ then D^* is called an optimal level-α test. For a given number α it may be impossible to find a number C such that $P_0(Z \geq C) = \alpha$. There will always be a randomized decision rule which

achieves this but these will not be considered here. See [15] for a discussion of randomized rules.

Decision rules having the form of D^* are optimal also in the sense of minimal Bayes Risk. The Bayes Risk for a rule D is

$$\pi\ \alpha(D) + (1 - \pi)\ \beta(D) \tag{1.2}$$

where π is the prior probability of the distribution being P_0. This assumes the 0-1 cost (or loss) function whereby a cost of 1 is incurred when an error of either type is made. Now the expected cost is the probability (under the appropriate hypothesis) of making an error and when this is averaged over the two hypotheses according to the prior probability π, (1.2) results.

For fixed π the Bayes Risk is minimized by D^* with C having the value $\pi/(1 - \pi)$, i.e.

$$\inf_D[\pi\ \alpha(D) + (1 - \pi)\ \beta(D)] = \int[\pi \wedge (1 - \pi)\ Z]\ dP_0$$

and the infimum is achieved at $D_\pi = \{Z \geq \pi/(1 - \pi)\}$. This is proved easily using the Lebesque Decomposition as follows. First,

$$\pi\ \alpha(D_\pi) + (1 - \pi)\ \beta(D_\pi) = \int_{\{Z \geq \pi/(1-\pi)\}} \pi\ dP_0 + \int_{\{Z < \pi/(1-\pi)\}} (1 - \pi)\ Z\ dP_0$$

$$= \int[\pi \wedge (1 - \pi)\ Z]\ dP_0,$$

and for any D

$$\pi\,\alpha(D) + (1-\pi)\,\beta(D) = \int_D \pi\, dP_0 + \int_{D^c} (1-\pi)\, Z\, dP_0 + (1-\pi)\, P_1(D^c \cap \{Z = \infty\})$$

$$\geq \int_D \pi\, dP_0 + \int_{D^c} (1 - \pi)\, Z\, dP_0$$

$$\geq \int [\pi \wedge (1 - \pi)\, Z]\, dP_0.$$

1.2 Sequential Tests of Hypotheses

Wald's Sequential Probability Ratio Test (SPRT) is a procedure for testing

$$H_0 : P_0 \text{ vs. } H_1 : P_1$$

based on a sequence $X_1, X_2, \ldots$ of independent identically distributed (i.i.d.) random variables having distribution either P_0 or P_1. If $X_1, \ldots X_k$ are observed the SPRT uses the statistic

$$Z_k = \prod_{i=1}^{k} Z(X_i) \tag{1.3}$$

where Z is the likelihood ratio dP_1/dP_0. This is reasonable because Z_k is the likelihood ratio of the distribution of $X_1, \ldots, X_k$ under P_1 with respect to the distribution under P_0. The SPRT proceeds as follows:

if $Z_k \geq A$ then H_1 is accepted

if $Z_k \leq B$ then H_0 is accepted

if $B < Z_k < A$ then another observation is taken,

with A and B satisfying $0 < B \leq 1 \leq A < \infty$. This procedure can be expressed in terms of the stopping rule

$$T = \inf \{k: Z_k \geq A \text{ or } Z_k \leq B\} \tag{1.4}$$

and the decision rule

$$D = \{Z_T \geq A\}, \tag{1.5}$$

where Z_T denotes the value of Z_k when $T = k$.

An immediate question arises: can it happen that Z_k never crosses the boundaries determined by A and B? To answer this question probability distributions, corresponding to P_0 and P_1, for infinite sequences $X_1, X_2, \ldots$ must be used. These are the infinite product measures denoted Q_0 and Q_1. The questions above is answered in the negative by

$$Q_0(B < Z_k < A, \quad k = 1,2,\ldots) = 0.$$

$$Q_1(B < Z_k < A, \quad k = 1,2,\ldots) = 0.$$

These statements are implied by the stronger results

$$Z_k \to 0 \text{ a.s. under } Q_0$$

$$Z_k \to \infty \text{ a.s. under } Q_1.$$

A proof of these uses the Strong Law of Large Numbers applied to the sequence $\{-a \vee \log Z(X_i)\}_{i=1}^{\infty}$, where $X \vee Y$ is defined to be the larger of

X and Y. Note that the SPRT could equally well be defined with $\log Z_k$ in place of Z_k. The Strong Law of Large Numbers yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{1}^{n} (-a \vee \log Z(X_i)) = \overline{E}_0(-a \vee \log Z(X_1)) \text{ a.s. } (Q_0).$$

where $\overline{E}_0$ denotes expected value under Q_0. By Jensen's Inequality, provided P_0 and P_1 are distinct in the sense that $P_0(Z = 1) < 1$,

$$E_0(\log Z) = \int \log Z \, dP_0 < \log \int Z \, dP_0 \leq 0.$$

For large enough a then

$$\overline{E}_0(-a \vee \log Z(X_1)) = E_0(-a \vee \log Z) < 0$$

and

$$\lim_{k\to\infty} \sum_{i=1}^{k} (-a \vee \log Z(X_i)) = -\infty \text{ a.s. } (Q_0)$$

$$\lim_{k\to\infty} \sum_{i=1}^{k} \log Z(X_i) = -\infty \text{ a.s. } (Q_0)$$

$$\lim_{k\to\infty} Z_k = 0 \text{ a.s. } (Q_0).$$

By symmetry, $\lim_{k\to\infty} \frac{1}{Z_k} = 0$ a.s. (Q_1) and so $\lim_{k\to\infty} Z_k = \infty$ a.s. (Q_1).

We have just seen that the conditions for stopping in (1.4) will be met eventually, i.e., T is a.s. finite. Now let us compare the SPRT to other sequential tests of H_0 vs. H_1. A sequential test in general

consists of a stopping rule and a decision rule. A stopping rule is a random variable T taking values in the positive integers such that the set $\{T = k\}$ depends only on $X_1,\ldots,X_k$. The decision rule of a sequential test is a set D which depends only on the observations X_i up until the random time T, i.e., for each k, $D \cap \{T = k\}$ is determined by $X_1,\ldots,X_k$.

Criteria for comparing sequential tests include error probabilities and the Average Sample Number (ASN) which is the expected value of the stopping rule.

Only tests with finite ASN will be considered worthwhile; this implies that the conditions for stopping will almost surely be met eventually. The error probabilities are defined exactly as for non-sequential tests,

$$\alpha(T, D) = Q_0(D)$$

and

$$\beta(T, D) = Q_1(D^c).$$

The SPRT has the following optimality property.

Optimality Property of the SPRT. Let $\alpha = Q_0(D)$ and $\beta = Q_1(D^c)$ be the error probabilities of the SPRT defined in (1.4) and (1.5). If (T', D') is any other sequential test of $H_0:P_0$ vs. $H_1:P_1$ with smaller error probabilities,

$$Q_0(D') \leq \alpha;\ Q_1(D'^c) \leq \beta$$

then the SPRT has smaller ASN under both hypotheses,

$$\overline{E}_0(T') \geq \overline{E}_0(T)$$

and $$\overline{E}_1(T') \geq \overline{E}_1(T).$$

(As for $\overline{E}_0$, $\overline{E}_1$ denotes expectation under Q_1).

There have been four strategies for proving this result. The original is due to Wald and Wolfowitz, [23], another is due to Lehmann (see [15] or [11]) and two others ([3], [20]) first prove that SPRTs are Bayes procedures. This latter result is important enough to be stated as a separate result. The Bayes Risk of a sequential test of $H_0:P_0$ vs. $H_1:P_1$ which uses stopping rule T and decision rule D is

$$\rho(T,D;\pi) = \pi(Q_0(D) + c\,\overline{E}_0(T)) + (1 - \pi)(Q_1(D^c) + c\,\overline{E}_1(T)) \qquad (1.6)$$

where π is the prior probability of the true distribution being P_0 and c is the cost per observation. Just as for the Bayes Risk in (1.2) the 0-1 cost function is employed here.

Bayes Optimality of the SPRT: There exist constants A and B which depend on π such that the Bayes Risk (1.6) is minimized by the SPRT which has stopping boundaries A and B.

This property is proved in [20]; in Section 2.2 we will demonstrate how to adapt the argument given there to the continuous time setting.

1.3 Weak Convergence of Likelihood Ratios

For testing simple hypotheses, results based on the likelihood ratio are good tests as measured by the optimality properties we have just seen. For testing a simple null hypothesis against a composite

alternative a reasonable approach consists of choosing one element of the alternative, thus forming a new simple alternative. For example let $X_1,\ldots,X_n$ be a random sample from the $N(\theta,1)$ distribution and consider testing $H_0:\theta = 0$ vs. $H_1:\theta > 0$. One test is based on the likelihood ratio for $H_0:\theta = 0$ vs. $H_1:\theta = \theta_0$ for some fixed $\theta_0 > 0$. In this case the likelihood ratio is

$$Z(x) = \frac{\frac{1}{\sqrt{2\pi}}e^{-(x-\theta_0)^2/2}}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}} = e^{\theta_0 x - \theta_0^2/2} \tag{1.7}$$

and the likelihood ratio based on $X_1,\ldots,X_k$ is

$$Z_k = \prod_1^k Z(X_i) = \exp(\theta_0 \sum_1^k X_i - k\,\theta_0^2/2)$$

Another example involves the parameter θ in the $\exp(1+\theta)$ distribution. The likelihood ratio for $H_0:\theta = 0$ vs. $H_1:\theta = \theta_0$ is

$$Z(x) = \frac{(1+\theta_0)\,e^{-(1+\theta_0)x}}{e^{-x}} = (1+\theta_0)\,e^{-\theta_0 x}, \tag{1.8}$$

and the likelihood ratio based on a random sample $X_1,\ldots X_k$ is

$$Z_k = \prod_1^k Z(X_i) = (1+\theta_0)^k \exp(-\theta_0 \sum_1^k X_i). \tag{1.9}$$

One way of comparing two tests of $H_0:\theta = 0$ vs. $H_1:\theta > 0$ is to look

at their performance for testing H_0 vs. simple alternatives and let the alternatives approach H_0. A common choice for alternative is $H_{1,n}: \theta_0/\sqrt{n}$ where n is the sample size. The reason for this choice is the desire for the test statistic to have a non-degenerate limiting distribution under the alternative. This enables one to calculate limiting (or asymptotic) power and it is by this criterion that tests will be compared. Tests which perform well according to this are considered sensitive to small departures from $\theta = 0$.

For making asymptotic power calculations the limiting distribution of the likelihood ratio is useful in two ways:

1. When measuring the performance of tests based on the likelihood ratio its limiting distribution is essential.
2. The limiting distribution, under the alternative, of other statistics can be found from the joint limiting distribution, under the null, of the statistic and the likelihood ratio.

These two uses are explored in Chapter 2 and Chapter 3. In Chapter 2 tests based on Z are shown to have certain asymptotic optimality properties. In Chapter 3 weak convergence of Z is used to find the limiting distribution of Pearon's chi-square statistic for goodness of fit tests.

Let us examine the asymptotic distribution of the likelihood ratio in the exponential example introduced above. For each n there is a random sample $X_1^n, X_2^n, \ldots$ from the $\exp(1 + \theta)$ distribution. The hypothesis $H_{1,n}$ says that $\theta = \theta_0/\sqrt{n}$ for this sample. The likelihood ratio for $H_0: \theta = 0$ vs. $H_{1,n}$ is, from (1.7),

$$Z^n(x) = (1 + \frac{\theta_0}{\sqrt{n}})\ e^{-(\theta_0/\sqrt{n})\ x}$$

and the likelihood ratio statistic based on $X_1^n,\ldots,X_n^n$ is

$$Z_n^n = \prod_1^n Z^n(X_i^n) = (1 + \frac{\theta_0}{\sqrt{n}})^n \exp(\frac{-\theta_0}{\sqrt{n}} \sum_1^n X_i^n).$$

By a Taylor expansion of $\log(1 + x)$, $\log Z_n^n$ can be written

$$\log Z_n^n = n \log(1 + \frac{\theta_0}{\sqrt{n}}) - \frac{\theta_0}{\sqrt{n}} \sum_1^n X_i^n$$

$$= n(\frac{\theta_0}{\sqrt{n}} - \frac{\theta_0^2}{2n} + O(\frac{1}{n^{3/2}})) - \frac{\theta_0}{\sqrt{n}} \sum_1^n X_i^n$$

$$= \frac{-\theta_0}{\sqrt{n}} \sum_1^n (X_i^n - 1) - \frac{\theta_0^2}{2} + O(\frac{1}{\sqrt{n}})$$

i.e., the remainder term $O(1/\sqrt{n})$ is deterministic and converges to 0 at rate $1/\sqrt{n})$. Now under H_0, $\{X_i^n - 1\}_{i=1}^{\infty}$ is a sequence of i.i.d. mean 0, variance 1 random variables and hence $(1/\sqrt{n}) \sum_1^n (X_i^n - 1) \xrightarrow{d} N(0,1)$ by the Central Limit Theorem. Therefore

$$\log Z_n^n \xrightarrow{d} N(\frac{-\theta_0^2}{2},\ \theta_0^2) \quad \text{under } H_0. \tag{1.10}$$

Similarly under $H_{1,n}$

$$\{(1 + \frac{\theta_0}{\sqrt{n}})\ (X_i^n - \frac{1}{1 + \theta_0/\sqrt{n}})\}_{i=1}^{\infty}$$

is a sequence of i.i.d. mean 0, variance 1 random variables and thus, using the same Taylor expansion,

$$\log Z_n^n = \frac{-\theta_0}{(1 + \theta_0/\sqrt{n})} \frac{(1 + \theta_0/\sqrt{n})}{\sqrt{n}} \sum_1^n (X_i^n - \frac{1}{1 + \theta_0/\sqrt{n}}) - \frac{\theta_0/\sqrt{n}}{1 + \theta_0/\sqrt{n}}$$

$$+ \theta_0/\sqrt{n} - \frac{\theta_0^2}{2} + O(\frac{1}{\sqrt{n}})$$

$$= \frac{-\theta_0}{(1 + \theta_0/\sqrt{n})} \frac{(1 + \theta_0/\sqrt{n})}{\sqrt{n}} \sum_1^n (X_i^n - \frac{1}{1 + \theta_0/\sqrt{n}}) + \frac{\theta_0^2}{1 + \theta_0/\sqrt{n}}$$

$$- \frac{\theta_0^2}{2} + O(\frac{1}{\sqrt{n}}) \xrightarrow{d} -\theta_0\ N(0,1) + \theta_0^2 - \frac{\theta_0^2}{2}$$

where N(0,1) stands for a random variable having that distribution. Therefore

$$\log Z_n^n \xrightarrow{d} N(\frac{\theta_0^2}{2},\ \theta_0^2) \text{ under } H_{1,n}.$$

There is a connection between the limiting distributions of $\log Z_n^n$ under the null and alternative hypotheses with another hypothesis testing problem which is thought of as a "limiting problem." Given

$X \overset{d}{=} N(\theta,1)$ the likelihood ratio statistic for testing

$$H_0:\theta = 0 \text{ vs. } H_1:\theta = \theta_0 \neq 0$$

is

$$Z(X) = \exp(\theta_0 X - \frac{\theta_0^2}{2}).$$

Thus

$$\log Z(X) \overset{d}{=} N(\frac{-\theta_0^2}{2}, \theta_0^2) \qquad \text{under } H_0, \tag{1.11}$$

$$\log Z(X) \overset{d}{=} N(\frac{\theta_0^2}{2}, \theta_0^2) \qquad \text{under } H_1, \tag{1.12}$$

and hence

$$\log Z_n^n \overset{d}{\to} \log Z(X) \qquad \text{under } H_0 \text{ and under } H_{1,n}. \tag{1.13}$$

We will use this fact for evaluating the asymptotic properties of tests based on Z_n^n as the sample size n gets large. The broadest interpretation of (1.13) is that the parameter θ in the $\exp(1 + \theta/\sqrt{n})$ family of distributions plays the same role asymptotically as θ in the $N(\theta,1)$ family. We pursue this idea in Section 1.5.

For evaluating the performance of tests of the form $\{Z_n^n \geq K_n\}$ or equivalently $\{\log Z_n^n \geq C_n\}$ let P_1^n be the $\exp(1 + \theta_0/\sqrt{n})$ distribution and P_0^n be the $\exp(1)$ distribution. In the notation introduced in

Section 1.1, $Z_n^n = dP_1^n/dP_0^n$. Now if the test mentioned previously is to have asymptotic level α, i.e.,

$$P_0^n(\log Z_n^n \geq C_n) \to \alpha \qquad \text{as } n \to \infty$$

then (1.11) and (1.13) imply that

$$\frac{C_n + \theta_0^2/2}{\theta_0} \to Z_\alpha \qquad \text{as } n \to \infty$$

or

$$C_n \to \theta_0 Z_\alpha - \frac{\theta_0^2}{2} \qquad \text{as } n \to \infty$$

where Z_α is the $100(1 - \alpha)$ percentile of the standard normal distribution. Note that from (1.11) and (1.13) it follows that

$$\frac{\log Z_n^n + \theta_0^2/2}{\theta_0} \xrightarrow{d} N(0,1) \qquad \text{under } H_0 \text{ (under } P_0^n).$$

Now by (1.12) and (1.13) the asymptotic power of this test is

$$\lim_{n\to\infty} P_1^n(\log Z_n^n \geq C_n) = \lim_{n\to\infty} P_1^n\left(\frac{\log Z_n^n - \theta_0^2/2}{\theta_0} \geq \frac{C_n - \theta_0^2/2}{\theta_0}\right) \qquad (1.14)$$

$$= 1 - \Phi(Z_\alpha - \theta_0)$$

where Φ is the distribution function of the standard normal

distribution. How does this compare to the asymptotic power of other tests of H_0 vs. $H_{1,n}$ which have asymptotic level α? This can be answered by considering the power of level α tests of $H_0: \theta = 0$ vs. $H_1: \theta = \theta_0$ based on X having distribution $N(\theta,1)$. Now $1 - \Phi(Z_\alpha - \theta_0)$ is the power of the test $\{\frac{\log Z(X) - \theta_0^2/2}{\theta_0} \geq Z_\alpha - \theta_0\}$ or $\{\log Z(X) \geq \theta_0 Z_\alpha - \theta_0^2/2\}$ and by the Neyman-Pearson lemma this test has the greatest possible power.

From this it can be shown (see [10]) that the test $\{\log Z_n^n \geq \theta_0 Z_\alpha - \theta_0^2/2\}$ has the greatest asymptotic power among all tests having asymptotic level α. This is demonstrated in general in [10]; the particular form of the likelihood ratios is not important. We will produce a similar result in the sequential testing situation in Chapter 2. For this purpose it is necessary to study the functional convergence of the likelihood ratio viewed as a stochastic process. We take up this topic next.

1.4 Functional Convergence of Likelihood Ratios

In Section 1.2 sequential procedures for testing simple hypotheses were examined and the SPRT was seen to be optimal in certain ways. The question of asymptotic power against alternatives tending to the null leads to the study of the limiting distribution of the likelihood ratio considered as a process with time measured by observations of the data points. The data $X_1, X_2, \ldots$ are i.i.d. observations from

distribution either P_0 or P_1. The likelihood ratio process $\{Z_k : k=1,2,\ldots\}$ is defined in (1.3).

In the non-sequential case of the previous section a sequence of alternative hypotheses was indexed by the sample size and the alternative grew closer to the null as the sample size increased. With a larger number of observations smaller departures from the null hypothesis can be detected with equal power. In the sequential situation, to detect nearby alternatives many observations are required on average and so it is reasonable to approximate the likelihood ratio process by a continuous time process.

Based on an i.i.d. sequence $X_1^n, X_2^n, \ldots$ with distribution either P_0^n or P_1^n one continuous time version of the likelihood ratio process is

$$Z^n(t) = \prod_{i=1}^{[nt]} Z^n(X_i^n) \tag{1.15}$$

where $Z^n = dP_1^n/dP_0^n$ as defined in (1.1) and $[nt]$ denotes the integer part of nt. Symbolically, the observations X_i^n are associated with the time points i/n.

Example. If $X_1^n, X_2^n, \ldots$ are independent $N(\theta,1)$ random variables and H_0 specifies $\theta = 0$ while $H_{1,n}$ specifies $\theta = \theta_0/\sqrt{n}$ then

$$\log Z^n(t) = \frac{\theta_0}{\sqrt{n}} \sum_{1}^{[nt]} X_i^n - \frac{[nt]}{2n}\,\theta_0^2. \tag{1.16}$$

It is known that processes of this form converge weakly to a Brownian motion (e.g., see Corollary 6 of [16]); in this case we have

$$\log Z^n(t) \xrightarrow{w} \theta_0 B(t) - \frac{\theta_0^2}{2} t \qquad \text{under } H_0 \tag{1.17}$$

and

$$\log Z^n(t) \xrightarrow{w} \theta_0 B(t) + \frac{\theta_0^2}{2} t \qquad \text{under } H_{1,n} \tag{1.18}$$

where $\{B(t):t \geq 0\}$ is a standard Brownian motion. The convergence $\xrightarrow{w}$ takes place in the space $D([0,\infty))$ of right continuous functions with left limits with the Skorohod metric (see [1]). However, because $B(t)$ has continuous sample paths we can use the alternative formulation

$$f(\log Z^n(\cdot)) \xrightarrow{d} f(\theta_0 B(\cdot) - \frac{\theta_0^2}{2}(\cdot))$$

for functionals f continuous with respect to the sup-norm (uniform) metric ([1]).

As in the non-functional case the limiting distributions of the log-likelihood under the null and alternative hypotheses are the distributions of the log-likelihood process for a "limiting hypothesis testing problem." This fact will be used for computation of asymptotic power and the derivation of asymptotically optimal sequential procedures in Chapter 2.

Conditions which guarantee the weak convergence in (1.17) and (1.18) for general likelihood ratios are explored in [12]. One result

states that

$$\log Z^n(t) \xrightarrow{w} B(t) - \frac{1}{2}\lambda t \qquad \text{under } H_0 \tag{1.19}$$

and

$$\log Z^n(t) \xrightarrow{w} B(t) + \frac{1}{2}\lambda t \qquad \text{under } H_1 \tag{1.20}$$

if and only if

$$n \int \left(\sqrt{f_1^n(x)} - \sqrt{f_0^n(x)}\right)^2 dx \to \frac{\lambda}{4} \tag{1.21}$$

and

$$n \int_{A_n(\varepsilon)} \left(\sqrt{f_1^n(x)} - \sqrt{f_0^n(x)}\right)^2 dx \to 0 \tag{1.22}$$

as $n \to \infty$ where $A_n(\varepsilon) = \{x: \left|\sqrt{f_1^n(x)/f_0^n(x)} - 1\right| \geq \varepsilon\}$. Here $\{B(t): t \geq 0\}$ is a Brownian motion with variance λ per unit time (i.e., Var $(B(t)) = \lambda t$.)

More general processes can arise as the limit of a log-likelihood ratio process, including processes with independent normally distributed increments. If we have independent and identically distributed observations $X_1^n, X_2^n, \ldots$ such that (1.21) and (1.22) hold then the limiting process can only be a Brownian motion. The reason for this is clear; if $X_1^n, X_2^n, \ldots$ are i.i.d. then $\log Z^n(t)$ has stationary independent increments because it is formed from partial sums of i.i.d. random variables. If the limiting process $\log Z(t)$ say, has stationary, independent, normally distributed increments it must be a Brownian motion. The limiting processes in (1.19) and (1.20) are Brownian

motions both with variance λ per unit time and with drifts $-\lambda/2$ and $\lambda/2$ per unit time respectively.

1.5 Contguity and Convergence of Experiments

The concept of nearness of null and alternative hypotheses or of families of probability measures is made precise by the notions of contiguity and convergence of experiments.

A sequence $\{P_1^n\}$ of probability measures is said to be contiguous to another sequence $\{P_0^n\}$ (written $P_1^n \triangleleft P_0^n$) if for any sequence of events $\{A^n\}$

$$\lim_{n\to\infty} P_0^n(A^n) = 0 \text{ implies } \lim_{n\to\infty} P_1^n(A^n) = 0.$$

Discussion of contiguity and its uses can be found in [12] and [21]. Contiguity has a close relationship with weak convergence of the likelihood ratio $Z^n = dP_1^n/dP_0^n$. In the case that Z^n has a limiting distribution under the null hypothesis contiguity is equivalent to the existence of a limiting distribution for Z^n under the alternative hypothesis ([12]).

In order that asymptotic power be non-degenerate the sequence of alternatives must be contiguous to the sequence of null hypotheses. Typically in the absence of contiguity there will exist tests with arbitrarily small error probabilities for sufficiently large sample

sizes. This was the case in Section 1.2 where the likelihood ratio had a degenerate limit because both the null and alternative did not change.

When a composite hypothesis is specified the testing problem cannot be described in terms of contiguity or the likelihood ratio of two sequences of probability measures. A means of comparing more than two sequences of probabilities at one time is needed. Convergence of experiments describes the nearness of families of probability distributions. An experiment refers to a family $E = \{P_\theta\}$ of probability distributions. A sequence of experiments $E^n = \{P^n_\theta\}$ is said to converge to E (written $E_n \to E$) if for every finite set $\{\theta_1,\ldots,\theta_m\}$ the vector $(dP^n_{\theta_1}/d\mu^n,\ldots,dP^n_{\theta_m}/d\mu^n)$ converges in distribution, under μ^n, to $(dP_{\theta_1}/d\mu,\ldots,dP_{\theta_m}/d\mu)$ under μ, where $\mu^n = \sum_1^m P^n_{\theta_i}$ and $\mu = \sum_1^m P_{\theta_i}$.

In the case of binary experiments (those that contain two distributions) convergence of experiments coincides with weak convergence of the likelihood ratio.

Convergence of experiments is the essential hypothesis of the Hajek-LeCam minimax theorem ([17]). This is one example of its application to composite hypothesis testing.

An example of convergence of experiments is given by the family $E^n = \{\exp(1 + \theta/\sqrt{n}):\theta\epsilon R\}$ which has limiting experiment $E = \{N(\theta,1):\theta\epsilon R\}$. This fact is suggested (but not proven) by the one-dimensional weak convergence in (1.13).

An interesting way of thinking about convergence of experiments is

as an extension of the likelihood principle. The likelihood principle (see [6]) says that all inference about the family $\{P_\theta\}$ should be based on the likelihood function

$$L(\theta) = \frac{dP_\theta}{d\mu}$$

when there is a measure μ such that $P_\theta \ll \mu$ for all θ. An extension of this might say that when $\{P_\theta^n\} \to \{P_\theta\}$ (in the sense defined above) all inference about $\{P_\theta^n\}$ should be based on the likelihood function for $\{P_\theta\}$ when n is large.

CHAPTER 2

ASYMPTOTIC OPTIMALITY OF SEQUENTIAL PROCEDURES

2.1 Wald's Criterion

Let X_1^n, X_2^n,... be i.i.d. random variables with common distribution either P_0^n or P_1^n and let the likelihood ratio process $\{Z^n(t):t \geq 0\}$ be given by

$$Z^n(t) = \prod_{i=1}^{[nt]} \frac{dP_1^n}{dP_0^n}(X_i^n)$$

with $Z^n = \frac{dP_1^n}{dP_0^n}$ defined by (1.1). Assume that the process Z^n has the asymptotic behaviour discussed in Section 1.4, namely

$$\log Z^n(t) \xrightarrow{w} B(t) - \frac{\lambda}{2} t \qquad \text{under } P_0^n, \qquad [2.1]$$

and

$$\log Z^n(t) \xrightarrow{w} B(t) - \frac{\lambda}{2} t \qquad \text{under } P_1^n, \qquad [2.2]$$

where $\{B(t):t \geq 0\}$ is a Brownian Motion with variance λ per unit time (i.e., Var $B(t) = \lambda t$).

A test will be defined using the limiting process and this test will be shown to be asymptotically optimal when applied to testing $H_0:P_0^n$ vs. $H_{1,n}:P_1^n$. This is an extension to the sequential testing situation of the similar result discussed in Section 1.3.

As a first step we recognize the limiting process in (2.1) and (2.2) as likelihood ratios. Let P_0 and P_1 be the distributions on $C([0,\infty))$ of the processes $\{B(t):t \geq 0\}$ and $\{B(t) + \lambda t:t \geq 0\}$ respectively and let

$$Z(t) = \frac{dP_{1,t}}{dP_{0,t}}$$

where $P_{0,t}$ and $P_{1,t}$ are the restrictions of P_0 and P_1 to $C([0,t])$. It is shown in [10] that

$$\log Z(t) \stackrel{d}{=} B(t) - \frac{\lambda}{2} t \quad \text{under } P_0 \tag{2.3}$$

$$\log Z(t) \stackrel{d}{=} B(t) + \frac{\lambda}{2} t \quad \text{under } P_1. \tag{2.4}$$

If H_0 represents P_0^n and P_0 and H_1 represents P_1^n and P_1 then we have the weak convergence of the processes

$$Z^n \stackrel{w}{\to} Z \qquad \text{under } H_0 \text{ and under } H_1. \tag{2.5}$$

The Sequential Probability Ratio Test (SPRT) for testing H_0 vs. H_1 uses the stopping rule

$$T^* = \inf\{t: Z(t) \leq B \text{ or } Z(t) \geq A\} \tag{2.6}$$

and decision rule

$$D^* = \{Z(T^*) \geq A\}. \tag{2.7}$$

It can be shown ([8]) that T^* is finite under both hypotheses; thus when the event D^* does not occur $Z(T^*) \leq B$ and H_0 is accepted. The SPRT has the same optimality property in continuous time as it does in discrete time. A sequential procedure for testing H_0 vs. H_1 consists in general of a stopping rule T which takes values in $[0,\infty]$ such that the event $\{T \leq t\}$ is determined by $\{B(s):0 \leq s \leq t\}$ and a decision rule D which must be such that $D \cap \{T \leq t\}$ is determined by $\{B(s):0 \leq s \leq t\}$, for each $t \in [0,\infty]$.

Optimality Property of Continuous Time SPRT ([8]): Assume that for each n, Z^n has a continuous distribution under P_0^n. If a sequential test (T,D) of $H_0:P_0$ vs. $H_1:P_1$ has smaller error probabilities than (T^*,D^*) defined by (2.6) and (2.7), i.e.,

$$P_0(D) \leq P_0(D^*) \text{ and } P_1(D^c) \leq P_1(D^{*c})$$

then (T,D) must have higher average sample numbers (ASN),

$$E_0(T) \geq E_0(T^*) \text{ and } E_1(T) \geq E_1(T^*).$$

We will now prove a result (stated more precisely below) which says that this optimality property is preserved in the limit when the SPRT is applied to the discrete time setting. Consider the procedure (T_n^*, D_n^*) given by

$$T_n^* = \inf\{t: Z^n(t) \geq A \text{ or } Z^n(t) \leq B\}.$$

and $$D_n^* = \{Z^n(T_n^*) \geq A\}.$$

To study the asymptotic properties of (T_n^*, D_n^*) the following results are used

$$T_n^* \xrightarrow{d} T^* \quad \text{under } H_0 \text{ and under } H_{1,n} \tag{2.8}$$

$$Z^n(T_n^*) \xrightarrow{d} Z(T^*) \quad \text{under } H_0 \text{ and under } H_{1,n} \tag{2.9}$$

These follow from the fact that T^* and $Z(T^*)$ are continuous functionals of $\{Z(t):t \geq 0\}$ relative to the sup-norm metric and the weak convergence (2.5) holds with respect to this metric (see Section 1.4). From (2.9) it follows immediately that the asymptotic error probabilities of (T_n^*, D_n^*) are equal to the error probabilities of the SPRT (T^*, D^*), i.e.

$$\lim_{n\to\infty} P_0^n(D_n^*) = P_0(D^*) \tag{2.10}$$

and $$\lim_{n\to\infty} P_1^n(D_n^{*c}) = P_1(D^{*c}). \tag{2.11}$$

The same result for the average sample numbers requires the uniform integrability of $\{T_n^*\}$; this is demonstrated in Appendix 1, thus

$$\lim_{n\to\infty} E_0^n(T_n^*) = E_0(T^*) \tag{2.12}$$

and $$\lim_{n\to\infty} E_1^n(T_n^*) = E_1(T^*). \tag{2.13}$$

The asymptotic optimality result can now be stated.

Asymptotic Optimality Property of (T_n^*, D_n^*): Assume that $P_1(D) > 0$. If (T_n, D_n) is any sequential test of H_0 vs. H_1 satisfying

$$\overline{\lim_{n\to\infty}} P_0^n (D_n) \leq P_0(D^*) \tag{2.14}$$

and

$$\overline{\lim_{n\to\infty}} P_1^n (D_n^c) \leq P_1(D^*) \tag{2.15}$$

then

$$\underline{\lim_{n\to\infty}} E_0^n(T_n) \geq E_0(T^*) \tag{2.16}$$

and

$$\underline{\lim_{n\to\infty}} E_1^n(T_n) \geq E_1(T^*). \tag{2.17}$$

A proof of this result will now be given. First we find a SPRT which has the same error probabilities as (T_n, D_n). This is where the assumption of continuity of the distribution of Z^n is needed; it implies the existence of the required SPRT. We state the needed result from [24]:

Lemma. Assume that $Z^n = dP_1^n/dP_0^n$ has a continuous distribution. If α_0 and α_1 are non-negative numbers with $\alpha_0 + \alpha_1 \leq 1$ there exist A_n and B_n such that the SPRT with stopping boundaries A_n and B_n has error probabilities α_0 and α_1.

In order that the lemma applies the error probabilities of (T_n, D_n) must satisfy the constraint $\alpha_0 + \alpha_1 \leq 1$. Since we have assumed that $P_0(D^*) + P_1(D^{*c}) < 1$, (2.14) and (2.15) imply that for large n we will have $P_0^n(D_n) + P_1^n(D_n^c) < 1$ as required. Since only the tail of the sequence affects (2.16) and 2.17) we can assume without loss of generality that this inequality holds for all n.

Now let (T_n', D_n') be the SPRT determined by the Lemma, that is

$$T_n' = \inf\{t: Z^n(t) \geq A_n \text{ or } Z^n(t) \leq B_n\},$$

$$D_n' = \{Z^n(T_n') \geq A_n\}.$$

By the optimality property of (T_n', D_n') it must have lower ASN than $T_n, D_n)$; thus it will suffice to show (2.16) and (2.17) with (T_n', D_n') in place of (T_n, D_n). Because of the inequalities (2.14) and (2.15) for (T_n', D_n'), the sequences $\{A_n\}$ and $\{B_n\}$ must be bounded. If, say, $\{A_n\}$ was not bounded above then $\lim_{n\to\infty} P_1^n(D_n') = 0$ and this contradicts (2.15). By considering a subsequence if necessary assume $\{A_n\}$ and $\{B_n\}$ converge, say $\lim A_n = A'$, $\lim B_n = B'$. Now if one of (2.16), (2.17) did not hold the SPRT (T', D') with stopping boundaries A' and B' would be better than the optimal procedure (T^*, D^*) i.e., $P_0(D') \leq P_0(D^*)$, $P_1(D'^c) \leq P_1(D^{*c})$, $E_0(T') \leq E_0(T^*)$ and $E_1(T') \leq E_1(T^*)$ with strict

inequality in one of the last two inequalities. This contradicts the optionality property of (T^*, D^*).

2.2 Bayes Risk Criterion

In this section a different criterion for comparing sequential testing procedures is used, the Bayes risk. We begin with the set up described in the first paragraph of Section 2.1. For a sequential test of $H_0:P_0^n$ vs. $H_1:P_1^n$, say (T,D), we define the Bayes Risk, just as in (1.6), by

$$\rho_n(T,D;\pi) = \pi(P_0^n(D) + cE_0^n(T)) + (1 - \pi)(P_1^n(D^c) + cE_1^n(T)), \qquad (2.22)$$

where π is the prior probability of the distribution being P_0 and c is the cost per observation.

For a sequential test (T,D) of $H_0:P_0$ vs. $H_1:P_1$, where P_0 and P_1 are as in the previous section, the Bayes Risk is

$$\rho(T,D;\pi) = \pi(P_0(D) + cE_0(T)) + (1 - \pi)(P_1(D^c) + cE_1(T)), \qquad (2.23)$$

Here c represents the cost per unit time of observing the likelihood ratio process Z(t).

We will now solve the problem of minimizing ρ over all continuous time sequential tests. Our derivation will mimic the strategy used in [20] for deriving the same result in discrete time; the appropriate theory for the continuous time case corresponding to Snell's envelope is given in [20] and also in [9]. The solution will be a particular SPRT. Although the solution is derived here only for the special case that

P_0 and P_1 are distributions of Brownian motions the same argument will work for more general situations. In particular it will work under the general conditions of [8] which are used there for obtaining the previous optimality property of continuous time SPRTs given in Section 2.1.

To begin it will be necessary to consider the equivalent problem of minimizing

$$\rho^{(r)}(T,D;\pi) = \pi(P_0(D) + cE_0(T)) + r(1 - \pi)(P_1(D^c) + cE_1(T)),$$

allowing the new parameter r to vary. The first step consists of fixing a stopping rule and finding the best decision rule to go with it.

Lemma. If T is fixed

$$\min_D[\pi\, P_0(D) + r(1 - \pi)\, P_1(D^c)] = E_0(\pi \wedge r(1 - \pi)\, Z(T))$$

and the minimum is achieved at $D_* = \{Z(T) \geq \pi/r(1 - \pi)\}$.

Proof: First we note that $D \cap \{T \leq t\}$ is determined by $\{B(s):0 \leq s \leq t\}$, for all t, and $Z(T)$ equals dP_1/dP_0 on the σ-field of such events. By the Lebesque Decomposition,

$$\pi\, P_0(D) + r(1 - \pi)\, P_1(D^c) = \int_D \pi\, dP_0 + \int_{D^c} r(1 - \pi)\, Z(T)\, dP_0$$

$$\geq \int[\pi \wedge r(1 - \pi)\, Z(T)]\, dP_0.$$

It is straightforward to check that there is equality here for $D = D_*$.

The problem is now reduced to minimizing

$$\pi \, c \, E_0(T) + r(1 - \pi) \, c \, E_1(T) + E_0(\pi \wedge r(1 - \pi) \, Z(T))$$

$$= E_0(\pi \, c \, T + r(1 - \pi) \, c \, T \, Z(T) + \pi \wedge r(1 - \pi) \, Z(T))$$

$$= E_0(Y^{(r)}(T))$$

where the process $\{Y^{(r)}(t):t \geq 0\}$ is defined by

$$Y^{(r)}(t) = \pi \, c \, t + r(1 - \pi) \, c \, t \, Z(t) + \pi \wedge r(1 - \pi) \, Z(t).$$

According to Theorem 7.3 in [20] or Theorem 4 in [9] this can be done by finding the largest positive sub-martingale, say $\{V^{(r)}(t):t \geq 0\}$ dominated by $\{Y^{(r)}(t):t \geq 0\}$ and then forming the stopping rule

$$T^* = \inf\{t: Y^{(r)}(t) = V^{(r)}(t)\}. \tag{2.24}$$

The process $V^{(r)}$ is given by

$$V^{(r)}(t) = \operatorname{essinf} E(Y^{(r)}(T) \,|\, B(s):0 \leq s \leq t) \tag{2.25}$$

where the essinf is taken over all stopping rules T which satisfy $T \geq t$.

Since $Z(0) = 1$, the initial value $V^{(r)}(0)$ is deterministic and

$$V^{(r)}(0) = \inf E(Y^{(r)}(T)) \equiv h(r). \tag{2.26}$$

Note that h is an increasing concave function because it is the infimum of such functions. This fact will be important for determining the nature of the solution.

For any stopping rule T satisfying $T \geq t$

$$E(Y^{(r)}(T)\,|\,B(s):0 \le s \le t) = E(\pi c(T-t) + r((1-\pi)\,Z(t)\,c(T-t)\,\frac{Z(T)}{Z(t)}$$

$$+ \pi \wedge r(1-\pi)\,Z(t)\,\frac{Z(T)}{Z(t)}\,|\,B(s):0 \le s \le t)$$

$$+ \pi\, ct + r(1-\pi)\, ct\, Z(t) \qquad (2.27)$$

where we have used the fact that $E(Z(T)\,|\,B(s):0 \le s \le t) = Z(t)$ (i.e., Z is a martingale). Using the representation $Z(t) = e^{B(t) - \frac{\lambda}{2} t}$.

$$\frac{Z(u)}{Z(t)} = e^{B(u) - B(t) - \frac{\lambda}{2}(u-t)}$$

and since $\{B(t)\}$ has stationary independent increments, the process $\{\frac{Z(u)}{Z(t)}:u \ge t\}$ is independent of $\{B(s) :0 \le s \le t\}$ and has the same distribution as the process $\{Z(u):u \ge 0\}$. Therefore the conditional expectation in (2.27) is minimized exactly as for the case $t = 0$ in (2.26) but with r replaced by r Z(t), i.e.,

$$V^{(r)}(t) = \text{essinf}\, E(Y^{(r)}(T)\,|\,B(s):s \le t)) = h(r\,Z(t)) + \pi ct$$

$$+ r(1-\pi)\, ctZ(t).$$

Now the stopping rule T^* is given by

$$T^* = \inf\{t:Y^{(r)}(t) = V^{(r)}(t)\} = \inf\{t:h(rZ(t)) = \pi \wedge r(1-\pi)\,Z(t)\}.$$

In order for the Bayes Risk given in (2.23) to be minimized by T^*, r is now set to 1. Thus

$$T^* = \inf\{t: h(Z(t)) = \pi \wedge (1 - \pi)\, Z(t)\}.$$

Since h is increasing and concave, T^* has the form

$$T^* = \inf\{t: Z(t) \geq A \text{ or } Z(t) \leq B\}.$$

for constants A and B illustrated below.

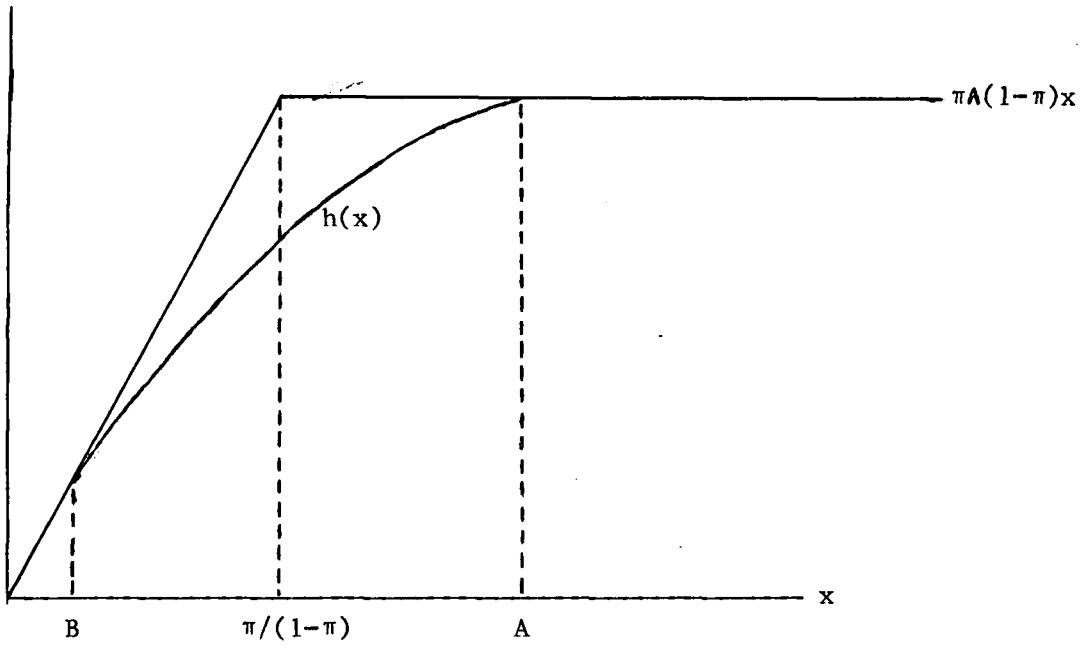

Fig. 1. Graph of h which determines stopping boundaries A, B of optimal SPRT.

If T^* is the stopping rule employed, the decision rules $\{Z(T^*) \geq A\}$ and $\{Z(T^*) \geq \frac{\pi}{1-\pi}\}$ (recall the lemma, pg. 33) are equivalent due to the inequality $B \leq \frac{\pi}{1-\pi} \leq A$. Also, the cases $B \geq 1$ and $A \leq 1$ correspond to $T^* \equiv 0$ in which the initial decision based only on the prior probability is optimal, having Bayes Risk $\pi \wedge (1 - \pi)$.

As in the previous section the optimal procedure for the continuous time problem will be applied to the discrete time setting; a procedure which minimizes the asymptotic Bayes Risk results. Define the stopping rule

$$T_n^* = \inf\{t: Z^n(t) \geq A \text{ or } Z^n(t) \leq B\}$$

where A and B are the stopping boundaries of the SPRT which minimizes the Bayes Risk (2.23) and the decision rule

$$D_n^* = \{Z^n(T_n^*) \geq A\}.$$

Thus (T_n^*, D_n^*) has the asymptotic optimality property given by (2.14) - (2.17). Here it will be shown to have the following property.

Asymptotic Bayes Optimality Property of (T_n^*, D_n^*): The asymptotic Bayes Risk of (T_n^*, D_n^*) is

$$\lim_{n\to\infty} \rho_n(T_n^*, D_n^*; \pi) = \rho(T^*, D^*; \pi). \tag{2.28}$$

If (T_n, D_n) is any sequential test of H_0 vs. H_1 then

$$\lim_{n\to\infty} \inf_{(T,D)} \rho_n(T, D; \pi) \geq \rho(T^*, D^*; \pi). \tag{2.29}$$

This will say that (T_n^*, D_n^*) has the smallest possible asymptotic Bayes Risk and the value is the Bayes Risk of the procedure (T^*, D^*).

The proof of (2.28) is achieved by the application of (2.10), (2.11), (2.12) and (2.13) which state that all of the components of the Bayes Risk $\rho_n(T_n^*, D_n^*; \pi)$ converge to the corresponding components of $\rho(T^*, D^*; \pi)$.

The first step in proving (2.29) is to compute the minimum value of ρ_n. From the discussion preceding the derivation of T^* it is known that ρ_n is minimized by a SPRT with some stopping boundaries, say A_n and B_n, that is

$$\inf_{T,D} \rho_n(T, D; \pi) = \rho_n(T_n, D_n; \pi)$$

where

$$T_n = \inf\{t: Z^n(t) \geq A^n \text{ or } Z^n(t) \leq B^n\}$$

and

$$D_n = \{Z^n(T_n) \geq A^n\}.$$

Assume now that along a subsequence of the integers $\{n'\}$ the limit

$$\lim_{n'} \rho_{n'}(T_{n'}, D_{n'}; \pi)$$

exists and is less than $\rho(T^*, D^*; \pi)$. Within this subsequence there is a further subsequence (also called $\{n'\}$) such that the limits $\lim_{n'} A^{n'} = A'$ and $\lim_{n'} B^{n'} = B'$ exist, possibly infinite. Finally we can

repeat the argument at the end of the previous section, to show that the continuous time SPRT with stopping boundaries A' and B' has lower Bayes Risk than the SPRT which uses A and B. This of course contradicts the fact that A and B were derived to minimize the Bayes Risk (2.23).

CHAPTER 3

POWER OF CHI-SQUARE TESTS

The focus of this section is the asymptotic power of Pearson's chi-square statistic for testing goodness of fit against a certain clas of alternatives. These alternatives are contiguous to the null hypothesis in the sense defined in Section (1.4).

We will reproduce a derivation of the limiting distribution of Pearson's chi-square statistic under the null hypothesis ([7], [19]). In [5] the limiting distribution under a class of alternatives is computed, whereby the asympotic power can be computed. We will give a different development of this result which uses the weak convergence of the likelihood ratio. This highlights the usefulness of the likelihood ratio as a tool for studying hypothesis testing problems. In [18] the limiting alternative distribution is found for situations where a parameter must be estimated.

Also we will compare the asymptotic power of the chi-square test and of a test based on the likelihood ratio. We know from Section 1.3 that the test based on the likelihood ratio must win; the extent of the difference is of interest.

Let $\underline{N} = (N_1,\ldots,N_k)$ be a multinomial random vector which records the numbers of data points which fall into each of k classifications. Let the total number of data points be n and the probability of any one falling into the ith category be P_i. The probability function of $\underline{N}$ is

$$P(N_1=n_1,\ldots,N_k=n_k) = \frac{n!}{n_1!..n_k!} \prod_1^k P_i^{n_i} \qquad (n_i \in Z_+, \sum_1^k n_i=n) \qquad (3.1)$$

A common question asks whether the probability vector $\underline{P} = (P_1,\ldots,P_k)$ belongs to a parametric family $\{\underline{P}(\theta):\theta \in H\}$. This question reflects on the distribution of the underlying data which is usually the source of interest. For example when testing whether a sample $X_1,\ldots,X_n$ came from the Normal distribution, categories $E_i = (a_{i-1}, a_i]$ $(i = 1,\ldots,k)$ could be formed and the numbers $N_i = \#\{X_j \in E_i\}$ of observations falling into these intervals recorded. under the normal distribution the probability vector $\underline{P}$ would be given by

$$P_i = \int_{a_{i-1}}^{a_i} \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} dx$$

$$= \Phi(\frac{a_i - \mu}{\sigma}) - \Phi(\frac{a_{i-1} - \mu}{\sigma})$$

The hypothesis of normality for $X_1,\ldots,X_n$ is also specified by the particular parametric form for $\underline{P}$.

In general a test of the composite hypothesis

$$H_0:\underline{P} = \underline{P}(\theta) \text{ for some } \theta \in H$$

requires estimation of θ. This situation is treated in [18]. We will consider only the simple hypothesis

$$H_0:\underline{P} = \underline{P}(\theta_0)$$

for some specified $\theta_0 \in H$. This is also written as

$$H_0: P_i = P_i^0 \qquad (i=1,\ldots,k) \tag{3.2}$$

where $\underline{P}(\theta_0) = (P_1^0,\ldots,P_k^0)$.

Pearson's chi-square statistic for testing H_0 is

$$X^2(n) = \sum_{i=1}^{k} \frac{(N_i - nP_i^0)^2}{nP_i^0}. \tag{3.3}$$

It will be shown that $X^2(n)$ has a limiting ($n \to \infty$) chi-square distribution with k-1 degrees of freedom. This fact is used for computing critical values of the test. The proof is based on a simple multivariate Central Limit Theorem ([7]) applied to the sequence of random vectors $\underline{V}_n$ given by

$$V_{n,i} = \frac{N_i - nP_i^0}{\sqrt{nP_i^0}}. \tag{3.4}$$

A simple computation produces the covariance matrix of $\underline{V}_n$

$$\mathrm{Cov}(\underline{V}_n) = I_k - \underline{q}\,\underline{q}'$$

where $\underline{q} = (\sqrt{P_1^0},\ldots,\sqrt{P_k^0})'$. Because the N_i are sums of independent identically distributed (Bernoulli) random variables the multivariate

CLT can be applied to yield

$$\underline{V}_n \xrightarrow{d} N_k(\underline{0}, \Lambda) \text{ under } H_0 \tag{3.5}$$

as $n \to \infty$ where $\Lambda = I_k - \underline{q}\,\underline{q}'$. In view of the relation

$$X^2(n) = \underline{V}_n' \underline{V}_n$$

the following result is needed ([7]), [19]).

Proposition 1. If $\underline{Y} \overset{d}{=} N_K(\underline{0}, \Lambda)$ and Λ is idempotent with rank r then

$$\underline{Y}' \underline{Y} \overset{d}{=} \chi^2_r.$$

An application of this to $\underline{V}_n$ (recall (3.5)) gives

$$X^2(n) \xrightarrow{d} \chi^2_{k-1} \tag{3.6}$$

since the covariance matrix $\Lambda = I_k - q\,q'$ of $\underline{V}_n$ is idempotent with rank k-1. Here we have used the continuity of the mapping $\underline{X} \to \underline{X}'\,\underline{X}$.

The statistic $X^2(n)$ is not designed with any specific alternatives to H_0 in mind. The asymptotic power of $X^2(n)$ against the sequence of alternatives

$$H_{1,n}: P_i = P_i^0 + \frac{C_i}{\sqrt{n}}, \tag{3.7}$$

where $\sum_{1}^{k} C_i = 0$, can be calculated. The limiting distribution of $X^2(n)$ under the sequence of alternatives $H_{1,n}$, is non-central chi-square as stated in the following result. The notation $\chi'^2_r(\Delta)$ represents the distribution of $(Z_1 + \sqrt{\Delta})^2 + Z_2^2 + \ldots + Z_r^2$ where $Z_1,\ldots,Z_r$ are i.i.d. standard normal random variables.

Theorem

$$X^2(n) \xrightarrow{d} \chi'^2_{k-1}\left(\sum_{1}^{k} \frac{C_i^2}{p_i^0}\right) \text{ as } n \to \infty. \tag{3.8}$$

One possible proof (as in [18]) uses a multivariate CLT for triangular arrays which establishes

$$\underline{V}_n \xrightarrow{d} N_K(\underline{\delta}, \Lambda) \qquad \text{under } H_{1,n} \qquad \text{as } n \to \infty \tag{3.9}$$

with Λ as in (3.5) and

$$\underline{\delta} = \left(\frac{C_1}{\sqrt{p_1^0}}, \ldots, \frac{C_k}{\sqrt{p_k^0}}\right).$$

This must be combined with the following fact about the multivariate normal distribution which generalizes Proposition 1.

Proposition 2. ([17]) If $\underline{Y} \overset{d}{=} N_k(\underline{\delta}, \Lambda)$ and Λ is idempotent with rank r and $\underline{\delta}$ is in the range (column space) of Λ then

$$\underline{Y}' \underline{Y} \overset{d}{=} \chi_r'^2 (\underline{\delta}' \underline{\delta}).$$

Using this result it is immediate that the Theorem follows from (3.9). A different proof of (3.9) will now be given; it will be based on the likelihood ratio for the simple hypothesis testing problem H_0 vs. $H_{1,n}$. This likelihood ratio is simply a ratio of multinomial probabilities defined by (3.1), namely

$$Z^n = \frac{\prod_1^k (P_i^0 + C_i/\sqrt{n})^{N_i}}{\prod_i^k (P_i^0)^{N_i}} \tag{3.10}$$

In order that Z^n can be used to find the limiting distribution of $\underline{V}_n$ there must be established a relationship between the two. This is done by taking logarithms and using a Taylor expansion as follows:

$$\log Z^n = \sum_1^k N_i \log(1 + \frac{C_i}{P_i^0 \sqrt{n}})$$

$$= \Sigma N_i (\frac{C_i}{P_i^0 \sqrt{n}} - \frac{C_i^2}{2(P_i^0)^2 n} + O(\frac{1}{n^{3/2}}))$$

$$= \frac{1}{\sqrt{n}} \Sigma \frac{C_i}{P_i^0} N_i - \frac{1}{n} \Sigma \frac{C_i^2}{2(P_i^0)^2} N_i + O_p(\frac{1}{\sqrt{n}})$$

$$= \frac{1}{\sqrt{n}} \Sigma \frac{C_i}{P_i^0} (V_{n,i} \sqrt{nP_i^0} + nP_i^0) - \frac{1}{2} \Sigma \frac{C_i^2}{P_i^{0\,2}} (P_i^0 + \frac{N_i}{n} - P_i^0) + O_p(\frac{1}{\sqrt{n}})$$

$$= \Sigma \frac{C_i}{\sqrt{P_i^0}} V_{n,i} - \frac{1}{2} \Sigma \frac{C_i^2}{P_i^0} + O_p(1)$$

$$= \underline{\delta}' \underline{V}_n - \frac{1}{2} \underline{\delta}' \underline{\delta} + O_p(1). \tag{3.11}$$

where $O_p(\frac{1}{\sqrt{n}})$ is a term which converges to 0 in probability at rate $1/\sqrt{n}$, and $O_p(1)$ converges to 0 in probability. The latter term arises because $N_i/n - P_i^0 = O_p(1)$.

Now we use the following strategy. There is a "limiting" simple hypothesis testing problem which approximates the multinomial testing problem H_0 vs. $H_{1,n}$ in such a way that there is a quantity which assumes the role of $\underline{V}_n$. The distribution of this quantity, under null and under alternative, is the limiting distribution of $\underline{V}_n$ under H_0 and under $H_{1,n}$, respectively. The limiting problem is

$$H_0: N_k(\underline{0}, \Lambda) \text{ vs. } H_1: N_k(\underline{\delta}, \Lambda).$$

The likelihood ratio based on a single observation $\underline{X}$ is (see Appendix A.2)

$$Z(\underline{X}) = \exp(\underline{\delta}' \underline{X} - \frac{1}{2} \underline{\delta}' \underline{\delta}).$$

Comparing the equation

$$\log Z(\underline{X}) = \underline{\delta}' \underline{X} - \frac{1}{2} \underline{\delta}' \underline{\delta} \tag{3.12}$$

with (3.11) it is seen that $(Z^n, \underline{V}_n)$ are related by the same equation as $(Z(\underline{x}), \underline{x})$, except for the error term $O_p(1)$. Also

$$\log Z^n \xrightarrow{d} \log Z \quad \text{under } H_0 \tag{3.13}$$

as $n \to \infty$. The limiting distribution of $\log Z^n$ is calculated from (3.11) and (3.5) as

$$\begin{aligned} \log Z^n &\xrightarrow{d} \underline{\delta}' N_k(\underline{0}, \Lambda) - \frac{1}{2} \underline{\delta}' \underline{\delta} \quad \text{under } H_0 \\ &\stackrel{d}{=} N(-\frac{1}{2} \underline{\delta}' \underline{\delta}, \underline{\delta}' \underline{\delta}) \end{aligned} \tag{3.14}$$

(since the variance of $\underline{\delta}' N_k(\underline{0}, \Lambda)$ is $\underline{\delta}' \Lambda \underline{\delta} = \underline{\delta}'(I_k - \underline{q}\,\underline{q}') \underline{\delta} = \underline{\delta}' \underline{\delta}$.) The distribution of log Z under H_0 is easily seen from (3.12) to be the same.

Finally, we show that (3.9) follows from (3.5), (3.11) and (3.12). Two lemmas are required; their proofs are found in the Appendix A.3.

Lemma 1. Let Z be a likelihood ratio and $\{Z^n\}$ a sequence of likelihood ratios. If a sequence of statistics X_n satisfies

$$(X_n, Z^n) \xrightarrow{d} (X, Z)$$

under the null hypothesis then

$$X_n \xrightarrow{d} X$$

under the alternative. Note that the distribution of X under the alternative is not the same as the distribution of X under the null hypothesis.

Lemma 2. Let X^n and Y^n be sequences of random quantities which converge in distribution say

$$X^n \xrightarrow{d} X,\ Y^n \xrightarrow{d} Y.$$

If there is a continuous function H and random quantities ε^n such that

$$Y = H(X),$$

$$Y^n = H(X^n) + \varepsilon^n$$

and $$\varepsilon^n \to 0 \text{ in probability}$$

then

$$(X^n, Y^n) \xrightarrow{d} (X, Y).$$

Lemma 1 reduces the proof of (3.9) to showing joint convergence of $\underline{V}_n$ and Z^n or equivalently of $\underline{V}_n$ and $\log Z^n$; this follows from Lemma 2 and (3.11). Since $\underline{V}_n \to \underline{X}$ where $\underline{X} \overset{d}{=} N_k(\underline{0}, \Lambda)$ under the null and

$\underline{X} \overset{d}{=} N_k(\underline{\delta}, \Lambda)$ under the alternative it follows that the latter is the limiting alternative distribution of $\underline{V}_n$.

Finally we turn to a comparison of the test based on $X^2(n)$ with the test based on Z^n.

The comparison will be done via the limiting distributions of the statistics. Denoting $\underline{\delta}' \underline{\delta}$ by Δ the limiting distributions of $X^2(n)$ are χ^2_{k-1} under H_0 and $\chi'^2_{k-1}(\Delta)$ under H_1, and the limiting distributions of $\log Z^n$ are $N(-\frac{1}{2}\Delta, \Delta)$ under H_0 and $N(\frac{1}{2}\Delta, \Delta)$ under H_1. The asymptotic power of the test based on $\log Z^n$ is given in (1.14); replacing θ_0^2 there by Δ the asymptotic power is

$$1 - \Phi(Z_\alpha - \sqrt{\Delta}) \tag{3.15}$$

where Z_α is the $100(1-\alpha)$ percentile and Φ is the distribution function of the standard normal distribution.

Now for levels $\alpha = .05$ and $\alpha = .01$ we find the values of required for the asymptotic power of the $X^2(n)$ test to be .85, .90 and .95. These values of Δ solve the equations

$$P(\chi'^2_{k-1}(\Delta) > \chi^2_{k-1,\alpha}) = .85, .90, .95.$$

where $\chi^2_{k-1,\alpha}$ is the $100(1-\alpha)$ percentile of the χ^2_{k-1} distribution and they can be read from Table 25 of [2]. From Δ the power in (3.15) is computed and this is then compared to the relevant power for the

chi-square test. In Table 2 below the results are displayed for various values of k-1 the degrees of freedom for the chi-square statistic.

Table 1. Asymptotic power $1 - \Phi(Z_\alpha - \sqrt{\Delta})$ of the test based on Z^n for values of size α, power β and degrees of freedom k-1 of the chi-square test.

	$\alpha = .05$			$\alpha = .01$		
k-1\β	.85	.90	.95	.85	.90	.95
1	.911	.945	.975	.901	.937	.971
2	.952	.972	.989	.945	.968	.987
3	.969	.983	.994	.968	.980	.993
4	.978	.989	.996	.976	.987	.995
5	.984	.992	.997	.982	.991	.997
6	.988	.994	.998	.987	.994	.998
7	.991	.996	.999	.990	.995	.999
8	.993	.997	.999	.992	.996	.999
9	.994	.997	.999	.994	.997	.999
10	.995	.998	1.000	.995	.998	.999
15	.998	.999	1.000	.998	.999	1.000

Note that the power seems to converge to 1 as k gets large; this is also expressed by the fact that the non-centrality parameter Δ must increase with the degrees of freedom in order that the chi-square test have constant power. For a large number of cells k the chi-square statistic $X^2(n)$ has greater difficulty in detecting a particular alternative because it attempts to detect alternatives in many directions. It should be mentioned that under alternatives other than that specified by Z_n (i.e., $P_i^0 + C_i/\sqrt{n}$), Z^n may have smaller asymptotic power than $X^2(n)$. Thus a trade-off exists between increased power from using the likelihood ratio and the risk of using the wrong likelihood ratio.

BIBLIOGRAPHY

1. Billingsley, P. (1968). Convergence of Probability Measures. Wiley, New York.

2. Biometrika Tables for Statisticians, Vol. II. (1972). Eds.: E.S. Pearson and H.O. Hartley. Cambridge University Press.

3. Chernoff, H. (1972). Sequential Analysis and Optimal Design. S.I.A.M., Philadelphia.

4. Chernoff, H. and Petkau, A.J. (1981). "Sequential medical trials involving paired data." Biometrika, 68, 1, 119-132.

5. Cochran, W.G. (1952). "The χ^2 test of goodness of fit." Ann. Math. Stat., 23, 315-345.

6. Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. Chapman and Hall, London.

7. Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

8. Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1953). "Sequential decision problems for processes with continuous time parameter. Testing hypotheses." Ann. Math. Stat., 24, 254-264.

9. Fakeev, A.G. (1970). "Optimal stopping rules for stochastic processes with continuous parameter." Thy. Prob. Appl. Vol. 15, No. 1, 324-331.

10. Freedman, D. (1971). Brownian Motion and Diffusion. Holden-Day, San Francisco.

11. Ghosh, B.K. (1970). Sequential Tests of Statistical Hypotheses. Addison-Wesley, Reading, Mass.

12. Greenwood, P.E. and Shiryayev, A.N. (1985). Contiguity and the Statistical Invariance Principle. Gordon and Breach, New York.

13. Lai, T.L., Siegmund, D. and Robbins, H. (1983). "Sequential design of comparative clinical trials." In Recent Advances in Statistics, Eds.: M.H. Risvi, J.S. Rustagi, D. Siegmund, 51-68. Academic Press, New York.

14. Lamperti, J. (1966). Probability. Benjamin/Cummings, Reading.

15. Lehmann, E.L. (1959). Testing Statistical Hypotheses. Wiley, New York.

16. Lipcer, R. and Shiryayev, A.N. (1980). "A functional central limit theorem for semimartingales." Thy. Prob. Appl. Vol. 25, No. 4, 667-688.

17. Millar, P.W. (1983). The Minimax Principle in Asymptotic Statistical Theory. Unpublished notes.

18. Mitra, S.K. (1958). "On the limiting power function of the frequency chi-square test." Ann. Math. Stat., 29, 1221-1233.

19. Moore, D.S. (1983). "Chi-square tests." Studies in Mathematics, Vol. 19: Studies in Statistics, 66-106, Ed.: R.V. Hogg, Mathematical Association of America.

20. Neveu, J. (1974). Discrete Parameter Martingales. Holden-Day, San Francisco.

21. Roussas, G. (1972). Contiguity of Probability Measures: Some Applications in Statistics. Cambridge University Press.

22. Thompson, M.E. (1971). "Continuous parameter optimal stopping problems." Z. Wahrscheinlichkeitstheorie, 19, 302-318.

23. Wald, A. and Wolfowitz, J. (1948). "Optimum character of the sequential probability ratio test." Ann. Math. Stat., 19, 326-339.

24. Wijsman, R.A. (1963). "Existence, uniqueness and monotonicity of sequential probability ratio tests." Ann. Math. Stat., 34, 1541-1548.

APPENDIX

A.1 Uniform Integrability of a Sequence of Stopping Times

The convergence of Average Sample Numbers ((2.12), (2.13)) which is used in Sections 2.1 and 2.2 requires the uniform integrability of the sequence of stopping times $\{T_n^*\}$ given by

$$T_n^* = \inf\{t: Z^n(t) \geq A \text{ or } Z^n(t) \leq B\}.$$

This we will establish now using the set-up described in the first paragraph of Chapter 2.

The uniform integrability must be established under both sequences of probabilities $\{P_0^n\}$ and $\{P_1^n\}$. In doing this for both sequences at once we will let P^n denote either of the sequences. Let F_n be the distribution function of T_n^* under P^n,

$$F_n(t) = P^n(T_n^* \leq t).$$

Let t be an integer. Then

$$\begin{aligned} 1 - F_n(t) &= P^n(T_n^* > t) \\ &= P^n(B < Z^n(s) < A \text{ for all } s \leq t) \\ &= P^n(\log B < \log Z^n(s) < \log A \text{ for all } s \leq t) \\ &= P^n(\log B < \log Z_k^n < \log A \text{ for all } k \leq tn) \end{aligned}$$

$$\leq P^n(\log B < \log Z_n^n < \log A,\ \log B < \log Z_{2n}^n < \log A, \ldots,$$

$$\log B < \log Z_{tn}^n < \log A)$$

$$\leq P^n(\left|\log Z_n^n\right| < C,\ \left|\log Z_{2n}^n - \log Z_n^n\right| < C, \ldots,$$

$$\left|\log Z_{tn}^n - \log Z_{(t-1)n}^n\right| < C)$$

where $C = \left|\log A\right| + \left|\log B\right|$. Since $\log Z_n^n$, $\log Z_{2n}^n - \log_n^n, \ldots,$ $\log Z_{tn}^n - \log Z_{(t-1)n}^n$ are i.i.d.

$$1 - F_n(t) \leq [P(n)]^t$$

with

$$P(n) = P^n(\left|\log Z_n^n\right| < C).$$

Now since $\log Z_n^n = \log Z^n(1) \overset{d}{\to} \log Z(1)$ we have

$$p(n) \to P(\left|\log Z(1)\right| < C) < 1$$

and thus we can assume without loss of generality that

$$P(n) \leq \gamma < 1 \qquad \text{for every } n$$

Therefore $1 - F_n(t) \leq \gamma^t$ for integers t, so for any t

$$1 - F_n(t) \leq 1 - F_n([t]) \leq \gamma^{[t]} \leq \gamma^{t-1} \tag{A1.1}$$

holds for every n.

Now by integration by parts

$$2\int_0^\infty (1 - F_n(t))\ t dt = (1 - F_n(t))\ t^2/_0^\infty + \int_0^\infty t^2\ dF_n(t)$$

$$= \int_0^\infty t^2\ dF_n(t)$$

using the inequality (A1.1). Therefore

$$E^n(T_n^*)^2 = 2\int_0^\infty (1 - F_n(t))\ t dt$$

$$\leq 2\int_0^\infty t\ \gamma^{t-1}\ dt < \infty.$$

It now follows that $\{T_n^*\}$ is uniformly integrable.

A.2 The Likelihood Ratio of Singular Multivariate Normal Distributions

It is required to find the likelihood ratio of the distributions $N_k(\underline{\delta}, \Lambda)$ and $N_k(\underline{0}, \Lambda)$. Consider the representation ([19])

$$\Lambda^{1/2}\ \underline{Z} + \underline{\delta}$$

for the $N_k(\underline{\delta}, \Lambda)$ distribution, where $\underline{Z}$ is a vector of i.i.d. N(0,1) variables and $\Lambda^{1/2}$ is the square-root of Λ, a symmetric matrix satisfying $\Lambda^{1/2}\ \Lambda^{1/2} = \Lambda$. In our situation Λ is idempotent with rank r so that $\Lambda^{1/2} = \Lambda$; also

$$\Lambda = B\begin{bmatrix} I_r & 0 \\ \hline 0 & 0 \end{bmatrix} B' = B\ D\ B' \qquad \text{(A2.1)}$$

with B an orthogonal matrix, a representation which will be used below. Now if $\underline{\delta}$ is in the range (column space) of Λ the vector $\Lambda\,\underline{Z} + \underline{\delta}$ will remain in the range of Λ and the two distributions $N_k(\underline{\delta}, \Lambda)$ and $N_k(\underline{0}, \Lambda)$ will have the same support. This will make their likelihood ratio meaningful. In Chapter 3 we had $\Lambda = I_k - \underline{q}\,\underline{q}'$ and $\underline{\delta}$ was orthogonal to $\underline{q}$ so that $\Lambda\,\underline{\delta} = \underline{\delta}$ thus ensuring that $\underline{\delta}$ is in the range of Λ.

Now let Q_0, Q_1 be the probability measures on R^k corresponding to the $N_k(\underline{0}, \Lambda)$, $N_k(\underline{\delta}, \Lambda)$ distributions respectively. With D as in (A2.1)

$$\underline{X} \overset{d}{=} N_k\ (\underline{0}, D) \Rightarrow B\underline{X} \overset{d}{=} N_k(\underline{0}, \Lambda) \qquad \text{(A2.2)}$$

Now let P_0, P_1 be the probability measures corresponding to $N_k(\underline{0}, D)$, $N_k(\underline{\mu}, D)$. From (A2.2) and (A2.3), $Q_0 = P_0B^{-1}$ and $Q_1 = P_1B^{-1}$ where the notation means

$$Q_0(A) = P_0(B^{-1}A),\ Q_1(A) = P_1(B^{-1}A)$$

for Borel sets A in R^k.

The likelihood ratio dP_1/dP_0 is simple to find and the following lemma shows how it relates to dQ_1/dQ_0, the desired likelihood ratio.

Lemma. Let P_0, P_1 be probability measures on a measure space (X,F) and $f:(X,F) \to (Y,G)$ be measureable and 1-1 with measurable inverse $f^{-1}:(Y,G) \to (X,F)$. Define Q_0, Q_1 on (Y,G) by

$$Q_0(A) = P_0(f^{-1} A),\ Q_1(A) = P_1(f^{-1} A). \quad (A \in G).$$

Then if $P_1 \ll P_0$ and $Q_1 \ll Q_0$ then

$$\frac{dQ_1}{dQ_0}(y) = \frac{dP_1}{dP_0}(f^{-1}(y)) \quad (y \in Y).$$

Proof: Let $A \in G$.

$$\int_A \frac{dP_1}{dP_0}(f^{-1}(y))\ dQ_0(y) = \int_A \frac{dP_1}{dP_0}(f^{-1}(y))\ dP_0\ f^{-1}(y)$$

$$= \int_{f^{-1}A} \frac{dP_1}{dP_0}(x)\ dP_0(x)$$

(by the change of variables formula with $y = f(x)$.)

$$= P_1(f^{-1}(A)) = Q_1(A).$$

The use of this lemma requires dP_1/dP_0. But P_0 is the distribution of a vector $(X_1,\ldots,X_r, 0,\ldots,0)'$ of r i.i.d. N(0,1) variables and P_1 is the distribution of this vector with the added mean vector $\underline{\mu} = (\mu_1,\ldots,\mu_r, 0,\ldots,0)'$. Therefore for any $\underline{x} = (x_1,\ldots,x_r, 0,\ldots 0)$

$$\frac{dP_1}{dP_0}(\underline{x}) = \frac{\exp\{-\frac{1}{2}\sum_1^r (x_i - \mu_i)^2\}}{\exp\{-\frac{1}{2}\sum_1^r x_i^2\}}$$

$$= \exp\{\sum_1^r \mu_i x_i - \frac{1}{2}\sum_1^r \mu_i^2\} = \exp\{\underline{\mu}' \underline{x} - \frac{1}{2}\underline{\mu}' \underline{\mu}\}$$

$$= \exp\{-\tfrac{1}{2}(\underline{x} - \underline{\mu})' (\underline{x} - \underline{\mu}) + \tfrac{1}{2}\underline{x}' \underline{x}\}.$$

Therefore, by the lemma, since the linear map B is 1-1 from the range of D to the range of Λ, for each y in the range of Λ

$$\frac{dQ_1}{dQ_0}(\underline{y}) = \frac{dP_1}{dP_0}(B^{-1}\underline{y}) = \frac{dP_1}{dP_0}(B'\ \underline{y})$$

$$= \exp\{\underline{\mu}' B' \underline{y} - \tfrac{1}{2}\underline{\mu}' \underline{\mu}\}$$

$$= \exp\{(B'\underline{\delta})' > B'\underline{y} - \tfrac{1}{2}(B'\underline{\delta})'(B'\underline{\delta})\}$$

$$= \exp\{\underline{\delta}'\underline{y} - \tfrac{1}{2}\underline{\delta}'\underline{\delta}\}.$$

and this was the formula used to obtain (3.12).

Note: A further use of this calculation is made for the application of the SPRT to the problem of testing the mean vector of a multivariate normal distribution. Note that only alternatives which specify a mean vector in the range of the covariance matrix can be tested.

A.3 Two Lemmas on Weak Convergence

In this section proofs of Lemma 1 and Lemma 2 are provided. Lemma 1 is first restated more precisely.

Lemma 1. Let P_0 and P_1 be probability measures with $P_1 \ll P_0$ and $Z = dP_1/dP_0$ be their likelihood ratio. For each n=1,2,... let P_0^n and P_1^n be probability measures with $P_1^n \ll P_0^n$ and $Z^n = dP_1^n/dP_0^n$. If there are random elements X, X^n such that

$$(X^n, Z^n) \xrightarrow{d} (X, Z) \text{ under } P_0^n, P_0$$

then

$$X^n \xrightarrow{d} X \text{ under } P_1^n, P_1.$$

Proof: If f is bounded and continuous on the space where X^n and X lie

$$\begin{aligned} \int f(X^n)\, dP_1^n &= \int f(X^n)\, Z^n\, dP_0^n \\ &= \int h(X^n, Z^n)\, dP_0^n \\ &\to \int h(X, Z)\, dP_0 \end{aligned}$$

(since $h(x, z) = f(\underline{X})\, z$ is continuous on the product space.)

$$\begin{aligned} &= \int f(X)\, Z\, dP_0 \\ &= \int f(X)\, dP_1. \end{aligned}$$

Lemma 2. Let P_0 and $P_0^n (n=1,2,\ldots)$ be probability measures and let X, Y, X^n and Y^n be random elements such that

$$X^n \xrightarrow{d} X \qquad \text{under } P_0^n,\ P_0$$

and

$$Y^n \xrightarrow{d} Y \qquad \text{under } P_0^n,\ P_0.$$

If there is a continuous function H and random elements ε^n such that

$$Y = H(X)$$

$$Y^n = H(X^n) + \varepsilon^n$$

$$\varepsilon^n \to 0 \text{ in probability under } P_0^n,$$

then

$$(X^n, Y^n) \xrightarrow{d} (X, Y) \qquad \text{under } P_0^n,\ P_0.$$

Proof: Since $\varepsilon^n \to 0$ in probability it suffices to prove that

$$(X^n, H(X^n)) \xrightarrow{d} (X, H(X)).$$

(see [1]). For this let f be continuous and bounded on the product space where (X,Y) lives. Then

$$\begin{aligned} \int f(X^n, H(X^n))\, dP_0^n &= \int g(X^n)\, dP_0^n \\ &\to \int g(X)\, dP_0 \\ &= \int f(X, H(X))\, dP_0 \end{aligned}$$

Since $g(x) = f(x, H(x))$ is continuous.