University of Alberta

Sequential Testing of Multiple Hypotheses

by

© Jesse Coull

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of

Master of Science
in
Statistics

Department of Mathematical Sciences
Edmonton, Alberta
Spring 1997

University of Alberta

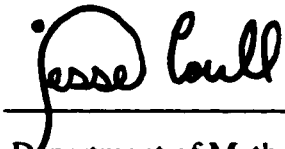
Library Release Form

Name of Author: Jesse Coull

Title of Thesis: Sequential Testing of Multiple Hypotheses

Degree: Master of Science

Year this Degree Granted: 1997



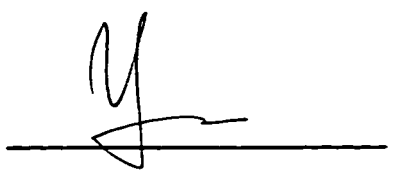
Department of Mathematical Sciences
University of Alberta
Edmonton, Alberta, Canada
T6G 2G1

Date: Jan. 23, 1997

University of Alberta

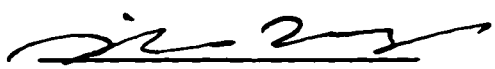
Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Sequential Testing of Multiple Hypotheses** submitted by Jesse Coull in partial fulfillment of the requirements for the degree of Master of Science in Statistics.



Dr. Y. Wu

Dr. R. Karunamuni



Dr. M. Meng

Date: January 9, 1997

Abstract

An overview of hypothesis testing is given with emphasis on the identification of optimal decision making procedures. These results are then extended to the multiple hypothesis testing problem. Next, Wald's Sequential Probability Ratio Test is presented along with discussion about the development of optimal sequential procedures for the multiple hypothesis case. Finally, numerical simulation is used to evaluate the optimality of one such procedure.

Sequential Testing of Multiple Hypotheses

List of Tables

List of Figures

1. Introduction

Within many scientific fields of study, hypothesis testing has long been one of the cornerstones of statistical inference. Scientists use such tests to indicate whether or not research data provides evidence in favour of, or in opposition to, proposed theories. Hence, it is of extreme importance that the statistical methodology that these tests employ be both accurate and efficient. One neither wants the conclusions reached to be inaccurate, nor does one want to needlessly waste time, energy, and money collecting unnecessary data. With these desires in mind, statisticians have strived to develop optimal tests; ones that make maximal use of the data that they are provided.

Originally, and in its simplest form, statistical hypothesis testing provides a means to accurately classify, after observing the characteristics of a sample, the characteristics of the population from which this sample was drawn. Two classical examples are the inspection of the quality of products in a batch, and the comparison with a new procedure or treatment with the standard one. In the first example, a number of products are inspected and their quality characteristics are measured in order to decide if we should accept or reject the batch from whence they came. In the second example, some items or subjects are treated with the new drug or procedure and others with the existing one, with the differences being measured to provide statistical evidence indicating whether the new procedure is effective or not.

In either of these situations, there are two major decisions which must be made after the hypothesis to be tested has been defined. First, one must decide on a sampling plan. It must be decided both how many items are to sampled, and how such sampling is going to take place. For example, instead of the usual practice of taking all samples before analysis of the data, they could be taken sequentially, one by one, with the decision to continue sampling contingent upon the observed items at hand. Second, it must be decided how these sampled items are to be used. The adopted testing procedure should

indicate when the proposed null hypothesis is to be accepted as possibly true, and when it is to be rejected as most likely false.

After these two decisions have been made, and the hypothesis test carried out, the obvious question is, how good were these decisions? In addition, if the choices made were not optimal, what could be done to improve upon them? It is for the answers to these questions that statisticians have searched and, in some cases, found.

The purpose of this thesis is to present a multiple hypothesis testing procedure, after reviewing the statistical inference materials that led to its development. The thesis will first outline the manner in which questions of optimality have been approached, and provide the answers to these questions when they are available. Then, after introducing some basic concepts, most importantly the fundamental Neyman-Pearson Lemma and Wald's Likelihood Identity, their application within the Monotone Likelihood Ratio and exponential families will be focussed upon. What follows will be the extension of some these results to the multiple hypothesis situation. In the next section, some of the difficulties which are commonly encounter when trying to employ the theoretical results are outlined, along with some of the simple solutions which have been proposed to deal with them. In the final section, we discuss the application of sequential procedures with initial focus falling upon Wald's likelihood ratio test. Finally, the generalization of the optimal Sequential Probability Ratio Test (SPRT) within the framework of multiple hypothesis tests is explored. In particular, a simple sequential multiple hypothesis testing procedure is discussed and its properties investigated.

2. The Core of Statistical Inference

The material presented in this section is standard and can be found in Lehmann (1986) or Berger (1980). However, the proofs and presentations are different and, in many cases, much simpler.

2.1 Initial Framework

We begin by restricting ourselves to those situations where there is only one choice to be made - either we accept or reject the proposed hypothesis based upon the data provided to us. In other words, we consider only one alternative hypothesis. Additionally, for the moment we restrict ourselves to procedures in which the sample size is fixed beforehand; essentially, the data from which the decision is made has already been collected and the option to go out and collect more is not a viable one. For this type of situation, much is known about the answers to the basic question of "what is an optimal procedure?" It is these situations which are explored first.

We introduce the following standard notations which will be used throughout the thesis. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a sample of size n from the population with probability distribution family $\{P_\theta(\cdot), \theta \in \Omega\}$ where Ω is the parameter space. It will be assumed, for technical reasons, that all the probability distributions under consideration are mutually absolutely continuous with respect to one another in the standard probability sense. With this notation, the testing problem can be formalized as one of -

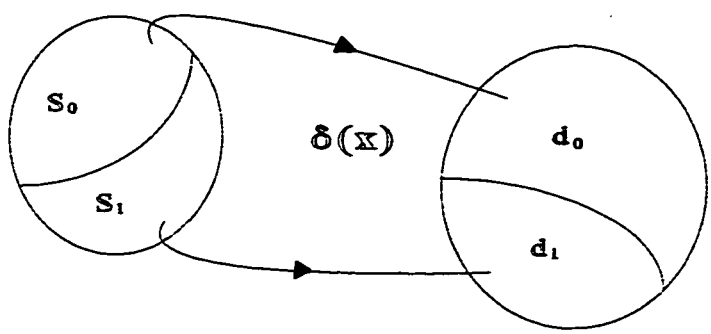
$$H_0 : \theta \in \Omega_0 \text{ vs. } H_1 : \theta \in \Omega_1 \quad \text{where } \Omega_0 \cup \Omega_1 = \Omega \text{ and } \Omega_0 \cap \Omega_1 = \varnothing \ .$$

The task is to select a "decision rule" indicating whether H_0 or H_1 is true. A decision rule, δ, is a function of $\mathbf{x} \in S$, the sample space, such that $\delta(\mathbf{x}) = d_0$ if and

only if H_0 is to be accepted, and $\delta(\mathbf{x}) = d_1$ if and only if H_1 is to be accepted.

One particularly naive decision rule which immediately comes to mind with this setup arises if one separates S as $S = S_0 \cup S_1$ with $S_0 \cap S_1 = \varnothing$ and lets $\delta(\mathbf{x}) = d_0$ (d_1) if, and only if, $\mathbf{x} \in S_0$ (S_1). In such cases, there is really only one optimality decision to be made - "What is the best possible choice of S_0? "

Figure 1.1 The simplest decision function



Before answering this question, and others, we first need some notion of what the word "best" implies. By evaluating the probability of type I (the probability of rejecting the null hypothesis, making decision d_1, when it is indeed true, $\theta \in \Omega_0$) and type II (probability of accepting the null hypothesis, d_0, when it is not true, $\theta \in \Omega_1$) errors of the chosen testing procedure, two measures of the quality of these choices are available; both of which we would like to try to minimize. Alternatively, one can also think of the minimization of these error probabilities as an effort to maximize the power of the test. That is, we wish to maximize the probability that our test can successfully declare the true state of nature. If the cost associated with making the two types of errors are fixed at c_0 and c_1

respectively, then one can further express these ideas with the following loss function of our simplistic test δ -

$$L[\theta,\delta(x)] = c_0 \cdot I_{[\theta \in \Omega_0, \delta(x)=d_1]} + c_1 \cdot I_{[\theta \in \Omega_1, \delta(x)=d_0]} \quad .$$

There is a loss of c_0 if the decision of d_1 is incorrect, of c_1 if the decision of d_0 is incorrect, and there is no loss if the choice of either d_0 or d_1 is correct. When the true population parameter is θ, we can also define the risk function to be the expected value of the loss function -

$$R(\theta,\delta) = E_\theta[L[\theta,\delta(x)]] \quad .$$

Then, one reasonable measure of "best" would be that choice of S_0, and thus of δ in our current simplistic setup, for which this risk function is minimized. Unfortunately, as indicated by its notation, this risk function is dependent on the true parameter, θ. For different values of this parameter, it is expected that there will be a different decision rule which minimizes the risk. Thus, it may be impossible to identify a globally optimal decision rule which is "best" for all possible values of the true parameter which is, after all, unknown.

2.2 Optimal Test Characteristics

To help restrict our search for the optimal rule, we consider two criteria that seem natural. First, for any rule under serious consideration, there should not exist another decision rule which is superior for all possible values of the true parameter. Such a rule is referred to as being admissible (Wald, 1939) and mathematically, we have -

$$\delta \text{ st. } \nexists\, \delta' \text{ with } R(\theta,\delta') \le R(\theta,\delta) \;\; \forall\, \theta \in \Omega \text{ with inequality for some } \theta\text{'s.}$$

Second, if $E_\theta[L(\theta',\delta)]$ is considered a function of θ' for θ given, then this function achieves its minimum for $\theta'=\theta$. In other words, if the value of the true parameter is θ (arbitrary), we wish the expected losses of our test to be at a

minimum at that value. Rules satisfying this criteria are referred to as risk unbiased and mathematically, we have -

$E_\theta[L(\theta',\delta)] \geq E_\theta[L(\theta,\delta)]$ all $\theta' \neq \theta$, any θ.

Proposition 2.2a:

δ is unbiased if, and only if,

$$P_\theta[\delta(X)=d_1] \leq \frac{c_1}{c_0+c_1}, \qquad \text{for all } \theta \in \Omega_0, \text{ and}$$

$$P_\theta[\delta(X)=d_1] \geq \frac{c_1}{c_0+c_1}, \qquad \text{for all } \theta \in \Omega_1 .$$

Proof of 2.2a:

$\Leftarrow$: ①
$$E_\theta[L(\theta,\delta)] = c_0 \cdot P_\theta[\delta(X)=d_1] \cdot I_{[\theta\in\Omega_0]} + c_1 \cdot P_\theta[\delta(X)=d_0] \cdot I_{[\theta\in\Omega_1]}$$

②
$$E_\theta[L(\theta',\delta)] = c_0 \cdot P_\theta[\delta(X)=d_1] \cdot I_{[\theta'\in\Omega_0]} + c_1 \cdot P_\theta[\delta(X)=d_0] \cdot I_{[\theta'\in\Omega_1]}$$

$\therefore$ ② - ① $\geq$

$$c_0 \cdot \frac{c_1}{c_0+c_1} - c_1 \cdot \frac{c_0}{c_0+c_1} + c_1 - (c_0+c_1) \cdot P_\theta[\delta(X)=d_1] \cdot I_{[\theta\in\Omega_0,\theta'\in\Omega_1]}$$

and, when $\theta \in \Omega_0$, $P[\delta(X)=d_1] \leq c_1 / (c_0 + c_1)$ so the third term is $\leq c_1$. Therefore the whole equation must be non-negative and our test unbiased as was set out to show.

$\Rightarrow$: Similarly, if δ is unbiased then, by ② - ①, it is true that

$$-c_0 \cdot p + c_1 \cdot (1-p) \geq 0 \text{ when } \theta \in \Omega_0$$

and (where $p = P_\theta[\delta(X)=d_1]$)

$$c_0 \cdot p + c_1 \cdot (1-p) \geq 0 \text{ when } \theta \in \Omega_1$$

which together give the required result.■

Proposition 2.2a indicates that for a decision to be unbiased, the Type I error must be consistently less than a constant $\alpha = c_1 / [c_0 + c_1])$, and that 1 - (Type II error) must always be larger than this same α.

One problem with the approach adopted is that the cost structure may be very hard to estimate for many testing problems. Thus, the previously determined cost structure of an unbiased test is often useless. Instead, we refer to a test as being of significance level α when -

$$sup_{\theta \in \Omega}\{P_\theta[\delta(X)=1]\} \leq \alpha$$

and we replace the idea of type II error by that of a power function -

$\beta(\theta) = P_\theta[\delta(X)=1]$.

Then an unbiased test is one for which the power is always larger than the significance level - $\beta(\theta) \geq \alpha$.

This definition of unbiasedness is much simpler and easier to deal with.

So far, the decision rules under consideration have been non random; the decision made was entirely dependent upon the value of $\mathbf{x}$. However, in a more general setup, we allow our decision rule to be random. A random decision rule or test, ϕ, is a function of the observations such that $\phi(\mathbf{x})$ gives the probability that we should reject H_0, and $1-\phi(\mathbf{x})$ gives the probability that we should accept H_0. Thus, the experimenter performs an additional independent Bernoulli trial with probability of success $\phi(x)$ and either rejects (makes decision d_1), or accepts (makes decision d_0), H_0 based upon the success, or failure of this trial.

Now, the space of all possible randomized tests will be denoted D; obviously the set of non-random tests (where $\phi(\mathbf{x}) \in \{0,1\}$) is a subset of D. By our earlier definition of the power function, we have than $\beta(\theta) = E_\theta[\phi(\mathbf{x})]$. Often, rather than just minimizing the risk function introduced earlier, it is more common to minimize type II error, while restraining the type I error below an acceptable level. Mathematically, we would like an α-level test, $\phi(\mathbf{x})$, so that for any other α-level test, $\phi'(\mathbf{x}) \in D$, -

$$E_\theta[\phi(\mathbf{x})] \geq E_\theta[\phi'(\mathbf{x})] \text{ for all } \theta \in \Omega_1.$$

In other words, when the parameter does not satisfy the null hypothesis, we would like our test to have the highest probability of rejecting that hypothesis. Such a test is referred to as being Uniformly Most Powerful (UMP) in D.

Proposition 2.2b:

If $\phi^*(\mathbf{x})$ is UMP in D, then it must be unbiased.

Proof 2.2b:

The test $\phi(\mathbf{x}) = \alpha$ is in D, has $\beta(\theta) = \alpha$, and is obviously unbiased. If $\phi^*(\mathbf{x})$ is UMP, then $\beta^*(\theta) \geq \beta(\theta) = \alpha$ for all $\theta \in \Omega_1$, and thus ϕ^* is unbiased as claimed.■

As simple as it is, proposition 2.2b is none the less very important. It says that if a UMP test exists in D, it is necessarily unbiased. Thus, when looking for a UMP test, we need only consider those which are unbiased; for this reason, we will denote the set of all unbiased tests as $D_U \subset D$. Unfortunately, finding a Uniformly Most Powerful Unbiased (UMPU) test does not guarantee that the test is also UMP in D; there may not exist a UMP test. However, in many situations, it is possible to identify a UMPU test in D which, in many respects, is the best possible test available.

2.3 Some Optimal Tests

Consider the simple test case where $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$. Before finding a UMP test of this hypothesis, we first adopt the notation of denoting the density function of P_θ as $f_\theta(\cdot)$, and the likelihood function for θ as -

$$l(\theta) = f_\theta(\mathbf{x}) = f_\theta(x_1,...,x_n) \ .$$

As well, we use Fisher's definition (Fisher, 1925) and say that $T(\mathbf{x})$ is a sufficient statistic for θ if -

$$l(\theta) = g_\theta(T(\mathbf{x}))\cdot h(\mathbf{x})$$

for some functions $g_\theta(\cdot)$ and $h(\cdot)$. Finally, we will need the simple, yet powerful, fact that -

$$E_{\theta_1}[g(X)] = E_{\theta_0}\left[g(X)\cdot\frac{f_{\theta_1}}{f_{\theta_0}}(X)\right]$$

This is the Likelihood Ratio Identity (LRI).

Proposition 2.3a:

The UMP, α-level test of $H_0: \theta = \theta_0$ vs. $H_1: \theta = \theta_1$ is given by

$$\phi^*(x) = I\left[\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c\right]$$

where c is chosen such that $E_0[\phi^*(X)]=\alpha$.

Proof 2.3a:

Consider any other α-level test $\phi(\mathbf{x})$ (i.e., $E_0[\phi(X)] \le \alpha$). Then, by the LRI, we have that -

$$E_1[\phi^*(X)-\phi(X)] = E_0\left[(\phi^*(X)-\phi(X))\cdot\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}\right]$$

$$\geq E_0\left[(\phi^*(X)-\phi(X))\cdot\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}-c\right)\right] \geq 0$$

and thus our test, ϕ^*, is more powerful.■

In this case, the decision function is very simple, in fact completely non-random. In addition, the proof of its optimality can also be easily extended to the problem of a slightly more complex null hypothesis.

Proposition 2.3b:

The UMP, α-level test of H_0: $\theta \in \{\theta_1,...,\theta_K\}$ vs. H_1: $\theta = \theta^*$ is given by -

$$\phi^*(x) = I\left[\frac{f_{\theta^*}(x)}{\sum_{i=1}^{K} k_i \cdot f_{\theta_i}(x)} > c\right]$$

where c and $k_i \geq 0$ are chosen so that $\sum_i k_i = 1$ and $E_i[\phi^*(X)] = \alpha$.

Proof 2.3b:

Similarly to the proof for 2.3a, we use the LRI to see that

$$E_{\theta^*}[\phi^*(X) - \phi(X)] = E_{\theta_i}\left[(\phi^*(X) - \phi(X))\cdot\frac{f_{\theta^*}(X)}{f_{\theta_i}(X)}\right]$$

$$\geq E_{\theta_i}\left[(\phi^*(X) - \phi(X))\cdot\left(\frac{f_{\theta^*}(X)}{f_{\theta_i}(X)} - c\cdot k_i\right)\right]$$

the last inequality being true since the final term being subtracted off is a known positive quantity (c and k_i both positive, and again $E_i[\phi^*(\mathbf{X})] = \alpha \geq E_i[\phi(\mathbf{X})]$). Additionally, notice that if -

$$\frac{f_{\theta^*}(x)}{\sum_{i=1}^{K} k_i \cdot f_{\theta_i}(x)} > c \qquad \text{then} \qquad \frac{f_{\theta^*}(x)}{k_i \cdot f_{\theta_i}(x)} > c$$

as the inequality on the right can only have a smaller denominator. This is all that is needed to proceed as in proposition 2.3a.■

Thus, there is a best test when Ω_0 contains any finite number of points. Naturally, we would like a best test for when Ω_1 is similarly complex and, even more generally, when either Ω_0 or Ω_1 contains any number of points, perhaps even infinite.

First, we consider the case where H_0: $\{\theta \leq \theta_0\}$ and H_1: $\{\theta > \theta_0\}$.

Definition 2.3a:

The family of densities $\{f_\theta(\mathbf{x}), \theta \in \Omega\}$ has the monotone likelihood ratio (MLR) property if, for any $\theta_1 > \theta_2$,

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = g_{\theta_1,\theta_2}[T(x)]$$

is a monotone increasing function of $t=T(x)$, a one dimensional statistic. Obviously, for MLR families, $T(x)$ is a sufficient statistic for θ; this can be seen by fixing θ_1.

Proposition 2.3c:

For an MLR family (Karlin and Rubin, 1956), a UMP α-level test of the hypothesis -

H_0: $\{\theta \leq \theta_0\}$ vs. H_1: $\{\theta > \theta_0\}$

is given by $\phi^*(x) = I_{[T(x)>c]}$, where c is chosen so that $E_0[\phi^*(\mathbf{X})] = \alpha$.

Proof 2.3c:

Consider first the simple hypothesis test of H_0: $\{\theta=\theta_0\}$ vs. H_1: $\{\theta=\theta_1\}$ for some $\theta_1>\theta_0$. We already know that the UMP α-level test for this problem is given by ϕ^*. Now, as the choice of c is independent of θ_1, this test must also be UMP for H_0: $\{\theta=\theta_0\}$ vs. H_1: $\{\theta>\theta_0\}$. Additionally, since any UMP test is unbiased, we must also have that $\beta(\theta_1) > \beta(\theta_0)$ for any $\theta_1>\theta_0$. Thus, $\beta(\theta)$ is monotonically increasing in θ and reaches its maximum under the null hypothesis of our original problem at θ_0. Therefore, ϕ^* is indeed UMP for H_0: $\{\theta\leq\theta_0\}$ vs. H_1: $\{\theta>\theta_1\}$ as was set out to show.■

This result is of tremendous importance as there exists a fairly wide class of densities possessing the MLR property. In fact, the entire class of densities referred to as the exponential family all possess the MLR property. We now discuss some characteristics of this important family.

Definition 2.3b:

$\{f_\theta(\mathbf{x}), \theta\in\Omega\}$ is a one parameter exponential family member if -

$$f_\theta(\mathbf{x}) = \exp\{\theta\cdot T(\mathbf{x}) - c(\theta)\} \cdot f_0(\mathbf{x})$$

and $\Omega = \{\theta \mid c(\theta) < \infty\}$ contains at least one point other than 0.

Here, $f_0(\mathbf{x})$ is commonly referred to as the generating density.

Proposition 2.3d:

$\{f_\theta(\mathbf{x})\}$ is MLR with $T(\mathbf{x})$ as a sufficient statistic.

Proof 2.3d:

This follows easily from the definitions of the MLR property and of the exponential family parameterization introduced above.■

Therefore, we have the hypothesis testing results for MLR possessing densities shown earlier. There is one additional important result.

Proposition 2.3e:

For the hypothesis testing problem H_0: $\{\theta \le \theta_1 \text{ or } \theta \ge \theta_2\}$ vs. H_1: $\{\theta_1 < \theta < \theta_2\}$, an α-level UMP test is given by $\phi^*(x) = I_{[b<T(x)<d]}$, where b and d are such that $E_1[\phi^*(\mathbf{X})] = E_2[\phi^*(\mathbf{X})] = \alpha$.

Proof 2.3e:

First, consider the testing problem of H_0: $\theta \in \{\theta_1, \theta_2\}$ vs. H_1: $\theta = \theta'$ for $\theta_1 < \theta' < \theta_2$. By proposition 2.3b, we know that the UMP test is as follows -

$$\phi^*(x) = I_{\left[\frac{f_{\theta'}(x)}{p f_{\theta_1}(x) + (1-p) f_{\theta_2}(x)} > c\right]}$$

for appropriately chosen values of p and c dependent upon θ_1, θ_2, and θ'.

Now, for the exponential family, the left hand side of the bracketed quantity greatly simplifies to the inverse of -

$$p \cdot \exp\{-(\theta' - \theta_1) \cdot T(x)\} + (1-p) \cdot \exp\{(\theta_2 - \theta') \cdot T(x)\}$$

Since this is a strictly convex function of T(x), there exist b and d so that -

$$\phi^*(\mathbf{x}) = I_{[b<T(x)<d]} \qquad \text{and} \qquad E_1[\phi^*(\mathbf{X})] = E_2[\phi^*(\mathbf{X})] = \alpha .$$

Now, the values of b and d are independent of θ' so our test is actually UMP for the testing problem of H_0: $\theta \in \{\theta_1, \theta_2\}$ vs. H_1: $\theta' \in (\theta_1, \theta_2)$.

Consider now the testing problem of H_0: $\theta \in \{\theta_1, \theta_2\}$ vs. H_1: $\theta = \theta'$, but this time for $\theta' < \theta_1$ or $\theta' > \theta_2$. It is once again easy to show that ϕ^* is the optimal test minimizing $\beta(\theta)$ under H_0. Therefore, $\beta(\theta) \le \alpha$ and our testing function is UMP for the more general case of our H_0: $\theta \notin (\theta_1, \theta_2)$. Combining these two results proves the proposition.■

2.4 Unbiasedness and Optimality in Multiple Hypothesis Testing

As was pointed out before, the main purpose of hypothesis testing is often one of classification; one wishes to distinguish between two competing hypotheses. More generally, this idea can be extended to a greater number of possible classifications and the multiple hypothesis testing problem naturally arises. How are our ideas of a "good" test carried over in this situation?

Suppose that $H_i : \theta = \theta_i$ for i = 1,...,k are the competing hypotheses and that -

$\varphi(x) = (\varphi_1(x),...,\varphi_k(x))$ is a test with $\sum_{i=1}^{k} \varphi_i(x) = 1$

where $\varphi_i(x) = P[H_i \text{ is the decision}]$. Then, denoting by c_{ij} the cost of accepting H_j when H_i is true, we have the following generalization for the unbiasedness of $\varphi(x)$.

Definition 2.4a:

$\varphi(x)$ is, as proposed by Lehmann, L-unbiased if -

$$\sum_{j \neq k} c_{kj} \cdot P[H_j|H_i] \geq \sum_{j \neq i} c_{ij} \cdot P[H_j|H_i] \qquad \text{for all k and i.}$$

Proposition 2.4a:

When $c_{ij}=1$, L-unbiasedness implies that $P[H_i|H_i] \geq \max_{\{k \neq i\}} P[H_k|H_i]$. This seems the natural extension of our original definition of unbiasedness.

Proof 2.4a:

$\sum_{j \neq k} P[H_j|H_i] \geq \sum_{j \neq i} P[H_j|H_i] \rightarrow 1 - P[H_k|H_i] \geq 1 - P[H_i|H_i]$

$\Rightarrow$ $P[H_i|H_i] \geq P[H_k|H_i]$, which is true for every $k \neq i$, and thus

$\Rightarrow$ $P[H_i|H_i] \geq \max_{k \neq i} P[H_k|H_i]$ as required.■

How about optimality? As in the simple testing situation, we would naturally like a test which minimizes risk. Letting $r_\varphi(j)$ denote the loss when H_j is true, then -

$$r_\varphi(j) = \sum_{k \neq j} c_{jk} \cdot P[H_k|H_j]$$

and it seems natural that φ^* should be considered optimal if, for any other test φ -

$$\sum_{j=1}^{k} r_{\varphi^*}(j) \leq \sum_{j=1}^{k} r_\varphi(j)$$

Proposition 2.4b:

Considering again the most simple case (when k=2 and c_{ij}=1) -

$$\varphi_2^*(x) = I_{[f_{\theta_2}(x) > f_{\theta_1}(x)]}$$

is, in this new sense, optimal.

Proof 2.4b:

Consider any other test φ. We have that -

$$r_{\varphi^*} - r_\varphi = P_{\varphi^*}[H_2|H_1] - P_\varphi[H_2|H_1] + P_{\varphi^*}[H_1|H_2] - P_\varphi[H_1|H_2]$$

$$= E_{\theta_1}[\varphi_2^*(x) - \varphi_2(x)] + E_{\theta_2}[(1 - \varphi_2^*(x)) - (1 - \varphi_2(x))]$$

$$= E_{\theta_1}\left[(\varphi_2^*(x) - \varphi_2(x)) \cdot \left(1 - \frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}\right)\right] \leq 0$$

and thus, φ^* has smaller risk.■

Unfortunately, these multiple hypothesis extensions of unbiasedness and optimality are not without problems.

Example 2.4a:

Taking -

$$f_1(x) = 0.3 \cdot I_{[0,1)}(x) + 0.7 \cdot I_{[1,2]}(x) \qquad \text{and}$$

$$f_2(x) = 0.2 \cdot I_{[0,1)}(x) + 0.8 \cdot I_{[1,2]}(x)$$

the optimal test is given by -

$$\varphi_1^*(x) = I_{[0,1]}(x) \qquad \text{and}$$

$$\varphi_2^*(x) = I_{[1,2]}(x)$$

But then -

$$P[H_1|H_1] = 0.3 < 0.7 = P[H_2|H_1]$$

and φ^* is not, even in this simple case, L-unbiased.

Fortunately, Van der Waerden was able to develop a more useful definition of unbiasedness for the multiple hypothesis testing problem.

Definition 2.4b:

$\varphi(\mathbf{x})$ is, in the sense of Van der Waerden, W-unbiased if -

$$P[\overline{H_j}|H_j] \leq P[\overline{H_j}|H_i]$$

for all $i \neq j$, where -

$$\overline{H_j} = \bigcup_{i \neq j} H_i$$

Proposition 2.4c:

The criteria for W-unbiasedness is equivalent to -

$$\max_{i \neq j} P[H_j|H_i] \leq P[H_j|H_j] \quad \text{for } j=1,\ldots,k.$$

Proof 2.4c:

Suppose, for some i and j, that this is not true. Then -

$P[H_j|H_i] > P[H_j|H_j]$, so -

$$P[\overline{H_j}|H_j] > P[\overline{H_j}|H_i]$$

which is a contradiction. A similar argument proves the reverse.■

With this definition, it is easily shown that our optimal multiple hypothesis test introduced above, φ^*, must be W-unbiased when $c_{ij} = 1$.

Proposition 2.4d:

The optimal, W-unbiased test when $c_{ij} = 1$ will have the following form -

$$\varphi_j^*(x) = I_{[f_{\theta_j}(x) \geq \max_{i \neq j} f_{\theta_i}(x)]}$$

Proof 2.4d:

Consider any other test φ -

$$r_{\varphi^*} - r_{\varphi} = P_{\varphi^*}[\overline{H_j}|H_j] - P_{\varphi}[\overline{H_j}|H_j] + P_{\varphi^*}[H_j|\overline{H_j}] - P_{\varphi}[H_j|\overline{H_j}]$$

$$= E_{\theta_j}[(1 - \varphi_j^*(x)) - (1 - \varphi_j(x))] + \sum_{i \neq j} E_{\theta_i}[\varphi_j^*(x) - \varphi_j(x)]$$

$$= E_{\theta_j}\left[(\varphi_j^*(x) - \varphi_j(x)) \cdot \left(\frac{\sum_{i \neq j} f_{\theta_i}(x)}{f_{\theta_j}(x)} - 1 \right)\right] < 0$$

since $\varphi^*_j(\mathbf{x}) \leq \varphi_j(\mathbf{x})$ implies that -

$$f_{\theta_j}(x) < max_{i \neq j} f_{\theta_i}(x) \quad \Rightarrow \quad \frac{\sum f_{\theta_i}(x)}{f_{\theta_j}(x)} - 1 > \frac{f_{\theta_j}(x)}{f_{\theta_j}(x)} - 1 = 0$$

and $\varphi^*_j(\mathbf{x}) > \varphi_j(\mathbf{x})$ implies that -

$$f_{\theta_j}(x) \geq max_{i \neq j} f_{\theta_i}(x) \quad \Rightarrow \quad \frac{\sum f_{\theta_i}(x)}{f_{\theta_j}(x)} - 1 \leq \frac{(k-1) \cdot f_{\theta_j}(x)}{f_{\theta_j}(x)} - 1 = k-2 \leq 0$$

for $k \geq 2$. Therefore, our test has the least risk.■

3. The Many Difficulties

Example 2.4a hints at the difficulties of the multiple hypothesis testing situation. However, there are also numerous difficulties in the single alternative situation as well. Lehmann (1955) and Wald (1943) both provide good overviews of the material in this section.

3.1 The First Roadblock

Recall that in section 2.3 an optimal test of H_0: $\{\theta \le \theta_1 \text{ or } \theta \ge \theta_2\}$ vs. H_1: $\{\theta_1 < \theta < \theta_2\}$ within the exponential family framework was identified. Unfortunately, there are problems when the hypothesis test is reversed, namely for -

H_0: $\theta_1 < \theta < \theta_2$ vs. H_1: $\theta < \theta_1$ or $\theta > \theta_2$,

or the limiting case of -

H_0: $\theta = \theta_0$ vs. H_1: $\theta \neq \theta_0$.

Proposition 3.1a:

For these above two sided testing problems in the exponential family, no UMP test exists.

Proof 3.1a:

We consider the limiting case only. Suppose that ϕ^* is a UMP test of size α. This test is necessarily unbiased by proposition 2.2b. Thus, by differentiating, we know that -

$$E_{\theta_0}[\phi^*(X)] = \alpha \qquad \text{and} \qquad \frac{\partial}{\partial\theta} E_{\theta_0}[\phi^*(X)] = 0$$

On the other hand, we consider the UMP test for the following hypothesis testing problem - H_0: $\theta \le \theta_0$ vs. H_1: $\theta > \theta_0$. Since the exponential family

possess MLR properties, we know that $E_\theta[\phi(X)]$ is strictly increasing and thus that -

$$\frac{\partial}{\partial\theta}E_\theta[\phi(X)] > 0$$

However, since ϕ^* is UMP, we know that -

$$E_\theta[\phi^*(X)] \geq E_\theta[\phi(X)]$$

for all $\theta \geq \theta_0$. So -

$$\frac{\partial}{\partial\theta}E_\theta[\phi^*(X)] \geq \frac{\partial}{\partial\theta}E_\theta[\phi(X)]$$

for some $\theta \geq \theta_0$. This is a contradiction, and our proof is complete.■

In this case, no best test even exists. How do we proceed? Although it seems a little like cheating, we just redefine our notion of best. As one might have guessed from our earlier discussion about unbiasedness, we will attempt to find a UMP test not over D, but instead over D_U - the set of all unbiased tests.

Proposition 3.1b:

$\phi^*(x) = I_{[T(x)<b \text{ or } >d]}$ where b and d are again chosen such that -

$E_1[\phi^*(X)] = E_2[\phi^*(X)] = \alpha$ is UMP in D_U, or UMPU.

Proof 3.1b:

Consider $D^\alpha = \{\phi \in D \mid E_1[\phi(X)] = E_2[\phi(X)] = \alpha\}$. Since $\beta(\theta)$, the power function of ϕ, is continuous, $D_U \subset D^\alpha \subset D$, it would be more than sufficient to show that $\phi^*(x)$ is UMP in D^α. However, the proof of this is identical to the proof of proposition 2.3e with Ω_0 replaced with Ω_1.■

It should be noted that when selecting b and d in the limiting case of -

$H_0: \theta=\theta_0$ vs. $H_1: \theta\neq\theta_0$

we need to ensure that $E_0[\phi^*(\mathbf{X})]=\alpha$ and $E_0[T(\mathbf{X})\cdot\phi^*(\mathbf{X})] = n\cdot E_0[T(\mathbf{X})]$ (for unbiasedness).

Example 3.1a:

Consider $X_1,\ldots,X_n$ which are iid $N(0,\sigma^2)$. For the testing problem of $H_0: \sigma=\sigma_0$ vs. $H_1: \sigma\neq\sigma_0$, proposition 3.1b indicates that the following test is UMPU -

$$\phi^*(x) = I_{[T(x)<b^* \text{ or } >d^*]} \Leftrightarrow I\left[\frac{T(x)}{\sigma_0^2}<b \text{ or } >d\right]$$

Here, $T(x)=\Sigma_i x_i^2$ is the sufficient statistic and $T(x)/\sigma_0^2$ has a χ^2 distribution with n degrees of freedom and therefore b and d are such that

$$\int_b^d \frac{T(x)}{\sigma_o^2}dx = 1-\alpha$$

and, for unbiasedness,

$$\int_b^d \frac{xT(x)}{\sigma_o^2}dx = n\cdot(1-\alpha) \quad .$$

Equivalently, the second statement implies that -

$$\left(\frac{b}{d}\right)^n = e^{(b-d)}$$

These two equations can be solved iteratively to yield appropriate values of b and d as required. For example, for n=20, α=0.05, the solution of b≈9.96 and d≈35.23 was obtained.

3.2 Other Difficulties

With the results presented so far, it would appear that today's statisticians are in an excellent position to provide optimal (UMP or UMPU) tests, especially when it is an exponential family population from which the sample is drawn. Unfortunately, this optimality often hinges upon some of the exponential families "nice" properties. In practice, this will often not be the case. In fact, the common hypothesis testing problem involving mixture density families provides an absurd example of the difficulties that can arise.

Proposition 3.2a:

Suppose the $f_\theta(x) = \theta \cdot f(x) + (1-\theta) \cdot g(x)$, for some $\theta \in [0,1]$, and that we wish to test the hypothesis that - H_0: $\theta \le \theta_0$ or $\theta \ge \theta_1$ vs. H_1: $\theta_0 < \theta < \theta_1$. Then, the trivial test $\phi^*(\mathbf{x}) = \alpha$ is a UMP test.

Proof 3.2a:

Consider any other α-level test $\phi(x)$. Necessarily, $E_0[\phi(\mathbf{X})] \le \alpha$ and $E_1[\phi(\mathbf{X})] \le \alpha$. But then, for $\theta \in (\theta_0, \theta_1)$ -

$$E_\theta[\phi(\mathbf{X})] = \theta \cdot E_0[\phi(\mathbf{X})] + (1 - \theta) \cdot E_1[\phi(\mathbf{X})]$$
$$\le \theta \cdot \alpha + (1 - \theta) \cdot \alpha = \alpha$$

so our trivial test, ϕ^*, is at least as powerful.■

This proposition seems to indicate that there is no benefit to be gained from sampling at all; the statistician would be best to base a choice of null or alternative hypothesis on chance alone. However, our search for a best test in this, and other,

situations is not completely hopeless. Once again, all we really need to do is redefine the notion of what "best" really means.

3.3 Locally Optimum Tests

Since all earlier notions of best rely on a comparison of the power functions of competing tests, is seems only reasonable that the notions of best for these new situations follow suit. However, focus now falls on the local behaviour of these power functions, not on the global behaviour as before (Fraser, 1968).

Consider the one sided test of H_0: $\theta = \theta_0$ vs. H_1: $\theta > \theta_0$. In this case, we consider all α-level test functions and try to maximize the derivative of the power function at θ_0, namely $\beta'(\theta_0)$.

Proposition 3.3a:

The locally most powerful (LMP) α-level test is given by

$$\phi^*(x) = I\left[\frac{\partial \ln f_{\theta_0}(x)}{\partial \theta_0} > c\right] = I_{[U.(\theta_0) > c]}$$

where c is chosen so that $\beta^*(\theta_0)=\alpha$ and $U.(\theta_0)$ is the score statistic -

$$U.(\theta_0) = \frac{\sum_{i=1}^{n} \frac{\partial \log f_\theta(x_i)}{\partial \theta_0}}{n}$$

Proof 3.3a:

Consider any other level α test, $\phi(\mathbf{x})$. As before, we use the LRI to see that -

$$\beta^{*\prime}(\theta_0) - \beta'(\theta_0) = E_{\theta_0}\left[(\phi^*(X) - \phi(X)) \cdot \frac{\partial \ln f_{\theta_0}(X)}{\partial \theta_0}\right]$$

$$\geq E_{\theta_0}\left[(\phi^*(X) - \phi(X)) \cdot \left(\frac{\partial \ln f_{\theta_0}(X)}{\partial \theta_0} - c\right)\right] \geq 0$$

and thus $\phi^*(x)$ is LMP as we set out to show.■

Unfortunately, this proposition cannot be extended to the two-sided testing situation. The test can only guarantee maximal power on one side. It would make sense to use -

$$\phi(x) = I_{[|U.(\theta_0)| > c]}$$

with c chosen so that $\beta(\theta_0) \leq \alpha$. However, such a modification will not even guarantee that the test is unbiased at θ_0. Instead, we hope to find a test for which the derivative of the power function vanishes at θ_0, yet has maximum curvature at that point.

Proposition 3.3b:

If $f_\theta(x)$ is sufficiently smooth within a neighbourhood of θ_0, the LMPU test of H_0: $\theta=\theta_0$ vs. H_1: $\theta \neq \theta_0$ for which $\beta(\theta_0) = \alpha$, $\beta'(\theta_0) = 0$, and $\beta''(\theta_0)$ is maximized is given by -

$$\phi^*(x) = I_{\left[\frac{\partial^2 f_{\theta_0}(x)}{\partial \theta_0^2} \geq c \cdot \frac{\partial f_{\theta_0}(x)}{\partial \theta_0} + d \cdot f_{\theta_0}(x)\right]}$$

where c and d are chosen appropriately to satisfy the listed constraints.

Proof 3.3b:

Consider any other unbiased, α-level test $\phi(\mathbf{x}) \in D_U$. By definition, $\beta(\theta_0) \leq \alpha$ and $\beta'(\theta_0)=0$. So,

$$\beta^{*''}(\theta_0)-\beta''(\theta_0) = \frac{\partial^2}{\partial\theta_0^2}\left[E_{\theta_0}[\phi^*(X)-\phi(X)]\right] = \frac{\partial^2}{\partial\theta_0^2}\left[\int(\phi^*(x)-\phi(x))\cdot f_\theta(x)dx\right]$$

$$= \int(\phi^*(x)-\phi(x))\cdot\frac{\partial^2 f_{\theta_0}(x)}{\partial\theta_0^2}dx = E_{\theta_0}\left[(\phi^*(X)-\phi(X))\cdot\frac{1}{f_{\theta_0}(X)}\cdot\frac{\partial^2 f_{\theta_0}(X)}{\partial\theta_0^2}\right]$$

$$\geq E_{\theta_0}\left[(\phi^*(X)-\phi(X))\cdot\left(\frac{1}{f_{\theta_0}(X)}\cdot\frac{\partial^2 f_{\theta_0}(X)}{\partial\theta_0^2} - c\cdot\frac{1}{f_{\theta_0}(X)}\cdot\frac{\partial f_{\theta_0}(X)}{\partial\theta_0} - d\right)\right] \geq 0$$

for any $c,d \geq 0$. Why? Well, we have subtracted off -

$$c\cdot\frac{\partial}{\partial\theta_0}E_{\theta_0}[\phi^*(X)-\phi(X)] = c\cdot[\beta^{*'}(\theta_0)-\beta'(\theta_0)] = 0$$

and

$$d\cdot E_{\theta_0}[\phi^*(X)-\phi(X)] \geq d\cdot(\alpha-\alpha) = 0$$

thus a necessarily non-negative quantity. Therefore, $\phi^*(\mathbf{x})$ does indeed maximize $\beta''(\theta_0)$ and will be the locally most powerful unbiased test as was set out to show.■

3.4 The Problem of Nuisance Parameters

There are other problems which are frequently encountered when trying to apply the theoretical results already presented. For instance, when we were successful in identifying best tests, the α-level of the test was governed by the boundary

points of Ω_0; this property being the result of the convexity of the power function within the exponential family. However, if the boundary points of a Ω_0 do not dominate the type I error for all $\theta \in \Omega_0$, then the problem of finding a best unbiased test is, as demonstrated by example 3.1a, difficult indeed. As a second example, sometimes both Ω_0 and Ω_1 are of the same dimension as Ω and thus there are many boundaries to consider. In other problematic cases, the underlying population densities under the competing hypotheses are from different families altogether.

One commonly encountered problem that can be dealt with involves the presence of nuisance parameters. In these cases, the parameter space can be separated as follows -

$$\Omega = \Omega_\psi \times \Omega_\lambda \qquad \text{and} \qquad \Omega_0 = \Omega_{\psi_0} \times \Omega_\lambda$$

where λ, possibly multi-dimensional, is the nuisance parameter and ψ is the parameter of interest upon which our inferences are based.

Obviously, it would simplify the problem greatly if we could identify tests whose Type I error rate is independent of the nuisance parameters. When testing $H_0: \psi = \psi_0$ vs. $H_1: \psi \neq \psi_0$, such a test is referred to as α-level similar and mathematically -

$$E_{(\psi_0,\lambda)}[\phi(X)] = \alpha \qquad \text{for all } \lambda \in \Omega_\lambda.$$

In trying to find such a test, it is natural to first find a sufficient statistic, S_λ, for λ which is independent of ψ. Then, the conditional density function $f_{(\psi,\lambda)}(x \mid S_\lambda = s)$ will be independent of λ and our optimal test statistic can be based on it rather than the usual density function. As an example, the testing of the mean of a

normal distribution when the variance is unknown involves the use of the conditional student t statistic, rather than the unconditional z-score statistic when the variance is known.

Another method which often simplifies the search for an optimal test in the presence of nuisance parameters involves the principle of invariance (Fraser, 1968 or Zacks, 1971).

Definition 3.4a:

Consider $G = \{g: S \rightarrow S\}$, a group of transformations on the sample space S. The induced group, under the requirement that -

$$P_{\theta}[g(X) \in A] = P_{g^*(\theta)}[X \in A]$$

for any $A \subset S$, of transformations on the parameter space Ω is $G^* = \{g^*: \Omega \rightarrow \Omega\}$. The testing problem $H_0: \theta \in \Omega_0$ vs. $H_1: \theta \in \Omega_1$ is invariant under G if the two parameter spaces, Ω_0 and Ω_1, are invariant under G^*. Similarly, a test, $\phi(\mathbf{x})$, is invariant under G if $\phi(g(\mathbf{x})) = \phi(\mathbf{x})$ for all $\mathbf{x} \in S$ and $g \in G$, and a statistic, $T(\mathbf{x})$, is invariant under G if $T(\mathbf{x})=T(g(\mathbf{x}))$. Finally, $T(\mathbf{x})$ is maximal invariant under G if $T(\mathbf{x}) = T(\mathbf{x}') \rightarrow \mathbf{x}' = g(\mathbf{x})$ for some $g \in G$.

Intuitively, the ideas of invariance correspond to certain ideas of symmetry which are often inherent to the problem. For example when modelling, if we shift our observations in a direction parallel to the projection plane, we expect the length of any model based residuals to remain constant and any statistics or tests for these models should reflect this.

It is easy to show that if $M(\mathbf{x})$ is a maximal invariant statistic, then any invariant test must be a function, $h[M(\mathbf{x})]$ of it. Thus, finding such a statistic often

simplifies the problem of finding an optimal, invariant test. In fact, after this statistic has been identified the testing problem is often simplified to the simple hypothesis case. In some situations, there are a number of invariant transformation groups and we can perform this reduction by invariance for each of them consecutively. In addition, these two methods, reduction by sufficiency and reduction by invariance, are often both applicable to the testing problem.

With the final addition of these results, the choice of a "best" testing procedure, in many situations, is an obvious one. The sole remaining challenge for the statistician lies in the fact that the critical regions of our test procedures have, up until this point, remained somewhat undefined. In order to achieve a desired level of significance, the statistician is faced with the problem of calculating tail probabilities of the test statistic under the null hypothesis. This problem can be a difficult one and much work has focused on the approximation of these tail probabilities.

3.5 Generalize Likelihood Ratio Test and Tail Probabilities

There is one important observation associated with the problem of approximating tail probabilities. This observation is that each of the test statistics for our UMP tests are in fact simple functions of the generalized likelihood ratio statistic for the hypothesis in question. Knowing that, under certain regularity conditions, twice the log likelihood ratio statistic converges, as $n \to \infty$, to a χ^2 distribution, the problem of finding tail probabilities is really quite easy. Unfortunately, this approximation, especially for small sample sizes, is quite crude.

Example 3.5a:

In the case of H_0: $\mu=0$ vs. H_1: $\mu \neq 0$ for observations from $N(\mu,\sigma^2)$, the generalized log likelihood ratio statistic $(\max_{\mu,\sigma} L(\mu,\sigma) - \max_{\sigma} L(0,\sigma))$ is -

$$W = n \cdot \log\left(1 + \frac{T^2}{n-1}\right)$$

where T is the students t_n statistic usually used for testing (Lawley, 1956). Thus, for t=4 and n=6, a standard t-table will yield an upper tail probability, P[W > t], of about 0.102. However, the χ^2 approximation yields a value of about 0.06; this is a very poor estimate.

To improve the chi-squared approximation, replace (Bartlett, 1953) W by $W' = W / E[W]$. This has the effect of improving the approximation from order $1/n$ to order $1/\sqrt{n}$. In our example, we use Taylor expansion to get -

$$W = n \cdot \left(\frac{T^2}{n-1} - \frac{T^4}{2 \cdot (n-1)^2}\right) + O_P\left(\frac{1}{n^2}\right)$$

Thus, $E[W] = 1 + 3/2 \cdot n + o(1/n)$ and, for example 3.5a, $P[W' > 4] \approx 0.093$. This correction provides significant improvement.

For the one sided hypothesis testing problem, twice the signed log-likelihood ratio statistic is used (Jensen, 1992) -

$$R = \text{sign}(\theta^* - \theta_0) \cdot \{2 \cdot [L(\theta^*) - L(\theta_0)]\}^{1/2}$$

The asymptotic approximation of the distribution of this statistic is, under the null hypothesis, normal and tail probabilities can be calculated accordingly. Again, improvements to this normal approximation have been investigated with Edgeworth Expansion (Barndorff-Nielsen, 1986) and Saddlepoint Methods (Reid, 1980) having been the most successful.

4. Sequential Hypothesis Testing

Wald's "Sequential Analysis" (Wald, 1947) provides an excellent theoretical introduction of this topic, while Wetherill and Glazebrook (1986) focuses more on the application of these techniques.

4.1 The Sequential Probability Ratio Test (SPRT)

Underlying all of the work presented to this point is the assumption that the sampling plan is fixed before the process of sampling begins. However, as alluded to in the introduction, there are other alternatives, most importantly that of sequential sampling. Although this sampling plan can be exploited to provide operational efficiencies in almost all hypothesis testing situations, it should be noted that this framework is often dictated by the problem at hand. As one example, when one is trying to detect changes in quality control, the relevant data is not observed all at once, but instead reveals itself one observation at a time and this fact should not be ignored. After all, it would obviously be unwise for the statistician to collect a weeks worth of production line data before running a hypothesis test on Friday afternoon, the results of which dictate that he tell the plant manager that all products made since noon on Monday are defective!

Under the most general decision framework, the sample is taken sequentially, and after each observation, the informed observer has the opportunity to either terminate sampling and make a decision based upon the information already available, or delay this decision and choose instead to view yet another observation. By considering such methods, one can hope to identify tests which would be more efficient when compared to our already identified optimal fixed sample size methods. They would hopefully achieve the same level of power with smaller sample sizes.

As before, we begin with the simplest hypothesis testing problem. Suppose that the set of possible distribution functions numbers only two, namely F_1 and F_2, and that the problem is one of determining which of these two distributions the sampled values are coming from (i.e., a hypothesis test with H_0: F_1 and H_1: F_2). In proposition 2.3a, the optimal fixed sampling plan test procedure for this problem was identified. This procedure, based on the likelihood ratio, minimized the probability of a Type II error (thus maximizing the power of the test) while holding the probability of Type I error below a suitably chosen value.

To search for an optimal sequential procedure, we need to return to the more general ideas of risk that were introduced earlier. However, for a sequential procedure the costs of sampling are not fixed and therefore the expected sample costs are also included as part of this risk. Namely,

Risk = (Cost of Type I Error × Probability of Type I Error) +
(Cost of Type II Error × Probability of Type II Error) +
(Cost per Observation × Expected Number of Observations)

For our problem, the chosen test procedure must provide a means of choosing between the three possible decisions: d_0 - in which we take an additional observation, and d_1 or d_2 - in which we terminate experimentation and declare that either F_1 or F_2 is the true distribution. Assuming that the losses of any terminal decision are given by $W(F_i,d_j) = W_{ij}$ which are greater than 0 when $i \neq j$ and equal to 0 when $i = j$, and that the costs of sampling are, by a suitable change of scale if necessary, equal to the number of observations, then Wald has shown that there exists a decision rule which minimizes risk. In fact, he provided the following risk minimizing decision rule - the sequential probability ratio test.

Let -

$$\ell_n = \prod_{i=1}^{n} \frac{f_{\theta_1}(x_i)}{f_{\theta_0}(x_i)} \quad .$$

Choose constants $B < A$ so that the Type I and Type II errors are restricted appropriately. Terminate experimentation, d_0, the first time, N, that $\ell_n \notin (B,A)$. Make decision d_1 if $\ell_N \leq B$, or decision d_2 if $\ell_N \geq A$.

Proposition 4.1a:

Choosing $B = \beta/(1-\alpha)$ and $A = (1-\beta)/\alpha$ within the SPRT will yield appropriate error probabilities.

Proof 4.1a:

$$\alpha = Pr_0[\ell_N \geq A] = E_0[I_{[\ell_N \geq A]}] = E_1[\frac{1}{\ell_N} \cdot I_{[\ell_N \geq A]}] \approx \frac{1}{A} \cdot E_1[I_{[\ell_N \geq A]}] = \frac{1-\beta}{A}$$

and similarly, $\beta \approx B \cdot (1 - \alpha)$. Together, these yield $B \approx \beta / (1-\alpha)$ and $A \approx (1-\beta) / \alpha$.■

Wald and Wolfowitz (1948) were able to show that the choices of A and B provided above are optimal. They provide, among all other tests with similarly restricted error probabilities, the test of minimal sample size. Thus, for this simplest hypothesis setup, the statisticians' choice of best is clear.

The efficiency of this test procedure is demonstrated with the following example.

Example 4.1a:

If testing H_0: $\theta = -0.5$ vs. H_1: $\theta = +0.5$, when $X \sim N(\theta,1)$, and with

$\alpha = \beta = 0.05$, then a fixed sample test would require a sample size of approximately 9.6. Alternatively, the expected sample size of the SPRT under either hypothesis is, as shown below, approximately 5.3 - an efficiency of more than 40%.

Proposition 4.1b:

The expected sample size for the hypotheses of example 4.1a is approximately 5.3.

Proof 4.1b:

Let $Z_n = \ln(f_1(x_i)) - \ln(f_0(x_i))$ so that $S_N = \sum_i Z_i = \ln(\ell_N)$. Then, $E[S_N] = E[N] \cdot E[Z_i]$ and thus -

$$E_0[N] = \frac{E_0[S_N]}{E_0[Z_1]} \approx \frac{(1 - \alpha) \cdot \ln(B) + \alpha \cdot \ln(A)}{-0.5} \approx 5.3$$

Similarly, $E_1[N] \approx 5.3$ as well.■

4.2 Generalizing the Optimal Bayes Solution

Following the notation of section 2.4, consider again the multiple hypothesis situation with H_i: $\theta = \theta_i$ for i=1,...,k. As in the single alternative case, a sequential testing procedure will need to select a stopping time, N, at which some decision, say δ_N, is made. There are several ways in which this might be done.

One way is to use likelihood ratio statistics, as the SPRT does.
Let -

$$N_j = \inf_{n \geq 0} \left[\frac{\max_{i \neq j} f_i(x_1,\ldots,x_m)}{f_j(x_1,\ldots,x_m)} \geq B \right]$$

be the time at which we reject H_j, and furthermore let $N_{(1)} \leq \ldots \leq N_{(K)}$ be the time ordered rejection times. Then, our decision rules has -

$$N = N_{(K-1)} \qquad \text{and} \qquad \delta_N = d_{k^*}.$$

where $k^* = \arg\max_i \{N_{(i)}\}$. It is easily shown that $P_k[N < \infty] \leq 1/B$ for all k and that -

Proposition 4.2a

$$E_k[N] \approx \frac{\log B}{\min_{j \neq k} D(f_k, f_j)} \qquad \text{as } B \to \infty$$

where -

$$D(f_i, f_j) = \int f_i \cdot \ln \frac{f_i}{f_j} dx \quad .$$

This is the Kullback-Leibler distance.

Proof 4.2a:

Similar to 4.2b shown below.■

A second alternative borrows from Bayesian techniques. Assuming linear cost structures for the sample size and decision errors, the SPRT was presented as the optimal sequential procedure in the single alternative situation. It should be noted that by modifying ℓ_n to reflect any *a priori* knowledge (i.e., $P[H_i \text{ is true}] = p_i$, $\sum_i p_i = 1$), this test will also provide the Bayes solution to the problem (Wald and Wolfowitz, 1950). When this approach is taken, the optimal test procedure is realized by stopping the sampling process as soon as -

$$p_n = \frac{p_0 \cdot \prod_{i=1}^{n} f_0(x_i)}{p_0 \cdot \prod_{i=1}^{n} f_0(x_i) + p_1 \cdot \prod_{i=1}^{n} f_1(x_i)}$$

falls outside of $(\underline{p}, \bar{p})$, and accepting either H_0 (if $p_n > \bar{p}$) or H_1 (if $p_n < \underline{p}$) accordingly. This test and the SPRT are equivalent as can be seen by setting -

$$A = (1-\underline{p})/\underline{p} \cdot p_1/(1-p_1) \quad \text{and} \quad B = (1-\bar{p})/\bar{p} \cdot p_1/(1-p_1), \quad \text{or}$$

$$\bar{p} = p_1 \cdot (1-\alpha) / [p_1 \cdot (1-\alpha-\beta) + \beta] \quad \text{and} \quad \underline{p} = p_1 \cdot \alpha / [1-\beta-p_1 \cdot (1-\alpha-\beta)].$$

Generalizing for the multiple hypothesis situation, let $\{\pi_j\}$ and $\{p^j_n = P[H_j|x_1,...,x_n]\}$ be the *a prior* and *a posteriori* distributions for $\{H_j\}$ respectively with -

$$p_n^j = \frac{\pi_j \cdot \prod_{i=1}^{n} f_j(x_i)}{\sum_{k=1}^{K} \pi_k \cdot \prod_{i=1}^{n} f_k(x_i)} \quad .$$

For constants $\{A_j : j = 1,...,k\}$, define -

$$N_{A_j} = inf_{n \geq 1}\left\{p_n^j > \frac{1}{1+A_j}\right\} = inf_{n \geq 1}\left\{\frac{P[\overline{H_k}|x_1,...,x_n]}{P[H_k|x_1,...,x_n]} < A_j\right\} \quad .$$

Here, $\{A_j\}$ are chosen so that - $\alpha_j = P[\text{reject } H_j] = \sum_{i=1,...,k} \pi_i \cdot \alpha_{ij} \leq \pi_j \cdot A_j$.
Then, our decision rule is given by -

$$N = \min[N_{A_1},...,N_{A_k}] \quad and \quad \delta_N = d_{k^*} .$$

$$where \quad k^* = arg\min\{N_{A_1},...,N_{A_k}\} \quad .$$

Proposition 4.2b:

We can show that (Baum and Veeravalls, 1994) -

$$E_j[N] \approx \frac{-\log A_j}{\min_{i \neq j} D(f_j, f_i)}$$

as max $A_j \rightarrow 0$.

Proof 4.2b:

Consider that -

$$N_{A_j} = inf_{n \geq 1} \left\{ -\frac{1}{n} \cdot log\left(\sum_{i \neq j} exp\left[-n\left[\frac{1}{n} \cdot \sum_{k=1}^{n} log \frac{f_j(x_k)}{f_i(x_k)} + \frac{1}{n} \cdot log \frac{\pi_j}{\pi_i}\right]\right]\right) > \frac{-\log A_j}{n} \right\}$$

and that under H_j -

$$\frac{1}{n} \cdot \sum_{k=1}^{n} log \frac{f_j(x_k)}{f_i(x_k)} + \frac{1}{n} \cdot log \frac{\pi_j}{\pi_i} \rightarrow D(f_j, f_i)$$

by the Law of Large Numbers.

Thus, -

$$-\frac{1}{n} \cdot log\left(\sum_{i \neq j} exp\left[-n\left[\frac{1}{n} \cdot \sum_{k=1}^{n} log \frac{f_j(x_k)}{f_i(x_k)} + \frac{1}{n} \cdot log \frac{\pi_j}{\pi_i}\right]\right]\right) \rightarrow min_{i \neq j} D(f_j, f_i)$$

as only the term of i corresponding to $\min_{i \neq j} D(f_j, f_i)$ dominates the convergence. Thus -

$$N_{A_j} \sim \frac{-\log A_j}{\min_{i \neq j} D(f_j, f_i)}$$

Next, consider that -

$$P_{f_j}\left[\left| \frac{N}{-log A_j} - \frac{1}{\min_{i \neq j} D(f_j, f_i)} \right| > \epsilon\right]$$

$$\leq P_{f_j}\left[\left|\frac{N_{A_j}}{-\log A_j} - \frac{1}{\min_{i \neq j} D(f_j, f_i)}\right| > \epsilon\right] + P_{f_j}[\overline{H_j}]$$

$$\leq P_{f_j}\left[\left|\frac{N_{A_j}}{-\log A_j} - \frac{1}{\min_{i \neq j} D(f_j, f_i)}\right| > \epsilon\right] + \pi_j \cdot A_j \rightarrow 0$$

as max $A_j \rightarrow 0$.■

As has already been stated, when k=2, Wald and Wolfowitz were able to prove the optimality of the SPRT; it minimizes, among all procedures with identical error probabilities, the expected sample size under both the null and alternative hypotheses. Now, neither of the two multiple hypotheses procedures introduced above can guarantee that the error probabilities satisfy the exact requirements. This is not a major drawback however as, like the SPRT, they do allow us to construct conservative tests. Unfortunately, the optimality enjoyed by the SPRT process is not guaranteed by either of the two procedures which have been introduced. In fact, an optimal procedure would require dynamic programming to solve for the optimal stopping boundaries. This is a significant drawback. Instead, we hope to adopt a simple procedure which is asymptotically optimal (Wald, 1941) as the desired error probabilities approach zero.

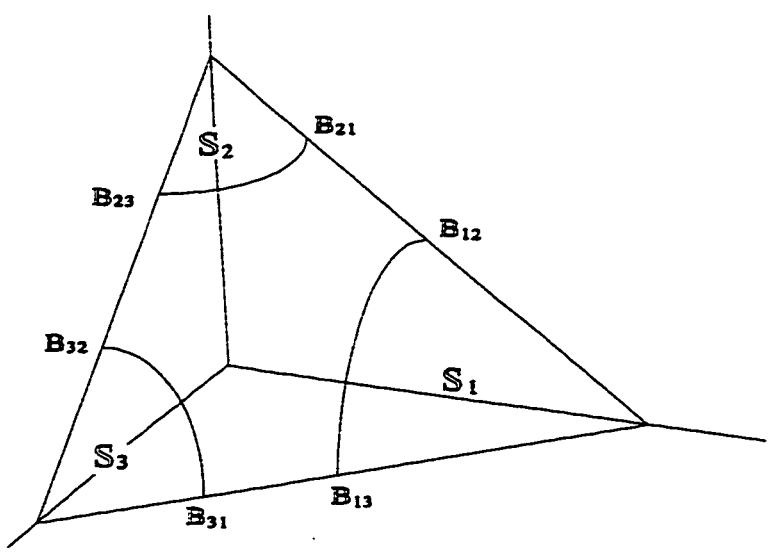
Consider the same general framework as our SPRT problem, but with three possible distribution functions, F_1, F_2, and F_3. The set of all possible *a priori* distributions for these three hypothesis can be represented by the points (ξ_1, ξ_2, ξ_3) of a triangular plane, T, in $\mathbb{R}^3$. After m observations have been made, the *a posteriori* distribution of our three hypotheses is given by the following -

$$\xi_m^i = \frac{\xi_0^i f_i(x_1) f_i(x_2) \cdots f_i(x_m)}{\sum_{j=1}^{3} \xi_0^j f_j(x_1) f_j(x_2) \cdots f_j(x_m)} \quad .$$

Together, these three probabilities (i=1,2,3) are also represented by a point, ξ_m, of our triangle. Obviously, if this point is one of the vertices of triangle T, say V_i, then the true distribution function of the random variables is F_i. In fact, it makes sense that if the *a posteriori* point is close to a vertex of T, then the true distribution function is likely that represented by this vertex.

Thus, a reasonable decision rule might involve a sequential procedure whereby one would construct a subset around each vertex, compute after each observation the point ξ_m, continue to take observations until this point lies within one of the constructed subsets, and conclude that the true distribution is F_i if at the subset S_i is the first one entered.

Figure 4.1 The three hypothesis situation



Wald and Wolfowitz were able to show not only that there exists three closed and convex subsets which make the proposed procedure optimal, but also that these subsets are independent of the *a priori* distribution. Thus, as was the case with choosing the constants A and B in the situation involving two probability distributions, constructing the three subsets, S_1, S_2, and S_3, in the situation involving three probability distributions depends only on the values of -

W_{ij}, the losses of concluding that F_j is the true distribution when F_i is, and

c, the cost of obtaining an additional observation.

In fact, Wald and Wolfowitz were able to use Bellman's optimality rule (Bellman, 1957) to determine the boundary values of these three subsets as they lie along the edges of triangle T, and the tangents of the boundaries at these edge points. Bellman's optimality rule states that if -

$$A_J(\xi) = \int J\left(\frac{\xi^1 f_1(x)}{\sum_j \xi^j f_j(x)},\ldots,\frac{\xi^k f_k(x)}{\sum_j \xi^j f_j(x)}\right) \cdot \sum_{j=1}^{k} \xi^j f_j(x)\, dx$$

where $J(\xi)$ is the minimum cost using the stopping rule with prior probability ξ, then -

$$J(\xi) = \min\{(1 - \xi^1),\ldots, (1-\xi^k), c + A_J(\xi)\} \qquad \text{- with c the sampling cost}$$

Unfortunately, an explicit formulation of the boundaries of these three sets would require dynamic programming . This is unfortunate as the decision rule described above is readily extendable to any countable number, k, of probability distributions through the construction of k similar subsets in $\mathbb{R}^k$, and the sequential checking of whether or not the calculated ξ_m were in one of these subsets. We let $N_j = \inf_n\{\xi^j_n \in S_j\}$ where S_j is a convex set containing corner point e_j and our decision rule has $N^* = \min\{N_1,\ldots,N_k\}$ and $\delta^* = \arg\min\{N_1,\ldots,N_k\}$.

4.3 Calculation of Boundary Points

Since one knows some aspects of the boundary of the subsets, it seems reasonable to try to approximate these boundaries. The most general approximation would provide a function which coincided not only with the known edge points, but also with the known tangents of the boundary set at these edge points. In the three hypothesis problem for example, if the boundary of S_i is given by $f_i(x)$ for $x \in [x_{i0}, x_{i1}]$ and $f_i'(x_{i0})$ and $f_i'(x_{i1})$ are given, then the polynomial, $p_i(x)$, constructed as follows will coincide exactly with $f_i(x)$ and its derivatives at the points where it meets the boundary of T.

$$p_i(x) = a + b(x - x_{i0}) + c(x - x_{i0})^2 + d(x - x_{i0})^2(x - x_{i1})$$

where $a = f_i(x_{i0})$

$b = f_i'(x_{i0})$

$c = [f_i(x_{i1}) - a - b(x_{i1} - x_{i0})] / (x_{i1} - x_{i0})^2$

$d = [f_i'(x_{i1}) - b - 2c(x_{i1} - x_{i0})] / (x_{i1} - x_{i0})^2$

A less complex approximation can be achieved by just joining the two known edge values of each vertex with a straight line; a first degree, rather than third degree, polynomial approximation.

The logical question is: How good are these approximations of the boundaries?

Before answering this question, we must first calculate the appropriate values. To reiterate, Wald and Wolfowitz's paper describes a testing procedure in which *a posterior* probabilities are plotted sequentially in $\mathbb{R}^3$ and the decision is based upon that points inclusion in or exclusion from three sets around the vertices of triangle T, the set of all possible *a posterior* probabilities. They describe how, given loss functions W_{ij} and assuming a linear, unit (which we can always be

achieved by scaling appropriately) cost of sampling, one can calculate the boundary points of these sets.

Unfortunately, the SPRT described in section 4.1, and the one which is likely to be used in practice, has critical regions based on error probabilities, not the costs of these errors and the sampling costs. It would be preferable if the critical regions of the test were based on error probabilities, not costs.

Although Wald and Wolfowitz's discussion of critical regions is based upon the knowledge of these costs, it should be apparent that the specification of the loss functions is equivalent to specifying of the probability of the various possible errors for the chosen test procedure. In the case of two dimensions for example, specifying the ratio of Type I and Type II errors is identical to specifying the relative loss functions for the two possible errors. As a Type I error becomes more expensive relative to a Type II error, the probability of making this more expensive Type I error should decrease in our chosen sequential test. As well, the selection of the combined probabilities of the various error types is equivalent to specifying the relative cost of sampling. As the combined probability of making an incorrect decision is allowed to increase, the expected number of items that will need to be sampled before reaching such a decision will be reduced.

Consider the determination of the boundary point B_{12} lying on the boundary of S_1 (see Figure 2). Wald and Wolfowitz argue that as the probability that H_3 is true is zero along this side of the triangle, an optimal sequential test will ignore this hypothesis altogether. Thus, when testing H_1 vs. H_2, B_{12} and B_{21} provide the critical values for this two sided test. Given α_{12} and α_{21}, this corresponds to selecting lower bound, $\underline{p}$, and upper bound, $\bar{p}$, as described in section 4.2.

Example 4.3a:

For $\alpha_{12} = 0.05$, $\alpha_{21} = 0.10$, and no *a priori* information ($p_1 = p_2 = 0.5$), we calculate $\underline{p} \approx 0.0526$ and $\bar{p} \approx 0.905$ and thus $B_{12} \approx (0.947, 0.0526, 0)$ and $B_{21} \approx (0.095, 0.905, 0)$. In similar fashion, we can determine the location of B_{13}, B_{31}, B_{23}, and B_{32}.

4.4 Simulation Study Results

The decision making procedure that has been chosen is asymptotically optimal. As error probabilities tend to zero, the average number of observations required by this procedure will be smaller than any other. However, the linear boundary approximations that have been made will have an effect on this optimality. To assess the effect, a number of simulation studies of different situations were conducted (see Appendix A). For each trial, one of the three competing hypotheses was randomly selected. Then, the sequential testing procedure, with appropriate error probabilities, was conducted with random observations generated from the distribution associated with the selected hypothesis. Based upon the results of these sequential tests, empirical error probabilities and sample size distributions (based upon 10,000 trials) were calculated.

The first question which needs to be answered is - "Has our test exhibited the desired levels of error?" In making a linear approximation to the convex shaped critical regions, we have constructed a conservative test. The first half of each table below provides an assessment of the degree of this conservation by comparing the theoretical and empirical error probabilities. Within the table, α_{ij} refers to the theoretical error probability of the sequential test, and e_{ij} refers to the experimental error probability that was observed. We present e_{ij} / α_{ij} which provides the proportion (%) of the actual error rate that was observed during simulation. Thus, a figure of 80% would indicate that experimental error

probability was about 20% smaller that the theoretical one.

As was pointed out in section 4.1, the sequential probability ratio test, in the simple hypothesis situation, can provide a considerable sampling costs savings. The second question which needs to be answered is "Does the sequential nature of our tests provided much benefit?"

For any sequential procedure with error probabilities α_{ij}, we have the following (Simons, 1968).

Proposition 4.4a:

$$E_i[N] \geq \max_{j \neq i} \left[\frac{\sum_{r=1}^{k} \alpha_{ir} \cdot \ln \frac{\alpha_{ir}}{\alpha_{jr}}}{D(f_i, f_j)} \right]$$

Proof 4.4a:

Using Wald's identity and conditional probabilities, we know that -

$$E_i\left[\sum_{m=1}^{N} \ln \frac{f_j(X_m)}{f_i(X_m)}\right] = E_i[N] \cdot E_i\left[\ln \frac{f_j(X)}{f_i(X)}\right] = \sum_r \alpha_{ir} E_i^r\left[\sum_{m=1}^{N} \ln \frac{f_j(X_m)}{f_i(X_m)}\right]$$

where $E_i^r[\cdot]$ denotes the expectation under H_i when we accept H_r, for i,r = 1,...,k.. By Jensen's inequality -

$$E_i^r\left[\sum_{m=1}^{N} \ln \frac{f_j(X_m)}{f_i(X_m)}\right] \geq \ln E_i^r\left[\prod_{m=1}^{N} \frac{f_j(X_m)}{f_i(X_m)}\right]$$

and by change or measure -

$$E_i^r\left[\prod_{m=1}^{N}\frac{f_j(X_m)}{f_i(X_m)}\right] = \frac{\alpha_{jr}}{\alpha_{ir}} \quad .$$

Thus, $E_i[N]\cdot E_i[\ln(f_i(X)/f_j(X)] \geq \sum_r \alpha_{ir}\cdot\ln(\alpha_{ir}/\alpha_{jr})$, and the proof is complete.■

Proposition 4.4a would allow us to measure the effects of boundary approximation on optimal sample size. However, our real desire is to compare the sample sizes for our empirical procedure against those of equivalent non-sequential procedures.

Consider our first simulation where we were faced with the problem of testing, with $X_i \sim$ Normal(μ,1), H_1: $\mu = -0.5$ vs. H_2: $\mu = 0$ vs. H_3: $\mu = +0.5$. If one were only testing H_1 vs. H_2 with $\alpha_{12}=\alpha_{21}=0.05$, then we would select -

$$n_{12}=\frac{[\Phi^{-1}(1-\alpha_{12})-\Phi^{-1}(\alpha_{21})]^2}{[\theta_1-\theta_2]^2}\approx\frac{(1.645--1.645)^2}{(-0.5-0)^2}\approx 43.3$$

To achieve the desired levels of Type I and Type II error. Similarly, we would have $n_{23}\approx 43.3$ and $n_{13}\approx 10.8$ to respectively test H_2 vs. H_3 and H_1 vs. H_3. Taking the maximum of these three sample sizes, a non-sequential procedure with equivalent error probabilities would require at least 44 observations to differentiate between these three hypotheses. The second half of each table below presents the average sample sizes that were obtained during simulation. Beside this is the percentage of savings that this represents over an equivalent non-sequential procedure. Thus, a figure of 30% would indicate that the sequential procedure required, on average, only 70% of the observations that an optimal non sequential procedure would have. The final column presents the percentage of

decision procedures which had sample sizes below that of an equivalent non sequential procedure. A figure of 90% would indicate that only one in ten sequential procedures required a sample size larger than that of a non sequential procedure with equivalently restricted error probabilities.

Consider the following hypothesis testing situation.

$X_i \sim \text{Normal}(\mu,1)$ $\quad H_1: \mu = -0.5$ vs. $H_2: \mu = 0$ vs. $H_3: \mu = +0.5$

Table 4.1 Least likely errors

	$\alpha_{21} = \alpha_{23} = 0.05$		$\alpha_{12} = \alpha_{32} = 0.05$		changing		Sample Size	% Savings	%-tile
$\alpha_{13} = \alpha_{31} = 0.01$	84.3	81.7	37.8	41.3	<0.1	1.5	30.1	30.5	81
$\alpha_{13} = \alpha_{31} = 0.05$	77.3	75.7	36.3	38.5	0.1	0.1	30.2	30.3	81
$\alpha_{13} = \alpha_{31} = 0.10$	76.0	71.9	32.9	36.4	<0.1	<0.1	30.4	29.8	81
$\alpha_{13} = \alpha_{31} = 0.20$	74.9	77.8	36.9	39.9	<0.1	<0.1	30.2	30.3	81
$\alpha_{13} = \alpha_{31} = 0.30$	70.7	81.6	37.2	44.6	<0.1	0.1	30.1	30.5	81

With these simulations, we attempt to see what effect a change in the least likely errors will have. In all cases, the observed error rates for α_{13} and α_{31} were almost nil. As a result of this, increases in the theoretical error probabilities had no consistent effect on either the other error rates, or the average sample sizes. Because the first row of this table presents what, in practice, might be a fairly typical situation, it is interesting to note that the observed error rates were all at least 15% below the theoretical probabilities. However, though conservative, the procedure still provides a significant sample size savings of about 30%.

Table 4.2 Increasing probabilities of error when middle hypothesis true

	$\alpha_{21}=\alpha_{23}=0.05$		changing		$\alpha_{13}=\alpha_{31}=0.01$		Sample Size	% Savings	%-tile
$\alpha_{12}=\alpha_{32}=0.01$	78.8	68.6	39.3	44.6	<0.1	<0.1	36.2	42.7	88
$\alpha_{12}=\alpha_{32}=0.05$	84.4	81.7	37.8	41.3	<0.1	1.5	30.1	30.5	81
$\alpha_{12}=\alpha_{32}=0.10$	76.3	77.5	38.0	40.7	<0.1	<0.1	27.1	21.0	75
$\alpha_{12}=\alpha_{32}=0.20$	93.2	76.9	41.0	41.1	<0.1	<0.1	23.3	5.7	65
$\alpha_{12}=\alpha_{32}=0.30$	91.5	97.8	39.6	62.6	<0.1	<0.1	20.4	5.6	65

Within these simulations, we attempt to see the effects of an increase in the probability of making an incorrect decision when the middle of the three ordered hypotheses is true. The increase in α_{12} and α_{32} has no consistent effect on the relative error rates that were observed; most were about 40% the size of the desired probability. However, these increases resulted in a fairly consistent increase in each of the two complementary error rates. As would be expected, as these probabilities increase sample sizes are noticeably smaller, however the relative savings of the sequential procedure over a non sequential one decrease.

Table 4.3 Curving the boundaries when middle hypothesis true

	$\alpha_{21}=\alpha_{23}=0.05$		changing		$\alpha_{13}=\alpha_{31}=0.01$		Sample Size	% Savings	%-tile
α_{12}=0.01 α_{32}=0.30	66.8	99.0	56.9	22.2	<0.1	<0.1	29.9	52.6	93
α_{12}=0.03 α_{32}=0.25	78.9	94.5	43.3	27.4	<0.1	<0.1	27.6	44.5	90
α_{12}=0.05 α_{32}=0.20	68.9	89.9	37.2	33.3	3.0	<0.1	27.2	37.2	86
α_{12}=0.08 α_{32}=0.15	89.7	72.6	41.0	33.4	2.9	<0.1	26.6	28.5	81
α_{12}=0.10 α_{32}=0.10	76.3	77.5	38.0	40.7	<0.1	<0.1	27.1	21.0	75

With these simulations, we again vary the probabilities of making an incorrect decision when the middle hypothesis is true. However, by increasing one of these

probabilities while decreasing the other, we create varying degrees of curvature in the S_1 and S_3 boundaries. The table shows that as α_{32} decreases, the relative rate observed for it increases, while that of its complimentary error decreases. However, the increase in α_{12} does not result in consistent decrease in its observed relative error rate, or an increase in its compliments. It appears as though the linear approximation only becomes a factor affecting relative error rates at higher probability levels (>5%). The effect on sample sizes is interesting. As we proceed down the rows of table 4.3, the total probability of error decreases and sample sizes should increase. However, the effect on sample size is minimal and, if anything, decreases. In addition, the relative efficiency of the sequential procedure over a non sequential decreases quite rapidly. This is in opposition to the effect of increasing total error rates seen in table 4.2.

Table 4.4 Increasing probabilities or error when middle hypothesis selected

	changing		$\alpha_{12}=\alpha_{32}=0.05$		$\alpha_{13}=\alpha_{31}=0.01$		Sample Size	% Savings	%-tile
$\alpha_{21}=\alpha_{23}=0.01$	72.3	66.3	33.9	34.4	<0.1	<0.1	39.7	37.1	88
$\alpha_{21}=\alpha_{23}=0.05$	84.4	81.7	37.8	41.3	<0.1	1.5	30.1	30.5	81
$\alpha_{21}=\alpha_{23}=0.10$	71.6	73.6	34.2	39.1	<0.1	3.0	25.5	25.7	75
$\alpha_{21}=\alpha_{23}=0.20$	65.1	59.5	42.5	44.4	6.1	12.0	21.5	13.0	61
$\alpha_{21}=\alpha_{23}=0.30$	53.3	58.1	35.6	41.4	20.9	23.7	18.8	13.0	64

Here, we look at the effects of increasing the probability of incorrectly selecting the second hypothesis when either the first or third is true. As they increase, their relative error rates generally decrease. The complimentary error rates remain consistently around 40%. Interesting change occurs in the α_{13} and α_{31} relative error rates which rise, for the first time, above 10%. As was seen in Table 4.2, the increasing error probabilities result in decreasing sample sizes, but reductions in the savings over non sequential procedures.

Table 4.5 Curving the boundary when middle hypothesis selected

	changing		$\alpha_{12}=\alpha_{32}=0.05$		$\alpha_{13}=\alpha_{31}=0.01$		Sample Size	% Savings	%-tile
α_{21}=0.01 α_{23}=0.30	53.8	52.8	44.1	40.0	17.9	<0.1	29.4	53.4	81
α_{21}=0.03 α_{23}=0.25	85.4	59.3	36.7	42.3	15.0	<0.1	26.7	46.3	85
α_{21}=0.05 α_{23}=0.20	70.6	67.6	52.9	42.9	11.6	<0.1	26.0	40.0	86
α_{21}=0.08 α_{23}=0.15	74.8	68.0	42.0	31.6	9.1	<0.1	25.1	32.5	87
α_{21}=0.10 α_{23}=0.10	71.6	73.6	34.2	39.1	<0.1	3.0	25.5	25.7	87

In this set of simulations, we again change the probability of incorrectly selecting the second hypothesis when either the first or third is true. Like in table 4.3, we increase one probability while decreasing the other. However in this situation, the results only affect the curvature of a single critical region, S_2. As in table 4.3, the changes at lower probability levels do not result in consistent changes in related relative error rates. But again, decreases at higher levels (>5%) are reflected in higher relative error rates. As in Table 4.2, decreases in total error probabilities again result in puzzling decreases in sample size. These decreases again result in decreases in the relative efficiency of our sequential procedure over non sequential ones.

Finally, we consider the following hypothesis testing situation -

$X_i \sim \text{Normal}(\mu,\sigma)$

$H_1: \mu = 0, \sigma = 1$ vs. $H_2: \mu = 0.5, \sigma = 1$ vs. $H_3: \mu = 0, \sigma = 1.5$

This hypothesis testing problem differs from the first under consideration in that the three hypothesis cannot be directly ordered.

Table 4.6 Non ordered hypotheses, increasing probabilities of error

	e_{21}/α_{21}	e_{23}/α_{23}	e_{12}/α_{12}	e_{32}/α_{32}	e_{13}/α_{13}	e_{13}/α_{31}	Sample Size	% Savings	%-tile
$\alpha_{ij} = 0.01$	55.8	21.1	45.8	3.1	69.3	47.0	40.8	79.6	100
$\alpha_{ij} = 0.05$	51.3	24.1	35.3	11.6	33.5	37.4	27.6	70.6	99
$\alpha_{ij} = 0.10$	51.5	29.9	41.9	11.6	39.7	37.4	21.4	60.4	97
$\alpha_{ij} = 0.20$	54.3	35.8	42.2	18.8	41.6	35.9	14.3	28.5	78

The increase in error probabilities has little consistent effect on most of the observed relative error rates, but does result in a considerable reduction in sample size. The relative efficiency of our sequential procedure over a non sequential ones reduces greatly as the error probabilities decrease, but even at high levels, the savings are substantial.

4.5 Conclusions

The results of the simulation study indicate that our sequential procedure is an impressive decision maker. It succeeded in providing sample size reductions, ranging from slight (about 5%) to extremely impressive (almost 80%), while continuing to restrict error probabilities below the required levels. The linear boundary region approximations employed resulted in an extremely conservative test with more than half the relative error rates being less than 50%. The third degree polynomial approximation which can be developed would hopefully improve upon this and provide even greater sample size reductions. However, as discussed in section 4.2, this requires the use of costs of loss to determine the critical regions of our test, instead of error probabilities that were used.

Considering the simplicity of our sequential procedure, its use should be considered whenever the multiple hypothesis testing problem presents itself.

4.5 Future Work

As pointed out in section 4.4, our sequential procedure is, at times, quite conservative. It would be interesting to investigate the characteristics of a sequential procedure with boundary regions based upon third degree polynomial approximations rather than the linear ones that our procedure employed. It is conjectured that such a procedure would provide a considerable reduction in the level of conservatism, and further impressive savings in sample size.

Although Wald was able to show that our sequential hypothesis test will, with probability one, terminate, our simulations show that the number of sampled observations is at times far greater than that which would be required by an equivalent non sequential procedure. It is for this reason that when sequential methods are employed in practice, the sampling process will often be truncated at some premature limit if a decision has not yet been reached. This practice will have an effect on the error probabilities of the procedure and these effects will of course be dependent upon the truncation point. Wald was able to provide crude upper bounds for these error probabilities (Wald, 1947) and his methods are directly transferable to our situation. It would be interesting to simulate a truncated multiple-hypothesis test and assess the accuracy of these bounds. It is suspected that, in most situations, the level of "safety" obtained by employing linear critical region boundary approximations within our sequential procedures would far exceed the level of "danger" incurred by truncating the procedure at most reasonable sample sizes. However, this conjecture remains to be investigated.

5. Bibliographical Notes

The idea of sufficiency was introduced by Fisher (1925) who is often considered the founder of modern statistics. Tests based on decision making were established by Neyman and Pearson (1933), with the more general approach of decision theory forwarded by Wald (1950). The initial materials in section 2 are standard and can be found in Lehmann (1986). The proofs and presentations given here are different and much simpler. Propositions about the monotone likelihood ratio and exponential families can be found in Karlin and Rubin (1956) and Barndorff-Nielsen (1978) respectively. The materials of section 2.4 are recent and extend the ideas of unbiasedness to the multiple hypothesis testing problem.

Section 3 outlines a number of advanced topics and related references include Lehmann (1986), Cox and Hinkley (1974), Zacks (1971), Reid (1988), and Fraser (1968).

Section 4 discusses the basic theory of sequential testing procedures where Wald (1947) is the most notable monograph. Important references include Wald and Wolfowitz (1948, 1950), Arrow, Blackwell, and Girshick (1949), and Siegmund (1985). Section 4.2 extends the results to multiple hypotheses cases and builds upon Simons (1968) and Baum and Veeravalls (1994).

6. References

Arrow, K.J., Blackwell, D., and Girshick, M.A., 1949. Bayes and minimax solutions of sequential decision problems. *Econometrica* **17**, 213-244.

Barndorff-Nielsen, O.E., 1986. Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307-322.

Bartlett, M.S., 1953. Approximate confidence intervals. *Biometrika* **40**, 12-19.

Baum, C.W. and Veeravalls, V.V., 1994. A sequential procedure for multiple hypothesis testing. *IEEE - Inference Theory* **40**, 1994-2007.

Bellman, R., 1957. *Dynamic Programming*, Princeton University Press, New Jersey.

Berger, James, 1980. *Statistical Decision Theory, Foundations, Concepts, and Methods*, Springer-Verlag, New York.

Cox, D.R. and Hinkley, D., 1974. *Theoretical Statistics*, Chapman and Hall, London.

Fisher, R.A., 1925. Theory of statistical estimation. *Cambridge Philosophical Society Proceedings* **22**, 700-725.

Fraser, D.A.S., 1968. *The Structure of Inference*, Wiley, New York.

Jensen, J.L., 1992. The modified signed likelihood statistic and saddle point approximations. *Biometrika* **79**, 693-703.

Karlin, S. and Rubin, H., 1956. The theory of decision procedures for distributions with monotone likelihood ratio. *Annals of Mathematical Statistics* **27**, 272-299.

Lawley, D.N., 1956. A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* **43**, 295-303.

Lehmann, E.L., 1955. Ordered families of distributions. *Annals of Mathematical Statistics* **26**, 399-419.

Lehmann, E.L., 1986. *Testing Statistical Hypotheses*, John Wiley and Sons, New York.

Neyman, J. and Pearson, E.S., 1933. *On the problem of the most efficient tests of statistical hypotheses*. Philosophical Transactions of the Royal Society - Series A **231**, 289-337.

Reid, N., 1980. *Saddlepoint methods and statistical inference*. Statistical Science **3**, 213-238.

Siegmund, D., 1985. *Sequential Analysis Tests and Confidence Intervals*, Springer-Verlag, New York.

Simons, G. 1968. Lower bounds for average sample number of sequential multihypothesis tests. *Annals of Mathematical Statistics* **38**, 1343-1364.

Wald, A., 1939. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics* **10**, 299-326.

Wald, A., 1941. Asymptotically most powerful tests of statistical hypotheses. *Annals of Mathematical Statistics* **12**, 1-19.

Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large, *American Mathematical Society Transactions* **54**, 426-482.

Wald, A., 1947. *Sequential Analysis*, John Wiley and Sons, New York.

Wald, A., 1950. *Statistical Decision Functions*, John Wiley and Sons, New York.

Wald, A. and Wolfowitz, J., 1948. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* **19**, 326-339.

Wald, A. and Wolfowitz, J., 1950. Bayes solutions of sequential decision problems. *Annals of Mathematical Statistics* **21**, 82-99.

Wetherill, G. Barrie and Glazebrook, Kevin D., 1986. *Sequential Methods in Statistics*, Chapman and Hall, New York.

Appendix A - Simulation Program Code

SAS program to simulate sequential multiple hypothesis tests and produce summary of results.

```
LIBNAME THESIS 'C:\THESIS\';

%LET ALPHA12=0.01;
%LET ALPHA13=0.01;
%LET ALPHA21=0.01;
%LET ALPHA23=0.01;
%LET ALPHA31=0.01;
%LET ALPHA32=0.01;
%LET MU1=0;     %LET SIGMA1=1;
%LET MU2=0.5;   %LET SIGMA2=1;
%LET MU3=0;     %LET SIGMA3=1.5;
%LET CNT1=200;
%LET CNT2=50;

DATA BOUNDARY;
 B12X=(1-&ALPHA12.)/(1-&ALPHA12.+&ALPHA21.); B12Y=1-B12X; B12Z=0;
 B21X=&ALPHA12./(1+&ALPHA12.-&ALPHA21.); B21Y=1-B21X; B21Z=0;
 B13X=(1-&ALPHA13.)/(1-&ALPHA13.+&ALPHA31.); B13Y=0; B13Z=1-B13X;
 B31X=&ALPHA13./(1+&ALPHA13.-&ALPHA31.); B31Y=0; B31Z=1-B31X;
 B23X=0; B23Y=(1-&ALPHA23.)/(1-&ALPHA23.+&ALPHA32.); B23Z=1-B23Y;
 B32X=0; B32Y=&ALPHA23./(1+&ALPHA23.-&ALPHA32.); B32Z=1-B32Y;
 OUTPUT;
RUN;

%MACRO TRIALS(COUNTER1,COUNTER2);

%DO I=1 %TO &COUNTER1.;

 DATA THESIS.SIM1(KEEP=H CHOICE N);
  SET BOUNDARY;
  FORMAT H CHOICE 1.  N 3.;
  SEED=-1; PI=ARCSIN(1)*2;
```

```
%DO J=1 %TO &COUNTER2.;
 CALL RANTBL(SEED,1/3,1/3,1/3,H);
 SELECT(H);
  WHEN(1) DO;
   MU=&MU1.; SIGMA=&SIGMA1.;
  END; /* WHEN H=1 */
  WHEN(2) DO;
   MU=&MU2.; SIGMA=&SIGMA2.;
  END; /* WHEN H=2 */
  WHEN(3) DO;
   MU=&MU3.; SIGMA=&SIGMA3.;
  END; /* WHEN H=3 */
 END; /* SELECT H */
 FLAG=0; P1=1/3;P2=1/3;P3=1/3; N=0;
 DO WHILE(NOT(FLAG));
  OBS=SIGMA*NORMAL(SEED)+MU; N=N+1;
  TEMP1=P1*1/(SQRT(2*PI)*&SIGMA1.)*EXP(-(((OBS-&MU1.)/&SIGMA1.)**2)/2);
  TEMP2=P2*1/(SQRT(2*PI)*&SIGMA2.)*EXP(-(((OBS-&MU2.)/&SIGMA2.)**2)/2);
  TEMP3=P3*1/(SQRT(2*PI)*&SIGMA3.)*EXP(-(((OBS-&MU3.)/&SIGMA3.)**2)/2);
  P1=TEMP1/(TEMP1+TEMP2+TEMP3);
  P2=TEMP2/(TEMP1+TEMP2+TEMP3);
  P3=TEMP3/(TEMP1+TEMP2+TEMP3);
  S1=(P1-B12X-(B13X-B12X)*(P3-0)/(B13Z-0))/(P1-1-(B13X-B12X)*(P3-0)/(B13Z-0));
  X1=P1+(1-P1)*S1; Y1=P3+(0-P3)*S1;
  DIST1=SQRT((P1-1)**2+(P3-0)**2); TEST1=SQRT((X1-1)**2+(Y1-0)**2);
  IF DIST1<TEST1 THEN DO;
   FLAG=1; CHOICE=1;
  END; /* IF INSIDE CRITICAL REGION 1 */
  S2=(P2-B23Y-(B21Y-B23Y)*(P1-0)/(B21X-0))/(P2-1-(B21Y-B23Y)*(P1-0)/(B21X-0));
  X2=P2+(1-P2)*S2; Y2=P1+(0-P1)*S2;
  DIST2=SQRT((P2-1)**2+(P1-0)**2); TEST2=SQRT((X2-1)**2+(Y2-0)**2);
  IF DIST2<TEST2 THEN DO;
   FLAG=1; CHOICE=2;
  END; /* IF INSIDE CRITICAL REGION 2 */
  S3=(P3-B31Z-(B32Z-B31Z)*(P2-0)/(B32Y-0))/(P3-1-(B32Z-B31Z)*(P2-0)/(B32Y-0));
  X3=P3+(1-P3)*S3; Y3=P2+(0-P2)*S3;
  DIST3=SQRT((P3-1)**2+(P2-0)**2); TEST3=SQRT((X3-1)**2+(Y3-0)**2);
  IF DIST3<TEST3 THEN DO;
```

```
      FLAG=1; CHOICE=3;
     END; /* IF INSIDE CRITICAL REGION 3 */
    END; /* DO WHILE DECISION HAS NOT BEEN MADE */
    OUTPUT;
  %END; /* DO LOOP */
 RUN;

 %IF &I.=1 %THEN %DO;
   DATA THESIS.SUMM3;
    SET THESIS.SIM1;
 %END; /* CREATE DATA SET IF FIRST SET OF ITERATIONS */
 %ELSE %DO;
  DATA THESIS.SUMM3;
   SET THESIS.SUMM3 THESIS.SIM1;
 %END; /*ELSE APPEND TO PREVIOUS SET OF ITERATIONS */
%END; /* DO LOOP OF ITERATIONS */

%MEND; /* END OF MACRO PROCEDURE TRIALS */

%TRIALS(&CNT1,&CNT2);
RUN;

PROC FREQ DATA=THESIS.SUMM3;
 TABLES H*CHOICE / NOCOL NOPERCENT;
RUN; /* FREQUENCY COUNT OF TRUE HYPOTHESIS, AND DECISION REACHED */

PROC SORT DATA=THESIS.SUMM3;
 BY H CHOICE;

PROC MEANS DATA=THESIS.SUMM3;
 BY H CHOICE;
 VAR N;
RUN; /*SAMPLE SIZE COUNTS */

PROC UNIVARIATE DATA=THESIS.SUMM3;
 VAR N;
 OUTPUT OUT=OUT1 PCTLPTS=50 TO 100 PCTLPRE=PERC;
RUN; /* DISTRIBUTION OF SAMPLE SIZES */
```

```
PROC PRINT DATA=OUT1;
RUN;
```