A THEORY OF SIGNIFICANCE TESTING

By

RICHARD DANIEL FINLEY

Bachelor of Science
Northeastern State College
Tahlequah, Oklahoma
1958

Master of Science
Colorado State University
Fort Collins, Colorado
1960

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
July, 1968

A THEORY OF SIGNIFICANCE

TESTING

Thesis Approved:

Thesis Adviser

Dean of the Graduate College

ACKNOWLEDGEMENT

I wish to express my utmost appreciation and thanks to Dr. J. Leroy Folks for serving as chairman of my graduate committee, for suggesting this thesis topic and for the guidance and valuable assistance he lent in directing the preparation of this thesis.

I also express gratitude to Dr. Carl E. Marshall, Director, Statistical Laboratory, for serving on my advisory committee and for his counseling and guidance during my stay at Oklahoma State University; and to Dr. David L. Weeks, Dr. David E. Bee, and Dr. Roy Gladstone for serving on my advisory committee.

Finally, I would like to express further appreciation to Professors Folks, Marshall, Weeks, and Bee for the educational opportunities which have been afforded to me through their superior professional capabilities as teachers.

Acknowledgement is also due the National Aeronautics and Space Administration for providing financial assistance for this thesis under grant number NGR 37-002-031.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

The name "test of significance" appears to have been first introduced by Fisher (5) in 1925 to describe a procedure for assessing the conformity or "goodness of fit" of a set of observations to a "null hypothesis", that is, the significance of an apparent discrepency between the observations and the hypothesis. According to Anscombe (1), the first clear use of such a procedure was as early as 1735 when Bernoulli (2) considered the closeness of the orbital planes of the planets to one another and to the equatorial plane of the Sun. Anscombe further claims that the first clear proposal for the general use of such a procedure in a broad class of situations was by Karl Pearson (12) in 1900, culminating in his χ^2 goodness of fit tests, but that the concept of significance tests was first considered in a general way by Fisher. In these and other early developments of the subject, such as Student's t-test (14), statistical tests do not appear to have been regarded as strict formal decision rules, and in fact, Fisher argued strongly against regarding a test of significance as a formal decision rule.

In the basic paper of Neyman and Pearson (11) in 1933, they derive a general theory for finding "most efficient tests of statistical hypotheses" and this theory is based on the development of decision rules for accepting or rejecting the hypothesis in question. In describing these test procedures they use the term "test of significance" in a decision making context. In a later paper by Fisher (6), he claims that Neyman had no concept of a test of significance simply as a means of learning but conceived of it only in the form of an acceptance procedure. Thus, when Neyman and Pearson thought they were correcting and improving his earlier works on tests of significance, they were in fact reinterpreting them in terms of that "technological and commercial apparatus known as an acceptance procedure", i.e., formal decision rule.

However, in spite of Fisher, since Neyman and Pearson published their paper in 1933, the concept of a statistical test as a decision making device has gained almost universal acceptance in statistical textbooks. It is generally presented as a formal decision rule for deciding which of two hypotheses to accept. The development of statistical decision theory in recent years has emphasized a more general approach, but it has not affected the presentation of statistical tests as decision making devices.

Taking Fisher's point of view, a distinction will be made between a test of hypothesis and a test of significance. A test of hypothesis is a formal decision rule in which one is committed to taking an action, that being either accepting or rejecting the null hypothesis. A test of significance simply consists of calculating the tail probabilities of a test statistic, called the test criterion, under a given null hypothesis. This tail probability is called the significance level and is such that the smaller the significance level, the more inconsistent with the data is the null hypothesis judged to be. The significance test consists only of evaluating the significance level and the experimenter is in no way forced to take any action for or against the null hypothesis.

Although significance testing is seldom described in statistical textbooks, a casual survey of some research journals indicates that it is an extremely common and apparently useful procedure. However, the accompanying description of such tests is nearly always in terms of the decision making ideas of hypothesis testing. Applied statisticians therefore seem to be in the unfortunate position of giving in to decision making ideas and concepts while actually subscribing to practice that which might better be described and recognized as data analysis and description techniques, such as significance testing.

Since significance testing seems to be widely practiced it seems that a theory of significance testing which is independent of decision rule techniques is needed. The immediate concern in developing such a theory is that of choosing a test criterion or test statistic to use in performing the significance test. Dempster and Schatzoff (3) state that perhaps the reason Neyman and Pearson formalized statistical tests into decision rules for accepting or rejecting the null hypothesis was in order to compare different tests. That is, in some way compare tests so as to arrive at a "best" test for a particular hypothesis problem. As is well known, the Neyman-Pearson theory compares tests by looking at operating characteristics of these decision rules under alternative hypotheses, in particular by looking at the power function.

It seems appropriate at this time to give a more formal definition of a test of significance as viewed in this presentation, to compare the significance test with the classical Neyman-Pearson theory involving size and power of a test, and to discuss the possibilities of choosing a test statistic for a test of significance.

Description of a Test of Significance

Let X denote a random variable, either vector or scalar, and assume that X has a probability density function $f_\theta(x)$, or cumulative distribution function, hereafter abbreviated c.d.f., given by $F_\theta(x)$, where θ is a parameter, vector or scalar, belonging to some parameter space Ω. Let the null hypothesis in question be given by

$$H_o: \theta \in \textcircled{H}_o \quad \text{where} \quad \textcircled{H}_o \subset \Omega .$$

Thus, the hypothesis may be either simple or composite. Let $T(X)$ denote a test statistic calculated from X with c.d.f. $G_\theta(t)$. It will be required that $G_\theta(t)$ be completely specified when $\theta \in \textcircled{H}_o$.

Anscombe (1) suggests some alternatives to the above requirement. He suggests finding the distribution of T conditional on the parameters, i.e., $\theta \in \textcircled{H}_o$, and then introducing a prior distribution for the parameters involved. The resulting significance test would then relate jointly to H_o and to this prior distribution. He also suggests that some kind of bounds for the aggregate of conditional distributions of T, again conditional on the parameters involved, might be used, rather than the conditional distribution itself.

The approach taken here means that the null hypothesis must be a simple hypothesis as far as T is concerned. For $\theta \in \textcircled{H}_o$ then, denote $G_\theta(t)$ by $G_o(t)$. Suppose further that T is chosen in such a way that small values are inconsistent with the null hypothesis. Then the significance level associated with T, denoted by $SL(T)$, is defined by

$$SL(T) = G_o(T) .$$

An observed value for the significance level, say α , is computed by

$$\alpha = G_o(t) .$$

A test of significance of the given hypothesis problem then consists of observing a value of T , say t , and computing $\alpha = G_o(t)$.

Now, the significance level, $SL(T)$, is a random variable and hence has a c.d.f. which will be denoted by $H_\theta(\alpha)$. Thus,

$$H_\theta(\alpha) = P_\theta\left[G_o(T) \leq \alpha\right] = P_\theta\left[SL(T) \leq \alpha\right] .$$

A further discussion of the above notation seems appropriate at this time. If two different statistics, say T and S , or $T^{(1)}$ and $T^{(2)}$, are under consideration, then the respective c.d.f.s will be denoted by

$$H_\theta^T(\alpha) = P_\theta\left[SL(T) \leq \alpha\right] \text{ and } H_\theta^S(\alpha) = P_\theta\left[SL(S) \leq \alpha\right] ,$$

or

$$H_\theta^{(1)}(\alpha) = P_\theta\left[SL(T^{(1)}) \leq \alpha\right] \text{ and } H_\theta^{(2)}(\alpha) = P_\theta\left[SL(T^{(2)}) \leq \alpha\right] .$$

Whenever no confusion arises as to what statistic is involved, no superscript will be used. Similarly, for specific values of θ , say θ_i , $H_{\theta_i}(\alpha)$ may be denoted by $H_i(\alpha)$. Also, for $\theta \in \text{Ⓗ}_o$, $H_\theta(\alpha)$ will be denoted by $H_o(\alpha)$.

If T is a continuous random variable, then under the null hypothesis,

$$H_o(\alpha) = \alpha \text{ for } 0 \leq \alpha \leq 1 .$$

That is, SL(T) is a uniform random variable distributed between zero and one. More generally,

$$H_o(\alpha) = P_o\left[SL(T) \leq \alpha\right] \leq \alpha \quad \text{for} \quad 0 \leq \alpha \leq 1 .$$

The question of whether or not it is necessary to specify the alternative hypothesis for a test of significance has been studied (1). Even though the question seems to remain open, it is a fact that tests are in use where the alternative hypothesis is not clearly specified, prime examples being goodness of fit tests. However, it is also true that in certain situations specification of the alternative hypothesis seems like an essential ingredient. For example, it seems necessary to specify the alternative hypothesis in order to choose between the one tailed and two tailed Student's t-test.

Here the point of view that an alternative hypothesis is necessary will be adopted and it will be given by

$$H_A: \theta \in \Theta_A \quad \text{where} \quad H_A \subset \Omega ,$$

with $\Theta_o \cap \Theta_A = \phi$ and $\Theta_o \cup \Theta_A \subset \Omega$. Thus, the alternative hypothesis may be either simple or composite.

A Correspondence Between Significance Testing and Hypothesis Testing

Suppose one has given a test of hypothesis of a given size α for the hypothesis problem

$$H_o: \theta \in \Theta_o \quad \text{versus} \quad H_A: \theta \in \Theta_A .$$

Let X denote the observation, either vector or scalar. The test of hypothesis will be assumed to be non-randomized. Thus, there exists a size α critical region R such that H_0 is rejected if $X \in R$ and H_0 is accepted if $X \notin R$.

For the above hypothesis then, a test of significance could be constructed by defining the significance level for an observed X to be

$$SL(X) = \alpha \text{ , if } X \in R \text{ ,}$$
$$= 1 \text{ , if } X \notin R \text{ .}$$

Next, suppose there is a whole class, say A , of sizes available, so that the given test of hypothesis could be performed at any size α for which $\alpha \in A$. Then, corresponding to A , there exists a class of critical regions $\mathcal{A} = \left\{R_\alpha \mid \alpha \in A\right\}$. That is, each α determines uniquely a critical region R_α. For this situation there appears to be a number of ways in which one could define a test of significance. The most appropriate seems to be as follows: For observed X , define SL(X) to be the smallest value α such that X belongs to every critical region of size greater than or equal to α, that is,

$$SL(X) = \text{Min}\left\{\alpha \mid X \in R_{\alpha^*} \text{ for all } \alpha^* \geq \alpha, \alpha^* \text{ and } \alpha \in A\right\}.$$

Note that for the one and two tailed tests usually encountered in statistical methods the above value of SL(X) gives that value of α at which the observation, or data, "would have been significant". Another feature of this significance test is that the set of possible values for SL(X) is the same as the set A of admissable sizes.

Conversely, suppose a test of significance, with associated test statistic T , is given. If one were to adopt the decision rule to reject the null hypothesis if the observed significance level is less than or equal to some pre-specified α , then

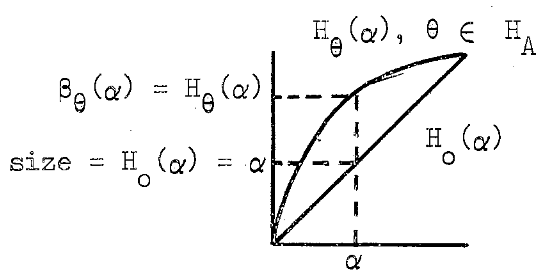
$$\Pr\left[\text{rejecting } H_o \mid H_o \text{ true}\right] = P_o\left[SL(T) \leq \alpha\right] = H_o(\alpha) ,$$

and since $H_o(\alpha) \leq \alpha$, equality if T is continuous, this constitutes a test of hypothesis of size α .

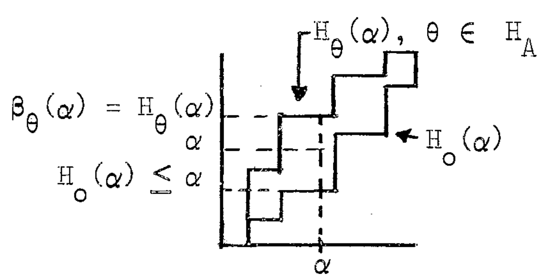
With this association between a decision rule and a significance test, an interpretation of the power of a test of hypothesis in the test of significance context can be obtained. Thus, the power is given by

$$\begin{aligned} \beta_\theta(\alpha) &= \Pr\left[\text{rejecting } H_o \mid H_o \text{ false}\right] \\ &= P_\theta\left[\text{rejecting } H_o\right] , \quad \theta \in \Theta_A , \\ &= P_\theta\left[SL(T) \leq \alpha\right] , \quad \theta \in \Theta_A , \\ &= H_\theta(\alpha) , \quad \theta \in \Theta_A . \end{aligned}$$

In thinking about the situation, the following sketches may be helpful.



Continuous Case



Discrete Case

In hypothesis testing attention is focused on a specific value of the significance level, say α, and $H_\theta(\alpha)$ is considered as θ varies over Θ_A. In significance testing $H_\theta(\alpha)$ must be considered as a function of both θ and α simultaneously. That is, the family of c.d.f.s $H_\theta(\alpha)$, rather than the values of this family for a specified α, are of interest.

Comparison of Test Statistics

The discussion in the preceding section was given merely to establish a correspondence between significance tests and the classical theory of hypothesis testing, and in order to do so it was necessary to regard a significance test as a decision making device. It should be emphasized again that a test of significance is not to be viewed as a formal decision rule but more as a means of data analysis and data interpretation. The significance test consists solely of evaluating the significance level corresponding to the observed value of the test statistic used with no commitment as to what use, if any, is to be made of the computed significance level.

In performing a significance test one may use any test criterion he chooses, that is, any statistic could be regarded as a test statistic for the given hypothesis problem. A particular statistic might be chosen over another because it seems to measure some characteristic of the observations which is of interest or perhaps there is simply some intuitive basis which suggests its use. Of immediate concern then, is the problem of providing a more precise method for comparing test statistics, even though no predictable use may be made of the resulting significance level. Thus, exactly what test statistic should be used in order to evaluate the significance level?

Several criteria are available to help answer the above question. In simple situations it may be possible simply to graph the $H_\theta(\alpha)$ curves for various statistics and compare these curves as θ and α vary over Θ_A and the unit interval respectively. Dempster and Schatzoff (3) investigate the properties of the single criterion "expected significance level", defined by

$$\mathrm{ESL} = \int_0^1 \alpha \, dH(\alpha) \ ,$$

as a basis for comparison of test statistics. They also suggest that a theory of tests might be built on the basis of comparing $H_\theta(\alpha)$ curves which parallels the Neyman-Pearson theory of hypothesis testing. Rigorous developments of the Neyman-Pearson theory of hypothesis testing may be found in Fraser (9) or Lehmann (10) .

Statement of the Problem

The main purpose of this investigation is to study the problem of, and difficulties which arise in, providing a rationale for significance testing which avoids commitment to any decision rules. The approach taken will be that suggested by Dempster and Schatzoff, namely, to build a theory which parallels the Neyman-Pearson theory of hypothesis testing.

In Chapter II a criterion for comparing test statistics is developed and the problem of evaluating the significance level for a problem involving a simple null hypothesis and a simple alternative hypothesis is discussed. In later chapters the scope of the study is extended to include different types of composite alternatives. Most of these extensions are restricted to one-parameter families of distributions so that much is still left to be desired as far as providing a complete rationale for significance testing is concerned.

There appear to be a number of tools available in the Neyman-Pearson hypothesis testing theory that help the theory flow smoothly, for example, the randomized decision rule and the trivial test function, which do not seem to have analogs in the test of significance context. It may certainly be that Neyman and Pearson realized these limitations and thus purposely "mis-interpreted" Fisher's concept of a significance test.

CHAPTER II

COMPARABLE STATISTICS AND SIMPLE HYPOTHESES

Criteria for Comparing Test Statistics

Consider the comparison of two test statistics $T^{(1)}$ and $T^{(2)}$ in the case of a simple alternative hypothesis $H_A: \theta = \theta_1$. It might seem that $T^{(1)}$ should be deemed preferable to $T^{(2)}$ if $H_1^{(1)}(\alpha) \geq H_1^{(2)}(\alpha)$ for all α with strict inequality for at least one α. This seems reasonable if both statistics are continuous random variables, for they would both admit the same set of possible values for the significance level, namely, $0 \leq \alpha \leq 1$. But if either or both of the statistics are discrete it is not obvious that this procedure is optimal.

As an example, consider two discrete statistics $T^{(1)}$ and $T^{(2)}$ whose corresponding $H_1^{(1)}(\alpha)$ and $H_1^{(2)}(\alpha)$ graphs are in Figure 1.

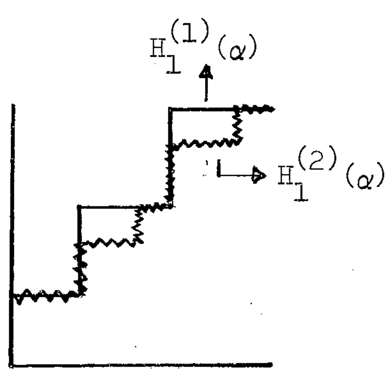


Figure 1

It does not seem clear that $T^{(1)}$ should be judged preferable to $T^{(2)}$ even though $H_1^{(1)}(\alpha) \geq H_1^{(2)}(\alpha)$ for all α with strict inequality for at least one α. The statistic $T^{(2)}$ might be more useful to the experimenter because more distinct significance levels are actually achievable.

Consider as a second illustration statistics $T^{(1)}$ and $T^{(2)}$ with $H_1^{(1)}(\alpha)$ and $H_1^{(2)}(\alpha)$ curves as given in Figure 2.

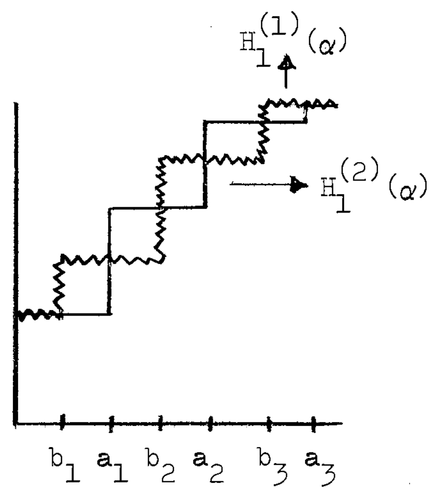


Figure 2

Note that $T^{(1)}$ admits possible significance levels b_1, b_2, and b_3 while $T^{(2)}$ admits significance levels a_1, a_2, and a_3. Both statistics give rise to the same number of achievable significance levels but have no achievable significance levels in common. Should $T^{(1)}$ be regarded as better than $T^{(2)}$ since $H_1^{(1)}(\alpha) \geq H_1^{(2)}(\alpha)$ for all achievable significance levels of $T^{(1)}$, or should $T^{(2)}$ be regarded as better since $H_1^{(2)}(\alpha) \geq H_1^{(1)}(\alpha)$ at all achievable significance levels of $T^{(2)}$?

In order to avert situations such as those indicated in the previous illustrations it appears that the comparison of two statistics must be restricted to those which admit the same set of achievable significance levels.

<u>Definition 2.1</u> Two statistics $T^{(1)}$ and $T^{(2)}$ are said to be comparable if and only if they have the same set of achievable significance levels.

As mentioned earlier, no difficulty arises in the comparison of two statistics if they are both continuous. That is, by the definition, any two continuous statistics are comparable. As a further consequence of the above definition, if $T^{(1)}$ and $T^{(2)}$ are comparable, then

$$H_o^{(1)}(\alpha) = H_o^{(2)}(\alpha) \tag{2.1}$$

for all α , $0 \leq \alpha \leq 1$. Moreover, each member of Equation (2.1) is equal to α for all achievable α .

<u>Definition 2.2</u> If $T^{(1)}$ and $T^{(2)}$ are comparable statistics, then $T^{(1)}$ is said to be more sensitive than $T^{(2)}$ if

$$H_1^{(1)}(\alpha) \geq H_1^{(2)}(\alpha) \tag{2.2}$$

for all α , with strict inequality holding for at least one α .

<u>Definition 2.3</u> If a statistic T is a most sensitive test statistic for all $\theta \in \textcircled{H}_A$, then T is said to be a uniformly most sensitive test statistic.

In choosing a statistic for evaluating the significance level for a problem involving a simple alternative, say H_A: $\theta = \theta_1$, one would thus hope to find a statistic which is a most sensitive test statistic. If the alternative is composite, say $H_A: \theta \in \Theta_A$, then a uniformly most sensitive test statistic would be desirable, although such may not always exist, as will be seen later.

The preceding definitions involving most sensitive test statistics do not assure a unique most sensitive test statistic for a given problem. In particular, for discrete statistics the property of being most sensitive applies only to a class of comparable statistics. In this sense there may be many most sensitive test statistics which are simply not comparable. For the continuous case however, since all continuous statistics are comparable, a most sensitive test statistic is unique in that any other statistic which is also most sensitive must actually be equivalent.

Simple Hypotheses

As mentioned by Dempster and Schatzoff (3), it is an immediate consequence of the Neyman-Pearson fundamental lemma of hypothesis testing applied to the whole range of sizes α , $0 \leq \alpha \leq 1$, that the likelihood ratio statistic is more sensitive than any other statistic, in the sense of Definition 2.2, for testing a simple null hypothesis against a simple alternative hypothesis. However, this requires one to interpret a significance test as a decision rule. It seems desirable then, to state a theorem and give the proof in the present context of significance testing.

Theorem 2.1 Neyman-Pearson Lemma for Significance Testing: The likelihood ratio statistic is a most sensitive test statistic for evaluating the significance level for a simple null hypothesis versus a simple alternative hypothesis.

<u>Proof</u>: Let the likelihood ratio statistic be given by

$$LR = T^{(1)}(x) = f_o(x)/f_1(x)$$

where f_o and f_1 are densities, $f_o \not\equiv f_1$, with respect to a measure μ , under the null hypothesis and alternative hypothesis respectively. Let $T^{(2)}(x)$ be any other statistic comparable with LR , i.e., such that

$$H_o^{(1)}(\alpha) = H_o^{(2)}(\alpha)$$

for all α . To prove the theorem it is necessary to show

$$H_1^{(1)}(\alpha) \geq H_1^{(2)}(\alpha) ,$$

or

$$P_1\left[SL(T^{(1)}) \leq \alpha\right] \geq P_1\left[SL(T^{(2)}) \leq \alpha\right] ,$$

or equivalently

$$P_1\left[G_o(T^{(1)}) \leq \alpha\right] \geq P_1\left[F_o(T^{(2)}) \leq \alpha\right] \qquad (2.3)$$

for all α , where G_o and F_o are the corresponding c.d.f.s of $T^{(1)}$ and $T^{(2)}$ under the null hypothesis. If $T^{(1)}$ and $T^{(2)}$ are continuous, G_o and F_o have inverses for all α . If they are discrete, define G_o and F_o only at the points of positive probability, that is, those points that admit achievable α values. Thus, G_o^{-1} and F_o^{-1} exist and (2.3) is equivalent to

$$P_1\left[T^{(1)}(x) \leq G_o^{-1}(\alpha)\right] \geq P_1\left[T^{(2)}(x) \leq F_o^{-1}(\alpha)\right] ,$$

or

$$P_1\left[f_o(x)/f_1(x) \le G_o^{-1}(\alpha)\right] \ge P_1\left[T^{(2)}(x) \le F_o^{-1}(\alpha)\right] .$$

Let

$$S_1 = \left\{x:\ f_o(x)/f_1(x) \le G_o^{-1}(\alpha)\right\}$$

and

$$S_2 = \left\{x:\ T^{(2)}(x) \le F_o^{-1}(\alpha)\right\} .$$

Then

$$P_1\left[G_o(T^{(1)}) \le \alpha\right] - P_1\left[F_o(T^{(2)}) \le \alpha\right]$$

$$= \int_{S_1} f_1(x)\ d\mu(x) - \int_{S_2} f_1(x)\ d\mu(x)$$

$$= \int_{S_1-S_2} f_1(x)\ d\mu(x) - \int_{S_2-S_1} f_1(x)\ d\mu(x)$$

$$\ge \frac{1}{G_o^{-1}(\alpha)} \left\{\int_{S_1-S_2} f_o(x)\ d\mu(x) - \int_{S_2-S_1} f_o(x)\ d\mu(x)\right\}$$

$$= \frac{1}{G_o^{-1}(\alpha)} \left\{\int_{S_1} f_o(x)\ d\mu(x) - \int_{S_2} f_o(x)\ d\mu(x)\right\}$$

$$= \frac{1}{G_o^{-1}(\alpha)} \left\{H_o^{(1)}(\alpha) - H_o^{(2)}(\alpha)\right\}$$

$$= 0 .$$

Thus,

$$H_1^{(1)}(\alpha) \ge H_1^{(2)}(\alpha) \qquad (2.4)$$

for all achievable α. However, if both statistics are discrete, $H_1^{(1)}(\alpha)$ and $H_1^{(2)}(\alpha)$ are step functions so that (2.4) holding for all achievable α would imply it holds for all α, $0 \le \alpha \le 1$.

The Neyman-Pearson lemma for significance testing gives a sufficient condition for a most sensitive test statistic. The question naturally arises as to whether or not it also gives a necessary condition for a most sensitive test statistic. That is, given a statistic T which is known to be a most sensitive test statistic, is this statistic equivalent to the likelihood ratio statistic? The answer, of course, is negative. The statistic T could be most sensitive among a class of statistics which does not contain the likelihood ratio statistic, i.e., the statistics in this class do not have the same achievable significance levels as the likelihood ratio statistic. The likelihood ratio statistic would be the "unique" most sensitive test statistic within the class of all statistics which are comparable to it in the sense that it is at least as sensitive as any other statistic in this class.

In hypothesis testing an immediate result of the Neyman-Pearson lemma is that the power β of the most powerful level α test for a simple null hypothesis versus a simple alternative hypothesis is such that $\beta > \alpha$. An analogous result for significance testing is given in the following corollary.

Corollary 2.1 For a simple null hypothesis versus a simple alternative hypothesis the likelihood ratio statistic is such that

$$H_1(\alpha) > H_0(\alpha)$$

for all α, $0 < \alpha < 1$, and

$$H_1(\alpha) > \alpha$$

for all achievable α.

<u>Proof</u>: Let the likelihood ratio statistic be given by

$$LR = T(x) = f_o(x)/f_1(x) ,$$

where f_o and f_1 are densities, $f_o \not\equiv f_1$, with respect to a measure μ , under the null hypothesis and alternative hypothesis respectively. Let G_o denote the c.d.f. of T under the null hypothesis. It is necessary then to prove that

$$P_1\left[SL(T) \leq \alpha\right] > P_o\left[SL(T) \leq \alpha\right] ,$$

for all α , $0 < \alpha < 1$, or equivalently

$$P_1\left[G_o(T) \leq \alpha\right] > P_o\left[G_o(T) \leq \alpha\right] , \tag{2.5}$$

for all α , $0 < \alpha < 1$. If T is discrete, define G_o only at the points of positive probability, that is, points which give achievable α values. Thus, G_o^{-1} exists, and the inequality in (2.5) is equivalent to

$$P_1\left[T(x) \leq G_o^{-1}(\alpha)\right] > P_o\left[T(x) \leq G_o^{-1}(\alpha)\right] ,$$

or

$$P_1\left[\frac{f_o(x)}{f_1(x)} \leq G_o^{-1}(\alpha)\right] > P_o\left[\frac{f_o(x)}{f_1(x)} \leq G_o^{-1}(\alpha)\right] .$$

Let

$$S = \left\{x \,\middle|\, \frac{f_o(x)}{f_1(x)} \leq G_o^{-1}(\alpha)\right\} .$$

If $0 < G_o^{-1}(\alpha) < 1$, then

$$\int_S \left[f_1(x) - f_o(x)\right] d\mu(x) \geq \int_S \left[\frac{f_o(x)}{G_o^{-1}(\alpha)} - f_o(x)\right] d\mu(x)$$

$$= \left[\frac{1}{G_o^{-1}(\alpha)} - 1\right] \int_S f_o(x)\, d\mu(x) > 0 \text{ .}$$

If $G_o^{-1}(\alpha) \geq 1$, then

$$\int_{\overline{S}} \left[f_1(x) - f_o(x)\right] d\mu(x) < \int_{\overline{S}} \left[f_1(x) - G_o^{-1}(\alpha)\, f_1(x)\right] d\mu(x)$$

$$= \left[1 - G_o^{-1}(\alpha)\right] \int_{\overline{S}} f_1(x)\, d\mu(x) \leq 0$$

implies

$$\int_{\overline{S}} \left[f_1(x) - f_o(x)\right] d\mu(x) < 0 \text{ .}$$

Thus, since the above integral plus the integral over S sums to zero,

$$\int_S \left[f_1(x) - f_o(x)\right] d\mu(x) > 0 \text{ .}$$

Therefore, for all $G_o^{-1}(\alpha)$, and since G_o was defined only at points which gave achievable values of α ,

$$H_1(\alpha) > H_o(\alpha) = \alpha \tag{2.6}$$

for all achievable α . But, if T is discrete $H_o(\alpha)$ and $H_1(\alpha)$ are step functions and (2.6) holding for all achievable α implies it holds for all α , $0 < \alpha < 1$. Hence, in general

$$H_1(\alpha) > H_o(\alpha) \text{ ,}$$

for all α , $0 < \alpha < 1$, and

$$H_1(\alpha) > \alpha$$

for all achievable α , as was to be proved.

When applying the Neyman-Pearson fundamental lemma for finding a most sensitive test statistic for evaluating the significance level, one might generally find it more convenient to use some function of LR rather than LR itself.

Theorem 2.2 Let T be a most sensitive test statistic for evaluating the significance level for $H_o: \theta \in \Theta_o$ versus $H_A: \theta \in \Theta_A$. Then any increasing function of T is also most sensitive for the given hypothesis problem.

Proof: Let T have c.d.f. $G_o(t)$ under H_o, and let $S = f(T)$ be any increasing function of T with c.d.f. under H_o of $F_o(s)$. Thus, $T = f^{-1}(S)$ implies $F_o(s) = G_o(f^{-1}(s)) = G_o(t)$ so that

$$\begin{aligned} H_\theta^T(\alpha) - H_\theta^S(\alpha) &= P_\theta\left[SL(T) \leq \alpha\right] - P_\theta\left[SL(S) \leq \alpha\right] \\ &= P_\theta\left[G_o(T) \leq \alpha\right] - P_\theta\left[G_o(T) \leq \alpha\right] \\ &= 0 . \end{aligned}$$

Therefore,

$$H_o^T(\alpha) = H_o^S(\alpha)$$

for all α and

$$H_\theta^S(\alpha) \geq H_\theta^T(\alpha)$$

for all α and $\theta \in \Theta_A$ implies that S is a most sensitive test statistic since T was given as most sensitive.

An immediate result of the theorem is the following corollary.

Corollary 2.2 For a simple null hypothesis versus a simple alternative hypothesis, any increasing function of the likelihood ratio statistic is also most sensitive.

To illustrate previous definitions and theorems, consider the example of five independent binomial trials with

$$H_o: p = .2 \quad \text{versus} \quad H_A: p = .4 .$$

Application of Theorem 2.1 and Corollary 2.2 yields

$$T^{(1)}(X) = -X$$

as a most sensitive test statistic. Thus, the significance level when x is observed is given by

$$G_o(-x) = P_o\left[-X \leq -x\right] = P_o\left[X \geq X\right] .$$

A calculation of the significance level for all possible values of x along with the corresponding $H_1^{(1)}(\alpha)$ values is given in Table 1 .

Table 1

x	f(x;p=.2)	f(x;p=.4)	α	$H_1^{(1)}(\alpha)$
5	.00032	.01024	a_1 = .00032	.01024
4	.00640	.07680	a_2 = .00672	.08704
3	.05120	.23040	a_3 = .05792	.31744
2	.20480	.34560	a_4 = .26272	.66304
1	.40960	.25920	a_5 = .67232	.92224
0	.32768	.07776	a_6 = 1.0000	1.00000

The lack of uniqueness of a most sensitive test statistic was discussed earlier. To emphasize this point consider a second test statistic, say

$$T^{(2)}(x) = |x - 3| .$$

Calculations of the significance level and the corresponding values of $H_1^{(2)}(\alpha)$ are given in Table 2.

Table 2

$t = \|x - 3\|$	$f(t;p=.2)$	$f(t;p=.4)$	α	$H_1^{(2)}(\alpha)$
0	.05120	.23040	b_1 = .05120	.23040
1	.21120	.42240	b_2 = .26240	.65280
2	.40992	.26944	b_3 = .67232	.92224
3	.32768	.07776	b_4 = 1.0000	1.00000

A graphical comparison of $H_1^{(1)}(\alpha)$ and $H_1^{(2)}(\alpha)$ is represented in Figure 3.

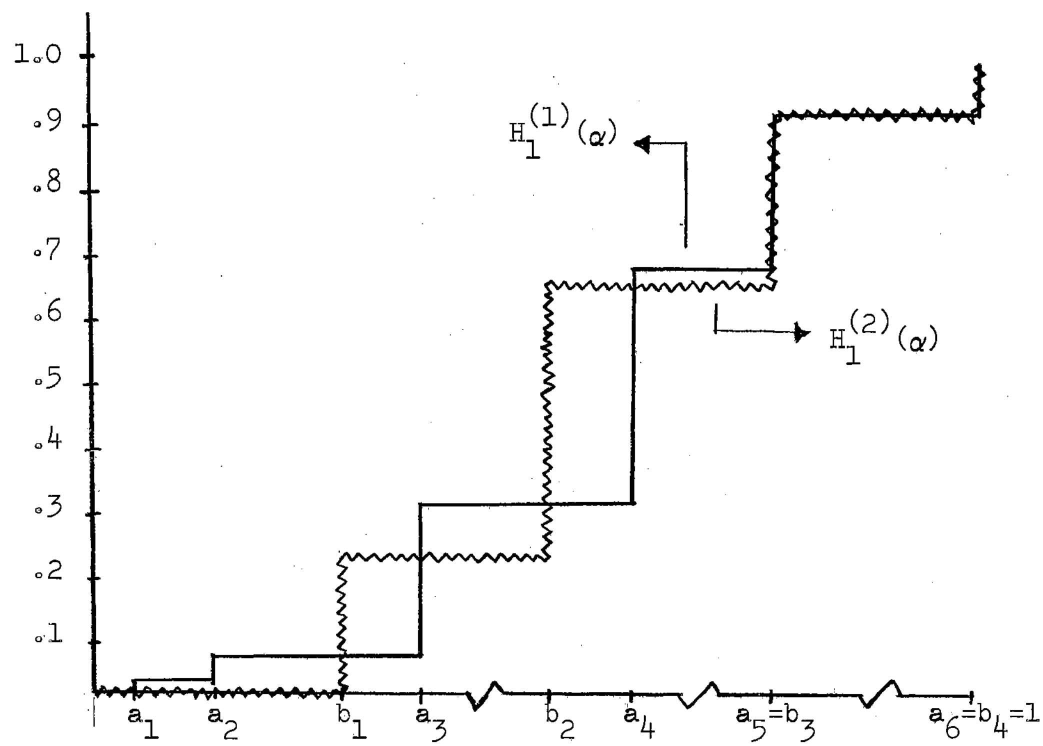


Figure 3

Although $T^{(1)}$ is the likelihood statistic and as such is a most sensitive test statistic, it is not evident that it is better than $T^{(2)}$. That is, $T^{(1)}$ and $T^{(2)}$ are not comparable. Certainly $H_1^{(1)}(\alpha)$ is not greater than or equal to $H_1^{(2)}(\alpha)$ for all α, for example, $.0512 < \alpha < .05792$ and $.2624 < \alpha < .26272$. It is true however, that $H_1^{(1)}(\alpha) > H_1^{(2)}(\alpha)$ for all α which are achievable values of $SL(T^{(1)})$.

Another difficulty here seems to be that the class of statistics comparable to the likelihood ratio contains only the likelihood ratio itself, or equivalent statistics, and in this respect there may be some apprehension about using the likelihood ratio statistic as the optimum test criterion for evaluating the significance level. One advantage of the likelihood ratio statistic in this example is that no other statistic yields more possible achievable significance levels.

The likelihood ratio statistic does possess some properties which in a sense make it more "admissable" as a test criterion than any other statistic, even though the statistics may not be comparable. Consider, for example, the value $\alpha = b_1$ for $T^{(2)}$. Now if it happened that $H_1^{(2)}(b_1) > H_1^{(1)}(a_3)$ this would certainly present some doubt as to the validity of the likelihood ratio as a best statistic. However, note that for all $b_k < a_j$, $H_1^{(2)}(b_k) \leq H_1^{(1)}(a_j)$. This result, which was first brought to the attention of this writer by Oscar Kempthorne, is presented in the following theorem.

<u>Theorem 2.3</u> Suppose $T = f_o(x)/f_1(x)$ is discrete with possible significance levels $a_1 < a_2 < \ldots$ and that S is any other discrete

statistic with significance levels $b_1 < b_2 < \dots$. Then for no $b_k < a_j$ is $H_1^S(b_k) > H_1^T(a_j)$.

<u>Proof</u>: Let t_j denote that value of T which gives rise to a_j and s_k that value of S giving rise to b_k and G_o and F_o the corresponding c.d.f.s of T and S . Thus,

$$H_1^T(a_j) - H_1^S(b_k) = P_1\left[G_o(T) \leq a_j\right] - P_1\left[F_o(S) \leq b_k\right]$$

$$= P_1\left[T \leq G_o^{-1}(a_j)\right] - P_1\left[S \leq F_o^{-1}(b_k)\right]$$

$$= P_1\left[T \leq t_j\right] - P_1\left[S \leq s_k\right] .$$

Let $S_1 = \left\{x \mid T \leq t_j\right\}$ and $S_2 = \left\{x \mid S \leq s_k\right\}$.

Then

$$P_1\left[T \leq t_j\right] - P_1\left[S \leq s_k\right] = \sum_{S_1} f_1(x) - \sum_{S_2} f_1(x)$$

$$\geq \frac{1}{t_j}\left[\sum_{S_1} f_o(x) - \sum_{S_2} f_o(x)\right]$$

$$= \frac{1}{t_j}\left[H_o^T(a_j) - H_o^S(b_k)\right]$$

$$= \frac{1}{t_j}\left[a_j - b_k\right] .$$

Thus for $b_k < a_j$, $H_1^T(a_j) \geq H_1^S(b_k)$ and the theorem is proved.

Hence, even though the two statistics may not be comparable, it seems like the likelihood ratio statistic is more admissable in that even though S might admit to smaller significance levels the corresponding $H_1^S(\alpha)$ value can not be as large as that associated with any larger significance level of the likelihood ratio. If there are common achievable significance levels for the two statistics then the likelihood ratio gives values of $H_1(\alpha)$ which are at least as large as the $H_1(\alpha)$ values for the statistic S .

CHAPTER III

ONE SIDED ALTERNATIVES

Introduction

The case where both the hypothesis and the class of alternatives are simple is mainly of theoretical interest since problems arising in the applications typically involve a parametric family of distributions depending on one or more continuous parameters. In this chapter, consideration is restricted to the situation where the distributions involved depend only on a single real valued parameter θ .

Although, for the one parameter case, one is restricted in significance testing to a simple null hypothesis he may wish to consider a composite alternative hypothesis, say $H_A: \theta \in Ⓗ_A$. Here, attention will be confined to one sided alternatives, and hence, hypotheses of the form

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta > \theta_o ,$$

or

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta < \theta_o .$$

A uniformly most sensitive test statistic will not always exist for the above hypothesis problems. For example, if the alternative $\theta > \theta_o$ is considered, one will generally be unable to find a statistic which is independent of $\theta > \theta_o$, in which case one would not really have a statistic. A uniformly most sensitive test statistic will exist if additional assumptions are made about the distributions involved.

Distributions with Monotone Likelihood Ratio

<u>Definition 3.1</u> A real parameter family of distributions is said to have strict monotone likelihood ratio if densities exist and if there exists a function $T(X)$ such that for $\theta_1 < \theta_2$, the ratio $f_{\theta_2}(X)/f_{\theta_1}(X)$ is an increasing function of $T(X)$ on the set of X for which the ratio exists.

<u>Theorem 3.1</u> Let θ be a real parameter and let the random variable X have probability density $f_\theta(X)$ with strict monotone likelihood ratio in $T(X)$. Then,

(i-a) For all θ_o, $-T(X)$ is a uniformly most sensitive test statistic for evaluating the significance level of the hypothesis problem

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta > \theta_o .$$

(i-b) $H_\theta^{-T}(\alpha)$ is strictly increasing in θ for all points θ for which $H_\theta^{-T}(\alpha) < 1$, for all α, $0 < \alpha < 1$.

(ii-a) For all θ_o, $T(X)$ is a uniformly most sensitive test statistic for evaluating the significance level of the hypothesis problem

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta < \theta_o .$$

(ii-b) $H_\theta^{T}(\alpha)$ is strictly decreasing in θ for all points θ for which $H_\theta^{T}(\alpha) < 1$, for all α, $0 < \alpha < 1$.

<u>Proof</u>: (i-a) Let θ_o be arbitrary and consider

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta = \theta_1$$

where θ_1 is arbitrary except that $\theta_1 > \theta_o$. Then by Theorem 2.1, $f_o(X)/f_1(X)$ is a most sensitive statistic for θ_o versus θ_1. But since $\theta_1 > \theta_o$, $f_1(X)/f_o(X)$ is increasing in $T(X)$ so that $f_o(X)/f_1(X)$ is increasing in $-T(X)$. Thus, by Corollary 2.2, $-T(X)$ is most sensitive. Furthermore, $-T(X)$ is independent of θ_1 (as long as $\theta_1 > \theta_o$), hence, $-T(X)$ is a uniformly most sensitive test statistic for evaluating the significance level of H_o versus H_A. Thus, for observed X, say x,

$$\begin{aligned} SL(-T(x)) &= P_o\left[-T(X) \leq -T(x)\right] \\ &= P_o\left[T(X) \geq T(x)\right]. \end{aligned}$$

(ii-b) Let θ_i and θ_j be any two values of θ with $\theta_i < \theta_j$. Consider

$$H_o: \theta = \theta_i \quad \text{versus} \quad H_A: \theta = \theta_j.$$

Then, by Corollary 2.1, $\alpha = H_{\theta_i}^{-T}(\alpha) < H_{\theta_j}^{-T}(\alpha)$ for all α. Thus, since θ_i and θ_j were arbitrary, $H_{\theta}^{-T}(\alpha)$ is strictly increasing in θ.

The proof of (ii) is omitted since it is analogous to that of (i).

An important class of families of distributions that satisfy the assumptions of Theorem 3.1 is the class of one-parameter exponential families.

Corollary 3.1 Let θ be a real parameter, and let X have probability density, w.r.t. some measure μ ,

$$f_\theta(X) = C(\theta)\, e^{Q(\theta)\, T(X)}\, h(X) ,$$

where Q is strictly monotone.

(i) Consider $H_o: \theta = \theta_o$ versus $H_A: \theta > \theta_o$;

a) If Q is increasing in θ , $-T(X)$ is uniformly most sensitive.

b) If Q is decreasing in θ , $T(X)$ is uniformly most sensitive.

(ii) Consider $H_o: \theta = \theta_o$ versus $H_A: \theta < \theta_o$;

a) If Q is increasing in θ , $T(X)$ is uniformly most sensitive.

b) If Q is decreasing in θ , $-T(X)$ is uniformly most sensitive.

Proof (i-a): Consider the ratio

$$\frac{f_o(X)}{f_\theta(X)} = \frac{C(\theta_o)}{C(\theta)}\, e^{T(X)\left[Q(\theta_o) - Q(\theta)\right]} .$$

If Q is increasing in θ , then $Q(\theta_o) - Q(\theta) < 0$, so that the ratio is strictly increasing in $-T(X)$. Applying Theorem 3.1(i-a) gives $-T(X)$ as uniformly most sensitive.

The rest of the Corollary is proved similarly by direct application of Theorem 3.1.

As examples, consider two of the well known hypothesis problems involving one parameter normal distributions.

Example 3.1: Consider a random sample of size n from $N(0,\sigma^2)$, and the hypothesis problem

$$H_o: \sigma^2 = \sigma_o^2 \text{ versus } H_A: \sigma^2 > \sigma_o^2 .$$

Here, $f_\sigma(X)/f_{\sigma_o}(X)$ is increasing in $T(X) = \sum_{i=1}^{n} X_i^2/\sigma_o^2$, for $\sigma^2 > \sigma_o^2$. Hence $-T(X)$ is uniformly most sensitive for evaluating the significance level and for observed $x = (x_1, \cdots, x_n)$ one obtains the observed significance level

$$SL(-T(x)) = P_o\left[\chi_n^2 \geq T(x)\right] .$$

Example 3.2: Consider a random sample of size n from $N(\mu,\sigma^2)$, σ^2 known, and the hypothesis problem

$$H_o: \mu = \mu_o \text{ versus } H_A: \mu < \mu_o .$$

Here $f_\mu(X)/f_{\mu_o}(X)$ is increasing in $\overline{X}$ for $\mu < \mu_o$ so that $\overline{X}$ is uniformly most sensitive, hence, $T(X) = \sqrt{n}(\overline{X} - \mu_o)/\sigma$ is uniformly most sensitive. Thus, for an observed X, say $x = (x_1, \cdots, x_n)$, the observed significance level is given by

$$SL(T(x)) = \Pr\left[N(0,1) \text{ variate} \leq T(x)\right] .$$

Locally Most Sensitive Statistics

Significance tests developed earlier for one sided alternatives involved a family of distributions where the distributions were sufficiently well behaved, i.e., a family of distributions which had strict monotone likelihood ratio. Suppose the random variable X has density $f_\theta(X)$, w.r.t. a measure μ , but does not have monotone likelihood ratio and consider

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta > \theta_o \ .$$

Let T^* be a statistic with distribution G_o under H_o and let $R = \left\{x: G_o(T^*) \leq \alpha\right\}$. Then

$$H_\theta^*(\alpha) = \int_R f_\theta(x)\, d\mu(x) \ .$$

Suppose $H_\theta^*(\alpha)$ possesses continuous derivatives w.r.t. θ and that the differentiation can be carried out under the integral so that

$$H_\theta^{*'}(\alpha) = \int_R f_\theta'(x)\, d\mu(x) \ ,$$

where primes indicate differentiation w.r.t. θ . If T^* were to be most sensitive it would be necessary that

$$H_o^{*'}(\alpha) > H_o^{S'}(\alpha)$$

for any other statistic which is comparable with T^* . However, the inequality would not imply T^* most sensitive, but it would imply T^* most sensitive in some one sided neighborhood of θ_o , say $[\theta_o \, , \ \theta_o + \epsilon)$. In this case T^* will be said to be locally most sensitive.

<u>Theorem 3.2</u> Let $T^*(X) = -f_o'(X)/f_o(X)$, and let S be any other statistic which is comparable with T^* . Let G_o and F_o denote the c.d.f.s under H_o for T^* and S respectively. Then, for all α , $0 \leq \alpha \leq 1$,

$$H_o^{*'}(\alpha) \geq H_o^{S'}(\alpha) \ .$$

Proof: Define G_o only at points of positive probability so that G_o^{-1} exists. Let

$$S_1 = \left\{X : G_o(T^*) \leq \alpha\right\} = \left\{X : T^* \leq G_o^{-1}(\alpha)\right\}$$

$$= \left\{X : f_o'(X) \geq - G_o^{-1}(\alpha)\ f_o(X)\right\}$$

and

$$S_2 = \left\{X : F_o(S) \leq \alpha\right\} .$$

Then

$$H_o^{*'}(\alpha) - H_o^{S'}(\alpha) = \int_{S_1} f_o'(x)\ d\mu(x) - \int_{S_2} f_o'(x)\ d\mu(x)$$

$$= \int_{S_1 - S_2} f_o'(x)\ d\mu(x) - \int_{S_2 - S_1} f_o'(x)\ d\mu(x)$$

$$\geq -G_o^{-1}(\alpha)\left[\int_{S_1 - S_2} f_o(x)\ d\mu(x) - \int_{S_2 - S_1} f_o(x)\ d\mu(x)\right]$$

$$= -G_o^{-1}(\alpha)\left[\int_{S_1} f_o(x)\ d\mu(x) - \int_{S_2} f_o(x)\ d\mu(x)\right]$$

$$= -G_o^{-1}(\alpha)\left[H_o^*(\alpha) - H_o^S(\alpha)\right]$$

$$= 0 .$$

Therefore $H_o^{*'}(\alpha) \geq H_o^{S'}(\alpha)$ for all α.

A similar theorem could be given for $H_o: \theta = \theta_o$ versus $H_A: \theta < \theta_o$ in which case $T^* = f_o'(X)/f_o(X)$ is such that $H_o^{*'}(\alpha) \leq H_o^{S'}(\alpha)$ for all α.

CHAPTER IV

TWO SIDED ALTERNATIVES

Introduction

When considering composite alternatives for significance tests attention has previously been restricted to one sided alternatives. In this chapter a one parameter family of distributions is considered where the hypothesis is of the form

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta \neq \theta_o \, .$$

Special attention will be given the one parameter exponential family. Since the distribution of the test statistic T is required to be completely determined under the null hypothesis other types of two sided problems such as

$$H_o: \theta \leq \theta_1 \text{ or } \theta \geq \theta_2 \text{, where } \theta_1 < \theta_2 \text{, versus } H_A: \theta_1 < \theta < \theta_2$$

and

$$H_o: \theta_1 \leq \theta \leq \theta_2 \quad \text{versus} \quad H_A: \theta < \theta_1 \text{ or } \theta > \theta_2$$

would have no place in the present context of significance testing.

The ultimate goal in determining a test statistic for evaluating the significance level for the hypothesis $\theta = \theta_o$ versus $\theta \neq \theta_o$ would be, of course, to find a statistic which is uniformly most sensitive. However, in general, such a statistic will not always exist. For example, let the observable random variable X have the density

$$f_\theta(X) = C(\theta)\, e^{\theta T(X)}\, h(X)$$

where X and $T(X)$ are continuous random variables. For the given hypothesis, suppose that T^* is a uniformly most sensitive test statistic. But also, from Corollary 3.1, for $\theta > \theta_o$ the statistic $-T$ is such that

$$H_\theta^{-T}(\alpha) \geq H_\theta^*(\alpha) .$$

Hence, $-T$ and T^* would have to be equivalent. Similarly for $\theta < \theta_o$, Corollary 3.1 would imply T and T^* equivalent, thus $-T$ and T equivalent, which leads to a contradiction that T^* is most sensitive.

So in determining what statistic to use for evaluating the significance level for a two sided hypothesis problem, since a uniformly most sensitive statistic will not exist, one may wish to find a statistic which is "best" among a smaller class of test statistics.

Unbiased Significance Tests

In hypothesis testing a simple condition that one sometimes imposed on a test of hypothesis of the form

$$H_o: \theta \in \Theta_o \quad \text{versus} \quad H_A: \theta \in \Theta_A$$

was that for no alternative in H_A was the probability of rejection less than the size of the test, that is, the power of the test was such that

$$\beta(\theta) \leq \alpha \ \text{ if } \ \theta \in \Theta_o \ \text{ and } \ \beta(\theta) \geq \alpha \ \text{ if } \ \theta \in \Theta_A .$$

A test satisfying this condition was called unbiased. Thus, for some hypothesis testing problems, although a uniformly most powerful test might not exist, one could possibly find a test which was uniformly most powerful among the smaller class of unbiased tests.

With the correspondence between hypothesis testing and significance testing established earlier the above definition would translate immediately to

$$H_\theta(\alpha) \leq \alpha \quad \text{for} \quad \theta \in \Theta_0 \quad \text{and} \quad H_\theta(\alpha) \geq \alpha \quad \text{for} \quad \theta \in \Theta_A \; .$$

As tests of significance have been formulated the first of these conditions is satisfied for all α, $0 \leq \alpha \leq 1$. In defining the concept of unbiasedness for significance testing it might be presumed that the second condition should hold for all α as well. But this condition is somewhat too strong, since if a statistic T is discrete $H_\theta^T(\alpha)$ is only meaningful at points α which are achievable values of $SL(T)$, i.e., points of increase of $H_0^T(\alpha)$. It would be possible to have $H_\theta^T(\alpha) \geq \alpha$, $\theta \in \Theta_A$, for all achievable α and still have $H_\theta^T(\alpha) < \alpha$, $\theta \in \Theta_A$, for some α which was not an achievable α.

<u>Definition 4.1</u> For the hypothesis problem $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_A$ the significance test corresponding to a statistic T is said to be unbiased if

$$H_\theta^T(\alpha) \geq \alpha \quad \text{for} \quad \theta \in \Theta_A$$

for all achievable α, or equivalently,

$$H_\theta^T(\alpha) \geq H_0^T(\alpha)$$

for all α, $0 \leq \alpha \leq 1$.

Throughout this investigation the statistic used for an unbiased significance test will be referred to as an unbiased statistic. This terminology should not be confused with the usual definition of an unbiased estimator.

Whenever a uniformly most sensitive test statistic exists for a given hypothesis problem then this statistic is unbiased. For the most sensitive test statistics derived in Corollary 3.1 this is easily seen by referring to parts (i-b) and (ii-b) of Theorem 3.1.

One Parameter Exponential Families

Special attention will now be given to the one parameter exponential family. Let θ be a real parameter and X a random observable vector with probability density

$$f_\theta(X) = C(\theta)\, e^{\theta T(X)}\, h(X)\ , \tag{4.1}$$

with respect to a measure μ. It might be recalled here that $T(X)$ is sufficient for θ. It was seen earlier that a uniformly most sensitive test statistic for the hypothesis

$$H_o\colon \theta = \theta_o \quad \text{versus} \quad H_A\colon \theta \neq \theta_o \tag{4.2}$$

does not exist. This presents the need to investigate the existence of unbiased significance tests.

Let T^* be a test statistic with distribution $G_o(T^*)$ under H_o, where for convenience if T^* is discrete G_o is defined only at points of positive probability. Then for T^* to be unbiased one needs

$$H_o^*(\alpha) = \alpha \quad \text{and} \quad H_\theta^*(\alpha) \geq \alpha \quad \text{for} \quad \theta \neq \theta_o$$

for all achievable α . Now $SL(T^*) = G_o(T^*)$ and

$$H^*_\theta(\alpha) = P_\theta\left[G(T^*) \leq \alpha\right] = P_\theta\left[T^* \leq t^*\right]$$

where $t^* = G_o^{-1}(\alpha)$. Let $R = \left\{X \mid G_o(T^*) \leq \alpha\right\} = \left\{X \mid T^* \leq t^*\right\}$.
Then

$$H^*_\theta(\alpha) = \int_R C(\theta)\, e^{\theta T(X)}\, h(X)\, d\mu(X) \ .$$

It will be assumed further that $H^*_\theta(\alpha)$ possesses continuous derivatives with respect to θ and that the differentiation can be carried out under the integral. Thus,

$$\begin{aligned} \frac{\partial}{\partial\theta} H^*_\theta(\alpha) &= H^{*'}_\theta(\alpha) \\ &= \int_R \left[C(\theta)\, T(X)\, e^{\theta T(X)}\, h(X) + C'(\theta)\, e^{\theta T(X)}\, h(X)\right] d\mu(x) \\ &= \int_R T(X)\, f_\theta(X)\, d\mu(X) + \frac{C'(\theta)}{C(\theta)} \int_R f_\theta(X)\, d\mu(X) \\ &= E_\theta\left[T \mid G_o(T^*) \leq \alpha\right] \left\{P_\theta\left[G_o(T^*) \leq \alpha\right]\right\} \\ &\quad + \frac{C'(\theta)}{C(\theta)} H^*_\theta(\alpha) \ ; \end{aligned} \tag{4.3}$$

or

$$H^{*'}_\theta(\alpha) = H^*_\theta(\alpha)\left\{E_\theta\left[T \mid G_o(T^*) \leq \alpha\right] + \frac{C'(\theta)}{C(\theta)}\right\} . \tag{4.4}$$

But $E_\theta(T) = - C'(\theta)/C(\theta)$ so that (4.4) becomes

$$H^{*'}_\theta(\alpha) = H^*_\theta(\alpha)\left\{E_\theta\left[T \mid SL(T^*) \leq \alpha\right] - E_\theta(T)\right\} \tag{4.5}$$

or

$$H^{*'}_\theta(\alpha) = H^*_\theta(\alpha)\left\{E_\theta\left[T \mid T^* \leq t^*\right] - E_\theta(T)\right\} . \tag{4.6}$$

Equation (4.5) holds for all achievable α and (4.6) for all possible values t^* in the range of T^* if T^* is continuous or all values t^* which have positive probability if T^* is discrete.

Now if T^* is to be unbiased, then

$$H_o^*(\alpha) = \alpha \quad \text{and} \quad H_\theta^*(\alpha) \geq \alpha$$

for all achievable α. That means, as a function of θ, $H_\theta^*(\alpha)$ must have its minimum value at $\theta = \theta_o$ for all achievable α. Thus, it is necessary that

$$\frac{\partial}{\partial\theta} H_\theta^*(\alpha) \Big|_{\theta=\theta_o} = H_o^{*'}(\alpha) = 0 . \tag{4.7}$$

Applying condition (4.7) to (4.5) and (4.6) gives

$$E_o\left[T \mid SL(T^*) \leq \alpha\right] = E_o(T)$$

for all achievable α and

$$E_o\left[T \mid T^* \leq t^*\right] = E_o(T)$$

for all possible values t^* of T^*. These results are stated in the following theorem.

<u>Theorem 4.1</u> For the density

$$f_\theta(X) = C(\theta)\, e^{\theta T(X)}\, h(X)$$

and the hypothesis problem

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta \neq \theta_o$$

a necessary condition for a statistic T^* to be unbiased is that

$$E_o\left[T \mid SL(T^*) \leq \alpha\right] = E_o(T) \tag{4.8}$$

for all achievable α and

$$E_0\left[T \mid T^* \leq t^*\right] = E_0(T) \tag{4.9}$$

for all possible values t^* of T^*.

It should be noted here that if in the density used in Theorem 4.1 θ were replaced by a strict monotonic function of θ, say $q(\theta)$, the theorem will again follow. In this case $E_\theta(T) = -C'(\theta)/q'(\theta)\, C(\theta)$, but condition (4.7) again reduces to equations (4.8) and (4.9).

Since equation (4.9) must hold for all t^* a more direct relationship between the sufficient statistic T and the unbiased statistic T^* can be obtained which may sometimes help in determining likely candidates for an unbiased statistic. In particular, T and T^* must be uncorrelated. In order to show this the following theorem is needed.

<u>Theorem 4.2</u> Let X and Y be random variables and suppose that

$$E\left[X \mid Y \leq y\right] = E\left[X\right]$$

for all y. Then X and Y are uncorrelated.

<u>Proof</u>: The proof is given for X and Y continuous random variables. The theorem can be proved for the discrete case by induction on the y values. Furthermore, the existence of the proper partial derivatives of the c.d.f.s $F(x)$, $F(y)$, and $F(x,y)$ will be assumed so that all necessary densities exist. Thus

$$E\left[X \mid Y \leq y\right] = \int_{-\infty}^{\infty} x \int_{-\infty}^{y} f(x,y)\, dy\, dx \,/\, P(Y \leq y)$$

$$= \frac{1}{P(Y \leq y)} \int_{-\infty}^{y} f(y) \int_{-\infty}^{\infty} x\, \frac{f(x,y)}{f(y)}\, dx\, dy$$

$$= \frac{1}{P(Y \leq y)} \int_{-\infty}^{y} E\left[X \mid y\right] f(y)\, dy$$

$$= E(X)$$

implies

$$\int_{-\infty}^{y} E\left[X \mid y\right] f(y)\, dy = E\left[X\right] F(y)$$

for all y. Taking partials with respect to y gives

$$E\left[X \mid y\right] f(y) = E\left[X\right] f(y) \quad \text{for all} \quad y$$

or

$$E\left[X \mid Y = y\right] = E\left[X\right] \quad \text{for all} \quad y .$$

Hence

$$E\left[X \mid Y\right] = E\left[X\right] .$$

Therefore

$$\begin{aligned} E\left[X\ Y\right] &= E\left\{E\left[X\ Y \mid Y\right]\right\} \\ &= E\left\{Y\left[E(X \mid Y)\right]\right\} \\ &= E\left\{Y\left[E(X)\right]\right\} \\ &= E(X)\ E(Y) . \end{aligned}$$

Hence, X and Y are uncorrelated.

Applying the results of this theorem to the conditions in Theorem 4.1 gives:

Theorem 4.3 For the density

$$f_\theta(X) = C(\theta)\ e^{\theta T(X)}\ h(X)$$

and the hypothesis problem

$$H_o: \theta = \theta_o \quad \text{versus} \quad H_A: \theta \neq \theta_o$$

a necessary condition for a statistic T^* to be unbiased is that T and T^* be uncorrelated.

It is useful to note that the condition for unbiasedness, namely $E_o\left[T \mid T^* \leq t^*\right] = E_o(T)$ for all possible values t^*, implies that

$$E_o\left[T \mid T^* > t^*\right] = E_o(T) ,$$

and also

$$E_o\left[T \mid T^* = t^*\right] = E_o(T)$$

for all possible t^*.

The conditions in Theorems 4.1 and 4.2 are necessary conditions on the test statistic for an unbiased test of significance. The problem remains to determine a sufficient condition and a more exact nature of the statistic T^*. At times it might help to look at the second derivative of $H_\theta^*(\alpha)$ with respect to θ. Thus

$$\frac{\partial^2}{\partial\theta^2} H_\theta^*(\alpha) = H_\theta^{*''}(\alpha)$$

$$= \int_R C''(\theta)\, e^{\theta T(x)}\, h(x)\, d\mu(x) + 2\int_R C'(\theta)\, T(x)\, e^{\theta T(x)}\, h(x)\, d\mu(x)$$

$$+ \int_R C(\theta)\, T^2(x)\, e^{\theta T(x)}\, h(x)\, d\mu(x)$$

$$= \frac{C''(\theta)}{C(\theta)} P_\theta\left[T^* \leq t^*\right] + 2\frac{C'(\theta)}{C(\theta)} E_\theta\left[T \mid T^* \leq t^*\right] P_\theta\left[T^* \leq t^*\right]$$

$$+ E_\theta\left[T^2 \mid T^* \leq t^*\right] P_\theta\left[T^* \leq t^*\right]$$

or

$$\frac{H_\theta''(\alpha)}{P_\theta\left[T^* \leq t^*\right]} = 2E_\theta^2(T) - E_\theta(T^2) - 2E_\theta(T)\, E_\theta\left[T \mid T^* \leq t^*\right]$$

$$+ E_\theta\left[T^2 \mid T^* \leq t^*\right] .$$

Now if the necessary condition is satisfied, i.e., $H_o'(\alpha) = 0$ for all achievable α, then at $\theta = \theta_o$, the above reduces to

$$\frac{H_o''(\alpha)}{H_o(\alpha)} = E_o\left[T^2 \mid T^* \leq t^*\right] - E_o^2\left[T \mid T^* \leq t^*\right] - \left\{E_o(T^2) - E_o^2(T)\right\} ,$$

or

$$\frac{H_o''(\alpha)}{H_o(\alpha)} = Var_o(T \mid T^* \leq t^*) - Var_o(T) .$$

The statistic T^* then must be chosen so that $H_o'(\alpha) = 0$ for all achievable α and such that

$$Var_o(T \mid T^* \leq t^*) > Var_o(T) \tag{4.10}$$

for all possible t^*.

In looking for a uniformly most sensitive unbiased test statistic, if the theory is to parallel the Neyman Pearson theory, one would perhaps try finding T^* as a function of the sufficient statistic T. Recall that in hypothesis testing for any test function there always exists a test based on the sufficient statistic which has the same power as the given test function. Although this writer has not been able to derive a proper analog to this theorem, the search for T^* will none the less be restricted to a function of T. If T^* is a function of T the unbiased condition does put some immediate restrictions on the nature of T^*.

Let t^* be any value of T^* and let $R = \left\{T \mid T^* \leq t^*\right\}$. Then

$$H_o^*(\alpha) = \int_R p_o(t)\, dt$$

where $p_o(t)$ is the density of t under H_o, and one must have

$$E_o(T \mid T^* \leq t^*) = \frac{\int_R t\, p_o(t)\, dt}{\int_R p_o(t)\, dt} = E_o(T) .$$

For every achievable α , or equivalently for all possible t^* , the corresponding set R must be of the form $R = R_1 \cup R_2$ where R_1 and R_2 each have positive probability and where R_1 contains points t such that $t < E_o(T)$ and R_2 contains points t such that $t > E_o(T)$. If this were not the case, that is if for some t^* the set R contained only points greater than $E_o(T)$ or only points less than $E_o(T)$, then one would have either

$$E_o(T \mid T^* \leq t^*) > E_o(T) \quad \text{or} \quad E_o(T \mid T^* \leq t^*) < E_o(T)$$

and the unbiased condition would not be satisfied.

Furthermore suppose that α_1 and α_2 are any two achievable values of $SL(T^*)$ corresponding to values t_1^* and t_2^* respectively, where $\alpha_1 < \alpha_2$. Then

$$H_\theta^*(\alpha_1) = \int_R p_\theta(t)\, dt = \int_{R_1} p_\theta(t)\, dt + \int_{R_2} p_\theta(t)\, dt$$

and

$$H_\theta^*(\alpha_2) = \int_S p_\theta(t)\, dt = \int_{S_1} p_\theta(t)\, dt + \int_{S_2} p_\theta(t)\, dt ,$$

where R_1 and R_2 and S_1 and S_2 are decompositions of the sets R and S as indicated in the preceding paragraph. Then $R_1 \subset S_1$ and $R_2 \subset S_2$ where $\subset$ denotes "proper subset of". This last statement can be proved as follows: Obviously $R_1 \subseteq S_1$ and $R_2 \subseteq S_2$. Suppose $R_1 = S_1$ and $R_2 \subset S_2$. The unbiased condition

$$E_o(T \mid T^* \leq t_1^*) = E_o(T \mid T^* \leq t_2^*) = E_o(T)$$

implies that

$$\frac{\int_R t\, p_o(t)\, dt}{\int_R p_o(t)\, dt} = \frac{\int_S t\, p_o(t)\, dt}{\int_S p_o(t)\, dt}$$

or

$$\frac{\int_{R_1} t\, p_o(t)\, dt + \int_{R_2} t\, p_o(t)\, dt}{\int_{R_1} p_o(t)\, dt + \int_{R_2} p_o(t)\, dt}$$

$$= \frac{\int_{R_1} t\, p_o(t)\, dt + \int_{R_2} t\, p_o(t)\, dt + \int_{S_2-R_2} t\, p_o(t)\, dt}{\int_{R_1} p_o(t)\, dt + \int_{R_2} p_o(t)\, dt + \int_{S_2-R_2} p_o(t)\, dt}$$

which becomes, upon simplification,

$$E_o(T \mid T^* \leq t_1^*) = \frac{\int_{S_2-R_2} t\, p_o(t)\, dt}{\int_{S_2-R_2} p_o(t)\, dt} . \qquad (4.11)$$

But, for $t \in S_2-R_2$ one has $t > E_o(T)$, hence (4.11) implies

$$E_o(T \mid T^* \leq t_1^*) > E_o(T) ,$$

which contradicts the unbiased condition. Thus, $R_1 \subset S_1$ and $R_2 \subset S_2$.

Moreover it seems reasonable that T^* should in fact be a unimodal function of T , since for earlier one-sided alternatives large values of T , small values of $-T$, indicated evidence against H_o when the alternative was $\theta < \theta_o$ and small values of T gave evidence against H_o when the alternative was $\theta > \theta_o$. So, although complete justification is not given here, in looking for an unbiased test statistic attention will be focused on unimodal functions of T , that is, two tailed tests. As will be seen however, there will not always exist an unbiased test statistic which is a two tailed function of T . This situation arises when T is discrete. When T is continuous it appears that a unimodal function of T should exist which is uniformly most sensitive unbiased. This conjecture is formally stated as follows: For the continuous one-parameter exponential family, equation (4.1), there exists a uniformly most sensitive unbiased test statistic $T^* = g(T)$, where g is a unimodal function of T , for evaluating the significance level of the hypothesis problem $H_o: \theta = \theta_o$ versus $H_A: \theta \neq \theta_o$.

This writer has not been able to prove the above conjecture in the direct context of significance testing as has been adopted in this paper. It can be proved however, by considering the significance level as a decision making device as follows:

Suppose the statement is false. Then for some α , say α_o , there is no unimodal function T^* of T such that both

$$H^*_\theta(\alpha_o) \geq H^*_o(\alpha_o) \text{ , (unbiased condition) ,} \qquad (4.12)$$

and

$$H^*_\theta(\alpha_o) \geq H^{(1)}_\theta(\alpha_o) \text{ , (most sensitive condition),} \qquad (4.13)$$

for any other continuous statistic $T^{(1)}$. Consider a unimodal function T^* and reject H_o if $SL(T^*) \leq \alpha_o$. This test has the form reject if $T < C_1$ or $T > C_2$, since T^* is unimodal. But it was assumed that not both (4.12) and (4.13) held simultaneously, i.e., no uniformly most powerful unbiased size α test has this form. This contradicts Lehman's proof (10) that there exists a uniformly most powerful unbiased test of the form reject if $T < C_1$ or $T > C_2$ for the hypothesis problem $\theta = \theta_o$ versus $\theta \neq \theta_o$.

For the discrete case the situation is summed up in the following theorem.

Theorem 4.4 Consider the exponential family (4.1) where T is discrete. Then, in general, there does not exist a unimodal function of T which is an unbiased test statistic for evaluating the significance level for the hypothesis problem $\theta = \theta_o$ versus $\theta \neq \theta_o$, for all θ_o.

The theorem can be proved by exhibiting a counter example that there does exist a unimodal function of T which is unbiased for $H_o: \theta = \theta_o$ versus $H_A: \theta \neq \theta_o$.

Example 4.1 Consider n trials of a point binomial and the hypothesis $H_o: p = p_o$ versus $H_A: p \neq p_o$. Then $X = \sum_{i=1}^{n} X_i$ is sufficient and

$$f_p(X) = \binom{n}{X} p^X (1 - p)^{n-X} .$$

This is a form of the one parameter exponential (n known) with $p = \theta$, $q(\theta) = \log p/1-p$ and $T(X) = X$. Suppose T^* is a unimodal function of X and denote an arbitrary achievable significance level by α_{jk} so that

$$H_p^*(\alpha_{jk}) = \sum_{x=o}^{j} \binom{n}{x} p^x (1-p)^{n-x} + \sum_{n-j+k}^{n} \binom{n}{x} p^x (1-p)^{n-x} .$$

Writing the above sums in terms of the Beta distribution gives

$$H_p^*(\alpha_{jk}) = 1 - \frac{\Gamma(n+1)}{\Gamma(j+1)\Gamma(n-j)} \int_0^p t^j (1-t)^{n-j-1}\, dt$$

$$+ \frac{\Gamma(n+1)}{\Gamma(n-j+k)\Gamma(j-k+1)} \int_0^p t^{n-j+k-1}(1-t)^{j-k}\, dt \; .$$

Now differentiating with respect to p and simplifying gives

$$H_p^{*'}(\alpha_{jk}) = \frac{\Gamma(n+1)p^j(1-p)^{j-k}}{(j-k)!(n-j+k-1)!} \Big[p^{n+k-2j-1}$$

$$- \frac{(n-j+k-1)!(j-k)!}{j!(n-j-1)!} (1-p)^{n+k-2j-1} \Big] \; .$$

Setting $H_p^{*'}(\alpha_{jk}) = 0$ gives the solution (other than the trivial solutions $p = 0 , 1$)

$$p = \frac{h}{1+h} \quad \text{where} \quad h = \left[\frac{(n-j+k-1)!(j-k)!}{j!(n-j-1)!}\right]^{\frac{1}{n+k-2j+1}} \; .$$

So, in general, the solution p is a function of j and k. That is, the value of p that minimizes $H_p^*(\alpha_{jk})$ depends on the particular significance level and for different α there is not one value p_o that will minimize $H_p^*(\alpha)$ for every achievable α. Therefore, in general, there can not exist an unbiased two tail test statistic for $H_o\colon p = p_o$ versus $H_A\colon p \neq p_o$ for arbitrary p_o.

If $k = 0$, then one obtains $p = ½$ which is independent of j and does give a minimum for $H_p^*(\alpha)$. Thus, if one uses an equal tails test there does exist an unbiased test of significance for $H_o\colon p = ½$ versus $H_A\colon p \neq ½$.

The preceding results are stated in the following theorem and corollary.

Theorem 4.5 For the binomial distribution with n known and parameter p, for $p_o \neq 0, \frac{1}{2}, 1$, there does not exist a unimodal function of X which is an unbiased test statistic for evaluating the significance level of the hypothesis problem $H_o: p = p_o$ versus $H_A: p \neq p_o$.

Corollary 4.1 For the binomial distribution there does exist an unbiased unimodal test statistic for the hypothesis problem $H_o: p = \frac{1}{2}$ versus $H_A: p \neq \frac{1}{2}$, and $T^*(X) = -\left|X - \frac{n}{2}\right|$ is such a statistic.

Consider now another example which illustrates the non-existence of a unimodal unbiased test of significance.

Example 4.2 Let $(x_1, \cdots, x_n)$ denote a random sample from a Poisson distribution with parameter λ. Then for $X = \sum_{i=1}^{n} X_i$,

$$f(X) = \frac{(n\lambda)^X e^{-n\lambda}}{X!}, \quad X = 0, 1, 2, \cdots,$$

is of exponential form with $T(X) = X$ and $q(\lambda) = \log n\lambda$. The hypothesis problem is $H_o: \lambda = \lambda_o$ versus $H_A: \lambda \neq \lambda_o$.

Because of the nature of the range on X, for any test of significance based on $T(X) = X$ to be two tailed and unbiased it would have to admit to a finite number of achievable significance levels (see the discussion associated with equation (4.11)). Suppose there are $R+1$ attainable values of X which are less than $E_o(T) = n\lambda_o$ and that for the j-th achievable significance level, say α_j,

$$H_\lambda(\alpha_j) = \sum_{x=0}^{j} \frac{(n\lambda)^x e^{-n\lambda}}{x!} + \sum_{2R+2-j}^{\infty} \frac{(n\lambda)^x e^{-n\lambda}}{x!} , \quad j = 0, \cdots, R+1 .$$

This form would also give the maximum number of achievable significance levels. Writing the sums in terms of the Gamma distributions gives

$$H_\lambda(\alpha_j) = \int_{n\lambda}^{\infty} \frac{1}{j!} z^j e^{-z} dz + 1 - \int_{n\lambda}^{\infty} \frac{1}{\Gamma(2R+2-j)} z^{2R+1-j} e^{-z} dz .$$

Then differentiation with respect to λ yields

$$H_\lambda'(\alpha_j) = n(n\lambda)^j e^{-n\lambda} \left[-\frac{1}{j!} + \frac{(n\lambda)^{2R-2j+1}}{(2R - j + 1)!} \right] .$$

Setting the derivative equal to zero and solving for λ gives the unique solution

$$\lambda = \frac{1}{n} \left[\frac{(2R - j + 1)!}{j!} \right]^{\frac{1}{2R-2j+1}} ,$$

and this value of λ does in fact minimize $H_\lambda(\alpha_i)$. Notice, however, that λ depends on j so that the same value of λ does not minimize $H_\lambda(\alpha)$ for all achievable α . Thus, the corresponding significance test can not be unbiased.

In the preceding examples the nonexistence of a unimodal unbiased test statistic resulted from the fact that $\frac{\partial}{\partial\theta} H_\theta(\alpha) = 0$ did not give the same solution of θ for every achievable α . One final example seems appropriate, that being a continuous situation where there does exist a unimodal test statistic T^* such that $\frac{\partial}{\partial\theta} H_\theta^*(\alpha) = 0$ admits the same solution of θ for every α , $0 \leq \alpha \leq 1$. Furthermore, it will be seen that the test statistic T^* which is used is, in fact, an unbiased statistic for the given two sided hypothesis problem.

<u>Example 4.3</u>: Let $X = (X_1, \cdots, X_n)$ denote a random sample of size n from $N(0, \sigma^2)$. The hypothesis is

$$H_o: \sigma^2 = \sigma_o^2 \quad \text{versus} \quad H_A: \sigma^2 \neq \sigma_o^2 .$$

The density of X is then

$$f(X) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-T/2\sigma^2}$$

where $T = \sum_{i=1}^{n} X_i^2$. This is a one parameter form of the exponential with $T = \Sigma X_i^2$ and $\theta = \sigma^2$ where $q(\theta) = -1/2\sigma^2$. Now T has probability density $(1/\sigma^2)\, f_n(t/\sigma^2)$ where

$$f_n(t) = \frac{1}{2^{n/2}\Gamma(n/2)} t^{n/2 - 1} e^{-t/2}, \quad t > 0 ,$$

is the χ^2 distribution with n degrees of freedom, i.e., χ_n^2. Under H_o, T/σ_o^2 has a χ_n^2 distribution.

If T^* is to be a unimodal unbiased function of T then for every α, that means every possible value t^*, there must exist values t_1 and t_2 such that

$$H_o(\alpha) = P_o(T^* \leq t^*) = P_o(T \leq t_1) + P_o(T \geq t_2) = \alpha$$

or

$$P_o(T^* > t^*) = P_o(t_1 < T < t_2) = P_o(t_1/\sigma_o^2 < \chi_n^2 < t_2/\sigma_o^2) .$$

Consider the unbiased condition

$$E_o(T \mid T^* > t^*) = E_o(T) = \sigma_o^2\, n$$

or equivalently

$$E_o(T \mid T^* > t^*) = \sigma_o^2 \left\{ E(\chi_n^2 \mid t_1/\sigma_o^2 < \chi_n^2 < t_2/\sigma_o^2 \right\} = \sigma_o^2\, n$$

which reduces to

$$\int_{t_1/\sigma_o^2}^{t_2/\sigma_o^2} t\, f_n(t)\, dt = n \int_{t_1/\sigma_o^2}^{t_2/\sigma_o^2} f_n(t)\, dt\ . \qquad (4.14)$$

Consider the integral

$$\int_a^b y^{n/2}\, e^{-y/2}\, dy\ .$$

Then integrating by parts, letting $u = y^{n/2}$ and $dv = e^{-y/2}\, dy$,

yields

$$\int_a^b y^{n/2}\, e^{-y/2}\, dy = -2y^{n/2}\, e^{-y/2} \Big|_a^b + n \int_a^b y^{n/2\,-1}\, e^{-y/2}\, dy\ . \qquad (4.15)$$

Applying this result to the left hand member of equation (4.14)

gives

$$\frac{-2t^{n/2}\, e^{-t/2}}{2^{n/2}\Gamma(n/2)} \Bigg|_{t_1/\sigma_o^2}^{t_2/\sigma_o^2} + n \int_{t_1/\sigma_o^2}^{t_2/\sigma_o^2} f_n(t)\, dt = n \int_{t_1/\sigma_o^2}^{t_2/\sigma_o^2} f_n(t)\, dt\ ,$$

and hence

$$t_1^{n/2}\, e^{-t_1/2\sigma_o^2} = t_2^{n/2}\, e^{-t_2/2\sigma_o^2}\ . \qquad (4.16)$$

Therefore, the necessary condition for unbiasedness implies that T^* be chosen so that equation (4.16) holds. The statistic

$$T^* = T^{n/2}\, e^{-T/2\sigma_o^2} \qquad (4.17)$$

would satisfy this condition.

It remains to show then, that for T^* as defined in equation (4.17), the equation $\frac{\partial}{\partial\sigma^2} H^*_{\sigma^2}(\alpha) = 0$ is satisfied at $\sigma^2 = \sigma_o^2$, for all α, and that this value σ_o^2 does actually minimize $H^*_{\sigma^2}(\alpha)$. Minimizing

$$H^*_{\sigma^2}(\alpha) = P_{\sigma^2}\left[T^* \leq t^*\right]$$

is equivalent to maximizing the function

$$G(\sigma^2) = 1 - H^*_{\sigma^2}(\alpha) = P_{\sigma^2}\left[T^* > t^*\right]. \tag{4.18}$$

But, by choice of T^*, an observed value t^* uniquely determines values t_1 and t_2 such that equation (4.16) holds and

$$G(\sigma^2) = P_{\sigma^2}\left[t_1 < T < t_2\right] = P_{\sigma^2}\left[t_1/\sigma^2 < T/\sigma^2 < t_2/\sigma^2\right]. \tag{4.19}$$

Now, when the parameter is σ^2, T/σ^2 has a χ^2_n distribution, hence,

$$G(\sigma^2) = \int_{t_1/\sigma^2}^{t_2/\sigma^2} \frac{t^{n/2-1}\, e^{-t/2}}{\Gamma(n/2)\, 2^{n/2}}\, dt\ .$$

Differentiating w.r.t. σ^2 and simplifying gives

$$G'(\sigma^2) = \frac{1}{\Gamma(n/2)2^{n/2}(\sigma^2)^{n/2+1}}\left[t_1^{n/2}\, e^{-t_1/2\sigma^2} - t_2^{n/2}\, e^{-t_2/2\sigma^2}\right]. \tag{4.20}$$

On comparison with equation (4.16), it is seen that $G'(\sigma_o^2) = 0$. Thus, also, $H^*_{\sigma_o^2}(\alpha) = 0$. Furthermore,

$$G''(\sigma^2) = \frac{(n/2+1)}{\Gamma(n/2)2^{n/2}(\sigma^2)^{n/2+2}}\left[t_2^{n/2}\, e^{-t_2/2\sigma^2} - t_1^{n/2}\, e^{-t_1/2\sigma^2}\right]$$

$$+ \frac{1}{\Gamma(n/2)2^{n/2}(\sigma^2)^{n/2+3}}\left[t_1^{n/2+1}\, e^{-t_1/2\sigma^2} - t_2^{n/2+1}\, e^{-t_2/2\sigma^2}\right]. \tag{4.21}$$

Evaluating equation (4.21) at $\sigma^2 = \sigma_o^2$ gives

$$G''(\sigma_o^2) = \frac{1}{\Gamma(n/2)2^{n/2}(\sigma_o^2)^{n/2+3}} \left[t_1^{n/2+1} \, e^{-t_1/2\sigma_o^2} - t_2^{n/2+1} \, e^{-t_2/2\sigma_o^2} \right]. \tag{4.22}$$

Therefore, $G''(\sigma_o^2) < 0$, since $t_1 < t_2$ and equation (4.16) imply the difference in equation (4.22) is negative. Also, the results $G'(\sigma_o^2) = 0$ and $G''(\sigma_o^2) < 0$ are independent of the value t^* , by choice of T^* , thus independent of α . Thus, $G(\sigma^2)$ is maximized at $\sigma^2 = \sigma_o^2$ for all α , i.e., $H^*_{\sigma^2}(\alpha)$ is minimized at $\sigma^2 = \sigma_o^2$. Hence,

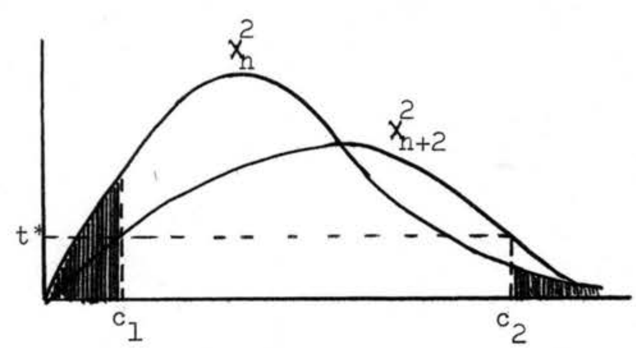
$$H^*_{\sigma^2}(\alpha) \geq H^*_o(\alpha) \text{ , for all } \alpha \text{ ,}$$

so that T^* is unbiased.

The conjecture made earlier would seem to suggest that T^* is uniformly most sensitive unbiased for the given hypothesis problem. For observed T^* , say t^* , t^* uniquely determines constants c_1 and c_2 such that the corresponding observed significance level is

$$\alpha^* = P_o(T^* \leq t^*) = \Pr(\chi_n^2 \leq c_1) + \Pr(\chi_n^2 \geq c_2) \text{ .}$$

The following graphical representation of how an observed significance level is obtained may be helpful. The value α^* is represented by the shaded area.



It might also be noted in passing that T^* and T do satisfy Theorem 4.3. That is, T^* and T are uncorrelated.

CHAPTER V

EXTENSIONS

After completion of this study many questions still remain unanswered and there is certainly a need for further investigation of this topic. There appear to be numerous possibilities for extension of the theory, but at the same time the theory seems to have some limitations.

In the consideration of composite alternatives, this investigation has been restricted to hypothesis problems involving a single real parameter. The immediate need would be to extend the development to multi-parameter situations. For example, it would be of special interest to show that the test statistics ordinarily used in the familiar normal theory tests of hypothesis problems considered in elementary statistical methods courses are also the optimum statistics for significance testing. For example, how would one justify the use of the t statistic for evaluating the significance level for a hypothesis involving μ when σ^2 is an unknown nuisance parameter?

The role of the sufficient statistic in significance testing seems to be a mystery. Although the significance tests developed in this investigation for the exponential distribution were based on a sufficient statistic, the sufficiency condition was really not used. In hypothesis testing the use of a test function based on the sufficient statistic S is justified by the fact that for any test function $\phi(X)$

there exists a test function, namely $E(\phi(X)|\,s)$, based on the sufficient statistic which has the same power function as $\phi(X)$. The analog to this theorem for significance testing would appear to be: If S is a sufficient statistic and T is any other statistic, then there exists a function of S , say $f(S)$, such that $H_\theta^f(\alpha) \geq H_\theta^T(\alpha)$, $\theta \in \Theta_A$, for all achievable α . However, this statement can readily be proved false for the statistics T and S discrete. For the continuous case, the natural candidates for $f(S)$, namely $f(S) = P_o(T \leq t|S)$ and $f(S) = E_o(T|S)$ do not give the desired result.

When widening the scope of the Neyman-Pearson development to multi-parameter situations a number of new concepts and devices are introduced; for example, similar tests, tests of Neyman structure, invariant tests and stringency. Perhaps analogs to these concepts can be given and used in extending the theory of significance testing as described in this paper. In conclusion, one such analog will now be considered.

Similarity and Neyman Structure

In hypothesis testing a test function $\phi(x)$ was said to be similar of size α if $\beta_\phi(\theta) = \alpha$ for all distributions of X belonging to a given family $\mathcal{P}^X = \left\{P_\theta^X , \theta \in \Theta_o\right\}$ of distributions, i.e., $\theta \in \Theta_o$. That is

$$P_\theta\left[\text{rejecting } H_o\right] = \alpha$$

for all $\theta \in \Theta_o$. Furthermore, if T was a sufficient statistic for $\mathcal{P}^X$, and $\mathcal{P}^T$ denoted the family $\left\{P_\theta^T , \theta \in \Theta_o\right\}$ of distributions of T as θ ranged over Θ_0 , then any test satisfying

$$E\left[\phi(x) \mid t\right] = \alpha , \quad \text{a.e. } \mathcal{P}^T ,$$

where a.e. $\mathcal{P}^T$ means the equation holds except on a set N with $P^T(N) = 0$ for all $P^T \in \mathcal{P}^T$, was similar with respect to $\mathcal{P}^X$. A test satisfying this condition was said to have Neyman structure with respect to T .

With the correspondence established earlier between hypothesis testing and significance testing, a test statistic T^* , or equivalently the test of significance corresponding to T^* , would be similar if $SL(T^*)$ was such that

$$P_\theta\left[SL(T^*) \leq \alpha\right] = \alpha , \quad \theta \in \Theta_0 ,$$

that is,

$$H_0^*(\alpha) = \alpha .$$

Now, if T^* is continuous this condition is automatically satisfied for all α . If T^* is discrete the condition is satisfied for all achievable α . Thus, similarity was essentially incorporated into the formulation of significance testing when it was required that the distribution of a test statistic be completely specified under the null hypothesis. As far as using similarity then to extend the theory of significance testing it appears to be a somewhat "empty" concept since similarity was actually built into the definition of a significance test to begin with.

Suppose next that T is sufficient for $\theta \in \Theta_0$. Then a test statistic T^* would be said to have Neyman structure with respect to

T if

$$P_\theta\left[SL(T^*) \leq \alpha \mid t\right] = \alpha \ , \quad \theta \in \Theta_o \ , \quad \text{a.e. } t \ ,$$

that is if

$$H_o^*(\alpha \mid t) = \alpha \ , \quad \text{a.e. } \mathcal{P}^T \ .$$

Since, by the formulation of significance tests given here, every test statistic is similar it would follow trivially that any test statistic which is of Neyman structure with respect to a sufficient statistic T is also similar.

Lehman (10) proves the following theorem: Let X be a random variable with distribution $P \in \mathcal{P} = \left\{P_\theta^X \mid \theta \in \Theta_o\right\}$ and let T be a sufficient statistic for $\mathcal{P}$. Then a necessary and sufficient condition for all similar tests to have Neyman structure with respect to T is that T be boundedly complete.

Consider the analogous result for significance tests.

<u>Theorem 5.1</u> If T is sufficient and boundedly complete for the family of distributions $\left\{P_\theta^X \mid \theta \in \Theta_o\right\}$ then all test statistics are of Neyman structure with respect to T.

<u>Proof</u>: Let T^* be any test statistic with distribution G_o under $H_o : \theta \in \Theta_o$ and T sufficient for $\theta \in \Theta_o$. Then

$$H_o^*(\alpha) = P_o\left[G_o(T^*) \leq \alpha\right] = P_o\left[SL(T^*) \leq \alpha\right] = \alpha$$

for all achievable α and any T^*. Let $A = \left\{X \mid SL(T^*) \leq \alpha\right\}$ and define the function

$$I_A(X) = 1 \quad \text{if} \quad X \in A \ ,$$

$$= 0 \quad \text{if} \quad X \notin A \ .$$

Then

$$H_o^*(\alpha) = E_o\left[I_A(X)\right] = \alpha$$

implies that

$$E_o\left[I_A(X)\right] = E_o\left\{E_o\left[I_A(X) \mid t\right]\right\} = \alpha ,$$

hence

$$E_o\left[E_o(I_A(X) \mid t) - \alpha\right] = 0 .$$

Now $E_o\left[I_A(X) \mid t\right]$ is actually independent of $\theta \in \Theta_o$, since the integral representation of $E_o\left[I_A(X) \mid t\right]$ involves only $I_A(X)$ and the conditional distribution of X given t , which is independent of θ since T was sufficient for $\theta \in \Theta_o$. Thus, let

$$f(t) = E\left[I_A(X) \mid t\right] - \alpha ,$$

which is strictly a function of t and also is bounded. Therefore

$$E\left[f(t)\right] = 0$$

implies

$$f(t) \equiv 0 ,$$

since T was boundedly complete. Thus, one has

$$E\left[I_A(X) \mid t\right] = \alpha$$

or

$$H_o^*(\alpha \mid t) = \alpha$$

for all achievable α .

The converse of Theorem 5.1 would be: If all test statistics have Neyman structure with respect to a sufficient statistic T , then T is boundedly complete.

The writer has not been able to prove the above conjecture but at the same time could not exhibit a counter example.

CHAPTER VI

SUMMARY

In this paper a study of providing a rationale for significance testing which avoids any commitment to decision rules is made. The approach taken is to attempt to build a theory of significance tests which parallels the classical Neyman-Pearson theory of hypothesis testing; that is, define a criterion for comparing significance tests, or equivalently statistics used in performing the significance test, then find a statistic which is in some sense "best" for a problem involving a simple null hypothesis and a simple alternative hypothesis, and then to extend the scope of the theory to include special classes of composite hypotheses.

In Chapter II a determination of a criterion for comparing test statistics is given and the concept of a most sensitive test statistic is discussed. It is shown that for a simple null hypothesis and a simple alternative hypothesis the familiar likelihood ratio statistic is a most sensitive test statistic. Other desirable properties of the likelihood ratio are considered.

In Chapter III consideration is given to one sided alternative hypotheses. Special attention is given to distributions with monotone likelihood ratios. Uniformly most sensitive test statistics are obtained for this class of distributions and in particular for the one parameter exponential family of distributions. Statistics which are

most sensitive in a neighborhood of the hypothesized value of the parameter are also discussed.

The theory is extended to two sided alternatives in Chapter IV. The concept of unbiased significance tests is discussed and a number of necessary conditions for the existence of an unbiased significance test are given with special attention again being given to the one parameter exponential distribution. Uniformly most sensitive unbiased test statistics are also discussed.

Chapter V concludes the investigation with some suggestions for extensions and further study.

A SELECTED BIBLIOGRAPHY

(1) Anscombe, F. J. "Tests of Goodness of Fit." Journal of the Royal Statistical Society. Series B. Vol. 25 (1963), 81-94.

(2) Bernoulli, D. "Recherches Physiques et Astronomiques..." Pieces qui ont Remporte la Prix Double de l'Academie Royale des Sciences. Paris, (1734), 93-122.

(3) Dempster, A. P. and Schatzoff, M. "Expected Significance Level as a Sensitivity Index for Test Statistics." Journal of the American Statistical Association. Vol. 60 (1965), 420-436.

(4) Ferguson, Thomas S. Mathematical Statistics, A Decision Theory Approach. Academic Press Inc. New York, N. Y. (1967).

(5) Fisher, R. A. Statistical Methods for Research Workers. Oliver and Boyd. Edinburgh and London (1925).

(6) Fisher, R. A. "Statistical Methods and Scientific Induction." Journal of the Royal Statistical Society. Series B. Vol. 17 (1955), 69-78.

(7) Fisher, R. A. Statistical Methods and Scientific Inference. Hafner Publishing Co. New York, N. Y. (1956).

(8) Fisz, Marek. Probability Theory and Mathematical Statistics. 3rd ed. John Wiley and Sons Inc. New York, N. Y. (1963).

(9) Fraser, D. A. S. Nonparametric Methods in Statistics. John Wiley and Sons Inc. New York, N. Y. (1957).

(10) Lehmann, E. L. Testing Statistical Hypotheses. John Wiley and Sons Inc. New York, N. Y. (1959).

(11) Neyman, J. and Pearson, E. S. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." Transactions of the Royal Society of London. Series A. Vol. 231 (1933), 289-337.

(12) Pearson, K. "On the Criterion that a Given System of Deviations from the Probable... is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling." Philosophical Magazine. Series 5. Vol. 50 (1900), 157-175.

(13) Rao, C. Radhakrishna. Linear Statistical Inference and Its Applications. John Wiley and Sons Inc. (1965).

(14) "Student". "The Probable Error of a Mean." Biometrika. Vol. 6 (1908), 1-25.

VITA

Richard Daniel Finley

Candidate for the Degree of

Doctor of Philosophy

Thesis: A THEORY OF SIGNIFICANCE TESTING

Major Field: Mathematics and Statistics

Biographical:

Personal Data: Born in New Albany, Indiana, August 30, 1936, the son of Demi D. and Katherine V. Finley.

Education: Attended grade school in Robinson, Illinois; graduated from Robinson High School in 1954; received the Bachelor of Science degree from Northeastern State College, Tahlequah, Oklahoma, with a major in Mathematics in May, 1958; received the Master of Science degree from Colorado State University, Fort Collins, Colorado, with a major in Mathematics in August, 1960; completed requirements for the Doctor of Philosophy degree in July, 1968, at Oklahoma State University.

Professional Experience: Graduate Assistant in the Department of Mathematics, Colorado State University, 1958-1960; Instructor in the Department of Mathematics, University of Wyoming, Laramie, Wyoming, 1960-1964. Graduate Assistant in the Department of Mathematics and Statistics, Oklahoma State University, Stillwater, Oklahoma, 1964-1968.

Professional Organizations: American Statistical Association, Mathematical Association of America.