IN THE NAME OF GOD

SHIRAZ UNIVERSITY

FACULTY OF SCIENCES

PH.D. THESIS IN MATHEMATICAL STATISTICS

TESTING SEPARATE FAMILIES OF HYPOTHESES

BY

ALIAKBAR RASEKHI

SUPERVISED BY

S. M. SADOOGHI-ALVANDI

JUNE 2008

دانشکده علوم

پایان نامه دکتری در رشته‌ی آمار

آزمون فرض خانواده‌های جدا

توسط

علی‌اکبر راسخی

استاد راهنما

دکتر محمد صدوقی الوندی

خرداد ماه ۱۳۸۷

IN THE NAME OF GOD

TESTING SEPARATE FAMILIES OF HYPOTHESES

BY

ALIAKBAR RASEKHI

THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY (Ph.D.)

IN

MATHEMATICAL STATISTICS
SHIRAZ UNIVERSITY
SHIRAZ
ISLAMIC REPUBLIC OF IRAN

EVALUATED AND APPROVED BY THE THESIS COMMITTEE AS: EXCELLENT

S. Sadooghi ....SADOOGHI-ALVANDI, S. M., Ph.D., ASSOC. PROF. OF STATISTICS (SUPERVISOR AND CHAIRMAN, SHIRAZ UNIVERSITY)

J. Behboodian ....BEHBOODIAN, J., Ph.D., PROF. OF STATISTICS (SHIRAZ UNIVERSITY)

N. Nematollahi ....NEMATOLLAHI, N., Ph.D., ASSOC. PROF. OF STATISTICS (ALLAMEH TABATABA'I UNIVERSITY)

A. R. Nematollahi ....NEMATOLLAHI, A. R., Ph.D., ASSOC. PROF. OF STATISTICS (SHIRAZ UNIVERSITY)

Z. Shishehbor ....SHISHEHBOR, Z., Ph.D., ASSOC. PROF. OF STATISTICS (SHIRAZ UNIVERSITY)

JUNE 2008

To the memory of my late mother

the first and greatest teacher in my life

ACKNOWLEDGEMENT

I would like to express my gratitude to my supervisor, Professor S.M. Sadooghi-Alvandi, for his inspiration, guidance, patience and support. I learnt a lot from him, far beyond statistics.

I also wish to thank the members of my committee, Professor J. Behboodian, Dr. N. Nematollahi, Dr. A. R. Nemotollahi and Dr. Z. Shishehbor for their helpful comments and correction leading to a better version of this thesis.

The comments and helps provided by Professors N. Reid and S. Wang are greatly appreciated.

Many thanks also go to my friends Dr M. Kharrati, A. Ardalan, H. A. Mardanifard and M. R. Akhond for their helps.

I am grateful to the Department of Statistics at Shiraz University for their assistance and support during my MSc and PhD programs.

Last but not least, I would like to thank my wife and our families for their helps and supports.

ABSTRACT

TESTING SEPARATE FAMILIES OF HYPOTHESES

BY

ALIAKBAR RASEKHI

For testing separate families of hypotheses, the likelihood ratio test does not have the usual asymptotic properties. This thesis, considers the asymptotic distribution of the ratio of maximized likelihoods (RML) statistic in the special case of testing separate scale or location-scale families of distributions.

We derive saddlepoint approximations to the density and tail probabilities of the log of the RML statistic. These approximations are based on the expansion of the log of the RML statistic up to the second order, which is shown not to depend on the location and scale parameters.

The resulting approximations are applied in several cases, including Rayleigh versus exponential, normal versus Laplace, normal versus Cauchy, and Weibull versus log-normal.

Our results show that the saddlepoint approximations are satisfactory even for fairly small sample sizes, and are more accurate than normal approximations and Edgeworth approximations, especially for tail probabilities which are the values of main interest in hypothesis testing problems.

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

CHAPTER 1
Introduction

In standard parametric inference, the family of probability models is completely specified except for a limited number of unknown parameters and the problem is to make inferences about the value of parameters: Suppose that we have a random sample $X_1,\ldots,X_n$ from a distribution with density $f(x;\theta)$, where θ is an unknown (vector-valued) parameter which ranges over parameter space Θ. The general testing problem may then be formulated as testing the null hypothesis $H_0:\theta\in\Theta_0$ versus the alternative hypothesis $H_1:\theta\in\Theta_1$, where Θ_0 is a subset of the parameter space and $\Theta_1=\Theta-\Theta_0$ is its complement. For this type of formulation, there are many well-known results for both small and large sample cases. The main point about this formulation is that the distributions under both null and alternative hypotheses belong to the same family. There is, however, a more general formulation in which the distributions under null and alternative hypotheses belong to *separate* (different) families. As noted by Box and Hunter (1965): "Most of statistical discussions begin by assuming that a model is known even though in practice, the model is usually known and the main problem is to build such a suitable model. The science of model-building has been a field neglected by most statistical authors. A notable exception is the pioneering work of Cox on tests of separate families of hypotheses."

Let $X_1,\ldots,X_n$ be a random sample from a continuous distribution with unknown density function $h(x)$, and consider the problem of testing

$$H_0:h(x)=f(x;\theta)\quad \text{vs}\quad H_1:h(x)=g(x;\lambda),$$

where $f(x;\theta)$ and $g(x;\lambda)$ are density functions depending on unknown, possibly vector-valued, parameters θ and λ; for example, to test lognormal distribution versus exponential distribution, that is

$$H_0 : f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x-\mu)^2/2\sigma^2} \quad \text{vs} \quad H_1 : f(x;\theta) = \theta e^{-\theta x}.$$

If $f(x;\theta)$ and $g(x;\lambda)$ are members of the same family of densities, then the standard likelihood ratio test is usually applicable and has the well known asymptotic properties. In particular, minus twice the log of the ratio of maximized likelihoods (RML) has an asymptotic chi-square distribution. However, these properties do not hold if $f(x;\theta)$ and $g(x;\lambda)$ are *separate* families (Cox, 1961), in the sense that no member of the first family can be obtained as the limit of members in the second family. Therefore, special methods have been developed for testing such hypotheses. It should be noted that in the literature, the problem of testing separate families is also known as the problem on testing *non-nested* hypotheses.

The problem of testing separate families of hypotheses arises in many areas, such as *biology* (testing between two quantal response curves, Cox 1962); *economics* (testing non-nested economic models, Pesaran 1982); *literature* (dating the works of Plato, Cox and Brandwood 1959); and *political science* (testing non-nested models of international relations, Clarke 2001).

Note that in this formulation, the two families are not treated symmetrically. There is an alternative formulation where the two families are treated symmetrically. In that case, the problem is that of model selection or discrimination—rather than hypothesis testing. Both formulations have been considered in the literature.

In this thesis, we restrict attention to testing separate scale families

$$f(x;\theta) = \frac{1}{\theta} f_0(\frac{x}{\theta}) \quad \text{vs} \quad g(x;\lambda) = \frac{1}{\lambda} g_0(\frac{x}{\lambda}), \tag{1.1}$$

or separate location-scale families

$$f(x;\theta)=\frac{1}{\theta^{(1)}}f_0\left(\frac{x-\theta^{(0)}}{\theta^{(1)}}\right) \quad \text{vs} \quad g(x;\lambda)=\frac{1}{\lambda^{(1)}}g_0\left(\frac{x-\lambda^{(0)}}{\lambda^{(1)}}\right), \tag{1.2}$$

with $\theta=(\theta^{(0)},\theta^{(1)})$, $\lambda=(\lambda^{(0)},\lambda^{(1)})$, where f_0 and g_0 are known density functions.

The general theory of testing separate families of hypotheses (also known as non-nested hypotheses) was initiated by Cox (1961, 1962). Let

$$T_n=n^{-1}[\ell_f(\hat{\theta})-\ell_g(\hat{\lambda})] \tag{1.3}$$

denote the log of the RML statistic, where

$$\ell_f(\theta)=\sum_{i=1}^{n}\log f(x_i;\theta) \quad \text{and} \quad \ell_g(\lambda)=\sum_{i=1}^{n}\log g(x_i;\lambda)$$

denote the log-likelihood functions and $\hat{\theta}$ and $\hat{\lambda}$ denote the maximum likelihood estimators under H_0 and H_1, respectively. Cox proposed a test based on a modification of the Neyman-Pearson likelihood ratio, the statistic

$$T_n^C=T_n-E_{\hat{\theta}}(T_n),$$

where $E_\theta(.)$ denotes expectation with respect to $f(x;\theta)$. This statistic compares the observed difference between maximized log-likelihoods with an estimate of its expected value under H_0. Thus a large negative value of this statistic indicates departure from H_0. The statistic T_n^C is asymptotically normal, but the normal approximation may be satisfactory only for large sample sizes (Chen, 1980). For small samples, the normal approximation may not work well for calculation of *tail* probabilities, which are the values of main interest in hypothesis-testing. (In discrimination, the most common procedure is to select the model with the higher likelihood; i.e. to select $f(x;\theta)$ if $T_n>0$ and to select $g(x;\lambda)$ if $T_n<0$. Hence, we usually need the approximation near the *center* of the distribution, where normal approximation is more satisfactory. For using normal approximation in discrimination problems, see Bain and Engelhardt (1980), Fearn and Nebenzahl (1991) and Gupta and Kundu (2003, 2004).

For testing location-scale families, the distribution of RML statistic T_n does not depend on the parameters; see Dumonceaux, et al. (1973). Therefore, in this case, $E_\theta(T_n)$ is a constant, and Cox's test is equivalent to using the statistic T_n directly: reject H_0 if $T_n < t_\alpha$, where t_α is the critical value for a test of size α. In this case we can obtain the critical values t_α by simulation for fixed value of sample size n (Dumonceaux, et al., 1973).

For testing separate scale or location-scale families, invariance considerations lead to a most powerful invariant (MPI) test; see Lehmann, (1986, Ch. 6) and Hájek and Šidák (1967). For scale families (1.1), if both densities are either zero for $x < 0$ or symmetric about zero, then the MPI test statistic is

$$\frac{\int_0^\infty v^{n-1} \prod_{i=1}^{n} f_0(vx_i)\, dv}{\int_0^\infty v^{n-1} \prod_{i=1}^{n} g_0(vx_i)\, dv}$$

(Lehmann, 1986, p.354, Lehmann, 2006); and for location-scale families (1.2), if both densities are symmetric, the MPI test statistic is

$$\frac{\int_{-\infty}^\infty \int_0^\infty v^{n-2} \prod_{i=1}^{n} f_0(vx_i + u)\, dv du}{\int_{-\infty}^\infty \int_0^\infty v^{n-2} \prod_{i=1}^{n} g_0(vx_i + u)\, dv du}$$

(Lehmann, 1986, p.338, Hájek and Šidák, 1967, p.51). The MPI test rejects H_0 for small values of the test statistics. However, as noted by Ducharme and Frichot (2003), the calculations are often intractable and "the MPI test has been confined to a limited pairs of densities".

In this thesis, we apply saddlepoint techniques to approximate the distribution of T_n. In general, saddlepoint approximations are more accurate than normal approximations and Edgeworth approximations, especially for tail probabilities (which are the values of main interest in hypothesis testing problems). This is illustrated by several examples, which show that the saddlepoint approximations are satisfactory even for fairly small sample sizes.

The organization of this thesis is as follows: In Chapter 2, we review the most important approaches for testing separate families without restriction on location-scales family. In Chapter 3, we first review Edgeworth and saddlepoint approximation and then derive these approximations for distribution of the RML statistics, as well as normal approximation. In Chapter 4, we apply our approximations for some special and important scale or location-scale families, including Rayleigh versus exponential, normal versus Laplace, normal versus Cauchy, and extreme value versus normal (which is identical to Weibull versus log-normal). We compare the approximations with the exact results obtained by simulation; and also compare saddlepoint, Edgeworth and normal approximations. We also compare RML and MPI test statistics. In Chapter 5, we present our conclusions and recommendations. The Appendix contains details of derivation of the saddlepoint approximation and the computer programs that have been used.

CHAPTER 2

Literature Review

There are several methods proposed in the literature for testing separate families of hypotheses for general densities $f(x;\theta)$ and $g(x;\lambda)$. The RML statistics is the most important statistics which is the basis of Cox's, Williams' and Loh's methods. Another method discussed in this chapter is Epps' method which is based on the empirical generating function.

In this chapter we assume that $X_1,\ldots,X_n$ is a random sample from a distribution with unknown density function $h(x)$, and we consider the problem of testing

$$H_0 : h(x) = f(x;\theta) \quad \text{vs} \quad H_1 : h(x) = g(x;\lambda),$$

and

$$T_n = n^{-1}[\ell_f(\hat{\theta}) - \ell_g(\hat{\lambda})]$$

is the log of the RML statistic.

2.1 Cox's Method

The problem of testing separate families of hypotheses was initiated by Cox (1961). He stressed the existence of a class of problems that have not received much attention in the literature; outlined a general method for tackling this problem based on the likelihood ratio; and applied his results to a few special cases. The large sample properties were also discussed in Cox (1962) by a slightly different argument with some more comments and examples.

Cox (1961, 1962) proposed a test based on the statistic

$$T_n^C = T_n - E_{\hat{\theta}}(T_n),$$

where $E_\theta(.)$ denotes expectation with respect to $f(x;\theta)$. This statistic compares the observed difference between maximized log-likelihoods with an estimate of its expected value under H_0. Thus a large negative value of this statistic indicates departure from H_0. He showed that the statistics T_n^C is asymptotically equivalent to

$$\hat{T}_n^C = n^{-1}\sum_{i=1}^{n}\left\{(\ell_i - \ell_i^*) - E_\theta(\ell_1 - \ell_1^*) - \sum_{k,l} E_\theta[\ell_{1,k}(\ell_1 - \ell_1^*)]\delta^{kl}\ell_{i,l}\right\}$$

where

$$\ell_i = \log f(X_i;\theta), \quad \ell_{i,r} = \frac{\partial}{\partial\theta^{(r)}}\log f(X_i;\theta), \quad \ell_i^* = \log g(X_i;\lambda_\theta)$$

and $[\delta^{kl}]$ is the inverse of $[E_\theta(\ell_{1,k}\ell_{1,l})]$. The quantity λ_θ is the almost-sure limit of $\hat{\lambda}$ (the maximum likelihood estimator of λ under H_1) when H_0 holds with parameter θ (see the Appendix). Then, he showed that, under H_0, T_n^C is asymptotically normal with mean 0 and variance

$$n^{-1}\left[\operatorname{var}_\theta(\ell_1 - \ell_1^*) - \sum_{kl}\gamma_k\delta^{kl}\gamma_l\right], \tag{2.1}$$

where $\gamma_k = E_\theta[\ell_{1,k}(\ell_1 - \ell_1^*)]$.

As an alternative approach, Cox also suggested combining the two families $f(x;\theta)$ and $g(x;\lambda)$ in a more general family such that each family is a special case of this general family. The density of the general family can be taken as

$$h(x;\theta,\lambda,p) = \frac{\{f(x;\theta)\}^p\{g(x;\lambda)\}^{1-p}}{\int\{f(y;\theta)\}^p\{g(y;\lambda)\}^{1-p}\,dy}.$$

Inference about p is then made in the usual way. Atkinson (1970) adopted this approach and he derived a statistic based on

$$T_n^A = n^{-1}\left\{\ell_f(\hat{\theta}) - \ell_g(\lambda_{\hat{\theta}}) - E_{\hat{\theta}}[\ell_f(\hat{\theta}) - \ell_g(\lambda_{\hat{\theta}})]\right\}$$

where $\lambda_{\hat{\theta}}$ is the value of λ_θ evaluated at $\hat{\theta}$. Under the null hypothesis, T_n^C and T_n^A are asymptotically normally distributed with mean zero and same variance. Pereira (1977) investigated the probability limits of these statistics under the alternative hypothesis and their behavior in finite samples under the null hypothesis. He concluded that Cox's statistics is on the whole preferable. He showed that T_n^A may provide an inconsistent test and gave an example (a test of H_0 against a class of alternatives H_1 is said to be consistent if, when any member of H_1 holds, the probability of rejecting H_0 tends to one as the sample size tends to infinity). In addition, in a simulation study T_n^A always showed a better agreement for the first two moments while T_n^C always showed a better agreement for the third and fourth moments. Therefore, from practical point of view, T_n^C is generally recommended because corrections for the lower order moments are more easily obtained.

Cox did not consider regularity conditions for normality of T_n^C. Regularity conditions and a rigorous proof of the asymptotic normality of Cox's statistic were given in White (1982). We now review these regularity conditions. In the following, the conditions stated in term of f will also be understood to apply to g, and parenthetical material indicates appropriate correspondence.

(1) The independent random variables $X_1,\ldots,X_n$ have common distribution H on Ω, a measurable Euclidian space, with measurable density $h = dH/d\nu$.

(2) The distribution function $F(x;\theta)$ ($G(x;\lambda)$) has density $f(x;\theta)$ ($g(x;\lambda)$) which is measurable in x for every θ in Θ (λ in Λ), a compact subset of p-dimensional (q-dimensional) Euclidian space, and continuous in θ (λ) for every x in Ω. The minimal support of f (g) does not depend on θ (λ).

(3) (a) $|\log f(x;\theta)| \le m(x)$ for all θ in Θ, where m is integrable with respect to H; and (b) $E[\log f(X;\theta)]$ has a unique maximum in Θ.

(4) $\partial \log f(x;\theta)/\partial\theta_i$, $i = 1,\ldots,p$ are measurable functions of x for each θ in Θ and continuously differentiable functions of θ for each x in Ω.

(5) $|\partial^2 \log f(x;\theta)/\partial\theta_i\partial\theta_j|$ and $|\partial \log f(x;\theta)/\partial\theta_i . \partial \log f(x;\theta)/\partial\theta_j|$, $i, j = 1,\ldots,p$ are dominated by functions integrable with respect to H for all x in Ω and θ in Θ.

(6) The true value parameter θ_0 is interior to Θ and $A(\theta_0)$ and $B(\theta_0)$ are non-singular, where

$$A(\theta) = \left\{ E\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta_i \partial\theta_j} \right] \right\}$$

and

$$B(\theta) = \left\{ E\left[\frac{\partial \log f(X;\theta)}{\partial\theta_i} . \frac{\partial \log f(X;\theta)}{\partial\theta_j} \right] \right\}.$$

(7) $\{\log[f(x;\theta)/g(x;\lambda)]\}^2$ is dominated by a measurable function integrable with respect to H for all θ, λ in $\Theta \times \Lambda$.

(8) $|\partial \log[f(x;\theta)/g(x;\lambda)] f(x;\lambda)/\partial\theta_i|$ and $|\partial \log[f(x;\theta)/g(x;\lambda)] f(x;\lambda)/\partial\lambda_i|$, $i = 1,\ldots,p$ $j = 1,\ldots,q$ are dominated for all θ, λ in $\Theta \times \Lambda$ by functions integrable with respect to ν.

(9) $A(\theta_0) = -B(\theta_0)$ under H_0.

It should be noted that although assumption (2) requires the parameter spaces Θ (Λ) to be a compact set, we can assume that the parameter space is an open set and then consider a compact subset of it which contains $\hat{\theta}$ almost surely; see Serfling (1980, p. 144) and Fearn and Nebenzahl (1991, p. 591).

By using computer simulation, Chen (1980) showed that Cox's test should be used only when the sample size is sufficiently large. Bain and Engelhardt (1980) used T_n in choosing between a Weibull and a Gamma model. Fearn and Nebenzahl (1991) applied T_n for finding the sample size required for deciding between two overlapping families, Weibull and gamma. Gupta and Kundu (2003, 2004) used an asymptotically equivalent statistics to T_n (see Appendix) in discrimination between some distributions.

As noted in the introduction, for testing location-scale families, the distribution of RML statistic T_n does not depend on the parameters; see Dumonceaux, et al. (1973), Antle and Bain (1969) and Fisher (1934). Therefore, in this case, Cox's test is equivalent to using the statistic T_n directly. Dumonceaux, et al. (1973) applied this statistic to testing some location-scale families and obtained the critical values t_α by simulation for different sample sizes.

2.3 Williams's Method

Williams (1970) observed that the conditions for validity of Cox's test did not hold in his problem and proposed directly simulating the distribution of T_n assuming $\theta = \hat{\theta}$. That is, for sufficiently large integer B, B sets

$$\left\{(x_{1k}^*, \dots, x_{nk}^*),\ k = 1, \dots, B\right\}$$

of artificial data are drawn from the population with density $f(x^*, \hat{\theta})$. From the k th set, the maximum likelihood estimators $\hat{\theta}_k^*$ and $\hat{\lambda}_k^*$ and

$$T_{nk}^* = n^{-1} \sum_{i=1}^{n} [\log f(x_{ik}^*; \hat{\theta}_k^*) - \log g(x_{ik}^*; \hat{\lambda}_k^*)]$$