COMPARING HIGHER ORDER LIKELIHOOD INFERENCE FOR LOCATION-SCALE MODELS

YONGXIU SHE

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO
AUGUST 2013

COMPARING HIGHER ORDER LIKELIHOOD INFERENCE FOR LOCATION-SCALE MODELS

by **Yongxiu She**

a dissertation submitted to the Faculty of Graduate Studies of York University in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

COMPARING HIGHER ORDER LIKELIHOOD INFERENCE FOR LOCATION-SCALE MODELS

by **Yongxiu She**

Examination Committee Members:

1. Professor Augustine Wong
2. Professor Steven Wang
3. Professor Xin Gao

Abstract

For parametric models, the third order asymptotic theories for approximating tail probabilities are extremely accurate even for small sample size. These methods only require the likelihood function and the observed sample. Two third order asymptotic methods developed by Skovgaard in 1996 and Fraser and Reid in 1999 are compared and applied to location-scale family model in this dissertation. The Fraser and Reid method and the Skovgaard method have similar ideas except the canonical parameterization is different.

Based on the special structure of location-scale model, a simple and accurate method is developed by transforming all the scale parameters into location type parameters. However, the general formulas to calculate the confidence intervals for location or scale parameter in the Fraser and Reid method and the Skovgaard method are also derived. The Behrens-Fisher problem with an assumption that the ratio of the two variances is known which is first considered by Schechtman and Sherman (2007) is revisited. Our proposed third order methods exhibit significant advantage over some existing first order methods especially for small sample size. All of these results will be illustrated through numerical studies.

Acknowledgements

I would like to express my special appreciation and thanks to my supervisor, Professor Augustine Wong, for his guidance, continuous support, inspirations and patience. You have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on my research has been invaluable.

I am heartily thankful to my supervisory committee member, Professor Xin Gao and Professor Steven Wang, who are always generous to offer their knowledge and patience and make my doctoral study smooth and rewarding.

Financial support from the Government of Ontario and the Department of Mathematics and Statistics at York University is appreciated.

I owe my deepest gratitude to my beloved parents, my husband and my kids for their support, care and love. Without their understanding and sacrifice, this thesis would not be possible.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

Inference for scalar parameter of a parametric model is a mainstay in statistical inference and is introduced in most, if not all, introductory statistics books. Parametric likelihood inference is often based on first order approximations to standard summary statistics from the likelihood, such as the signed log-likelihood ratio statistic, the Wald statistic and the Score statistic. From these statistics, we are able to derive the p-value, as well as the confidence interval for the scalar parameter. Details for these first order approximations are reviewed in Section 1.2.3.

Although the first order methods are widely used, they generally do not give accurate approximation especially for small sample size. In recent literature, a more accurate asymptotic method was derived by Fraser and Reid in 1995. Rekkas and Wong (2008) implemented the Fraser and Reid method for fat-tailed distribution. Wong, Chang and Rekkas (2013) applied the Fraser and Reid method to time series model. Another accurate asymptotic method developed by Skovgaard in 1999 but is not as frequently mentioned in statistics literature as the Fraser and Reid method. In my thesis, I will compare the accuracy and computation efficiency of these two asymptotic methods for location-scale models. Finally, it is illustrated in this thesis that, for the Behrens-Fisher problem, the two methods give identical numerical results.

1.2 Literature Review

This section reviews some key concepts and definitions in parametric statistical inference. In Section 1.2.1, likelihood function and maximum likelihood estimation defined by Fisher (1922) are reviewed. We introduce the general terminologies and notation that will be used throughout this thesis in Section 1.2.2. In Section 1.2.3, the first order asymptotic techniques are examined. The development of saddlepoint approximation is introduced in Section 1.2.4. The Lagrange Multiplier technique is reviewed in Section 1.2.5.

1.2.1 Definitions of Likelihood Function and MLE

Assume $y = (y_1, \cdots, y_n)'$ is a random sample obtained from a population with the parameter θ, where $\theta = (\theta_1, \cdots, \theta_k)' \in \Theta$ is a k-dimensional parameter. The **Like-**

lihood Function of the sample is defined as:

$$L(\theta) = L(\theta; y_1, \cdots, y_n) = cf(y_1, \cdots, y_n; \theta)$$

for values of θ within a given domain, where $c > 0$ is a multiplicative constant, and $f(y_1, \cdots, y_n; \theta)$ is the value of the joint probability distribution function for n independent identically distributed variables or the joint probability density function of the random variables $Y_1, \cdots, Y_n$ evaluated at $Y_1 = y_1, \cdots, Y_n = y_n$ (Miller and Miller (2003)).

The **Log-Likelihood Function** is

$$\ell(\theta) = \ell(\theta; y_1, \ldots, y_n) = a + \sum_{i=1}^{n} \log(f(y_i; \theta)) \tag{1.1}$$

where $a \in \mathbb{R}$ is an additive constant. Without lost of generality, a is taken to be 0 hereafter.

It is more convenient to work with log-likelihood function. The reason is that logarithm is a monotonically increasing function, and therefore the logarithm of a function achieves its maximum value at the same point as the function itself. Moreover, it is easier to calculate the maximum value by taking the derivative of a logarithm function. For example, in many statistical applications, the likelihood function is a collection of statistically independent observations (product of independent probability density functions for continuous distribution, once taking logarithm, then it becomes the sum of individual log density function), and it is easier to compute the derivatives of sum than the derivatives of product.

In statistics, **Maximum Likelihood Estimation** is a method of estimating the parameters of a statistical model. In general, for a fixed set of data from an underlying statistical model, the method of maximum likelihood selects values of the model parameters that maximize the likelihood function.

Given a statistical model $\{f(y_1,\ldots,y_n;\theta) : \theta \in \Theta\}$ with log-likelihood function $\ell(\theta)$, the **Score function** U is defined to be the gradient of $\ell(\theta)$:

$$U(\theta) = \nabla\ell(\theta) = \frac{\partial\ell(\theta)}{\partial\theta} = \ell_\theta(\theta). \tag{1.2}$$

By setting $U(\theta)$ to 0, we have the likelihood equations:

$$U(\theta) = \left(\frac{\partial\ell(\theta)}{\partial\theta_1},\ldots,\frac{\partial\ell(\theta)}{\partial\theta_k}\right)' = (0,\ldots,0)'.$$

The maximum likelihood estimate (MLE) $\hat{\theta}$ of the parameter vector θ can usually be found by solving the likelihood equations.

1.2.2 Terminologies and Notation

The followings are the notation that will be used throughout this dissertation.

- $\theta = (\psi, \lambda')'$, parameter of model with size k;
- ψ, scalar parameter of interest;

- λ, nuisance parameter with size $k-1$;

- $\ell(\theta)$, log-likelihood function, $\ell(\theta) = \ell(\theta; y_1, \ldots, y_n) = \log f(y_1, \cdots, y_n; \theta)$; when $Y_1, \cdots, Y_n$ are independent and identically distributed random variables with probability density function $f(\cdot; \theta)$, then $\ell(\theta)$ can be simplified to $\ell(\theta) = \sum_{i=1}^{n} \log\left[f(y_i; \theta)\right]$;

- $\ell_\theta(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$, score function;

- $\ell_{\theta\theta'}(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'}$;

- $j_{\theta\theta'}(\theta) = -\ell_{\theta\theta'}(\theta)$, the observed full information matrix;

- $j_{\lambda\lambda'}(\theta) = -\ell_{\lambda\lambda'}(\theta)$, the observed nuisance information matrix;

- $i_{\theta\theta'}(\theta) = \mathrm{E}\left[j_{\theta\theta'}(\theta)\right]$, the full expected Fisher information matrix;

- $\hat{\theta}$, the overall maximum likelihood estimate which maximizes $\ell(\theta)$, i.e., $\hat{\theta}$ is obtained by solving the likelihood equations: $\ell_\theta(\hat{\theta}) = \left.\frac{\partial \ell(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$;

- $\tilde{\ell}(\theta)$, tilted log-likelihood function, $\tilde{\ell}(\theta) = \ell(\theta) + \hat{\kappa}(\psi(\theta) - \psi_0)$ where ψ_0 is a given ψ value (See Section 1.2.5);

- κ, Lagrange multiplier (See Section 1.2.5);

- $\hat{\kappa}$, the Lagrange multiplier which maximized the tilted log-likelihood function (See Section 1.2.5);

- $\hat{\theta}_\psi$, the constrained MLE which maximizes $\ell(\theta)$ for a given ψ, i.e., $\hat{\theta}_\psi$ is generally obtained by solving the estimating equation $\ell_\lambda(\hat{\theta}_\psi) = \frac{\partial \ell(\theta)}{\partial \lambda}\Big|_{\theta=\hat{\theta}_\psi} = 0$, with λ explicitly available. However, if λ is not explicitly available, we could apply Lagrange Multiplier technique (See Section 1.2.5) to obtain $\hat{\theta}_\psi$;

- $\tilde{\jmath}_{\theta\theta'} = \tilde{\jmath}_{\theta\theta'}(\hat{\theta}_\psi) = -\tilde{\ell}_{\theta\theta'}(\hat{\theta}_\psi)$;

- $\varphi = \varphi(\theta)$, canonical parameter of an exponential family model.

Moreover, the following mild regularity conditions given in Wilks (1938) are assumed to be true throughout this dissertation:

- $f(y;\theta) > 0$ is twice differentiable in a neighborhood of θ;

- $E[\ell_\theta(y;\theta)\ell'_\theta(y;\theta)]$ exists and is nonsingular;

- $\int \sup_{\theta\in N} |f_\theta(y;\theta)|dy < \infty$ and $\int \sup_{\theta\in N} |f_{\theta\theta'}(y;\theta)|dy < \infty$;

- $\int \sup_{\theta\in N} |\ell_{\theta\theta'}(y;\theta)|dy < \infty$.

1.2.3 The First Order Asymptotic Methods

Consider $y = (y_1, \cdots, y_n)'$ from a statistical model with log-likelihood function $l(\theta)$. Under the mild regularity conditions discussed in Section 1.2.2, by applying Central Limit Theorem and Taylor expansion, the followings can be obtained:

- $\ell'_\theta(\theta)\left\{\mathrm{var}[\ell_\theta(\theta)]\right\}^{-1}\ell_\theta(\theta) \xrightarrow{d} \chi^2_k$, where $\mathrm{var}(\ell_\theta(\theta)) = i_{\theta\theta'}(\theta)$ (Rao 1947);

- $(\hat\theta-\theta)'[\mathrm{var}(\hat\theta)]^{-1}(\hat\theta-\theta) \xrightarrow{d} \chi^2_k$, where $\mathrm{var}(\hat\theta) \approx i^{-1}_{\theta\theta'}(\theta)$ (Wald 1943);

- $2\left[\ell(\hat\theta)-\ell(\theta)\right] \xrightarrow{d} \chi^2_k$ (Wilks 1938).

Note that although $i_{\theta\theta'}(\theta)$ can be difficult to obtain, it can be approximated by $j_{\theta\theta'}(\hat\theta)$. Hence the following test statistics are usually used when θ is a scalar parameter:

- the Wald statistic

$$w = w(\theta) = j(\hat\theta)^{\frac{1}{2}}(\hat\theta-\theta), \tag{1.3}$$

- the signed log-likelihood ratio statistic

$$r = r(\theta) = sign(\hat\theta-\theta)\{2[l(\hat\theta)-l(\theta)]\}^{\frac{1}{2}}, \tag{1.4}$$

- the Score statistic

$$s = s(\theta) = j(\hat\theta)^{-\frac{1}{2}}l_\theta(\theta). \tag{1.5}$$

And all of them are asymptotically distributed as $N(0,1)$ (Cramer (1946, Section 33) and Lehmann (1983, Chapter 6)). Based on Taylor expansion, the accuracy of all of the three methods is $O(n^{-\frac{1}{2}})$ and they are referred to as the first order methods.

If θ is a vector parameter with ψ being the scalar parameter of interest, and λ being an explicitly known nuisance parameter, then the three test statistics become:

- the Wald statistic

$$w = w(\psi) = (var(\hat{\psi}))^{-\frac{1}{2}}(\hat{\psi} - \psi), \tag{1.6}$$

- the signed log-likelihood ratio statistic

$$r = r(\psi) = sign(\hat{\psi} - \psi)\{2[l(\hat{\theta}) - l(\hat{\theta}_{\psi})]\}^{\frac{1}{2}}, \tag{1.7}$$

- the Score statistic

$$s = s(\psi) = (var(\hat{\theta}_{\psi}))^{\frac{1}{2}} l_{\psi}(\hat{\theta}_{\psi}) \tag{1.8}$$

where

$$var(\hat{\psi}) \approx \psi_{\theta}(\hat{\theta}) j_{\theta\theta'}^{-1}(\hat{\theta}) \psi_{\theta}(\hat{\theta})'$$

$$\psi_{\theta}(\hat{\theta}) = \left.\frac{\partial\psi(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}$$

$$l_{\psi}(\hat{\theta}_{\psi}) = \left.\frac{\partial l(\psi, \lambda)}{\partial\psi}\right|_{\lambda=\hat{\lambda}_{\psi}}.$$

For first order asymptotic theory, Barndorff-Nielsen and Sorenen (1994) suggested to use the observed information rather than the expected Fisher information as the estimate of the inverse of the variance of the MLE.

The **Significance(P-value) Function** p: $\Omega \to [0,1]$ is taken to be

$$p(\theta) = p(\hat{\theta} < \hat{\theta}^0; \theta).$$

$p(\theta)$ is also called **confidence distribution function** since all possible confidence intervals can be obtained by inverting $p(\theta)$.

The p-value functions based on (1.3), (1.4) and (1.5) are defined as

$$p(\psi) = \begin{cases} \Phi(w) \\ \Phi(r) \\ \Phi(s) \end{cases} \tag{1.9}$$

where $\Phi(\cdot)$ is the cumulative distribution function of N(0,1).

A $(1-\alpha) * 100\%$ **Confidence Interval** for θ is

$$\left(min\left\{p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right)\right\}, max\left\{p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right)\right\}\right). \tag{1.10}$$

The first order asymptotic theory of likelihood-based methods is studied in many texts on statistical theories. See, for example, Cox and Hinkley (1974, Chapter 9), Sen and Singer (1993, Chapter 5), Ferguson (1996, Part 4), Schervish (1997, Chapter 7) and Severini (2000, Chapter 4). The first two approximations are more often used. Recent researchers showed that (1.4) gives more powerful and accurate test than (1.3), but (1.3) is more popular in applied statistical analysis since it is simple to apply (Neymann and Pearson (1933), Doganaksoy and Schmee (1993)).

Although the three first order methods are widely used in hypothesis testing, they do not perform well when the sample size is small or when the underlying distribution is far away from the normal distribution. Some researchers working on likelihood inference focused on how to improve the accuracy of these first order asymptotic methods. In the rest of the thesis, I will introduce some higher order asymptotic methods and compare the accuracies with those obtained by the first order methods.

1.2.4 Saddlepoint Approximation

In order to provide a better approximation to the true density function for the average of n independent and identically distributed random variables, Daniels in 1954 introduced saddlepoint method. The method is very accurate but can be extremely complicated in terms of computation.

1.2.4.1 Saddlepoint Approximation for Mean

Let $Y_1, \ldots, Y_n$ be independent, identically distributed random vectors from a model with density $f_Y(\cdot;\theta)$. The moment generating function and cumulant generating function are defined by

$$M(t) = E[e^{t'Y}]$$

$$K(t) = \log(M(t))$$

respectively.

The saddlepoint approximation for the density of the mean of n independent, identically distributed random variables, $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, is given by Daniels (1954):

$$f_{\bar{Y}}(\bar{y}) = (2\pi)^{-k/2}\left\{\frac{n}{\left|\mathrm{K}_{tt'}(\hat{t})\right|}\right\}^{1/2} \exp\left[n\{\mathrm{K}(\hat{t}) - \hat{t}'\bar{y}\}\right]\{1 + O(n^{-1})\} \qquad (1.11)$$

where

$$K_t(t) = \frac{\partial K(t)}{\partial t}$$
$$K_{tt'}(t) = \frac{\partial K^2(t)}{\partial t \partial t'}$$

and $\hat{t}$ is known as the **saddlepoint** and is defined by

$$K_t(\hat{t}) = \bar{y}.$$

One way to derive (1.11) is by combining the exponential tilting and the Edgeworth expansion techniques. For detailed review of the derivation, please refer to Barndorff-Nielsen (1979), Barndorff-Nielsen and Cox (1979, 1989, 1994) and Reid (1988). Note that (1.11) has a relative error of $O(n^{-1})$.

Durbin (1980) showed that if $(2\pi)^{-k/2}$ is replaced by a re-normalizing constant a, the error term of the saddlepoint approximation in (1.11) is reduced to $O(n^{-3/2})$. Then the saddlepoint approximation for the density of the mean becomes

$$f_{\bar{Y}}(\bar{y}) = a\left\{\frac{n}{\left|\mathrm{K}_{tt'}(\hat{t})\right|}\right\}^{1/2} \exp\left[n\{\mathrm{K}(\hat{t}) - \hat{t}'\bar{y}\}\right]\{1 + O(n^{-3/2})\}. \qquad (1.12)$$

In general, a has to be obtained numerically.

1.2.4.2 Saddlepoint Approximation for MLE

There is another important development from the saddlepoint approximation which is an approximation for the density of the maximum likelihood estimate. Suppose $y = (y_1, \dots, y_n)'$ is a random sample from exponential family model with density

$$f_Y(y;\theta) = \exp\{\theta' y - a(\theta) - b(y)\}.$$

From the above model, the log-likelihood function can be written as

$$l(\theta) = n\theta'\bar{y} - na(\theta).$$

The cumulant generating function is

$$K(t) = \log(M(t)) = a(\theta + t) - a(\theta).$$

For the above exponential model, there exists a one-to-one correspondence between the minimum sufficient statistic $\bar{Y}$ and the maximum likelihood estimate $\hat{\theta}$

$$K_t(\hat{\theta}) = \bar{Y}.$$

From these, Barndorff-Nielsen (1980, 1983) showed that the saddlepoint approximation of the density function of $\hat{\theta}$ is

$$f(\hat{\theta};\theta) = c|j_{\theta\theta'}(\hat{\theta})|^{1/2} \exp\left\{\ell(\theta) - \ell(\hat{\theta})\right\} \{1 + O(n^{-3/2})\} \tag{1.13}$$

where c is a normalized constant which is generally obtained numerically. Note that (1.13) has a relative error of $O(n^{-3/2})$.

For a general model, if an ancillary statistic is available, there is one-to-one correspondence between the minimum sufficient statistic and an ancillary statistic. The construction of the ancillary statistic is discussed in Barndorff-Nielsen (1980). For a general model, the conditional distribution of the maximum likelihood estimate $\hat{\theta}$ given the ancillary statistic A is derived by Barndorff-Nielsen and Cox (1984)

$$f(\hat{\theta}|A;\theta) = a(\theta, A)|j_{\theta\theta'}(\hat{\theta})|^{1/2} \exp\left\{\ell(\theta) - \ell(\hat{\theta})\right\} \{1 + O(n^{-3/2})\}\,. \tag{1.14}$$

The above approximation has a relative accuracy of $O(n^{-3/2})$. Detailed discussion of the saddlepoint method and its application in statistics can be found in Barndorff-Nielsen (1983), McCullagh (1987), Fraser (1988), Reid (1988) and Barndorff- Nielsen and Cox (1989). The main concern about this approach is that the ancillary statistic may not always exists nor unique.

1.2.5 Lagrange Multiplier Technique

Lagrange Multiplier is a method in mathematical optimization which provides a strategy for finding the local extrema of a function subject to equality constraints.

For instance, consider the optimization problem: maximize $f(y)$ subject to $g(y) = c$. We introduce a new variable κ, called **Lagrange Multiplier** and study the **Lagrange function** defined by

$$H(y, \kappa) = f(y) + \kappa(g(y) - c).$$

Let $(\hat{y}, \hat{\kappa})$ satisfies

$$\left.\frac{\partial H(y,\kappa)}{\partial y}\right|_{(\hat{y},\hat{\kappa})} = 0,$$

and

$$\left.\frac{\partial H(y,\kappa)}{\partial \kappa}\right|_{(\hat{y},\hat{\kappa})} = 0.$$

Then $(\hat{y}, f(\hat{y}))$ is the extrema of $f(y)$ subject to $g(y) = c$. For detailed review of Lagrange Multiplier technique, please refer to Lang (1973).

In our case, we want to maximize $\ell(\theta)$ subject to $\psi(\theta) = \psi_0$, the Lagrange function is

$$H(\theta, \kappa) = \ell(\theta) + \kappa(\psi(\theta) - \psi_0),$$

then $(\hat{\theta}_\psi, \hat{\kappa})$ satisfies

$$\left.\frac{\partial H(\theta,\kappa)}{\partial \theta}\right|_{(\hat{\theta}_\psi,\hat{\kappa})} = 0,$$

$$\left.\frac{\partial H(\theta,\kappa)}{\partial \kappa}\right|_{(\hat{\theta}_\psi,\hat{\kappa})} = 0.$$

The tilted log-likelihood function is defined as

$$\tilde{\ell}(\theta) = \ell(\theta) + \hat{\kappa}(\psi(\theta) - \psi_0).$$

The tilted log-likelihood is useful whenever the nuisance parameter is not explicitly available or does not exist in closed form. Note that, for a given constraint $\psi(\theta) = \psi_0$, $\tilde{\ell}(\hat{\theta}_\psi) = \ell(\hat{\theta}_\psi)$, i.e., the tilted likelihood has the same maximum as the original likelihood at the constrained MLE.

1.3 Summary

We briefly reviewed the three first order methods. As it is shown in the following two Chapters, the first order methods are pretty easy to use but not that accurate, especially for small sample size. We also reviewed the saddlepoint method.

Barndorff-Nielsen and Cox (1979) highlighted the usefulness and accuracy of the saddlepoint approximation of the density function of the mean of n independent identically distributed random variables. They also pointed out that to approximate the cumulative distribution function of $\bar{Y}$, generally, does not have closed form and hence, numerically integrating the saddlepoint density is required. Higher order methods will be introduced in the following Chapter. We will see how accurate they are especially for small sample size.

2 Higher Order Likelihood Asymptotic

2.1 Methodology

Followed by the idea of saddlepoint approximation and using complex analysis, Lugannani and Rice (1980) calculated the cumulative distribution function of $\bar{Y}$, the mean of n independent identically distributed random variables:

$$F_{\bar{Y}}(\bar{y}) = P(\bar{Y} \leq \bar{y}) = \left[\Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{Q}\right)\right]\left(1 + O(n^{-3/2})\right), \qquad (2.1)$$

where

$$\begin{aligned} r &= \operatorname{sgn}(\hat{t})\sqrt{2n[\hat{t}\bar{y} - \mathrm{K}(\hat{t})]} && (2.2) \\ Q &= \hat{t}\sqrt{nK_{tt'}(\hat{t})}. && (2.3) \end{aligned}$$

Note that $\hat{t}$ is the saddlepoint as defined in Chapter 1 satisfied $K_t(\hat{t}) = \bar{y}$, $K(t)$ is the cumulant generating function and $K_t(\hat{t})$ and $K_{tt'}(\hat{t})$ are the first and second derivatives of the cumulant generating function with respect to t evaluated at the saddlepoint. Barndorff-Nielsen (1986) derived alternative approximation that incorporates the correction term into the quantile of the normal cumulative distribution:

$$F_{\bar{Y}}(\bar{y}) = P(\bar{Y} \le \bar{y}) = \Phi(r^*)\left(1 + O(n^{-3/2})\right) \tag{2.4}$$

where

$$r^* = r - \frac{1}{r}\log\frac{r}{Q}. \tag{2.5}$$

Assume $(y_1, \ldots, y_n)$ is a random sample obtained from the natural exponential family with probability density function

$$f(y;\theta) = \exp\{\theta y - c(\theta)\}\, h(y)$$

where θ is a scalar canonical parameter, $c(\theta)$ and $h(y)$ are known functions. The moment generating function and the cumulant generating function are

$$\begin{aligned} M(t) &= \exp\{c(t+\theta) - c(\theta)\} \\ K(t) &= c(t+\theta) - c(\theta) \end{aligned}$$

respectively. The log-likelihood function is

$$\begin{aligned} l(\theta) &= \log \prod_{i=1}^{n} f(y_i;\theta) \\ &= \theta \sum_{i=1}^{n} y_i - nc(\theta). \end{aligned}$$

According to the definition of MLE in the previous Chapter, we have

$$\ell_\theta(\theta)\bigg|_{\theta=\hat{\theta}} = \sum_{i=1}^{n} y_i - nc_\theta(\theta)\bigg|_{\theta=\hat{\theta}} = 0$$

i.e.

$$c_\theta(\theta)\bigg|_{\theta=\hat\theta} = \frac{\sum_{i=1}^n y_i}{n} = \bar y.$$

From the definition of saddlepoint in Chapter 1, we have

$$K_t(\hat t) = c_t(t+\theta)\bigg|_{t=\hat t} = \bar y.$$

It is equivalent to say

$$\frac{dc(\theta)}{d\theta}\bigg|_{\theta=\hat\theta} = \frac{dc(t+\theta)}{ct}\bigg|_{t=\hat t}.$$

This implies that $\hat\theta = \hat t + \theta$. Therefore,

$$\begin{aligned} K(\hat t) &= c(\hat t+\theta) - c(\theta) \\ &= c(\hat\theta) - c(\theta). \end{aligned}$$

Thus r and Q in equations (2.2) and (2.3) have simple forms for natural exponential family case

$$\begin{aligned} r &= \operatorname{sgn}(\hat\theta-\theta)\left\{2[l(\hat\theta)-l(\theta)]\right\}^{1/2} && (2.6) \\ Q &= (\hat\theta-\theta)j_{\theta\theta'}^{1/2}(\hat\theta). && (2.7) \end{aligned}$$

Note r and Q coincide with the signed log-likelihood ratio statistic and the standardize maximum likelihood departure for a scalar parameter situation. Hence, the p-value function of parameter θ with relative error $O(n^{-3/2})$ approximated by Lugannani-Rice (1980) and Barndorff-Nielsen (1986) formulas are

$$p(\theta) = \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{Q}\right) \tag{2.8}$$

$$p(\theta) = \Phi(r^*) \tag{2.9}$$

respectively.

Fraser (1990) showed that these two methods are equivalent up to third order accuracy. Both (2.8) and (2.9) are referred to as third order methods. For exponential family models and transformation models, (2.8) and (2.9) could be obtained easily. However for general models, Q could be difficult or impossible to obtain since it is based on the existence of ancillary statistic. In practice, there is no accessible procedure available for the construction of an ancillary in a general context. In this Chapter, asymptotic approaches developed by Fraser and Reid in 1999 (FR method), and Skovgaard in 1996 and 2001 will be discussed and compared.

2.2 Fraser and Reid Method (1999)

This Section details the mechanics developed by Fraser and Reid in 1999 of the likelihood based third order method for the natural exponential model (canonical exponential model), then extends the methodology to general statistical model (Fraser (1990)).

2.2.1 Natural Exponential Model

Consider a natural exponential family model

$$f(y;\theta) = \exp\left\{\theta' y - c(\theta)\right\} h(y)$$

where $\theta = (\psi, \lambda')'$ is the canonical parameter with ψ being the scalar parameter of interest. It is easy to see that y is a sufficient statistic. For any random sample from the above model, the sign log-likelihood ratio statistic is:

$$r = r(\psi) = \text{sgn}(\hat{\psi} - \psi)\left\{2[l(\hat{\theta}) - l(\hat{\theta}_{\psi})]\right\}^{1/2}. \tag{2.10}$$

By using the the sequential saddlepoint procedure (Fraser, Reid and Wong (1991)) and taken into consideration of eliminating the nuisance parameter λ, a measure of the standardized maximum likelihood estimate departure calculated in the canonical parameter space is

$$Q = Q(\psi) = (\hat{\psi} - \psi) \left\{ \frac{\left| j_{\theta\theta'}(\hat{\theta}) \right|}{\left| j_{\lambda\lambda'}(\hat{\theta}_{\psi}) \right|} \right\}^{1/2}, \tag{2.11}$$

where $j_{\theta\theta'}(\hat{\theta})$ is the observed overall information matrix evaluated at the overall MLE $\hat{\theta}$ and $j_{\lambda\lambda'}(\hat{\theta}_{\psi})$ is the observed nuisance information matrix evaluated at the constrained MLE $\hat{\theta}_{\psi}$. Detailed derivation of (2.11) is discussed in Fraser, Reid and Wong (1991).

Hence, the p-value function, $p(\psi)$, can be obtained by using either the Lugannani-Rice (LR) approximation (2.8) or the Barndorff-Nielsen (BN) approximation (2.9) with r and q defined as above. And the third order confidence interval of ψ can be obtained from (1.10).

2.2.1.1 *Example*: Approximation to Gamma distribution

To illustrate the application and accuracy of the third order method for natural exponential family, let us examine the following example. Suppose $(y_1, \cdots, y_n)$ is a sample from exponential distribution with density function

$$f(y;\theta) = \theta \exp\{-y\theta\}, y > 0, \theta > 0.$$

Then

$$l(\theta) = n\log(\theta) - \sum_{i=1}^{n} y_i\theta.$$

To obtain $\hat{\theta}$, the first derivative of the likelihood function is:

$$l_\theta(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} y_i$$

and by equating it to be 0, we have

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} y_i}.$$

Since

$$\begin{aligned} l_{\theta\theta'}(\theta) &= \frac{\partial l(\theta)}{\partial\theta\partial\theta'} = -\frac{n}{\theta^2} \\ j_{\theta\theta'}(\theta) &= -l_{\theta\theta'}(\theta) = \frac{n}{\theta^2}, \end{aligned}$$

the signed log-likelihood ratio statistic and the standardized maximum likelihood estimate departure given in (2.6) and (2.7) can be easily obtained

$$\begin{aligned} r &= \operatorname{sgn}(\hat{\theta} - \theta) \left\{ 2[l(\hat{\theta}) - l(\theta)] \right\}^{1/2} \\ Q &= (\hat{\theta} - \theta) \left(\frac{n}{\hat{\theta}^2} \right)^{1/2}. \end{aligned}$$

Hence, we can approximate p-value function for θ by (2.8) or (2.9).

Moreover if $(y_1, \cdots, y_n)$ is sample from exponential distribution model, it is equivalent to say that the sample is from (Gamma$(\theta, 1)$). Due to the additivity property for the Gamma distribution, $T = \sum_{i=1}^{n} Y_i \sim$ Gamma(θ, n). From distribution theory, the exact p-value function for θ is

$$p(\theta) = 1 - P(T \leq t; \theta).$$

To compare the accuracy of our proposed methods, let us consider three simulated data sets from exponential distribution with rate 3 and sample size to be 1, 10 and 100. The data sets are given in Table 2.1.

Table 2.2 shows the exact 90% confidence interval, two first order 90% confidence intervals and two third order 90% confidence intervals. Figure 2.1, 2.2 and 2.3 show the p-value functions obtained from the exact distribution, $\Phi(r)$, $\Phi(Q)$, (2.8) and (2.9), labelled as Exact, r, Wald, LR and BN respectively for the three data sets. The two horizontal lines indicate the nominal levels, 0.05 and 0.95 for the 90% confidence interval of the three data sets.

It is obvious that the third order methods give us remarkable accuracy comparing to the first order methods. And we can see the outstanding performance of the third order methods especially when the sample size is extremely small.

Table 2.1: Three simulated data sets

Data Set	Sample Size n	Observations y
1	1	0.6218
2	10	0.1349, 0.0489, 0.5769, 0.0298, 0.2223, 0.3581, 0.5039, 0.4381, 0.0522, 0.2484
3	100	0.4144, 0.2246, 0.5301, 0.3607, 0.2655, 0.4818, 1.4973, 0.5678, 0.2068, 0.1188, 0.2296, 0.2775, 0.1115, 0.5300, 0.0266, 0.1837, 0.3632, 0.0013, 0.2069, 1.6207, 0.0636, 0.1906, 0.1419, 0.0601, 0.0897, 0.2463, 0.2653, 0.3116, 1.0407, 0.6212, 0.6178, 0.1703, 0.2818, 0.0930, 0.3652, 0.1147, 1.3563, 0.4325, 0.6620, 0.0748, 0.1555, 0.3511, 0.0878, 0.2765, 0.8559, 0.7396, 0.4226, 0.2803, 0.0639, 0.0341, 0.4719, 1.3479, 0.1575, 0.3955, 0.0496, 0.2169, 0.2091, 0.0255, 0.0460, 0.9841, 0.5178, 0.1715, 0.0419, 0.3045, 0.1291, 0.3456, 0.6784, 0.0053, 0.3189, 0.1104, 0.0224, 0.2119, 0.6228, 0.4304, 0.6046, 0.2783, 0.0863, 0.2978, 0.3854, 0.3133, 0.2731, 0.3787, 0.0910, 0.2766, 0.0183, 0.2871, 0.2269, 0.0104, 0.2936, 0.1007, 0.8131, 0.0031, 0.8989, 0.2401, 0.0896, 0.3636, 0.4987, 0.1003, 0.0293, 0.1321

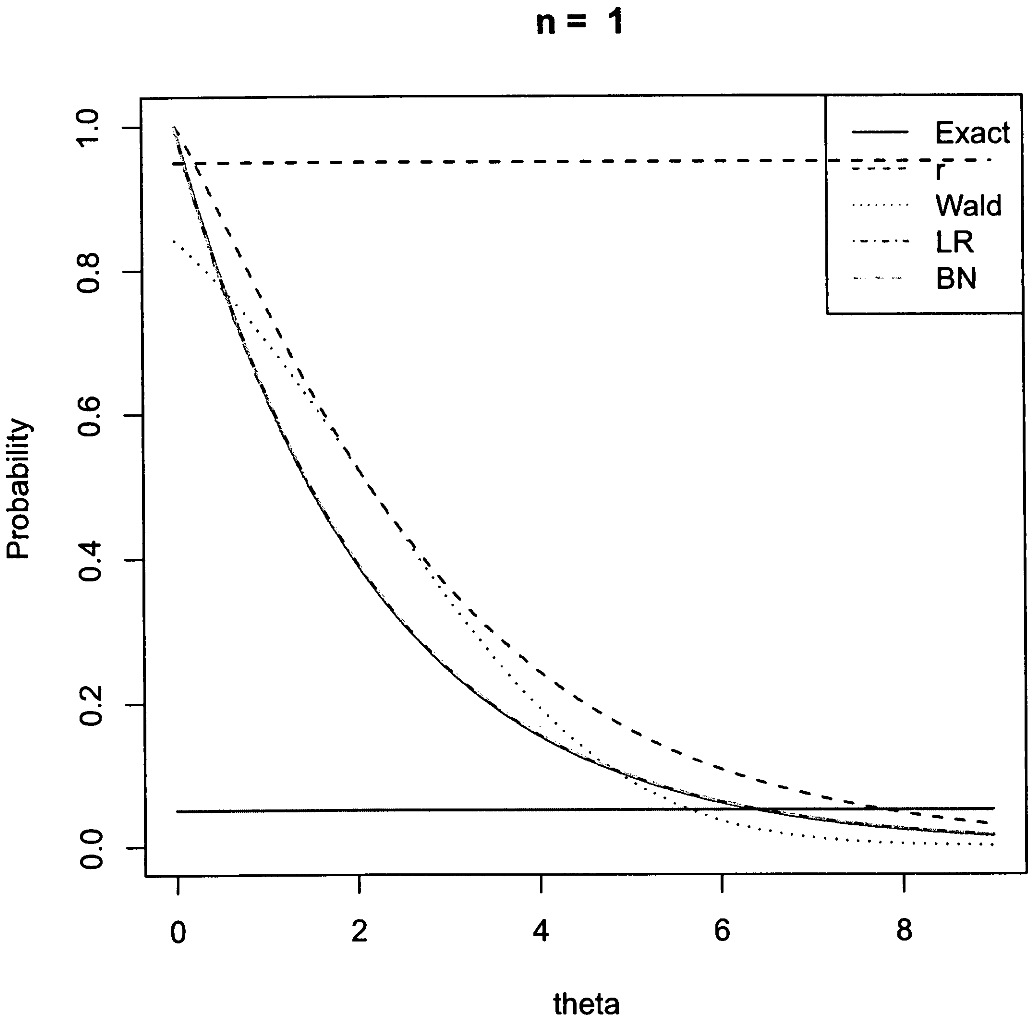


Figure 2.1: using Data set 3 in Table 2.1

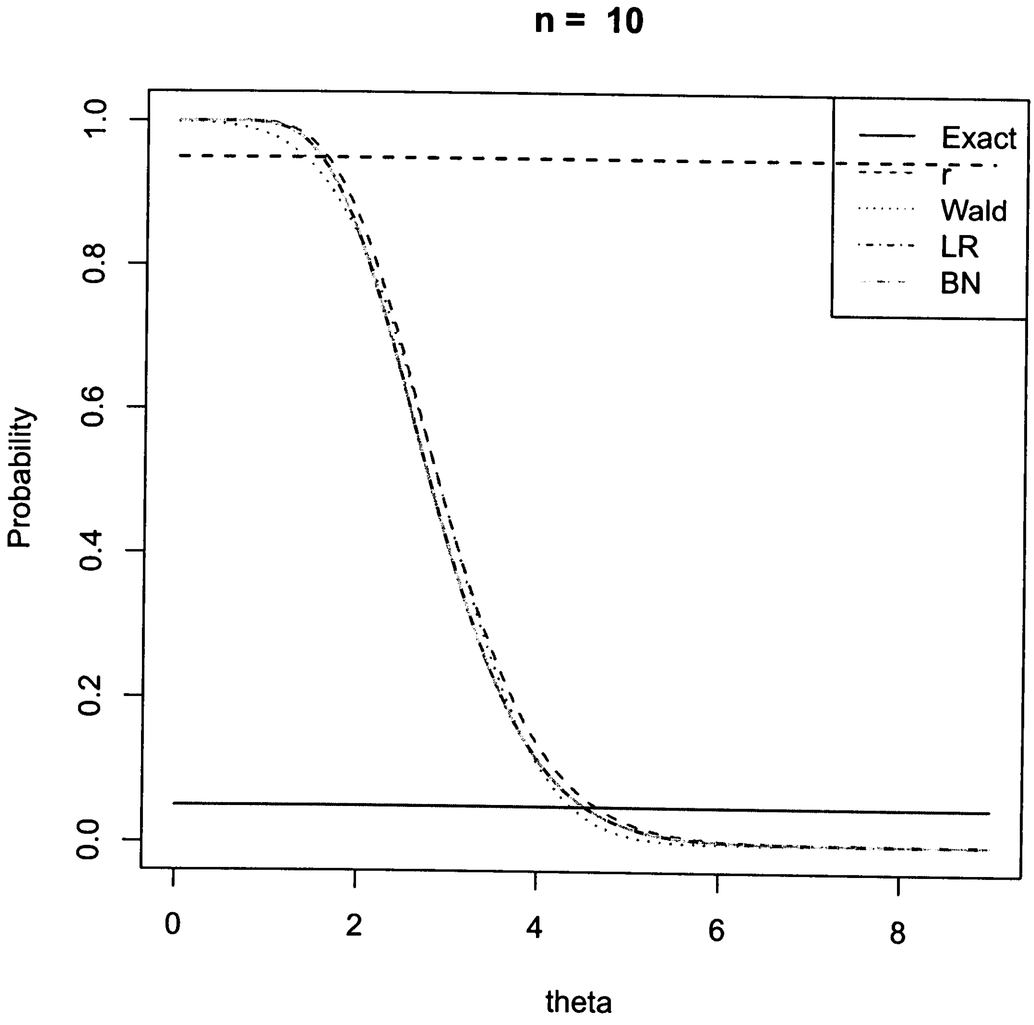


Figure 2.2: using Data set 2 in Table 2.1

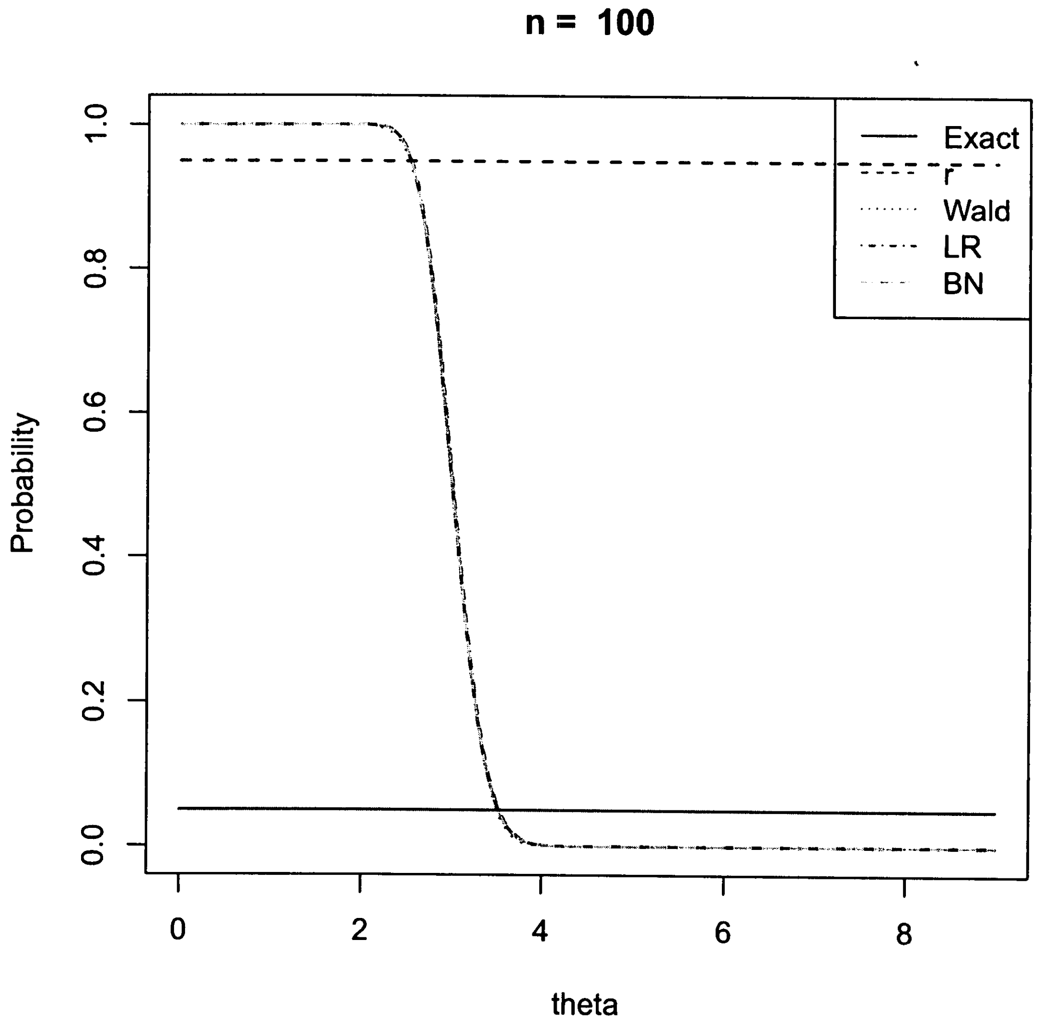


Figure 2.3: using Data set 1 in Table 2.1

Table 2.2: 90% confidence interval

Method	$n = 1$	$n = 10$	$n = 100$
Exact	(0.0803, 4.8203)	(2.0803, 6.0103)	(2.5003, 3.4803)
Wald	(0.0000, 4.2503)	(1.8403, 5.8103)	(2.4903, 3.4703)
r	(0.1603, 5.8603)	(2.1603, 6.1703)	(2.5103, 3.4903)
LR	(0.0803, 4.8303)	(2.0803, 6.0103)	(2.5003, 3.4803)
BN	(0.0903, 4.8403)	(2.0803, 6.0103)	(2.5003, 3.4803)

2.2.2 General Exponential Model

Consider a general exponential model

$$f(y;\theta) = \exp\left\{\varphi'(\theta)t(y) - c(\theta)\right\} h(y)$$

where $\varphi(\theta)$ and $t(y)$ are the canonical parameter and the canonical variable. Let $\theta = (\psi, \lambda')'$ where the scalar parameter of interest is $\psi(\theta) = \psi$ and λ is nuisance parameter vector.

Note that a vector exponential family is said to be curved if the dimension of $\theta = (\theta_1, \cdots, \theta_k)'$ is less than the dimension of the vector $\varphi(\theta) = (\varphi_1(\theta), \cdots, \varphi_s(\theta))'$. That is, if the dimension of the parameter vector is less than the number of functions of the parameter vector in the above representation of the probability density function.

Since most common distributions in the exponential family are not curved, and many algorithms designed to work with any member of the exponential family implicitly or explicitly assume that the distribution is not curved, we restrict our attention to full ranked exponential family model, i.e., the rank of our canonical parameter and the rank of the natural parameter are the same. For more details about the curved exponential family model, please see Barndorff- Nielsen (1978).

The signed log-likelihood ratio statistic, $r = r(\psi)$, is invariant to the reparameterization and is

$$r = r(\psi) = \text{sgn}(\hat{\psi} - \psi)\left\{2[l(\hat{\theta}) - l(\hat{\theta}_{\psi})]\right\}^{1/2}$$

which is the same as in (2.10). Therefore the only thing we need to focus on is how to express $Q = Q(\psi)$ in the canonical parameter, $\varphi(\theta)$, scale.

For the general exponential set up above,

$$\ell(\theta; y) = \varphi'(\theta)t(y) - c(\theta).$$

The reparameterization $\varphi = \varphi(\theta)$ typically does not have the parameter of interest ψ as a separate component, so it is necessary to extract a linear surrogate for ψ from the new parameter φ. We obtain this by constructing a scalar parameter $\chi(\theta)$ that is an orthogonal combination of the coordinates of $\varphi(\theta)$ and is tangential to $\psi(\theta)$ at

$\hat{\theta}_\psi$:

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat{\theta}_\psi)}{|\psi_{\varphi'}(\hat{\theta}_\psi)|}\varphi(\theta).$$

The gradient $\psi_{\varphi'}(\theta)$ of $\psi(\theta)$ with respect to $\varphi(\theta)$ is calculated as

$$\psi_{\varphi'}(\theta) = \left\{\frac{\partial\psi(\theta)}{\partial\theta'}\right\}\left\{\frac{\partial\varphi(\theta)}{\partial\theta'}\right\}^{-1} = \psi_{\theta'}(\theta)\varphi_{\theta'}^{-1}(\theta)$$

and is evaluated at the constrained maximum likelihood value $\hat{\theta}_\psi$. This scalar parameter $\chi(\theta)$ operates as a canonical parameter in a one-dimensional marginal model used to access the value ψ.

For simplicity in terms of calculation, we have the formula for the scalar parameter of interest in $\varphi(\theta)$ scale

$$\chi(\theta) = \frac{\psi_{\theta'}(\hat{\theta}_\psi)\varphi_{\theta'}^{-1}(\hat{\theta}_\psi)}{|\psi_{\theta'}(\hat{\theta}_\psi)\varphi_{\theta'}^{-1}(\hat{\theta}_\psi)|}\varphi(\theta). \tag{2.12}$$

Basically, the calibrated version, $\chi(\theta)$ of $\varphi(\theta)$ is a vector from the space spanned by the columns the $\varphi(\theta)$ and its direction depends on the constrained MLE for given $\varphi(\theta)$. Hence $|\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)|$ is a measure of departure of $\hat{\psi}$ from ψ in $\varphi(\theta)$ scale.

Fraser, Reid and Wu (1999) obtained an estimated variance for $\left(\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)\right)$ in $\varphi(\theta)$ scale:

$$\widehat{\text{var}}(\chi(\hat{\theta}) - \chi(\hat{\theta}_\psi)) = \frac{|j_{(\lambda\lambda')}(\hat{\theta}_\psi)|}{|j_{(\theta\theta')}(\hat{\theta})|}.$$

The full nuisance information determinant is recalibrated on the $\varphi(\theta)$ scale:

$$|j_{(\theta\theta')}(\hat{\theta})| = |j_{\theta\theta'}(\hat{\theta})||\varphi_{\theta'}(\hat{\theta})|^{-2}$$

and nuisance information defined on the canonical parameter space

$$|j_{(\lambda\lambda')}(\hat{\theta}_{\psi})| = |j_{\lambda\lambda'}(\hat{\theta}_{\psi})||\varphi_{\lambda'}(\hat{\theta}_{\psi})\varphi'_{\lambda'}(\hat{\theta}_{\psi})|^{-1}.$$

The quantity Q is then a standardized maximum likelihood departure in the surrogate parameterization $\chi(\theta)$:

$$Q = sign(\hat{\psi} - \psi)\frac{|\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})|}{\left\{\widehat{\text{var}}(\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi}))\right\}^{\frac{1}{2}}}.$$

Combine the calculation above, we have the following formula to calculate Q:

$$Q = sign(\hat{\psi} - \psi)|\chi(\hat{\theta}) - \chi(\hat{\theta}_{\psi})|\left\{\frac{|j_{\lambda\lambda'}(\hat{\theta}_{\psi})||\varphi_{\lambda'}(\hat{\theta}_{\psi})\varphi'_{\lambda'}(\hat{\theta}_{\psi})|^{-1}}{|j_{\theta\theta'}(\hat{\theta})||\varphi_{\theta'}(\hat{\theta})|^{-2}}\right\}^{-\frac{1}{2}}. \tag{2.13}$$

Or from simple calculation,

$$Q = sign(\hat{\psi} - \psi)\frac{|\varphi(\hat{\theta}) - \varphi(\hat{\theta}_{\psi}), \varphi_{\lambda}(\hat{\theta}_{\psi})|}{|\varphi_{\theta}(\hat{\theta})|}\frac{|j_{\theta\theta'}(\hat{\theta})|^{1/2}}{|j_{\lambda\lambda'}(\hat{\theta}_{\psi})|^{1/2}}. \tag{2.14}$$

Therefore, the p-value function for ψ can be obtained by using either the Lugannani-Rice approximation (2.8) or the Barndorff-Nielsen approximation (2.9) with r and Q being defined in (2.10) and (2.13), respectively. Thus a $(1 - \alpha) \times 100\%$ confidence interval for ψ can be obtained.

2.2.2.1 Approximating the Cumulative Distribution Function of the t Distribution

In the above Section, the p-value function for a scalar parameter of interest from a general exponential family model was obtained. We now consider a random sample $x = (x_1, \dots, x_n)$ from a normal distribution with mean μ and variance σ^2. If the parameter of interest is the mean parameter, $\psi(\theta) = \mu$, it is well known that the exact p-value function of μ is

$$p(\mu) = F_{t_{n-1}}(t),$$

where $F_{t_{n-1}}()$ is the cumulative distribution function of the Student t distribution with $(n-1)$ degrees of freedom and

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

By applying the Fraser and Reid method, this model is an exponential family model with log-likelihood function given by

$$\ell(\theta) = \ell(\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i - \mu)^2.$$

A convenient version of the canonical parameter is

$$\varphi(\theta) = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right)'.$$

It is easy to obtain the overall maximum likelihood estimation of θ:

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)' = (\bar{x}, (n-1)s^2/n)' = \left(\frac{\sum x_i}{n}, \frac{\sum (x_i - \bar{x})^2}{n}\right)' \quad \text{and} \quad |j_{\theta\theta'}(\hat{\theta})| = \frac{n^2}{2\hat{\sigma}^6}$$

and the constrained maximum likelihood estimation of θ:

$$\hat{\theta}_\mu = (\mu, \hat{\sigma}^2_\mu)' = \left(\mu, \frac{\sum (x_i - \mu)^2}{n}\right)' \quad \text{and} \quad |j_{\sigma^2\sigma^2}(\hat{\theta}_\mu)| = \frac{n}{2\hat{\sigma}^4_\mu}.$$

The signed log-likelihood ratio statistic can then be simplified to

$$r(\mu) = \operatorname{sgn}(\bar{x} - \mu)\left\{n \log\left(1 + \frac{t^2}{n-1}\right)\right\}^{1/2}. \tag{2.15}$$

Moreover, with

$$\varphi_\theta(\theta) = (\varphi_\mu(\theta), \varphi_{\sigma^2}(\theta)) = \begin{pmatrix} 0 & -\frac{1}{\sigma^4} \\ \frac{1}{\sigma^2} & -\frac{\mu}{\sigma^4} \end{pmatrix},$$

$Q(\mu)$ can be simplified to

$$Q(\mu) = \sqrt{n(n-1)}\left[\frac{t}{(n-1) + t^2}\right] \tag{2.16}$$

and $r^*(\mu)$ can be obtained from (2.5). Thus the p-value function of μ, or equivalently the cumulative distribution function of the Student t distribution with $(n-1)$ degrees of freedom can be approximated by $\Phi(r^*(\mu))$.

Finally by re-indexing the above result, the cumulative distribution function for the Student t distribution with ν degrees of freedom can be approximated by using the Barndorff-Nielsen formula

$$F_{t_\nu}(t) = \Phi\left(r - \frac{1}{r}\log\frac{r}{Q}\right),$$

or by using the asymptotically equivalent Lugannani and Rice formula

$$F_{t_\nu}(t) = \Phi(r) + \phi(r)\left\{\frac{1}{r} - \frac{1}{Q}\right\},$$

where

$$r = \mathrm{sgn}(t)\left\{(\nu+1)\log\left(1+\frac{t^2}{\nu}\right)\right\}^{1/2} \quad \text{and} \quad Q = \sqrt{\nu(\nu+1)}\left(\frac{t}{\nu+t^2}\right), \tag{2.17}$$

with $O(n^{-3/2})$ accuracy.

In the Fraser and Reid method for natural exponential model Section, we compare the Fraser and Reid method with the traditional first order methods. Here to illustrate the accuracy of our proposed method, we compare it with some recent approximations. Jing, Shao and Zhou (2004) applied the saddlepoint method without using the moment generating functions to approximate the cumulative distribution function of the Student t distribution function. They provide numerical comparison of their approximations with the exact Student t distribution with 5 degrees of freedom and get pretty accurate result. But the exact form of their result is very complicated. Table 1 contains the results from Jing, Shao and Zhou (2004) and the results from our approximations using both the Barndorff-Nielsen (BN) and Lugannani and Rice (LR) formulas. In Figure 2.4, we plot the relative errors of the three approximations. From Table 2.3 and Figure 2.4, we observe that the Jing, Shao and Zhou's method and our method are almost indistinguishable around the center of the distribution, but our approximations are much better towards the tail of the distribution which is crucial for inference purposes. In Figure 2.5 we plot our proposed approximations for the extreme case of the Student t distribution with 1 degree of freedom. Even for this extreme case, our approximations give remarkably accurate approximations, especially so using the Lugannani and Rice approximation. But it is obvious that Jing, Shao and Zhou's method is not accurate in the extreme case.

Table 2.3: Comparisons for $1 - F_{t_5}(t)$

t	Exact	Jing	BN	LR
0.1001	0.4620	0.4621	0.4618	0.4618
0.2010	0.4243	0.4244	0.4238	0.4238
0.3034	0.3869	0.3872	0.3861	0.3861
0.4082	0.3500	0.3505	0.3489	0.3490
0.5164	0.3138	0.3146	0.3125	0.3126
0.6290	0.2785	0.2797	0.2771	0.2771
0.7473	0.2443	0.2460	0.2427	0.2427
0.8729	0.2113	0.2136	0.2097	0.2097
1.0078	0.1799	0.1829	0.1782	0.1783
1.1547	0.1502	0.1539	0.1485	0.1486
1.3171	0.1225	0.1268	0.1208	0.1209
1.5000	0.0970	0.1010	0.0954	0.0955
1.7107	0.0739	0.0793	0.0725	0.0727
1.9604	0.0536	0.0592	0.0524	0.0525
2.2678	0.0363	0.0417	0.0353	0.0355
2.6667	0.0223	0.0271	0.0215	0.0217
3.2271	0.0116	0.0154	0.0112	0.0113
4.1295	0.0045	0.0070	0.0043	0.0044

Figure 2.4: Relative error for approximations to 1-$F_{t_5}(t)$

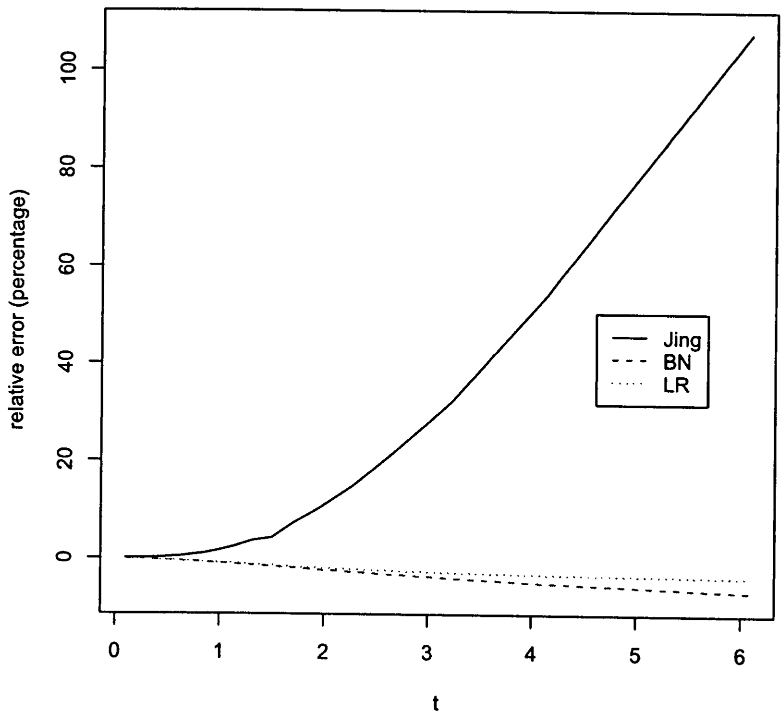
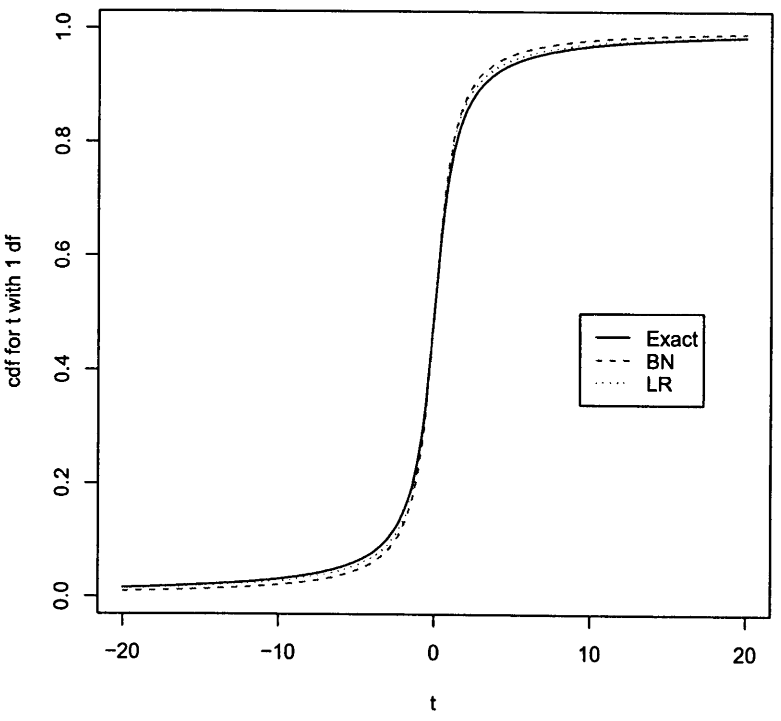


Figure 2.5: Approximations to $F_{t_1}(t)$



2.2.2.2 Approximating the Cumulative Distributions Functions of the χ^2 Distribution

Consider a random sample $x = (x_1, \ldots, x_n)$ from the normal distribution with mean 0 and variance σ^2 for which the parameter of interest is the variance parameter, $\psi(\theta) = \sigma^2$. A convenient canonical parameter is $\varphi(\theta) = 1/\sigma^2$. The exact p-value function of σ^2 is given by

$$p(\sigma^2) = F_{\chi^2_n}(x^2),$$

where $F_{\chi^2_n}()$ is the cumulative distribution function of the χ^2 distribution with n degrees of freedom and

$$x^2 = \frac{\sum_{i=1}^n x_i^2/n}{\sigma^2}.$$

By applying the Fraser and Reid method, we can obtain

$$\begin{aligned} r &= \operatorname{sgn}(x^2 - n)\sqrt{(x^2 - n) + n \log \frac{n}{x^2}} && (2.18) \\ Q &= \frac{x^2 - n}{\sqrt{2n}} && (2.19) \end{aligned}$$

and r^* can be obtained from (2.5). Thus the p-value function of σ^2, or equivalently the cumulative distribution function of the χ^2 distribution with n degrees of freedom can be approximated by $\Phi(r^*)$.

In other words, the cumulative distribution function of the χ^2 distribution with ν degrees of freedom can be approximated by using the Barndorff-Nielsen formula

$$F_{\chi^2_\nu}(x^2) = \Phi\left(r + \frac{1}{r}\log\frac{Q}{r}\right), \tag{2.20}$$

or by using the asymptotically equivalent Lugannani and Rice formula

$$F_{\chi^2_\nu}(x^2) = \Phi(r) + \phi(r)\left\{\frac{1}{r} - \frac{1}{Q}\right\}, \tag{2.21}$$

where

$$r = \operatorname{sgn}(x^2 - \nu)\left[(x^2 - \nu) + \nu \log\frac{\nu}{x^2}\right]^{1/2} \quad \text{and} \quad Q = \frac{x^2 - \nu}{\sqrt{2\nu}} \tag{2.22}$$

with $O(n^{-3/2})$ accuracy.

Lin (1988) provides a very simple approximation to the cumulative distribution of the χ^2 distribution. Figures 2.6 and 2.7 plot approximations to the cumulative distribution function $F_{\chi^2_\nu}(x^2)$ for $\nu = 5$ and 1, respectively. We observe that the Lin (1988) approximation is not at all satisfactory. We also observe that, even for the extreme case of the χ^2 distribution with 1 degree of freedom, the proposed approximations give remarkably accurate approximations.

Figure 2.6: Approximations to $F_{\chi_5^2}(\chi^2)$

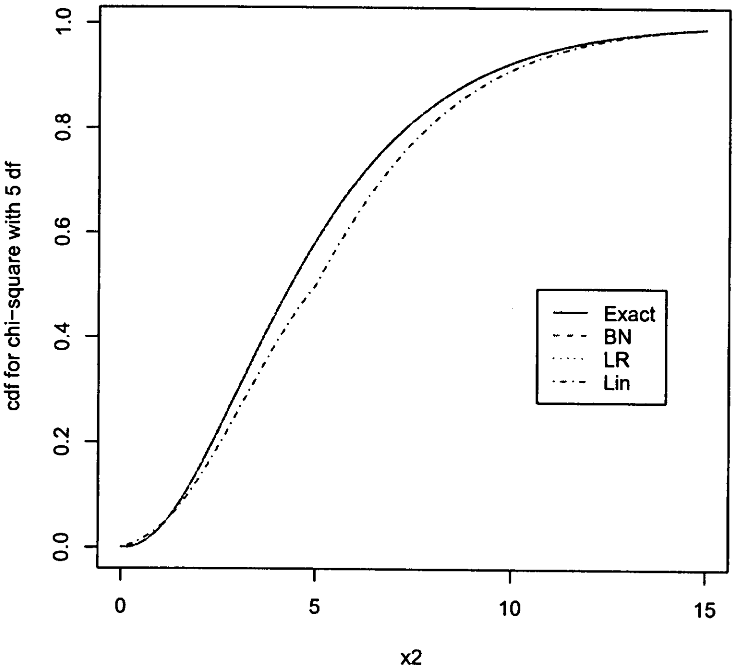
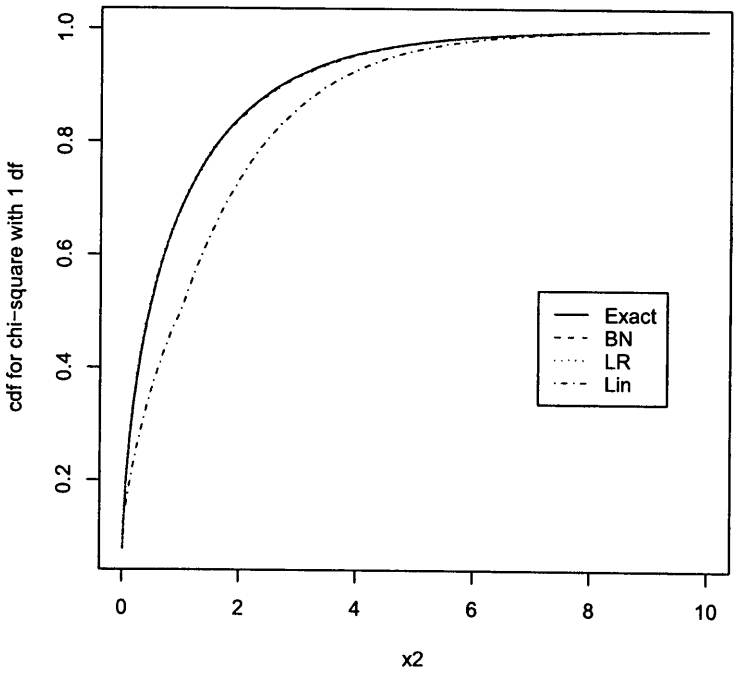


Figure 2.7: Approximations to $F_{\chi_1^2}(\chi^2)$



2.2.3 General Statistical Model

If the dimension of the variable and the dimension of the parameter are the same, as may occur after a reduction by sufficiency in exponential families, approximate p-values for testing a component of the canonical parameter can be obtained from appropriate density approximation. In the case when sufficiency and ancillarity do not reduce the dimension of the variable to that of the parameter, some alternative reduction method such as approximate ancillarity seems needed in order to apply available methods. An approximate ancillary can be developed using likelihood ratio statistics for testing the full model (Barndorff-Nielsen (1986); Barndorff-Nielsen and Wood (1998)). However the feasible methods are lacking for tail probability approximation (see, for example, the discussion in Pierce and Peters (1992)). In 1995, Fraser and Reid indicated that it is possible to find an approximating exponential family model, the tangent exponential model, by using both ancillary direction and observed likelihood to construct an approximate ancillary statistic and the subsequent derivation of significance probabilities having third order accuracy for scalar parameter of interest. The tangent exponential model has the same observed log-likelihood function as the original model and the same first derivative with respect to the data at the observed data point. The tangent exponential model at the data point y^0 is defined from the model $f(y;\theta)$ as

$$f_{TEM}(s;\theta)ds = \exp\{\varphi(\theta)'s + \ell(\theta;y^0)\}h(s)ds \tag{2.23}$$

where $s = s(y)$ is a nominal variable that can be viewed as a score variable $s(y) = \ell_\theta(\hat{\theta}^0;y)$, and $\ell(\theta;y^0)$ and $\varphi(\theta;y^0)$ are defined from the original model as

$$l(\theta;y^0) = \log(f(y^0;\theta)) \tag{2.24}$$

$$\varphi(\theta;y^0)' = \ell_{;V}(\theta;y^0) \tag{2.25}$$

where the notation $\ell_{;V}$ denotes differentiation in the sample space in directions given by the columns of a matrix V. The term ancillary direction $V = (v_1, \cdots, v_p)$ stands for the tangent direction to the ancillary surface at the observed data. It can be constructed easily from a first derivative ancillary based on a full-dimensional pivotal quantity. The pivotal quantity is typically straightforward and natural, and can be viewed as presenting how the variable measures the parameter. Next a general way of obtaining the ancillary direction are discussed.

The tangent vectors V are constructed using a vector $z = (z_1, \cdots, z_n)'$ of pivotal quantities $z_i = z_i(y_i;\theta)$ that has a fixed distribution. A simple choice is given by the successive distribution functions $z_i = F(y_i;\theta)$ for $(i = 1, \cdots, n)$ which are uniformly distributed. The array V is obtained from the pivotal $z(y;\theta)$ by

$$V = \frac{\partial y}{\partial \theta'}\bigg|_{(y^0,\hat{\theta})} = -\left(\frac{\partial z}{\partial y'}\right)^{-1}\left(\frac{\partial z}{\partial \theta'}\right)\bigg|_{(y^0,\hat{\theta})}$$

where y^0 is the observed data point and $\hat{\theta}$ is the maximum likelihood estimate.

If we choose cumulative distribution functions to be our pivotal quantities, the V becomes

$$V = \frac{\partial y}{\partial \theta'} = -F_{y'}^{-1}(y;\theta)F_{\theta'}(y;\theta)\bigg|_{(\hat{\theta},y^0)} \tag{2.26}$$

where $F_y(y;\theta)$ and $F_\theta(y;\theta)$ are the partial derivatives of $F(y;\theta)$ with respect to y and θ, respectively. For more details on the ancillary directions tangent to the surface of an approximate ancillary statistic see Fraser (1990), Fraser and Reid (1995, 1996).

Fraser and Reid (1993) introduced the tangent exponential model which is the exponential family model whose asymptotic expansion is closest to that of the true model. The advantage is that highly accurate approximations available for the exponential family can be extended to general models.

If the model is a conditional model which is conditioned on some ancillary statistic, then the conditional likelihood gradient becomes the full likelihood gradient. Fraser and Reid (1999) showed that the gradient of the conditional likelihood, $\varphi(\theta)$, is

$$\begin{aligned}\varphi(\theta) &= \left.\frac{\partial}{\partial V} l(\theta; y)\right|_{y^0} \\ &= l_y(\theta'; y^0)V \\ &= \left(\sum_{i=1}^{n} \frac{\partial}{\partial y_i} l(\theta; y^0) v_{i1}, \cdots, \sum_{i=1}^{n} \frac{\partial}{\partial y_i} l(\theta; y^0) v_{ik}\right). \qquad (2.27)\end{aligned}$$

This gradient $\varphi(\theta)$ gives a canonical reparameterisation, which is the canonical parameter of the tangent exponential model at the data point y^0 (Fraser and Reid (1995)). So the observed log-likelihood $l(\theta; y^0)$ and the ancillary direction V together will produce a locally defined canonical parameter $\varphi(\theta)$. The tangent exponential model provides the full third order p-values for the original model (Fraser and Reid (1999)).

Once we have the tangent exponential family model and locally defined canonical parameter, the methodology in Section 2.2.2 can be directly applied to approximating the tail probability by using either the Barndorff-Nielsen approximation (2.8) or the Lugannani-Rice approximation (2.9) with r and Q being defined in (2.10) and (2.13), respectively. Thus a $(1-\alpha) \times 100\%$ confidence interval for ψ has the same expression as in Section 2.2.2. We will apply the asymptotic method to all the examples and simulations in the rest of the thesis, and compare it with some recent methods to see the accuracy in location-scale family models.

2.3 Skovgaard Method

A different approach to higher order approximation was proposed by Skovgaard (1996), who obtained estimates of the directions of conditioning. Let $I(\theta;\theta_0)$ designate a mean log-likelihood function:

$$I(\theta;\theta_0) = E_{\theta_0}\{\ell(\theta;y)\} = \int \ell(\theta;y)f(y;\theta_0)dy \tag{2.28}$$

where the symbol E_{θ_0} means taking the expectation over the distribution with parameter θ_0. This function arises in studies of the robustness of likelihood inference, where it is called the Fraser Information (Kent (1982), Fraser and Reid (2010)). A new version of φ, say $\bar{\varphi}$, of the canonical parameter for the model (2.12) is defined by differentiating the function $I(\theta;\hat{\theta})$ instead of $\ell(\dot{\theta};y)$

$$\bar{\varphi}(\theta) = \frac{\partial}{\partial\theta_0}I(\theta;\theta_0)\big|_{\theta_0=\hat{\theta}} = \frac{\partial}{\partial\hat{\theta}}I(\theta;\hat{\theta}). \tag{2.29}$$

Averaging the log-likelihood in the calculation of $\bar{\varphi}$ eliminates dependence on the approximate ancillary, and also for many models, the calculation of $\bar{\varphi}$ is simpler than the calculation of φ. On the other hand, it reduces the accuracy of tail area approximations based on $\bar{\varphi}$ discussed in the following part of this Section. Broadly speaking, the φ version is easier to compute in transformation families and the $\bar{\varphi}$ is easier to compute in curved exponential families.

Example. Suppose y_i follows a one-parameter location model $f(y-\theta)$. A sample

$(y_1, \cdots, y_n)$ admits an exact ancillary statistic, $a = (a_1, \cdots, a_n) = (y_1 - \hat{\theta}, \cdots, y_n - \hat{\theta})$, and the $n \times 1$ vector V from the pivotal $z_i = y_i - \theta$ is simply a vector of 1s. Thus

$$\begin{aligned} \varphi(\theta) &= \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \log f(y_i - \theta) \\ &= -\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(y_i - \theta). \end{aligned}$$

From (2.27),

$$\begin{aligned} I(\theta; \theta_0) &= E_{\theta_0}\{\ell(\theta; y)\} \\ &= E_{\theta_0}\left\{\sum_{i=1}^{n} \log f(y_i - \theta)\right\} \\ &= n \int \log f(y_i - \theta) f(y_i; \theta_0) dy_i \\ &= n \int \log f(y_i - \theta) f(y_i - \theta_0) dy_i. \end{aligned}$$

Then

$$\bar{\varphi}(\theta) = n \int \frac{\partial}{\partial \hat{\theta}} \log f(y - \theta) f(y - \hat{\theta}) dy.$$

On dividing the above expressions for $\varphi(\theta)$ and $\bar{\varphi}(\theta)$ by n, we see that $\varphi(\theta)$ is the nonparametric bootstrap estimate of the expected value of $\frac{\partial}{\partial \theta} \log f(y - \theta)$ and $\bar{\varphi}(\theta)$ is the parametric bootstrap estimate of the same quantity.

For higher order log-likelihood inference, Skovgaard (1996, 2001) derived another expression for r^* from Barndorff-Nielsen (1986) for tail area approximation in a wellbehaved parametric model with the same expression of r but different expression of Q as follows:

$$Q = \left[\hat{S}^{-1}\hat{q}\right]_{\psi} |j_{\theta\theta}(\hat{\theta})|^{1/2} |j_{\lambda\lambda}(\hat{\theta}_{\psi})|^{-1/2} |i_{\theta\theta'}(\hat{\theta})|^{-1} |\hat{S}| \tag{2.30}$$

where $[...]_{\psi}$ denotes the ψ coordinate of the vector, $i_{\theta\theta'}(\hat{\theta})$ is the expected Fisher information matrix, $j_{\theta\theta}(\hat{\theta})$ and $j_{\lambda\lambda}(\hat{\theta}_{\psi})$ are the observed Fisher information matrices for θ and λ, respectively, $\hat{q}$ and $\hat{S}$ are defined as follows:

$$\hat{q} = cov(l_{\theta}(\hat{\theta}), l(\hat{\theta}) - l(\hat{\theta}_{\psi})) \quad \text{and} \quad \hat{S} = cov(l_{\theta}(\hat{\theta}), l_{\theta}(\hat{\theta}_{\psi}))$$

respectively. The first component of the vector in square brackets assumes that ψ is the first component of θ.

The approximate $(1-\alpha)100\%$ confidence interval for ψ has the same expression as in Section 2.2.2.

The idea of the Skovgaard method came from Barndorff-Nielson's likelihood approximation formula. Barndorff-Nielson in 1986 and 1991 derived Q as follows:

$$Q = \left\{[l_{\theta}(\hat{\theta}) - l_{\theta}(\hat{\theta}_{\psi})]' l_{\theta\theta}^{-1}(\hat{\theta}_{\psi})\right\}_{\psi} |j(\hat{\theta})|^{-1/2} |l_{\theta\theta}(\hat{\theta}_{\psi})| |j(\hat{\theta}_{\psi})|^{-1/2}.$$

It involves the sample space derivatives $l_{\theta}(\hat{\theta}) - l_{\theta}(\hat{\theta}_{\psi})$ and $l_{\theta\theta}(\hat{\theta}_{\psi})$ and the sample space derivatives are only defined when an ancillary statistic is specified and $\hat{\theta}$ is sufficient. Even so, the computation may be difficult. We can still calculate it in full exponential models and in simple transformation models. But what if it is not available to calculate? Skovgaard (1996) estimated the sample space derivatives by covariances as follows:

$$\begin{aligned}[l_\theta(\hat{\theta}) - l_\theta(\hat{\theta}_\psi)]' &\cong \hat{q}'\hat{\imath}^{-1}\hat{\jmath}\\ l_{\theta\theta}(\hat{\theta}_\psi) &\cong \hat{S}'\hat{\imath}^{-1}\hat{\jmath}\end{aligned}$$

where

$$\begin{aligned}\hat{S} &= cov_{\hat{\theta}}(l_\theta(\hat{\theta}), l_\theta(\hat{\theta}_\psi))\\ \hat{q} &= cov_{\hat{\theta}}(l_\theta(\hat{\theta}), l_\theta(\hat{\theta}) - l_\theta(\hat{\theta}_\psi)).\end{aligned}$$

Then Q could be estimated by (2.30).

The Skovgaard method expressions do not require specification of the ancillary statistic or its tangent vectors V. But the expected Fisher information matrix, together with the two covariances defined above are of the same computational complexity. Alternative expressions for $\hat{q}$ and $\hat{S}$ are in terms of derivatives of the Kullback-Leibler distance

$$\begin{aligned}KL(\theta, \theta_1) &= E_\theta\{\log f(y;\theta) - \log f(y;\theta_1)\}\\ &= E_\theta\{\ell(\theta) - \ell(\theta_1)\}\end{aligned}$$

from which we obtain

$$\begin{aligned}
\chi_{10}(\theta, \theta_1; \theta) &= cov_\theta \{\ell_\theta(\theta), \ell(\theta) - \ell(\theta_1)\} \\
&= \frac{\partial}{\partial \theta} KL(\theta, \theta_1), \\
\chi_{11}(\theta, \theta_1; \theta) &= cov_\theta \{\ell_\theta(\theta), \ell(\theta_1)\} \\
&= -\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta_1} KL(\theta, \theta_1).
\end{aligned}$$

Then we have estimation of $\hat{q}$, $\hat{S}$ and $i(\hat{\theta})$ as follows:

$$\begin{aligned}
\hat{q} &= \chi_{10}(\hat{\theta}, \hat{\theta}_\psi; \hat{\theta}) \\
\hat{S} &= \chi_{11}(\hat{\theta}, \hat{\theta}_\psi; \hat{\theta}) \\
i_{\theta\theta'}(\hat{\theta}) &= \chi_{11}(\hat{\theta}, \hat{\theta}).
\end{aligned}$$

These are the methods for estimating the three complex terms in the Skovgaard method. But Severini (1999, 2000) showed that in the model that he considered, the estimation of $\hat{S}$ and $\hat{q}$ are numerically unstable.

From the definition of $\bar{\varphi}(\theta)$ at (2.29), we could see that the Skovgaard method of Q is identical to (2.14) with different canonical parameter $\bar{\varphi}$ as

$$\bar{\varphi}(\theta) = cov_{\hat{\theta}}\{\ell_\theta(\hat{\theta}), \ell(\theta)\} \tag{2.31}$$

$$\bar{\varphi}_\theta(\theta) = cov_{\hat{\theta}}\{\ell_\theta(\hat{\theta}), \ell_\theta(\theta)\} \tag{2.32}$$

where $cov_{\hat{\theta}}$ means taking the covariance over the distribution with parameter $\hat{\theta}$. Note $\bar{\varphi}_\theta(\hat{\theta}) = i_{\theta\theta'}(\hat{\theta})$, and Skovgaard (1996) also noted that the determinant in the numerator of (2.14) can be expressed as

$$|\varphi_\theta(\hat{\theta}_\psi)|[\varphi_\theta^{-1}(\hat{\theta}_\psi)\left\{\varphi(\hat{\theta}) - \varphi(\hat{\theta}_\psi)\right\}]_1$$

where the choice of the first component of the vector in square brackets assumes that ψ is the first component of θ (Fraser and Reid (2010)).

Let us consider a simple example where $y = (y_1; \cdots ; y_n)$ is a random sample from a normal distribution with mean μ and variance σ^2. The parameter of interest is $\psi = \mu$. Here are some facts that we need to calculate Q in the Skovgaard method:

$$
\begin{aligned}
\ell(\theta) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \\
\ell_\mu(\theta) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\
\ell_{\sigma^2}(\theta) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 \\
\ell_{\mu\mu}(\theta) &= -\frac{n}{\sigma^2} \\
\ell_{\mu\sigma^2}(\theta) &= -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu) \\
&= \ell_{\sigma^2\mu} \\
\ell_{\sigma^2\sigma^2}(\theta) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i-\mu)^2 \\
\hat{\theta} &= (\hat{\mu}, \hat{\sigma^2}) \\
&= (\bar{y}, \frac{(n-1)s^2}{n}) \\
j_{\theta\theta}(\theta) &= \begin{bmatrix} -\ell_{\mu\mu} & -\ell_{\mu\sigma^2} \\ -\ell_{\sigma^2\mu} & -\ell_{\sigma^2\sigma^2} \end{bmatrix} \\
j_{\theta\theta}(\hat{\theta}) &= \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{bmatrix} \\
\left| j_{\theta\theta}(\hat{\theta}) \right| &= \frac{n^2}{2\hat{\sigma}^6} \\
\hat{\theta}_\psi &= (\mu, \hat{\sigma}^2_\mu) \\
&= \left(\mu, \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n}\right)
\end{aligned}
$$

$$\begin{aligned} j_{\lambda\lambda}(\theta) &= j_{\sigma^2\sigma^2}(\theta) \\ &= -\ell_{\sigma^2\sigma^2}(\theta) \\ &= \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{n}{2\sigma^4} \\ \left|j_{\lambda\lambda}(\hat{\theta}_\psi)\right| &= \frac{n}{2\hat{\sigma}_\mu^4} \end{aligned}$$

As discussed above, we use Kullback-Leibler divergence to derive the two covariance $\hat{S}$, $\hat{q}$ and the expected Fisher information matrix $i_{\theta\theta'}(\hat{\theta})$.

Consider two univariate normal models F_0 and F with parameter $\theta_0 = (\mu_0, \sigma_0^2)$ and $\theta = (\mu, \sigma^2)$. After some calculations from the formula for KL divergence, we have

$$KL(\theta_0, \theta) = \frac{1}{2}\left\{\frac{(\mu-\mu_0)^2}{\sigma^2} + \frac{\sigma_0^2}{\sigma^2} - \log\left(\frac{\sigma_0^2}{\sigma^2}\right) - 1\right\}.$$

For details of the calculations, see Strimmer (2010). Then it is straightforward to obtain $\hat{S}$, $\hat{q}$ and the expected Fisher information matrix $i_{\theta\theta'}(\hat{\theta})$. Note that for our simple example, the expected Fisher information matrix for normal distribution is as follows:

$$i_{\theta\theta'}(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4}, \end{bmatrix}$$

therefore

$$|i_{\theta\theta'}(\hat{\theta})|^{-1} = 2\hat{\sigma}^6.$$

Results from the above simulated data show that the Skovgaard method and the Fraser and Reid method give indistinguishable estimation accuracy. Both are much better than the first order methods especially for small sample size (n = 3 for the following figure).

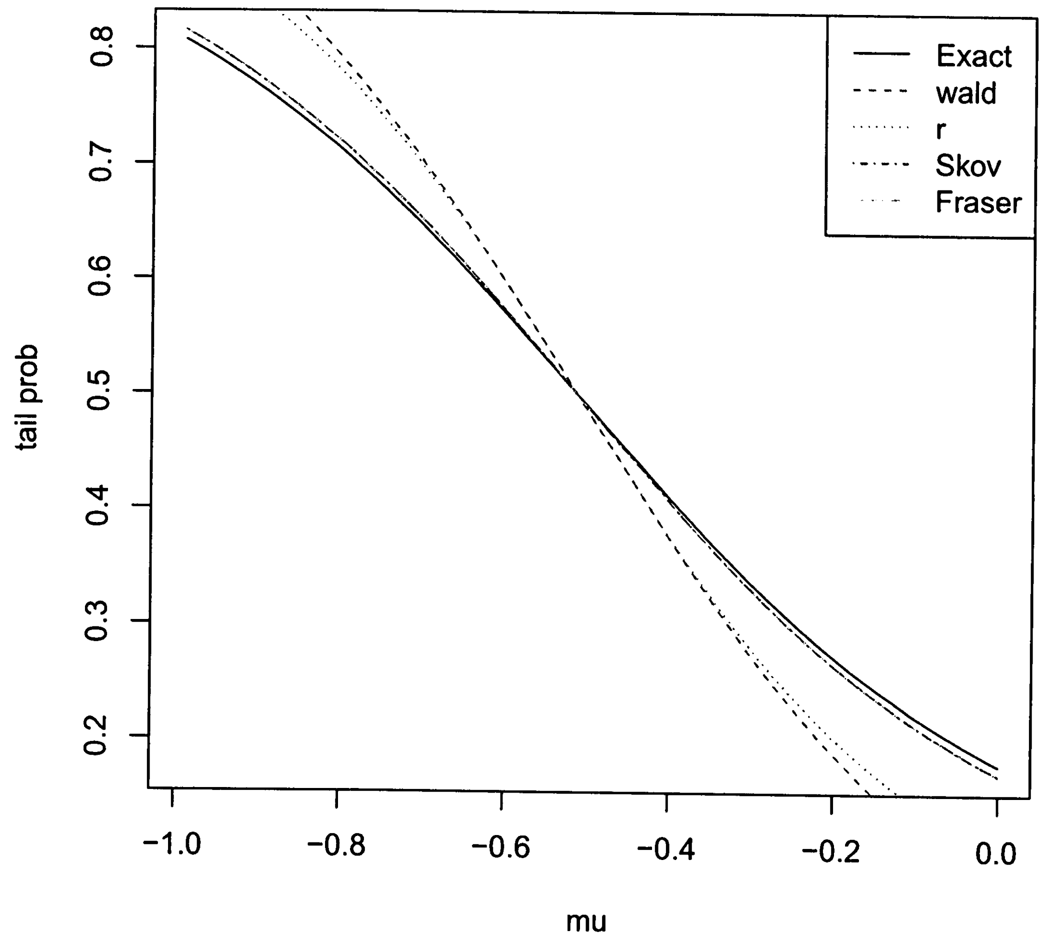


Figure 2.8: Approximation to the mean of Normal distribution

2.4 Summary

We introduced the third order methods to calculate the p-value function of a parameter of interest approximated by Lugannani-Rice and Barndorff- Nielsen with third order relative error. Fraser developed the method from natural exponential, exponential to general models. Based on the property for exponential family models, we derive simpler formula to calculate the p-value function for the parameter of interest. But for the general models, we need to specify the ancillary statistic and its tangent vectors. Skovgaard developed another approach to calculate the the confidence intervals for the parameter of interest without specifying the ancillary statistic and its tangent vectors. But it is difficult, or sometimes impossible, to calculate the expected Fisher information matrices and covariances. Also Fraser (2010) established a simple connection between the higher order approximation due to Skovgaard (1996) and that of Fraser and Reid (1999). He shows that the Skovgaard approximation to p-value function can be obtained by using the exponential family model, but with a different canonical parameter. From my numerical comparison, Jing, Shao and Zhou (2004) method for Student t distribution is not satisfactory toward the tail of the distribution which is crucial for inference purpose. Although Lin (1988) provided a very simple approximation to the cumulative distribution of the χ^2 distribution, but the Fraser and Reid method and the Skovgaard method give more accurate approximations especially for the extreme case. Formulas for r and Q of Student t distribution and χ^2 distribution are explicitly calculated for the Fraser and Reid method. Formulas for r and Q of normal distribution are numerically calculated for the Skovgaard method.

3 Inference on Location-Scale Family

This Chapter is devoted to the Fraser and Reid method and the Skovgaard method applied to the location-scale model to obtain the third order approximation to the p-value function for either the location or scale parameter. In Section 3.1, we will start from a simple location model, and then extend to general location model, finally convert scale parameter to be another location type of parameter to calculate the confidence interval. In Section 3.2, general formulas to calculate the confidence intervals for location or scale parameters are derived.

3.1 A Simple and Accurate Approach for Location-Scale Model

In this Section, based on the special structure of the location-scale model, a simple, efficient and accurate numerical procedure is first developed for the location model and then extended to location-scale model.

For statistical purposes, the approximation of the cumulative distribution function of $\bar{Y}$, the mean of n independent identically distributed random variables, is more convenient if based on the likelihood function for an embedding exponential model. In Chapter 2, we introduced the invariant tail probability formula which can be calculated by (2.1), (2.6) and (2.7) or (2.11) or (2.13) based on this approach. The approximation can also be embedded in a location model. The two tail area approximations are special cases of different invariant versions of the Lugannani and Rice formula (1980). The invariant version are due to Barndorff-Nielsen (1988, 1990) and Fraser (1990). Fraser's (1990) invariant version uses a data dependent parameter that is obtained as the sample space derivative of the observed likelihood and is given by (2.1) with r in (2.6) and Q specified as follows:

$$Q = (l_y(\hat{\theta}) - l_y(\theta)) l_{y\theta}^{-1}(\hat{\theta}) \left\{ j_{\theta\theta'}(\hat{\theta}) \right\}^{1/2}. \tag{3.1}$$

3.1.1 Inference for Simple Location Model

Consider the simple location model:

$$f(y; \mu) = f(y - \mu),$$

the observed log-likelihood function is

$$l(\mu) = \sum_{i=1}^{n} \log f(y_i - \mu).$$

The three first order statistics for location model are as follows:

- the Wald statistic

$$w(\mu) = j_{\mu\mu}(\hat{\mu})^{1/2}(\hat{\mu} - \mu),$$

- the signed log-likelihood ratio statistic

$$r(\mu) = sign(\hat{\mu} - \mu)\left\{2[l(\hat{\mu}) - l(\mu)]\right\}^{1/2},$$

- the Score statistic

$$s(\mu) = j_{\mu\mu}(\hat{\mu})^{-1/2} l_{\mu}(\mu).$$

Then the p-value could be calculated from (2.9).

For the Fraser and Reid method, from some simple calculation, the Q for location model from (3.1) is

$$Q = l_{\mu}(\mu)\left\{j_{\mu\mu}(\hat{\mu})\right\}^{-1/2}. \tag{3.2}$$

The same expression can be easily derived from the Skovgaard formula as well. That means, the Fraser and Reid method and the Skovgaard method give the same expression of Q for simple location model.

Note that since this is a location model and there is no nuisance parameter, the statistic Q is the Score statistic. Therefore, the significance function for location model can be obtained by using either the Barndorff-Nielsen approximation (2.8) or the Lugannani-Rice approximation (2.9) with $r = r(\mu)$ and Q being defined above.

The following simple example illustrates the accuracy of this numerical procedure even for a very small sample size. However, for simple location model, to obtain the p-value function, numerical interaction of the density may be preferred. This third order approximation to obtain the p-value function could serve as the basic procedure which allows us to extend to location-scale model and avoid complex integrals from these models.

Example: Consider a Cauchy distribution with location parameter μ and scale 1, which has density

$$f(x;\mu) = \pi^{-1}(1 + (x - \mu)^2)^{-1}.$$

Let us consider an extreme case with the sample size one and the observed data is $x^0 = 1$. Therefore the observed log-likelihood function is

$$l(\mu) = -\log(1 + (1 - \mu)^2).$$

The exact p-value function can be evaluated by direct integration and is equal to

$$p(\mu) = \pi^{-1}[arctan(1 - \mu) + \pi/2]$$

for $\mu \in (-\infty, \infty)$. The following figure plots the p-value functions obtained from the Wald statistic, signed log-likelihood statistic, our proposed method and the exact significance value. It is obvious to see that even in such a extremely small sample size ($n = 1$), our proposed method still out performed the other 3 asymptotic methods. Also note that the unusual behavior of the p-value function obtained by the Score statistic is due to the non-monotonicity of the score function for the Cauchy distribution.

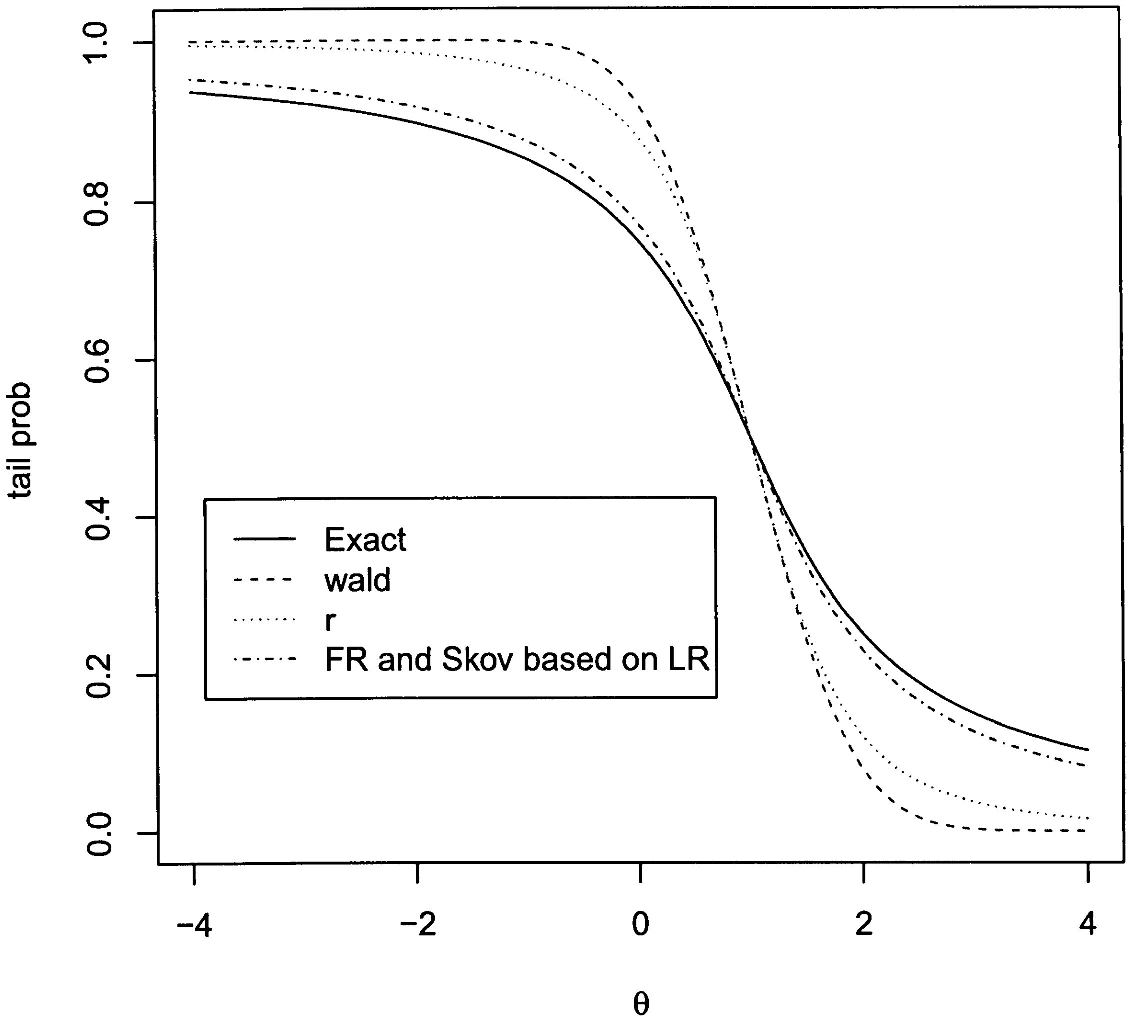


Figure 3.1: $p(\mu)$ for Cauchy distribution with sample size 1

3.1.2 Inference on General Location Model

Now let us consider the general location model which has density

$$f(y_1, \cdots, y_n; \mu_1, \cdots, \mu_n) = f(y_1 - \mu_1, \cdots, y_n - \mu_n).$$

If we are interested in a particular location parameter, say μ_i, then we need the marginal density of $(y_i - \mu_i)$, which should be obtained by integrating out the other ys. To avoid the high dimensional integrations, we could use the numerical procedure described in the previous Section which only needs the marginal log-likelihood function for μ_i. However the marginal log-likelihood function may not be easy to obtain in an explicit form. DiCiccio, Field and Fraser (1990), Fraser, Lee and Reid (1990) and Fraser and Reid (1990) derived and justified the approximated observed marginal log-likelihood function.

Let $\theta = (\psi, \lambda')'$, where ψ is the location parameter that we are interested, and λ is the vector containing the remaining location parameters, or we can say λ is the nuisance parameter. Then the observed marginal log-likelihood function can be approximated by

$$l_m(\psi) = l(\hat{\theta}_\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\hat{\theta}_\psi)| \tag{3.3}$$

where $\hat{\theta}_\psi = (\psi, \hat{\lambda}'_\psi)'$, $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ for a fixed ψ, $l(\psi, \hat{\lambda}_\psi)$ is the original observed log-likelihood function evaluated at $\hat{\lambda}_\psi$, and $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) =$

$-\partial^2 l(\psi,\lambda)/\partial\lambda\partial\lambda'$ evaluated at $\hat{\lambda}_\psi$.

Thus we can get Q in (3.2) by using equation (3.3) as input. Therefore, the p-value function for general location model can be obtained.

Skovgaard (2001) derived the Q explicitly for full exponential model and simple transformation models. For general location model, the formula to calculate Q is as follows:

$$Q = \ell_\psi(\hat{\theta}_\psi)\left\{\frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}\right\}^{-\frac{1}{2}}. \tag{3.4}$$

3.1.3 Inference for Location-Scale Model

Now let us consider the location-scale model. There exists a random variable Z that is

$$Z = \frac{Y-\mu}{\sigma}$$

which has the same distribution under all values of the parameter. The density of the location-scale model has the following form

$$f(y;\mu,\sigma) = \frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$$

where μ is the location parameter and σ is the scale parameter. The joint density can also be rewritten as

$$f(y_1,\cdots,y_n;\mu,\sigma) = \prod f(\frac{y_i-\mu}{s}e^{\log(s)-\log(\sigma)})e^{-\log(\sigma)}$$

where μ is a location parameter and $\tau = \log(\sigma)$ is a location type parameter. Therefore we could use our numerical procedure for the general location model to obtain the p-value function for location-scale model. The observed log-likelihood function can be written as

$$l(\mu, \tau) = -n\tau + \sum_{i=1}^{n} \log f((y_i - \mu)e^{-\tau}).$$

For inference concerning either μ or $\tau = \log(\sigma)$, we need marginal log-likelihood function for the Fraser and Reid method which can be approximated by (3.3). Once we have the observed marginal log-likelihood function, the numerical procedure introduced in the previous Section will give the approximated significance function. For the Skovgaard method, we can calculate Q explicitly. The following example will demonstrate how this numerical procedure works.

Example: The Lieblein and Zelen (1956) data which recorded the lifetimes, t, until failure of 23 deep-grove ball bearings in millions of revolutions are

17.88	28.92	33.00	41.52	42.12	45.60
48.48	51.84	51.96	54.12	55.58	67.80
68.64	68.64	68.68	84.12	93.12	98.64
105.12	105.82	127.92	128.04	173.40	

Wong (1992) analyzed this data set by using the log-normal analysis. Let $y = \log(t)$, then $y \sim N(\mu, \sigma)$. Note normal distribution belongs to location-scale family with location parameter μ and scale parameter σ. The density function of y is as follows

$$f(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}.$$

And the observed log-likelihood function can be expressed as:

$$\begin{aligned} l(\mu,\sigma) &= \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i-\mu}{\sigma})^2} \\ &= -n\log\sigma - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{n}(y_i-\mu)^2. \end{aligned}$$

With reparameterization, $\tau = \log(\sigma)$, it is changed to location model with location parameter μ and location type of parameterτ. Then the observed log-likelihood function becomes

$$l(\mu,\tau) = -n\tau - \frac{1}{2}e^{-2\tau}\sum_{i=1}^{n}(y_i-\mu)^2. \tag{3.5}$$

Now we are ready to apply our numerical procedure discussed in the previous Sections. First we need the observed marginal log-likelihood functions for our location parameter or location type of parameter (depends on which parameter we are interested in). Here are something we need to obtain the observed marginal log-likelihood functions for μ and τ in the Fraser and Reid method and to calculate Q in the Skovgaard method

$$\begin{aligned}
l_\mu(\mu,\tau) &= e^{-2\tau}(n\bar{y} - n\mu) \\
l_\tau(\mu,\tau) &= -n + e^{-2\tau}\sum_{i=1}^{n}(y_i - \mu)^2 \\
l_{\mu\mu}(\mu,\tau) &= -ne^{-2\tau} \\
l_{\tau\tau}(\mu,\tau) &= -2e^{-2\tau}\sum_{i=1}^{n}(y_i - \mu)^2 \\
\hat{\mu} &= \bar{y} \\
\hat{\tau} &= -\frac{1}{2}log(n-1)s^2 \\
\hat{\tau}_\mu &= -\frac{1}{2}\log\frac{n}{\sum_{i=1}^{n}(y_i - \mu)^2} \\
\hat{\mu}_\tau &= \bar{y}
\end{aligned}$$

where $\bar{y}$ is the sample mean and s is the sample standard deviation.

Then the marginal log-likelihood functions for μ and τ are

$$\begin{aligned}
l_m(\mu) &\approx -\frac{n}{2}\log\frac{n}{\sum_{i=1}^{n}(y_i - \mu)^2} \\
&\approx -\frac{n}{2}\log[(n-1)s^2 + n(\bar{y} - \mu)^2], \qquad (3.6) \\
l_m(\tau) &\approx -(n-1)\tau - \frac{n-1}{2}s^2e^{-2\tau}. \qquad (3.7)
\end{aligned}$$

Then

$$\begin{aligned}
l_{m_\mu}(\mu) &\approx n^2(\bar{y}-\mu)[(n-1)s^2+n(\bar{y}-\mu)^2]^{-1} \\
l_{m_{\mu\mu}}(\mu) &\approx n^2[(n-1)s^2+n(\bar{y}-\mu)^2]^{-2}[n(\bar{y}-\mu)^2-(n-1)s^2] \\
\left[-l_{m_{\mu\mu}}(\mu)\right]^{-1/2} &\approx \frac{[(n-1)s^2+n(\bar{y}-\mu)^2]}{n}[n(\bar{y}-\mu)^2-(n-1)s^2]^{-1/2} \\
\left[-l_{m_{\mu\mu}}(\hat{\mu})\right]^{-1/2} &\approx \frac{(n-1)^{1/2}}{n}s.
\end{aligned}$$

So if the parameter of interest is μ, the location parameter of normal distribution, the Q needed in the Fraser and Reid method is as follows:

$$\begin{aligned}
Q &= \ell_{m_\mu}(\mu)\left[j_{m_{\mu\mu}}(\hat{\mu})\right]^{-1/2} \\
&= \frac{n(n-1)^{1/2}(\bar{y}-\mu)s}{(n-1)s^2+n(\bar{y}-\mu)^2} \qquad (3.8)
\end{aligned}$$

where the sample standard deviation s has value 0.5333, and sample mean $\bar{y}=4.1503$ for this example.

If our parameter of interest is τ, the log scale parameter of normal distribution, following the same steps, we can get exact formula to calculate Q as well.

$$\begin{aligned}
l_{m_\tau}(\tau) &\approx (n-1)(s^2e^{-2\tau}-1) \\
l_{m_{\tau\tau}}(\tau) &\approx -2(n-1)s^2e^{-2\tau} \\
[-l_{m_{\tau\tau}}(\tau)]^{-1/2} &\approx [2(n-1)]^{-1/2}s^{-1}e^{\tau} \\
[-l_{m_{\tau\tau}}(\hat{\tau})]^{-1/2} &\approx 2^{-1/2}(n-1)^{-1}s^{-2}.
\end{aligned}$$

So the Q needed in the Fraser and Reid method for log scale parameter of interest in normal distribution is as follows:

$$\begin{aligned} Q &= \ell_{m_\tau}(\tau)\,[j_{m_{\tau\tau}}(\hat{\tau})]^{-1/2} \\ &= 2^{-1/2}(e^{-2\tau} - s^{-2}). \qquad (3.9) \end{aligned}$$

For the Skovgaard method, Q could be calculated explicitly as follows:

If our parameter of interest is μ, then

$$\begin{aligned}
\ell_\mu(\theta) &= e^{-2\tau}(n\bar{y}-n\mu) \\
\ell_\mu(\hat{\theta}_\tau) &= \left[(n-1)s^2+n(\bar{y}-\mu)^2\right](n\bar{y}-n\mu) \\
j_{\lambda\lambda}(\theta) &= j_{\tau\tau}(\theta) \\
&= -\ell_{\tau\tau}(\theta) \\
&= 2e^{-2\tau}\sum_{i-1}^{n}(y_i-\mu)^2 \\
j_{\lambda\lambda}(\hat{\theta}_\tau) &= 2n \\
j_{\theta\theta}(\theta) &= \begin{bmatrix} -\ell_{\mu\mu} & -\ell_{\mu\tau} \\ -\ell_{\tau\mu} & -\ell_{\tau\tau} \end{bmatrix} \\
&= \begin{bmatrix} ne^{-2\tau} & 2e^{-2\tau}(n\bar{y}-n\mu) \\ 2e^{-2\tau}(n\bar{y}-n\mu) & 2e^{-2\tau}\sum_{i=1}^{n}(y_i-\mu)^2 \end{bmatrix} \\
j_{\theta\theta}(\hat{\theta}) &= \begin{bmatrix} n(n-1)s^2 & 0 \\ 0 & 2(n-1)^2s^4 \end{bmatrix} \\
|j_{\theta\theta}(\hat{\theta})| &= 2n(n-1)^3s^6
\end{aligned}$$

since $s^2=\frac{1}{n-1}\sum_{n=1}^{n}(y_i-\bar{y})^2$. Then we are able to calculate Q as follows:

$$Q=\left[(n-1)s^2+n(\bar{y}-\mu)^2\right](n\bar{y}-n\mu)(n-1)^{-3/2}s^{-3}. \tag{3.10}$$

If our parameter of interest is log scale parameter τ, then

$$\begin{aligned}
l_\tau(\theta) &= -n + e^{-2\tau}\sum_{i=1}^{n}(y_i-\mu)^2 \\
l_\tau(\hat{\theta}_\tau) &= -n + e^{-2\tau}(n-1)s^2 \\
|j_{\theta\theta}(\hat{\theta})| &= 2n(n-1)^3s^6 \\
|j_{\lambda\lambda}(\theta)| &= |j_{\mu\mu}(\theta)| \\
&= |-l_{\mu\mu}| \\
&= ne^{-2\tau}
\end{aligned}$$

Then we are able to calculate Q as follows:

$$Q = 2^{-1/2}(n-1)^{-1/2}s^{-1}e^{-3\tau} - 2^{-1/2}n(n-1)^{-3/2}e^{-\tau}s^{-3}. \tag{3.11}$$

Therefore, the p-value functions for μ and τ can be obtained by using either the Barndorff-Nielsen approximation (2.8) or the Lugannani-Rice approximation (2.9) with r calculated by (2.6) and Q being defined above. The confidence intervals for parameters μ and τ can be obtained by (1.10).

The following two tables show that the numerical procedure give pretty accurate approximation compared to the exact confidence intervals for the location and the scale parameters.

Table 3.1: 90% confidence interval for μ

	Exact	Fraser and Reid Method	Skovgaard Method
90% CI	(3.959, 4.341)	(3.960, 4.341)	(3.961, 4.341)

Table 3.2: 90% confidence interval for τ

	Exact	Fraser and Reid Method	Skovgaard Method
90% CI	(0.429, 0.712)	(0.429, 0.713)	(0.429, 0.713)

3.2 General Formula for Location-Scale Model from the Fraser and Reid Method and the Skovgaard Method

Suppose Y belongs to the location-scale model defined as

$$Y = \mu + \sigma z \tag{3.12}$$

where z has known density function f. The parameter for location-scale model is $\theta = (\mu, \sigma')$. It is equivalent to say

$$z = \frac{y - \mu}{\sigma}.$$

The log-likelihood function of θ for the location-scale model is

$$\begin{aligned} l(\mu,\sigma) &= \sum_{i=1}^{n} l_i(\mu,\sigma) \\ &= -n\log(\sigma) + \sum_{i=1}^{n} \log f(\frac{y_i-\mu}{\sigma}) \qquad (3.13) \\ &= -n\log(\sigma) + \sum_{i=1}^{n} \log f(z_i). \end{aligned}$$

The overall maximum likelihood estimate $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ can be obtained by solving the following estimating equations:

$$l_\mu(\mu,\sigma) = -\sigma^{-1}\sum_{i=1}^{n} f'(z_i)f^{-1}(z_i) = 0 \qquad (3.14)$$

$$l_\theta(\mu,\sigma) = -n - \sum_{i=1}^{n} f'(z_i)f^{-1}(z_i)\sigma^{-1}(y_i-\mu) = 0. \qquad (3.15)$$

The observed full information matrix is

$$j_{\theta\theta'}(\mu,\sigma) = -\begin{pmatrix} l_{\mu\mu} & l_{\sigma\mu} \\ l_{\mu\sigma} & l_{\sigma\sigma} \end{pmatrix}$$

where

$$\begin{aligned} l_{\mu\mu}(\mu,\sigma) &= \sigma^{-2}\sum_{i=1}^{n}[f''(z_i)f^{-1}(z_i) - f'(z_i)f^{-2}(z_i)] \\ l_{\mu\sigma}(\mu,\sigma) &= \sigma^{-1}\sum_{i=1}^{n}[f'(z_i)f^{-1}(z_i) + f''(z_i)f^{-1}(z_i)(z_i) - f'(z_i)f^{-2}(z_i)(z_i)] \\ &= l_{\sigma\mu}(\mu,\sigma) \\ l_{\sigma\sigma}(\mu,\sigma) &= \sigma^{-1}\sum_{i=1}^{n}(y_i-\mu)[f'(z_i)f^{-1}(z_i) + f''(z_i)f^{-1}(z_i)(z_i) - f'(z_i)f^{-2}(z_i)(z_i)]. \end{aligned}$$

The determinant of the observed information matrix can be developed as follows:

$$|j_{\theta\theta'}(\hat{\theta})| = D(\hat{z})\hat{\sigma}^{-4} \tag{3.16}$$

where

$$\begin{aligned} D(\hat{z}) &= \left\{\sum_{i=1}^{n} g''(\hat{z}_i)\right\}\left\{n + \sum_{i=1}^{n} \hat{z_i}^2 g''(\hat{z}_i)\right\} - \left\{\sum_{i=1}^{n} \hat{z}_i g''(\hat{z}_i)\right\}^2 \\ g(\hat{z}_i) &= -\log f(\hat{z}_i) \\ \hat{z} &= (\hat{z}_1, \cdots, \hat{z_n}) = (\frac{y_1 - \hat{\mu}}{\hat{\sigma}}, \cdots, \frac{y_n - \hat{\mu}}{\hat{\sigma}}). \end{aligned}$$

Similarly, we can easily obtain the observed nuisance information matrix $j_{\lambda\lambda'} = -l_{\lambda\lambda'}$.

Whenever the nuisance parameter is not explicitly available or does not exist in close form, we use Lagrange multiplier technique to get tilted log-likelihood as introduced in Section 1.2.5. For testing $\psi(\mu, \sigma) = \psi_0$, first we need the Lagrange function to derive the constrained MLE and constrained observed nuisance information matrix:

$$H(\mu, \sigma, \lambda) = \ell(\mu, \sigma) + \lambda(\psi(\mu, \sigma) - \psi_0).$$

The constrained MLE $\hat{\theta}_{\psi_0} = (\hat{\mu}_{\psi_0}, \hat{\sigma}_{\psi_0}, \hat{\lambda}_{\psi_0})$ is calculated by the following estimating equations:

$$\begin{aligned} H_\mu(\mu, \sigma, \lambda) &= \ell_\mu(\mu, \sigma) + \lambda\psi_\mu(\mu, \sigma) = 0 \\ H_\sigma(\mu, \sigma, \lambda) &= \ell_\sigma(\mu, \sigma) + \lambda\psi_\sigma(\mu, \sigma) = 0 \\ H_\lambda(\mu, \sigma, \lambda) &= \psi(\mu, \sigma) - \psi_0 = 0. \end{aligned}$$

Then the tilted log-likelihood function is as follows:

$$\tilde{\ell}(\hat{\mu}_{\psi_0}, \hat{\sigma}_{\psi_0}, \hat{\lambda}_{\psi_0}) = \ell(\hat{\mu}_{\psi_0}, \hat{\sigma}_{\psi_0}) + \hat{\lambda}_{\psi_0}(\psi(\hat{\mu}_{\psi_0}, \hat{\sigma}_{\psi_0}) - \psi_0).$$

And the constrained observed nuisance information matrix is

$$\tilde{\jmath}_{\theta\theta'}(\mu, \sigma) = -\begin{pmatrix} \tilde{l}_{\mu\mu} & \tilde{l}_{\sigma\mu} \\ \tilde{l}_{\mu\sigma} & \tilde{l}_{\sigma\sigma} \end{pmatrix}$$

.

From (2.26), the ancillary direction

$$V_{ij} = -\frac{\partial F(y_i;\theta)/\partial\theta_j}{\partial F(y_i;\theta)/\partial y_i}\bigg|_{\hat{\theta}}$$

where i indicates the dimension of data, and j indicates the dimension of parameter θ. For location-scale model

$$V = \begin{pmatrix} 1 & \frac{y_1-\hat{\mu}}{\hat{\sigma}} \\ \vdots & \vdots \\ 1 & \frac{y_n-\hat{\mu}}{\hat{\sigma}} \end{pmatrix}. \tag{3.17}$$

From (2.27), the canonical parameter $\varphi(\theta)$ for location-scale model is

$$\begin{aligned} \varphi(\theta) &= (\sigma^{-1}\sum_{i=1}^{n} f'(z_i)f^{-1}(z_i), \sigma^{-2}\sum_{i=1}^{n} f'(z_i)f^{-1}(z_i)(y_i-\mu)) \\ &= (\varphi_1, \varphi_2). \end{aligned} \tag{3.18}$$

The scalar parameter of interest in $\varphi(\theta)$ scale is

$$\chi(\theta) = \frac{\psi_{\theta'}(\hat{\theta}_\psi)\varphi_{\theta'}^{-1}(\hat{\theta}_\psi)}{|\psi_{\theta'}(\hat{\theta}_\psi)\varphi_{\theta'}^{-1}(\hat{\theta}_\psi)|}\varphi(\theta)$$

where

$$\varphi_{\theta'}(\theta) = \begin{pmatrix} \frac{\partial\varphi_1}{\partial\mu} & \frac{\partial\varphi_1}{\partial\sigma} \\ \frac{\partial\varphi_2}{\partial\mu} & \frac{\partial\varphi_2}{\partial\sigma} \end{pmatrix}$$

$$[\varphi_{\theta'}(\theta)]^{-1} = (\frac{\partial\varphi_1}{\partial\mu}\frac{\partial\varphi_2}{\partial\sigma} - \frac{\partial\varphi_1}{\partial\sigma}\frac{\partial\varphi_2}{\partial\mu})^{-1} \begin{bmatrix} \frac{\partial\varphi_2}{\partial\sigma} & -\frac{\partial\varphi_1}{\partial\sigma} \\ \frac{\partial\varphi_2}{\partial\mu} & \frac{\partial\varphi_1}{\partial\mu} \end{bmatrix}.$$

Then we are able to calculate r and Q from formula (2.10) and (2.13) to further derive the confidence interval for our hypothesis test by using the Fraser and Reid method.

For the Skovgaard method, the determinant of the observed information matrix and the observed constrained information matrix could be easily got from the above. But the information matrix together with the two covariance terms are pretty complicated to calculate. The Fraser and Reid's and the Skovgaard's development in asymptotic methods provide accurate approximations for p-value function and thus confidence intervals for a scalar component parameter in location-scale model. Although, conceptual-wise, the Skovgaard method is pretty straightforward, it could still be complicated in terms of calculation.

3.3 Summary

A simple and accurate numerical procedure to obtain the p-value function is developed for the location model and then extended to location-scale model. For inference concerning location parameter or scale parameter, the Fraser and Reid numerical procedure only depends on the observed log-likelihood function which can be either full, marginal or conditional log-likelihood function. But this method can only be used once the parameter of interest is either the location parameter or the scale parameter. If the parameter of interest is a function of these two parameters, for example the ratio of the location parameter to the scale parameter, then we need to use other approximation methods such as the Skovgaard method (Skovgaard (2001)). General formulas to calculate the confidence intervals for location or scale parameter in the Fraser and Reid method and the Skovgaard method are derived in the second Section. So that we could apply these third order methods directly to any location-scale model. It makes our third order methods to be pretty straightforward. But it could still be complicated in terms of calculation.

4 Revisit Behrens-Fisher Problem Using Third Order Methods

Inference for the difference of two independent normal means has been widely studied in statistical literature. Typically, the variances are assumed to be unknown and must be estimated. When we assume equal variances, then a pooled estimate of the common variance is used and the test statistic is exactly distributed as a Student t distribution. However, without making the equality of variances assumption, the problem is then the well-known Behrens-Fisher problem, where no exact distribution of the test statistic is available. There exist many approximate solutions for this problem. Most statistical software packages use the Satterthwaite (1946) solution, where the test statistic is approximately distributed as a Student t distribution. Maity and Sherman (2006) considered the Behrens-Fisher problem with an additional assumption that one of the variances is known, and a Satterthwaite type solution is obtained. Wong and Wu (2008) examined the problem considered by Maity and Sherman (2006) and derived a likelihood based asymptotic solution, which has excellent coverage property. Schechtman and Sherman (2007) also considered the Behrens-Fisher problem but with an assumption that the ratio of the two variances is known. This problem arises in many practical situations. For example, when two instruments report averaged responses of the same object based on a different number of replicates, the ratio of the variances of the response is then known, and is the ratio of the number of replicates going into each response. In this chapter, we apply our third order methods developed by Fraser and Reid, and Skovgaard to Behrens-Fisher problem. The simulation results showed the excellent coverage property of our proposed method.

4.1 Main Result for the Fraser and Reid Method

Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_m)$ be random samples from two independent normal distribution with mean and variance (μ_x, σ_x^2) and (μ_y, σ_y^2), respectively. Assume σ_x^2 and σ_y^2 are unknown but with the ratio $\sigma_y^2/\sigma_x^2 = c$ known. Our parameter of interest is $\psi = \mu_x - \mu_y$.

The log-likelihood function can be derived as follows:

$$l(\theta) = -\frac{m+n}{2}\log(\sigma_x^2) - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(x_i - \psi - \mu_y)^2 - \frac{1}{2c\sigma_x^2}\sum_{j=1}^{m}(y_j - \mu_y)^2$$

where $\theta = (\psi, \mu_y, \sigma_x^2)'$.

Here are some facts that we need to calculate Q in the Fraser and Reid method:

$$\begin{aligned}
\ell_\psi(\theta) &= \frac{n(\bar{x}-\psi-\mu_y)}{\sigma_x^2} \\
\ell_{\mu_y}(\theta) &= \frac{n(\bar{x}-\psi-\mu_y)}{\sigma_x^2} + \frac{m(\bar{y}-\mu_y)}{c\sigma_x^2} \\
\ell_{\sigma_x^2}(\theta) &= -\frac{m+n}{2\sigma_x^2} + \frac{1}{2(\sigma_x^2)^2}\left(\sum_{i=1}^{n}(x_i-\psi-\mu_y)^2 + \frac{1}{c}\sum_{j=1}^{m}(y_j-\mu_y)^2\right).
\end{aligned}$$

Then we have MLE for this problem as follows:

$$\begin{aligned}
\hat{\psi} &= \bar{x}-\bar{y} \\
\hat{\mu}_y &= \bar{y} \\
\hat{\sigma}_x^2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{c}\sum_{j=1}^{m}(y_j-\bar{y})^2}{m+n}.
\end{aligned}$$

The overall observed information matrix can be calculated by the inverse of the second derivatives.

$$j_{\theta\theta'}(\theta) = \begin{pmatrix}
\frac{n}{\sigma_x^2} & \frac{n}{\sigma_x^2} & \frac{n(\bar{x}-\psi-\mu_y)}{(\sigma_x^2)^2} \\
\frac{n}{\sigma_x^2} & \frac{n}{\sigma_x^2} + \frac{m}{c\sigma_x^2} & \frac{n(\bar{x}-\psi-\mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y}-\mu_y)}{c(\sigma_x^2)^2} \\
\frac{n(\bar{x}-\psi-\mu_y)}{(\sigma_x^2)^2} & \frac{n(\bar{x}-\psi-\mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y}-\mu_y)}{c(\sigma_x^2)^2} & -\frac{m+n}{2(\sigma_x^2)^2} + \frac{1}{(\sigma_x^2)^3}(\sum(x_i-\psi-\mu_y)^2 + \frac{1}{c}\sum(y_j-\mu_y)^2)
\end{pmatrix}.$$

Here is the observed nuisance information matrix:

$$j_{\lambda\lambda'}(\theta) = \begin{pmatrix}
\frac{n}{\sigma_x^2} + \frac{m}{c\sigma_x^2} & \frac{n(\bar{x}-\psi-\mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y}-\mu_y)}{c(\sigma_x^2)^2} \\
\frac{n(\bar{x}-\psi-\mu_y)}{(\sigma_x^2)^2} + \frac{m(\bar{y}-\mu_y)}{c(\sigma_x^2)^2} & -\frac{m+n}{2(\sigma_x^2)^2} + \frac{1}{(\sigma_x^2)^3}(\sum(x_i-\psi-\mu_y)^2 + \frac{1}{c}\sum(y_j-\mu_y)^2)
\end{pmatrix}.$$

Then

$$
\begin{aligned}
j_{\theta\theta'}(\hat{\theta}) &= \begin{pmatrix} \frac{n}{\hat{\sigma}_x^2} & 0 & 0 \\ 0 & \frac{m}{c\hat{\sigma}_x^2} & 0 \\ 0 & 0 & \frac{m+n}{2(\hat{\sigma}_x^2)^2} \end{pmatrix} \\
|j_{\theta\theta'}(\hat{\theta})| &= \frac{mn(m+n)}{2c(\hat{\sigma}_x^2)^4}.
\end{aligned}
$$

Apply Lagrange Multiplier Technique introduced in Section 1.2.5. We have the tilted log-likelihood function:

$$\tilde{l}(\theta) = l(\theta) + \hat{\kappa}(\mu_x - \mu_y - \psi).$$

So that the constrained MLE can be calculated by setting the constrained score functions to be zero. Then

$$
\begin{aligned}
\hat{\mu}_{\psi_y} &= \frac{n}{n+m/c}\bar{x} + \frac{m/c}{n+m/c}\bar{y} - \frac{n}{n+m/c}\psi \\
\hat{\sigma}^2_{\psi_x} &= \frac{\sum(x_i - \psi - \hat{\mu}_{\psi_y})^2 + \frac{1}{c}\sum(y_j - \hat{\mu}_{\psi_y})^2}{m+n} \\
\hat{\kappa} &= \frac{m(\bar{y} - \hat{\mu}_{\psi_y})}{c\hat{\sigma}^2_{\psi_x}}.
\end{aligned}
$$

Now we are able to calculate the determinant of constrained nuisance information matrix

$$|j_{\lambda\lambda'}(\hat{\theta}_\psi)| = \frac{(m+n)(m+cn)}{2c(\hat{\sigma}^2_{\psi_x})^3}.$$

From the log-likelihood function, we can easily determine our canonical parameter $\varphi(\theta) = (\frac{\psi+\mu_y}{\sigma_x^2}, \frac{\mu_y}{\sigma_x^2}, \frac{1}{\sigma_x^2})'$. Then we have

$$\begin{aligned}
\varphi_\theta(\theta) &= \begin{pmatrix} \sigma_x^{-2} & \sigma_x^{-2} & -(\psi+\mu_y)\sigma_x^{-4} \\ 0 & \sigma_x^{-2} & -\mu_y\sigma_x^{-4} \\ 0 & 0 & -\sigma_x^{-4} \end{pmatrix} \\
|\varphi_\theta(\theta)| &= -\sigma_x^{-8} \\
\varphi_\lambda(\theta) &= \begin{pmatrix} \sigma_x^{-2} & -(\psi+\mu_y)\sigma_x^{-4} \\ \sigma_x^{-2} & -\mu_y\sigma_x^{-4} \\ 0 & -\sigma_x^{-4} \end{pmatrix} \\
\varphi^\psi(\theta) &= (\sigma_x^2, -\sigma_x^2, -\psi\sigma_x^2).
\end{aligned}$$

Now we have everything to get our signed log-likelihood ratio statistic r and the standardized maximum likelihood estimate departure calculated in the canonical parameter space Q in formula (2.13) as follows:

$$\begin{aligned}
r &= sgn(\hat{\psi}-\psi)\sqrt{2(m+n)log(\frac{\hat{\sigma}_{\psi_x}}{\hat{\sigma}_x})} \\
Q &= \sqrt{\frac{mn}{m+cn}}\frac{\hat{\sigma}_x^2}{\hat{\sigma}_{\psi_x}^3}(\hat{\psi}-\psi).
\end{aligned}$$

Therefore, the p-value function for ψ can be obtained by using either the Barndorff-Nielsen approximation (2.8) or the Lugannani-Rice approximation (2.9) with r and Q calculated above.

4.2 Main Result for the Skovgaard Method

The expression for r^* derived by Skovgaard in 1996 and 2001 has the same expression of r but different Q which has formula (2.30). From the calculation in the above Section, we know the observed Fisher information matrix $j_{\theta\theta'}(\hat{\theta})$ and constrained observed Fisher information matrix $j_{\lambda\lambda'}(\hat{\theta}_\psi)$. In order to get Q, we need to calculate the two covariance matrices, $\hat{S}$ and $\hat{q}$, and the expected Fisher information matrix $i(\hat{\theta})$. As discussed in Section 2.3, all of them could be estimated by Kullback-Leibler distance.

We know

$$\begin{aligned} KL(\theta_1, \theta_2) &= E_{\theta_1}\left[\log \frac{f(x,y;\theta_1)}{f(x,y;\theta_2)}\right] \\ &= E_{\theta_1}\left[\ell(\theta_1) - \ell(\theta_2)\right]. \end{aligned}$$

For the Behrens-Fisher problem, let $\theta_1 = (\psi_1, \mu_{y1}, \sigma_{x1}^2)$ and $\theta_2 = (\psi_2, \mu_{y2}, \sigma_{x2}^2)$, then we have

$$\begin{aligned} \ell(\theta_1) - \ell(\theta_2) = -\log \frac{\sigma_{x1}^2}{\sigma_{x2}^2} \quad & -\frac{1}{2\sigma_{x1}^2}\left[x^2 + (\psi_1 + \mu_{y1})^2 - 2(\psi_1 + \mu_{y1})x\right] \\ & -\frac{1}{2c\sigma_{x1}^2}\left[y^2 + \mu_{y1}^2 - 2\mu_{y1}y\right] \\ & +\frac{1}{2\sigma_{x2}^2}\left[x^2 + (\psi_2 + \mu_{y2})^2 - 2(\psi_2 + \mu_{y2})x\right] \\ & +\frac{1}{2c\sigma_{x2}^2}\left[y^2 + \mu_{y2}^2 - 2\mu_{y2}y\right]. \end{aligned}$$

Note

$$E_{\theta_1}[x] = \mu_{x1} = \psi_1 + \mu_{y1}$$

$$E_{\theta_1}[x^2] = \sigma^2_{x1} + (\psi_1 + \mu_{y1})^2$$

$$E_{\theta_1}[y] = \mu_{y1}$$

$$E_{\theta_1}[y^2] = c\sigma^2_{x1} + \mu^2_{y1}.$$

Therefore, we have the expression of Kullback-Leibler distance:

$$KL(\theta_1, \theta_2) = \frac{(\psi_1 + \mu_{y1} - \psi_2 - \mu_{y2})^2}{2\sigma^2_{x2}} + \frac{(\mu_{y1} - \mu_{y2})^2}{2c\sigma^2_{x2}} + \frac{\sigma^2_{x1}}{\sigma^2_{x2}} - \log\frac{\sigma^2_{x1}}{\sigma^2_{x2}} - 1.$$

As discussed in Section 2.3,

$$\begin{aligned} \chi_{10}(\theta_1, \theta_2; \theta_1) &= \frac{\partial}{\partial\theta_1} KL(\theta_1, \theta_2) \\ \chi_{11}(\theta, \theta_1; \theta) &= -\frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_2} KL(\theta_1, \theta_2). \end{aligned}$$

Then we have estimation of $\hat{q}$, $\hat{S}$ and $i_{\theta\theta'}(\hat{\theta})$ as follows:

$$\begin{aligned} \hat{q} &= \chi_{10}(\hat{\theta}, \hat{\theta}_\psi; \hat{\theta}) \\ \hat{S} &= \chi_{11}(\hat{\theta}, \hat{\theta}_\psi; \hat{\theta}) \\ i_{\theta\theta'}(\hat{\theta}) &= \chi_{11}(\hat{\theta}, \hat{\theta}). \end{aligned}$$

For the Behrens-Fisher problem,

$$
\begin{aligned}
\chi_{10}(\theta_1,\theta_2;\theta_1) &= \frac{\partial}{\partial\theta_1} KL(\theta_1,\theta_2) \\
&= \begin{pmatrix} \frac{1}{\sigma_{x2}^2}(\psi_1+\mu_{y1}-\psi_2-\mu_{y2}) \\ \frac{1}{\sigma_{x2}^2}(\psi_1+\mu_{y1}-\psi_2-\mu_{y2})+\frac{1}{c\sigma_{x2}^2}(\mu_{y1}-\mu_{y2}) \\ \frac{1}{\sigma_{x2}^2}-\frac{1}{\sigma_{x1}^2} \end{pmatrix} \\
\chi_{11}(\theta_1,\theta_2;\theta_1) &= -\frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_2} KL(\theta_1,\theta_2) \\
&= \begin{pmatrix} \frac{1}{\sigma_{x2}^2} & \frac{1}{\sigma_{x2}^2} & \frac{1}{(\sigma_{x2}^2)^2}(\psi_1+\mu_{y1}-\psi_2-\mu_{y2}) \\ \frac{1}{\sigma_{x2}^2} & \frac{c+1}{c\sigma_{x2}^2} & \frac{1}{(\sigma_{x2}^2)^2}(\psi_1+\mu_{y1}-\psi_2-\mu_{y2})+\frac{1}{c(\sigma_{x2}^2)^2}(\mu_{y1}-\mu_{y2}) \\ 0 & 0 & \frac{1}{(\sigma_{x2}^2)^2} \end{pmatrix}.
\end{aligned}
$$

Now we are able to calculate $\hat{q}$, $\hat{S}$ and the expected Fisher information matrix in terms of MLE. Then we have everything to calculate Q from equation (2.30) for the Skovgaard method.

From Figures 4.1, 4.2 and 4.3, we can observe that our third order methods are very accurate even when the sample sizes are extremely small.

Here is the algorithm to perform Monte Carlo simulation studies:

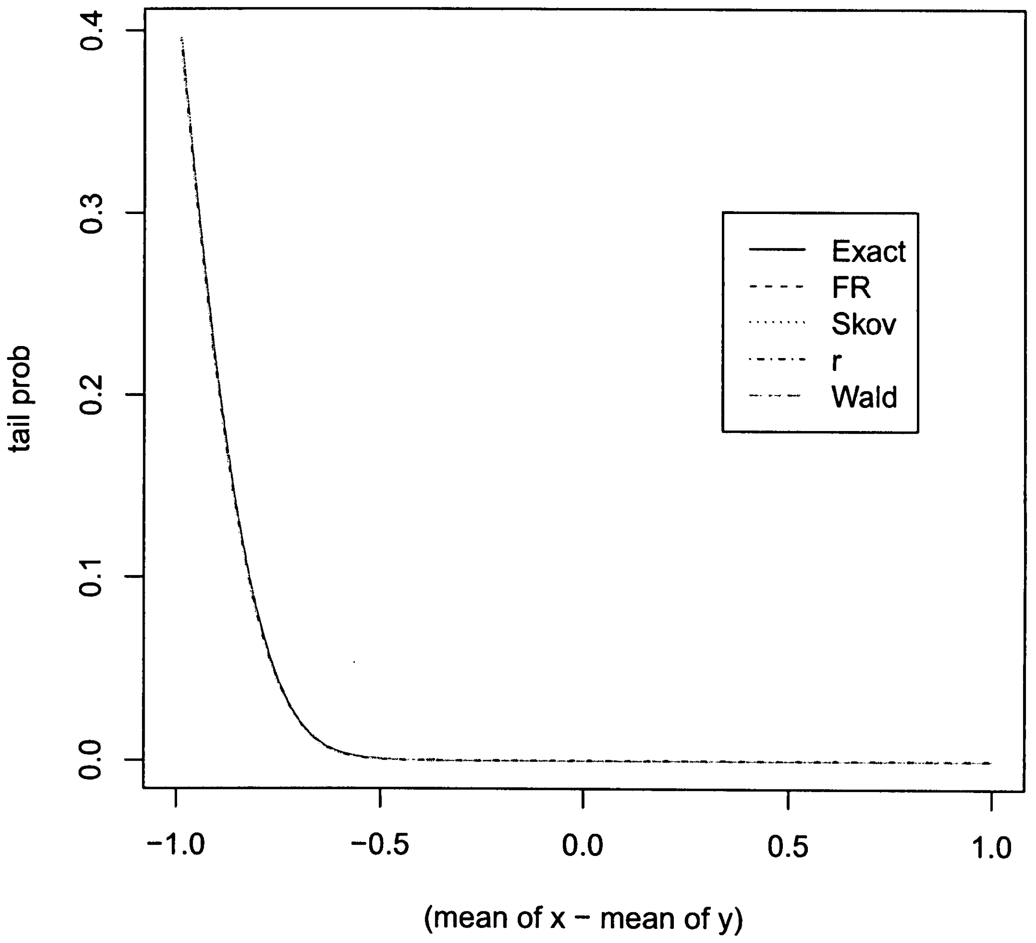


Figure 4.1: $p(\mu_x - \mu_y)$ for independent normal distribution with sample size n=100, m=150

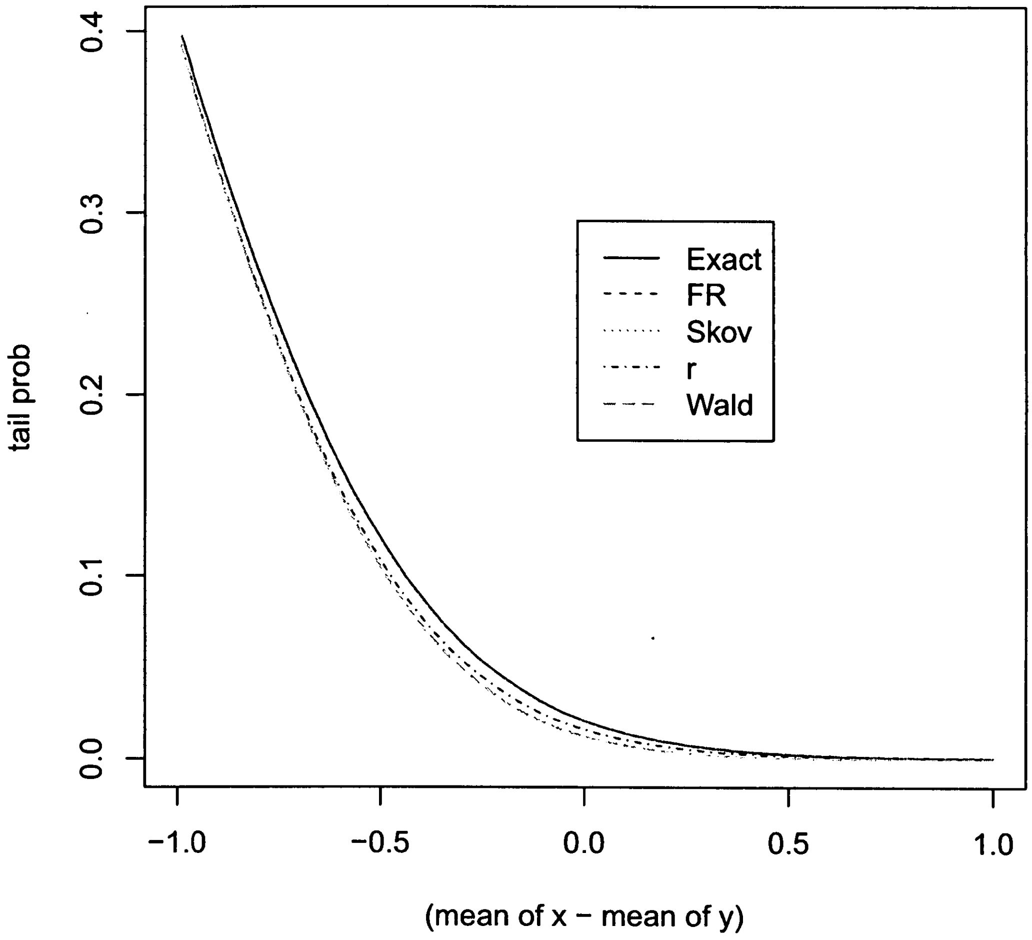


Figure 4.2: $p(\mu_x - \mu_y)$ for independent normal distribution with sample size n=10, m=15

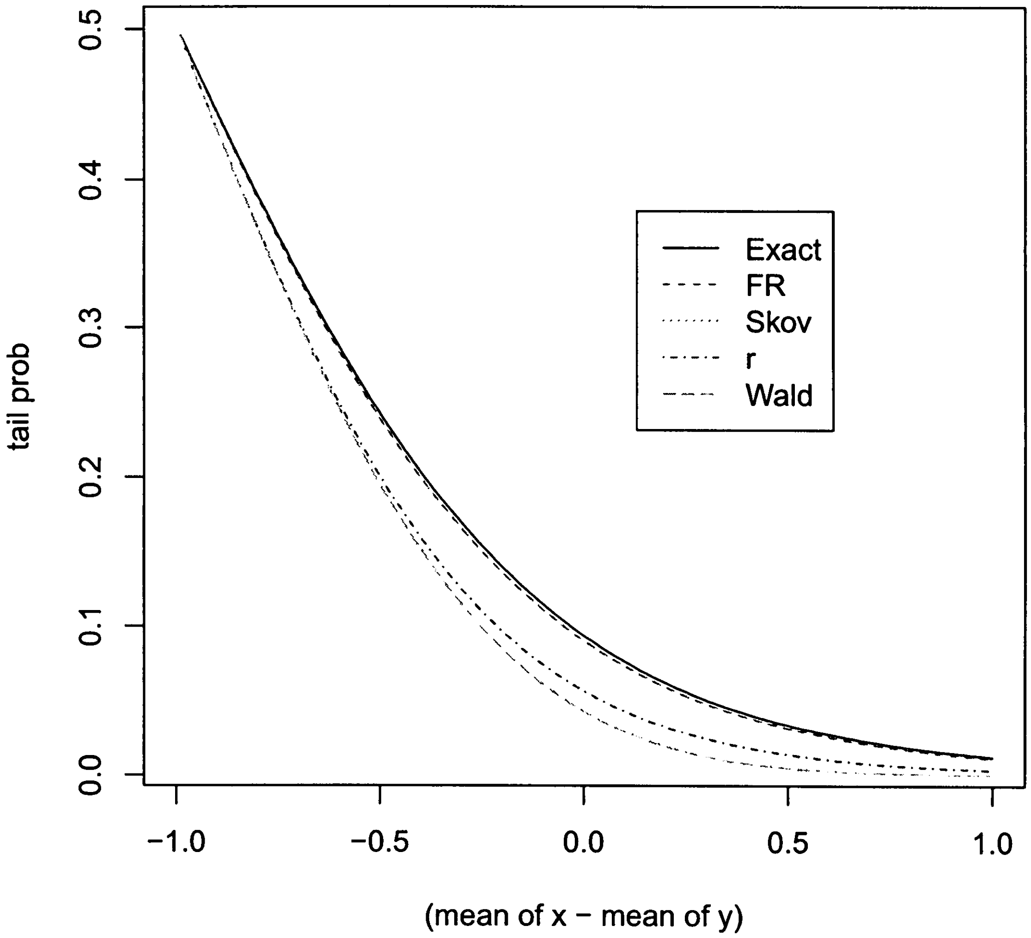


Figure 4.3: $p(\mu_x - \mu_y)$ for independent normal distribution with sample size n=3, m=5

Set up: $N = 10000, n = 100, m = 150, c = 0.5, \mu_x = 0, \mu_y = 1, \sigma_x^2 = 1, \sigma_y^2 = c * \sigma_x^2$.

Aim p-value for testing $\mu_x - \mu_y = \psi$.

Step 1 (a) Simulate sample of size n from normal distribution $N(\mu_x, \sigma_x^2)$.

(b) Simulate sample of size m from normal distribution $N(\mu_y, \sigma_y^2)$.

Step 2 (a) For the Fraser and Reid method: Calculate p-value for $\psi = -1$ from formula (2.8) or (2.9) with r calculated from (2.10) and Q calculated from (2.12) or (2.13). If p-value is less than 0.025, then lower error of the Fraser and Reid method = lower error of the Fraser and Reid method + 1. If p-value is greater than 0.975, then upper error of the Fraser and Reid method = upper error of the Fraser and Reid method + 1.

(b) For the Skovgaard method: Calculate p-value function from formula (2.8) or (2.9) with r calculated from (2.10) and Q calculated from (2.30). If p-value is less than 0.025, then lower error of the Skovgaard method = lower error of the Skovgaard method + 1. If p-value is greater than 0.975, then upper error of the Skovgaard method = upper error of the Skovgaard method + 1.

Step 3 Repeat step 1 and step 2 N times.

Step 4 Report lower error and upper error for both methods.

Run the algorithm multiple times by changing the sample size of n and m to be smaller. The following is the result for both methods: The following two tables show that both of the Fraser and Reid method and the Skovgaard method give pretty accurate approximation compared to the exact confidence intervals even for the sample size to be extremely small.

Table 1: $\mu_x = 0, \mu_y = 1, \sigma_x^2 = 1, \sigma_y^2 = c\sigma_x^2, n = 100$ and $m = 150$

c (known)	Method	Lower Error	Upper Error	Central Coverage
5	r	0.0287	0.0218	0.9495
	Fraser and Reid	0.0282	0.0213	0.9505
	Skovgaard	0.0282	0.0213	0.9505
3	r	0.0289	0.0211	0.9500
	Fraser and Reid	0.0283	0.0205	0.9512
	Skovgaard	0.0283	0.0205	0.9512
1	r	0.029	0.0226	0.9484
	Fraser and Reid	0.0287	0.0221	0.9492
	Skovgaard	0.0287	0.0221	0.9492
0.5	r	0.0306	0.0218	0.9476
	Fraser and Reid	0.0299	0.0212	0.9489
	Skovgaard	0.0299	0.0212	0.9489

Table 2: $\mu_x = 0, \mu_y = 1, \sigma_x^2 = 1, \sigma_y^2 = c\sigma_x^2, n = 10$ and $m = 15$

c (known)	Method	Lower Error	Upper Error	Central Coverage
5	r	0.0336	0.0314	0.9350
	Fraser and Reid	0.0266	0.0239	0.9495
	Skovgaard	0.0266	0.0239	0.9495
3	r	0.0337	0.0315	0.9348
	Fraser and Reid	0.0262	0.0240	0.9498
	Skovgaard	0.0262	0.0240	0.9498
1	r	0.0339	0.0302	0.9359
	Fraser and Reid	0.0273	0.0253	0.9474
	Skovgaard	0.0273	0.0253	0.9474
0.5	r	0.0333	0.0309	0.9358
	Fraser and Reid	0.0266	0.0256	0.9478
	Skovgaard	0.0266	0.0256	0.9478

Table 3: $\mu_x = 0, \mu_y = 1, \sigma_x^2 = 1, \sigma_y^2 = c\sigma_x^2, n = 3$ and $m = 5$

c (known)	Method	Lower Error	Upper Error	Central Coverage
5	r	0.0493	0.0499	0.9008
	Fraser and Reid	0.0260	0.0263	0.9477
	Skovgaard	0.0260	0.0263	0.9477
3	r	0.0498	0.05	0.9002
	Fraser and Reid	0.0263	0.0255	0.9482
	Skovgaard	0.0263	0.0255	0.9482
1	r	0.0517	0.0497	0.8986
	Fraser and Reid	0.026	0.0255	0.9485
	Skovgaard	0.026	0.0255	0.9485
0.5	r	0.0514	0.0492	0.8994
	Fraser and Reid	0.0256	0.025	0.9494
	Skovgaard	0.0256	0.025	0.9494

Monte Carlo simulation results show that the Fraser and Reid method and the Skovgaard method give indistinguishable results whereas the result from the first order method, especially for extremely small sample size, is not satisfactory.

4.3 Summary

The likelihood based third order methods to obtain inference for the difference of two independent normal means with known ratio of variances are proposed. Monte Carlo simulation results showed that the the Fraser and Reid method and the Skovgaard method give much better estimation when the sample sizes are small and they are almost indistinguishable. Schechtman & Sherman (2007) method is tailored made for this particular problem and cannot be applied to the case where the ratio of variances is unknown. However, the proposed methods can still be applied to the unknown ratio of variance case.

5 Discussion and Future Work

The third order likelihood based inferences on a scalar parameter of interest for location-scale family models are discussed and developed in this thesis. Most of researchers use the Fraser and Reid method, and some others use the Skovgaard method. But nobody really think why they choose one or the other. In my thesis, I compared these two methods theoretically and numerically for location-scale models. The Fraser and Reid method is easy to calculate, especially for full exponential models and in simple transformation models. This method is theoretically complicated, but computationally more efficient in terms of programming. But it involves the sample space derivatives and the sample space derivatives are only defined when an ancillary statistic is specified and the overall MLE is sufficient. And sometimes it is not available to calculate, for example, in discrete case. The Skovgaard method does not require specification of the ancillary statistic or its tangent vector V, so it is applicable to discrete distribution. And it is easier to understand. But the expected Fisher information matrix, together with the two covariances are of the same computational complexity. Also the Skovgaard method is not commonly used in current literature. Both proposed methodologies require reasonable computational complexity and exhibits high accuracy for relatively small data set. Both methods have advantages and disadvantages. And in terms of numerical accuracy in approximating p-value, the two methods give almost indistinguishable results when the model is a location-scale model.

Based on the current work, there are several possible directions that research could be extended to:

Firstly, the model I am dealing with is from a continuous distribution. However, when the distribution is of a discrete nature, the ancillary direction V in the Fraser and Reid method cannot be obtained by differentiation. Instead, we could try the Fraser and Reid method by differencing and compare the results with the Skovgaard method.

Secondly, models with Gaussian error structure are widely studied as normal distribution is a simple and reasonable choice for the error term. Alternative to the normal distribution, the error terms can be assumed to follow other distributions, such as Student t distribution. Therefore, similar studies can also be performed for models with non-Gaussian errors.

Thirdly, the third order methods could be applied to reliability problems. For example, $P(X < Y)$ where X and Y are independent random variables that follow known parametric distibutions. This kind of problem is widely considered in engineering, medical studies, economics and finance, for example, Wong (2012). We could use the Skovgaard method comparing with the Fraser and Reid method to illustrate the accuracy of our proposed methods.

Bibliography

[1] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

[2] Barndorff-Nielsen, O.E. (1980). Conditionality resolutions. *Biometrika*, **67**, 293-310.

[3] Barndorff-Nielen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimate. *Biometrika*, **70**, 343-365.

[4] Barndorff-Nielsen, O.E. (1986). Inference on full and partial parameters based on the standardized signed log-likelihood ratio. *Biometrika*, **73**, 307-322.

[5] Barndorff-Nielsen, O.E. and Chamberlin, S.R. (1991). An ancillary invariant modification of the signed log likelihood ratio. *Scandinavian Journal of Statistics*, **18**, 341-52.

[6] Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and saddlepoint ap-

proximations with statistical applications (with discussion). *Journal of the Royal Statistical Society B*, **41**, 279-312.

[7] Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman and Hall.

[8] Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Aspptotics.* London: Chapman and Hall.

[9] Barndorff-Nielsen, O.E. and Sorensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic process. *International Statistical Review*, **62**, 133-165.

[10] Barndorff-Nielsen, O.E. and Wood, A.T.A. (1998). On large deviations and choice of ancillary for p* and r*. *Bernoulli*, **4**, 35-63.

[11] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

[12] Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

[13] Daniels, H.E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, **25**, 631-635.

[14] Davison, A.C. and Hinkley, D.V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, **75**, 417-431.

[15] DiCiccio, T.J., Field, C.A. and Fraser, D.A.S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika*, **77**, 77-95.

[16] Doganaksoy, N. and Schmee, J. (1993). Comparisons of approximate confidence intervals for distributions used in life-data analysis, *Technometrics*, **35**, 175-184.

[17] Durbin, J. (1980). Approximations for densities for sufficient estimators. *Biometrika*, **67**, 334-487.

[18] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, **9**, 586-596.

[19] Ferguson, T.S. (1996). *A Course in Large Sample Theory.* Chapman-Hall, New York.

[20] Fisher, R.A. (1921). On the "probable error"of a coefficient of correlation deduced from a small sample. *Metron* 1 3-32. [CP14 in Bennett (1971), vol. 1.].

[21] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics.

Philos. Trans. Roy. Soc. London Ser. A 222 309-368. [CP18 in Bennett (1971), vol. 1.].

[22] Fraser, D.A.S. (1988). Normed likelihood as saddlepoint approximation. *Journal of Multivariate Analysis.* **27**, 181-193.

[23] Fraser, D.A.S. (1988). *Encyclopedia of Statistzcol Sciences.* Structural Inference. Wiley, New York.

[24] Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika,* **77**, 65-76.

[25] Fraser, D.A.S., Lee, H.S. and Reid, N. (1990). Nonnormal linear regression: an example of significance levels in high dimensions. *Biometrika*, **77**, 333-341.

[26] Fraser, D.A.S. and Reid, N. (1990). From multiparameter likelihood to tail probability for a scalar parameter. *Technical report*, 90-03, University of Toronto.

[27] Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematics*, **7**, 33-53.

[28] Fraser, D.A.S. and Reid, N. (1996). Ancillary information for statistical inference. *Technical report*, University of Toronto, Department of Statistics.

[29] Fraser, D.A.S. and Reid, N. (2010). Mean loglikelihood and higher-order approximations. *Biometrika*, **97**, 159-170.

[30] Fraser, D.A.S., Reid, N. and Wong, A. (1991). Exponential linear models: a two pass procedure for saddlepoint approximation. *Journal of the Royal Statistical Society B*, **53**, 483-92.

[31] Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika*, **86**, 249-264.

[32] Jing, B.Y., Shao, Q.M. and Zhou, W. (2004). Saddlepoint Approximation for Student t-Statistic With No Moment Conditions, *Annals of Statistics*, **32**, 2679-2711.

[33] Kent, J.T. (1982). Robust properties of likelihood ratio test. *Biometrika*, **69**, 19-27.

[34] Lang, S. (1973). *Calculus of Several Variables.* Reading, MA: Addison-Wesley.

[35] Lehmann, E.L. (1983). *Theory of point estimation.* Wiley, New York.

[36] Lieblein, J. and Zelen, M. (1956). Statistical Investigation of the Fatigue Life of Deep Groove Ball Bearings. *Journal of Research, Nat. Bur. of Standards*, **57**, 273-316.

[37] Lin, J.T. (1988). Approximating the Cumulative Chi-square Distribution and Its Inverse, *The Statistician*, **37**, 3-5.

[38] Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution function of the sum of independent variables. *Advanced Applied Probability*, **12**, 475-490.

[39] McCullagh, P. (1987). *Tensor Methods in Statistics.* London: Chapman and Hall.

[40] Maity, A. and Sherman, M. (2006). The two sample *t*-test with one variance unknown. *The American Statistician*, **60**, 163-166.

[41] Miller, I. and Miller, M. (2003). *John E. Freund's Mathematical Statistics with Applications (7th Edition).* Prentice Hall.

[42] Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, **231**, 289-337.

[43] Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journals of the Royal Statistical Society B*, **54**, 701-737.

[44] Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Camb. Phil. Soc.*, **44**, 50-7.

[45] Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, **3**, 213-227.

[46] Rekkas, M. and Wong, A. (2008). Implementing likelihood-based inference for fat-tailed distributions. *Finance Research Letters*, **5**, 32-46.

[47] Satterthwaite, F.E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **6**, 110114.

[48] Schechtman, E. and Sherman M. (2007). The two-sample t-test with a known ratio of variances. *Statistical Methodology*, **4**, 508-514.

[49] Schervish, M.J. (1997). *Theory of Statistics.* Springer, New York.

[50] Sen, P. K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications.* New York, NY: Chapman and Hall.

[51] Severini, T.A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika*, **86**, 23548.

[52] Severini, T.A. (2000). *Likelihood Methods in Statistics.* Oxford: Oxford University Press.

[53] Skovgaard, I.M. (1996). An Explicit Large-Deviation Approximation to One-Parameter Tests. *Bernoulli*, **2**, 145-165.

[54] Skovgaard, I.M. (2001). Likelihood asymptotics. *Scand. J. Statist.*, **28**, 332.

[55] Strimmer, K. (2010). *Statistical Thinking: Intrductioon to Probabilitic Data Analysis*, Leipzig: university of leipzig.

[56] Wald, A. (1943). Tests of Statistical Hypothesis When the Number of Observations is large. *Transaction of the A.M.S.*, **54**, 426-482.

[57] Wilks, S.S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.*, **1**, 60-2.

[58] Wong, A. (1992). Converting observed likelihood to levels of significance for transformation models. *Communications in Statistics*, **21**, 2809-2823.

[59] Wong, A. and Wu, Y. (2008). Likelihood analysis for the difference in means of two independent normal distributions with one variance unknown, *Journal of Statistical Research*, **42**, 17-35.

[60] Wong, A. (2012). Interval estimation of $P(Y < X)$ for generalized Pareto distribution, *Jounal of Statistical Planning and Inference*, **142**, 601-607.

[61] Wong, A., Chang, F. and Rekkas, M. (2013). Improved likelihood-based inference for the MA(1) model. *Journal of Statistical Planning and Inference*, **143**, 209-219.