SADDLEPOINT APPROXIMATION TO CERTAIN CUMULATIVE DISTRIBUTION FUNCTIONS

ALMUTAIRI ALYA OMAR

UNIVERSITI SAINS MALAYSIA

2015

SADDLEPOINT APPROXIMATION TO CERTAIN CUMULATIVE DISTRIBUTION FUNCTIONS

By

ALMUTAIRI ALYA OMAR

Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

February 2015

ACKNOWLEDGEMENTS

First, I would like to express my deepest appreciation to my supervisor, Professor Dr. Low Heng Chin and thank her for invaluable ideas, helpful comments, great support, continuous encouragement and kind patience. Without her help and advice, I cannot imagine completing my thesis. I feel very lucky to have worked with her and have benefited from her invaluable experience.

Sincere thanks to the Dean of the School of Mathematical Sciences and the staff in the Universiti Sains Malaysia especially the staff in the School of Mathematical Sciences for their assistance and cooperation. Finally, my most heartfelt thanks go to my dearest parents and husband, for their selfless love and support.

TABLE OF CONTENTS

2 LITERATURE REVIEW

3 ESTIMATION OF THE NEW CDFs USING SADDLEPOINT APPROXIMATIONS FOR LINEAR COMBINATION MODELS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

H_L	Alternative Hypothesis
Z'	Conjugate Vector for Z
Φ	Cumulative Standard Normal Distribution
F_X	Cumulative Distribution Function.
$f_X(x)$	Common Distribution
$k_1, k_2,...$	Cumulants
$cov(X_j, X_i)$	Covariance
$K_X(s)$	Cumulant-generating Function
I_x	Continuity-Corrected or Offset Value of k
μ	Expected Value (mean) of Random Variable X
■	End of Proof
N	Factorial
$\hat{P}_{r_1}$	First Approximation
$K'(\hat{s}, \hat{t})$	Gradient with Respect to both s and t
K'_s	Gradient with Respect to the Components s alone
τ_χ	Interior of the Convex Hull of the Support χ
I_χ,	Interior of the Support x of $f(x)$
$K(s,t)$	Joint Cumulant Generating
Lc_N	Linear Combinations of Random Variables
T_N	Linear Rank Statistic
Q_1	Lower quartile
U	Mann-Whitney Statistic
$M_X(s)$	Moment Generating Function
M	Median Test
$f_X^{*n}(s)$	*n*-fold Convolution
$F^{n^*}(r)$	*n*-fold Convolution of the Distribution $F(r)$
H_0	Null Hypothesis
R^m	Real numbers with *m*-dimensional
(X, Y)	Random Vector with a Distribution in R^m with $\dim(X) = m_x$, $\dim(y) = m_y$

$\hat{f}$	Saddlepoint Density / Mass Function
$\hat{P}_{r_2}$	Second Continuity Corrected Approximation
$\emptyset$	Standard Normal Probability Density
h	Step of the Lattice Distribution
$K''_{Y(t)}(\hat{s})$	Second Derivative of the Cumulant-Generating Function
Z	Standard normal distribution
$skew(X)$	Skewness
$\hat{P}_{r_3}$	Third Approximation
Q_3	Upper quartile
VA	Van der Waerden Test
σ^2	Variance of Random Variable *X*
W	Wilcoxon Rank-Sum Test Statistic
R_N	Weighted Random Sum
c_i	Weights or Scores
$X \sim$ F	*X* has Distribution *F*
$X \sim$ f	*X* has Distribution with Density *f*

LIST OF ABBREVIATIONS

CDF	Cumulative Distribution Function.
CF	Characteristic Function.
CGF	Cumulate Generating Function.
$sgn(\hat{s})$	Captures the Sign ± for $\hat{s}$.
CLT	Central Limit theorem
Dim	Dimensional
$E(X)$	Expected Value (mean) of Random Variable X
i.i.d.	Independent and Identically Distributed
IQR	Interquartile Range
MGF	Moment Generating Function
MSE	Mean Square Error
Min	Minimum
PDF	Probability Density Function
PGF	Probability Generating Function
P-value	Probability Value
PMF	Probability Mass Function
P_r (X)	Probability of Event X
RE	Relative Error
$V(X)$	Variance of Random Variable X
vs.	Versus
MSE (1)	The mean square error for the saddlepoint versus the exact
MSE (2)	The mean square error for the normal approximation versus the exact
RMA	Robust Multi-array Average

LIST OF PUBLICATIONS

1. Al Mutairi Alya O. and Low Heng Chin (2013). Saddlepoint Approximation to Cumulative Distribution Function for Poisson–Exponential Distribution. *Modern Applied Science journal, Canadian centre of science and education*, Vol. 7, No. 3, p 26-33, ISSN 1913-1844.

2. Al Mutairi Alya O. and Low Heng Chin (2013). Saddlepoint Method to Cumulative Distribution Function for Poisson-Binomial Model. *Modern Applied Science journal, Canadian centre of science and education,* Vol. 7, No. 6, p 101-105, ISSN 1913-1844.

3. Al Mutairi Alya O. and Low Heng Chin (2013). Saddle Point Approximation to Cumulative Distribution Function for Damage Process. *Journal of Asian Scientific Research, Asian Economic and Social Society.* Vol.3, No. 5, pp. 485-492.

4. Alya O. Al Mutairi & Heng Chin Low (2015) *P*-Value for two-Sample Linear Rank Tests Using Permutation Simulations and the Normal Approximation Method, *Communications in Statistics - Simulation and Computation*, 44:3, 819-826, DOI: 10.1080/03610918.2013.794287.

5. Al Mutairi Alya O. and Low Heng Chin (2015). Improved measures of the spread of data for some unknown complex distributions using saddlepoint approximations, *Communications in Statistics - Simulation and Computation*. ISSN 0361-0918. *Published Online*

6. Al Mutairi Alya O. and Low Heng Chin (2014). Estimations of the central tendency measures of the random-sum Poisson-Weibull distribution using saddlepoint approximation, *Journal of Applied Sciences*. Volume 14, Number 16, pp. 1889-1893 PISSN 1812-5654. EISSN 1812-5662.

7. Al Mutairi Alya O. and Low Heng Chin (2014). Saddlepoint Approximation to the Cumulative Distribution Function for Some Difficult and Known Linear Combination of Random Variables, *Canadian Journal of Pure and Applied Sciences*. Vol.8, No. 3. pp. 3081-3089

8. Al Mutairi Alya O. and Low Heng Chin (2015). Saddlepoint Approximation to the Cumulative Distribution Function for a Linear Combination of Convolution Gamma–Exponential Models, *Canadian Journal of Pure and Applied Sciences*. Vol.9, No. 1. pp. 3259-3265

ANGGARAN TITIK PELANA FUNGSI TABURAN LONGGOKAN TERTENTU

ABSTRAK

Anggaran adalah sangat penting kerana kadangkala adalah tidak mungkin untuk mendapatkan perwakilan yang tepat dari fungsi ketumpatan kebarangkalian (FKK) dan fungsi taburan longgokan (FTL). Melibatkan untuk benar (tepat) perwakilan yang mungkin pengiraan dalam beberapa kes memudahkan rawatan analisis. Dalam kajian ini, yang dikenali ekor anggaran, pengiraan kebarangkalian untuk kes univariat telah dilanjutkan, termasuk kes bersyarat univariat. Pendekatan pertama (jumlah rawak $S_{N(t)}$) berlaku untuk kedua-dua pemboleh ubah berterusan dan diskret (yang dibincangkan dalam perkiraan jumlah rawak Poisson-Chi kuasa dua dan pemboleh ubah rawak Poisson-eksponen, yang mempunyai pengedaran yang terus menerus, dan rawak diskret jumlah model binomial Poisson-Negatif), termasuk data sebenar daripada data. Pendekatan kedua (convolutions Gamma dan pembolehubah rawak eksponen L_N) juga berlaku untuk lanjutan convolutions ini, yang amat sukar untuk diperoleh. Ia juga dopat membandingkan pengiran eksponen dengan kadar yang tepat dan biasa. Kaedah ini digunakan untuk contoh-contoh tertentu, termasuk data sebenar diatur dari pembetulan latar belakang untuk Illumina BeadArray. Pendekatan ketiga, termasuk perkiraan anggaran bersyarat, menggunakan anggaran ganda, walaupun cara ini seolah-olah menjadi sangat sukar untuk didapatkan, jika mereka lebih layak. Pertengahan nilai-P telah diturunkan menggunakan saddlppoint bersyarat saddlppoint untuk statistik peringkat-tertib untuk dua masalah sampel, dengan mempertimbangkan contoh-contoh nyata dari Buku Data Kecil Set. Sebagai kesimpulan, kaedah anggaran memberikan pendekatan yang sangat tepat untuk CDF dan melampaui pendekatan biasa untuk tiga statistik sulit dan tidak diketahui yang dinyatakan di atas.

SADDLEPOINT APPROXIMATION TO CERTAIN CUMULATIVE DISTRIBUTION FUNCTIONS

ABSTRACT

Approximations are very important because it is sometimes not possible to obtain an exact representation of the probability density function (PDF) and the cumulative distribution function (CDF). Even when true (exact) representations are possible, approximations, in some cases, simplify the analytical treatments. In this study, the known saddlepoint tail probability approximations for univariate cases have been extended, to include univariate conditional cases. The first approximation (the random sum $S_{N(t)}$) is applied to both continuous and discrete variables (will discuss approximations within the random sum Poisson–Chi square and Poisson–Exponential random variables, which have a continuous distribution, and the discrete random sum Poisson–Negative Binomial model), including real data from the annual maximum daily rainfall data. The second approximation (convolutions of Gamma and Exponential random variable, L_N) is also applied to the extension of these convolutions, which are extremely difficult to obtain. These exponential computations are also compared with the exact and normal approximations. This method is applied to a real data set from a background correction for Illumina BeadArray. The third approximation, including conditional saddlepoint approximations, uses the double saddlepoint, although this method seems to be very difficult to obtain, even if they are feasible. The mid-*P*-value was derived using conditional saddlepoint approximation for rank-order statistics for two sample problems, considering real data. As a conclusion, the saddlepoint methods provide very accurate approximation for the CDFs and surpass normal approximation for the three difficult and unknown statistics mentioned above.

CHAPTER 1

INTRODUCTION

1.1 Introduction

The meaning of approximation refers, in a particular way, to the representation of "something" by "something else" that might be a useful replacement for the "something".

Approximations are very important because it is sometimes not possible to precisely represent such a "something". Even when a true (exact) representation is possible, the approximations sometimes simplify the analytical treatments. In the applied problems, approximations are used constantly. For example, many scientific works and nearly all statistical analyses are based on mathematical models that are essentially approximations.

Most statistically applied problems depend on approximations to PDFs or CDFs derived from asymptotic theory. Exact formulas are rarely obtained in a form that is simple enough to be used directly. Typically, the approximations to be used depend on certain calculations drawing upon the theory of probability. This theory can be combined with asymptotic methods from the analysis and development of asymptotic expansions. One such example is the saddlepoint approximation.

Compared with other asymptotic approximations, saddlepoint approximations (Daniels, 1954) have the advantage of always generating probabilities, being very accurate in the tails of the distribution, and being accurate with small samples, sometimes even with only one observation.

Recently, modern statistical methods, such as linear combinations of random variables, use models that require the computation of probabilities from complicated distributions, which can lead to intractable computations. These linear combinations occur in a wide range of fields. Unfortunately, in most cases, a closed, analytic expression for the probability distribution function (PDF) and the cumulative distribution function (CDF) have difficult and complex calculations; in some of other cases, the exact expression is not yet known.

In this study, saddlepoint approximation methods have been proven to be useful for a range of problems, and hence have attracted considerable interest. With possible estimations to ensure that the saddlepoint density method integrates to unity over its support, the accuracy of the approximation can sometimes be improved. In certain cases, this function can even reproduce the exact density.

The background of the study, problem identification, objectives of the study, significance of the study, scope of the study and organization of the thesis are presented in this chapter.

1.2 Background of the study

The need to analyze distributions of linear combinations of random variables arises in many fields of research, such as biology, seismology, risk theory, insurance application and health science. A mathematically linear combination is expressed as shown in the expression below:

$$Lc_N = c_1X_1 + c_2X_2 + c_3X_3 + \cdots + c_NX_N, \tag{1.1}$$

Here, a set of coefficients, c_1 through c_N is available, that are multiplied by the corresponding variables, X_1 through X_N. The first term, c_1 times X_1, is added to c_2 times X_2, and so on, up to the variable X_N (Ali and Obaidullah, 1982).

This process can be expressed as the sum of $i = 1,2,\ldots,N$ of the terms c_i times X_i. The selection of the coefficients c_1 through c_N very much depends on the application of interest and the types of scientific questions that are addressed.

1.3 The linear combination of random variables

This study, discuss saddlepoint approximations to cumulative distribution functions for the linear combination of random variables (1.1) in three different cases, as presented below.

1.3.1 The linear combination of the sum of independent random variables when N is a random variable

The distributions considered in this study result from the combination of two independent distributions in a particular way. When all $c_i = 1$, this process is termed "generalization" by some authors (Johnson *et al.*, 2005), though the term "generalized" is greatly overused in statistics. This distribution includes the sums of independent and identically distributed (i.i.d.) random variables, $\{X_i\}$, with random index N, independent of X_i's,.

Definition 1.1

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables with a common distribution $f_X(x)$. Let N be a discrete random variable that takes the values $1,2,3,\ldots$ and let $X_i's$ be independent of N, and c_i be non-negative real numbers. The sum

$$R_N = c_1X_1 + c_2X_2 + c_3X_3 + \cdots + c_NX_N \tag{1.2}$$

is called the weighted random sum (Kasparaviciute and Leonas, 2013).

Such sums have a wide range of applications in branching processes, damage processes and risk theory. A common application of the random sum is the total claim amount presented to an insurance company, where N is the number of claims and the $X_i's$ are the individual claims, which are assumed to be independent.

In general, random sums are extremely difficult to investigate; therefore, approximation techniques are frequently employed. Saddlepoint methods overcome this difficulty while providing us with an influential tool for obtaining precise expressions for distribution functions that are still unknown in the closed form. In addition, these methods roughly surpass other techniques, in terms of calculating expenses but exceed no other methods, in terms of accuracy.

This study, discuss as approximations of the unknown difficult random sum Poisson-Chi square and Poisson-Exponential random variables, which have a continuous distribution, and of the random sum Poisson-Negative Binomial random variable, which has a discrete distribution.

However, the study shows that the saddlepoint approximation method is not only quick, dependable, stable and accurate enough for general statistical inference but it is also applicable without deep knowledge of probability theory.

1.3.2 The linear combination of the sum of independent random variables when N is a constant

Linear combinations of convoluted random variables occur in a wide range of fields. In most cases, the exact distribution of these linear combinations is extremely difficult to determine, and the normal approximation usually performs very badly for these complicated distributions. A better method of approximating linear combination distributions involves the additional use of saddlepoint approximation.

Saddlepoint approximation is able to provide accurate expressions for distribution functions that are unknown in their closed forms. This method not only yields an accurate approximation near the centre of the distribution but also controls the relative error in the far tail of the distribution.

Definition 1.2

The probability distribution of the sum of two or more independent random variables is the convolution of their individual distributions. Consider the sum of two independent random variables, Z and Y. The distribution of their sum, $X = Z + Y$, is the convolution of these random variables. Now, let $X_1, X_2, \dots, X_N$ be i.i.d. random variables and $c_1, c_2, \dots, c_N$ be numbers. Thus, the random variable

$$L_N = c_1X_1 + c_2X_2 + c_3X_3 + \cdots + c_NX_N = \sum_{i=1}^{N} c_iX_i \qquad (1.3)$$

is called the linear combination of the convolution random variable.

This study, derive the saddlepoint approximation of the convolution $aZ + bY$, where $a > 0$ and $b > 0$ are real constants and Z, Y denote Gamma and Exponential random variables, respectively, while being distributed independently of each other. Then, the study discusses the generalization of this convolution, which is a linear combination of the convolution Gamma–Exponential, which in turn is still unknown and extremely difficult to obtain. The associated saddlepoint approximation CDFs are derived. The plots for the CDFs and some measures of spread and central tendency for the new distribution are also given.

1.3.3 The linear combination of sum of independent Bernoulli random variables when N is a constant and c_i are scores

The approximation for the distribution function of a test statistic is extremely important in statistics. Many statistical procedures that are applicable to the two sample problems are based on the rank order statistics for the combined samples, and many commonly used two-sample rank tests act as a linear combination of certain indicator Bernoulli random variables for the combined ordered samples.

For the approximation presented in this study, a saddlepoint formula is proposed assuming c_i , $i = 1,2, \dots, N$ are given constants called weights or scores and $\{X_i\}$ are Bernoulli variables.

Definition 1.3

Let $X_1, X_2, \dots, X_m$ and $Y_1, Y_2, \dots, Y_n$ be two independent random samples drawn from populations with the continuous cumulative distribution functions F_X and F_Y, respectively. Let $N = m + n$; then, the statistic

$$T_N = \sum_{i=1}^{N} c_i Z_i \tag{1.4}$$

is called a linear rank statistic, where the $\{c_i\}$ are given constants called weights or scores, $Z_i = 1$ if the i$^{\text{th}}$ sampled value in the combined ordered sample is X and $Z_i = 0$ if it is Y (Gibbons and Chakraborti, 2003). It is noteworthy to mention that the statistic T_N is a linear combination of independent indicator Bernoulli random variables $\{Z_i\}$.

This study examines mid-P-values from the null permutation simulations distributions of tests based on the rank-order statistics for the combined two samples. The permutation simulations may lead to intractable computations apart from small values for the sample size, and the normal approximation may not result in the desired accuracy, particularly when the sample size is small. Saddlepoint approximation can be used to overcome these problems. This method results in a highly accurate approximation without placing constraints or guidelines on the values of the sample sizes.

In the three cases of linear combinations involving random variables presented above, the study used the saddlepoint approximation formula proposed by Daniels (1954, 1987) that has the type developed by Lugannani and Rice (1980) to approximate these unknown difficult statistics based on their moment generating functions, which can be performed using theorems related to these unknown statistics. Then, the study derived the saddlepoint equations that, in some cases, can be solved using numerical methods. By performing some calculations and applying saddlepoint formulas, the cumulative distribution function CDF and probability density function PDF can be obtained for these unknown difficult statistics. Subsequently, the study finds the exact distributions using simulation methods and the mean square error (MSE) to investigate the performance of the saddlepoint approximation for the linear combination models in the three different cases.

1.4 Problem identification

To estimate the unknown distribution of CDFs in the linear combination of random variables, the saddlepoint approximation method is used. The identification of the problem is based on three different cases as follows:

(1) The random sum distribution plays a key role in both probability theory and its applications in biology, seismology, risk theory, meteorology and health science. The statistical significance of this distribution arises from its applicability to real-life situations, in which the researcher often observes only the total amount, say S_N, which is composed of an unknown random number N of random contributions, say $X's$.

In health science, the random sum plays a very important role in many real-life applications. For example, let the number of hotbeds of a contagious disease follow a Poisson distribution with a mean of ψ, and let the number of sick people within the hotbed follow a Negative Binomial distribution. If the goal is to find the probability that the total number of sick people is greater than 70, then the total number of sick people within the hotbed is

$$S_{N_1} = \sum_{i=1}^{N} X_i \tag{1.5}$$

where $X_i \backsim Negative\ Binomial\ (r, m)$ and $N \backsim Poisson(\psi)$.

Another practical application of the random sum is the number of times that it rains in a given time period, say N, which has a Poisson distribution with mean λ. If the amount of rain that falls has an Exponential distribution and if the rain falling in that time period is independent of N, then the total rainfall in the time period is

$$S_{N_2} = \sum_{i=1}^{N} Y_i \tag{1.6}$$

where $Y_i \backsim Exponential(\phi)$ and $N \backsim$ Poisson (λ) .

In fact, the total amounts of the random sums S_{N_1} and S_{N_2} are composed of an unknown random number N of other random contributions, say X or Y which are very complex to analyze. In most cases, the distribution of the random sum is still unknown; in other cases, it is already known but is too complex for the computation of the distribution function, which often becomes too slow for many problems (Johnson *et al.,* 1992). The saddlepoint approximation method can help us gain knowledge for these unknown difficult statistical behaviours.

This study discusses random sum Poisson-Chi square random variables and Poisson-Exponential random variables as continuous distributions and random sum Poisson-Negative Binomial random variables as discrete distributions.

(2) Saddlepoint approximation plays an important role in helping us gain knowledge about unknown difficult distributional behaviours, such as the linear combination of random variables. This study discusses the linear combination of the Gamma-Exponential distribution. This convolution model is widely used in background adjustment methods, including the multiplicative model based expression index (MMBE) (Li and Wong, 2001) and the robust multi-array average (RMA) method (Irizarry *et al.,* 2003). This convolution model is given by

$$L = S + Y \tag{1.7}$$

Assuming that S and Y are both independent. Following the RMA model, S is assumed to have an Exponential distribution with mean θ, and Y represents a Gamma distribution (α, β).

The exact distribution of the linear combination is derived when S and Y, respectively, are Exponential and Gamma random variables and are independent of each other. Nadarajah and Kotz (2005) found the exact expression for the CDF, but complications arose due to the special functions because the incomplete Gamma, complementary incomplete Gamma and error function can lead to intractable computations. However, in some cases, numerical computation suggests that this finding is incorrect because it differs substantially from the numerical convolution (Butler, 2007).

This thesis investigates the saddlepoint approximations of the convolution $aZ + bY$, where $a > 0$ and $b > 0$ are real constants and Z, Y denote Gamma and Exponential random variables, respectively, and are independently distributed. However, it discusses the linear combination of this convolution (i.e., its extension to convolution for Gamma - Exponential random variables), which is unknown and is extremely difficult to obtain.

(3) Hypotheses testing is one of the most important challenges in manipulating nonparametric statistics. Various nonparametric statistics have been proposed and discussed over the course of many years. Many statistical procedures that are routinely used in empirical science require the computation of P-values. The methods used to determine P-values from the null permutation distributions are

normal approximations and permutation simulations. Permutation simulations are very difficult to obtain, even apart from the small values for the sample size. Normal approximations may lead to inaccurate results, particularly when the sample sizes are small. In such cases, the exact P-value must be estimated using an approximation method. For the approximation presented in this study, the saddlepoint formule proposed by Daniels (1954, 1987) and the type developed by Lugannani and Rice (1980) are used.

In this study, linear combinations of the rank-order statistics for the two sample problems are considered, assuming $c_i, i = 1,2, \dots, N$ are given constants called weights or scores and $\{X_\mathrm{i}\}$ assumes a Bernoulli distribution.

1.5 Objectives of the Study

The objectives of the study are as follows:

(1) To develop new estimators (CDFs) using saddlepoint approximations for the following random sums:

i. Poisson-Chi square random variable, which has a continuous distribution.

ii. Poisson-Negative Binomial random variable, which has a discrete distribution.

(2) To derive the saddlepoint approximation (CDFs) of the convolutions:

i. $aZ + bY$, where $a > 0$ and $b > 0$ are real constants and Z, Y denote Gamma and Exponential random variables, respectively, and are distributed independently of each other.

ii. The linear combination of Gamma - Exponential convolution (i.e., an extension of the convolution Gamma - Exponential random variables).

(3) To find the mid-P-value using saddlepoint approximation for the linear combinations of rank-order statistics for the two sample problems.

1.6 Significance of the Study

Modern statistical methods use models that require the computation of probabilities from complicated distributions, which can lead to intractable computations, such as the linear combinations of random variables in the three different cases described in Section 1.5. These strategies have been widely used in many applications.

The difficulty of determining the exact CDF distributions for the sum of independent random variables appears in many applied problems. Three suggestions for computing the CDF are as follows:

(1) Use an empirical distribution,

(2) Use a normal approximation,

(3) Use a saddlepoint approximation.

Option (1) may lead to intractable computations apart from the small values of the sample size, and option (2) may not result in the desired accuracy, especially when the sample size is small. The saddlepoint option (3) is examined in this study and is shown to result in highly accurate approximations, which are accurate to the tails of the distribution. Further, option (3) is shown to be accurate with small samples, sometimes with only one observation. Another advantage of saddlepoint methods is that the required computational times are essentially negligible compared with those of simulations.

1.7 Scope of the study

In this study, the problem of finding the exact CDFs in the linear combinations of random variables in the three different settings is examined. Saddlepoint approximations are used to solve this problem from three different angles, such as hypothesis testing, random sums and convolutions of random variables, to investigate the performance of the saddlepoint approximation method in many different fields.

1.8 Organization of the thesis

An overview of the saddlepoint approximations and of the linear combinations of random variables in three different settings to address the problem of calculating exact CDFs for these linear combinations is given in Chapter 2. In this study, new estimators are developed. The development of these new estimators and their properties are presented in Chapter 3. Some real applications and numerical comparisons between these estimators and other estimators, such as normal approximations, are performed to investigate the performance of the new estimators, as presented in Chapter 4. New saddlepoint estimators' performances with exact and normal approximations are discussed in Chapter 5. Chapter 6 provides a summary and conclusion of the study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the problem of approximating the density or distribution function of a statistic that has a great practical importance. Saddlepoint approximations can be obtained for any statistic that recognizes a cumulant-generating function (CGF). In principle, knowledge of the CGF allows us to obtain the density and distribution functions using saddlepoint approximation theorems. However, in practice, the complexity of the associated integration can make it very costly (or impossible) to obtain an exact analytic result.

Saddlepoint approximation was introduced to statistical science by Daniels (1954) as a means of obtaining a very accurate approximation of the density of the mean of a sample of independent and identically distributed observations.

Saddlepoint methods provide approximations for densities and probabilities that are very accurate in a wide variety of settings. This accuracy is observed in numerical calculations. The resulting approximation is often more accurate than a normal approximation or even other methods of approximation.

This chapter is designed to describe an elementary motivation of the basic saddlepoint approximation technique with certain relevant information. So the reader may begin using the methods in practical settings with only a minimal theoretical understanding. Additionally, this chapter introduces linear combinations of random variables in three different settings (random sums, linear combinations of the convolutions of random variables and hypothesis testing). It also discusses the properties, such as the mean, variance, the moment generating function (MGF) and the probability generating function (PGF), of these linear combinations of random variables in the next chapter to better clarify the development of saddlepoint approximations.

2.2 Fundamentals of saddlepoint approximations

Among the many tools that have been developed for use in statistics and probability in the history of statistics, possibly the least understood and most important tool is saddlepoint approximation. Its importance is due to its capacity to provide probability approximations with a very high accuracy, surpassing the level suggested by the currently supported theory. Saddlepoint approximations are the least understood because of the difficulty of both the approximation methods themselves and the research papers and books that have been written on the subject (Butler, 2007).

It is noteworthy that Blackwell and Hodges (1959), as well as Barndorff-Nielsen and Cox (1979) considered using saddlepoint approximations for densities by expanding the complex moment-generating function in a Taylor series about a saddlepoint of that function. Although the approximation is very accurate, the formula is very complicated and difficult to apply. For this reason and others, this subject is least understood and often ignored by most researchers.

Recently, saddlepoint approximation has been developed and enhanced. This study aims to make this subject accessible to the widest possible audience by using graduated levels of difficulty without a deep knowledge of probability theory. In addition, it overcome certain problems for estimating CDFs from a different angle based on the same formula for saddlepoint approximation as given in Butler (2007), who introduced useful discussions on this subject.

The saddlepoint approximation was introduced into statistics in 1954 by Henry E. Daniels. The basic result for approximating the density function of the sample mean has been generalized to many situations. The accuracy of this approximation is very good. Daniels (1980) explained that the Normal, Gamma and Inverse Normal are the only cases for which the renormalized saddlepoint approximation can reproduce the exact density in the scalar case. These also happen for Normal and Inverse Normal densities without renormalization.

Lugannani and Rice (1980) proposed a saddlepoint approximation for a cumulative distribution function that is very accurate in the tails. Of course, this is of great interest in testing problems. Other approaches for approximating the tail areas were discussed by Barndorff-Nielsen (1991), and Daniels (1987) who provided a full review of this issue.

Routledge and Tsao (1997) proved that the derivation of Lugannani and Rice's asymptotic expansion for the cumulative distribution function is the same as Daniels's asymptotic approximation for the corresponding density function. A comprehensive review of the application of saddlepoint approximations can be found in Reid (1988, 1991), as well as Field and Ronchetti (1990), while an extensive discussion of saddlepoint methods was provided by Goustis and Casella (1999) and Butler (2007).

Saddlepoint approximations are constructed by performing various operations on the moment-generating function or, equivalently, the cumulant-generating function of a random variable. The Moment Generating Function (MGF) of a random variable X, is defined as

$$M_X(s) = E(e^{sX}) = \begin{cases} \sum e^{sx} P(X = x) & if\ X \quad is\ discrete \\ \int_{-\infty}^{\infty} e^{sx} f(x) dx & if\ X \quad is\ \text{continuous} \end{cases} \tag{2.1}$$

for all real s, if the expectation is defined (Hogg and Craig, 1978).

The cumulant-generating function (CGF) is given by

$$K_X(s) = ln\, M_X(s)\ . \tag{2.2}$$

This leads to $K_X(s) = ln\, E(e^{sX})$. (2.3)

2.2.1 Univariate density and mass function for saddlepoint approximation

Daniels (1954) introduced a saddlepoint formula for the univariate density and mass function, assuming that X is a continuous or discrete random variable with (density or mass) $f(x)$ with MGF $M_X(s)$ and CGF $K_X(s)$. This saddlepoint approximation for $f(x)$ is given as

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi K''(\hat{s})}} \exp\{\, K(\hat{s}) - \hat{s}x \,\} \quad , \tag{2.4}$$

where $\hat{s} = \hat{s}(x)$ is the unique solution to the saddlepoint equation $K'(\hat{s}) = x$ and primes denote derivatives. The approximation is meaningful for values of x that are interior points of $\{x: f(x) > 0\} = \chi$.

2.2.2 Univariate saddlepoint approximation for cumulative distribution function for a continuous random variable

Lugannani and Rice (1980) proposed a saddlepoint approximation to the cumulative distribution function for a continuous random variable X with CDF F, CGF K, and mean μ . The saddlepoint approximation for $F(x)$, as introduced by Lugannani and Rice (1980), is

$$\hat{F}(x) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})(1/\hat{w} - 1/\hat{u}) & if x \neq \mu \\ \dfrac{1}{2} + \dfrac{K'''(0)}{6\sqrt{2\pi K''(0)^{3/2}}} & if x = \mu \end{cases} \tag{2.5}$$

where ϕ and Φ denote the standard normal density and CDF, respectively, and

$$\hat{w} = sgn(\hat{s})\sqrt{2\{\hat{s}x - K(\hat{s})\}}\,, \quad \hat{u} = \hat{s}\sqrt{K''(\hat{s})} \tag{2.6}$$

are functions of x and saddlepoint $\hat{s}$. In this case, $\hat{s}$ is the implicitly defined function of x given as the unique solution to $K^{'}(\hat{s}) = x$, and $sgn(\hat{s})$ captures the sign ± for $\hat{s}$.

The bottom expression in Equation (2.5) defines the approximation at the mean of X or when $\hat{s} = 0$. In this case, $\hat{w} = 0 = \hat{u}$, and the last factor in the top expression of Equation (2.5) is undefined. As $x \to \mu$, the limiting value of the top expression is the bottom expression (Butler, 2007).

2.2.3 Univariate saddlepoint approximation for cumulative distribution function for a discrete random variable

Daniels (1987) presented two such continuity corrected modifications, which are mentioned below, because the discrete CDF approximation must be modified for the expressions for the continuous CDF approximation to achieve high levels of accuracy, assuming that X has CDF $F(k)$ with supportive integers and mean μ.

Rather than considering the CDF value $F(k)$, the right tail probability $P_r(X \geq k)$ or the left-continuous function, this value is approximated, which helps avoid what would otherwise be difficult notational problems.

2.2.3.1 First continuity correction

Suppose that $k \in I_{\chi}$, the interior of the support of X such that the saddlepoint equation can be solved at value k. The first approximation is

$$\hat{P}_{r_1}(X \geq k) = \begin{cases} 1 - \Phi(\hat{w}) - \phi(\hat{w})(1/\hat{w} - 1/\hat{u}_1) & if \quad k \neq \mu \\ \dfrac{1}{2} - \dfrac{1}{\sqrt{2\pi}}\left\{\dfrac{K'''(0)}{6K''(0)^{3/2}} - \dfrac{1}{2\sqrt{K''(0)}}\right\} & if \quad k = \mu \end{cases} \tag{2.7}$$

where

$$\hat{w} = sgn(\hat{s})\sqrt{2\{\hat{s}k - K(\hat{s})\}}, \qquad \hat{u}_1 = \{1 - exp(-\hat{s})\}\sqrt{K''(\hat{s})} \tag{2.8}$$

(Butler, 2007).

$\hat{s}$ solves $K'(\hat{s}) = k$. The expression for $\hat{w}$ in Equation (2.8) agrees with Equation (2.6) in the continuous setting. In Equation (2.8), $\hat{u}_1$ replaces $\hat{u}$ in Equation (2.6).

To understand the difference associated with using $\hat{u}_1$ instead of $\hat{u}$, consider the Taylor series expansion, as

$$e^{-\hat{s}} = \frac{(-\hat{s})^0}{0!} + \frac{(-\hat{s})^1}{1!} + \frac{(-\hat{s})^2}{2!} + \frac{(-\hat{s})^3}{3!} + \cdots$$

then
$$e^{-\hat{s}} = 1 - \hat{s} + \frac{\hat{s}^2}{2} - \frac{\hat{s}^3}{6} + \cdots$$

$$1 - e^{-\hat{s}} = \hat{s} - \frac{\hat{s}^2}{2} + \frac{\hat{s}^3}{6} - \cdots$$

from the continuous setting and the approximations to the first order, $1 - e^{-\hat{s}} \simeq \hat{s}$ then $\hat{u}_1 \simeq \hat{u}$.

However, to the second order, $\hat{u}_1 < \hat{u}$, which implies that $\hat{P}_{r_1}(X \geq k) > 1 - \hat{F}(k)$. Thus, the use of the smaller $\hat{u}_1$ in place of $\hat{u}$ adjusts the tail probability in a direction that is consistent with a continuity correction.

2.2.3.2 Second continuity correction

Define $k^- = k - 0.5 \in I_x$ as the continuity-corrected or offset value of k. The second approximation solves the offset saddlepoint equation

$$K^{'}(\tilde{s}) = k^- \tag{2.9}$$

for the continuity-corrected saddlepoint $\tilde{s}$. Saddlepoint $\tilde{s}$ and value of k^- are used to alter the inputs into the CDF approximation, according to

$$\widetilde{w}_2 = sgn(\tilde{s})\sqrt{2\{\tilde{s}k^- - K(\tilde{s})\}} \tag{2.10}$$

$$\tilde{u}_2 = 2\sinh(\tilde{s}/2)\sqrt{K''(\tilde{s})} \tag{2.11}$$

where $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$.

This leads to the second continuity corrected approximation

$$\hat{P}_{r_2}(X \geq k) = \begin{cases} 1 - \Phi(\widetilde{w}_2) - \phi(\widetilde{w}_2)(1/\widetilde{w}_2 - 1/\tilde{u}_2) & if \quad k^- \neq \mu \\ \dfrac{1}{2} - \dfrac{K'''(0)}{6\sqrt{2\pi}K''(0)^{3/2}} & if \quad k^- = \mu, \end{cases} \tag{2.12}$$

(Butler, 2007).

2.2.3.3 Third approximation

This approximation is denoted as $\hat{P}_{r_3}(X \geq k)$ and uses Equation (2.12) with $\widetilde{w}_2$ from Equation (2.10) and replaces $\tilde{u}_2$ with the following:

$$\tilde{u}_3 = \tilde{s}\sqrt{K^{''}(\tilde{s})} \tag{2.13}$$

(Butler, 2007).

The above approximation is based on the Normal distribution. However, using another basis for the distributions is also possible, as in Gamma and Inverse Normal distributions.

Saddlepoint approximations for conditional densities and mass functions are obtained for the use of two saddlepoint approximations, one for the joint density and the other for the marginal.

These conditional probability approximations are very important because they provide us with alternative methods of computation, perhaps based upon simulation methods, despite seeming to be very complex to present or not feasible. For further information on the methods of conditioning in statistical inference, see Reid (1991).

2.2.4 Double saddlepoint approximation, conditional density and distribution functions

Let (X, Y) be a random vector with a distribution in R^m with $\dim(X) = m_x$, $\dim(Y) = m_{\mathrm{y}}$, and $m_{\mathrm{y}} + m_x = m$. Let all components be continuous, and suppose that there is a joint density $f(x, y)$ with support $(X, Y) \in \chi \subseteq R^m$. The conditional density of Y at y given $X = x$ is defined as

$$f(y|x) = \frac{f(x, y)}{f(x)} \quad . \qquad (x, y) \in \chi \tag{2.14}$$

The double-saddlepoint density or mass function is

$$\hat{f}(y|x) = \frac{\hat{f}(x, y)}{\hat{f}(x)}, \qquad (x, y) \in \tau_\chi \tag{2.15}$$

where $\hat{f}$ is used to indicate a saddlepoint density/mass function and τ_χ is the interior of the convex hull of the support χ. The idea of using two saddlepoint approximations to recover a conditional mass function was introduced by Daniels (1958) and was fully elaborated, as in Equation (2.14), by Barndorff-Nielsen and Cox (1979). Double-saddlepoint densities given in Equation (2.15) may be expressed in terms of the joint CGF of (*X*, *Y*), as follows:

Let $K(s,t)$ denote the joint CGF for $(s,t) \in S \subseteq R^m$, where S is open and components s and t are associated with X and Y, respectively. Then, the saddlepoint approximation for $f(x,y)$ is given by Butler (2007) as follow:

$$\hat{f}(x,y) = \frac{1}{(2\pi)^{\frac{m}{2}} |K''(\hat{s},\hat{t})|^{0.5}} \exp(K(\hat{s},\hat{t}) - \hat{s}x - \hat{t}y), \quad x\epsilon\, \tau_{\chi}, \tag{2.16}$$

for $(x,y)\epsilon\, \tau_{\chi}$. Here, the *m*-dimensional saddlepoint $(\hat{s},\hat{t})$ solves the set of *m* equations

$$K'(\hat{s},\hat{t}) = (x,y), \tag{2.17}$$

where $K'(\hat{s},\hat{t})$ is the gradient with respect to both s and t. If $K''(\hat{s},\hat{t})$ is the corresponding Hessian, then the marginal saddlepoint density is given by Butler (2007) as below

$$\hat{f}(x) = \frac{1}{(2\pi)^{\frac{m_x}{2}} \left|K_{ss}''(\hat{s}_0,0)\right|^{0.5}} \exp(K(\hat{s},0) - \hat{s}x), \quad x\epsilon\, \tau_{\chi}, \tag{2.18}$$

where $\hat{s}_0$ is the m_x dimensional saddlepoint that solves

$$K'(\hat{s},0) = x\,. \tag{2.19}$$

Here, K' is the gradient with respect to the components s alone and K_{ss}'' is the corresponding Hessian. Also it can be shown that

$$\hat{f}(y|x) = \frac{1}{(2\pi)^{\frac{m_y}{2}} \left\{\frac{|K''(\hat{s},\hat{t})|}{\left|K_{ss}''(\hat{s}_0,0)\right|}\right\}^{0.5}} \exp[(K(\hat{s},\hat{t}) - \hat{s}x - \hat{t}y) - (K(\hat{s}_0,0) - \hat{s}_0 x)] \quad x\epsilon\, \tau_{\chi}, \tag{2.20}$$

This approximation is practically meaningful if y is an interior point of the conditional support for y, given $X = x$; see Barndorff-Nielsen and Cox (1979).

In the full generality of the independence assumption, $\hat{f}(y|x)$ can be written as

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)} = \frac{\hat{f}(x)\hat{f}(y)}{\hat{f}(x)} = \hat{f}(y) \tag{2.21}$$

Equation (2.21) represents the marginal saddlepoint density of y. The double-saddlepoint density of y given x is the same as the marginal or single-saddlepoint density of y. This finding could also be explained in terms of the reduction of Equation (2.20).

Under independence, the joint CGF separates into

$$K(s,t) = K(s,0) + K(0,t) \tag{2.22}$$

so that the saddlepoint equations are

$$K'_s(\hat{s},\hat{t}) = x \tag{2.23}$$

$$K'_t(\hat{s},\hat{t}) = y \tag{2.24}$$

The first of these equations is identical to the saddlepoint equation

$$K'(\hat{s},0) = x \tag{2.25}$$

by the uniqueness of saddlepoint solutions, $\hat{s}_0 = \hat{s}$. Using the same argument, the second of these equations reveals that $\hat{t}$ is the saddlepoint in the marginal saddlepoint density $\hat{f}(y)$. These facts lead to the single saddlepoint density $\hat{f}(y)$.

A double-saddlepoint CDF approximation for Y given $X = x$ when dim $(y) = 1$ was given by Skovgaard (1987). The following notation

$$F(y|x) = P(Y \leq y|X = x) \tag{2.26}$$

represents the true CDF.

The Skovgaard (1987) approximation when Y is a continuous variable for which $F(y|x)$ admits a density is

$$\hat{F}(y|x) = \Phi(\hat{w}) + \phi(\hat{w})(1/\hat{w} - 1/\hat{u}), \tag{2.27}$$

where

$$\hat{w} = sgn(\hat{t})\sqrt{2[\{K(\hat{s}_0, 0) - \hat{s}_0 x\} - \{K(\hat{s}, \hat{t}) - \hat{s}x - \hat{t}y\}]} \tag{2.28}$$

$$\hat{u} = \hat{t}\sqrt{\frac{|K''(\hat{s}, \hat{t})|}{|K_{ss}''(\hat{s}_0, 0)|}} \tag{2.29}$$

A discrete probability distribution concentrated on a set of points of the form $\gamma + nh$, where $h > 0$, γ is a real number and $n = 0, \pm 1, \pm 2, \ldots$ is called the lattice distribution. The number h is called the step of the lattice distribution. If the support for y is the integer lattice, then the continuity corrections to CDF, as introduced by Skovgaard (1987), should be used to achieve the greatest accuracy.

2.2.4.1 First continuity correction

Suppose $(\hat{s}, \hat{t})$ is the solution to $K'(\hat{s}, \hat{t}) = (j, k)$ required for the saddlepoint. Then,

$$\hat{P}_{r_1}(Y \geq k|X = j) = 1 - \Phi(\hat{w}) - \phi(\hat{w})(1/\hat{w} - 1/\tilde{u}_1), \tag{2.30}$$

where

$$\hat{w} = sgn(\hat{t})\sqrt{2[\{K(\hat{s}_0, 0) - \hat{s}_0 j\} - \{K(\hat{s}, \hat{t}) - \hat{s}j - \hat{t}k\}]} \tag{2.31}$$

$$\tilde{u}_1 = \left(1 - e^{-\hat{t}}\right)\sqrt{\frac{|K''(\hat{s}, \hat{t})|}{|K_{ss}''(\hat{s}_0, 0)|}} \quad , \tag{2.32}$$

where $K_s'(\hat{s}_0, 0)$ is the gradient with respect to s and $\hat{s}_0$ is obtaining by solving $K_s'(\hat{s}_0, 0) = j$; see Skovgaard (1987).