CAIRO UNIVERSITY

INSTITUTE OF STATISTICAL STUDIES AND RESEARCH, CAIRO UNIVERSITY

EGYPT

ON GOODNESS-OF-FIT TESTS BASED ON EMPIRICAL CHARACTERISTIC FUNCTION

By

Suzanne Abdel-Rahman Sayyid Allam

Under the Supervision of

Professor Samir Kamel Ashour
Professor of Mathematical Statistics-Institute of Statistical Studies and Research-Cairo University

Professor Ahmed Fouad Mohamed
Professor of Mathematical Statistics-Institute of Statistical Studies and Research-Cairo University

A Thesis Submitted to the Department of Mathematical Statistics in Partial Fulfillment of the Requirements for the Ph.D. in Statistics

2011

Approval Sheet

(On Goodness-of-Fit Tests Based on Empirical Characteristic Function)

By

Suzanne Abdel-Rahman Sayyid Allam

This Thesis for the PhD In Mathematical Statistics , Instiute Of Statistical Studies and Research ,Cairo University , and has been approved by

Name	Signature
S.K. Ashour	
M.E. Hadidi	M.E. Hadidi
Ahmed Fouad	Ahmed Fouad
E.A. Elsherpieny	E.A Elsherpieny

ACKNOWLEDGEMENTS

All gratitude is to Allah who guided and aided me through every single step taken in this research.

I wish to acknowledge my greatest debt to Professor Samir Ashour for his guidance, support, encouragement and valuable comments from the very beginning of this research. I'm extremely grateful to him for devoting so much of his time in supervising me during the preparation of this thesis and consistent availability throughout the research process and answers to all my questions.

Finally, I wish to express my deep thanks to my family for their support, patience, encouragement and understanding throughout the research process.

Table of Contents

Page

ABSTRACT …… vi

CHAPTER I: INTRODUCTION ……... 1

CHAPTER II: DEFINITIONS AND NOTATION ……... 3

(2.1) Methods of Estimation ……. 3

(2.1.1) Method of Moments …….. 3

(2.1.2) Method of Maximum Likelihood …….. 4

(2.2) Empirical Characteristic Functions ……... 5

(2.3) Goodness-of-Fit Tests ……..... 8

(2.3.1) Chi-Square Goodness of Fit Test ……... 10

(2.3.2) Tests Based on Empirical Distribution Function ……... 11

(2.3.3) Tests Based on Regression and Correlation Coefficients ... 16

(2.3.4) Power Concept of a Goodness-of-Fit Test …….. 18

(2.4) Some Important Distributions …… 19

(2.4.1) The Generalized Exponential Distribution …… 19

(2.4.2) The Cauchy Distribution …… 23

(2.4.3) Mixtures of Normal Distributions …….. 24

(2.5) Pearson System ……... 25

CHAPTER III: GOODNESS-OF-FIT TESTS BASED ON EMPIRICAL CHARACTERISTIC FUNCTION (A LITERATURE REVIEW) …… 27

(3.1) Tests based on unweighted integrals of the squared modulus of the difference between the ECF and the CF …… 27

Page

(3.2) Tests based on weighted integrals of the squared modulus of the difference between the ECF and the CF …….. 32

(3.2.1) Goodness-of-Fit Tests Based on ECF for the Cauchy Distribution ……... 32

(3.2.2) Goodness-of-Fit Tests Based on ECF for Mixtures of Normal Distributions ……..... 39

(3.3) Tests based on differences between real (imaginary) parts of the ECF and real (imaginary) parts of the CF ……... 42

CHAPTER IV: GOODNESS-OF-FIT TESTS FOR THE GENERALIZED EXPONENTIAL DISTRIBUTION …… 48

(4.1) Relevant Review ……. 48

(4.2) ECF Goodness of Fit Tests for the Generalized Exponential Distribution …….. 49

(4.3) EDF Goodness-of-Fit Tests for the Generalized Exponential Distribution …….. 53

(4.4) Power Study …… 56

(4.5) Effect of Estimation Method on the Power of the ECF Goodness-of-Fit Test …… 64

(4.5.1) Estimation of the Unknown Parameters …… 65

(4.5.2) Obtaining Critical Values …….. 66

(4.5.3) Power Comparison under Each Estimation Method for the ECF Test ……. 67

(4.6) Sampling Distribution of the ECF Test Statistic……. 75

(4.7) Conclusions …… 77

Page

APPENDICIES ……... 79

Appendix A – Tables ……... 80

Appendix B – Mathcad Programs ……... 102

REFERENCES …… 151

ARABIC SUMMARY ...…… 155

ABSTRACT

Characteristic functions (CF) were originally developed as a tool for the solution of problems in probability theory and admit many important applications in this branch of Mathematics as well as in Mathematical Statistics. The empirical characteristic function (ECF) is the sample counterpart of the CF. It was defined by Parzen (1962) and can be used in statistical inference. It can be used for parameter estimation and hypothesis testing.

In the literature, there are many studies which introduced goodness-of-fit tests based on the ECF using different methodologies. The basic idea of the ECF method is to compare the CF derived from the hypothesized model with the ECF obtained from the sample data.

In this thesis, goodness-of-fit tests based on the ECF are used for testing the fit of the generalized exponential distribution. In addition, a power comparison is conducted with other common goodness-of-fit tests that are the tests based on empirical distribution function (EDF). Also, the effect of the estimation method used for estimating the unknown parameters of the generalized exponential distribution on the power of the ECF test is studied. Finally, the sampling distribution for the ECF test statistic is obtained using Pearson system.

Chapter I
INTRODUCTION

A statistical problem encountered in many areas of research is the need to assess whether a sample of observations comes from a specified distribution. Typically such situations are known as 'Goodness of Fit' problems, that is, how well are the data modeled by a certain distribution? A goodness-of-fit test uses the properties of a hypothesized distribution to assess whether a sample of observations is generated from that distribution.

Characteristic functions (CF) were originally developed as a tool for the solution of problems in probability theory and admit many important applications in this branch of Mathematics as well as in Mathematical Statistics. The empirical characteristic function (ECF) is the sample counterpart of the CF. It was defined by Parzen (1962) and can be used in statistical inference. It can be used for parameter estimation and hypothesis testing.

In the literature, there are many studies which introduced goodness-of-fit tests based on the ECF using different methodologies. The basic idea of the ECF method is to compare the CF derived from the hypothesized model with the ECF obtained from the sample data.

The generalized exponential distribution was introduced by Gupta and Kundu (1999). This distribution can be used quite effectively in analyzing many lifetime data, particularly in place of the two-parameter gamma and Weibull distributions. It is observed that the generalized

exponential distribution can be considered for situations where a skewed distribution for a non-negative random variable is needed.

In this thesis, goodness-of-fit tests based on the ECF are used for testing the fit of the generalized exponential distribution. In addition, a power comparison is conducted with other common goodness-of-fit tests that are the tests based on empirical distribution function (EDF). Also, the effect of the estimation method used for estimating the unknown parameters of the generalized exponential distribution on the power of the ECF test is studied. Finally, the sampling distribution for the ECF test statistic is obtained using Pearson system. All the Algorithms in this thesis are implemented using the Mathcad (version 13) software.

This thesis is organized as follows. Some important definitions and notation are introduced in Chapter II. Chapter III includes a literature review about different goodness-of-fit tests that are based on ECF. Chapter IV, which is the main analytical chapter, introduces a power comparison between the ECF and EDF tests, for testing the fit of the generalized exponential distribution. In addition, the same chapter includes a study about the effect of parameter estimation method on the power of the ECF test for testing the fit of the generalized exponential distribution. Finally, Chapter IV contains a section about obtaining the sampling distribution of the ECF test statistic using Pearson system.

The results of all simulation experiment conducted in this thesis are organized in tables in appendix A. Also, appendix B contains all the Mathcad programs implemented in this thesis.

Chapter II
DEFINITIONS AND NOTATION

This chapter is devoted to some important definitions and notation that will be used in the present dissertation.

(2.1) Methods of Estimation

In the literature, there are different methods for estimating the unknown parameters of a statistical distribution. The most commonly used among these methods are the method of moments and method of maximum likelihood.

(2.1.1) Method of Moments

The method of moments (MM) is a technique for constructing estimators for the parameters which is based on matching the sample moments with the corresponding population moments. It is frequently called the method of moments because it is understood that, whenever possible, the parameter should be estimated by using moments, particularly, the lowest order moments that are convenient. The MM method provides estimators that are consistent but not as efficient as the maximum likelihood ones. It is often used because it leads to very simple computations, unlike maximum likelihood method which can become very cumbersome.

The MM method consists of equating the first few moments of a population to the corresponding moments of a sample, thus getting a number of equations that are needed to be solved in terms of the unknown parameters of the population. Therefore, if a population has k unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$, then the parameter estimates $\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_k$ can be obtained by solving the following system of simultaneous equations:

$$m_r' = \mu' \ \bar{\theta} \ \bar{\theta} \quad \bar{\theta}), \qquad r = 1,2,...,k,$$

where $m_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r$ is the r^{th} sample moment of a set of observations $x_1, x_2, ..., x_n$ and $\mu' \ \bar{\theta} \ \bar{\theta} \quad \bar{\theta}) = E(X^r)$ is the r^{th} population moment (see Miller and Miller, 1999).

(2.1.2) Method of Maximum Likelihood

The maximum likelihood (ML) method is one of the most important methods in the theory of estimation. The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data. ML estimation begins with writing a mathematical expression known as the likelihood function of the sample data. The likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the maximum likelihood estimates.

If $x_1, x_2, ..., x_n$ are the values of a random sample from a distribution having probability density function $f(x;\theta)$, the likelihood function $L(\theta)$ of the sample is given by:

$$L(\theta) = \quad \theta), \quad \theta \in \Omega,$$

where Ω is a given domain for the values of θ. Here $f(x_1, x_2, ..., x_n; \theta)$ is the value of the joint probability density of the random variables $X_1, X_2, ..., X_n$ at $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$. Thus, the method of ML consists of maximizing the likelihood function with respect to θ. The value of θ that maximizes the likelihood function is referred to as the maximum likelihood estimate of θ. In

other words, the maximum likelihood estimate of θ is the solution of the equation $\left.\frac{dL(\theta)}{d\theta}\right|_{\theta=\hat{\theta}} = 0$.

If the likelihood function contains *k* parameters, i.e. if:

$$L(\theta_1, \theta_2, ..., \theta_k) = f(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_k), \quad (\theta_1, \theta_2, ..., \theta_k) \in \Omega,$$

then the maximum likelihood estimates of the parameters $\theta_1, \theta_2, ..., \theta_k$ are the values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ in Ω which maximize $L(\theta_1, \theta_2, ..., \theta_k)$. These values are obtained by simultaneously solving the following *k* equations:

$$\left.\frac{\partial L(\theta_1, \theta_2, ..., \theta_k)}{\partial \theta_i}\right|_{\theta_i=\hat{\theta}_i} = 0, \qquad i = 1, 2, ..., k .$$

Since the maximum value of $L(\theta)$ will occur at the same points as the maximum value of $\ln[L(\theta)]$, it will be easier to work with the logarithm of the likelihood function (see Miller and Miller, 1999).

(2.2) Empirical Characteristic Functions

Characteristic functions were originally developed as a tool for the solution of problems in probability theory and admit many important applications in this branch of Mathematics as well as in Mathematical Statistics.

Let X be a random variable and let $F(x)$ be the distribution function of X given by $F(x) = \Pr[X \le x]$, $x \in R$. Then, the characteristic function (CF), $\Phi(t)$, of the random variable X [or of the distribution function $F(x)$] is a complex valued function given by:

$$\Phi(t) = E(e^{itx}) = \int_{-\infty}^{\infty} e^{itx} f(x)dx \,, \qquad (2.1)$$

where $t \in R$ and $i = \sqrt{-1}$. For discrete random variables the CF reduces to:

$$\Phi(t) = \sum_{r} e^{itr} P(X = r) \,.$$

Note that according to Euler's formula, for any real number z:

$$e^{iz} = \cos z + i \sin z \,,$$

therefore $\Phi(t)$ in (2.1) could be expressed as follows:

$$\Phi(t) = E(\cos tX) + iE(\sin tX) = \int_{-\infty}^{\infty} \qquad \int_{-\infty}^{\infty} \sin tx f(x)dx \,.$$

The main advantage of the CF over other transforms such as the probability generating function or the moment generating function is that the integral exists for any probability distribution. Characteristic functions have the following properties:

i. $\Phi(0) = 1$

ii. $|\Phi(t)| \leq 1$ for all t

iii. The characteristic function of $a + bX$ is $e^{iat}\Phi(bt)$.

iv. A characteristic function Φ is real valued if and only if the distribution of the corresponding random variable X is symmetric about zero, that is if and only if $P[X > z] = P[X < -z]$ for all $z \geq 0$.

v. The characteristic function of the sum of independent random variables is the product of the characteristic functions of each of the random variables.

vi. Two distribution functions $F_1(x)$ and $F_2(x)$ are identical if, and only if, their characteristic functions $\Phi_1(t)$ and $\Phi_2(t)$ are identical. In other words, a distribution function F is determined uniquely by its CF Φ.

For additional information on the above properties of characteristic functions see Lukacs (1970).

The Empirical characteristic function (ECF) is the sample counterpart of the CF. It was defined by Parzen (1962) as:

$$\Phi_n(t) = \frac{1}{n}\sum_{j=1}^{n} \qquad \frac{1}{n}\sum_{j=1}^{n} \qquad \left[-\sum \qquad \right], \tag{2.2}$$

where $X_1, X_2, ..., X_n$ is a random sample of independently and identically distributed random variables.

The ECF can be used in statistical inference. The method of model-fitting via the ECF was discussed by many researchers. The advantage of using this procedure is that one can avoid difficulties arising in calculating or maximizing the likelihood function. Thus, it is a desirable estimation method when the maximum likelihood approach encounters difficulties but the CF has a tractable expression. The basic idea of the ECF method is to compare the CF derived from the model with the ECF obtained from data.

The justification for the ECF method is that there is a one-to-one correspondence between the CF and its distribution function. In other words, a distribution function is determined uniquely by its CF. As a consequence, the ECF retains all information in the sample. This observation suggests that estimation and inference via the ECF should work as efficiently as the likelihood-based approaches. The general idea for ECF estimation is to minimize various distance measures between the ECF and CF (Yu, 2004).

Also, the idea of hypothesis testing using the ECF is based on measuring the distance between the ECF and the CF of a random variable under the null hypothesis.

Many studies presented different distance measures between the ECF calculated from a sample and the CF of a population. For example, Besbeas and Morgan (2004) used the integral of the squared modulus of the difference between the ECF and CF with a weight function for estimating the parameters in a mixture of normal densities. The parameter estimators are obtained by minimizing the following criterion:

$$I(\theta) = \int_{-\infty}^{\infty} \left| \Phi_n(\cdot) - \Phi(\cdot, \theta) \right|^2 dW(t)$$

with respect to θ, where θ is the vector of unknown parameters. $W(t)$ is some weight function selected to ensure convergence of the integral $I(\theta)$. In addition, Feuerverger and Mureika (1977) constructed a symmetry test based on the weighted integral of the squared difference between the imaginary part of $\Phi_n(t)$ and zero. Murota and Takeuchi (1981) used the studentized ECF to test the shape of the distribution. The studentized ECF is given by:

$$\Phi'_n(t) = \Phi_n\left(t / s\right),$$

where $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2$ and $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$.

(2.3) Goodness-of-Fit Tests

A statistical problem encountered in many areas of research is the need to assess whether a sample of observations comes from a specified distribution. Typically such situations are known as 'Goodness of Fit' problems, that is, how well are the data modeled by a certain distribution? If the data are univariate or multivariate, continuous or discrete, ordinal or nominal, researchers are