INFERENCE CONCERNING THE MEDIAN OF A GAMMA DISTRIBUTION

by

G.E.M.U.P.D. Ekanayake

Thesis submitted in partial fulfillment of requirement for the Degree of Master of Science in Applied Statistics

Department of statistics and computer science
Faculty of Graduate Studies
University of Sri Jayawardenepura
Nugegoda
Sri Lanka

2005

DECLARATION

"The work described in this thesis was carried out by me at the University of Sri Jayewardenepura, under the supervision of Dr. B.M.S.G. Banneheka, Senior Lecturer of Statistics in the Department of Statistics and Computer Science, University of Sri Jayewardenepura, and a report on this has not been submitted in whole or in part to any University or any other Institution for another Degree/Diploma".

Date : 29.04.2005

P · D ·

G.E.M.U.P.D. Ekanayake

GS/PS/942/2000

"I certify that the above statement made by the candidate is true and that this thesis is suitable for submission to the university for the purpose of evaluation".

Date ...29/04/2005...........

....

Dr. B.M.S.G. Banneheka

Department of Statistics
and Computer Science,
Faculty of Applied sciences,
University of Sri Jayawardenepura,
Nugegoda,
Sri Lanka.

ABSTRACT

The population median is considered as better than the population mean as the representative of the 'average' or central tendency of skewed distributions. The gamma distribution is often used as a model for positively skewed distributions. The inference regarding the mean of a gamma distribution is easy. However, the inference regarding the median of a gamma distribution is not that easy and it has not been studied fully.

In this research, we compare several estimators for the median of a gamma distribution. We found that the maximum likelihood estimator is superior to the other estimators considered. However, it needs intensive computations. We found an estimator, which is only slightly inferior to the maximum likelihood estimator, but far more easy to calculate even with a pocket calculator. This was found based on an approximation that we derived for the median. We also propose a method for constructing confidence interval using the same approximation. This method also works well, especially with relatively large samples.

ACKNOWLEDGEMENT

I would like to express my gratitude and appreciation to the Department of Statistics and Computer Science, University of Sri Jayewardenepura, for conducting the Master degree program in Applied statistics.

I also extend my sincere gratification to the staff, who conducted the lectures during the M.Sc.course and I appreciably believe that this course has made a distinctive advance in my academic pursuits.

I am deeply indebted to my supervisor Dr. B.M.S.G. Banneheka, Senior Lecturer of Statistics in the Department of Statistics and Computer Science, University of Sri Jayewardenepura, who directed me to undertake this study giving invaluable guidance and suggestions to make this study as success.

I am also much thankful to Mr. A. G. W. Nanayakkara, Director General, Department of census and statistics for granting me duty leave to follow this Msc. Course.

My sincere thanks goes to Mr. D.C.A. Gunawardena, Director, Mr. W.N.T. Perera and Mrs. J.A.S.S. Jayawardena, Senior Statisticians and Statisticians of prices and wages Division in the Department of census and statistics for giving me required facilities in this regard.

I am also much grateful to my colleagues for their kind cooperation to me during the period of study.

Finally, I express my great devotion to my parents, sister and brother for giving me inspiration in attaining my academic carrier.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The distribution of any quantitative random variable can be reasonably described by four characteristics: the average, dispersion, kurtosis and skewness. The average provides a good description of the central tendency or location of the distribution. The dispersion characterizes the spread of the distribution around its average. The skewness provides a measure of the location of the mode (or high point in the distribution) relative to the average. The kurtosis provides a measure of the "peaked-ness" of a distribution. Out of these four characteristics, people are often more interested in the 'average' and the dispersion.

The word 'average' is used to describe different parameters by different people. For example, a farmer may be interested in the 'average yield per acre' that can be obtained by cultivating a particular species of chillies. A provincial council may be interested in finding the 'average household income' in its area. A shoe store manager may be interested in finding the 'average size' of slippers sold of a particular type. As in these three examples, 'the average' may mean different characteristics of the variable of interest. While the farmer is interested in the population 'mean' of yield, the provincial council may be more interested in the population 'median' of income. The shoe store manager may be looking for the population 'mode' of sizes of slippers sold.

The intent of each parameter is to identify a score that might appropriately represent the typical score of that distribution. In general, these parameters identify a point near the centre of the distribution. Therefore, these have been called “measures of central tendency”. These measures of central tendency are the mean, median and mode. Although each measure of central tendency attempts to identify the most typical score in that distribution of scores, each measure has its own interpretation of the most typical score.

The mean defines central tendency as the mathematical average of all the scores. It has the interpretation that it is the amount that one unit would get if the whole population is divided equally among all units. It also has the interpretation as the ‘balancing point’ of the distribution. The median defines central tendency as the point where half the scores fall above that value and half the scores fall below it. Finally, the mode defines central tendency as the most frequently occurring score in that distribution of scores. The two most widely used measures of central tendency are the mean and the median. Although the mode is also a measure of central tendency, its use is usually limited to describing qualitative data. When one is to select a measure of central tendency, the choice is usually between the mean and the median. Which measure should be chosen?

Such questions often arise in statistics, since there is usually more than one statistical method available for dealing with a problem. However, this does not imply that all methods are equally acceptable for a given situation. The correct choice will depend, in part, on the type of data being analyzed (qualitative or quantitative), the shape of the distribution of scores, and the question being asked.

If the data being analyzed is qualitative, then the only measure of central tendency that can be reported is the mode. However, if the data is quantitative in nature (ordinal or interval/ratio) then the mode, median or mean can be used to describe the 'average'.

With quantitative data, the shape of the distribution of scores (symmetrical, negatively or positively skewed) plays an important role in determining the appropriateness of the specific measure of central tendency to accurately describe the 'average'. If the distribution of scores is symmetrical or nearly so, the median and mean (as well as the mode) will be very close to each other in value. Mathematically, it is relatively easier to make inferences about the mean, than about the median. Especially, when large samples are available, the central limit theorem can be used to make inference about the mean. Therefore, in this case, the mean is the value of central tendency that is usually considered. Consequently the variance or standard deviation is considered as the measure of dispersion.

However, if the distribution of scores is positively or negatively skewed, the mean will tend to either overestimate (in positively skewed distributions) or underestimate (in negatively skewed distributions) the true central tendency of the distribution. In extreme cases of skewed data, the mean can lie at a considerable distance from most of the scores. Therefore, in skewed distributions, the median will tend to be the more accurate measure to represent the 'average' of the distribution than the mean because the median can never have more than one half the scores above or below it. If the data were skewed, then the measure of variability that would be appropriate for that data would be the quartile deviation.

Gamma distribution is often used to model right skewed distributions. Inference concerning the mean of a gamma distribution is fairly easy because the sample mean $\bar{x}$ can be used as an estimator and the properties of this estimator are well known. The inference concerning the mean is easier for large samples because we can use the central limit theorem to construct confidence intervals and tests hypothesis. However, the inference concerning the median is not as easy. In this research, we focus our attention to the median of a gamma distribution. We explore some point estimators and interval estimators calculated by inverting the likelihood ratio test.

Chapter 2 presents some general facts about gamma distributions and two approximations that we derived for the median ν of a gamma distribution. Chapter 3 examines seven possible estimators for the median of a gamma distribution. Chapter 4 introduces a way to construct confidence interval for the median.

Chapter 2

GAMMA DISTRIBUTION

2.1 Introduction

The density function of the gamma distribution given by

$$f_x(x;\alpha,\beta)=\frac{e^{-x/\beta}x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} \; ; x>0,\ \alpha>0,\ \beta>0 \tag{2.1}$$

We denote this by writing $X \sim Gamma(\alpha,\beta)$. Here, $\alpha(>0)$ is called the shape parameter and $\beta(>0)$ is called the scale parameter of the gamma distribution. For this distribution,

$$\text{Mean} = E(X) = \alpha\beta = \mu \text{ (say)}$$

$$\text{Standard Deviation} = SD(X) = \sqrt{\alpha}\,\beta = \sigma \text{ (say)}$$

$$\text{Skewness} = 2\sqrt{\frac{1}{\alpha}} = \gamma_1 \text{ (say)}$$

$$\text{Kurtosis} = \frac{6}{\alpha} = \gamma_2 \text{ (say)}$$

Reparametricing as $\beta = \mu/\alpha$, the density can be written as

$$f_X^*(x;\alpha,\mu)=\frac{e^{-\alpha x/\mu}x^{\alpha-1}}{\Gamma(\alpha)(\mu/\alpha)^{\alpha}} \; ; x>0,\ \alpha>0,\ \mu>0. \tag{2.2}$$

The median ν is given as the solution of

$$\int_0^{\nu} f_x(x;\alpha,\beta)dx = 0.5$$

The median ν of a $Gamma(\alpha,\beta)$ = β (median of a $Gamma(\alpha,1)$).

Figure 2.1 – 2.5 show the density functions of several gamma distributions. The mean (μ) and median (ν) are marked by dark line and dotted line respectively.