STUDY OF SOME FUNCTIONS OF RANKS

BY

MOHAMED GAMAL HABIB

Supervised by

Prof. Mohamed Abdul-Hamid Taha
Prof. & Head; Dept. of Mathematics
Faculty of Science
Tanta University

Prof. Abdel-Azim Anis
Prof. of Mathematical Statistics
Faculty of Science
Ain Shams University

for

The Master Degree of Science
Major Subject Mathematical Statistics

1975

Ain Shams University
Faculty of Science

STUDY OF SOME FUNCTIONS OF RANKS

BY

MOHAMED GAMAL HABIB
B. Sc. 1969

A Thesis Submitted in Partial Fulfillment of the Requirements for the Award of the M.Sc. Degree in Mathematical Statistics

بسم الله الرحمن الرحيم

TO MY PARENTS

ACKNOLEOGEMENT

I am deeply grateful to Professor M.A.H. Taha, Professor and Head of Department of Mathematics, Faculty of Science, Tanta University and to Professor A. Anis, Professor of Mathematical Statistic, Faculty of Science, Ain Shams University, for their kind supervision, help and encouragement during the preparation of this thesis.

I am deeply grateful also to Professor A.H. El Mawaziny, Head of the Department of Mathematical Statistics, Institute of Statistical STudies and Research for his continuous encouragement and help.

My sincere thanks are also due to Professor A.E. Sarhan, Dean of the Institute of Statistical Studies and Research, CAiro University, for his encouragement and the facilities given to me during the preparation of this thesis.

Thanks are also due to the staff members and colleagues of the Institute of Statistical Studies and Research and The Computer Center for their help and cooperation.

APPROVAL SHEET

This thesis for the Degree of Master of Science in Mathematical Statistics has been approved by:

Dr. ....

Dr. ....

Dr. ....

CONTENTS

INTRODUCTION

The work presented in this thesis is devoted to the exploration of certain aspects of testing nonparametric statistical hypotheses. The case dealt with is of the following type: Given two samples $X_m = (x_1, \ldots, x_m)$ and $Y_n = (y_1, \ldots, y_n)$ drawn from continuous populations, it is required to test the hypothesis that $F(x) = G(x)$ where $F(x)$ and $G(x)$ are the comulative distribution functions (CDF's) of X and Y respectively.

This problem is approached by applying different tests which belong to the family of nonparametric tests. In particular the discussions are centered around the rank tests, specially those that depend on the powers of ranks.

The work included in this thesis is divided into Four chapters:

Chapter (I) is devoted to the basic concepts and results for rank functions and the general linear rank tests.

Chapter (II) covers a group of rank tests. The first section is devoted to tests for location while the second section discusses some related tests for scale alternative.

In section three the author introduces certain modifications for some rank tests with a numerical study on the efficiency of the modified rank test with respect to the locally most powerful rank test (LMPRT). This study shows the increase of the efficiency of the modified test by increasing the rank power to a certain degree beyond which the efficiency decreases.

Chapter (III) contains a discussion of the difficulties involued in obtaining LMPRT. These difficulties may be avoided by using the polynomial of ranks, which proved to be almost as good as LMPRT. This chapter is divided into seven sections in section one and two the general polynomial form and the asymptotic relative efficiency (ARE) with respect to the LMPRT are given. Section three covers the ARE of the test for symmetric distributions, two new relationships are given in this section. A study of the relationships between certain known tests and the polynomial form are shown in section four. Section five presents evidence of the robustness of the test against changes in the polynomial coefficients, except in a narrow neighbourhood to the point of zero efficiency. The relationships between the efficiency and the skewness of the underlying distribution is also studied. Section six is devoted to a numerical study of the robustness of the test. The last section is an appendex containing the FORTRAN programmes used in the numerical studies for the robustness.

In chapter (I V) a new approach to the polynomial approximation is given using the orthogonal polynomial in formulating the test. The suggested approach yields the same maximum efficacy as that given by Taha. The essential difference between the suggested method and that of Taha is that while the first requires the inverse of a diagonal matrix, the last requires the inverse of an ill - conditioned Hilbert matrix. This shows that the efficacy of the test is always non-decreasing as the polynomial degree increases. According to the studies given in the chapter, it is conjectured that our formula requires a smaller sample size [illegible] than the formula given by Taha.

CHAPTER (I)

LINEAR RANK TESTS

This chapter is devoted to the study of the characteristics of the linear rank statistics (LRS). The first section is concerned with the basic definitions and notations, the mean and the variance of LRS are also derived. The second section deals with the behaviour of LRS particularly its asymptotic normality. The conditions under which LRS will be locally most powerful rank test (LMPRT) is discussed in the third section. The fourth section studies the efficiency, in general, of any two tests which applies to the problem when using LRS.

1.1. <u>Linear Rank Statistics (LRS).</u>

Consider two sets of independent observations $x_1, \dots, x_m$ and $y_1, \dots, y_n$ of the continuous random variables X and Y which follows the commulative distribution functions (CDF's) F(x) and G(x) respectively. Pooling the two samples and arranging the deduced one in an ascending order; denote the new sample by $Z = (z_1, \dots z_N)$ such that $z_1 < z_2 \dots < z_N$; $N = m+n$. From which a new sequence,

z_{N_1} , z_{N_2} , ... , z_{NN} , is formed by putting

$$z_{N_i} = \begin{cases} 1 & \text{if } z_i \text{ is an X - observation} \\ 0 & \text{if } z_i \text{ is an Y - observation} \end{cases}$$

<u>The Rank Function:</u>

The rank function of the X observations is defined as follows:

$$R_i = \begin{cases} i & \text{if } z_{N_i} = 1 \\ 0 & \text{if } z_{N_i} = 0 \end{cases}$$

where $i = 1,2, \ldots , N.$

<u>Linear Statistics:</u>

A linear statistic is defined as one of the form

$$S = \sum_{i=1}^{N} c_i \, a(R_i) \qquad (1.1)$$

where the c_i's are constants called the regression constants and the a( )'s are functions of R_i called the scores.

The problem of interest in this thesis is to test the null hypothesis

$$H_0 : F(x) = G(x) \qquad (1.2)$$

against the alternatives

$$H_1 : F(x) \neq G(x) \qquad (1.3)$$

within the alternative hypotheses, we are interested in the following two special cases:

$$H_{11} : G(x) = F(x - \theta) \qquad , \theta \neq o$$

$$H_{12} : G(x) = F(x / \theta) \qquad , \theta \neq 1$$

H_{11} with non zero θ is the location alternative two sided case and with $\theta > o$ (or $\theta < o$) is the one side location alternative. Similarly, H_{12} with $\theta \neq 1$ is the scale alternative of the two-sided, whereas $\theta > 1$ (or $\theta < 1$) is the one side scale alternative.

For the location alternative, if F(x) is the normal distribution we have the t-test which is the locally most powerful test (LMPT) for this case. Generally, if the functional form of F(x) is known, we can deduce the suitable parametric test. But in so many cases we have no [illegible] form of F(x) and this makes [illegible] which is independent of the form of F(x). Another important case is when X cannot be measured numerically but can be arranged or evaluated in the form of scores (for example; pain, taste, ...). Fortunately, there exists several procedure which are called nonparametric or distribution - free methods. The rank tests are of great importance among the class of nonparametric tests. Some characteristics of linear rank tests are given in the following:

1.2. Mean and Variance of Linear Rank Tests:

For the linear rank test (LRT) given by equation (1.1) the mean and the variance can be obtained as follows:

$$E(S) = \sum_{i=1}^{N} c_i \, E\left\{a(R_i)\right\},$$

since $\quad E\left\{a(R_i)\right\} = \sum_{j=1}^{N} a(j)\, P(R_i = j), \quad i=1,2,\ldots,N$

as $R = (R_1, \ldots, R_N)$ is uniformly distributed i.e.

[illegible] $\frac{1}{N}$ [illegible] $= 1, \ldots, N.$

[illegible]

[illegible]

$$\text{Var}(S) = \sum_{i=1}^{N} c_i^2 \, \text{Var}\left\{a(R_i)\right\}$$

$$+ \sum_{i \neq j}^{N} c_i c_j \, \text{cov}\left\{a(R_i),\, a(R_j)\right\}$$