TESTS BASED ON EMPIRICAL DISTRIBUTION FUNCTION

Submitted in partial fulfillment of the requirements
for the award of the degree of

MASTER OF SCIENCE
IN
MATHEMATICS AND COMPUTING

Submitted by

Gurpreet Kaur

Roll no:-301403004

Under the guidance of

Dr. Anil Gaur

Assistant Professor

School of Mathematics

Thapar University

Patiala-147004(PUNJAB)

INDIA

July, 2016

Certificate

I hereby declare that the work which is being presented here in the dissertation entitled **"TESTS BASED ON EMPIRICAL DISTRIBUTION FUNCTIONS"** in partial fulfillment of the requirement for the award of degree of **Master of Science** in **Mathematics and Computing** submitted in School of Mathematics, Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Anil Gaur**, Assistant Professor, SOM and refer other researcher's work which are duly listed in the reference section.

The matter presented in this thesis has not been submitted in any other University/Institute for the award of my degree.

Dated: 02/07/2016

Gurpreet Kaur

Roll no.- 301403004

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dated: 02/07/2016

Dr. Anil Gaur

Assistant Professor

SOM, Thapar University

Countersigned By:-

Dr. A.K. Lal

Head, SOM

Thapar University

Patiala-147004

Dr. S.S.Bhatia

Dean of Academic Affairs

Thapar University

Patiala-147004

Acknowledgements

First of all, I would like to thank the almighty for granting perseverance. I would like to express my gratitute to my honorable supervisor **Dr. Anil Gaur, Assistant Professor, SOM, Thapar University, Patiala,** for their patient guidence and support throughout this work. I was truly very fortunate to have the opportunity to work under him as a student. It was both an honor and privilege to work with him. He also provides help in technical writing and presentation style and I found this guidence to be extremely valuable.

I take this opportunity to express my sincere thanks to **Dr. A.K. Lal, Head, SOM,Thapar University, Patiala,** for their valuable support and help without which it would not have been possible for me to complete this work.

I would like to thank my beloved parents for their years of unyielding love and encouragement. They have always wanted the best for me and I admire my parent's determination and sacrifice to put me through college.

Finally, I am also thankfull to all my friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work.

Gurpreet Kaur.

Patiala

July, 2016

Gurpreet Kaur

Abstract

The chapter-wise summary of the thesis is as follows:

Chapter 1 includes introduction about the Tests based on Empirical Distribution Functions. The main focus of this chapter is on Non-parametric tests. This chapters includes basic concepts, definitions, and brief discription about the Goodness-of-fit problem. Goodness-of-fit tests are used to check the compatibility of a set of observed sample values with a normal distribution or any other distribution. These tests are designed for a null hypothesis which is the statement about the form of probability function or cumulative distribution function of the parent population from which the sample is drawn. Here χ^2 Goodness-of-fit test and its applications are described in details.

In **Chapter 2** the second goodness-of-fit test the Kolmogorov-Smirnov test is discussed in details. The Kolmogorov-Smirnov statistics are used as general goodness-of-fit tests which are known to be more sensitive to location than to scale alternatives. This test is based on vertical deviation between observed and expected cumulative distribution functions. In this chapter the Kolmogorov-Smirnov one-sample statistic, the Kolmogorov-Smirnov two-sample statistic and their applications are discussed.

Then, in **Chapter 3** the two-sample, distribution-free statistics of Smirnov (1939) are used to define a new statistic. While the Smirnov statistics are used as a general goodness-of-fit test, a distribution-free scale test based on this new statistic is developed. It is shown that this new test has higher power than the two-sided Smirnov statistic in detecting differences in scale for some symmetric distributions with equal means/medians.

Table of Contents

List of Tables

Chapter 1

Introduction

Let X be a random variable, continuous or discrete, with probability mass function $p(x)$ or probability density function $f(x)$ respectively. If for every real number x there exists a probability that the value assumed by the random variable X does not exceed x, called the cumulative distribution function (*cdf*) of random variable X denoted by $F_X(x)$ and is defined as

$$F_X(x) = P[X \leq x] = \begin{cases} \sum\limits_{x_i \leq x} p(x_i) & \text{if X is discrete} \\ \\ \int\limits_{-\infty}^{x} f(t)dt & \text{if X is continuous} \end{cases}$$

1.1 The Empirical Distribution Function

For a random sample from the distribution F_X, the empirical distribution function (*edf*), denoted by $S_n(x)$, is simply the proportion of sample values less than or equal to the specified value x, that is,

$$S_n(x) = \frac{number\ of\ sample\ values\ \leq\ x}{n}$$

Suppose that the n sample observations are distinct and arranged in increasing order so that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest,..., and $X_{(n)}$ is the largest. The *edf* $S_n(x)$ is defined as

$$S_n(x) = \begin{cases} 0 & if\ x < X_{(1)} \\ \frac{i}{n} & if\ X_{(i)} \leq x < X_{(i+1)}, \quad i = 1, 2, ..., n \\ 1 & if\ x \geq X_{(n)}. \end{cases}$$

Suppose that the sample observations with sample size n=5 are given by 9.4, 11.2, 11.4, 12 and 12.6. Here $S_n(x)$ is a step function, with jumps occurring at the distinct ordered sample values, where the height of each jump is equal to the reciprocal of the sample size, i.e. 1/5 or 0.2.

1.2 Goodness-of-fit Problem

The main problem in statistics is to obtain information about the form of the population from which sample is drawn. The focus of investigation might be to find the shape of this distribution. Alternatively, some inference concerning a particular form of the population may be of the primary interest. In classical statistics, information about the form of the population generally must be incorporated or postulated in the null hypothesis to perform an exact parametric type of inference. For example, consider we have a small number of observations from an unknown population with unknown variance and the hypothesis of interest concerns the value of population mean. The traditional parametric test, based on t-distribution, is derived under the assumption of normal population. The exact distribution theory and the probabilities of both type of errors depends on form of the underlying population. Therefore it might be desirable to check on the reasonableness of the normality assumption before forming any conclusions based on t-distribution. If normality assumption appears not to be justified, some type of nonparametric inference for location might be more appropriate with a small sample size.

A goodness-of-fit type of test can be used to check the compatibility of a set of observed sample values with a normal distribution or any other distribution. These tests are designed for a null hypothesis which is the statement about the form of probability

function or cumulative distribution function (*cdf*) of the parent population from which the sample is drawn. Ideally, the hypothesized distribution is completely specified, including all the parameters. Since the alternative is necessarily quite broad, including differences only in scale, location, other parameters, form or any combination thereof, the rejection of null hypothesis does not provide much specific information. Goodness-of-fit tests are mainly used only when the form of the population is in question, with the hope that the null hypothesis will be found acceptable.

Classically, there are two types of goodness-of-fit tests. The first type is designed for null hypothesis concerning a discrete distribution and compares the observed frequencies with the expected frequencies under the null hypothesis. This test is known as the χ^2-test proposed by Karl Pearson early in the history of statistics. The second type of goodness-of-fit test the Kolmogorov-Smirnov test is designed for null hypothesis concerning a continuous distribution and compares the observed cumulative relative frequencies with those expected under the null hypothesis.

1.3 The χ^2 Goodness-of-Fit Test

Let X be a random sample of size n which is drawn from a population with unknown cumulative distribution function F_X and the hypothesis-testing situation

$$H_0 : F_X(x) = F_0(x) \quad \text{for all x}$$

where $F_0(x)$ is completely specified against the general alternative

$$H_A : F_X(x) \neq F_0(x) \quad \text{for some x.}$$

To apply the χ^2-test in this situation, firstly, the sample data must be grouped according to some scheme in order to form a frequency distribution. In the case of qualitative data, where the hypothesized distribution would be discrete, the categories would be the relevant numerical or verbal classifications. For example, in tossing a coin, the categories would be the numbers of heads; in tossing a die, the categories would be the numbers of spots; in surveys of brand preferences, the categories would be the brand names considered. When the sample observations are quantitative, the categories

would be the numerical classes chosen by the experimenter. In this case, frequency distribution is not unique and some information is necessarily lost. Even though the hypothesized distribution is mostly continuous with the measurement data, the data must be categorized for analysis by the χ^2-test.

When the population distribution is completely specified by the null hypothesis, one can calculate the probability that a random observation will be classified into each of the chosen or fixed categories. These probabilities multiplied by n give the frequencies for each category which would be expected if the null hypothesis were true. Except for the sampling variation, there should be close agreement between these expected and observed frequencies if the sample data are compatible with the specified $F_0(x)$. The corresponding observed and expected frequencies can be compared visually using a histogram, a frequency polygon, or a bar chart. The χ^2 goodness-of-fit test provides a probability basis for the comparison and deciding whether the lack of agreement is too great to have occurred by chance.

Assume that the n observations have been grouped into k mutually exclusive categories and denote the observed and expected frequencies for the i^{th} class by f_i and e_i re respectively, $i = 1, 2, ..., k$. The decision regarding fit is to be based on the deviation $f_i - e_i$. The sum of these k deviations is zero except for rounding. The test criterion suggested by Pearson (1900) is the sum of squares of these deviations, normalized by the expected frequency, or

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{1.3.1}$$

A large value of Q would reflect an incompatibility between the observed and expected relative frequencies, and therefore the null hypothesis should be rejected. The exact probability distribution of the random variable Q is quite complicated, but for large samples its distribution is approximately χ^2 with $k - 1$ degrees of freedom. The theoretical basis for this can be argued briefly as follows.

The only random variables of concern are the class frequencies $F_1, F_2, ..., F_k$, which constitute a set of random variables from a k-variate multinomial distribution with k possible outcomes, the i^{th} outcome being the i^{th} category in the classification system. With $\theta_1, \theta_2, ..., \theta_k$ denoting the probabilities of the respective outcomes and $f_1, f_2, ...f_k$

denoting the observed outcomes, the likelihood function of the sample then is

$$L(\theta_1, \theta_2, ...\theta_k) = \frac{n!}{f_1! f_2! ... f_k!} \prod_{i=1}^{k} \theta_i^{f_i} \; ; \quad \sum_{i=1}^{k} f_i = n; \; \sum_{i=1}^{k} \theta_i = 1. \tag{1.3.2}$$

The null hypothesis was assumed to specify the population distribution completely, from which the θ_i can be calculated. This hypothesis is actually concerned only with the values of these parameters and can be equivalently stated as

$$H_0 : \theta_i^0 = \frac{e_i}{n} \quad for \;\; i = 1, 2, ..., k.$$

It is easily shown that the maximum-likelihood estimates of the parameters in 2.1.2 are $\hat{\theta}_i = f_i/n$. The likelihood-ratio statistic for this hypothesis then is

$$T = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{L(\theta_1^0, \theta_2^0, ..., \theta_k^0)}{L(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)} = \prod_{i=1}^{k} \left(\frac{\theta_i^0}{\hat{\theta}_i} \right)^{f_i}$$

The distribution of the random variable $-2 \ln T$ can be approximated by the χ^2 distribution. The degrees of freedom are $k - 1$, since the restriction $\sum_{i=1}^{k} \theta_i = 1$ leaves only $k - 1$ parameters in Ω to be estimated independently. We have here

$$-2 \ln T = -2 \sum_{i=1}^{k} f_i \left(\ln \theta_i^0 - \ln \frac{f_i}{n} \right). \tag{1.3.3}$$

The expression in eq. (3) can be used as a test criterion for goodness-of-fit. We shall now show that it is asymptotically equivalent to the expression for Q given in 2.1.1. The Taylor series expansion of $\ln \theta_i$ about $f_i/n = \hat{\theta}_i$ is

$$\ln \theta_i = \ln \hat{\theta}_i + (\theta_i - \hat{\theta}_i) \frac{1}{\hat{\theta}_i} + \frac{(\theta_i - \hat{\theta}_i)^2}{2!} \left(-\frac{1}{\hat{\theta}_i^{\,2}} \right) + \epsilon$$

so that

$$\begin{aligned} \ln \theta_i^0 - \ln \frac{f_i}{n} &= \left(\theta_i^0 - \frac{f_i}{n} \right) \frac{n}{f_i} - \left(\theta_i^0 - \frac{f_i}{n} \right)^2 \frac{n^2}{2 f_i^2} + \epsilon \\ &= \frac{(n\theta_i^0 - f_i)}{f_i} - \frac{(n\theta_i^0 - f_i)^2}{2 f_i^2} + \epsilon \end{aligned} \tag{1.3.4}$$

where ϵ represents the sum of terms alternating in sign

$$\sum_{j=3}^{\infty}(-1)^{j+1}\left(\theta_i^0 - \frac{f_i}{n}\right)^j \frac{n^j}{j!f_i^j}$$

Substituting 3.1.1 in 2.1.3, we have

$$\begin{aligned} -2\ln T &= -2\sum_{i=1}^{k}(n\theta_i^0 - f_i) + \sum_{i=1}^{k}\frac{(n\theta_i^0 - f_i)^2}{f_i} + \sum_{i=1}^{k}\epsilon' \\ &= 0 + \sum_{i=1}^{k}\frac{(f_i - e_i)^2}{f_i} + \epsilon'' \end{aligned}$$

By the law of large numbers $\frac{F_i}{n}$ is known to be a consistent estimator of θ_i or

$$\lim_{n\to\infty}\left[\frac{1}{n}P(|F_i - n\theta_i| > \epsilon)\right] = 0 \qquad \text{for every } \epsilon > 0$$

Thus we see that $-2lnT$ approaches the statistic Q and the probability distribution of Q converges to that of $-2\ln T$, which is χ^2 distribution with $k-1$ degrees of freedom. An approximate α-level test then is obtained by rejecting H_0 when Q exceeds the $(1-\alpha)$th quantile point of the χ^2 distribution, denoted by $\chi^2_{k-1,\alpha}$. This approximation can be used with confidence as long as every expected frequency is at least equal to 5. For any e_i smaller than 5, the usual procedure is to combine adjacent groups in the frequency distribution until this restriction is satisfied. The number of degrees of freedom then must be reduced to correspond to the actual number of categories used in the analysis. However this rule of 5 should not be considered inflexible. It is conservative, and the χ^2 approximation is often reasonably accurate for expected cell frequencies as small as 1.5.

Any case where the θ_i are completely specified by the null hypothesis is thus easily handled. The more typical situation, however, is where the null hypothesis is composite, i.e., it states the form of the distribution but not all the relevant parameters. For example, when we wish to test whether a sample is drawn from some normal population, μ and σ would not be given. However, in order to calculate the expected frequencies under H_0, μ and σ must be known. If the expected frequencies are estimated from the data as $n\hat{\theta}_i^0$ for $i = 1, 2, ..., k$ the random variable for the goodness-of-fit test in 2.1.1

becomes

$$Q = \sum_{i=1}^{k} \frac{(F_i - n\hat{\theta}_i^0)^2}{n\hat{\theta}_i^0} \tag{1.3.5}$$

The asymptotic distribution of Q then may depend on the method employed for estimation. When the estimates are found by the method of maximum likelihood for the grouped data, the $L(\hat{\omega})$, where the $\hat{\theta}_i^0$ are the MLEs of the θ_i^0 under H_0. The derivative of distribution of T and therefore Q goes through exactly as before except that the dimension of the space ω is increased. The degrees of freedom for Q goes through exactly as before except that the dimension of the space is increased. The degrees of freedom for Q then are $k - 1 - s$, where s is the number of independent parameters in $F_0(x)$ which has to be estimated from the grouped data in order to estimate all the θ_i^0. In the normal goodness-of-fit test, for example, the μ and σ parameter estimates would be calculated from grouped data and used with tables of normal distribution to find the $n\hat{\theta}_i^0$, and the degrees of freedom for k categories would be $k - 3$. When the original data are ungrouped and the MLEs are based on the likelihood function of all the observations, the theory is different. In practice, however, the statistic in 1.3.5 is often treated as a chi-square variable.

Example 1.1

A die was rolled 30 times with the results shown below:

Number of spots	1	2	3	4	5	6
Frequency(x_i)	1	4	9	9	2	5

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, them what is the value of chi-square statistic and decision reached?

Solution In this problem, the null hypothesis is

$$H_0 : p_1 = p_2 = ... = p_6 = \frac{1}{6}$$

The alternative hypothesis is that not all p_i's are equal to $\frac{1}{6}$. The test will be based on

30 trails, so $n = 30$. The test statistic

$$Q = \sum_{i=1}^{6} \frac{(x_i - e_i)^2}{e_i}$$

where $p_1 = p_2 = ... = p_6 = \frac{1}{6}$. Thus

$$e_i = np_i = (30)\frac{1}{6} = 5$$

and

$$\begin{aligned} Q &= \sum_{i=1}^{6} \frac{(x_i - e_i)}{e_i} = \sum_{i=1}^{6} \frac{(x_1 - 5)^2}{5} \\ &= \frac{1}{5}[16 + 1 + 16 + 16 + 9] = 11.6 \end{aligned}$$

The tabulated χ^2 value for $\chi^2_{0.95}(5)$ is given by

$$\chi^2_{0.95}(5) = 11.07.$$

Since

$$11.6 = Q > \chi^2_{0.95}(5) = 11.07.$$

the null hypothesis $H_0 : p_1 = p_2 = ... = p_6 = \frac{1}{6}$ should be rejected.

Example 1.2

A quality control engineer has taken 50 samples of size 13 each from a production process. The numbers of defective for these samples are recorded below. Test the null hypothesis at level 0.05 that the number of defective follows

1. The Poisson distribution.

2. The Binomial distribution.

Number of defects	0	1	2	3	4	5	6 or more
number of samples	10	24	10	4	1	1	0

Solution Since the data are grouped and both of the hypothesized null distributions are discrete, the chi-square goodness-of-fit test is appropriate. Since no parameters are specified, they must be estimated from data in order to carry out the test in both (1) and (2).

(1) The Poisson distribution is $f(x) = \frac{\exp^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, ...$, where λ here is the mean number of defectives in a sample of size 13. The maximum-likelihood estimate of λ is the mean number of defectives in the 50 samples, that is,

$$\hat{\lambda} = \frac{0(10) + 1(24) + 2(10) + 3(4) + 4(1) + 5(1)}{50} = \frac{65}{50} = 1.3$$

We use this value in $f(x)$ to estimate the probabilities as $\hat{\theta}_i$ and to compute $\hat{e_i} = 50\hat{\theta}_i$. The final $\hat{\theta}$ is not for exactly 5 defects but rather for 5 or more; this is necessary to make $\sum\hat{\theta} = 1$. The final $\hat{e}$ is less than one, so it is combined with the category above before calculating Q. The final result is

Table 1.1: Calculation of Q for Poisson distribution

Defects	f	$\hat{\theta}$	$\hat{e}$		$(f-\hat{e})^2/\hat{e}$
0	10	0.2725	13.625		0.9644
1	24	0.3543	17.715		2.2298
2	10	0.2303	11.515		0.1993
3	4	0.0998	4.990		0.1964
4	1	0.0324	1.620	} 2.155	0.0111
5 or more	1	0.0107	0.535		
					3.6010

$Q = 3.6010$ with 3 degrees of freedom; we start out with $k-1 = 5$ degree of freedom and lose one for estimating θ and one more for combining the last two categories.But the 0.05 critical value for the chi-square distribution with 3 degrees of freedom is 7.81. Our $Q = 3.6010$ id smaller than this value, so we cannot reject the null hypothesis. Thus our conclusion about the Poisson distribution is that we cannot reject the null hypothesis.

(2) The null hypothesis is that the number of defectives in each sample of 13 follows the binomial distribution with $n = 13$ and p is the probability of a defective in any sample. The maximum-likelihood estimate of p is the total number of defectives, which we found

in (1) to be 65, divided by the $50(13) = 650$ observations, or $p = 65/650 = 0.1$. This is the value we use in the binomial distribution to find $\hat{\theta}$ and $\hat{e} = 50\hat{\theta}$ in Table 1.2.

Table 1.2: Calculation of Q for Binomial distribution

Defects	f	$\hat{\theta}$		$\hat{e}$		$(f-\hat{e})^2/\hat{e}$
0	10	0.2542		12.710		0.5778
1	24	0.3671		18.355		1.7361
2	10	0.2448		12.240		0.4099
3	4	0.0997		4.986		0.1950
4	1	0.0277	0.0342	1.385	1.710	0.0492
5 or more	1	0.0065		0.325		
						2.9680

The final result is $Q = 2.9680$, again with 3 degrees of freedom, so the critical value at 0.05 level is again 7.81. Our conclusion about the binomial distribution is that we cannot reject the null hypothesis.

This example illustrate a common result with chi-square goodness-of-fit tests, i.e., that each of the two (or more) different null hypothesis may be accepted for the same data set. Obviously, the true distribution cannot be both binomial and Poisson at the same time. Thus, the appropriate conclusion on the basis of a chi-square goodness-of-fit test is that we do not have enough information to distinguish between these two distributions.

Chapter 2

The Kolmogorov-Smirnov Test

In the χ^2 goodness-of-fit test, the comparison between expected and observed class frequencies is made for a set of k groups. Only k comparisons are made even though there are n observations, where $k \leq n$. If the n sample observations are values of continuous random variable, as opposed to strictly categorical data, comparisons can be made between observed and expected cumulative relative frequencies for each of different observed values. The cumulative distribution function of the sample or the empirical distribution is an estimate of the population *cdf*. Several goodness-of-fit test statistics are functions of the deviations between the observed cumulative distribution and the corresponding cumulative probabilities expected under the null hypothesis. The function of these deviations used to perform a test might be the sum of squares, or absolute values, or the maximum deviation, to name only a few. The best-known test is the Kolmogorov-Smirnov statistic.

2.1 The Kolmogorov-Smirnov one-sample Statistics

The Kolmogorov-Smirnov one sample statistic is based on differnces between the hypothesized cumulative distribution function $F_0(x)$ and the empirical distribution function of the sample $S_n(x)$ for all x. $S_n(x)$ provides a consistent point estimator for the true distribution $F_x(x)$. Further, by the Glivenko-Cantelli Theorem, we know that as n increases, the step function $S_n(x)$, with jumps occurring at the values of the order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ for the sample, approaches the true distribution function $F_X(x)$ for all x. Therefore, for large n, the deviations between the true function and its statistical image, $|S_n(x) - F_x(x)|$, should be small for all values for x. This result

suggests that the statistic

$$D_n = \sup_x |S_n(x) - F_X(x)| \tag{2.1.1}$$

is, for any n, a reasonable measure of the accuracy of our estimate.

This D_n statistic, called the Kolmogorov-Smirnov one-sample statistic, is particularly useful in nonparametric statistical inference because the probability distribution of D_n does not depend on $F_X(x)$ as long as F_X is continuous. Therefore, D_n is called a distribution-free statistic.

The directional deviations defined as

$$D_n^+ = \sup_x [S_n(x) - F_X(x)]$$

$$D_n^- = \sup_x [F_X(x) - S_n(x)]$$

are called the one-sided Kolmogorov-Smirnov statistics. These measures are distribution free.

Theorem 2.1.1.

The statistics $D_n,\ D_n^+,\ D_n^-$ are completely distribution free for any continuous F_X.

Proof

$$D_n = \sup_x |S_n(x) - F_X(x)| = \max_x (D_n^+, D_n^-)$$

defining the additional order statistics $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$,we can write

$$S_n(x) - \frac{i}{n} \quad \textit{for } X_{(i)} \leq x < X_{(i+1)},\ i = 0, 1, ..., n$$

Therefore, we have

$$\begin{aligned} D_n^+ = \sup_x [S_n(x) - F_X(x)] &= \max_{0 \leq i \leq n} \sup_{X_{(i)} \leq x < X_{(i+1)}} [S_n(x) - F_X(x)] \\ &= \max_{0 \leq i \leq n} \sup_{X_{(i)} \leq x < X_{(i+1)}} \left[\frac{i}{n} - F_X(x)\right] \end{aligned}$$

$$
\begin{aligned}
&= \max_{0\le i\le n}\left[\frac{i}{n} - \inf_{X_{(i)}\le x<X_{(i+1)}} F_X(x)\right] \\
&= \max_{0\le i\le n}\left[\frac{i}{n} - F_X(X_{(i)})\right] \\
&= \max\left\{\max_{0\le i\le n}\left[\frac{i}{n} - F_X(X_{(i)})\right], 0\right\}
\end{aligned}
\tag{2.1.2}
$$

Similarly

$$
\begin{aligned}
D_n^- &= \max\left\{\max_{0\le i\le n}\left[F_X(X_{(i)}) - \frac{i-1}{n}\right], 0\right\} \\
D_n &= \max\left\{\max\left\{\max_{0\le i\le n}\left[\frac{i}{n} - F_X(X_{(i)})\right], \max_{0\le i\le n}\left[F_X(X_{(i)}) - \frac{i-1}{n}\right]\right\}\right\}
\end{aligned}
\tag{2.1.3}
$$

Hence D_n, D_n^+, D_n^- all are distribution free as they depend only on the random variables $F_X(X_{(i)}),\ i = 1, 2, ..., n$ and independent of the particular F_X.

Theorem 2.1.2.

For $D_n = \sup_x |S_n(x) - F_X(x)|$ where $F_X(x)$ is any continuous *cdf*, we have

$P\big(D_n < \frac{1}{2n} + v\big)$

$$
= \begin{cases}
0 & \text{for v} \le 0, \\
\int_{1/2n-v}^{1/2n+v}\int_{1/3n-v}^{1/3n+v}\cdots\int_{(2n-1)/2n-v}^{(2n-1)/2n+v} \times f(u_1, u_2, ..., u_n)du_n...du_1 & \text{for } 0 < v < \frac{2n-1}{2n} \\
1 & \text{for } v \ge \frac{2n-1}{2n},
\end{cases}
$$

where

$$f(u_1, u_2, ...u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < ...u_n < 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

Proof

$$\begin{aligned} P\left(D_n < \frac{1}{2n} + v\right) &= P\left[\sup_x |S_n(x) - x| < \frac{1}{2n} + v\right] \\ &= P\left[|S_n(x) - x| < \frac{1}{2n} + v, \text{for all x}\right] \\ &= P\left[\left|\frac{i}{n} - x\right| < \frac{1}{2n} + v, for X_{(i)} \leq x < X(i+1), \text{for all } i = 0, 1, ...n\right] \end{aligned}$$

$$= P\left[\frac{i}{n} - \frac{1}{2n} - v < x < \frac{i}{n} + \frac{1}{2n} + v, \ for\ X_{(i)} \leq x < X(i+1)\right], \quad \text{for all } i = 0, 1, ..., n$$

$$= P\left[\frac{2i-1}{2n} - v < x < \frac{2i+1}{2n} + v, \text{for } X_{(i)} \leq x < X(i+1)\right], \quad \text{for all } i = 0, 1, ..., n$$

Consider any two consecutive values of i. For any $0 \leq i \leq n-1$, both

$$A_i : \left\{\frac{2i-1}{2n} - v < x < \frac{2i+1}{2n} + v, \text{for } X_{(i)} \leq x \leq X(i+1)\right\}$$

and

$$A_{(i+1)} : \left\{\frac{2i-1}{2n} - v < x < \frac{2i+3}{2n} + v, \text{for } X_{(i+1)} \leq x \leq X(i+2)\right\}$$

Since $X_{(i+1)}$ is the random variable common to both events and common set of x is $(2i+1)/2n - v < x < (2i+1)/2n + v$ for $V \geq 0$, the event $A_{(i)} \bigcap A_{(i+1)}$ for any

$0 \leq i \leq n-1$ is

$$\frac{2i-1}{2n} - v < X_{(i+1)} < \frac{2i+1}{2n} + v \quad \text{for all } v \geq 0$$

In other words,

$$\frac{2i-1}{2n} - v < X_{(i+1)} < \frac{2i+1}{2n} + v \quad \text{for } X_{(i)} \leq x \leq X(i+1), \text{ for all } i = 0, 1, ..., n$$

if and only if

$$\frac{2i+1}{2n} - v < X_{(i+1)} < \frac{2i+1}{2n} + v \quad \text{for all } i = 0, 1, ..., n-1; \; v \geq 0$$

The joint probability distribution of order statistics is

$$f_{X_{(1)},X_{(2)},...,X_{(n)}}(x_1, x_2, ..., x_n) = n! \quad \text{for } x_1 < x_2 < ... < x_n < 1$$

Putting all this together now, we have

$$\begin{aligned} & P\left(D_n < \frac{1}{2n} + v\right) \qquad \text{for all } -\frac{1}{2n} < v < \frac{2n-1}{2n} \\ = P & \left(\frac{2i+1}{2n} - v < X_{(i+1)} < \frac{2i+1}{2n} + v \quad \text{for all } i = 0, 1, ..., n-1\right) \\ & 0 \leq v < \frac{2n-1}{2n} \\ = P & \Big[\left(\frac{1}{2n} - v < X_{(1)} < \frac{1}{2n} + v\right) \bigcap \left(\frac{3}{2n} - v < X_{(2)} < \frac{3}{2n} + v\right) \\ & \times \bigcap ... \bigcap \left(\frac{2n-1}{2n} - v < X_{(n)} < \frac{2n-1}{2n} + v\right) \Big] \\ & \text{for all } 0 \leq v < \frac{2n-1}{2n} \end{aligned}$$

which is equivalent to the stated integral.

Consider $n = 2$, for all $0 \leq i \leq 3/4$,

$$P(D_2 < 1/4 + v) = 2! \int_{1/4-v}^{1/4+v} \int_{3/4-v}^{3/4+v} du_2 du_1 \quad 0 < u_1 < u_2 < 1$$

The limits overlap when $1/4 + v \geq 3/4 - v$, or $v \geq 1/4$. When $0 \leq v < 1/4$, we have $u_1 < u_2$ automatically. Therefore, for $0 \leq v < 1/4$,

$$P(D_2 < 1/4 + v) = 2 \int_{1/4-v}^{1/4+v} \int_{3/4-v}^{3/4+v} du_2 du_1 = 2(2v)^2$$

But for $1/4 \leq v \leq 3/4$, the region of integration is as figure. Dividing the integral into two pieces, we have for $1/4 \leq v < 3/4$,

$$\begin{aligned} P(D_2 < 1/4 + v) &= 2 \left[\int_{3/4-v}^{1/4+v} \int_{u_1}^{1} du_2 du_1 + \int_{0}^{3/4-v} \int_{3/4-v}^{1} du_2 du_1 \right] \\ &= -2v^2 + 3v - 1/8. \end{aligned}$$

Collecting the results for all v,

$$P(D_2 < 1/4 + v) = \begin{cases} 0 & \text{for } 0 < u_1 < u_2 < ...u_n < 1 \\ 2(2v)^2 & \text{for } 0 < v < 1/4 \\ -2v^2 + 3v - 0.125 & \text{for } 1/4 \leq v < 3/4 \\ 1 & \text{for } v \geq 3/4. \end{cases}$$

The inverse procedure is to find that number $D_{n,\alpha}$ such that $P(D_n, D_{n,\alpha}) = \alpha$. In our numerical example with $n = 2, \alpha = 0.05$,we find v such that

$$P(D_2 > 1/4 + v) = 0.05 \quad \text{or} \quad P(D_2 < 1/4 + v) = 0.95$$

and then set $D_{2,0.05} = 1/4 + v$. From the previous evaluation of the D_2 sampling distribution, either

$$2(2v)^2 = 0.95 \quad \text{and} \quad 0 < v < 1/4$$

or

$$-2v^2 + 3v - 0.125 = 0.95 \quad \text{and} \quad 1/4 \leq v < 3/4$$

The first result has no solution, but the second yields the solution $v = 0.5919$. Therefore, $D_{2,0.05} = 0.8419$.

Theorem 2.1.3.

If F_X is any continuous distribution function, then for every $d > 0$,

$$\lim_{n\to\infty} P\left(D_n \leq d/\sqrt{n}\right) = L(d)$$

where

$$L(d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2d^2}$$

Some of the results for the asymptotic approximation to $D_{n,\alpha} = d_\alpha/\sqrt{n}$ are:

$P(D_n > d_\alpha/\sqrt{n})$	0.20	0.15	0.10	0.05	0.01
d_α	1.07	1.14	1.22	1.36	1.63

The approximation has been found to be close enough for practical application as long as n exceeds 35. A comparison of exact and asymtotic values of $D_{n,\alpha}$ for $\alpha = 0.01$ and 0.05.

Theorem 2.1.4.

For $D_n^+ = \sup_x [S_n(x) - F_X(x)]$ where $F_X(x)$ is any continuous *cdf*, we have

$$P\big(D_n^+ < c\big) \;=\; \begin{cases} 0 & c \;\leq\; 0, \\ \int_{1-c}^{1} \int_{n-1/n-c}^{u_n} \ldots \int_{2/n-c}^{u_3} \int_{1/n-c}^{u_2} f(u_1, u_2, ..., u_n) du_1 ... du_n & 0 < c < 1 \\ 1 & \text{for } c \geq 1, \end{cases}$$

where

$$f(u_1, u_2, ...u_n) \;=\; \begin{cases} n! & \text{for } 0 < u_1 < u_2 < ...u_n < 1 \\ 0 & \text{otherwise} \end{cases}$$

Proof Assume that F_X is the uniform distribution on (0,1). Then we can write

$$D_n^+ = \max\left[\max_{1\leq i\leq n}\left(\frac{i}{n} - X_{(i)}\right), 0\right]$$

For all $0 < c < 1$, we have

$$\begin{aligned} P(D_n^+ < c) &= P\left[\max_{1\leq i\leq n}\left(\frac{i}{n} - X_{(i)}\right) < c\right] \\ &= P\left(\frac{i}{n} - X_{(i)} < c \quad \text{for all } i = 1, 2, ..., n\right) \\ &= P\left(X_{(i)} > \frac{i}{n} - c \quad \text{for all } i = 1, 2, ..., n\right) \\ &= \int_{1-c}^{\infty} \int_{(n-1)/n-c}^{\infty} \ldots \int_{2/n-c}^{\infty} \int_{1/n-c}^{\infty} f(x_1, x_2, ..., x_n) dx_1 ... dx_n \end{aligned}$$

where

$$f(x_1, x_2, ..., x_n) = \begin{cases} n! & \text{for } 0 < x_1 < x_2 < ...x_n < 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

which is equivalent to the stated integral.

Theorem 2.1.5.

If F_X is any continuous distribution function, then for every $d \geq 0$

$$\lim_{n\to\infty} P\big(D_n^+ < d/\sqrt{n}\big) = 1 - e^{-2d^2}$$

As a result of this theorem, χ^2 tables can be used for the distribution of a function of D_n^+ because of the following corollary.

Corollary

If F_X is any continuous distribution function, then for every $d \geq 0$, the limiting distribution of $V = 4nD_n^{+2}$, as $n \to \infty$, is the chi-square distribution with 2 degrees of freedom.

Proof We have $D_n^+ < d/\sqrt{n}$ if and only if $4nD_n^{+2} < 4d^2$ or $V < 4d^2$. Therefore

$$\lim_{n\to\infty} P(V < 4d^2) = \lim_{n\to\infty} P\big(D_n^+ < d/\sqrt{n}\big) = 1 - e^{-4d^2/2}$$

$$\lim_{n\to\infty} P(V < c) = 1 - e^{-c/2} \qquad for\ all\ c > 0$$

The right-hand side is the cdf of a χ^2 distribution with 2 degrees of freedom.

The procedure is to set $4nD_{n,0.05}^{+2} = 5.99$ and solve to obtain

$$D_{n,0.05}^+ \sqrt{1.4975/n} = 1.22/\sqrt{n}$$

2.1.1 Application of The Kolmogorov-Smirnov one-sample Statistics

The statistical use of the Kolmogorov-Smirnov statistic in a goodness-of-fit type of problem is obvious. Assume we have the random sample $X_1, X_2, ..., X_n$ and the hypothesis-testing situation $H_0 : F_X(x) = F_0(x)$ for all x, where $F_0(x)$ is a completely specified continuous distribution function.
Since $S_n(x)$ is the statistical image of the population distribution $F_X(x)$, the difference between $S_n(x)$ and $F_0(x)$ should be small for all x except for sampling variation, if the null hypothesis is true. For the usual two-sided goodness-of-fit alternative.

$$H_1 : F_X(x) \neq F_0(x) \quad \text{for some } x$$

large absolute values of these deviations tend to discredit the hypothesis. Therefore, the Kolmogorov-Smirnov goodness-of-fit test with significance level α is to reject H_0 when $D_n > D_{n,\alpha}$. From the Glivenko-Cantelli theorem, we know that $S_n(x)$ converges to $F_X(x)$ with probability 1, which implies consistency.
The value of the Kolmogorov-Smirnov goodness-of-fit statistics D_n in 2.1.1 can be calculated using 2.1.3 if all n observations have different numerical values. However, the expression below is considerably easier for algebraic calculation and applies when ties are present. The formula is

$$D_n = \sup_x |S_n(x) - F_0(x)| = \max_x [|S_n(x) - F_0(x)|, |S_n(x - \varepsilon) - F_0(x)|]$$

where ε denotes any small positive number.

Example 2.1

The data on the heights of 12 infants are given below:
18.2, 21.4, 22.6, 17.4, 17.6, 16.7, 17.1, 21.4, 20.1, 17.9, 16.8, 23.1. Test the hypothesis that the data came from some normal population at a significance level $\alpha = 0.1$.

Solution Here, the null hypothesis is

$$H_0 \ : \ X \ \sim \ N(\mu, \sigma^2).$$

First we estimate μ and σ^2 from the data. Thus, we get

$$\bar{x} = \frac{230.3}{12} = 19.2.$$

and

$$s^2 = \frac{4482.01 - \frac{1}{12}(230.3)^2}{12-1} = \frac{62.17}{11} = 5.65.$$

Hence $s = 2.38$. Then by the hypothesis

$$F(x_{(i)}) = P\left(Z \leq \frac{x_{(i)} - 19.2}{2.38}\right)$$

where $Z \sim N(0,1)$ and $i = 1, 2, ..., n$. Next we compare the Kolmogorov-Smirnov statistic D_n the given sample size 12 using the following tabular form.

i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{12} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{12}$
1	16.7	0.1469	-0.0636	0.1469
2	16.8	0.1562	0.0105	0.0729
3	17.1	0.1894	0.0606	0.0227
4	17.4	0.2236	0.1097	-0.0264
5	17.6	0.2514	0.1653	-0.0819
6	17.9	0.2912	0.2088	-0.1255
7	18.2	0.3372	0.2461	-0.1628
8	20.1	0.6480	0.0187	0.0647
9	21.4	0.8212	0.0121	0.0712
10	21.4			
11	22.6	0.9236	-0.0069	0.0903
12	23.1	0.9495	0.0505	0.0328

Thus

$$D_{12} = 0.2461.$$

From the tabulated value, we see that $d_{12} = 0.34$ for significance level $\alpha = 0.1$. Since D_{12} is smaller than d_{12} we accept the null hypothesis $H_0 : X \sim N(\mu, \sigma^2)$. Hence the data came from the normal population.

Example 2.2

Let $X_1, X_2, ..., X_{10}$ be a random sample from a distribution whose probability density function is

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Based on the observed values 0.62, 0.36, 0.23, 0.76, 0.65, 0.09, 0.55, 0.26, 0.38, 0.24, test the hypothesis $H_0 : X \sim UNIF(0,1)$ against $H_a : X \nsim UNIF(0,1)$ at a significance level $\alpha = 0.1$.

Solution The null hypothesis is $H_0 : X \sim UNIF(0,1)$. Thus

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \\ x & \text{if } 0 \leq x < 1 \\ \\ 1 & \text{if } x \geq 1 \end{cases}$$

Hence

$$F(x_{(i)}) = x_{(i)} \quad \text{for } i = 1, 2, ..., n.$$

Next we compute the Kolmogorov-Smirnov statistic D_n the given sample of size 10 using the following tabular form.

i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{10} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{10}$
1	0.09	0.09	0.01	0.09
2	0.23	0.23	-0.03	0.13
3	0.24	0.24	0.06	0.04
4	0.26	0.26	0.14	-0.04
5	0.36	0.36	0.14	-0.04
6	0.38	0.38	0.22	-0.12
7	0.55	0.55	0.15	-0.05
8	0.62	0.62	0.18	-0.08
9	0.65	0.65	0.25	-0.15
10	0.76	0.76	0.24	-0.14

Thus

$$D_{10} = 0.25.$$

From the tabulated value, we see that $d_{10} = 0.37$ for significance level $\alpha = 0.1$. Since D_{10} is smaller than d_{10} we accept the null hypothesis

$$H_0 : X \sim UNIF(0, 1).$$

Example 2.3

The 20 observations below were chosen randomly from the continuous uniform distribution over (0,1), recorded to four significant figures, and rearranging in increasing order of the magnitude. Determine the values of D_n, and test the null hypothesis that the square roots of these numbers also have the continuous uniform distribution over (0,1).

0.0123	0.1039	0.1954	0.2621	0.2802
0.3217	0.3645	0.3919	0.4240	0.4814
0.5139	0.5846	0.6275	0.6541	0.6889
0.7621	0.8320	0.8871	0.9249	0.9634

Solution The calculation needed to find D_n are shown in Table 2.1. The entries in first column, labeled x, are not the observations above, but their respective square roots, because the null hypothesis is concerned with the distribution of these square roots. The $S_n(x)$ are the proportions of observed values less than or equal to each different observed x. The hypothesized distribution here is $F_0(x) = x$, so the third column is exactly the same as the first column. The fourth column is the difference $F_n(x - \varepsilon) - F_0(x)$, that is, the difference between the S_n value for a number slightly smaller than an observed x and the F_0 value for that observed x. Finally the sixth and seventh columns are the absolute values of the differences of the numbers in the fourth and the fifth columns. The supremum is the largest entry in either of the last two columns; its value here in $D_n \geq 0.352$, so we reject the null hypothesis that these numbers are uniformly distributed.

The theoretical justification behind this example is as follows:

Let Y have the continuous uniform distribution on (0,1) so that $f_Y(y) = 1$ for $0 \leq y \leq 1$.

Table 2.1: Calculation of D_n for Example 2.3

x	$S_n(x)$	$F_0(x)$	$S_n(x) - F_0(x)$	$S_n(x-\varepsilon) - F_0(x)$	$\lvert S_n(x) - F_0(x)\rvert$	$\lvert S_n(x-\varepsilon) - F_0(x)\rvert$
0.11	0.05	0.11	-0.06	-.011	0.06	0.11
0.32	0.10	0.32	-0.22	-0.27	0.22	0.27
0.44	0.15	0.44	-0.29	-0.34	0.29	0.34
0.51	0.20	0.51	-0.31	-0.36	0.31	0.36
0.53	0.25	0.53	-0.28	-0.33	0.28	0.33
0.57	0.30	0.57	-0.27	-0.32	0.27	0.32
0.60	0.35	0.60	-0.25	-0.30	0.25	0.30
0.63	0.40	0.63	-0.23	-0.28	0.23	0.28
0.65	0.45	0.65	-0.20	-0.25	0.20	0.25
0.69	0.50	0.69	-0.19	-0.24	0.19	0.24
0.72	0.55	0.72	-0.17	-0.22	0.17	0.22
0.76	0.60	0.76	-0.16	-0.21	0.16	0.21
0.79	0.65	0.79	-0.14	-0.19	0.14	0.19
0.81	0.70	0.81	-0.11	-0.16	0.11	0.16
0.83	0.75	0.83	-0.08	-0.13	0.08	0.13
0.87	0.80	0.87	-0.07	-0.12	0.07	0.12
0.91	0.85	0.91	-0.06	-0.11	0.06	0.11
0.94	0.90	0.94	-0.04	-0.09	0.04	0.09
0.96	0.95	0.96	-0.01	-0.06	0.01	0.06
0.98	1.00	0.98	-0.02	-0.03	0.02	0.03

Then the pfd of $X = sqrtY$ can be shown to be $f_X(x) = 2x$ for $0 \leq x \leq 1$, which is not uniform.

ONE-SIDED TESTS

With the statistics D_n^+ and D_n^-, it is possible to use Kolmogorov-Smirnov statistics for a one-sided goodness-of-fit test which would detect directional differences between $S_n(x)$ and $F_0(x)$. For the alternative

$$H_{1,+} : F_X(x) \geq F_0(x) \qquad \text{for all } x$$

the appropriate rejection region is $D_n^+ > D_{n,\alpha}^+$, for the alternative

$$H_{1,-} : F_X(x) \leq F_0(x) \qquad \text{for all } x$$

H_0 is rejected when $D_n^- > D_{n,\alpha}^-$. Both of these tests are consistent against their respective alternatives.

Most tests of goodness of fit are two-sided. However, it is useful to know that the tail

probabilities for the one-sided statistics are approximately one-half of the corresponding tail probabilities for the two-sided statistics.

2.2 The Kolmogorov-Smirnov Two-Sample Test

The Kolmogorov-Smirnov statistic is another one-sample test that can be adapted to the two-sample problem. In the two-sample case , the comparison is made between the empirical distribution functions of the two samples.

The order statistics corresponding to two random samples of size m and n from continuous populations F_X and F_Y, are $X_{(1)}, X_{(1)}, ..., X_{(m)}$ and $Y_{(1)}, Y_{(1)}, ..., Y_{(n)}$. Their respective empirical distribution functions, denoted by $S_m(x)$ and $S_n(x)$, are defined as before:

$$S_m(x) \;=\; \begin{cases} 0 & \text{if } x < X_{(1)} \\ k/m & \text{if } X_{(k)} \le x < X_{(k+1)} \quad \text{for } k = 1, 2, ..., m-1 \\ 1 & \text{if } x \ge X_{(m)} \end{cases}$$

and

$$S_n(x) \;=\; \begin{cases} 0 & \text{if } x < Y_{(1)} \\ k/n & \text{if } Y_{(k)} \le x < Y_{(k+1)} \quad \text{for } k = 1, 2, ..., n-1 \\ 1 & \text{if } x \ge Y_{(n)} \end{cases}$$

In a combined ordered arrangement of the $m + n$ sample observation, $S_m(x)$ and $S_n(x)$ are the respective proportions of X and Y observations which do not exceed the specified value x.

If the null hypothesis

$$H_0 : F_Y(x) \;=\; F_X(x) \quad \text{for all } x$$

is true, the population distributions are identical and we have two samples from the same population. The empirical distribution functions for the X and Y samples are reasonable estimates of their respective population *cdf*. Therefore, allowing for sampling

variation, there should be reasonable agreement between the two empirical distributions if indeed H_0 is true; otherwise the data suggest that H_0 is not true and therefore should be rejected. This is the intuitive logic behind most two-sample tests, and the problem is to define what is a reasonable agreement between the two empirical *cdf*'s. In other words, how close do the two empirical *cdf*'s have to be so that they could be viewed as not significantly different, taking account of the sampling variability. The two-sided Kolmogorov-Smirnov two-samle test criterion, denoted by $D_{m,n}$, is based on the maximum absolute difference between the two empirical distributions

$$D_{m,n} = \max_x |S_m(x) - S_n(x)|$$

Since here only the magnitudes, and not the directions, of the deviations are considered, $D_{m,n}$ is appropriate for a general two-sided alternative

$$H_A : F_Y(x) \neq F_X(x) \quad \text{for some } x$$

and the rejection region is in the upper tail, defined by

$$D_{m,n} \geq c_\alpha$$

where

$$P(D_{m,n} \geq c_\alpha | H_0) \leq \alpha$$

Because of the Glivenko-cantelli theorem, the test is consistent for this alternative. The P value is

$$P(D_{m,n} \geq D_0 | H_0)$$

where D_0 is the observed value of the two-sample K-S ' test statistic. As with the one-sample Kolmogorov-Smirnov statstic, $D_{m,n}$ is completely distribution free for any continuous common population distribution since order is preserved under a monotone transformation. That is, if we let $z - F(x)$ for the common continuous *cdf* F, we have $S_m(z) - S_m(x)$ and $S_n(z) - S_n(x)$, where the random variable Z, corresponding to z, has the uniform distribution on the unit interval.

ONE-SIDED ALTERNATIVES

A one-sided two-sample maximum-unidirectional-deviation test can also be defined, based on the staistics

$$D^{+}_{m,n} \;=\; \max_{x}[S_m(x) - S_n(x)]$$

For an alternative that the X random variables are stochastically smaller than the Y's

$$H_1 \;:\; F_Y(x) \leq F_X(x) \qquad \text{for all } x$$

$$F_Y(x) < F_X(x) \qquad \text{for some } x$$

the rejection region should be

$$D^{+}_{m,n} \geq c_\alpha$$

The one-sided test based on $D^{+}_{m,n}$ is also distribution free and consistent against the alternative H_1. Since either sample may be labeled the X sample, it is not necessary to define another one-sided statistic for the alternative that X is stochastically larger than Y.

2.2.1 Application of The Kolmogorov-Smirnov two-sample Statistics

Example 2.4

A random sample of size 9, $X_1, ..., X_9$ is obtained from one population, and a random sample of size 15, $Y_1, ...Y_{15}$ is obtained from a second population. A graph of their empirical distribution functions is given in figure 2.2.1 .

The null hypothesis is that the two populations have identical distribution functions. If the respective distribution functions are denoted by $F(x)$ and $G(x)$, then the null hypothesis may be written as

$$H_0 : F(x) \;=\; G(x) \quad \text{for all } -\infty < x < \infty$$

x

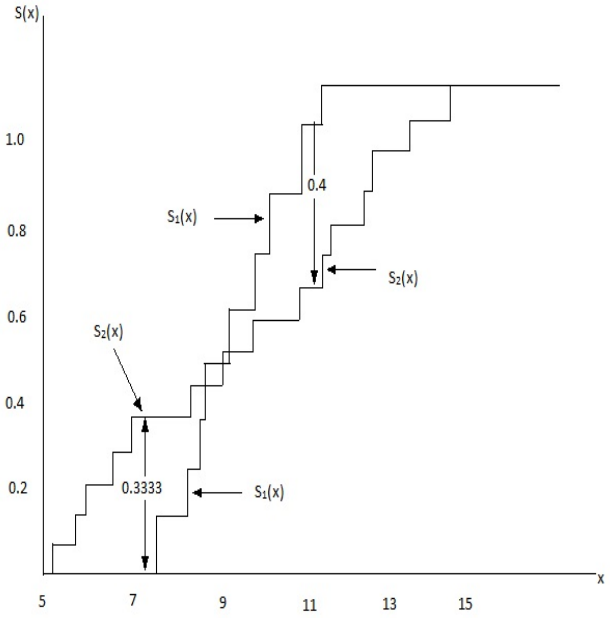

Figure 2.2.1: Graphs of $S_1(x)$, the e.d.f of Y, showing the maximum distance between them

The alternative hypothesis may be stated as

$$H_1 : F(x) \neq G(x) \quad \text{for at least one value of } x$$

The test statistic for the two-sided test is given as

$$\begin{aligned} T_1 &= \sup |S_1(x) - S_2(x)| \\ &= \frac{2}{5} = 0.400 \end{aligned}$$

the largest absolute difference between $S_1(x)$ and $S_2(x)$, which happens to occur between $x = 11.2$ and $x = 11.3$. The value of 0.400 for T_1 could also have been determined graphically in figure. From the graph one can easily see that the difference $S_1(x)$ and $S_2(x)$ changes only at those observed values $x = X_i$ or $x = Y_i$, and that is why it is sufficient to compute $S_1(x) - S_2(x)$ only at the observed sample values, as done here.

From table we see that the 0.95 quantile of T_1, for the two-sided test and for $n = 9 = N_1$ and $m = 15 = N_2$, is given as $w_{0.95} = \frac{8}{15}$. For these data T_1 value may be estimated as

X_i	Y_i	$S_1(x) - S_2(x)$	X_i	Y_i	$S_1(x) - S_2(x)$
	5.2	0 - 1/15 = -1/15		9.8	5/9 - 8/15 = 1/45
	5.7	0 - 2/15 = -2/15	9.9		6/9 - 8/15 = 2/15
	5.9	0 - 3/15 = -1/5	10.1		7/9 - 8/15 = 11/45
	6.5	0 - 4/15 = -4/15	10.6		8/9 - 8/15 = 16/45
	6.8	0 - 5/15 = -1/3		10.8	8/9 - 9/15 = 13/45
7.6		1/9 - 5/15 = -2/9	11.2		1 - 9/15 = 2/5
	8.2	1/9 - 6/15 = -13/45		11.3	1 - 10/15 = 1/3
8.4		2/9 - 6/15 = -8/45		11.5	1 - 11/15 = 4/15
8.6		3/9 - 6/15 = -1/15		12.3	1 - 12/15 = 1/5
8.7		4/9 - 6/15 = 2/4		12.5	1 - 13/15 = 2/15
	9.1	4/9 - 7/15 = -1/15		13.4	1 - 14/15 = 1/15
9.3		5/9 - 7/15 = 4/45		14.6	1 - 1 = 0

slightly larger than 0.20.

For the sake of comparison, the approximate 0.95 quantile based on the asymptotic distribution is found to be

$$w_{0.95} \cong 1.36\sqrt{\frac{m+n}{mn}} = 0.573$$

which is slightly larger than the exact value of $\frac{8}{15} = 0.533$. This illustrates the tendency of the asymptotic approximation to furnish a conservative test.

Chapter 3

A two-sample distribution-free scale test of the Smirnov type

Let $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ be two random samples from populations with absolutely continuous distribution functions F_1 and F_2, respectively, having equal means/medians. If D^+ and D^- are usual one-sided Smirnov statistics, $D = \max(D^+, D^-)$ is the two-sided statistic that is used as a test against general alternatives. The statistic defined by

$$G = \min(D^+, D^-)$$

is a useful statistic against scale alternatives. Since G is available any time D is computed, both D and G should be utilized; even if D is nonsignificant, a large value of G is possible which indicates a scale change from the null hypothesis of identical populations.

Let F_m and F_n be the usual empirical distribution functions for X and Y, respectively; and $W_{(i)}$ $i = 1, 2, ..., m+n$ be the pooled order statistics where each $W_{(i)}$ is either an X or Y observation. With $U_0 = 0$ and for $i = 1, 2, ..., m+n$ Let

$$U_{(i)} = \begin{cases} U_{(i-1)} + \frac{1}{m} \equiv F_m(W_{(i)}) - F_n(W_{(i)}), & \text{if } W_{(i)} \text{ is from the } X \text{ sample} \\ \\ U_{(i-1)} - \frac{1}{n} \equiv F_m(W_{(i)}) - F_n(W_{(i)}), & \text{if } W_{(i)} \text{ is from the } Y \text{ sample} \end{cases}$$

Then

$$D^+ = \max_{1 \le i \le m+n} (U_{(i)}), \qquad D^- = \max_{1 \le i \le m+n} (-U_{(i)}), \quad \text{and}$$

$$D = \max(D^+, D^-)$$

Gideon and Mueller(1978) used a somewhat different notation and multiplied the $U_{(i)}$ by mn in order to obtain a simple addition-of-integer calculation of the Smirnov statistic; they also demonstrated the result given in the following paragraph.

An informative graph is obtained by plotting the numbers $U_{(i)}$ $i = 1, 2, ..., m + n$ at the points $\big(F_m(W_{(i)}),\ F_n(W_{(i)})\big)$. The absolute values of the $U_{(i)}$ give the vertical distance of the point $\big(F_m(W_{(i)}),\ F_n(W_{(i)})\big)$ from the diagonal line from $(0, 0)$ to $(1, 1)$.

EXAMPLE

An example from Conover (1980), p. 370 is given as
A random sample of size 9, $X_1, ..., X_9$ is obtained from one population, and a random sample of size 15, $Y_1, ...Y_{15}$ is obtained from a second population. A graph of their empirical distribution functions is given as

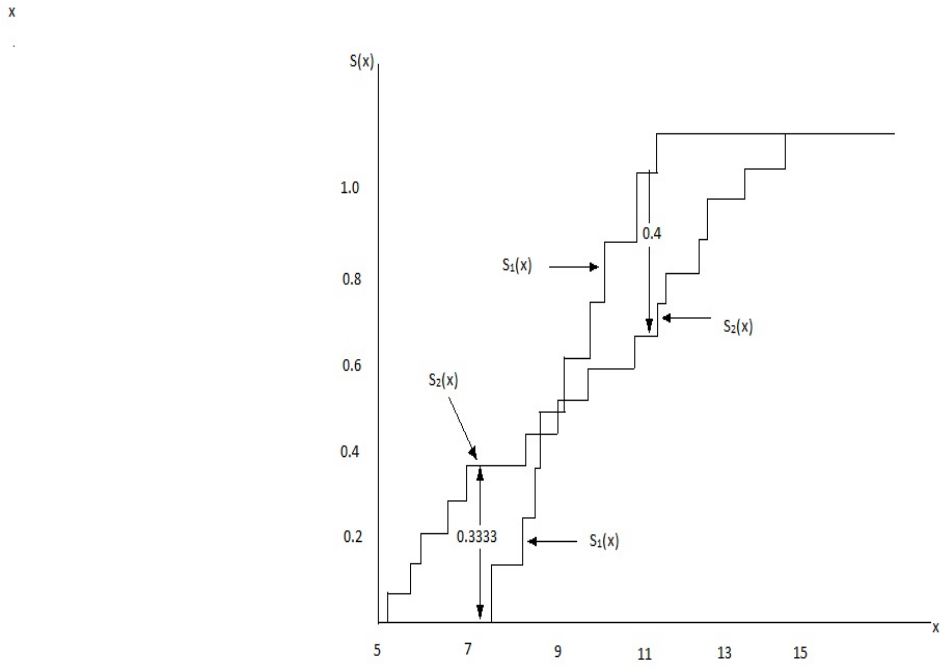

Figure 3.0.1: Graphs of $S_1(x)$, the e.d.f of Y, showing the maximum distance between them

In this example D is not significant but G is significant is used to illustrate the statistics. In order to keep the table of the G statistics in integer form, this section uses mnD^+

and mnD^- which have integer values. In the pooled ordered data given below, the X variable data($m = 9$) are starred whie the Y variable data($n = 15$) are not.

$W_{(i)}, i = 1, 2, ..., 24, =$ (5.2, 5.7, 5.9, 6.5, 6.8, 7.6*, 8.2, 8.4*, 8.6*, 8.7*, 9.1, 9.3*, 9.8, 9.9*, 10.1*, 10.6*, 10.8, 11.2*, 11.3, 11.5, 12.3, 12.5, 13.4, 14.6). The corresponding 9.15 $U_{(i)}, i = 1, 2, ..., 24, =$ (-9, -18, -27, -36, -45, -30, -39, -24, -9, 6, -3, 12, 3, 18, 33, 48, 39, 54, 45, 36, 27, 18, 9, 0). The X and Y variables are independent and the Smirnov test was used to test $H_0 : F_1(x) = F_2(x)$ for all x against a general alternative. It is easily seen that $mnD^- = 135D^- = 45$ and $mnD^+ = 135D^+ = 54$. Thus $135D = 54$. From tables of D for the Smirnov statistic for $\alpha = 0.05$, $P(135D > 72) = 0.05$ so that D is nonsignificant.

However, $mnG = 135G = 45$; and from the Table of Critical Values for G, $P(135G \geq 39) = 0.0168$. Since $45 > 39$, G is significant. Thus while D was found to be nonsignificant, the significance of G indicates the data contract H_0 in the direction of a difference in scale rather than in location.

3.1 Critical Values

The null distribution of mnD^+ and mnD^- are identical since for each arrangement of Xs and Ys that gives rise to a particular value of mnD^+ the reverse ordering gives that same value for mnD^-. For example with $m = 2$ and $n = 3$, $XYYXY$ gives $6D^+ = 3$ and $6D^- = 1$ while the reverse arrangement $YXYYX$ gives $6D^+ = 1$ and $6D^- = 3$. Thus the following relationship was used to obtain the tail area probabilities for mnG :

$$P(mnG \geq v) = 2P(mnD^+ \geq v) - P(mnD \geq v). \tag{3.1.1}$$

An exact critical value necessarily implies a randomized test since mnG is a discrete statistic. A modified FORTRAN subroutine that was utilized in Mueller(1978) was used to compute $P(mnD^+ \geq v)$ and $P(mnD \geq v)$. These routines were many times more efficient than those that appear in Kim and Jennrich(1973). Given α and an assumable value of G, v was selected so that $P(mnG \geq v) \geq \alpha$ and $P(mnG > v) \leq \alpha$.

Thus an exact $\alpha-$level test (0.01 and 0.05) can be constructed from the following Table by the randomization technique.

The limiting distribution of G is obtained from the relationship 3.1.1 and the Known

limiting distribution of D^+ and D^-, Smirnov(1939):

$$\lim_{\substack{m,n\to\infty \\ n/m\to q>0}} P\left(\sqrt{\frac{mn}{m+n}}G > z\right) = 2\sum_{i=2}^{\infty}(-1)^i \exp(-2i^2z^2)$$

For large sample sizes this limiting distribution leads to the approximate critical values given in Table 3.1. The table is used in exactly the same way as the Kolmogorov-Smirnov large sample tables are used.

Table 3.1: Right-Tail Probability for G based on the Limiting Distribution

$P\left(G > z_\alpha \frac{mn}{m+n}\right)$	0.20	0.10	0.05	0.025	0.01	0.005
z_α	0.5293	0.6094	0.6781	0.7397	0.8137	0.8654

3.2 Monte Carlo Studies

The studies consist of two parts. The first focuses on the statistics D and G; the second refers to an earlier study which examines the rejection regions of these two statistics and relates their power to that of four nonparametric scale tests.

A simulation study was performed to examine the Type I error and power properties of statistics D and G when sampling from a standard standard normal, or uniform (-0.5,0.5) distribution. Equal sample sizes of 7 & 13, 13 & 7, 15 & 25 and 25 & 15. The first sample in each pair is designated as X and the second as Y. Throughout the study, X is related Y by $X = \sigma Y$ where Y is generated from one of the four distributions given above and the scale factor σ is 0.5, 1.0 (the null hypothesis), 1.5,2, or 3. For each situation, 2000 Monte Carlo trails formed. Since the probability distributions of D and G are discrete, assumable alpha levels closest to $\alpha = 0.05$ were used. In all cases, the difference in the alpha level used for G was within 0.004 of that used for D.

The results for equal sample sizes are given in Table 3.2 and for unequal sample sizes in Table 3.3. When $\sigma = 1$, the null hypothesis is true and it is possible o compare the actual sizes of the test with the nominal values. In general, the differences between these values appear to be due to random variation; no inherent pattern over either distributions or samples sizes is apparent. An exception to this occurs for the D statistic for sample sizes $m = n = 30$ and $m = n = 40$. In both instances the actual test size is always below the nominal level: the relative errors range from -31% to-16% in the first case

and from -50% to -39% in the second. These discrepancies are most likely due to the fact that the table by Kim and Jennrich (1973) gives for sample sizes greater than 25 the critical value of mnD whose actual size of Type I error is less than or equal to the stated nominal level.

The distribution in Table 3.2 and 3.3 are listed in the order that the power of G, in general, increases for a fixed scale factor. The most notable exception occurs for $\sigma = 1.5$ where the power of G is lower for the normal exception occurs for $\sigma = 1.5$ where the power of G is lower for the normal distribution than for the Laplace distribution in five of the eight cases. Due to the symmetry of the problem when $m = n$, the results for the scale factors 0.5 and 2 are essentially the same. In all cases G is more powerful than D, often sudstantially so. Only in the case of large, equal sample sizes (30 & 40) and the uniform distribution for $m = n = 20$ in given figure.

In an earlier Monte Carlo study by Rothan (1982) involving the Cauchy and normal distributions and samples of sizes 6,6; 5,7; 4,8; 10,10; 9,11; 8,12; 7,13; 20,20; 19,21; 18,22; and 15,15 with randomized tests, a preliminary examination of the output appeared to indicate that under the null hypothesis of identical distributions the statistics D and G were rejecting different types of samples. Define $D \cup G$ as a test which rejects if D exceeds its critical value or G exceeds its critical value (or if both exceed their respective critical values). For the null distributions studied, the rejection region was essentially partitioned into two disjoint regions. At the $\alpha = 0.01$ level, the overlap was no greater than 1.77% ($m = n = 20$, normal distribution) and the mean overlap was 0.51%. Thus D and G approximately additive for the null distribution and $D \cup G$ rejects roughly at the 2α level when D and G each reject at the α level, for $\alpha \leq 0.05$. This additivityof D and G for the null distribution supports the claim that D and G are sensitive to different types of alternatives.

In this same study, the following nonparametric scale tests were included in the power comparisons: (A) Freund-Ansari-Bradley (1957,1960), (S) Siegel-Tukey (1960), and (Z) Koltz normal scores (1962). For the normal distribution the $Z-$test was most powerful. For the Cauchy distribution, the Z-test tends to be less powerful than the A and S-tests and even less powerful than G for some scale factors less than 1. One typical example for the Cauchy distribution follows: for $m = n = 20$, $\sigma = 4.0$and $\alpha = 0.01$,the powers of the various tests were 0.08 (D), 0.33 (G), 0.40 (Z), 0.54 (A and S). Although the power

of G is often less than that of A, S, or Z, users of D have G at no additional computing expense whereas the other statistics must be computed in a different manner.

In the event that symmetric distributions do not have equal means/medians, a test proposed by Blair and Thomson (1992) could be applied. Their statistics B_W and B_S are shown to be competitive with the $Siegel - Tukey$ when there is a difference in location and have substantially greater power than that of B_W and B_S in some cases when the assumption of equal means/medians holds (Blair and Thomson: Table 6, 7, and 8 with $k = 0$).

3.3 Concluding Remarks

Anytime that the Smirnov statistics are computed, the statistic G is automatically available. The only requirement for its implementation is a table of its critical values. This study has shown that G is more sensitive to scale shifts than is D. This agrees with Capon (1965) who developed a lower bound for the asymptotic relative efficiency of the *Smirnov* test with respect to the likelihood ratio test. The lower bound was much lower for scale alternatives than for location alternatives for both the normal and Cauchy distributions. By considering G as well as D, the practitioner if often able to gain additional insight with no extra labor. Of course, it is always possible to apply a location test such as Mann-Whitney-Wilcoxon test and then a scale test. However, the Freund-Ansari-Bradley test, the Siegel-Tukey test, and the Koltz normal scores test all require different ranking procedures from those required by the Mann-Whitney-Wilcoxon test. In addition, the linear rank statistics cannot be adequately graphed. The Smirnov statistics can be, and the graph itself combined with the scale and location concepts developed in this work may be used to gain a deeper understanding of possible departures from the null hypothesis.

Table 3.2: Comparison of the Type I error and power properties of the G and D tests, equal sample sizes and σ ranging from 0.5 to 3

G-alpha =0.052	D-alpha = 0.052									
10 & 10	0.5									
	G	D	G	D	G	D	G	D	G	D
Normal	0.2440	0.0945	0.0510	0.0575	0.1120	0.0650	0.2455	0.1055	0.4310	0.1730
Uniform	0.3730	0.1355	0.0585	0.0510	0.1840	0.0710	0.3860	0.1425	0.5715	0.2405
G-alpha =0.080	D-alpha = 0.081									
20 & 20	0.5									
	G	D	G	D	G	D	G	D	G	D
Normal	0.5005	0.2490	0.0730	0.0960	0.2500	0.1490	0.5070	0.2480	0.7895	0.5275
Uniform	0.7330	0.4195	0.0845	0.0720	0.4445	0.1765	0.7420	0.3915	0.9085	0.7645
G-alpha =0.070	D-alpha = 0.071									
30 & 30	0.5									
	G	D	G	D	G	D	G	D	G	D
Normal	0.6440	0.3280	0.0675	0.0840	0.3030	0.1375	0.6570	0.3080	0.9040	0.7255
Uniform	0.8890	0.5880	0.0685	0.0695	0.5990	0.2155	0.8950	0.5685	0.9845	0.9550
G-alpha =0.054	D-alpha = 0.05									
40 & 40	0.5									
	G	D	G	D	G	D	G	D	G	D
Normal	0.7295	0.2320	0.0525	0.0305	0.3500	0.0765	0.7320	0.2105	0.9605	0.6660
Uniform	0.9490	0.5390	0.0475	0.0260	0.6860	0.1265	0.9495	0.5135	0.9970	0.9560

Table 3.3: Comparison of the Type I error and power properties of the G and D tests, unequal sample sizes and σ ranging from 0.5 to 3

G-alpha =0.058	D-alpha = 0.054									
7 & 13	0.5		1.0		1.5		2		3	
	G	D	G	D	G	D	G	D	G	D
Normal	0.2075	0.0610	0.0645	0.0475	0.1430	0.0950	0.2665	0.1575	0.4515	0.3040
Uniform	0.3300	0.0900	0.0460	0.0455	0.2310	0.1375	0.4025	0.2295	0.5920	0.4365

G-alpha =0.058	D-alpha = 0.054									
13 & 7	0.5		1.0		1.5		2		3	
	G	D	G	D	G	D	G	D	G	D
Normal	0.2595	0.1815	0.0565	0.0485	0.0910	0.0530	0.2290	0.0725	0.4285	0.1150
Uniform	0.4040	0.2660	0.0630	0.0530	0.1585	0.0525	0.3350	0.0765	0.5880	0.1615

G-alpha =0.063	D-alpha = 0.062									
15 & 25	0.5		1.0		1.5		2		3	
	G	D	G	D	G	D	G	D	G	D
Normal	0.4885	0.1510	0.0675	0.0695	0.1910	0.1090	0.4135	0.2175	0.7175	0.4140
Uniform	0.7840	0.2280	0.0585	0.0625	0.3870	0.1485	0.6745	0.3015	0.3015	0.5485

G-alpha =0.063	D-alpha = 0.062									
25 & 15	0.5		1.0		1.5		2		3	
	G	D	G	D	G	D	G	D	G	D
Normal	0.4025	0.2055	0.0665	0.0655	0.2215	0.0940	0.4820	0.1530	0.7910	0.3280
Uniform	0.6765	0.2855	0.0595	0.0600	0.4195	0.0970	0.7715	0.2275	0.9410	0.5450

Bibliography

[1] Ansari, A. R. and Bradley, R. A.(1960) ,"Rank Sum Tests for Dispersions," *Ann. Math. Statist.,* 31, 1174-1189.

[2] Becker R. A., Chambers, J. M., and Wilks, A. R. (1988). *TheNewS* Language Pacific Grove CA: Wadsworth & Brooks/Cole, Advanced Books & Software.

[3] Blair, R. Clifford and Thompson, G. L. (1992). "A distribution-free rank-like test for scale with unequal population locations," *Commun. Statist. -Simul. Comp.,* 21, 353-371.

[4] Capon, J.(1965). "On the asymptotic efficiency of the Kolmogorov-Smirnov test," *J. Amer. Statist. Ass.,* 60, 843-853.

[5] Chambers, J. M. Cheveland, W. S. Kleiner, B. S., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Boston: Duxbury Press, division of Wadsworth, Inc.

[6] Conover, W. J. (1980). *Practical Nonparametric Statistics.* 2nd Ed. New York: John Wiley and Sons.

[7] Freund, J. E. and Ansari, A. R. (1957). Two-way rank sum test for variances. Technical Report ♯34. Virginia Polytechnic Insititute, Blackburg, Virginia.

[8] Gideon, R. and Mueller, D. E. (1978). "Computation of the two-sample Smirnov statistics," *The Amer. Statist.,* 32, 136-137.

[9] Gideon, R. and Rothan, A. M. (1991). "The Kolmogorov-Smirnov statistics in variation testing," *Commun. Statist.-Theor. Meth..,* 20, 73-86.

[10] JD Gibbons,S Chakraborti (2003), *Nonparametric Statistical Inference,Fourth Edition,* Marcel Dekker Inc., New York.

[11] Kim, P. J. and Jennrich, R. I. (1973). "Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov Criterion, $D_{m,n} : m \leq n$," *Selected Tables in Mathematical Statistics* 1(H. L. Harter and D. B. Owen, Eds.). Providence: American Mathematical Society, 79-170.

[12] Klotz, J. H. (1962). "Nonparametric tets for scale," *Ann. Math. Statist.,* 33, 498-512.

[13] Moses, L. E. (1963). "Rank-tests for Dispersions," *Ann. Math. Statist.,* 34, 973-983.

[14] Mueller, D. E. (1978). A geometric view of the Kolmogorov-Smirnov statistics with multi-sample generalizations. Unpublished doctoral dissertation. University of Montana, Missoula, Montana.

[15] Rothan, A. M. (1982). A distribution-free scale test of the Kolmogorov-Smirnov type. Unpublished doctoral dissertation.

[16] Siegel, S. and Tukey, J. W. (1960). "A nonparametric sum of ranks procedure for relative spread in unpaired samples," *J. Amer. Statist. Ass.,* 55, 429-445. Errate,

Ibid., 56, (1961), 1005.

[17] Smirnov, N. V. (1939). "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bull. Math. Univ.* Moscow, Serie Int., 2, 3-14.