Iranian Journal of Science & Technology, Transaction A, Vol. 32, No. A1
Printed in The Islamic Republic of Iran, 2008

"Research Note"

ON THE DISTRIBUTION OF Z-SCORES*

J. BEHBOODIAN[1**] AND A. ASGHARZADEH[2]

[1]Department of Mathematics, Islamic Azad University, Shiraz, I. R. of Iran
[2]Department of Statistics, Faculty of Basic Sciences, Mazandaran University, Babolsar, I. R. of Iran
Emails: behboodian@stat.susc.ac.ir, a.asgharzadeh@umz.ac.ir

Abstract – Let $X_1,...,X_n$ be a random sample from a distribution with sample mean $\overline{X}$ and sample variance S^2. In this paper we consider certain very general properties of the so-called "Z-scores" $(X_i - \overline{X})/S : i = 1,...,n$. A representation theorem is then given for Z-scores obtained from an underlying normal population, together with a theorem for their limiting distribution as the sample size tends to infinity. Finally, two applications involving grading and testing for an outlier are presented.

Keywords – Finite exchangeability, grading, outlier test, quadratic forms, Thompson's identity, Samuelson's inequality, Slutsky's theorem

1. INTRODUCTION

Let $X_1,...,X_n$ be independent and identically distributed (iid) random variables from a distribution with distribution function $F(x)$, finite mean μ, and finite variance $\sigma^2 > 0$. Then the standardized random variables $(X_i - \mu)/\sigma$, $i = 1,2,...,n$, with mean zero and variance one, are also iid. If the unknown parameters μ and σ are replaced by their estimators $\overline{X} = \sum_{i=1}^{n} X_i/n$ and $S = \sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)}$, that is, the sample mean and the sample standard deviation, the random variables $Z_i = (X_i - \overline{X})/S$, $i = 1,2,...,n$, are obtained. These random variables, which do not depend on the units of measurement, are usually referred to as Z-scores by behavioral scientists [1, 2] and are often used in education and psychology. They are no longer independent, but as we will show in this paper, they are finitely exchangeable with $\mathrm{E}(Z_i) = 0$ and $\mathrm{var}(Z_i) = (n-1)/n$. When $F(x)$ is a normal distribution, Z_i is called a normal Z-score.

However a literature survey, through journals, monographs, books, and the Internet, shows that there is not much theoretical work on the Z-scores as far as we know.

In an old article, Thompson [3] applied a normal Z-score for the rejection of outliers and discussed its distribution.

*Received by the editor October 28, 2006 and in final revised form June 22, 2008
**Corresponding author

Samuelson [4] and Oklin [5] proved that Z-scores are bounded. Lehmann and Casella [6], and also Shao [7], gave the density of a normal Z-score without any discussion. In this article, we obtain some new results about Z-scores which may not be found in the existing statistical literature.

In Section 2, we study the properties of Z-scores for random samples from a general distribution. Section 3 presents the exact distribution function of a normal Z-score and the asymptotic distribution of a general Z-score. Section 4 gives the moments and the kurtosis of a Z-score. Finally, in Section 5, two applications of a normal Z-score are given with numerical examples.

2. GENERAL BASIC PROPERTIES OF Z-SCORES

To study the properties of Z-scores in general, we use the *equal in distribution* technique which follows from the following definition.

Definition 1. Two random vectors $V = (V_1, V_2, ..., V_n)$ and $W = (W_1, W_2, ..., W_n)$ are said to be *equal in distribution*, denoted by $V \overset{D}{=} W$, if they have the same distribution.
For any two random vectors *V* and *W* that are equal in distribution, it is known that $g(V) \overset{D}{=} g(W)$ for any measurable function $g(t)$ from R^n to R^k [8].

Definition 2. The random variables $Y_1, Y_2, ..., Y_n$ are said to be *finitely exchangeable* if for every permutation $i_1, i_2, ..., i_n$ of the integers 1,2,...,*n*, $(Y_1, Y_2, ..., Y_n) \overset{D}{=} (Y_{i_1}, Y_{i_2}, ..., Y_{i_n})$.
Now, if the Y_i's are iid then they are exchangeable, but the converse is not true. Using $g(t_1, t_2, ..., t_n) = t_1$ we conclude that $Y_1 \overset{D}{=} Y_{i_1}$, i.e. exchangeable random variables are identically distributed. Similarly, $(Y_1, Y_2) \overset{D}{=} (Y_{i_1}, Y_{i_2})$, $(Y_1, Y_2, Y_3) \overset{D}{=} (Y_{i_1}, Y_{i_2}, Y_{i_3})$, and so on.

In what follows we give some of the basic properties of Z-scores assuming that $n \geq 2$:

(**A**) $Z_1, Z_2, ..., Z_n$ are finitely exchangeable.

To prove this property we use $(X_1, X_2, ..., X_n) \overset{D}{=} (X_{i_1}, X_{i_2}, ..., X_{i_n})$ and the function $g(t_1, t_2, ..., t_n) = (\frac{t_1 - \bar{t}}{s_t}, \frac{t_2 - \bar{t}}{s_t}, ..., \frac{t_n - \bar{t}}{s_t})$, with $\bar{t} = \sum_{i=1}^{n} t_i / n$ and $s_t^2 = \sum_{i=1}^{n} (t_i - \bar{t})^2 / (n-1)$. Because of exchangeability, $Z_1, Z_2, ..., Z_n$ are identically distributed. Thus, in what follows, we consider $Z_1 = (X_1 - \overline{X})/S$ as being a representative of all the Z-scores.

(**B**) Z_1 is ancillary when μ and σ^2 are unknown.

The proof of this property follows on noting that the distribution of the vector $\big((X_1 - \mu)/\sigma, (X_2 - \mu)/\sigma, ..., (X_n - \mu)/\sigma\big)$ does not depend on the parameters μ, σ^2, and Z_1 is a function of this vector. This can be shown by expanding Z_1 in terms of the components of the vector as follows:

$$Z_1 = \frac{X_1 - \overline{X}}{S} = \frac{[(X_1 - \mu) - (\overline{X} - \mu)]/\sigma}{S/\sigma}$$

Therefore the distribution of Z_1 does not depend on μ, σ^2, i.e. Z_1 is ancillary.

(**C**) Z_1 is bounded.

To prove this properly, we delete X_1 from the sample and consider

$$\overline{X}_* = \frac{1}{n-1} \sum_{i=2}^{n} X_i \quad , \quad S_*^2 = \frac{1}{n-2} \sum_{i=2}^{n} (X_i - \overline{X}_*)^2. \tag{1}$$

Now, it is easy to show that

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=2}^{n}(X_i-\bar{X}_*)^2+\frac{n}{n-1}(X_1-\bar{X})^2. \tag{2}$$

This identity is called *Thompson's identity* and can be found in [3, 5]. From (2), we have

$$(n-1)S^2=(n-2)S_*^2+\frac{n}{n-1}(X_1-\bar{X})^2, \tag{3}$$

$$|Z_1|=\left|(X_1-\bar{X})/S\right|\le(n-1)/\sqrt{n}. \tag{4}$$

Inequality (4) is often referred to as *Samuelson's inequality* [4, 5].

(**D**) $\mathrm{E}(Z_1)=0$, $\mathrm{var}(Z_1)=(n-1)/n$.

To prove these two results, we use the facts that $Z_1,Z_2,...,Z_n$ are identically distributed, bounded, and the relations

$$\sum_{i=1}^{n}(X_1-\bar{X})/S=\sum_{i=1}^{n}Z_i=0 \quad \text{and} \quad \sum_{i=1}^{n}\left((X_1-\bar{X})/S\right)^2=\sum_{i=1}^{n}Z_i^2=n-1.$$

If we use $S_b^2=(n-1)S^2/n$ (a biased estimator of σ^2) instead of S^2, we have $\mathrm{var}\left((X_1-\bar{X})/S_b\right)=1$. Therefore, $(X_1-\bar{X})/S_b$ with mean zero and standard deviation this can be converted to $T_1=A(X_1-\bar{X})/S_b+B$ with mean B and standard deviation $A>0$.

(**E**) $\mathrm{cov}(Z_1,Z_2)=-1/n$.

Using $(Z_1,Z_2)\overset{D}{=}(Z_i,Z_j)$ for $1\le i<j\le n$, $\mathrm{var}(Z_1)=(n-1)/n$, and $\mathrm{var}\left(\sum_{i=1}^{n}Z_i\right)=0$, we obtain $\mathrm{cov}(Z_1,Z_2)=-1/n$ and the coefficient of correlation $\rho(Z_1,Z_2)=-1/(n-1)$. We conclude that although the Z_i's are identically distributed, they are clearly not independent.

(**F**) The distribution of Z_1 is symmetric about zero if the distribution of the X_i is symmetric about μ. This follows from the fact that $(X_1-\mu,X_2-\mu,...,X_n-\mu)\overset{D}{=}(\mu-X_1,\mu-X_2,...,\mu-X_n)$. Now, using the function $g(t_1,t_2,...,t_n)=(t_1-\bar{t})/s_t$, we have $(X_1-\bar{X})/S\overset{D}{=}(\bar{X}-X_1)/S$ or $Z_1\overset{D}{=}-Z_1$.

3. TWO IMPORTANT THEOREMS FOR THE NORMAL CASE

Theorem 1. (*A representation theorem*) Let $X_1,...,X_n$ be a random sample of size $n\ge3$ from the $N(\mu,\sigma^2)$ distribution and $Z_1=(X_1-\bar{X})/S$ be a Z-score with distribution function $F_{Z_1}(z)$. Then

$$\textbf{(a)}\quad Z_1^2\overset{D}{=}\frac{(n-1)^2}{n}\frac{1}{1+(n-2)F_{(n-2,1)}}, \tag{5}$$

where $F_{(m_1,m_2)}$ is a random variable from the *F*-distribution with m_1 and m_2 degrees of freedom.

$$\textbf{(b)}\quad F_{Z_1}(z)=\frac{1}{2}+\frac{\delta(z)}{2}P\left(F_{(1,n-2)}\le\frac{n(n-2)z^2}{(n-1)^2-nz^2}\right), \tag{6}$$

where $0\le|z|\le\dfrac{n-1}{\sqrt{n}}$ and $\delta(z)=\begin{cases}1 & z>0\\ 0 & z=0\\ -1 & z<0\end{cases}$.

Proofs:

(a) Dividing both sides of (3) by σ^2, we obtain

$$\frac{(n-1)S^2}{\sigma^2} = \frac{(n-2)S_*^2}{\sigma^2} + \frac{n}{n-1}\frac{(X_1-\bar{X})^2}{\sigma^2},$$

or, more succinctly, $Q = Q_1 + Q_2$. If $\chi^2_{(k)}$ denotes a random variable from the chi-square distribution with k degrees of freedom, we have

$$Q = \frac{(n-1)S^2}{\sigma^2} \overset{D}{=} \chi^2_{(n-1)}\ ,\ \ Q_1 = \frac{n}{n-1}\frac{(X_1-\bar{X})^2}{\sigma^2} \overset{D}{=} \chi^2_{(1)}\ ,\ Q_2 = \frac{(n-2)S_*^2}{\sigma^2} \overset{D}{=} \chi^2_{(n-2)},$$

which are three quadratic forms in *n* iid normal random variables.

By definition the rank of a quadratic form $x'Mx$, with a vector $x = (x_1, x_2, ..., x_n)'$ and an $n \times n$ symmetric matrix M, is the rank of M. It is easy to show that $\text{rank}(Q) = \text{rank}(Q_1) + \text{rank}(Q_2)$. Therefore, we conclude that Q_1 and Q_2 are independent [9]. Thus, we have

$$Z_1^2 = \frac{(X_1-\bar{X})^2}{S^2} = \frac{(n-1)(X_1-\bar{X})^2}{(n-1)S^2} = \frac{Q_1(n-1)^2/n}{Q_1+Q_2} = \frac{(n-1)^2}{n}\frac{1}{1+(n-2)R},$$

where $R = \left[Q_2/(n-2)\right]/(Q_1/1) \overset{D}{=} F_{(n-2,1)}$. Now, using the function $g(t) = \frac{(n-1)^2}{n}\frac{1}{1+(n-2)t}$ on both sides of $R \overset{D}{=} F_{(n-2,1)}$, we obtain (5).

(b) It is easy to show that for any continuous random variable Z with a distribution that is symmetric about zero, the distribution of Z can be written in the following form:

$$F_Z(z) = \frac{1}{2} + \frac{\delta(z)}{2}P(Z^2 \le z^2). \tag{7}$$

From property F, we know that for a normal case the distribution of Z_1 is symmetric about zero. Now, using (5) and (7), we obtain (6).

Corollary 1. From the proof of (a) we have $Q_1/\left[Q_2/(n-2)\right] \overset{D}{=} F_{(1,n-2)} \overset{D}{=} T^2_{(n-2)}$, and as a result,

$$\left|\frac{X_1-\bar{X}}{S_*}\right| \overset{D}{=} \sqrt{\frac{n-1}{n}}\left|T_{(n-2)}\right|, \tag{8}$$

where $T_{(n-2)}$ denotes a random variable from the *t*-distribution with $n-2$ degrees of freedom.

Theorem 2. Let $X_1, ..., X_n$ be a random sample of size *n* from a distribution with finite mean μ and finite variance $\sigma^2 > 0$. Also, let $\bar{X}_n$ and S_n^2 denote $\bar{X}$ and S^2 for such a sample of size n. Then, as $n \to \infty$:

(**a**) The limiting distribution of $Z_1 = \frac{X_1-\bar{X}_n}{S_n}$ is the same as the distribution of the standard variable $\frac{X_1-\mu}{\sigma}$.

(**b**) In the normal case, the density of Z_1 converges to the standard normal density.

Proofs:

(a) $Z_1 = \frac{X_1-\bar{X}_n}{S_n} = \frac{(X_1-\mu)/\sigma}{S_n/\sigma} - \frac{\bar{X}_n-\mu}{S_n}$. Now, by the law of large numbers, $\bar{X}_n - \mu \overset{P}{\to} 0$, and by Slutsky's theorem, $S_n \to \sigma$. As a result, $Z_1 \overset{D}{\to} \frac{X_1-\mu}{\sigma}$ [10].

(b) Differentiating (6) and using the density of $F_{(1,n-2)}$, for $n > 4$ the density of Z_1 is given by

$$g_n(z) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} \frac{\sqrt{n}}{n-1}\left[1-\frac{nz^2}{(n-1)^2}\right]^{\frac{n-4}{2}} \quad , \quad 0 \le |z| \le \frac{n-1}{\sqrt{n}}. \tag{9}$$

Now, as $\lim_{n\to\infty}\left[1-\frac{nz^2}{(n-1)^2}\right]^{\frac{n-4}{2}} = e^{-z^2/2}$ and, using Stirling's formula, $\lim_{n\to\infty}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}\frac{\sqrt{n}}{n-1} = \frac{1}{\sqrt{2}}$, we obtain $\lim_{n\to\infty} g_n(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$.

Remark 1. Density (9) can be derived using the fact that the ancillary $(X_1 - \bar{X})/S$ is independent of the complete sufficient statistic $(\bar{X}, S^2)$ and, as a result, from S [6]. But this approach is involved and does not illustrate the distributional structure of Z_1.

Remark 2. It is observed that part (a) of Theorem 2 implies $Z_1 \xrightarrow{D} Z \sim N(0,1)$ in the normal case. The same result can also be obtained from part (a) of Theorem 1 using the distribution of $T_{(n-2)}$, and the facts that $F_{(n-2,1)} \overset{D}{=} 1/T^2_{(n-2)}$ and $T_{(n-2)} \xrightarrow{D} Z \sim N(0,1)$. But statement (b) of Theorem 2 concerning the $\lim_{n\to\infty} g_n(z)$ is stronger than the statement $Z_1 \xrightarrow{D} Z \sim N(0,1)$, according to Scheffe's useful Convergence Theorem [10].

4. MOMENTS AND KURTOSIS OF NORMAL Z-SCORES

Differentialing the density $g_n(z)$, $n > 4$, with respect to z, we observe that it has one maximum point at zero. Therefore (9) is unimodal, symmetric, and bell-shaped.

Theoretically it is interesting to notice that (a) for $n = 2$, Z_1 is discrete with $P(Z_1 = \pm\sqrt{2}) = \frac{1}{2}$ (b) for $n = 3$, Z_1 is continuous with a minimum point at zero and (c) for $n = 4$, and Z_1 has a uniform density on the interval $(-\frac{3}{2}, \frac{3}{2})$.

Figure 1 graphs (9) for $n = 3, 4, 5, 8, 15$, together with the standard normal density. It can be observed from this plot that (9) is reasonably well approximated by the standard normal distribution for a sample size of $n = 15$.

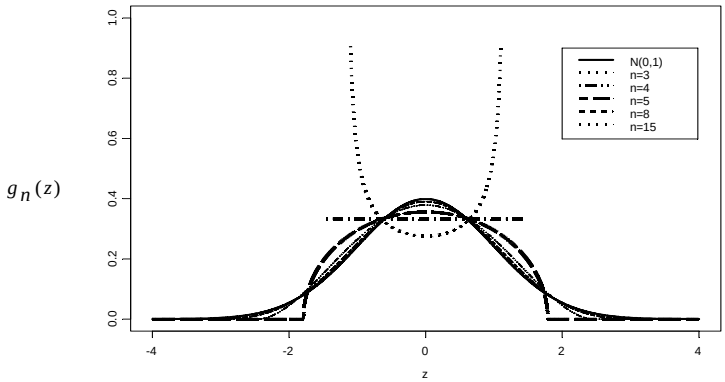

Fig. 1. Density of the normal Z-score for $n = 3, 4, 5, 8, 15$ compared with the normal density

All of the moments of Z_1 exist, since Z_1 is bounded (Property C). All the odd moments of a normal Z-score are zero because its distribution is symmetric about zero. To find the even moments, we use the moments of a beta variable B and the relation between the beta and F-distributions. Casella and Berger, [11]. From (5) we have

$$Z_1^2 = \frac{(X_1 - \overline{X})^2}{S^2} \stackrel{D}{=} \frac{(n-1)^2}{n} \frac{F_{(1,n-2)}}{F_{(1,n-2)} + (n-2)} \stackrel{D}{=} \frac{(n-1)^2}{n} B, \tag{10}$$

where B has a beta distribution with $\alpha = 1/2$, $\beta = (n-2)/2$. Using the moments of B, we obtain

$$E(Z_1^{2k}) = \frac{(n-1)^{2k}}{n^k} \frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(k+\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}, \quad k = 1,2,.... \tag{11}$$

This result can also be deduced from Basu's theorem [6] since Z_1 is ancillary and $(\overline{X}, S^2)$ is a complete sufficient statistic for (μ, σ^2)

From (11) we find $E(Z_1^2) = (n-1)/n$ and $E(Z_1^4) = 3(n-1)^3/n^2(n+1)$. Therefore, the coefficient of kurtosis for Z_1 is

$$\frac{E(Z_1^4)}{\left[E(Z_1^2)\right]^2} = \frac{3(n-1)}{n+1}. \tag{12}$$

We observe that, as $n \to \infty$, $(12) \to 3$, the coefficient of kurtosis for the standard normal density. Note that the convergence to this limit is fairly swift as the coefficient of kurtosis takes the values 1, 3/2 and 2 for $n = 2, 3$ and 5, respectively.

5. TWO APPLICATIONS OF Z-SCORES

I. Grading

Z-scores are often employed in grading and testing. Users usually assume that $(X_1 - \overline{X})/S$ has a standard normal distribution when the underlying population is normal. However, as we have shown in Section 3, the distribution of a Z-score is only approximately normal when the sample size is large. To illustrate this point, we consider the following example.

Example 1. A teacher gave a test to a class of 20 students and found that the class average and standard deviation had been 16 and 4, respectively. Assuming normality, what is the probability that a student in the class obtained a mark of less than 15 out of 20.

Using the distribution function (6) and the fact that Z_1 and $(\overline{X}, S^2)$ are independent, we have

$$\begin{aligned} P(X_1 < 15 \mid \overline{X} = 16 \text{ and } S = 4) &= P\left(\frac{X_1 - \overline{X}}{S} < \frac{15 - \overline{X}}{S} \mid \overline{X} = 16 \text{ and } S = 4\right) \\ &= P\left(\frac{X_1 - \overline{X}}{S} < \frac{15-16}{4}\right) = P\left(Z_1 < -\frac{1}{4}\right) \\ &= \frac{1}{2} - \frac{1}{2} P(F_{(1,18)} < 0.0625) \\ &= 1 - P(T_{(18)} < 0.25) = 0.4027. \end{aligned}$$

The equivalent probability obtained, assuming (erroneously) that Z_1 is normally distributed, is 0.4043. Given our previous comments regarding the similarity of (9) to the standard normal density for a sample size of $n = 20$, the closeness of the two probabilities is, perhaps, not surprising.

II. Outlier detection

During data analysis it is generally advisable to examine data for outliers, i.e. extreme observations. For example, in practice, a Z-score greater than 3 is often considered as providing evidence that the observation in the question is an outlier.

More formally, in a random sample $X_1,...,X_n$ from the $N(\mu,\sigma^2)$ distribution, X_1 can be classified as an outlier at the $0 < \alpha < 1$ level of significance if $P\left(\left|\frac{X_1-\bar{X}}{S}\right| \geq k_\alpha\right) = P\left(|Z_1| \geq k_\alpha\right) \leq \alpha.$ In order to identify k_α we can use part (a) of Theorem 1 or, for large *n*, a normal approximation.

Alternatively, for small sample sizes we can use the statistic $R = \left|\frac{X_1-\bar{X}}{S^*}\right|$ as a means of investigating whether X_1 is an outlier or not. Clearly, if X_1 is an outlier then $|X_1-\bar{X}|$ will be large and, by deleting X_1 from the sample, S^* will be smaller than *S*. Consequently, *R* should be large. Denoting the observed value of *R* by r, the probability of obtaining such an extreme value of *R* is given by

$$P\left(\sqrt{1-\frac{1}{n}}\left|T_{(n-2)}\right| > r\right). \tag{13}$$

Example 2. The scores of a vocabulary test, out of 4, were:

3.4, 2.2, 3.4, 2.5, 3.3, 1.6, 3.1, 3.8, 3.5, 2.6, 2.9, 3.9, 3.7, 3.3, 3.0, 3.4.

A chi-square goodness-of-fit test showed that the normal distribution, with $\hat{\mu} = \bar{x} = 3.1$, fits the data reasonably well.

Given that the value 1.6 is rather small in comparison with the other results, we investigate whether it might be considered to be an outlier. Deleting 1.6 from the scores, we obtain $s_* = 0.4870$ and $r = |(x_1-\bar{x})/s_*| = 3.0803.$ Now, from (13), we obtain

$$P\left(\sqrt{1-\frac{1}{16}}\left|T_{(14)}\right| > 3.0803\right) < 0.01.$$

Therefore, there is strong evidence that the score 1.6 is indeed an outlier.

Acknowledgments- The authors would like to thank the editor and the final referee for his very constructive suggestions.

REFERENCES

1. Lockhart, R. S. (1997). *Introduction to Statistics and Data Analysis for the Behavioral Sciences.* New York, Freeman and Company.
2. Abdi, H. (2007). *Z-scores*. California: Encyclopedia of Measurement and Statistics, Edited by Neil Salkind, Sage Thousand Oaks.
3. Thompson, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation, *The Annals of Mathematical Statistics*. *6*, 215-219.

4. Samuelson, P. A. (1968). How deviant can you be? *Journal of the American Statistical Association, 63*, 1522-1525.
5. Olkin, I. (1992). A matrix formulation on how deviant an observation can be. *The American Statistician*. *46*, 205-209.
6. Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation.* New York, Springer, 2nd Ed.
7. Shao, J. (1998). *Mathematical Statistics*. New York, John Wiley.
8. Randles, R. H. & Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York, John Wiley.
9. Roussas, G. G. (1997). *A Course in Mathematical Statistics*. San Diago, California: Academic Press, 2nd Ed.
10. Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Landon, Chapman and Hall.
11. Casella, G. & Berger, R. L. (1990). *Statistical Inference*. Pacific Grove, California: Wadsworth & Brooks.