

The Central Limit Theorem,
Edgeworth Expansions
and an
Introduction to Asymptotic Theory

Derek Sonderegger
Department of Mathematical Sciences
Montana State University
August 3, 2004

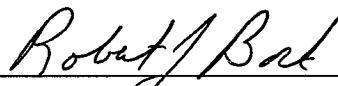
A writing project submitted in partial fulfillment
of the requirements for the degree
Master of Sciences in Statistics

APPROVAL
of a writing project submitted by

Derek Sonderegger

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

8/13/04
Date


Robert Boik
Writing Project Director

1 Introduction

Asymptotic results have long been important in statistical theory. Before computing power was inexpensive, asymptotic theory provided approximations for computationally tedious tasks. For example if $X \sim \text{Bin}(n, p)$ then for large values of n , X can be well approximated by the normal distribution $N(np, np(1-p))$. Reference books contained hundreds of approximations so that statisticians could perform their work without excessive tedious calculation. Fortunately cheap computing has removed the need for many of those approximations, but other limiting results are so fundamental to the study of statistics that introductory asymptotic theory still should be covered at the graduate level.

The Central Limit Theorem is most often presented in terms of estimating the mean of a distribution. Suppose the random variables X_i are independently and identically distributed with mean μ and variance σ^2 . Then the Central Limit tells us that the sample mean $\bar{X} = \sum_1^n X_i/n$ has a distribution that is *approximately* $N(\mu, \sigma^2/n)$. It is relatively straightforward to prove the CLT using moment generating function, but it also is interesting to examine how quickly the distribution converges to normality. Furthermore, it is possible to make adjustments to increase the accuracy of the approximation. This is of particular importance in small sample situations. Edgeworth expansions are one method of using information about higher order moments to increase accuracy. Edgeworth expansions were introduced by Edgeworth (1905). Introductions to the Edgeworth expansion can be found in Wallace (1958), Chambers (1967), Hall (1992), Chi (2001), and Boik (2004).

2 Taylor Expansions

Recall from calculus that if the convergence criteria are met, then the Taylor expansion of a function f about the point x_0 is

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)(x - x_0)^2}{2!} + \dots,$$

or written more compactly

$$f(x) = \sum_{i=0}^{\infty} f^{(i)}(x_0) \frac{(x - x_0)^i}{i!}.$$

One Taylor series that is often used in the study of Characteristic Functions is the series representation of $f(x) = e^{itx}$, where $i = \sqrt{-1}$. Since

the derivative of e^{itx} with respect to x is ite^{itx} then the Taylor series, when expanded around $x = 0$ is

$$e^{itx} = 1 + itx + \frac{(itx)^2}{2!} + \frac{(itx)^3}{3!} + \dots$$

In numerical methods, it is very common to only use the first several terms in the series as an approximation to the actual function. It then becomes important to keep track of the order of magnitude of the terms that are discarded. Consider the Taylor Series expansion of $f(x) = \sin(x)$ around $x = 0$:

$$\sin(x) = x - \frac{(x)^3}{3!} + \frac{(x)^5}{5!} - \dots + (-1)^n \frac{(x)^{2n+1}}{(2n+1)!}.$$

Since this is an alternating series it is easy to show that keeping the first n components of the sum and discarding the remaining will result in an error term that is no greater than the first term that is dropped. As $x \rightarrow 0$ the exponents get very small and the factorials get very large. Both of the above Taylor series are valid, but the expansion of \sin appears to converge much more rapidly. It is useful to be able to compare how rapidly the two series converge. The appropriate tool is $O()$ notation, which is pronounced Big O.

3 Big O Notation

Bishop, Fienberg and Holland (1975) make the following definition:

Definition 1 If $\{a_n\}$ and $\{b_n\}$ are two sequences of real numbers then $a_n = O(b_n)$ if $|a_n/b_n|$ is bounded for large n .

The idea behind $O()$ notion is to compare the relative size of $\{a_n\}$ to $\{b_n\}$, typically with $\{a_n\}$ being the sequence of interest and with $\{b_n\}$ being the comparison sequence. For example, if two matrices are of size $N \times N$ then adding the matrices requires N^2 operations and matrix addition is said to be a $O(N^2)$ operation. Matrix multiplication requires N^3 operations and is a $O(N^3)$ operation.

In the \sin function example, suppose $s_2(x)$ is a function composed of the sum of the first two elements of the Taylor series of $\sin(x)$. Then the error term is

$$e(x) = \sin(x) - s_2(x) = \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

As $x \rightarrow 0$, $e(x)/x^5$ is bounded so it is said that the error for the s_2 approximation is $O(x^5)$ as $x \rightarrow 0$. Adding more terms to the approximation

makes the error term go to zero even faster, and in practice only 4 or 5 terms are necessary for good precision around $x = 0$.

There are several conventions to be aware of when using this notation.

1. Constants are ignored because the researcher is primarily interested in the rate at which the series increases or decreases with respect to certain well known sequences such as $\{n \ln n\}$, $\{n\}$, $\{\sqrt{n}\}$, $\{1\}$, $\{n^{-1/2}\}$.
2. Terms of similar magnitude can be combined because constants are ignored. That is, $O(n) + O(n) = 2O(n) = O(n)$.
3. Terms of smaller magnitude can be combined with larger terms. For example, $O(n) + O(n^{1/2}) = O(n)$.

The last two rules are only valid if the number of terms being summed does not depend on n . The issue to avoid is having n terms of order $O(1/n)$ and claiming the sum is $O(1/n)$ when it should be $O(1)$.

4 Big O_p Notion

In order to use the $O()$ notation in a stochastic environment, Bishop, Fienberg and Holland (1975) makes the following modification to take into account probability. First define $O_p(1)$ as follows:

Definition 2 Let X_n be a stochastic sequence. $X_n = O_p(1)$ if for every $\eta > 0$, there exists $K(\eta)$ and $n(\eta)$ such that if $n > n(\eta)$ then

$$P\{|X_n| \leq K(\eta)\} \geq 1 - \eta.$$

Simply the definition means that for any η , a K and n can be chosen such that almost certainly X_m is less than K when $m > n$. This definition means that with arbitrary precision, X_n is bounded. Sometimes X_n is said to be “bounded in probability.”

For example suppose $X_i \sim \text{Gamma}(5, .2)$ then $X_i = O_p(1)$ because any individual observation is bounded in probability, but $W_n = \sum_{i=1}^n X_i \neq O_p(1)$ because the sequence W_n keeps growing with every additional element.

Definition 3 Suppose that X_n/b_n is $O_p(1)$. Then X_n is $O_p(b_n)$.

Of particular interest is the sequence $\epsilon_n = \bar{X}_n - \mu$, where \bar{X}_n is the sample mean using n observations from a population with mean μ and standard deviation σ . Since the standard deviation of \bar{X}_n is σ/\sqrt{n} , as $n \rightarrow \infty$ the

terms of the series get getting smaller and smaller. In order to use the definitions for O_p , first notice that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

by the Central Limit Theorem. The important thing to notice is that this quantity is bounded in probability and therefore

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = O_p(1).$$

It follows that

$$\bar{X}_n - \mu = O_p(\sigma/\sqrt{n}) = O_p(n^{-1/2}),$$

because the constant σ doesn't affect the order of magnitude.

5 Characteristic Functions

In most introductory probability courses, students are introduced to the moment generation function (MGF) as a device to uniquely identify a particular distribution. One problem with MGFs is that they don't always exist. The solution is to use the Characteristic Function, which always exists. The characteristic function is the complex extension of the MGF and is defined as $\phi_x(t) = E(e^{itX})$, where $i^2 = -1$.

The Characteristic Function shares many properties with the MGF. The Characteristic Function can be used to uniquely identify distributions, and can be used to show limiting results. If a sequence of functions $\phi_{X_n}(t) \rightarrow \phi_X(t)$ then the distribution of X_n is said to converge to the distribution of X .

5.1 Cumulants

Taking derivatives of the moment generating function and evaluating at $x = 0$ is one method of generating the moments of a distribution. Taking derivatives of the log moment generating function and evaluating at $x = 0$ generates a sequence of numbers called *cumulants*. Cumulants are of interest because there is a simple relationship between the moments of a distribution and its cumulants and that the *ith* cumulant of a sum of random variables is the sum of the *ith* cumulants. Clearly this is useful when describing the distribution of sums of random variables.

In general, suppose that X is a random variable with characteristic function

$$\phi_X(t) = E(e^{itx}) = \int e^{itx} f(x) dx.$$

Provided that the MGF exists then the expansion

$$\phi_X(t) = \exp \left\{ \sum_{n=1}^{\infty} \kappa_n (it)^n / n! \right\}$$

exists. Taking the natural log of ϕ_X yields

$$K(t) = \sum_{n=0}^{\infty} \frac{\kappa_n (it)^n}{n!},$$

where κ_j is the j th cumulant and $K(t)$ is the cumulant generation function. It is easy to show that

$$\kappa_j = \frac{K^{(j)}(0)}{i^j}.$$

Example. Consider $X \sim N(0, 1)$. Then

$$\begin{aligned} \phi(t) &= E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = \int_{-\infty}^{\infty} \frac{e^{itx - x^2/2 + t^2/2 - t^2/2}}{\sqrt{2\pi}} dx \\ &= e^{-t^2/2} \int_{-\infty}^{\infty} \frac{e^{t^2/2 + itx - x^2/2}}{\sqrt{2\pi}} dx = e^{-t^2/2} \int_{-\infty}^{\infty} \frac{e^{(t+xi)^2/2}}{\sqrt{2\pi}} dx \\ &= e^{-t^2/2} \int_{-\infty}^{\infty} \frac{e^{-(-x+ti)^2/2}}{\sqrt{2\pi}} dx = e^{-t^2/2}. \end{aligned}$$

Now the cumulant generating function $K(t)$ is $K(t) = \log(\phi(t)) = -t^2/2$.

The derivatives evaluated at $t = 0$ yields $\kappa_1 = 0$ and $\kappa_2 = 1$. All the rest of the derivatives are zero so $\kappa_i = 0$ for $i \geq 3$.

5.2 Properties of Cumulants

To show the important property that the cumulants for a sum of random variables is the sum of the cumulants, consider $S = X + Y$, where X and Y are independently distributed random variables. Then $\phi_S(t) = \phi_X(t)\phi_Y(t)$ therefore $K_S(t) = K_X(t) + K_Y(t)$.

It also is important to consider the cumulants of the random variable aX where a is a constant. The cumulants $\kappa_j(aX)$ are merely $a^j \kappa_j(X)$. This is shown by

$$\phi_{aX}(t) = \phi_X(at) \Rightarrow K_{aX}(t) = K_X(at) \Rightarrow \kappa_j(aX) = a^j \kappa_j(X).$$

With these two rules, it is possible to examine the cumulants of the sum and mean of n iid variables. Let $S = \sum_{i=1}^n X_i$ and $\bar{X} = S/n$. Then the cumulants are $\kappa_i(S) = \sum \kappa_i(X) = n\kappa_i(X)$ and

$$\kappa_i(\bar{X}) = \frac{\kappa_i(X)}{n^{i-1}}.$$

In particular notice that the mean $\kappa_1(\bar{X}) = \kappa_1(X)$ and variance $\kappa_2(\bar{X}) = \kappa_2(X)/n$ as is expected. Also it is useful to notice the higher order cumulants of the mean are $O_p(n^{-(i-1)})$.

6 Cumulants and Moments

Consider the following series expansion of $\phi_x(t)$:

$$\begin{aligned} \phi_x(t) &= \mathbb{E}(e^{itx}) = \int e^{itx} f(x) dx \\ &= \int \left\{ 1 + itx + \frac{(itx)^2}{2!} + \frac{(itx)^3}{3!} + \dots \right\} f(x) dx \\ &= \int \left\{ f(x) + itxf(x) + \frac{(it)^2 x^2 f(x)}{2!} + \frac{(it)^3 x^3 f(x)}{3!} + \dots \right\} dx \\ &= 1 + it\mathbb{E}(X) + \frac{(it)^2 \mathbb{E}(X^2)}{2!} + \frac{(it)^3 \mathbb{E}(X^3)}{3!} + \dots \end{aligned}$$

Since the cumulant expansion of $\phi_X(t)$ also involves terms of $(it)^j/j!$, relating coefficients yields the following relationships between cumulants and the central moments of the distribution:

$$\begin{aligned} \kappa_1 &= \mathbb{E}(X) \\ \kappa_2 &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ \kappa_3 &= \mathbb{E}(X^3) - 3\mathbb{E}(X^2)\mathbb{E}(X) + 2(\mathbb{E}X)^3 = \mathbb{E}(X - \mathbb{E}X)^3 \\ \kappa_4 &= \mathbb{E}(X^4) - 4\mathbb{E}(X^3)\mathbb{E}(X) - 3(\mathbb{E}(X^2))^2 + 12\mathbb{E}(X^2)(\mathbb{E}X)^2 - 6(\mathbb{E}X)^4 \\ &= \mathbb{E}(X - \mathbb{E}X)^4 - 3(\text{Var}(X))^2. \end{aligned}$$

In this manner κ_j can be thought of as a polynomial of degree j of the moments. The relationship can be inverted and the moments can be written as a polynomial of degree j of the cumulants.

7 Miscellaneous Tricks

7.1 Inversion Theorem

Suppose that Y is a scalar random variable with characteristic function $\phi_Y(t)$ and cdf $F_Y(y)$ that is continuous and differentiable, then

$$f_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \phi_Y(t) dt.$$

For proof, see Billingsly (2003, section 26). The inversion theorem is useful because it is a convenient way to write a pdf in terms of the characteristic function. Since the characteristic function is defined in terms of the pdf, it is now possible to switch back and forth between the two with only minimal effort.

7.2 Hermite Polynomials

Let $\varphi(x)$ be the standard normal pdf and $\phi(x)$ be the standard normal characteristic function. Define the r^{th} Hermite polynomial, $H_r(x)$ as

$$H_r(x) = \frac{(-1)^r d^{(r)}\phi(x)}{\psi(x) (dx)^r}, \text{ for } r = 1, 2, \dots$$

While there is a general formula for an arbitrary Hermite polynomial (see Pace and Salvan (1997, section 10.2). it will suffice to only deal with the first several polynomials which are

$$H_1(x) = x, \quad H_2(x) = x^2 - 1, \quad H_3(x) = x^3 - 3x, \quad H_4(x) = x^4 - 6x^2 + 3,$$

$$H_5(x) = x^5 - 10x^3 + 15x, \text{ and } H_6(x) = x^6 - 15x^4 + 45x^2 - 15.$$

8 Edgeworth Expansions

It is now possible examine how quickly

$$Z = \frac{n^{1/2}(\bar{X} - \mu)}{\sigma}$$

converges in distribution to $N(0, 1)$, where \bar{X} is the sample mean of n observations and μ and σ^2 are the mean and variance of the distribution in question.

Let $\phi_Z(t)$ be the characteristic function for Z , and $\phi(t)$ be the characteristic function of the distribution being sampled from. First notice that Z has characteristic function

$$\phi_Z(t) = \left\{ \phi \left(\frac{t}{\sqrt{n}\sigma} \right) \right\}^n \exp \left\{ \frac{-\sqrt{n}it\mu}{\sigma} \right\}.$$

And the cumulant generating function is

$$\begin{aligned} K_Z(t) &= n \ln \left\{ \phi \left(\frac{t}{\sqrt{n}\sigma} \right) \right\} - \frac{\sqrt{n}it\mu}{\sigma} = n \sum_{j=2}^{\infty} \left(\frac{it}{\sigma\sqrt{n}} \right)^j \frac{\kappa_j(X)}{j!} \\ &= \frac{t^2}{2} + \frac{(it)^3 \kappa_3(X)}{6\sqrt{n}\sigma^3} + \frac{(it)^4 \kappa_4(X)}{24n\sigma^4} + O(n^{-3/2}). \end{aligned}$$

Using the inversion formula yields

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \phi_Z(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \exp\{K_Z(t)\} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \exp \left\{ \frac{t^2}{2} + \frac{(it)^3 \kappa_3(X)}{6\sqrt{n}\sigma^3} + \frac{(it)^4 \kappa_4(X)}{24n\sigma^4} + O(n^{-3/2}) \right\} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} e^{-t^2/2} \left[1 + \frac{(it)^3 \rho_3}{6\sqrt{n}} + \frac{(it)^4 \rho_4}{24n} + \frac{(it)^6 \rho_3^2}{72n} + O(n^{-3/2}) \right] dt, \end{aligned}$$

where $\rho_j = \kappa_j(X)/\sigma^j$ is the standardized cumulant. This is a sum of terms of the form

$$(-1)^r \int_{-\infty}^{\infty} (-1)^r e^{-itz} e^{-t^2/2} (it)^r c,$$

where c is a constant that does not depend on t . Notice that

$$(-1)^r e^{-itz} (it)^r = \frac{d^r}{(dz)^r} e^{itz}.$$

If it is legitimate to exchange the order of integration and differentiation, then the terms can be written as

$$c(-1)^r \frac{d^r}{(dz)^r} \int_{-\infty}^{\infty} e^{-itz} e^{-t^2/2} dt.$$

Using the inversion theorem

$$\int_{-\infty}^{\infty} e^{-itz} e^{-t^2/2} dt = \varphi(z).$$

Taking derivatives yields terms of the form:

$$c(-1)^r \frac{d^r}{(dz)^r} \varphi(z) = cH_r(z).$$

Writing the first four terms of the sum and consolidating the remaining terms into an error term, yields

$$f_Z(z) = \varphi(z) \left[1 + \frac{\rho_3}{6\sqrt{n}} H_3(z) + \frac{\rho_4}{24n} H_4(z) + \frac{\rho_3^2}{72n} H_6(z) + O(n^{-3/2}) \right].$$

Now transforming Z back to \bar{X} yields

$$f_{\bar{X}}(\bar{x}) = \frac{\sqrt{n}}{\sigma} \varphi(z) \left[1 + \frac{\rho_3}{6\sqrt{n}} H_3(z) + \frac{\rho_4}{24n} H_4(z) + \frac{\rho_3^2}{72n} H_6(z) + O(n^{-3/2}) \right]$$

where $z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$ and $\varphi(z)$ is the standard normal pdf.

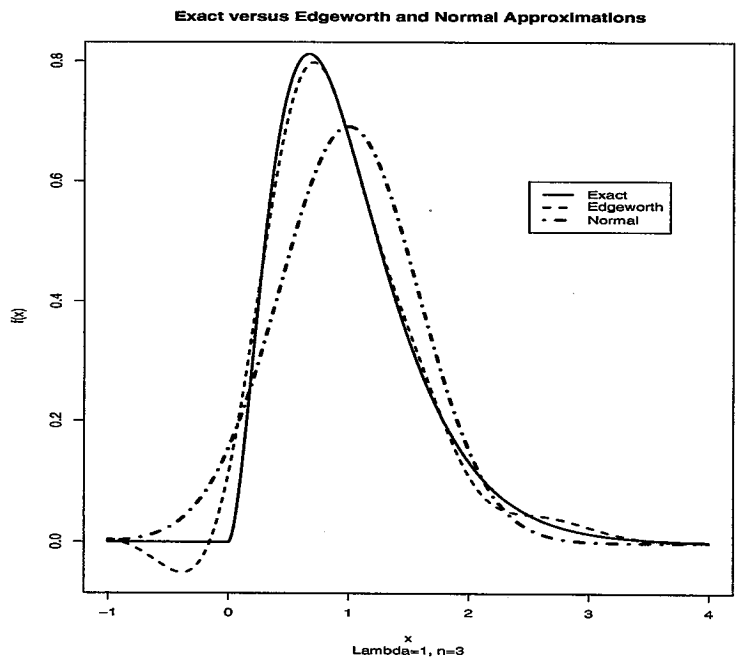
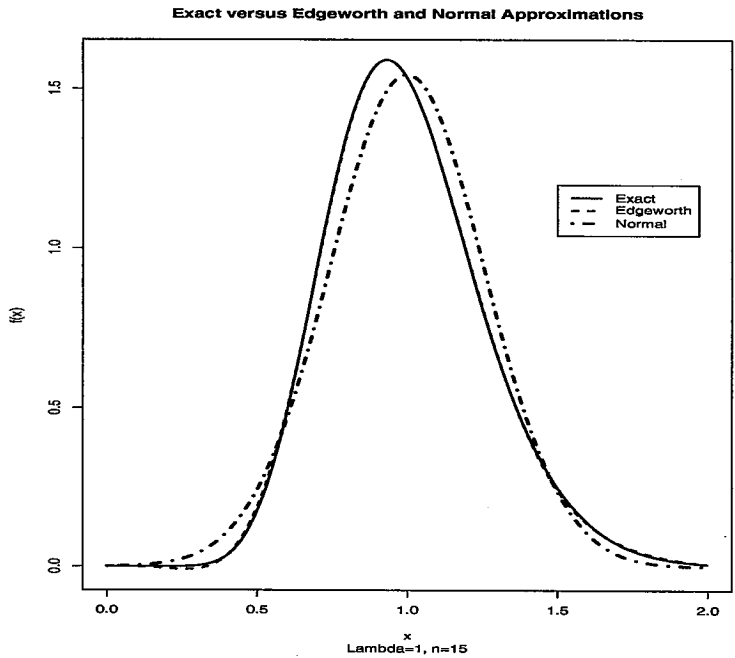
Example Suppose that a random sample of n observations is distributed $X_i \sim \text{Exp}(\lambda)$. The cumulants of the Exponential distribution are $\kappa_j = \lambda^j (j - 1)!$. Then examining the MGF, it is apparent that $\sum X_i \sim \text{Gamma}(n, \lambda)$ and that $\bar{X} \sim \text{Gamma}(n, \lambda/n)$.

Therefore, in the Edgeworth expansion for \bar{X} ,

$$\rho_j = \frac{\kappa_j(X)}{\sigma^j} = \frac{\lambda^j (j - 1)!}{\lambda^j} = (j - 1)!$$

$$f_{\bar{X}}(\bar{x}) = \frac{\sqrt{n}}{\lambda} \varphi(z) \left[1 + \frac{2!}{6\sqrt{n}} (z^3 - 3z) + \frac{3!}{24n} (z^4 - 6z^2 + 3) + \frac{(2!)^2}{72n} (z^6 - 15z^4 + 45z^2 - 15) + O(n^{-3/2}) \right]$$

Given $\lambda = 1$, below is two graphs comparing the actual distribution of the mean and the normal and Edgeworth expansions. In the first graph $n = 15$, and the second graph, $n = 3$.



9 Discussion

In the first graph, the Edgeworth approximation is almost indistinguishable from the true distribution, having accounted for the skewness and kurtosis

of the true distribution, while the normal approximation does not fit as well. However the second graph displays the negative aspect of the Edgeworth expansion. Since the Edgeworth uses a polynomial multiplied by a standard normal, it is possible to get bimodal shapes, and even negative probability densities. These problems can be address using Saddlepoint approximations and are discussed in Pace and Salvan (1997).

References

- [1] Billingsley, P. (1986). *Probability and Measure*, Second Edition, New York: John Wiley & Sons.
- [2] Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis*, Cambridge: MIT Press.
- [3] Boik, R.J. (2004). Lecture Notes: Statistics 550, Spring 2004. Montana State University.
- [4] Chi, Z. (2001). Course Notes for Statistics 30400, Distribution Theory. University of Chicago.
- [5] Edgeworth, F. Y. (1905). The Law of Error. *Cambridge Philos. Trans.* **20**, 36-66 and 113-141.
- [6] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York: Springer.
- [7] Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective*, Singapore: World Scientific.
- [8] Wallace, D.L. (1958). Asymptotic approximations to distributions. *The Annals of Mathematical Statistics*, **29**, 635-654.