Two central limit theorems and their application to the estimation of both parameters in the binomial distribution

by Willis John Alberda

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of DOCTOR OP PHILOSOPHY in Mathematics

Montana State University

Abstract:

While investigating experimentally the dynamics of a host-pathogen system, the statistical problem of estimating and finding confidence intervals for the probability of mutation from avirulence to virulence of the pathogen was encountered. Assuming the probability of mutation is constant during any experimental procedure, this problem may be considered as that of estimating the parameter p in the Binomial Distribution function B(z; n,p). Because of the nature of the pathogen, however, to obtain this estimator it was also necessary to estimate the total number of observations.

In the binomial setting this is equivalent to also estimating and finding confidence intervals for the parameter n. With this experimental situation as background, two central limit theorems which give a solution to these problems are proved. These theorems are proved under various assumptions which appear feasible in the experimental situation. Other than the sampling without replacement scheme, the techniques are distribution free in so far as no specific underlying distribution function is assumed. These theorems, then, are an approach to estimating and finding confidence intervals for both parameters in the Binomial Distribution.

A solution to the problem of finding the rate of convergence in each case is also given. This solution is obtained using the same assumptions used in establishing the central limit theorems.

The techniques suggested by these two theorems and the rates of convergence for each case are then applied to the experimental data obtained.

TWO CENTRAL LIMIT THEOREMS AND THEIR APPLICATION TO THE ESTIMATION OF BOTH PARAMETERS IN THE BINOMIAL DISTRIBUTION

by

WILLIS JOHN ALBERDA

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

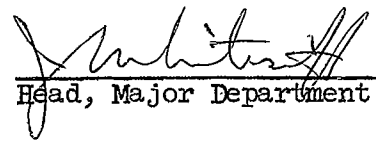
of

DOCTOR OF PHILOSOPHY

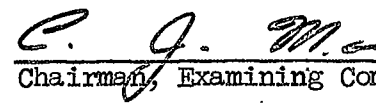
in

Mathematics

Approved:



Head, Major Department



Chairman, Examining Committee

Dean, Graduate Division

MONTANA STATE COLLEGE
Bozeman, Montana

August, 1964

VITA

The author, Willis John Alberda, was born in Bozeman, Montana, February 7, 1936, to Mr. and Mrs. Peter W. Alberda of Manhattan, Montana. He received his secondary education at the Manhattan Christian High School, Manhattan, Montana, for two years, and graduated from Western Christian High School, Hull, Iowa. In 1959 he received a Bachelor of Arts degree in education with a mathematics major from Calvin College, Grand Rapids, Michigan. In 1963 he received a Master of Science degree in Mathematics from Montana State College, Bozeman, Montana.

ACKNOWLEDGEMENT

The research reported in this thesis was supported by the United States Atomic Energy Commission, Division of Biology and Medicine Project AT(45-1) - 1729. The author is very grateful for the financial assistance received from this agency during the completion of this work and for the funds made available for the experimental investigations from which the data used in this paper was obtained.

The author wishes to thank Dr. C. J. Mode for his encouragement and the excellent advice he offered so willingly during the time this work was being completed. The author is also grateful to Mr. Gordon Chang who performed the experiments described in this paper and who contributed the descriptions of the experimental proceedures and the data he obtained. Thanks, too, are due Dr. C. P. Quesenberry for his advice in setting up the study program preceding this work, also for the corrections and excellent suggestions he and the other committee members made to improve the final copy.

A word of acknowledgment is also due Miss Marilin McElwee for her diligence in typing the final copy and the author's wife for her patience and work in typing the first draft.

TABLE OF CONTENTS

ABSTRACT

While investigating experimentally the dynamics of a host-pathogen system, the statistical problem of estimating and finding confidence intervals for the probability of mutation from avirulence to virulence of the pathogen was encountered. Assuming the probability of mutation is constant during any experimental procedure, this problem may be considered as that of estimating the parameter p in the Binomial Distribution function $B(z; n,p)$. Because of the nature of the pathogen, however, to obtain this estimator it was also necessary to estimate the total number of observations. In the binomial setting this is equivalent to also estimating and finding confidence intervals for the parameter n. With this experimental situation as background, two central limit theorems which give a solution to these problems are proved. These theorems are proved under various assumptions which appear feasible in the experimental situation. Other than the sampling without replacement scheme, the techniques are distribution free in so far as no specific underlying distribution function is assumed. These theorems, then, are an approach to estimating and finding confidence intervals for both parameters in the Binomial Distribution.

A solution to the problem of finding the rate of convergence in each case is also given. This solution is obtained using the same assumptions used in establishing the central limit theorems.

The techniques suggested by these two theorems and the rates of convergence for each case are then applied to the experimental data obtained.

CHAPTER I

INTRODUCTION

Experimental Setting

The statistical problems considered in this dissertation arose in an experimental investigation of the dynamics of a biological host-pathogen system. The solutions to these problems, not necessarily the only possible solutions, were obtained under conditions which appeared most feasible in this experimental situation.

In this experimental investigation, the barley mildew fungus was treated with a mutagenic agent, inoculated on eight resistant varieties of barley, and then screened for mutants. One of the main objectives of this experimental investigation was the estimation of the probability of mutation from avirulence to virulence at eight loci of this fungus.

The principal statistical problem arising from this work was that of finding an estimator and confidence intervals for the probability of mutation at each locus. The difficulties in finding this estimator and confidence intervals, however, arise from the fact that under ordinary conditions the fungus does not grow on the resistant varieties of barley. Moreover, when a pustule appears on a resistant variety it is considered a mutant. The estimator of the probability of mutation at each locus of this fungus, quite naturally, will be a ratio of the total number of mutants observed on some specific resistant variety of barley to the total number of spores germinated on this variety of barley. Therefore, since it is impossible to count the total number of spores germinated on a set of resistant varieties, it is necessary to estimate this total number of

spores germinated for each resistant variety. The statistical problems are then to find estimators and confidence intervals for the probability of mutation at each locus and for the total number of spores germinated on each resistant variety of barley.

The experimental procedure for estimating the total number of spores germinated on a set of resistant varieties may be described as follows. Whenever a set of pots of resistant varieties of barley was inoculated with mildew spores, a set of pots of seedlings of a susceptible variety was also inoculated. Let X_i represent the total number of spores germinated on leaf i of a resistant variety of barley. $X_1, X_2, \ldots, X_{\mathbb{N}}$ is then the population of the number of spores germinated on $\mathbb{N}$ leaves of the resistant variety of barley, where $\mathbb{N}$ is a known positive integer. The spores germinated on each resistant variety of barley could be represented in this way. As soon as pustules on the susceptible variety were visible, a count of the number of pustules on n leaves was made. Let x_i represent the number of pustules counted on leaf i of the susceptible variety. The sample $x_1, x_2, \ldots, x_n$ is the number of pustules which appeared and were counted on n leaves of the susceptible variety of barley, where n is also known and $n < \mathbb{N}$. The numbers $x_1, x_2, \ldots, x_n$ will be considered a random sample of size $n < \mathbb{N}$ taken without replacement from a population identical to the finite population $X_1, X_2, \ldots, X_{\mathbb{N}}$ of size $\mathbb{N}$. If $\bar{x}$ is the average number of pustules per leaf on the susceptible leaves, then $\mathbb{N}\bar{x}$ was used as an estimate of the total number of spores germinated on the resistant variety of barley or the total number of observations.

This procedure could be repeated indefinitely, say up to q stages, so that if $N_k\bar{x}_k$ represents the estimate of the total number of observations at the K-th stage, then

$$\hat{S}_q = \sum_{k=1}^{q} N_k \bar{x}_k$$

provided an estimate of the total number of observations over q stages.

If z_q represents the total number of mutant pustules observed over q stages, then

$$\hat{\pi}_q = \frac{z_q}{\hat{S}_q}$$

provided an estimate of the probability of mutation.

The theoretical problem consists of finding confidence intervals for the total number of observations, S_q, and the probability of mutation, π.

Having considered the experimental situation in this setting, however, before proceeding to the discussion of the theoretical problems it is appropriate at this point to mention several results from sampling without replacement theory as found in Wilks (1962). These results and relations will serve to motivate certain concepts and symbols used in the discussion of the theoretical problems.

Preliminary Results From Sampling Theory

For k = 1, 2, 3, ..., where k designates the k-th stage in this experimental procedure, let $X_{k1}, X_{k2}, \ldots, X_{kN_k}$ be a finite population of size N_k with mean μ_k and variance σ_k^2, where

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{ki},$$

and

$$\sigma_k^2 = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (X_{ki} - \mu_k)^2 .$$

Let $x_{k1}, x_{k2}, \cdots, x_{kn_k}$ be a sample of size n_k taken without replacement from this population with mean $\bar{x}_k$ and variance $\hat{\sigma}_k^2$, where

$$\bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki},$$

and

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2$$

With these definitions, it is shown in the reference of Wilks cited above that

$$E(\bar{x}_k) = \mu_k,$$

$$\mathrm{Var}(\bar{x}_k) = \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \sigma_k^2,$$

and

$$E(\hat{\sigma}_k^2) = \sigma_k^2.$$

From these results it follows immediately that

$$E(N_k \bar{x}_k) = S_{N_k},$$

where

$$S_{N_k} = \sum_{i=1}^{N_k} X_{ki} = N_k \mu_k,$$

and $$\text{Var}\,(N_k\bar{x}_k) = N_k^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right)\sigma_k^2 .$$

Also it is obvious that $E(\hat{S}_q) = S_q$, where

$$\hat{S}_q = \sum_{k=1}^{q} N_k\bar{x}_k,$$

and

$$S_q = \sum_{k=1}^{q} N_k\mu_k,$$

and that $E(\hat{s}_q^2) = s_q^2$, where

$$\hat{s}_q^2 = \sum_{k=1}^{q} N_k^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right)\hat{\sigma}_k^2 ,$$

and

$$s_q^2 = \sum_{k=1}^{q} N_k^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right)\sigma_k^2 .$$

Furthermore, if the assumption of independence between stages is imposed, then the following theorem of some importance in the sequel is also true.

Theorem 1.1: If s_q^2, $\hat{S}_q$, and S_q are defined as before and the $N_k\bar{x}_k$ form a sequence of independent random variables, then

$$\text{Var}(\hat{S}_q) = E(\hat{S}_q - S_q)^2 = s_q^2 .$$

Proof: Since $E(\hat{S}_q) = S_q$,

$$\text{Var}(\hat{S}_q) = E(\hat{S}_q - S_q)^2 .$$

$$E(\hat{S}_q - S_q)^2 = E\left[\sum_{k=1}^{q} N_k(\bar{x}_k - \mu_k)\right]^2,$$

$$= \sum_{k=1}^{q} N_k^2 E(\bar{x}_k - \mu_k)^2, \text{ by the assumption of independence,}$$

$$= \sum_{k=1}^{q} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right)\sigma_k^2,$$

$$= s_q^2 .$$

With these preliminary results, the problems as stated in the first section can now be described more precisely in mathematical terms.

Theoretical Setting

In order to find confidence intervals for the total number of observations, S_q, and for the probability of mutation, π, it is necessary to determine the distribution functions of the random variables used to estimate these parameters. In many cases, as for example in this situation, the exact distribution functions may be difficult to determine and not amenable to tabulation. When these situations arise, a solution can be obtained by determining a limiting distribution which is amenable to tabulation. Approximate confidence intervals can then be constructed from this limiting distribution. This is the approach used to obtain a solution to the problems presented here.

Under fairly general assumptions it can be shown that the limiting

distribution of the random variable

$$\frac{\hat{S}_q - S_q}{s_q}$$

is Normal with mean zero and variance one. Since in many cases it is possible to substitute the estimator of the variance and still obtain the same limiting distribution, the problem of finding confidence intervals for S_q, the total number of observations, appears quite logically to be that of determining the limiting distribution of the random variable

$$\frac{\hat{S}_q - S_q}{\hat{s}_q} .$$

Since s_q is assumed unknown, some estimator must be substituted. The choice of $\hat{s}_q$ seems natural from the unbiasedness of $\hat{s}_q^2$ and will indeed plan an important role in the solution of the problem.

Furthermore, under fairly general assumptions the limiting distribution of the random variable

$$\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}$$

is Normal with mean zero and variance one. Since $\hat{S}_q$ is an unbiased estimator of S_q, the natural extension in this case is, if possible, to determine the limiting distribution of the random variable

$$\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} .$$

From this it is possible to construct approximate confidence intervals for π, the probability of mutation.

Once the limiting distribution of these two random variables have been established and are amenable to tabulation, it is possible for a fixed q to construct approximate confidence intervals for S_q, the total number of observations, and π, the probability of mutation, respectively. The final problem then is to find a rate of convergence to the limiting distribution in each case.

Before proceeding to the solution of these problems it is necessary to make some assumptions under which it is possible to solve the problems stated above and to exhibit the existence of a probability space on which the random variables with these imposed conditions are defined. The existence of this probability space is necessary in order to give meaning to phrases or symbols such as almost sure convergence ($\xrightarrow{a.s.}$) of sequences of random variables, convergence in probability ($\xrightarrow{P}$) of sequences of random variables, convergence in distribution $\left(\mathcal{L}(X_n) \to \mathcal{L}(X)\right)$, and events $\left(\left[X_n \leq x\right]\right)$ as they appear in the sequel. (The symbol $\left[X_n \leq x\right]$ means the set of ω such that the random variable satisfies the condition stated, as for example, $X_n(\omega) \leq x$.)

Basic Assumptions and an Existence Theorem

The problems as described in the previous section are solved under the following assumptions.

1. For $q = 1, 2, 3, \ldots$, the population $(X_{q1}, X_{q2}, \ldots, X_{qN_q})$ is a finite population of distinct, positive integers with mean μ_q and

variance σ_q^2, both unknown. The distinctness implies that for $q = 1, 2, 3, \ldots$, $\sigma_q^2 \geq d_1 \geq 0$ and also the sample variance $\hat{\sigma}_q^2 \geq d_2 > 0$.

2. For $q = 1, 2, 3, \ldots$, and $i = 1, 2, \ldots, N_q$, $0 < m \leq X_{qi} \leq M < \infty$, with $M-m > 0$.

3. The N_q and n_q are known, positive, integers such that for $q = 1, 2, 3, \ldots$, $2 < N_q \leq N < \infty$ and $2 \leq n_q \leq N_q - 1$.

4. $N_q\bar{x}_q$, $q = 1, 2, 3, \ldots$, is a sequence of independent not necessarily identically distributed random variables.

5. For $q = 1, 2, 3, \ldots$, $z_{N_q} = \sum_{k=1}^{S_{N_q}} y_k$, the total number of mutants at stage q, and $N_q\bar{x}_q$, the estimator of the total number of observations at stage q, are independent random variables. S_{N_q} as defined before is the total number of observations at stage q. The random variables y_k defined by

$$y_k = 1 \quad \text{if mutation occurs,}$$

$$= 0 \quad \text{otherwise,}$$

with $P\left[y_k=1\right] = \pi$, for all k, are independent.

The following remarks and theorem exhibit the existence of a probability space on which the random variables satisfying these assumptions are defined.

For each q, the probability function of $N_q\bar{x}_q$ and the range, I_{1q}, of this random variable are determined by the sampling without replacement scheme. For $q = 1, 2, 3, \ldots$, set

$$P\left[N_q \bar{x}_q = i_{1q}\right] = p_{1q}(i_{1q}),$$

where i_{1q} is some element of I_{1q}, some finite set of real numbers. As stated, $p_{1q}(i_{1q})$ for all $i_{1q} \varepsilon I_{1q}$ is determined by the sampling without replacement probability function, but for use in the sequel need not be calculated. Similarly, for each q, by the definition of the z_{N_q} and the y_k in assumption 5, the probability function of z_{N_q} is given by

$$P\left[z_{N_q} = i_{2q}\right] = \binom{S_{N_q}}{i_{2q}} \pi^{i_{2q}} (1-\pi)^{S_{N_q} - i_{2q}} = p_{2q}(i_{2q}),$$

where i_{2q} is an element of I_{2q}, the range space of z_{N_q}, which is some finite set of integers 0, 1, 2, ..., S_{N_q}.

If $X_q = (N_q \bar{x}_q, z_{N_q})$, then the range space of the vector random variable X_q is some product space $I_q = I_{1q} X I_{2q}$, and by assumption 5, the probability function of the vector random variable X_q is

$$P\left[X_q = i_q\right] = p_{1q}(i_{1q}) p_{2q}(i_{2q}) = p_q(i_q),$$

where $i_q = (i_{1q}, i_{2q})$ is an element of I_q. With these remarks it is now possible to prove the following.

Theorem 1.2: (Existence Theorem) There exists a probability space $(\Omega, \mathcal{A}, P)$ with the sequence of independent vector random variables X_q defined on it and such that for each q the two elements of X_q are also independent.

Proof: In order to prove this theorem, it is sufficient to characterize

the space Ω, determine the σ-field $\mathcal{A}$, and define a set function P on the σ-field $\mathcal{A}$. Toward this end then, let Ω be the set of all sequences of the form $(i_1, i_2, i_3, \ldots)$ where $i_k \in I_k$. Define a cylinder $C(i_1, i_2, \ldots, i_N)$ for some fixed N to be the set of all sequences in Ω such that the fixed elements $i_1, i_2, \ldots, i_N$ appear in the first N places. Varying N over all positive integers and i_k over all elements of I_k, yields a class $\mathcal{C}$ of all such cylinders $C(i_1, i_2, \ldots, i_N)$. Then $\mathcal{A}$ is taken to be the minimal σ-field over the class $\mathcal{C}$. Now define the set function P on the class $\mathcal{C}$ as follows. For each $C(i_1, i_2, \ldots, i_N)$ assign the probability

$$PC(i_1, i_2, \ldots, i_N) = \prod_{t=1}^{N} p_t(i_t).$$

The probabilities defined in this way are consistent. Hence, by various theorems, for example Theorem A, Pg. 137 Loe've (1960), the set function on the class $\mathcal{C}$ may be extended to a probability measure on $\mathcal{A}$. Hence, it is possible to construct a probability space $(\Omega, \mathcal{A}, P)$. To define the random variables X_q on this probability space proceed as follows. If ω is the sequence (i_q), then set $X_q(\omega) = i_q$ for $q = 1, 2, 3, \ldots$. In this manner the vector random variables X_q are defined on this probability space and the assignment of the set function P makes them independent. The assignment of p_q makes the elements of the vector independent. These assignments determine all distribution functions on this space and the theorem is proved.

From this theorem then, since z_q and $\hat{S}_q$ are sums of independent

random variables defined on this space, they too are defined on this probability space. Their joint distribution function is determined by the sampling without replacement scheme and the binomial distribution of the y_k.

In the sequel all events $\left[X_n \leq x\right]$ will be elements of the σ-field $\mathcal{A}$ and a.s. convergence and convergence in probability of sequences of random variables will be convergence with respect to this probability space.

Boundedness and divergence of sequences of random variables, though not designated as such in various places in the context, is a.s. boundedness and a.s. divergence with respect to this probability space. In these cases, however, the sets of unboudedness and of convergence are empty sets.

With this introduction and these preliminary theorems it is possible to proceed with the solution of the problems presented in the previous section.

CHAPTER II

THE LIMITING DISTRIBUTION OF $(\hat{S}_q - S_q)/\hat{s}_q$

Using the assumptions made in the Introduction and the notation found there, the procedure used in finding the limiting distribution of the random variable $(\hat{S}_q - S_q)/\hat{s}_q$ is the following: first establish that

$$\mathcal{L}\left(\frac{\hat{S}_q - S_q}{s_q}\right) \to N(0,1),$$

second show that

$$P\left|\frac{\hat{S}_q - S_q}{s_q} - \frac{\hat{S}_q - S_q}{\hat{s}_q}\right| \to 0.$$

These two results are then sufficient to prove that

$$\mathcal{L}\left(\frac{\hat{S}_q - S_q}{\hat{s}_q}\right) \to N(0,1),$$

from which approximate confidence intervals for S_q can be constructed.

The first statement is shown by the following:

Theorem 2.1: If

$$Y_{qk} = \frac{N_k \bar{x}_k - S_{N_k}}{s_q},$$

then

$$\sum_{k=1}^{q} Y_{qk} = \frac{\hat{S}_q - S_q}{s_q}$$

and $$\mathcal{L}\left(\frac{\hat{S}_q - S_q}{s_q}\right) \to N(0,1).$$

Proof: Obviously $$\sum_{k=1}^{q} Y_{qk} = \frac{\hat{S}_q - S_q}{s_q}$$

by the definition of the Y_{qk}. To prove this theorem it is possible to use the "Bounded Case" Theorem, Pg. 277 Loeve (1960). It is then necessary to verify that

(i) $|Y_{qk}|$ is uniformly bounded for all k,

(ii) $s_q \to \infty$ as $q \to \infty$.

By the assumption of independence of the $N_k\bar{x}_k$, the Y_{qk} are also independent random variables, and since $E(\bar{x}_k) = \mu_k$ the Y_{qk} are centered at expectations so that the other conditions necessary to employ the "Bounded Case" Theorem are satisfied.

(i) Since $s_q^2 > 0$ and $S_{N_k} = N_k\mu_k$, $|Y_{qk}|$ is bounded for all k if $|N_k\bar{x}_k - N_k\mu_k|$ is bounded for all k. But

$$N_k|\bar{x}_k - \mu_k| \leq N_k(M-m) \text{ by assumption 2,}$$

$$\leq N(M-m) \text{ by assumption 3,}$$

hence, $|Y_{qk}|$ is uniformly bounded for all k.

(ii)
$$s_q = \sqrt{\sum_{k=1}^{q} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \sigma_k^2},$$

and since by assumption 1 $\sigma_k^2 \geq d_1 > 0$ for all k and by assumption 3 $2 \leq n_k < N_k \leq N$,

$$N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \sigma_k^2 \not\to 0$$

as $k \to \infty$. Hence,

$$\sum_{k=1}^{q} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \sigma_k^2 \to \infty$$

as $q \to \infty$ and therefore $s_q \to \infty$ as $q \to \infty$ and the theorem is proved.

The following relations are necessary to prove the next theorem and are also required in later theorems, hence, are given here as a lemma.

Lemma 2.1: If (i) s_q^2, $\hat{s}_q^2$, σ_k^2, and $\hat{\sigma}_k^2$ are defined as before,

(ii) $\sigma_k^2 \geq d_1 > 0$ and $\hat{\sigma}_k^2 \geq d_2 > 0$ (assumption 1),

(iii) $2 \leq n_k \leq N_k - 1$ (assumption 3),

then (1) $0 < d_1 \leq \sigma_k^2 \leq 2(M-m)^2$ for all k,

(2) $0 < d_2 \leq \hat{\sigma}_k^2 \leq 2(M-m)^2$ for all k,

$$(3) \qquad 0 < qd_1 \leq s_q^2 \leq 2qN^2(M-m)^2,$$

$$(4) \qquad 0 < qd_2 \leq \hat{s}_q^2 \leq 2qN^2(M-m)^2.$$

Proof: (1) By (ii) $\sigma_k^2 \geq d_1 > 0$ and

$$\sigma_k^2 = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (X_{ki} - \mu_k)^2 \leq \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (M-m)^2 \text{ by assumption 2,}$$

$$= \frac{N_k}{N_k - 1} (M-m)^2,$$

$$\leq 2(M-m)^2,$$

since by assumption 3, $2 < N_k$.

(2) By (ii) $\hat{\sigma}_k^2 \geq d_2 > 0$ and by the same reasoning as in (1),

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2 \leq \frac{n_k}{n_k - 1} (M-m)^2 \leq 2(M-m)^2,$$

since by assumption 3 also, $2 \leq n_k$.

$$(3) \quad s_q^2 = \sum_{k=1}^{q} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \sigma_k^2 \geq d_1 \sum_{k=1}^{q} N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \quad \text{by (1),}$$

$\geq qd_1$, since by (iii) and assumption 3 $N_k^2 \left(\frac{1}{n_k} - \frac{1}{N_k}\right) \geq 1$.

Also

$$s_q^2 \leq \sum_{k=1}^{q} 2N_k^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right)(M-m) \text{ by (2)},$$

$$\leq 2\,qN^2(M-m)^2 \text{ since by (iii) and assumption 3}$$
$$N_k^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right) \leq N^2.$$

(4) Replacing σ_k^2 by $\hat{\sigma}_k^2$ and d_1 by d_2 in the proof of (3), it follows immediately that

$$qd_2 \leq \hat{s}_q^2 \leq 2qN^2(M-m)^2,$$

and the lemma is proved.

These relations are useful in the next theorem which in turn gives the key to establishing the second statement in the introduction of this chapter.

Theorem 2.2: If s_q^2 and $\hat{s}_q^2$ are defined as before, then

$$\frac{s_q^2 - \hat{s}_q^2}{q} \xrightarrow{a.s.} 0.$$

Proof: Using Kolmogorov's Convergence Criterion, Loeve (1960); Pg. 238, on the random variables

$$Y_k = N_k^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right)(\sigma_k^2 - \hat{\sigma}_k^2),$$

it must be shown that

(i) $E(Y_k) = 0$ which then also proves that the Y_k are integrable,

(ii) the Y_k are independent,

(iii) $\sum_{k=1}^{\infty} \frac{\mathrm{Var}(Y_k)}{k^2} < \infty.$

(i) $E(Y_k) = 0$ since $E(\hat{\sigma}_k^2) = \sigma_k^2$.

(ii) The Y_k are independent by assumption 4.

(iii) $$\sum_{k=1}^{\infty} \frac{Var(Y_k)}{k^2} = \sum_{k=1}^{\infty} \frac{E(Y_k^2)}{k^2} = \sum_{k=1}^{\infty} \frac{N_k^4(\frac{1}{n_k} - \frac{1}{N_k})^2}{k^2} \; E(\sigma_k^2 - \hat{\sigma}_k^2)^2$$

$$= \sum_{K=1}^{\infty} \frac{N_k^4(\frac{1}{n_k} - \frac{1}{N_k})^2}{k^2} \; (E\hat{\sigma}_k^4 - \sigma_k^4),$$

$$< \sum_{k=1}^{\infty} \frac{N_k^4(\frac{1}{n_k} - \frac{1}{N_k})^2}{k^2} \; E\hat{\sigma}_k^4 ,$$

$$\leq \sum_{k=1}^{\infty} \frac{4N_k^4(M-m)^2}{k^2} \qquad \text{by Lemma 2.1,}$$

$$\leq 4N^4(M-m)^2 \sum_{k=1}^{\infty} \frac{1}{q^2} < \infty ,$$

since $\sum_{k=1}^{\infty} \frac{1}{q^2} < \infty$ and N and M-m are finite. This proves the theorem.

This result gives the key to proving the following

Theorem 2.3: If S_q, $\hat{S}_q$, s_q and $\hat{s}_q$ are defined as before, then

$$\frac{\hat{S}_q - S_q}{s_q} - \frac{\hat{S}_q - S_q}{\hat{s}_q} \xrightarrow{P} 0.$$

Proof:

By Markov's Inequality

$$P\left[\left|\hat{S}_q - S_q\right| \left|\frac{1}{s_q} - \frac{1}{\hat{s}_q}\right| \geq \epsilon\right] \leq E\left|\hat{S}_q - S_q\right| \left|\frac{1}{s_q} - \frac{1}{\hat{s}_q}\right| / \epsilon$$

for all $\epsilon \geq 0$ and by Schwarz's Inequality

$$E\left|\hat{S}_q - S_q\right| \left|\frac{1}{s_q} - \frac{1}{\hat{s}_q}\right| \leq \sqrt{E(\hat{S}_q - S_q)^2 \, E\left(\frac{1}{s_q} - \frac{1}{\hat{s}_q}\right)^2}$$

But $E(\hat{S}_q - S_q)^2 = s_q^2$ by Theorem 1.1 and

$$E\left(\frac{1}{s_q} - \frac{1}{\hat{s}_q}\right)^2 = E\left(\frac{\hat{s}_q - s_q}{s_q \hat{s}_q}\right)^2 = \frac{1}{s_q^2} E\left(\frac{\hat{s}_q - s_q}{\hat{s}_q}\right)^2 .$$

Moreover,

$$E\left(\frac{\hat{s}_q - s_q}{\hat{s}_q}\right)^2 = E\left[\frac{\left|\hat{s}_q - s_q\right| \left|\hat{s}_q - s_q\right|}{\hat{s}_q^2}\right] ,$$

$$\leq E\left[\frac{\left|\hat{s}_q - s_q\right| \left|\hat{s}_q + s_q\right|}{\hat{s}_q^2}\right] ,$$

$$= E\left|\frac{\hat{s}_q^2 - s_q^2}{s_q^2}\right| ,$$

$$\leq \frac{1}{q d_1} E\left|\hat{s}_q^2 - s_q^2\right| \text{ by Lemma 2.1.}$$

Thus

$$E(\hat{S}_q - S_q)^2 \, E\left(\frac{1}{s_q} - \frac{1}{\hat{s}_q}\right)^2 \leq \frac{1}{q\bar{d}_1} \, E|\hat{s}_q^2 - s_q^2| \quad .$$

But $|\hat{s}_q^2 - s_q^2| < \hat{s}_q^2 + s_q^2 \leq 4N^2q(M-m)^2$ by Lemma 2.1 so that

$$\left|\frac{\hat{s}_q^2 - s_q^2}{q}\right| < 4N^2(M-m)^2 < \infty \quad .$$

Hence, the random variables $\frac{\hat{s}_q^2 - s_q^2}{q}$ are integrable. Also by Theorem 2.2 this sequence of random variables converges almost surely to 0, hence, the sequence converges in probability to 0 and by the Dominated Convergence Theorem, Loe've (1960); Pg. 152,

$$E\left[\left|\frac{\hat{s}_q^2 - s_q^2}{q}\right|\right] \to 0$$

as $q \to \infty$ and the theorem is proved.

In order to prove the final theorem of this chapter, it is necessary at this point to employ a theorem of Loe've (1960); Pg. 168.

The following theorem stated in an equivalent manner as that in Loe've is given without proof for the general case.

Theorem 2.4: If

(i) $\mathcal{L}(X_n) \to \mathcal{L}(Y)$,

(ii) $X_n - Y_n \xrightarrow{P} 0$,

then

$$\mathcal{L}(Y_n) \to \mathcal{L}(Y).$$

Note: If $X_n - Y_n \xrightarrow{a.s.} 0$, the theorem also holds.

The Central Limit Theorem of interest in this chapter then follows immediately as shown by

Theorem 2.5: If S_q, $\hat{S}_q$ and $\hat{s}_q$ are defined as before, then

$$\mathcal{L}\left(\frac{\hat{S}_q - S_q}{\hat{s}_q}\right) \to N(0,1).$$

Proof: By Theorem 2.1

$$\mathcal{L}\left(\frac{\hat{S}_q - S_q}{s_q}\right) \to N(0,1),$$

and by Theorem 2.3

$$\frac{\hat{S}_q - S_q}{s_q} - \frac{\hat{S}_q - S_q}{\hat{s}_q} \xrightarrow{P} 0,$$

thus by applying Theorem 2.4 the result follows.

From this theorem then it follows that

$$\lim_q P\left[\left|\frac{\hat{S}_q - S_q}{\hat{s}_q}\right| \leq z\right] = \frac{1}{\sqrt{2\pi}} \int_{-z_\alpha}^{z_\alpha} e^{-t^2/2}\, dt \quad \text{for } z_\alpha \geq 0$$

so that for q "sufficiently" large an approximate confidence interval for S_q is given by

$$P\left[\hat{S}_q - z_\alpha \hat{s}_q \leq S_q \leq \hat{S}_q + z_\alpha \hat{s}_q\right] \cong 1 - \alpha.$$

CHAPTER III

THE LIMITING DISTRIBUTION OF $(z_q - \hat{S}_q\pi)/\sqrt{\hat{S}_q\pi(1-\pi)}$.

The Central Limit Theorem

In establishing the limiting distribution of $(z_q - \hat{S}_q\pi)/\sqrt{\hat{S}_q\pi(1-\pi)}$, it is necessary to employ methods differing from those in Chapter II. After having shown that

$$\mathcal{L}\left(\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}\right) \longrightarrow N(0,1),$$

attempts to show that

$$\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} - \frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} \xrightarrow{P} 0 \quad \text{failed.}$$

The theorem, however, is proved by showing that for all $z \in R$, R the set of real numbers,

$$\left| P\left[\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} < z\right] - P\left[\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right] \right| \to 0.$$

In order to justify the use of the pivotal quantity

$$\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}}$$

as the expression from which to work for confidence intervals for the probability of mutation $\pi > 0$, which is assumed to be constant over all stages, consider the following. If as in assumption 5 the y_i are defined

as

$$y_i = 1 \text{ if mutation occurs}$$

$$y_i = 0 \text{ otherwise}$$

where $P\left[y_i = 1\right] = \pi$, then the probability function of y_i is

$$h(y_i) = \pi^{y_i}(1-\pi)^{1-y_i} \quad .$$

If, as before, $z_q = \sum_{i=1}^{S_q} y_i$ is the total number of mutations over q stages

where the y_i are independent indicators as defined above, then the following two theorems are well known and are given without proof.

Note: S_q and the random variable z_q are non-negative integers and $z_q \leq S_q$.

Theorem 3.1: If the y_i's are as defined above and

(i) $\pi = P\left[y_i = 1\right]$,

(ii) $z_q = \sum_{i=1}^{S_q} y_i$,

then the probability function of z_q is

$$h(z_q) = \binom{S_q}{z_q} \pi^{z_q} (1-\pi)^{S_q - z_q} \quad .$$

Theorem 3.2: If the probability function of z_q is $h(z_q)$ defined above, then the maximum likelihood estimator of π is $\pi_q = z_q/S_q$ and

(i) $E(\pi_q) = \pi$,

(ii) $\mathrm{Var}(\pi_q) = \frac{\pi(1-\pi)}{S_q}$.

Now since

$$\frac{\pi_q - \pi}{\sqrt{\frac{\pi(1-\pi)}{S_q}}} - \frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}$$

it is possible to find confidence intervals for π if the limiting distribution of the random variable formed by replacing S_q by $\hat{S}_q$ can be shown to be $N(0,1)$.

Before proceeding to the theorems, the following lemma will be required in the proofs of these theorems.

Lemma 3.1: If S_q and $\hat{S}_q$ are defined as before, then

(i) $S_q \to \infty$ as $q \to \infty$,

(ii) $\hat{S}_q \to \infty$ as $q \to \infty$,

(iii) $0 < \frac{m}{M} \leq \frac{\hat{S}_q}{S_q} \leq \frac{M}{m} < \infty$.

Proof:

(i) $S_q = \sum_{k=1}^{q} S_{N_k} \geq \sum_{k=1}^{q} N_k m \geq q2m \to \infty$ as $q \to \infty$ by assumptions 2 and 3.

(ii) Similarly $\hat{S}_q = \sum_{k=1}^{q} N_k \bar{x}_k \geq \sum_{k=1}^{q} N_k m \geq q2m \to \infty$ as $q \to \infty$.

(iii) $$\frac{\hat{S}_q}{S_q} = \frac{\sum_{k=1}^{q} N_k \bar{x}_k}{\sum_{k=1}^{q} N_k \mu_k} \leq \frac{\sum_{k=1}^{q} N_k M}{\sum_{k=1}^{q} N_k m} = \frac{M}{m} < \infty$$ by assumption 2,

also,

$$\frac{\hat{S}_q}{S_q} \geq \frac{\sum_{k=1}^{q} N_k m}{\sum_{k=1}^{q} N_k M_k} = \frac{m}{M} > 0.$$

Thus,

$$0 < \frac{m}{M} \leq \frac{\hat{S}_q}{S_q} < \frac{M}{m} < \infty .$$

These results are useful in proving the following central limit theorem.

Theorem 3.3: If z_q, S_q, S_{N_k}, and z_{N_k} are defined as before and

$$Z_{qk} = \frac{z_{N_k} - S_{N_k}\pi}{\sqrt{S_q\pi(1-\pi)}} ,$$

then

$$\sum_{k=1}^{q} Z_{qk} = \frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}$$

and

$$\mathcal{L}\left(\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}\right) \to N(0,1) .$$

Proof: Obviously

$$\sum_{k=1}^{q} Z_{qk} = \frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}$$

by the definition of Z_{qk}. Also $E(Z_{N_k}) = S_{N_k}\pi$, hence, the random variables Z_{qk} are centered at expectations. Again using the "Bounded Case" Theorem,

Loeve (1960); Pg. 277, it is apparent from the definition of the z_{N_k} that the Z_{qk} are independent random variables and since $|z_{N_k} - S_{N_k}\pi| < N(M) < \infty$, the Z_{qk} are uniformly bounded. By Lemma 3.1 $S_q\pi(1-\pi) \to \infty$ as $q \to \infty$ so all conditions of the "Bounded Case" Theorem are satisfied and the theorem is proved.

Now in order to determine the limiting distribution of the random variable formed by replacing S_q by $\hat{S}_q$ it is necessary to introduce a random variable which can be utilized because of (iii) Lemma 3.1.

Definition 3.1: Let $\theta_q = \frac{\hat{S}_q/S_q}{M/m}$ where S_q, $\hat{S}_q$, M and m are defined as before.

This random variable introduced for the sake of convenience of notation gives the key to the proof of the principal central limit theorem of this chapter. From Lemma 3.1 it is obvious that $0 < (m/M)^2 \leq \theta_q \leq 1$. The following is also a true theorem.

Theorem 3.4: If θ_q is as defined above, then

(i) $E(\theta_q) = m/M$,

(ii) $\theta_q \xrightarrow{a.s.} m/M$.

Proof:

(i) $$E(\theta_q) = E\,\frac{\hat{S}_q/S_q}{M/m} = \frac{mE(\hat{S}_q)}{MS_q} = m/M.$$

(ii) By Tchebichev's Inequality, for all $\epsilon > 0$

$$P\left[|\theta_q - m/M| \geq \epsilon\right] \leq \frac{E(\theta_q - m/M)^2}{\epsilon^2}$$

But $E(\theta_q - m/M)^2 = E\theta_q^2 - (m/M)^2 = (m/M)^2 \frac{1}{S_q^2} E(\hat{S}_q^2) - (m/M)^2 = (m/M)^2 s_q^2/S_q^2$

since by Theorem 1.1 $E(\hat{S}_q^2) = s_q^2 + S_q^2$.

Thus by Lemmas 2.1 and 3.1

$$E(\theta_q - m/M)^2 \leq (m/M)^2 \frac{2N^2 (M-m)^2}{q4m^2} = \frac{K}{q} \text{ where } K = \frac{N^2(M-m)^2}{2m^2}$$

therefore $K/q \to 0$ as $q \to \infty$ and thus $\theta_q \xrightarrow{P} m/M$. Moreover since

$0 < (m/M)^2 \leq \theta_q \leq 1$,

$$|\theta_{q+r} - \theta_q| = \theta_q \left|\frac{\theta_{q+r}}{\theta_q} - 1\right| \leq \left|\frac{\theta_{q+r}}{\theta_q} - 1\right| = \left|\frac{m\hat{S}_{q+r}/MS_{q+r}}{m\hat{S}_q/MS_q} - 1\right|$$

$$= \left|\frac{S_q \hat{S}_{q+r}}{\hat{S}_q S_{q+r}} - 1\right| .$$

Now let $\hat{S}_r = \sum_{k=q+1}^{q+r} N_k \bar{x}_k$ and $S_r = \sum_{k=q+1}^{q+r} N_k \mu_k$ then

$$|\theta_{q+r} - \theta_q| \leq \left| \frac{S_q(\hat{S}_q + \hat{S}_r) - \hat{S}_q(S_q + S_r)}{\hat{S}_q(S_q + S_r)} \right| = \left|\frac{S_q \hat{S}_r - \hat{S}_q S_r}{\hat{S}_q(S_q + S_r)}\right|$$

$$\leq \left|\frac{S_q \hat{S}_r - \hat{S}_q S_r}{\hat{S}_q S_q}\right| \quad \text{since } 0 \leq S_r,$$

$$= \left|\frac{\hat{S}_r}{\hat{S}_q} - \frac{S_r}{S_q}\right| \to 0 \text{ as } q \to \infty,$$

since S_r and $\hat{S}_r$ are bounded and S_q and $\hat{S}_q$ diverge as $q \to \infty$ by Lemma 3.1. Thus the θ_q satisfy the Cauchy a.s. convergence criterion so that $\theta_q \xrightarrow{a.s.} \theta$ for some θ. However, since $\theta_q \xrightarrow{P} m/M$, there exists a subsequence

q' such that $\theta_{q'} \xrightarrow{a.s.} m/M$, but since the sequence itself also converges a.s. they must converge to the same limit m/M. Thus $\theta_q \xrightarrow{a.s.} m/M$ and the theorem is proved.

Note: Since $0 < (m/M)^2 \leq \theta_q \leq 1$, $1/\theta_q \xrightarrow{a.s.} M/m$ for

$$|\theta_q - m/M| = (m/M)\,\theta_q |M/m - 1/\theta_q| \geq (m/M)^3 |M/m - 1/\theta_q| \ .$$

With this result the central limit theorem follows from the following observation.

Theorem 3.5: For all $z \in R$, R the set of real numbers, and $\pi \geq 0$,

$$\left| P\left[\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} < z\right] - P\left[\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right] \right| \leq P\left[\theta_q \neq m/M\right] \ .$$

Proof:

$$P\left[\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} < z\right] = P\left[\frac{\pi_q - \pi}{\sqrt{\frac{\pi(1-\pi)}{S_q}}} < z\right]$$

and

$$P\left[\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right] = P\left[\frac{\hat{\pi}_q - \pi}{\sqrt{\frac{\pi(1-\pi)}{\hat{S}_q}}} < z\right]$$

where $\pi_q = z_q/S_q$ and $\hat{\pi} = z_q/\hat{S}_q$. Also, by the definition of θ_q,

$$\frac{\hat{\pi}_q - \pi}{\sqrt{\frac{\pi(1-\pi)}{\hat{S}_q}}} = \frac{\sqrt{\frac{M\theta_q}{m}}\left(\frac{m}{M\theta_q}\pi_q - \pi\right)}{\sqrt{\frac{\pi(1-\pi)}{S_q}}} \ .$$

Thus,

$$P\left[\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} < z\right] = P\left[\pi_q - \pi < z'\right],$$

and,

$$P\left[\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right] = P\left[\sqrt{\frac{M\theta_q}{m}}\left(\frac{m}{M\theta_q}\pi_q - \pi\right) < z'\right]$$

where $z' = z\sqrt{\frac{\pi(1-\pi)}{S_q}}$.

Now let $A_q = \left[\pi_q - \pi < z'\right]$,

$$B_q = \left[\sqrt{\frac{M\theta_q}{m}}\left(\frac{m}{M\theta_q}\pi_q - \pi\right) < z'\right],$$

$$C_q = \left[\theta_q = m/M\right],$$

and A_q^c, B_q^c and C_q^c be the complements of each of these events. Clearly if $\omega \in A_q^c C_q$, then $\omega \in B_q^c$, thus $A_q^c C_q \subset B_q^c$ and $B_q \subset A_q \cup C_q^c$ so that $PB_q \leq P(A_q \cup C_q^c) \leq PA_q + PC_q^c$. Thus $PB_q - PA_q \leq PC_q^c$ and $-PC_q^c \leq PA_q - PB_q$. Moreover, if $\omega \in B_q^c C_q$, then $\omega \in A_q^c$, so that $B_q^c C_q \subset A_q^c$ and $A_q \subset B_q \cup C_q^c$. Thus, $PA_q \leq P(B_q \cup C_q^c) \leq PB_q + PC_q^c$ and $PA_q - PB_q \leq PC_q^c$ which coupled with the previous result gives $|PA_q - PB_q| \leq PC_q^c$. This, when converted to the original notation, proves the theorem.

This bound on the difference of the probabilities of these two random variables for all z belonging to R together with the previous theorems of this chapter gives the following central limit theorem.

Theorem 3.6: If z_q, $\hat{S}_q$ and π, $\pi > 0$, are defined as before, then

$$\mathcal{L}\left(\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}}\right) \to N(0,1).$$

Proof: Let

$$F_q(z) = P\left[\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right],$$

$$H_q(z) = P\left[\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} < z\right],$$

and,

$$G(z) = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-t^2/2}\, dt, \text{ here } \pi = 3.1416\cdots .$$

The theorem is proved if it can be shown that for all $z \in R$ $|F_q(z) - G(z)| \to 0$. However, $|F_q(z) - G(z)| \leq |F_q(z) - H_q(z)| + |H_q(z) - G(z)|$ and by Theorem 3.3, $|H_q(z) - G(z)| \to 0$. From Theorem 3.5

$$|F_q(z) - H_q(z)| \leq P\left[\theta_q \neq m/M\right]$$

which approaches zero as $q \to \infty$ since by Theorem 3.4 $\theta_q \xrightarrow{a.s.} m/M$, and the theorem is proved.

From this theorem it follows that for $z > 0$

$$\lim_q P\left[\frac{|z_q - \hat{S}_q\pi|}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right] = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-t^2/2}\, dt$$

from which approximate confidence intervals for π, the probability of mutation, can be constructed. In the next section a discussion is given concerning the construction of these confidence intervals and some of their properties.

Properties of the Approximate Confidence Intervals for π.

It is of interest to investigate the properties of the confidence intervals for π derived from the previous theorem. In order to do this the following concept is of some importance.

Definition 3.2: An estimator θ_n is said to be an almost sure (a.s.) estimator of θ if, and only if, θ_n converges almost surely to θ.

Note: Almost sure estimators are consistent estimators and in fact the convergence is point wise convergence.

Theorem 3.7: If

(i) $\pi_q = z_q/S_q$,

(ii) $\hat{\pi}_q = z_q/\hat{S}_q$ with $z_q \overset{a.s.}{\leq} \hat{S}_q$,

then $\hat{\pi}_q$ is an a.s. estimator of π. $\pi_q \xrightarrow{a.s.} \pi$ and $\hat{\pi}_q$ is an asymptotically unbiased estimator of π.

Proof: By the definition of the z_{N_q}, they are independent and integrable and $\mathrm{Var}(z_{N_q}) = S_{N_q}\pi(1-\pi)$ so that

$$\sum_{q=1}^{\infty} \frac{\mathrm{Var}(z_{N_q})}{S_q^2} < \frac{MM\pi(1-\pi)}{m^2} \sum_{q=1} \frac{1}{q^2} < \infty .$$

Thus, by Kolmogorov's a.s. convergence criterion Pg. 238 Loe've (1960)

$$\pi_q = \sum_{q=1}^{q} \frac{z_{N_k}}{S_q} \xrightarrow{a.s.} \pi.$$

$$|\hat{\pi}_q - \pi| = |\hat{\pi}_q - \pi_q + \pi_q - \pi| \leq |\hat{\pi}_q - \pi_q| + |\pi_q - \pi|.$$

By the first part of the proof $|\pi_q - \pi| \xrightarrow{a.s.} 0$ and

$$|\hat{\pi}_q - \pi_q| = |\frac{z_q}{S_q} - \frac{z_q}{\hat{S}_q}| = \frac{z_q}{S_q} |\frac{m}{M\theta_q} - 1| \xrightarrow{a.s.} 0$$

since $z_q/S_q \xrightarrow{a.s.} \pi$ and $\theta_q \xrightarrow{a.s.} m/M$. Thus $\hat{\pi}_q$ is an a.s. estimator of π. Also, since $\hat{\pi}_q$ is a.s. bounded, it follows by the Dominated Convergence Theorem Pg. 152 Loe've (1960), that $E(\hat{\pi}_q) \to \pi$, and the theorem is proved.

The following lemmas are useful in establishing properties of the confidence intervals for π.

Lemma 3.2: In the equation with real coefficients $ax^2 + bx + c = 0$ if

(i) $a > 0$, $b < 0$, $c \geq 0$,

(ii) $2a + b > 0$, $b^2 - 4ac \geq 0$,

(iii) $a + b \geq -c$,

then the roots,

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a},$$

have the following characteristics:

$$0 \leq x_1 \leq x_2 \leq 1.$$

Proof: Since $b^2 - 4ac \geq 0$, the roots are real, and since $a > 0$, $b < 0$, then $-b/2a > 0$ so that

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \leq \frac{-b + \sqrt{b^2 - 4ac}}{2a} = x_2 .$$

$x_1 \geq 0$ if $-b - \sqrt{b^2 - 4ac} \geq 0$ since $a > 0$. But $ac \geq 0$ since $a > 0$ and $c \geq 0$, so $-4ac \leq 0$ and $b^2 - 4ac \leq b^2$. Thus $\sqrt{b^2 - 4ac} \leq -b$ since $-b > 0$ and therefore, $x_1 \geq 0$. If $-c \leq a + b$, then $-4ac \leq 4a^2 + 4ab$ and $b^2 - 4ac \leq 4a^2 + 4ab + b^2 = (2a + b)^2$, so that

$$\sqrt{b^2 - 4ac} \leq 2a + b \text{ since } 2a + b > 0.$$

Thus,

$$-b + \sqrt{b^2 - 4ac} \leq 2a,$$

or

$$x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \leq 1 \text{ since } a > 0,$$

and the lemma is proved.

Lemma 3.3: If y is any real number and x is a non-negative real number, then $|y| \leq x$ if, and only if, $y^2 \leq x^2$.

Proof: If $|y| \leq x$, then $|y|\,|y| \leq |y|\,x \leq x^2$ or $y^2 \leq x^2$.

If $y^2 \leq x^2$, then $y^2 - x^2 \leq 0$ or $(y - x)(y + x) \leq 0$, which implies

(i) $(y - x) \leq 0$ and $(y + x) \geq 0$, or

(ii) $(y - x) \geq 0$ and $(y + x) \leq 0$.

But,

(i) implies $y \leq x$ and $y \geq -x$ or $|y| \leq x$, and

(ii) implies $y \geq x > 0$ and $y < 0$ since $x > 0$.

This is impossible, hence, the only possible conclusion is that $|y| \leq x$.

The following theorem shows the equivalence between the pivotal quantity

$$\frac{|z_q - \hat{S}_q\pi|}{\sqrt{\hat{S}_q\pi(1-\pi)}} \leq z$$

and the fact that the end points of the confidence interval for π are roots of an appropriate quadratic equation and that they also satisfy certain conditions.

Theorem 3.8: For $0 < z < \infty$

(i) $$\frac{|z_q - \hat{S}_q\pi|}{\sqrt{\hat{S}_q\pi(1-\pi)}} \leq z$$

if, and only if, $a_q\pi^2 + b_q\pi + c_q \leq 0$ where

$a_q = \hat{S}_q(\hat{S}_q + z^2)$,

$b_q = -\hat{S}_q(2z_q + z^2)$,

$c_q = z_q^2$,

(ii) the roots of the equation $a\pi^2 + b_q\pi + c_q = 0$ are

$$\pi_{1q} = \frac{-b_q - \sqrt{b_q^2 - 4a_qc_q}}{2a_q}, \quad \pi_{2q} = \frac{-b + \sqrt{b_q^2 - 4a_qc_q}}{2a_q}$$

and if $\hat{S}_q \overset{a.s.}{\geq} z_q$ satisfy

(iii) $0 \overset{a.s.}{\leq} \pi_{1q} \overset{a.s.}{\leq} \pi_{2q} \overset{a.s.}{\leq} 1,$

(iv) $\pi_{1q} - \pi_{2q} \xrightarrow{a.s.} 0.$

(v) If $z_q = 0$, then $\pi_{1q} = 0.$

Proof: (i) From Lemma 3.3

$$\frac{|z_q - \hat{S}_q\pi|}{\sqrt{\hat{S}_q\pi(1-\pi)}} \leq z$$

if, and only if,

$$\frac{(z_q - \hat{S}_q\pi)^2}{\hat{S}_q\pi(1-\pi)} \leq z^2$$

or, equivalently, if, and only if,

$$z_q^2 - 2z_q\hat{S}_q\pi + \hat{S}_q^2\pi^2 \leq z^2\hat{S}_q\pi(1-\pi),$$

or,

$$\hat{S}_q(\hat{S}_q + z^2)\,\pi^2 - \hat{S}_q(2z_q + z^2)\,\pi + z_q^2 \leq 0,$$

which proves (i)

(ii) Clearly the roots of the equation

$$a_q\pi^2 + b_q\pi + c_q = 0$$

where

$$a_q = \hat{S}_q(\hat{S}_q + z^2),$$

$$b_q = -\hat{S}_q(2z_q + z^2),$$

$$c_q = z_q^2,$$

are $\pi_{1q} = \dfrac{-b_q - \sqrt{b_q^2 - 4a_q c_q}}{2a_q}$ and $\pi_{2q} = \dfrac{-b_q + \sqrt{b_q^2 - 4a_q c_q}}{2a_q}$.

(iii) Since it is assumed that $\hat{S}_q \overset{a.s.}{\geq} z_q$ and by definition $\hat{S}_q > 0$ and $z_q^2 \geq 0$ and $z^2 > 0$, it is evident that $a_q > 0$, $b_q < 0$, and $c_q \geq 0$. Since $\hat{S}_q \overset{a.s.}{\geq} z_q$

$$a_q + b_q = \hat{S}_q(\hat{S}_q - 2z_q) \overset{a.s.}{\geq} z_q(z_q - 2z_q) = -z_q^2 = -c_q$$

$$b_q^2 - 4a_q c_q = \hat{S}_q^2(4z_q^2 + 4z_q z^2 + z^4) - 4z_q^2\hat{S}_q(\hat{S}_q + z^2)$$

$$= 4z_q\hat{S}_q z^2(\hat{S}_q - z_q) + z^4\hat{S}_q^2 \overset{a.s.}{>} 0.$$

$$2a_q + b_q = 2\hat{S}_q(\hat{S}_q - z_q) + \hat{S}_q z^2 \overset{a.s.}{\geq} \hat{S}_q z^2 > 0.$$

Thus by Lemma 3.2, $0 \overset{a.s.}{\leq} \pi_{1q} \overset{a.s.}{\leq} \pi_{2q} \overset{a.s.}{\leq} 1$.

(iv)

$$\pi_{2q} - \pi_{1q} = \frac{\sqrt{b_q^2 - 4a_q c_q}}{a_q},$$

$$= \frac{\sqrt{4z_q\hat{S}_q z^2(\hat{S}_q - z_q) + z^2\hat{S}_q^2}}{\hat{S}_q(\hat{S}_q + z^2)},$$

$$\overset{a.s.}{\leq} \frac{\sqrt{4\hat{S}_q^3 z^2 + z^4\hat{S}_q^3}}{\hat{S}_q(\hat{S}_q + z^2)}, \text{ since } \hat{S}_q \overset{a.s.}{\geq} z_q \text{ and}$$

$\hat{S}_q \geq 1$ for q large.

Therefore,

$$\pi_{2q} - \pi_{1q} \overset{a.s.}{\leq} \frac{1}{\sqrt{\hat{S}_q}} \sqrt{\frac{4z^2 + z^4}{1 + z_q^2/S_q}} \to 0 ,$$

as $q \to \infty$ since $\hat{S}_q \to \infty$ as $q \to \infty$ and $\sqrt{4z^2 + z^4}$ is finite for z finite.

Thus, $\pi_{2q} - \pi_{1q} \xrightarrow{a.s.} 0$.

(v) If $z_q = 0$, then $c_q = 0$ and

$$\pi_{1q} = \frac{-b_q - \sqrt{b_q^2}}{2a_q} = \frac{-b_q - (-b_q)}{2a_q} = 0 \text{ since } b_q < 0,$$

and the theorem is proved.

These results prove that

$$P\left[\frac{|z_q - \hat{S}_q\pi|}{\sqrt{\hat{S}_q\pi(1-\pi)}} \leq z\right] = P\left[\pi_{1q} \leq \pi \leq \pi_{2q}\right] ,$$

so that

$$\lim_q P\left[\pi_{1q} \leq \pi \leq \pi_{2q}\right] = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-t^2/2}\, dt$$

where $z > 0$ and π_{1q} and π_{2q} have the characteristics given in the theorem above.

Furthermore for all $\pi \quad \left[\pi_{1q}, \pi_{2q}\right]$,

$$\frac{|z_q - \hat{S}_q\pi|}{\sqrt{\hat{S}_q\pi(1-\pi)}} \leq z$$

for all $z > 0$, and conversely. This is true since the function

$$y_q = a_q\pi^2 + b_q\pi + c_q$$

is a quadratic function in π and $a_q > 0$ implies the graph of the function is a parabola opening upward with π-intercepts at π_{1q} and π_{2q}.

CHAPTER IV

RATES OF CONVERGENCE

In this chapter an attempt is made to determine just how good the approximate confidence intervals in the preceding two chapters are or alternatively to determine the rate in terms of q at which each of these distribution functions approach the Normal Distribution with mean zero and variance one.

In Theorem 2.1 it was shown that

$$\mathcal{L}\left(\frac{\hat{S}_q - S_q}{s_q}\right) \to N(0,1),$$

and in Theorem 3.3 it was shown that

$$\mathcal{L}\left(\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}}\right) \to N(0,1).$$

Both of these random variables were shown to be sums of independent random variables centered at expectations and by the bounded assumption the third absolute moment is finite. They thus satisfy the conditions necessary to employ the theorem of Crame'r (1937); Pg. 77, for determining their rates of convergence to N(0,1).

In Theorem 3.5 it was shown that

$$\left|P\left[\frac{z_q - S_q\pi}{\sqrt{S_q\pi(1-\pi)}} < z\right] - P\left[\frac{z_q - \hat{S}_q\pi}{\sqrt{\hat{S}_q\pi(1-\pi)}} < z\right]\right| \leq P\left[\theta_q \neq m/M\right]$$

for all $z \in R$, R the set of real numbers. But since