Optimal Sample Size Determination in Seed Health Testing - A Simulation Study

Hari Krishna Susarla

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

June 2005

Your file Votre référence
ISBN: 0-494-10215-2
Our file Notre référence
ISBN: 0-494-10215-2

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis is prepared

By: Hari Krishna Susarla

Entitled: Optimal Sample Size Determination in Seed Health Testing - A Simulation Study

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair

______________________________ Examiner

______________________________ Examiner

______________________________ Supervisor

Approved by ______________________________
Chair of Department or Graduate Program Director

__________ 2005 ______________________________
Dean of Faculty

ABSTRACT

Optimal Sample Size Determination in Seed Health Testing - A Simulation Study

Hari Krishna Susarla

Selection of an appropriate sample size is an important prerequisite to any statistical investigation

In this thesis the problem of identifying the sample size for testing the seed health by noting the presence or the absence of pathogen(s) is considered. The cross-classified data of variety by seed by pathogen is collected for the purpose, which consists of N observations for each variety of seed. Here N is regarded as population size and the outcome is a Bernoulli random variable.

A simulation method for identifying the sample size is developed and is compared with five existing methods.

The simulation method is based on chi-square (χ^2) measure of goodness of fit of empirical distribution with that of a theoretical distribution. Here k repeated samples for each of the sample sizes n=10(10)50(25)100(100)500, using a simple random sampling without replacement (SRSWOR), are considered. For each of the k samples of size n, the chi-square (χ^2) measure of goodness of fit is computed. Since these k observed χ^2 values follow a theoretical χ^2 distribution, we considered the upper 0.05 quantile χ^2 (0.05-uq) and corresponding P-value for sample size determination. Thus for each sample size n, we have the 0.05-uq χ^2 and the corresponding P-value. Now the optimal sample size is determined as equal to the earliest instance of that sample size corresponding to which the P-value is non-significant at the desired level of significance.

Acknowledgements

I wish to thank my thesis supervisor, Dr. X. Zhou. His guidance and encouragement helped me stay on course. I am also grateful to Dr. Y.P. Chaubey for his help in getting installed SAS software in the department Graduate Computer Lab. Finally, I would like to thank Karuna Sri Godavari for her loving support throughout the writing of this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Importance of Sample Size

Determining an appropriate sample size for a study whether it is an agricultural field experiment, laboratory animal study or seed health testing, is an important step in the statistical investigation.

The following examples are taken from Desu and Raghava Rao (1990). Consider a gubernatorial election in which Candidate A has actually received 45% of electoral votes. If an exit poll of 100 voters is taken, then there is a 13.35% chance of his getting more than half of votes. If a sample of 2000 voters is taken, however then this chance is nearly zero. By taking a small sample of 100 voters, the candidate gets a false hope of winning the election, while the true picture emerges with a large sample of 2000 voters. Interviewing 2000 voters is expensive and time consuming. One would like to find a minimum required sample size that enables one to estimate the proportion of his preferred votes to the desired level of accuracy.

In the next example consider the case of a manufacturer of Model X cars with given equipment options who claims that those cars give an average gas mileage of at least 25 miles per gallon (mpg). Assume that the standard deviation of gas mileage delivered by such cars is 1 mpg. A consumer protection agency wants to disprove the manufacturer's claim using a test at α=0.05 level of significance. Using only four test cars, the agency has 26% chance of rejecting the manufacturer's claim, when the actual gas mileage delivered is 24.5mpg. Thus cars giving 0.5 mpg less than the claim have a 74% chance of meeting the manufacturer's claim. If 100 cars are tested this chance of accepting the manufacturer's claim becomes nearly 0%. Then once again there is an appropriate sample size that enables the experimenter even to retain hypothesis with a high probability when it fails.

An adequate sample size ensures reliable information regardless of the outcome of the study. Conducting of a study with an inadequate sample size is not only futile but also unethical. In clinical or laboratory studies exposing human subjects or animals to the risks of research is justifiable only when there is a realistic possibility of benefit to those subjects or may lead to substantial scientific progress.

Here we consider the problem of testing seed health by noting the presence or the absence of pathogen(s) on each seed of a given seed lot of sorghum variety. Absence of a pathogen indicates healthiness of a seed (free from disease) and presence indicates otherwise. Thereby one can dichotomize a given seed lot to be either healthy or diseased with a reasonably small degree of risk. For this study, a large number of seeds N=1050, is considered for each variety of seed lot, where the number of seed varieties is equal to five and each seed is tested for the presence or the absence of a particular pathogen and is scored as either 1 or 0 respectively. The number of pathogens identified is 15. Clearly each data point i.e., a particular pathogen-seed combination is a Bernoulli trial. The cross-classified data of variety by seed by pathogen is collected. Here N is assumed to be large enough to be considered as population.

The primary purpose of the study is to find the number of seeds to be tested for each variety of seed material, to identify the pathogens, in order to classify the seed material to be healthy or diseased across all the pathogens. Usually it is very expensive and time consuming to collect, maintain and test a large sample of seeds. So one may want to know a suitable sample size with which he can make a decision about the seed lot across all the pathogens involved with a reasonable degree of accuracy.

Objectives

1. The purpose of the study is to find an appropriate sample size which statistically determines the health of a given seed lot for a desired degree of accuracy.
2. To explore and compare different statistical methods available.
3. To do a simulation study (data driven approach) to find a suitable sample size and compare with other methods.

1.2 Materials and Methods

1.2.1 Seed Material

One-kilogram seeds of each of five sorghum seed varieties (IS-10392, IS-10757, IS-2742, IS-3025 and IS-8080) were collected in the year 2002 from the Quarantine unit of the International Crops Research Institute for the Semi Arid Tropics (ICRISAT), Hyderabad, India, gene bank, which was called as original sample (OS). Random samples of seeds of sizes 50, 100(100)400, were taken from each OS, and called as working samples (WS) for the purpose of seed health testing. The total number of seeds N=1050 in the WS were arranged ten seeds per plate in blotter method and replications were made in proportion to the size of WS. Then each seed from each sorghum variety was evaluated for the presence or the absence of each of 15 pathogens (list is given below) with a stereo-binocular microscope. Not all the varieties have shown all the pathogens. Some times a particular seed may show incidence of more than one pathogen. The number of seeds shown a particular pathogen was recorded and a proportion of incidence was calculated. This proportion is some times called as incidence rate. Table 1.1 shows the incidence of pathogens for different sorghum varieties.

1.2.2 Data

The data Y_{ij} , i=1...N, j=1...d consists of the presence or the absence of a pathogen, coded as (1or 0), for each of d pathogens corresponding to N seeds. The data Y_{ij} can be considered as independent Bernoulli random variables. Since the incidence of a pathogen on a seed is independent of the incidence of the same pathogen on any other seed. So these pathogen incidences can be considered as independent trials with probability of the presence or the absence of a pathogen being constant. So pathogen incidence on each seed is a Bernoulli trial.

The following is the list of pathogens identified under laboratory testing, which were coded for all the N seeds for the purpose of study.

1. Aspergillus niger (AN),
2. Aspergillus flavus (AF),
3. Rhizopus species (RH),
4. Curvularia lunata (CL),
5. Alternaria alternata (ALT),
6. Fusarium species (FS),
7. Penicillium species (PEN),
8. Epicoccum nigrum (EP),
9. Curvularia species (CUR),

10. Bipolaris sorghicola (BIP),
11. Exserohilum rostratum (EXS),
12. Phoma sorghina (PHO),
13. Nigrospora orizae (NIG),
14. Chaetomium globossium (CHA) and
15. Trichothecium species (TRI) .

The incidence of pathogens on different sorghum seed varieties is given in Table 1.1 and the incidence rates are given in Table 1.2 respectively. Note that from Table 1.1 not all pathogens were present in all the seed varieties. Also from Table 1.2 one can note that the pathogen incidence rates are very small except for the pathogens CL, AL and FU across all the seed varieties.

Table 1.1 Pathogen incidences for different sorghum varieties

	Variety				
Pathogen	IS-10392	IS-10757	IS-2742	IS-3025	IS-8080
AN		*			
AF	*				
RH					*
CL	*	*	*	*	*
AL	*	*	*	*	*
FU	*	*	*	*	*
PE	*				
EP	*	*	*	*	*
CU	*	*	*	*	*
BI	*	*	*	*	*
BH	*		*	*	*
NI	*				*
CH			*		*
TR					*
EX			*		

* Represents presence of pathogen

Table 1.2 Percentage incidence of pathogens in different sorghum varieties

		Sorghum variety									
		IS-10392		IS-10757		IS-2742		IS-3025		IS-8080	
S.No.	Pathogen	Prop	%inc	Prop	%inc	Prop	%inc	Prop	%inc	Prop	%inc
1	AN	0	0.00	0.001905	0.19	0	0.00	0	0.00	0	0.00
2	AF	0.000952	0.10	0	0.00	0	0.00	0	0.00	0	0.00
3	RH	0	0.00	0	0.00	0	0.00	0	0.00	0.000952	0.10
4	CL	0.597143	59.71	0.645714	64.57	0.534286	53.43	0.716190	71.62	0.850476	85.05
5	AL	0.104762	10.48	0.048571	4.86	0.075238	7.52	0.079048	7.90	0.105714	10.57
6	FU	0.052381	5.24	0.013333	1.33	0.371429	37.14	0.069524	6.95	0.005714	0.57
7	PE	0.007619	0.76	0	0.00	0	0.00	0.000952	0.10	0	0.00
8	EP	0.016190	1.62	0.009524	0.95	0.004762	0.48	0.008571	0.86	0.010476	1.05
9	CU	0.007619	0.76	0.004762	0.48	0.009524	0.95	0.009524	0.95	0.038095	3.81
10	BI	0.005714	0.57	0.002857	0.29	0.003810	0.38	0.003810	0.38	0.000952	0.10
11	PH	0.008571	0.86	0	0.00	0.001905	0.19	0.004762	0.48	0.003810	0.38
12	NI	0.002857	0.29	0	0.00	0	0.00	0	0.00	0.000952	0.10
13	CH	0	0.00	0	0.00	0.000952	0.10	0	0.00	0.002857	0.29
14	TR	0	0.00	0	0.00	0	0.00	0	0.00	0.000952	0.10
15	EX	0	0.00	0	0.00	0.002857	0.29	0	0.00	0	0.00

%inc: Percentage incidence of pathogen

Chapter 2

Survey of Literature

2.1 Population or Sample Size Estimation.

Let X_1, X_2,..., X_n be independent random variables with same probability density function (PDF) f(x|θ), where θ is the parameter of interest. Let X_i be a Bernoulli random variable and the number of observed 1's, M is a binomial variable with parameters n and θ. Kotz and Johnson (1986) consider a situation where estimation of sample size n is of considerable interest along with θ from observed values of M. In some other situations, n represents the population size, but the problem of estimation is similar in both the cases.

In some life testing experiments the total number of items being tested is known before hand and the test is carried out for a fixed amount of time, which results in a truncated sample. This type of situation may arise when, among the items put on a life test, there is certain unknown number of items with a specific defect identifiable only after failure. Blumenthal and Marcus (1975) study this situation, where the interest lies in estimating the number of defectives of a particular type after an initial period. Jelinski and Moranda (1972) consider a similar situation of estimating the total number of errors N in the testing program after running the program for a fixed period of time and thus obtaining time of detection for M distinct defective items.

Anscombe (1961) considers a situation of estimating the number of responses to an advertising campaign. Here n represents the number of responsive people among K contacted in the campaign. Wittes (1970) studies the problem of estimating the size of a subpopulation of persons who have a trait that occurs rarely in the population at large.

Binomial samples with unknown n, arise when in a sequence of Bernoulli trials only the successes are observable. Let M be the number of successes in n (unknown) Bernoulli trials with probability of success θ. Blumenthal and Dahiya (1981) and many other authors have investigated the situation of this unknown sample size. For the case of θ is known, Feldman and Fox (1968) examine asymptotic properties of the MLE (maximum likelihood estimation) and MME (modified moment estimation) of n and also examine two modifications of these estimators. Olkin, et al. (1981) propose some stabilized estimators. Draper and Guttman (1971) consider Bayes estimators, which we will not discuss here.

The following is the summary of results given by Blumenthal and Dahiya (1981).

Case: θ is known:

Let $X_1, X_2, \ldots, X_r$ be a random sample from a binomial (n, θ), where θ is known and we want to estimate n.

Then the likelihood is

$$L(n \mid \theta, x) = \prod_{i=1}^{n} \binom{n}{x_i} \theta^{x_i} (1-\theta)^{n-x_i} \,. \qquad (2.1)$$

Then the integer valued MLE $\hat{n}$ of n is given by [V], where V is the solution of

$$r \log_e (VQ) - \sum_{i=1}^{r} \log_e (V - x_i) = 0 \,,$$

and Q=1-θ. Also the above authors give asymptotic properties of $\hat{n}$ in the following theorem.

2.1.1 Theorem (Blumenthal and Dahiya (1981))

a. For any sample the estimator $\hat{n}$ is finite.

b. For n fixed $P(\hat{n}=n)\to 1$ as $r \to \infty$,

c. Let $n \to \infty$, and let r be either fixed or increasing, then

1. $$\frac{(\hat{n}-n)}{n^{\alpha}} \xrightarrow{P} 0 \text{ for } \alpha > ½ \text{ and } \log_e\left(\frac{r}{n}\right) \to 0,$$

2. $$\sqrt{\frac{r}{n}}(\hat{n}-n) \xrightarrow{L} N\left(0, \frac{1-\theta}{\theta}\right), \quad if\frac{r}{n} \to 0.$$

Recently Adcock (1997) reviews different sample size determination (SSD) methods. SSD for binomial distribution is one of the oldest and formal procedures available, for example one can look into Cochran (1963). The conventional starting point of SSD for the binomial distribution is to assume normal approximation for the sample proportion of successes, p

i.e., $$p \sim N(\theta, \theta(1-\theta)/n),$$

where θ is the population proportion. Let the probability of the absolute difference, e, between p and θ is at least 1- α, at the desired level of significance α . It can be written as,

$$P[|p-\theta| \le e] \ge 1-\alpha,$$

which leads to

$$n \ge \frac{\theta(1-\theta)Z_{\alpha/2}^2}{e^2}, \tag{2.2}$$

where Z_α is the upper α% value of standard normal distribution.

Desu and Raghava Rao (1990) describe the procedure of determination of sample size n to test the hypothesis $H_0{:}\theta=\theta_0$ vs. $H_1{:}\theta=\theta_1$, using the arcsine transformation

$$Z = 2\sqrt{n}[\mathrm{Sin}^{-1}(\sqrt{p}) - \mathrm{Sin}^{-1}(\sqrt{\theta})],$$

which follows approximately N(0,1). This leads to an expression for n as,

$$n = \frac{0.25(Z_\alpha + Z_\beta)^2}{\left(\mathrm{Sin}^{-1}\sqrt{\theta_0} - \mathrm{Sin}^{-1}\sqrt{\theta_1}\right)^2}, \tag{2.3}$$

which requires both θ_0 and θ_1 to be specified. Where Z_α and Z_β are the upper α% and β% values of standard normal distribution. If we write $\theta_1 = \theta_0 + e$ and expand the $Sin^{-1}\sqrt{\theta_1}$ about θ_0, by Taylor's theorem to terms of order e, we obtain,

$$n = \frac{\theta_0(1-\theta_0)(Z_\alpha + Z_\beta)^2}{e^2} \tag{2.4}$$

A Bayesian treatment of SSD for the binomial distribution is given by Adcock (1987), which we will not discuss here.

2.2 Survey of Existing Methods

In general, investigation of the sample size determination methods is difficult because of complex mathematical nature and multitude of different formula.

In classical hypothesis testing we have two hypothesis, null hypothesis and alternative hypothesis. The null hypotheses usually concludes that there is no significant difference between groups being compared with respect to the variable of interest. For example comparing the effects of two pesticides in a pest control experiment, the null hypothesis would be that the incidence of pest on the crop that receives one pesticide is the same as the incidence of pest in the crop that receives the other. To draw reliable conclusions, one has to define the null hypothesis prior to the beginning of the data collection. Mathematically the null hypothesis can be defined in many different ways. For example the proportion of plants has disease in each variety of pesticide-applied fields could be written as ($\theta_1 = \theta_2$), where θ_1 and θ_2 are population proportion values of diseased plants in both the pesticide applied fields respectively.

The alternative hypothesis refers that there is a significant difference between the treatments compared with respect to the variable of interest. The magnitude of this difference of treatments, i.e., (θ_2-θ_1), is generally referred to as effect size. This effect size plays an important role and choosing it is the first step in sample size determination. One usually designs a statistical study to find a minimal statistically significant effect size. If the study data really detects this effect size, then one is forced to change his usual practices in favor of the other. Determination of the minimal significant effect size is generally a non-statistical judgment based on the context and field of study. If one wants to find a smaller effect size then a larger sample size is needed. Generally time and resources do not permit to have a large sample size.

There are always some risks involved in drawing conclusions from an experimental data. These are called type-I and type-II errors. Type-I error is associated with falsely rejecting the null hypothesis when it is true. That is in other words falsely detecting a difference when in reality it does not exists. It is a type of false positive. The probability of this error is denoted by α.

Type-II error is associated with falsely accepting the null hypothesis and concluding that there is no difference, but in reality the difference exists. This is a false negative. The probability of this error is denoted by β.

Usually the type-II error is considered in terms of power of the study rather than in β alone. The power of the study is the probability of obtaining a statistically significant

P-value, if a true difference exists, that is equal to the effect size defined by the alternative hypothesis. Then the power of the study is 1-β. For any statistical study the power of the method is determined by sample size, effect size and by α.

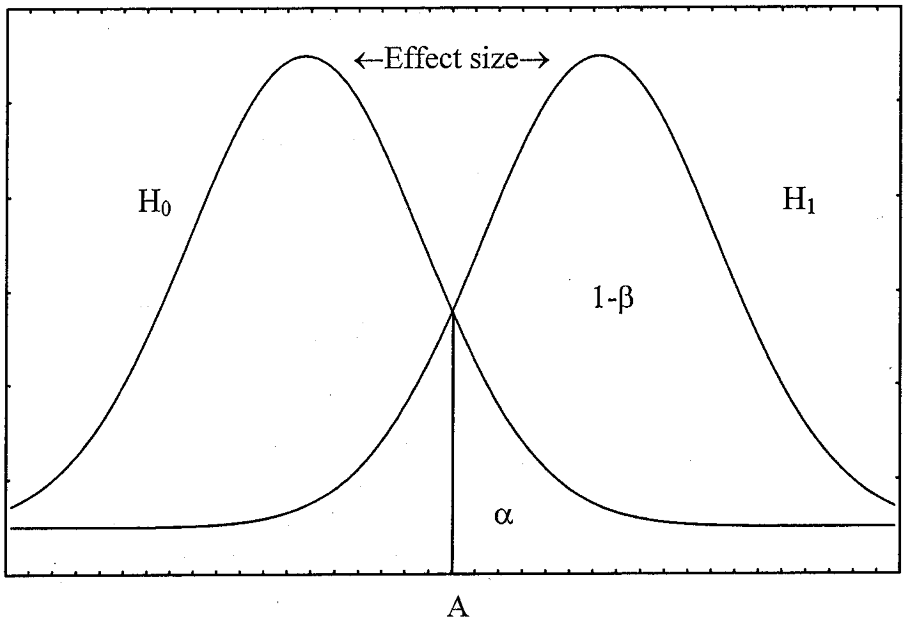
In the study of power, we have three parameters under control. They are any three of the sample size, the effect size, α and β. By fixing any three parameters together one can determine the fourth. For example if one fixes α, effect size and power then he can determine the sample size.

Graphically the power analysis can be shown as in Figure 2.1. Here we consider a normal distribution setup. Because when sample size is large Binomial distribution converges to normal distribution. The curve on the left hand side is the distribution of the values under null hypothesis, where as the curve on the right hand side is the distribution of values under the alternative hypothesis. Here the effect size is the difference between the peaks of the two curves. Suppose we draw a vertical line corresponding to a P-value equal to α, which touches the horizontal axis at A, then any value falling right of A results in rejection of the null hypothesis. Now the power of the test is the area under the second curve to the right of A which is 1-β.

If one chooses a small value of α then the area under the right curve decreases, which results in lower value of power. That is, setting a more stringent criterion for rejecting the null hypothesis results in increase of type-II error.

In literature for example see Roger (2000), one could find many mathematical approximations for sample size determination formulas and by nature they are all ratios. In all of them the outcome of interest is assumed to follow a normal distribution. Generally the numerator term is a function of Z_α and Z_β and the denominator term is proportional to the square of the effect size. Here Z_α and Z_β are standard normal critical values corresponding to α and β levels of significance.

Figure 2.1 Power analysis



Usually statistical investigations are planned to estimate a parameter of interest. For example the mean yield per hectare, mean IQ score in a psychological experiment, etc that are continuous variables and proportions such as proportion number of defectives in a quality control study, are discrete variables. In each of these cases one can measure the precision by considering a width of (1-α) % confidence interval. If one uses a larger sample size then he gets a smaller width for confidence interval. Therefore one may consider the width of the interval instead of effect size to determine an appropriate sample size. Sometimes this confidence width is termed as precision, which is generally a function of variance of data. Thus a statistical study can be planned to a desired precision without choosing or bothering about of the effect size.

One cannot plan a statistical investigation if the likely variability or variance of the outcome of the interest is not known. Such a variability is generally not under the control of the experimenter and is important in determining the sample size. Therefore one should estimate the variability, without that, it is impossible to estimate the sample size. To estimate the variance, one can conduct a pilot study to get a likely value of variance. A Number of authors have provided different approaches to conduct pilot studies, and information thus derived is incorporated into the final analysis to determine the sample size.

When the variance is poorly known, the likely effect size is difficult to predict. In such a case one can think of using a sequential trial based on maximum possible variance. If the variance in the real data is small, then that will reduce the required sample size and the trial will terminate earlier.

In the light of above discussions, we now consider the following approaches to find optimum sample size when the outcome of the interest is a Bernoulli variable.

1. Sampling from a Bernoulli distribution.
2. Power analysis
3. Coefficient of variation method.
4. Sequential analysis based on coefficient of variation..
5. Poisson method.

2.2.1 Sampling from a Bernoulli Distribution

Let X be a Bernoulli random variable with probability of success θ. It has a probability mass function as,

$$f(x;\theta) = \theta^{x}(1-\theta)^{1-x}, \qquad x=0,1. \qquad (2.5)$$

Then the mean and variance of this distribution are respectively θ and θ(1-θ). Now let $X_1,X_2, \ldots,X_n$ be a random sample on X. Then the sum of Bernoulli random variables,

$$Y = \sum_{i=1}^{n} X_i$$

has a binomial distribution with parameters n and θ, and its probability mass function is given as

$$f(y,n,\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}, \qquad y=0,1,2, \ldots n \qquad (2.6)$$

Then the mean and variance of Y are nθ and nθ(1-θ) respectively. Thus Y/n is an unbiased estimator of θ. We now consider the determination of sample size in estimation and testing of hypothesis on θ in the following sections.

Estimation of θ in Binomial Distribution

Let θ is estimated by $\hat{\theta}$=Y/n. One can determine sample size n by controlling the absolute error e, where $e = \left|\hat{\theta} - \theta\right|$ is the positive difference between $\hat{\theta}$ and θ, with a high probability, that satisfies the following relation,

$$P[|\hat{\theta} - \theta| \leq e] \geq 1-\alpha,$$

i.e., $$P\left[\left|\frac{Y}{n} - \theta\right| \leq e\right] \geq 1-\alpha, \qquad (2.7)$$

where α and e are pre-specified positive constants.

Now the left hand side term of (2.7) can be written as,

$$P\left[-e \le \frac{Y}{n} - \theta \le e\right] \ge 1-\alpha, \tag{2.8}$$

i.e., $$P[n(\theta - e) + 1 \le Y \le n(\theta + e)] = p[y_1 \le Y \le y_2]$$

$$= \sum_{y=y_1}^{y_2} \binom{n}{y} \theta^y (1-\theta)^{n-y}. \tag{2.9}$$

To obtain a solution of (2.7) one can use central limit theorem, which states that the distribution of $\frac{Y/n - \theta}{\sqrt{\theta(1-\theta)/n}}$ is asymptotically N(0,1).

Without loss of generality one can divide throughout the left hand side term of above (2.8) with $\sqrt{\theta(1-\theta)/n}$ and gets,

$$P\left[-\frac{e}{\sqrt{\theta(1-\theta)/n}} \le \frac{Y/n-\theta}{\sqrt{\theta(1-\theta)/n}} \le \frac{e}{\sqrt{\theta(1-\theta)/n}}\right] \ge 1-\alpha,$$

i.e., $$P\left[-\frac{e}{\sqrt{\theta(1-\theta)/n}} \le Z \le \frac{e}{\sqrt{\theta(1-\theta)/n}}\right] \ge 1-\alpha.$$

Where Z is a standard normal variable. Because normal distribution is a symmetric distribution, the areas to the left side and right side of Z are equal in magnitude. Using this symmetric property of normal distribution the above inequality can be written as,

$$P[Z \le Z_{\alpha/2}] \ge 1-\alpha/2,$$

where

$$Z_{\alpha/2} = \frac{e}{\sqrt{\theta(1-\theta)/n}} \tag{2.10}$$

is the upper ($\alpha/2$)% value of standard normal distribution.

By rearranging the terms in (2.10) one can get an expression for lower bound for sample size n as

$$n \geq \frac{\theta(1-\theta)Z_{\alpha/2}^2}{e^2}. \qquad (2.11)$$

Because θ is not known, one can concentrate on maximum value of the product θ(1-θ). However, since 0≤θ≤1, the product θ(1-θ) attains its maximum equal to 0.25 when θ=0.5. So one can use θ =0.5 to get a generous estimate of n as denoted by $[n^*]+1$ where

$$n^* \geq \frac{Z_{\alpha/2}^2}{4e^2} \qquad (2.12)$$

and $[n^*]$is the integer value of n^*. Thus one can also tabulate the required sample sizes for varying values of α, θ and e.

Now one has the following result [see Desu and Raghava Rao(1990)].

Result

An appropriate sample size n for estimating θ is given by $[n^*]+1$, where$n^*$is computed from (2.12).

3.2 Power Analysis (Tests of Hypothesis About θ)

Suppose one wants to test H_0:θ=θ_0 against the one sided alternative H_1:θ>θ_0 .

Then let the critical region be

$$Y > c, \qquad (2.13)$$

where c satisfies

$$P[Y > c \mid \theta = \theta_0] \leq \alpha . \qquad (2.14)$$

Also if one wants to have power to be at least (1-β) at $\theta=\theta_1(>\theta_0)$, then that can be mathematically expressed as

$$P[Y > c \mid \theta = \theta_1] \geq 1 - \beta . \qquad (2.15)$$

In normal distribution case the power analysis can be depicted as shown in Figure 3.1 (pp.14). Now an approximate solution to sample size n can be obtained by making the well-known arcsine transformation [see Desu and Raghava Rao (1990)] on $\hat{\theta} = Y/n$ that is defined as,

$$Z = 2\sqrt{n}[\mathrm{Sin}^{-1}\sqrt{\hat{\theta}} - \mathrm{Sin}^{-1}\sqrt{\theta}] \sim N(0,1)$$

asymptotically and expressing the probabilities in (2.14) and (2.15) in terms of standard normal distribution function as follows.

Rearranging the terms of inequality (2.14) and equating it to maximum risk α, one gets

$$P\left[\frac{Y}{n} > \frac{c}{n}\middle| \theta = \theta_0\right] = \alpha .$$

Now applying the arcsine transformation on Y/n, one will get

$$P\left[Z > 2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}} - \mathrm{Sin}^{-1}\sqrt{\theta_0}\right\}\right] = \alpha .$$

Which can be written as

$$1 - P\left[Z < 2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}} - \mathrm{Sin}^{-1}\sqrt{\theta_0}\right\}\right] = \alpha .$$

Therefore

$$1 - \varphi\left[2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}} - \mathrm{Sin}^{-1}\sqrt{\theta_0}\right\}\right] = \alpha ,$$

where ϕ is a standard normal distribution function.

Hence

$$\varphi\left[2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}}-\mathrm{Sin}^{-1}\sqrt{\theta_0}\right\}\right]=1-\alpha .$$

Finally one will get,

$$Z_\alpha=2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}}-\mathrm{Sin}^{-1}\sqrt{\theta_0}\right\}. \qquad (2.16)$$

Similarly rearranging the terms in inequality (2.15) and equating to the lower bound of power, that is equal to (1-β), one will get,

$$P\left[\frac{Y}{n}>\frac{c}{n}\middle|\theta=\theta_1\right]=1-\beta$$

Now again by applying the arcsine transformation on Y/n, one can get

$$P\left[Z>2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}}-\mathrm{Sin}^{-1}\sqrt{\theta_1}\right\}\right]=1-\beta .$$

Proceeding as above one gets,

$$1-\varphi\left[2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}}-\mathrm{Sin}^{-1}\sqrt{\theta_1}\right\}\right]=1-\beta .$$

Since β is the type II error defined under the alternative hypothesis and is a lower tail probability, the corresponding value of standard normal distribution is $-Z_\beta$ at β level of significance. Thus one will get

$$-Z_\beta=2\sqrt{n}\left\{\mathrm{Sin}^{-1}\sqrt{\frac{c}{n}}-\mathrm{Sin}^{-1}\sqrt{\theta_1}\right\},$$

and can be written as

$$Z_\beta = 2\sqrt{n}\left\{\text{Sin}^{-1}\sqrt{\theta_1} - \text{Sin}^{-1}\sqrt{\frac{c}{n}}\right\}. \qquad (2.17)$$

Then the solution to the sample size n can be found by adding the equations (2.16) and (2.17), which is

$$Z_\alpha + Z_\beta = 2\sqrt{n}\{\text{Sin}^{-1}\sqrt{\theta_1} - \text{Sin}^{-1}\sqrt{\theta_0}\}.$$

In the above expression the only unknown quantity is n. Then finally solving for n, one can get

$$n = \frac{(Z_\alpha + Z_\beta)^2}{4[\text{Sin}^{-1}\sqrt{\theta_1} - \text{Sin}^{-1}\sqrt{\theta_0}]^2}. \qquad (2.18)$$

In the above expression (2.18) if one writes $\theta_1 = \theta_0 + e$, where e is defined earlier to be equal to $\left|\frac{Y}{n} - \theta\right|$, and expand $Sin^{-1}\sqrt{\theta_1}$ about θ_0 by Taylor's expansion to the terms of order e, Desu and Raghava Rao (1990) obtain,

$$\text{Sin}^{-1}\sqrt{\theta_1} = \text{Sin}^{-1}\sqrt{\theta_0 + e}$$

$$= \text{Sin}^{-1}\sqrt{\theta_0} + \frac{d}{d\theta_0}\left(\text{Sin}^{-1}\sqrt{\theta_0}\right)e + \frac{d^2}{d\theta_0^2}\left(\text{Sin}^{-1}\sqrt{\theta_0}\right)\frac{e^2}{2!} + \ldots$$

and finally after differentiating and omitting the higher order terms (because they are negligible), one gets,

$$\text{Sin}^{-1}\sqrt{\theta_1} = \text{Sin}^{-1}\sqrt{\theta_0} + \frac{e}{\sqrt{4\theta_0(1-\theta_0)}} + \varepsilon \qquad (2.19)$$

where ε is a negligible quantity.

Now substitute (2.19) in (2.18) and solving for n, one obtains,

$$n = \frac{\theta_0(1-\theta_0)(Z_\alpha + Z_\beta)^2}{e^2}. \qquad (2.20)$$

Now (2.20) leads to the conservative rule similar to the inequality in (2.12) when 0.5 replaces θ_0, as

$$n = \frac{(Z_\alpha + Z_\beta)^2}{4e^2}. \tag{2.21}$$

Now one has the following result [see Desu and Raghava Rao (1990)].

Result

Consider a one sided α level test $H_0:\theta=\theta_0$ vs the alternative $H_1:\theta=\theta_1(>\theta)$. Then an approximate sample size, required to give power at least 1-β at the alternative is [n]+1, where n is given by (2.18).

2.2.3 Coefficient of Variation (CV) Method (Sukhatme and Sukhatme (1970))

One can use this method to determine the sample size at a desired CV, where CV is the coefficient of variation of the proportion parameter θ of a binomial distribution. Intuitively one can see that coefficient of variation is a function of sample size through the mean and the variance of random variable. We know that the coefficient of variation is defined as the ratio of standard deviation to mean. From general principles of sampling one knows that with the increase of sample size the standard deviation decreases which results in lower values of CV. Thus one may use CV as a criterion to identify an appropriate sample size.

Let Y be a random variable from a binomial distribution with parameters n and θ. Then we know that

$$E_\theta(Y/n) = \theta$$

and the standard error of (Y/n) is

$$SE_\theta(Y/n) = \sqrt{\theta(1-\theta)/n}\,.$$

Then we can define the coefficient of variation, C of Y/n, as

$$C = \frac{SE_\theta(Y/n)}{E_\theta(Y/n)}.$$

Then $$C = \frac{\sqrt{\theta(1-\theta)/n}}{\theta} = \sqrt{\frac{(1-\theta)}{n\theta}}. \tag{2.22}$$

Solving for n gives,

$$C^2 = \frac{1-\theta}{n\theta}$$

i.e., $$n = \frac{1-\theta}{\theta}\frac{1}{C^2}. \tag{2.23}$$

Thus the sample size n is a function of θ and C. By choosing a desired value of C for a particular θ, one can calculate the required sample size. For varying values of θ=0.1(0.1)0.9 and C=0.01(0.1)0.2 one could tabulate the values of sample size (see Table 4.2, pp.40).

2.2.4 Sequential Sampling Approach (Wald (1947))

In some practical situations such as life testing, clinical trials and destructive scientific experiments, sampling is both expensive and time consuming. Hence in these situations one may consider it to be more efficient to take samples sequentially rather than to take all at one time. The sampling procedure is terminated when a particular condition is met in the sample. That condition is usually called as a stopping rule. Hence one has to define a stopping rule to know when to terminate the sampling process.

In a single sampling situation, the entire sample is drawn at a single instance. Whereas in multisampling or sequential sampling, the samples are taken in successive stages based upon the results obtained from the previous sampling. Thus multistage or sequential sampling allows assessing the results at each stage and facilitates the possibility of stopping the process by reaching at an early decision. If the situation is clearly favorable or unfavorable (for example the quality of a particular lot is definitely good or bad), then terminating the sampling process usually saves both time and resources. One may continue with sampling process when the data obtained earlier is ambiguous, so that one can use additional information to take a better decision.

In order to apply sequential sampling procedure one needs to know the values of the parameters in the null and the alternative hypotheses respectively. Usually that's not the case. So he has to estimate the values from the data. Then based on the parameter values he can define a stopping rule. In the present situation, instead of considering the parameter values to define a stopping rule, one can use a measure of precision that is coefficient of variation, C. This method is described below.

In the present situation of seed health testing, seeds were selected at random, one after another (sequentially) and incidence (presence '1' or absence '0') of a particular pathogen was recorded as described earlier. Hence a total of j number of seeds was tested up to the j^{th} stage of sampling. Obviously here the outcome of each seed tested is a Bernoulli random variable with parameter θ.

Now let T_j be the cumulative number of diseased seeds observed up to the j^{th} stage of sampling. Clearly j and T_j are integers. Here $T_j \sim$ Binomial (j, θ).

Also let p_j be the proportion of diseased seeds based on T_j at the j^{th} stage of sampling defined as

$$p_j = \frac{T_j}{j}.$$

Then the expected value of p_j is given by

$$E_\theta(p_j) = \frac{E_\theta(T_j)}{j} = \frac{j\theta}{j} = \theta,$$

and similarly the standard error of p_j is given as,

$$SE_\theta(p_j) = \sqrt{V_\theta(p_j)} = \sqrt{V_\theta(T_j/j)} = \sqrt{\frac{1}{j^2}V(T_j)}\,.$$

Hence

$$SE_\theta(p_j) = \sqrt{\frac{1}{j^2}j\theta\theta(-\theta)} = \sqrt{\theta(1-\theta)/j}$$

Let the expected number of cumulative diseased seeds at the jth stage of sampling is defined as $E_\theta(T_j)$. We know that for binomial distribution,

$$E_\theta(T_j) = j\theta,$$

which can be written as,

$$E_\theta(T_j) = \frac{j}{\frac{\theta+(1-\theta)}{\theta}} = \frac{j}{1+\left(\frac{1-\theta}{\theta}\right)}$$

$$= \frac{j}{1+\frac{\theta(1-\theta)}{j}\frac{j}{\theta}\frac{1}{\theta}} = \frac{j}{1+\frac{[SE_\theta(p_j)]^2}{\theta^2}j}$$

where

$$SE(\theta E= \sqrt{\frac{\theta(1-\theta)}{j}}$$ was defined earlier.

Therefore

$$E_\theta(T_j) = \frac{j}{1+\left(\frac{SE_\theta(p_j)}{\theta}\right)^2 j},$$

and finally,

$$E_\theta(T_j) = \frac{j}{1+jC^2}, \tag{2.24}$$

where C is the coefficient of variation for θ (i.e., the pre-specified precision), which is defined as,

$$C = \frac{SE_\theta(p_j)}{\theta}.$$

Here $E_\theta(T_j)$ in (2.24) is a function of coefficient of variation C and j , which is free from sample proportion. By fixing the value of C at a desired level one can see that $E_\theta(T_j)$ converges as j increases. Now one can compare T_j and $E_\theta(T_j)$ at a desired level of precision (C), and can stop the sequential sampling process only when the observed number of diseased seeds T_j exceeds the expected number of diseased seeds $E_\theta(T_j)$, then compute θ as,

$$\theta = \frac{T_j}{j}.$$

Now one can plot the curves (j, $E_\theta(T_j)$) see Figure 4.1 (page 43), for different values of C=0.05(0.05)0.2. By observing the diagram one could notice that the curves stabilize rapidly with increasing in values of j. The curves become horizontal almost parallel to X-axis. Now one could select a point on a curve of specified precision (C), where the slope of the curve is reasonably small approximately equal to zero and the corresponding value on the horizontal axis will be the value for sample size n.

The above sequential method can be summarized as follows,

1. test each seed one after another and incidence of disease (presence:1 or absence:0) is noted.

2. At the end of each stage of testing (say the j^{th} stage), the cumulative number of diseased seeds T_j and the expected number of diseased seeds $E_\theta(T_j) = j/(1+jC^2)$ are computed at a desired level of precision C.

3. Then T_j and $E_\theta(T_j)$ are compared, and the following decision rule is applied,

   if $T_j > E_\theta(T_j)$ then stop the sampling process and set $\theta = T_j/j$.

   Otherwise continue the sampling process.

Note: To facilitate easy comparison, one could draw the curves (j, $E_\theta(T_j)$) for different pre-specified values of C. Then the observed T_j values are plotted against the curves. The decision to stop the sampling process is made at the first instance of $T_j > E_\theta(T_j)$.

4. Finally, the sample size n is the value on the horizontal axis (see Figure 4.1) corresponding to which the slope of the curve is negligible.

2.2.5 Poisson Procedure

Burstein (1971) explains the Poisson procedure based on the well-known fact that when $\theta \rightarrow 0$ and $n \rightarrow \infty$ and $\lambda = n\theta$ is finite then binomial distribution converges to a Poisson distribution. Here λ is the parameter of the Poisson distribution.

When the probability of success θ decreases, then the Poisson procedure becomes an increasingly accurate method of obtaining the sample size. The Poisson procedure is described below.

Assume that a sample of size n is drawn from an infinite population. Let θ be the probability of success of an event, where each event is a Bernoulli trial. Let Y be the number of such events in the sample. Then one can compute a value for observed proportion of events as Y/n, which can be considered to be an estimate of θ.

To determine the sample size n, one has to specify the maximum tolerable error $e = \bar{\theta} - \frac{Y}{n}$ at a desired level of confidence 1-α. Here

$$\bar{\theta} \approx \frac{\bar{\lambda}}{n + \frac{(\bar{\lambda} - Y)}{2}}, \tag{2.25}$$

is the upper confidence limit for θ calculated at confidence level (1-α)% given by Anderson and Burstein (1967). The upper confidence limit, $\bar{\lambda}$ for the parameter λ of Poisson distribution is given as

$$\bar{\lambda} = \frac{1}{2n} \chi^2_{2y+2,1-\alpha/2}, \tag{2.26}$$

which is derived using the mathematical relationship between Poisson distribution and chi-square distribution [see Evans, Hastings and Peacock (1993)]. The above relationship between the two distributions is expressed as

$$P[Y \leq y] = P[\chi^2 > 2\lambda\lambda] \tag{2.27}$$

where Y ~ Poisson(λ) and $\chi^2 \sim \chi^2_{2(1+y)}$ (chi-square distribution with degrees of freedom 2(y+1)).

Usually in real life one does not know the value of the proportion of events, but sometimes can anticipate it from the sample. This information may come from knowledge about the population under study, experience with similar populations and a pilot study etc. Let us define this proportion as the largest anticipated sample proportion and denote it by $\hat{\theta}$. Also denote the anticipated value of the upper confidence limit of θ, that is, $\bar{\theta}$ as $\hat{\bar{\theta}}$, which is computed as,

$$\hat{\bar{\theta}} = \hat{\theta} + e. \tag{2.28}$$

Now one can determine the sample size n by rearranging the terms in (2.25) as,

$$\bar{\theta}\left[n+\frac{(\bar{\lambda}-Y)}{2}\right]\approx\bar{\lambda},$$

then

$$\bar{\theta}[2n+\bar{\lambda}-Y]\approx 2\bar{\lambda},$$

that is

$$\bar{\theta}[2n-Y]\approx\bar{\lambda}[2-\bar{\theta}],$$

therefore it can be written as

$$n\bar{\theta}\left(2-\frac{Y}{n}\right)\approx\bar{\lambda}[2-\bar{\theta}].$$

It follows that

$$\bar{\lambda}\approx\frac{n\bar{\theta}\left(2-\frac{Y}{n}\right)}{(2-\bar{\theta})},$$

hence finally one gets,

$$\frac{\bar{\lambda}}{Y}\approx\frac{\bar{\theta}\left(2-\frac{Y}{n}\right)}{\left[\frac{Y}{n}(2-\bar{\theta})\right]}. \qquad (2.29)$$

Now by assuming the sample proportion $Y/n=\hat{\theta}$ and the upper confidence limit $\bar{\theta}=\hat{\bar{\theta}}$, then (2.29) will become

$$\frac{\bar{\lambda}}{Y}\approx\frac{\hat{\bar{\theta}}(2-\hat{\theta})}{[\hat{\theta}(2-\hat{\bar{\theta}})]}. \qquad (2.30)$$

We may now denote the left hand side and the right hand side of (2.30) separately by Q and $\hat{Q}$ respectively as

$$Q = \frac{\bar{\lambda}}{Y}, \tag{2.31}$$

$$\hat{Q} = \frac{\hat{\hat{\theta}}(2-\hat{\theta})}{[\hat{\theta}(2-\hat{\hat{\theta}})]}. \tag{2.32}$$

Now one may use Table.8 of Herman Burstein (1971) to determine the sample size that gives values for Y for various values of Q based on the relationship between Poisson and Chi-square distributions and at confidence level (1-α)%. Using the fact that Q $\approx \hat{Q}$, the sample size determination procedure is described as follows.

1. Guess the sample proportion $\hat{\theta}$ and the reasonable error e.
2. Calculate the anticipated upper confidence limit $\hat{\hat{\theta}}$ from (2.28),

   i.e., $\hat{\hat{\theta}} = \hat{\theta} + e$
3. Calculate $\hat{Q}$ from (2.32) as $\hat{Q} = \hat{\hat{\theta}}(2-\hat{\theta})/[\hat{\theta}(2-\hat{\hat{\theta}})]$
4. From Table 8 of Herman Burstein (1971) find Q nearest to $\hat{Q}$, for specified confidence value (1-α)%. Note that one need not have to know the value of $\bar{\lambda}$.
5. From the above Table 8 identify the value of Y corresponding to Q. Sometimes if necessary linearly interpolate between Q values to find Y as given below,

   $$Y = Y_a + \frac{(Y_b - Y_a)(Q_a - \hat{Q})}{(Q_a - Q_b)}$$

   where,

   Q_a is the bounding value larger than $\hat{Q}$,

   Q_b is the bounding value smaller than $\hat{Q}$,

Y_a is the value corresponding to Q_a and

Y_b is the value corresponding to Q_b.

6. Now calculate sample size n by using above Y and rearranging the terms in the known relation,

$$\frac{Y}{n} = \hat{\theta},$$

finally $$n = \frac{Y}{\hat{\theta}}.$$

Chapter 3

Simulation Approach

3.1 Simulation

Inferences based on just one sample of a particular size n could be misleading. Because this may not give an idea of the likely variation in the results had we drawn more samples of that size. Generally repeated samples provide an objective assessment of the degree of consistency and stability of results. One can conduct simulations through repeated sampling to find a lower optimal sample size.

As described by Chandra et al. (2001) a sample of size n that consistently does not reject $H_0 : p_j = \theta_j$, and θ_j is known (where p_j and θ_j are sample and population proportion values of incidence respectively of the j th pathogen), at a chosen level of significance α across all k-repeated samples is the safest sample size to use. The authors propose to use a measure of goodness-of-fit of the sample estimate to the population value in terms of χ^2 value for each of the k-repeated samples. They base their argument on the following reasons,

1. χ^2 measures the discrepancy between sample estimate and population value.
2. Additive property of χ^2 distribution, which states that the sum of χ^2 random variables follow a χ^2 distribution with corresponding sum of degrees of freedom.

Now for a sample of size n, they define the χ^2 measure for testing $H_0: p_j = \theta_j$, j=1, 2…d as,

$$\chi_j^2 = \frac{(p_j - \theta_j)^2}{\theta_j} \sim \chi^2_{(1)}$$

and
$$\chi^2 = \sum_{j=1}^{d} \chi_j^2 \sim \chi^2_{(d)},$$

at a level of significance α across all the d pathogens

Now for each sample of size n, there are k-repeated samples, which in turn result in k-observed χ^2 values. Under the null hypothesis $H_0: p_j = \theta_j$, for a given sample size n, these k-observed χ^2 values are iid random variables from the corresponding theoretical χ^2 distribution. This theoretical χ^2 distribution may provide a lower bound on the optimal sample size, if that exists.

To determine the optimal sample size, one can suitably choose a characteristic of the k-observed χ^2-values. Some possible characteristics are the maximum, 0.01-uq (uq: upper quantile above which there are 1% of observed values) 0.05-uq and a median of the observed distribution of the k values of χ^2. If one plots sample size n versus the chosen characteristic, a discernable pattern emerges with increasing values of n (see Figures 4.2 a-e). One expects that with increasing values of n the values of the chosen characteristic of χ^2 decrease. Because, generally with increase of sample size n, the sample proportion approaches the population proportion value. In other words the discrepancy between sample and population values decreases, which in turn results in lower values for χ^2.

Obviously the safest characteristic to use is the maximum, because it covers the maximum possible risk in terms of the largest possible discrepancy between the population and the sample values. However, theoretically the χ^2 can assume the maximum value to be infinity. Because of the randomness of the nature of the observed χ^2 values, the maximum χ^2 values show an erratic pattern, which they did, with increasing sample size n. The possibility for the erratic pattern could be explained as follows. In drawing k-repeated samples of size n, there is a possibility of drawing such a sample for which the sample proportion value could be either abnormally high or low, which could be an outlier. Though the probability of drawing such a sample is very small but is possible in practical situations. This single value of proportion coming from a sample, which is by nature a random variable, will inflate the max-χ^2 value. Thus the maximum χ^2 value could become an outlier and basing decision on this maximum χ^2 value could be misleading. That situation will make it difficult to clearly identify an optimal sample size.

However on the other hand, use of median, compared to using either the observed 0.05-uq or 0.01-uq χ^2 values covers much less risk. So, one needs to choose a suitable characteristic, which covers reasonable amount of risk, keeping in mind that moving towards maximum increases the risk of erratic-ness. Here the authors compared the values of χ^2 for minimum, 0.95-uq, 0.75-uq, 0.5-uq, 0.25-uq, 0.05-uq and maximum along with the P-values in their study of genetic relationships.

The authors have considered the use of 0.05-uq χ^2 value to be adequate to determine an optimal sample size. Since theoretically 0.05-uq is the 95th percentile point, above which the area under the theoretical distribution is 0.05. That is, the event of getting a χ^2 value above 0.05-uq has a risk of 5%. Thus any 0.05-uq χ^2 value that is non-significant at a chosen level of significance α, implies that, for the corresponding sample size, all samples of that size will consistently deliver non-significant χ^2 values 95% of the times, hence provide a good fit to the population. Also the more the P-value of the observed 0.05-uq χ^2 value exceeds the specified α, the less is the discrepancy between population and sample values.

From this perspective, one could chose an α-value other than the conventional values of 0.05 and 0.01 to further minimize the risk of selecting an inappropriate sample size. The upper quantile χ^2 values and the corresponding P-values can be summarized in a tabular or graphical form, provide an objective probabilistic basis to select a suitable sample size.

The following is the methodology of simulation approach adopted based on the above discussion.

1. Here we consider N=1050 seeds as large enough to be a population.
2. We have drawn k=5000 independent random samples, each of size n=10(10) 50(25)200(100) 500, using a simple random sampling without replacement (SRSWOR), from a total of N seeds.
3. For each of the k=5000 repeated samples of a particular sample size n, the pathogen incidence rate p_j, j=1...d is estimated.
4. Also the goodness-of-fit of the sample estimate to the population value is calculated in terms of χ^2 value as described above for each of the k=5000 samples.
5. Thus we have 5000 χ^2 values for each sample of size n. By definition they follow a theoretical χ^2 distribution. Since each χ^2 value measures the

discrepancy between sample and population proportion values, defined as chi-square measure of goodness of fit follows a χ^2 distribution.

6. For each sample of size n out of the k observed χ^2 values we identified minimum, 0.95-uq, 0.75-uq, 0.5-uq, 0.25-uq, 0.05-uq and maximum (uq: represents upper quantile) along with the corresponding P-values. They can be plotted and also can be expressed in a tabular form. See tables 4.5 a-e and Figure 4.3
7. We used 0.05-uq χ^2 values to determine the optimum sample size as follows. The sample size at which the 0.05-uq χ^2 values become non-significant at a chosen level of significance across all the seed varieties is the optimum sample size.
8. Further we have used Kolmogrov-Smirnov (K-S) test (Sokal and Rohlf (1981)) to test the goodness of fit of observed k χ^2 values with the corresponding theoretical χ^2 distributions for each sample of size n.

   The K-S test statistic D is defined as,

$$D = \underset{1 \le i \le k}{\mathrm{Max}} \left| F(\cdot) - \frac{i}{k} \right|,$$

   where $F(\cdot)$ is the theoretical cumulative distribution of the distribution being tested. In this case we are testing χ^2 distribution.

Chapter 4

Results

4.1 Results of Sampling from Binomial Distribution

The sample sizes for different pre-specified values of α, e and θ are summarized in Table 4.1 (pp.38) below.

An examination of the results in the Table 4.1 shows that for any given values of α and e, the sample sizes increase with increase in proportion value θ, reaching maximum at θ=0.5. After that the sample sizes progressively decrease to zero. So one may consider θ=0.5 to get a conservative sample size. Therefore for a sample to adequately represent the population with respect to any pathogen with incidence θ, ($0 \leq \theta \leq 1$) at e=0.05 and α=0.05, a sample of size about 384 randomly selected seeds is required. In practical situations the actual sample size required will be less than this conservative sample size.

Similarly if one considers α=0.01 and e=0.05, then he can observe from the table that the conservative sample size is 666 seeds.

4.2 Results of Power Analysis

Sample sizes for different pre-specified values of θ, α, β and e are summarized in Table 4.2 (pp.40). An Examination of the results in Table 4.2 shows that for any given values of α, β and e, the sample size values increase as θ increases, reaching maximum at θ =0.5. Beyond θ =0.5 the sample size values progressively decrease to zero. So, to identify a conservative sample size, again one needs to concentrate on θ =0.5. Here also one has to be careful in selecting the values of α, β and e.

Table 4.1 Results of sampling from Binomial distribution in estimation of θ.

θ	α			
	0.05		0.01	
	e		e	
	0.01	0.05	0.01	0.05
0.00	0	0	0	0
0.05	1825	73	3162	126
0.10	3457	138	5991	240
0.15	4898	196	8487	339
0.20	6147	246	10650	426
0.25	7203	288	12481	499
0.30	8067	323	13978	559
0.35	8740	350	15143	606
0.40	9220	369	15975	639
0.45	9508	380	16475	659
0.50	9604	**384**	16641	**666**
0.55	9508	380	16475	659
0.60	9220	369	15975	639
0.65	8740	350	15143	606
0.70	8067	323	13978	559
0.75	7203	288	12481	499
0.80	6147	246	10650	426
0.85	4898	196	8487	339
0.90	3457	138	5991	240
0.95	1825	73	3162	126
1.00	0	0	0	0

Suppose one selects the trio α, β and e to be 0.05, 0.15 and 0.05 respectively, then the conservative sample size needed is identified from the table as 1156 seeds. If β is changed to 0.1 keeping other parameters constant, then the minimum sample size needed will be 1299 seeds.

4.3 Results of Coefficient of Variation Method

The sample sizes for different pre-specified values of θ and coefficient of variation, C, are tabulated in Table 4.3 (pp.41).

An examination of results in Table 4.3 shows that the sample size values for a given C decrease as sample proportion θ increases. Theoretically the sample size approaches to infinity when θ=0. So, one has to be very careful in anticipating the value of the proportion θ. That knowledge may come from experience with similar studies or pilot samples etc. Suppose one chooses θ=0.1 and C to be either 15% or 20%, then from the table one could get the corresponding sample sizes required to be equal to 400 and 225 seeds respectively.

4.4 Results of Sequential Sampling

Results of the sequential approach to determine the sample size are shown in Figure.4.1 (pp.43). For different pre-specified values of coefficient of variation, C the points (j, $E_\theta(T_j)$) are plotted and are joined with smooth curves.

An examination of the diagram shows that the curves stabilize with increasing values of j, that is their slopes converge to zero. Hence the curves become almost parallel to X-axis. Suppose one wants to find a sample size for C=0.1 (i.e., at 10%), he can observe from the Figure 4.1 that the corresponding curve stabilizes at about j=500. So for this case the minimal sample size needed is 500 seeds for any proportion value θ.

Similarly if C is changed to 0.2 (i.e., 20%) then the minimal sample size needed is 300 seeds.

Table 4.2 Results of power analysis

α	β	e	θ										
			0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.01	0.05	0.01	0	18516	32917	43204	49376	51434	49376	43204	32917	18516	0
		0.05	0	741	1317	1728	1975	2057	1975	1728	1317	741	0
	0.1	0.01	0	16033	28503	37410	42754	44535	42754	37410	28503	16033	0
		0.05	0	641	1140	1496	1710	1781	1710	1496	1140	641	0
	0.15	0.01	0	14511	25797	33859	38695	40308	38695	33859	25797	14511	0
		0.05	0	580	1032	1354	1548	1612	1548	1354	1032	580	0
	0.2	0.01	0	13391	23807	31247	35711	37198	35711	31247	23807	13391	0
		0.05	0	536	952	1250	1428	1488	1428	1250	952	536	0
0.05	0.05	0.01	0	13829	24585	32268	36878	38415	36878	32268	24585	13829	0
		0.05	0	553	983	1291	1475	1537	1475	1291	983	553	0
	0.1	0.01	0	11695	20792	27289	31187	32487	31187	27289	20792	11695	0
		0.05	0	468	832	1092	1247	1299	1247	1092	832	468	0
	0.15	0.01	0	10401	18491	24269	27736	28891	27736	24269	18491	10401	0
		0.05	0	416	740	971	1109	1156	1109	971	740	416	0
	0.2	0.01	0	9457	16812	22066	25218	26269	25218	22066	16812	9457	0
		0.05	0	378	672	883	1009	1051	1009	883	672	378	0

Table 4.3 Sample sizes for Coefficient of variation method

θ	C				
	0.01	0.05	0.10	0.15	0.20
0.05	190000	7600	1900	844	475
0.10	90000	3600	900	400	225
0.15	56667	2267	567	252	142
0.20	40000	1600	400	178	100
0.25	30000	1200	300	133	75
0.30	23333	933	233	104	58
0.35	18571	743	186	83	46
0.40	15000	600	150	67	38
0.45	12222	489	122	54	31
0.50	10000	400	100	44	25
0.55	8182	327	82	36	20
0.60	6667	267	67	30	17
0.65	5385	215	54	24	13
0.70	4286	171	43	19	11
0.75	3333	133	33	15	8
0.80	2500	100	25	11	6
0.85	1765	71	18	8	4
0.90	1111	44	11	5	3
0.95	526	21	5	2	1
1.00	0	0	0	0	0

4.5 Results of Poisson Procedure

We will first discuss a simple example of calculating sample size before presenting the results. In this method one should supply the values of $\hat{\theta}$, e and α.

Example: Let $\hat{\theta}$=0.25, e=0.05 and α=0.05.

1. Calculate $\hat{\hat{\theta}} = \hat{\theta} + e$

$$= 0.25+0.05 = 0.03.$$

2. $$\hat{Q} = \frac{\hat{\hat{\theta}}(2-\hat{\theta})}{\hat{\theta}(2-\hat{\hat{\theta}})} = \frac{0.30\ (2-0.25)}{0.25\ (2-0.30)} = 1.235\,.$$

3. For $\hat{Q}$=1.235 and 1-α=95% the bounding values of Q in Table 8 of Herman Burstein (1971) are 1.245 and 1.229 with corresponding Y=80 and 90 respectively.

That is,

$$Q_a=1.245, \qquad Y_a=80,$$

$$Q_b=1.229, \qquad Y_b=90,$$

$$\hat{Q}=1.235,$$

then using interpolation we get

$$Y = 80 + \frac{(90-80)(1.245-1.235)}{(1.245-1.229)} = 86 \text{ (to nearest integer).}$$

4. Now the sample size n can be computed as,

$$n = \frac{Y}{\hat{\theta}} = \frac{86}{0.25} = 344\,.$$

The sample size values are summarized in Table 4.4 by applying the above procedure for different pre-specified values of $\hat{\theta}$=0.1(0.1)0.9 , e=0.01 and 0.05, and α=0.01 and 0.05 respectively.

Figure. 4.1 curves of (j, $E_\theta(T_j)$) for d

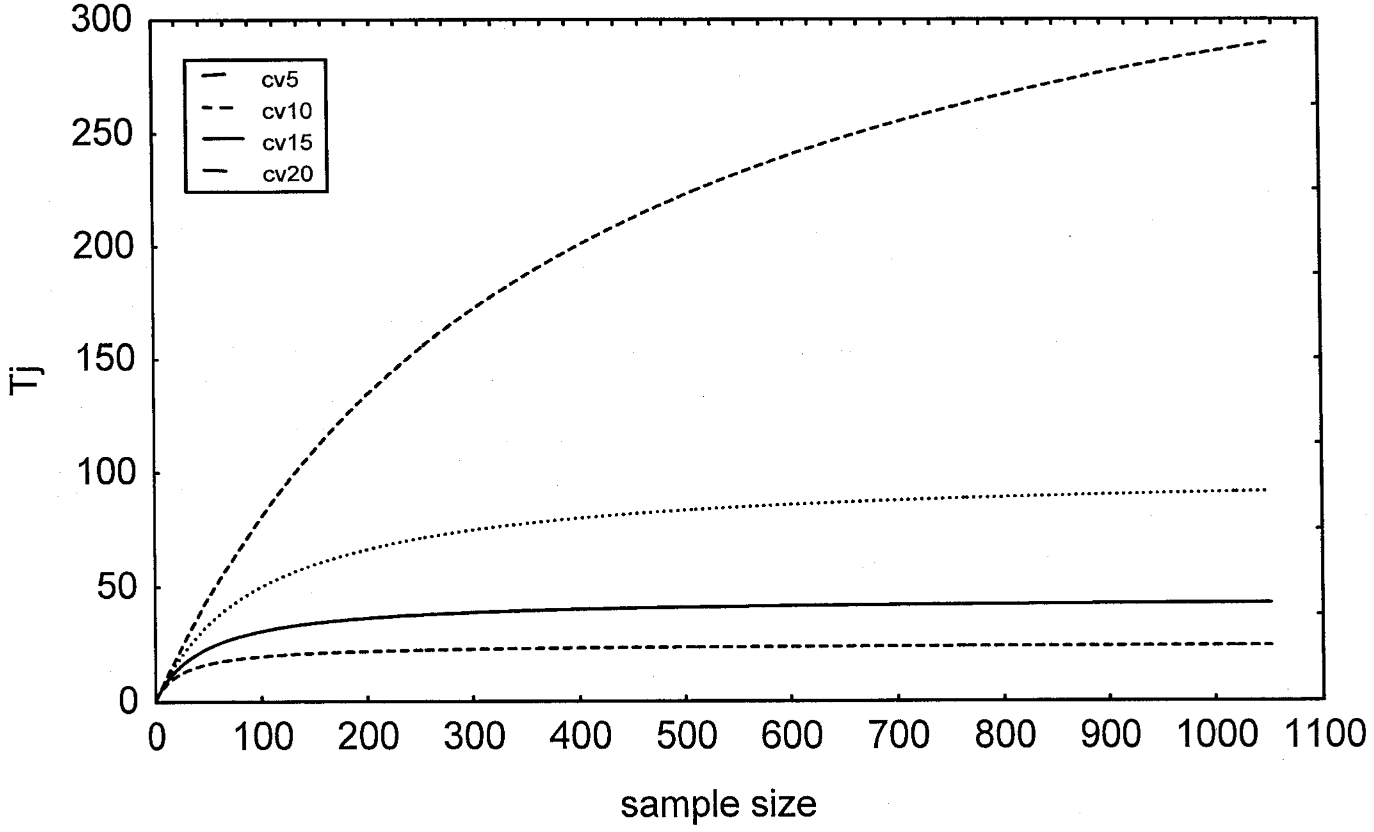


Table 4.4 Results of Poisson procedure

$\hat{\theta}$	e=0.05						e=0.01					
	$\hat{\hat{\theta}}=\hat{\theta}$ +e	$\hat{Q}$	α=0.05		α= 0.01		$\hat{\hat{\theta}}=\hat{\theta}$+e	$\hat{Q}$	α= 0.05		α= 0.01	
			Y	N	Y	n			Y	n	Y	n
0.1	0.15	1.541	20	200	33	330	0.11	1.106	385	3850	650	6500
0.2	0.25	1.286	60	300	100	500	0.21	1.056	1200	6000	2000	10000
0.3	0.35	1.202	110	367	192	640	0.31	1.039	2625	8750	4500	15000
0.4	0.45	1.161	170	425	292	730	0.41	1.031	4200	9000	7000	17500
0.5	0.55	1.138	235	470	390	780	0.51	1.027	6000	12000	9500	19000
0.6	0.65	1.161	170	425	292	730	0.61	1.031	4200	9000	7000	17500
0.7	0.75	1.202	110	367	192	640	0.71	1.039	2625	8750	4500	15000
0.8	0.85	1.286	60	300	100	500	0.81	1.056	1200	6000	2000	10000
0.9	0.95	1.541	20	200	33	330	0.91	1.106	385	3850	650	6500

4.6 Results of Simulation Method

The upper cumulative frequency distributions of the 5000 observed χ^2 values according to sample sizes n for different seed varieties are summarized in Tables 4.5a-4.5e (pp.38). As expected from the law of large numbers, the χ^2 values show generally a decreasing trend as the sample size increases. One can observe from Figure 4.2a-4.2e (pp.48) that the maximum χ^2 values show an erratic pattern as n increases, whereas the 0.05-uq χ^2 values follow a smooth decreasing curve. So in this study we consider the 0.05-uq χ^2 values to determine the sample size. Figure 4.3 (pp.51) graphically depicts the observed 0.05-uq χ^2 values and their corresponding P-values for different seed varieties. Closely observing Figure 4.3, one can see that the values of the 0.05-uq χ^2 values follow a smooth decreasing curve and stabilize as n increases. One can also see that the observed χ^2 values become non significant after certain n. An examination of Tables 4.5a-4.5e shows that the D (K-S statistic) value attains its minimum value at n=175 across all the seed varieties. It is significant with P<0.01. At this sample size the 0.05-uq χ^2 values become non significant across all the seed varieties. Thus this sample size n=175 seeds may serve as a lower bound on optimal n. So one may consider n=175 seeds as an adequate sample size to represent the population across all the sorghum varieties.

However, for α=0.05, the corresponding 0.05-uq χ^2 values for the five sorghum varieties are significant at n=100, 100, 125, 125, 125 having P-values equal to 0.0447, 0.0379, 0.0495, 0.0496 and 0.0402 respectively. But for the same α=0.05, at n=125, 125, 150, 150, 150, the P-values are 0.0671, 0.0508, 0.0666, 0.0631 and 0.0576 respectively, which exceed the α =0.05 and become non-significant for the first time. In real life situations one can choose a desired level of significance and accordingly he can select an appropriate sample size.

4.7 Conclusion and Discussion

Sample size plays an important role in any statistical investigation. Sometimes data collection could be very expensive and time consuming. So one wants to know it before hand to properly plan his resources to collect data. Especially in the case of Bernoulli outcomes one needs to be extra cautious in anticipating the proportion of success. Otherwise sample size determined may result in over estimation or under estimation.

Here we have presented and compared five different sample size determination methods along with the simulation method. The underlying assumptions for all the five methods are as follows,

1. the probability distribution under null hypothesis is normal, and
2. the inferences are based upon the large sample assumption.

If the proportion of success is very small then the normal approximation may not be suitable to identify sample size. So one may use Poisson procedure in such a case.

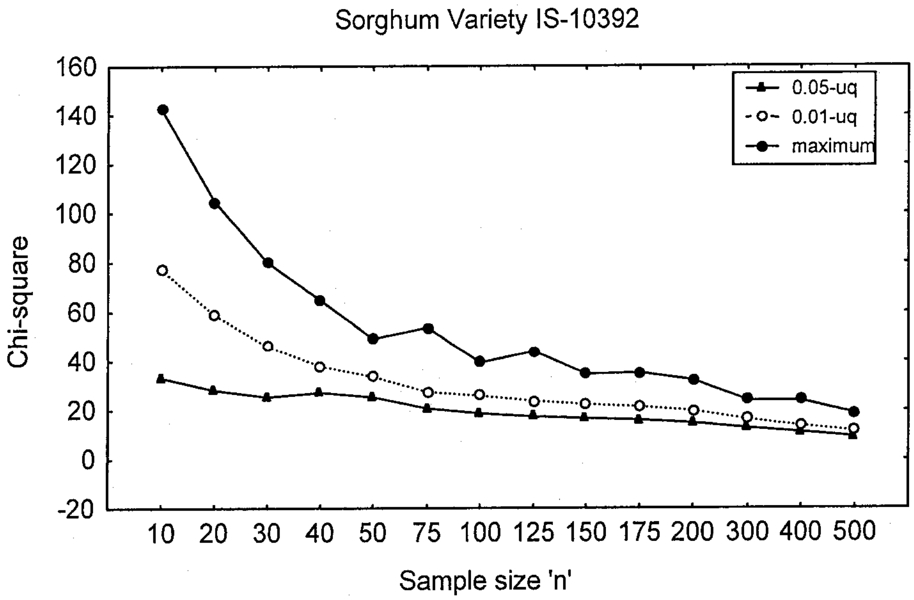
In all the five methods the sample size needed for a particular proportion of incidence (for a pathogen) is reported. But when we consider all the pathogens together then the level of significance would need to be modified according to Bonferroni correction. Then recalculate the sample size, which would be much higher than the sample sizes calculated for individual pathogens, because the new level of significance will be much lower than the earlier α.

In the simulation method, we used the additive property of χ^2 distributions. The resulting random variable again follows a χ^2 distribution. In other words, the individual

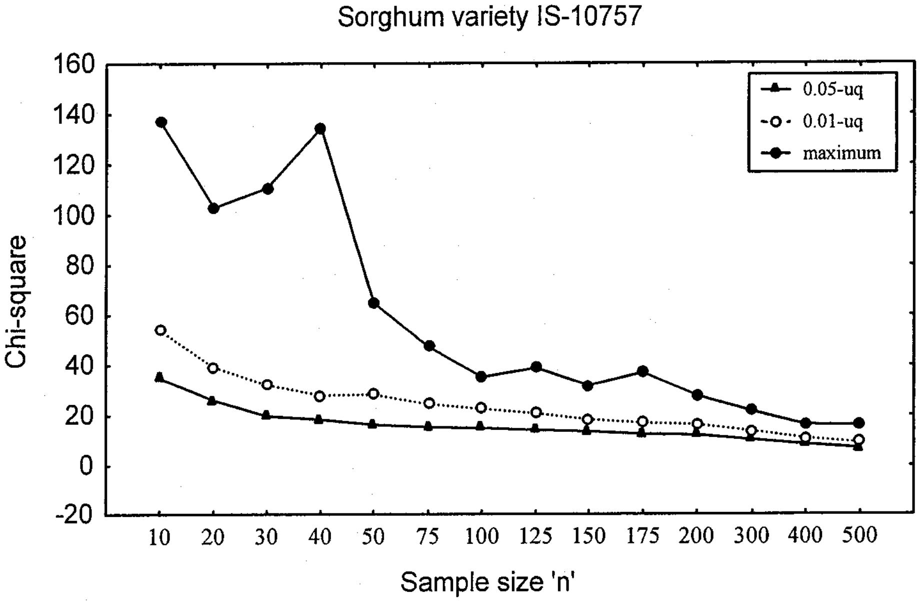
χ^2 measures of different pathogens were added together to form a pooled χ^2 value. This pooled value represents the information across all the pathogens, from which the inferences are drawn. The simulation method gives sample sizes across all the pathogens for different seed varieties as 125,125,150,150 and 150 respectively. These sample sizes from simulation method are much smaller than those were predicted by the other methods. The results from a simulation study are based on a particular data set and cannot be generalized but can serve as a tool and guide an experimenter in identifying a lower bound on the optimal sample size.

Figure 4.2 Comparison of maximum, 0.01-uq and 0.05-uq χ^2 values

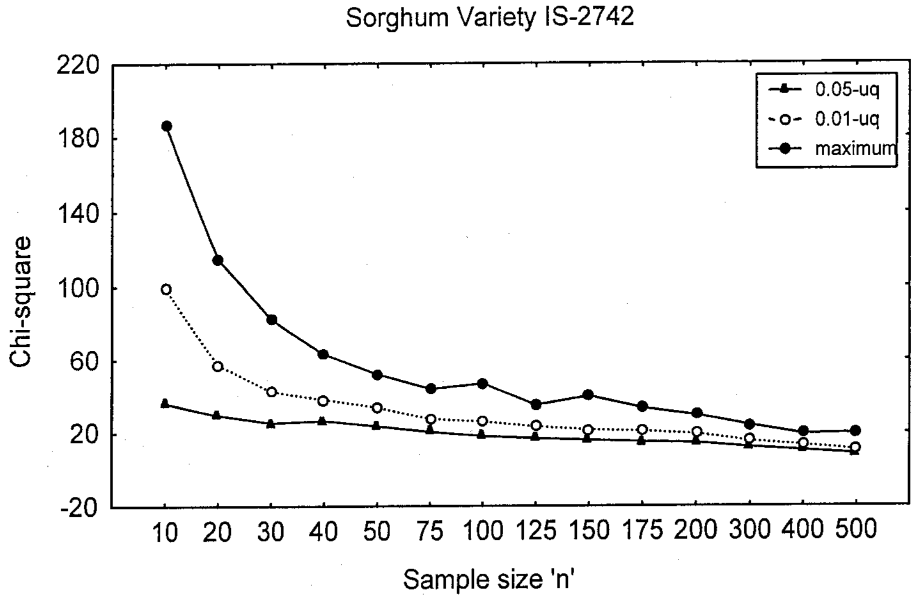
(a)



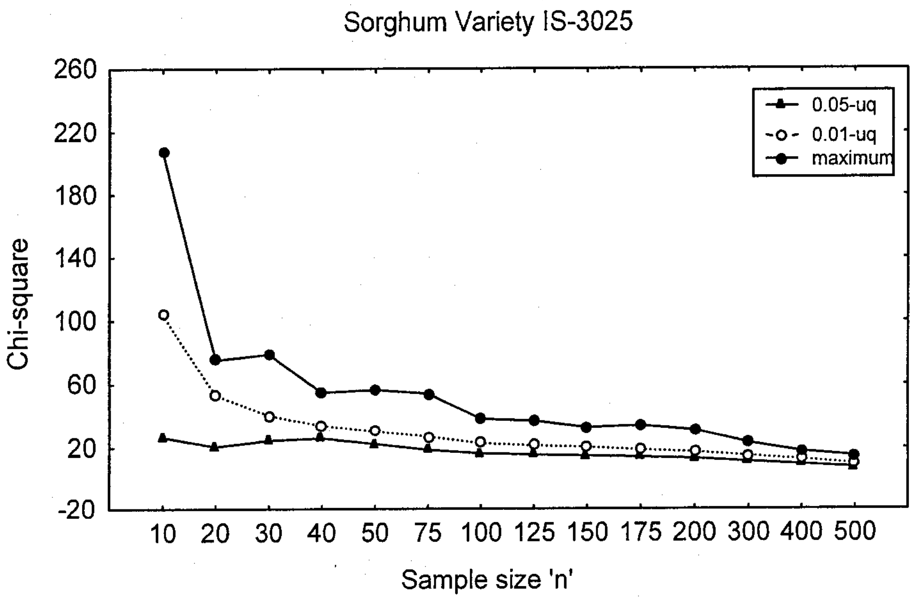
(b)



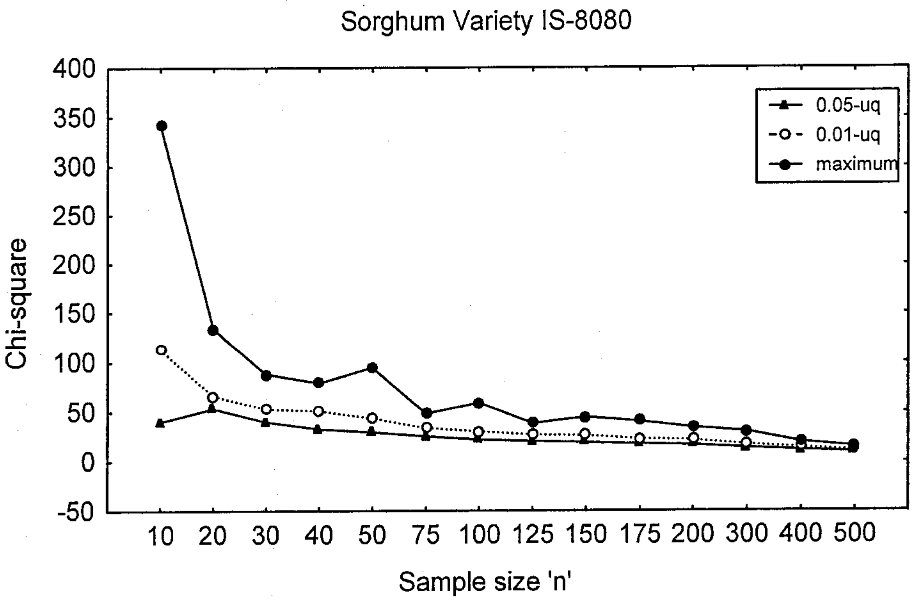
(c)


Sorghum Variety IS-2742
220
180
140
100
60
20
-20
Chi-square
0.05-uq
0.01-uq
maximum
10
20
30
40
50
75
100
125
150
175
200
300
400
500
Sample size 'n'

(d)


Sorghum Variety IS-3025
260
220
180
140
100
60
20
-20
Chi-square
0.05-uq
0.01-uq
maximum
10
20
30
40
50
75
100
125
150
175
200
300
400
500
Sample size 'n'

(e)



Sorghum Variety IS-8080
Chi-square
400
350
300
250
200
150
100
50
0
-50
0.05-uq
0.01-uq
maximum
10
20
30
40
50
75
100
125
150
175
200
300
400
500
Sample size 'n'

Figure 4.3 Comparison of 0.05-uq χ^2 values and P-values across different seed varieties.

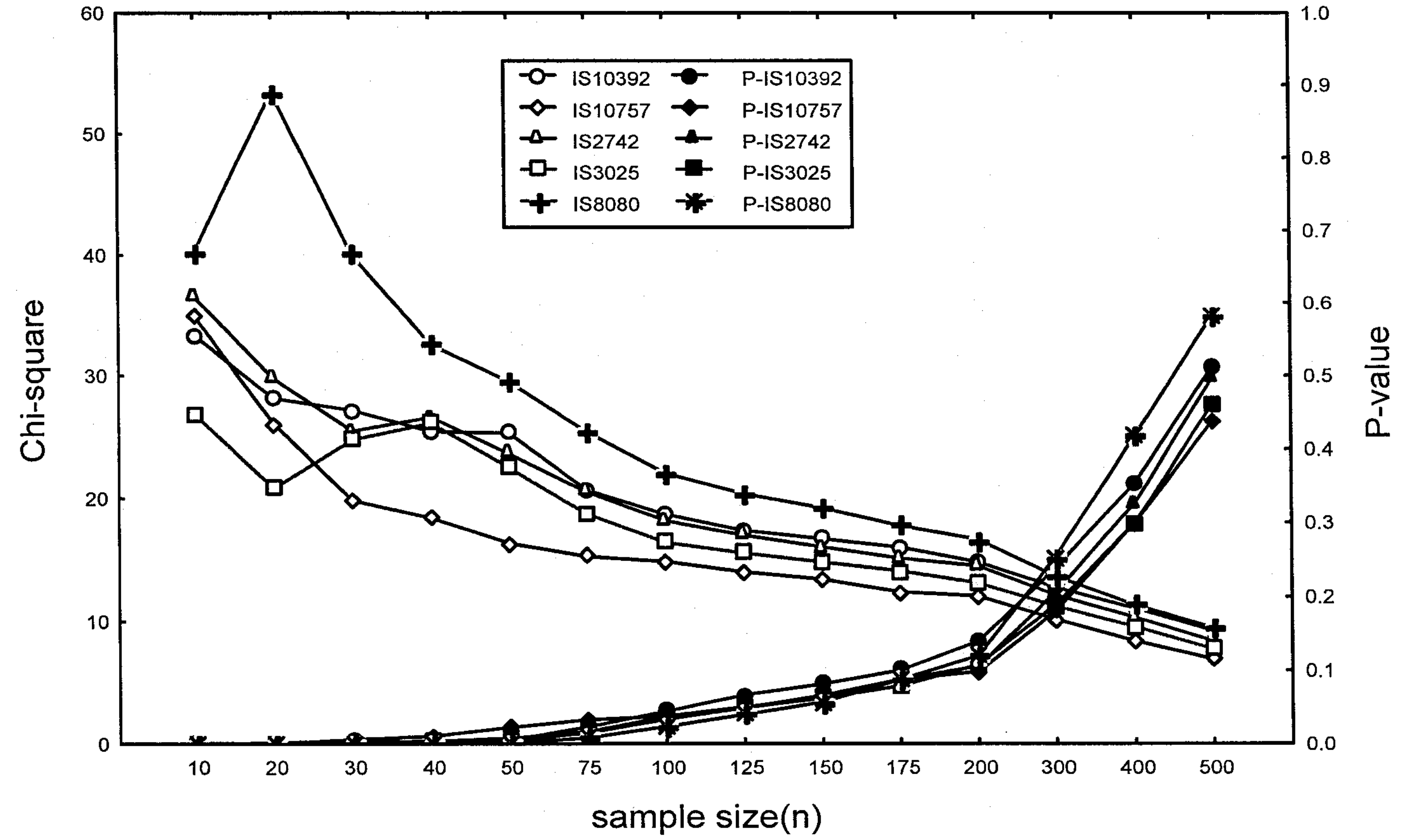


Table 4.5a Quantiles of observed 5000 χ^2 values of seed variety IS-10392[1]

n	min	0.95-uq	0.75-uq	0.50-uq	0.25-uq	0.05-uq	0.01-uq	max	$^{@}$KS-D
10	0.96 0.9999	1.06 0.9998	2.41 0.9921	5.21 0.8767	12.41 0.2586	33.29 0.0002	77.21 0	142.46 0	0.4164
20	1.01 0.9998	1.79 0.9977	3.58 0.9643	6.78 0.746	11.41 0.3265	28.16 0.0017	58.83 0	104.64 0	0.263
30	1.63 0.9985	2.23 0.9943	4.5 0.922	6.89 0.7358	11.68 0.307	25.42 0.0046	46.07 0	80.03 0	0.239
40	1.55 0.9988	2.87 0.9843	4.85 0.901	7.18 0.7083	12.24 0.2693	27.18 0.0024	37.93 0	64.71 0	0.2088
50	1.8 0.99766	3.22 0.97576	4.93 0.8958	7.44 0.68336	11.8 0.29866	25.29 0.00482	33.68 0.00021	49.12 0	0.1958
75	1.64 0.99843	3.22 0.97576	5.28 0.87171	7.7 0.65811	11.76 0.30144	20.71 0.02321	27.39 0.00226	53.6 0	0.165
100	1.24 0.99954	3.24 0.9752	5.3 0.87026	7.45 0.68239	11.39 0.32795	18.67 *0.04466*	25.86 0.00393	39.36 0.00002	0.1856
125	0.95 0.99986	3.42 0.96975	5.44 0.85992	7.69 0.65909	11.35 0.33091	17.34 *0.06717*	23.74 0.00832	43.89 0	0.1634
150	0.96 0.99986	3.26 0.97463	5.36 0.86587	7.68 0.66006	10.87 0.36774	16.66 0.08223	22.36 0.01337	34.68 0.00014	0.1612
175	1.17 0.99965	3.24 0.9752	5.4 0.86291	7.76 0.65227	10.62 0.38788	15.97 0.10049	21.39 0.01853	35.18 0.00012	0.158
200	1.2 0.99961	3.18 0.97686	5.36 0.86587	7.55 0.6727	10.03 0.43787	14.8 0.13953	19.78 0.0314	32.25 0.00036	0.196
300	1.13 0.9997	2.99 0.9817	4.79 0.9048	6.62 0.7608	8.89 0.5426	12.74 0.2386	16.17 0.0949	23.94 0.0078	0.2972
400	0.99 0.9998	2.7 0.9876	4.32 0.9318	5.83 0.8293	7.7 0.6581	11.05 0.3536	13.69 0.1876	23.9 0.0079	0.4124
500	1.09 0.9997	2.27 0.9938	3.64 0.9621	4.94 0.8951	6.51 0.7708	9.21 0.5123	11.48 0.3214	18.72 0.044	0.5418

Number of pathogens identified is10

[1] For each sample size and uq the χ^2 value and the corresponding P-value are provided.

$^{@}$ Kolmogrov -Smirnov statistic D value with P<0..001

Table 4.5b Quantiles of observed 5000 χ^2 values of seed variety IS-10757

N	min	0.95-uq	0.75-uq	0.50-uq	0.25-uq	0.05-uq	0.01-uq	max	$^{@}$KS-D
10	0.93 0.99588	0.93 0.99588	0.99 0.99499	1.88 0.96618	7.34 0.39436	34.86 0.00001	54.07 0	137.39 0	0.4964
20	0.66 0.99862	0.84 0.99701	1.93 0.96361	3.76 0.80697	8.3 0.30689	26.09 0.00049	39.27 0	102.98 0	0.3332
30	1.15 0.99203	1.21 0.99069	2.57 0.92173	4.03 0.77632	8.54 0.28739	19.83 0.00595	32.16 0.00004	110.79 0	0.2974
40	1.19 0.99116	1.46 0.98366	2.43 0.93228	4.81 0.68314	8.92 0.25845	18.35 0.01049	28 0.00022	134.58 0	0.222
50	1.21 0.99069	1.66 0.97625	2.76 0.90628	4.85 0.67826	8.89 0.26065	16.3 0.02251	28.3 0.00019	64.85 0	0.184
75	0.88 0.99654	1.8 0.97008	3.26 0.85995	5.23 0.63192	8.21 0.31444	15.25 0.03293	24.43 0.00096	47.27 0	0.1438
100	1.06 0.9938	1.9 0.96517	3.52 0.8331	5.43 0.60764	8.02 0.33083	14.85 0.03797	22.54 0.00205	35.2 0.00001	0.1152
125	0.98 0.99514	1.95 0.96256	3.61 0.82344	5.31 0.62219	7.74 0.35608	14.02 *0.05083*	20.72 0.00421	38.73 0	0.1316
150	0.97 0.9953	2.05 0.95702	3.64 0.82018	5.15 0.64166	7.48 0.38067	13.42 0.06251	18.27 0.01081	31.79 0.00004	0.1552
175	1.02 0.9945	2.13 0.9523	3.52 0.8331	5.03 0.6563	7.29 0.39932	12.4 0.08815	16.95 0.01772	36.92 0	0.1722
200	0.79 0.99754	2.04 0.95759	3.42 0.84363	4.94 0.66729	7.26 0.40232	12.06 0.0986	16.25 0.02293	27.81 0.00024	0.1732
300	0.52 0.99937	1.77 0.97147	3.13 0.87272	4.48 0.72312	6.35 0.49953	10.15 0.18023	13.59 0.05897	21.59 0.00299	0.2546
400	0.19 0.99998	1.47 0.98333	2.78 0.90458	4.02 0.77747	5.51 0.59798	8.39 0.29946	10.85 0.1453	16.26 0.02284	0.3496
500	0.29 0.99991	1.18 0.99138	2.35 0.93796	3.35 0.85085	4.67 0.70016	6.91 0.43831	9.08 0.24696	16.13 0.02395	0.4578

Number of pathogens identified is 10

Table 4.5c Quantiles of observed 5000 χ^2 values of seed variety IS-2742

N	min	0.95-uq	0.75-uq	0.50-uq	0.25-uq	0.05-uq	0.01-uq	max	$^{@}$KS-D
10	0.41 0.99999	0.59 0.99994	1.65 0.99587	3.27 0.95262	9.12 0.42627	36.45 0.00003	99.43 0	186.94 0	0.516
20	0.71 0.99986	0.83 0.99974	2.22 0.98749	4.75 0.85553	11.77 0.22659	29.8 0.00047	57.35 0	114.8 0	0.367
30	0.75 0.99983	1.17 0.99894	2.73 0.97405	5.4 0.79814	10.85 0.28613	25.46 0.0025	42.69 0	82.52 0	0.3178
40	0.97 0.9995	1.53 0.99692	3.14 0.95848	6.17 0.72279	10.86 0.28543	26.59 0.00163	38.33 0.00002	63.18 0	0.251
50	1.24 0.99866	1.76 0.99472	3.68 0.93118	6.52 0.68696	10.93 0.28054	23.62 0.00494	34.08 0.00009	52.21 0	0.2184
75	1.22 0.99874	2.21 0.98769	4.13 0.90265	6.59 0.67972	10.9 0.28263	20.54 0.01486	27.71 0.00107	44.12 0	0.198
100	1.48 0.99729	2.58 0.97865	4.27 0.89276	6.62 0.67662	10.62 0.30266	18.15 0.03347	26.13 0.00195	47.02 0	0.181
125	1.55 0.99676	2.64 0.97688	4.38 0.88467	6.66 0.67247	10.19 0.33532	16.95 *0.0495*	23.34 0.00548	35.17 0.00006	0.1782
150	1.48 0.99729	2.75 0.97339	4.53 0.87321	6.77 0.66105	9.96 0.35373	16.01 0.06667	21.62 0.01016	40.07 0.00001	0.1684
175	1.33 0.99822	2.84 0.9703	4.54 0.87243	6.78 0.66001	9.54 0.38899	15.1 0.08823	20.89 0.01315	33.62 0.0001	0.1642
200	1.29 0.99843	2.76 0.97306	4.59 0.86849	6.59 0.67972	9.18 0.42083	14.49 0.10593	19.39 0.02207	29.8 0.00047	0.1916
300	1.08 0.99923	2.63 0.97718	4.37 0.88542	5.96 0.74392	8.03 0.53113	12.13 0.20608	15.43 0.07978	24.08 0.00418	0.2826
400	0.71 0.99986	2.36 0.98441	3.84 0.92162	5.21 0.81563	6.91 0.64649	10.31 0.32598	13.18 0.15463	19.31 0.02268	0.3994
500	0.79 0.99979	1.97 0.99193	3.21 0.95538	4.35 0.8869	5.78 0.76172	8.37 0.49733	10.68 0.29828	19.84 0.01893	0.521

Number of pathogens identified is 9

Table 4.5d Quantiles of observed 5000 χ^2 values of seed variety IS-3025.

n	min	0.95-uq	0.75-uq	0.50-uq	0.25-uq	0.05-uq	0.01-uq	max	$^{@}$KS-D
10	0.49 0.99988	0.83 0.99911	1.63 0.99033	3.38 0.9083	10.03 0.26293	26.78 0.00077	103.94 0	207.06 0	0.4744
20	0.82 0.99915	1.02 0.99812	2.67 0.95333	5.11 0.74576	9.87 0.27427	20.86 0.00753	53.18 0	75.16 0	0.2662
30	0.94 0.9986	1.38 0.99453	3.17 0.92324	5.43 0.71078	9.53 0.29957	24.76 0.00171	40.32 0	78.75 0	0.232
40	1.16 0.99702	1.81 0.98629	3.27 0.91629	5.75 0.67521	9.16 0.32898	26.1 0.00101	34.08 0.00004	54.83 0	0.2058
50	1.47 0.99319	1.96 0.98221	3.75 0.87895	5.97 0.65059	9.25 0.32165	22.52 0.00404	30.45 0.00018	56.47 0	0.16
75	1.05 0.99791	2.31 0.97	3.95 0.86161	5.85 0.66403	9.68 0.28821	18.7 0.01655	27.11 0.00068	54.07 0	0.1688
100	1 0.99825	2.31 0.97	3.87 0.86866	5.87 0.66179	9.58 0.29575	16.37 0.03738	22.84 0.00358	38.19 0.00001	0.1746
125	0.93 0.99865	2.3 0.97041	3.9 0.86603	6.09 0.63715	9.52 0.30034	15.53 *0.04962*	21.37 0.00623	37 0.00001	0.1412
150	0.8 0.99922	2.17 0.97535	4.05 0.85258	6.17 0.6282	9.05 0.3381	14.8 0.06315	20.33 0.00916	32.8 0.00007	0.135
175	0.62 0.9997	2.2 0.97426	3.99 0.85802	5.94 0.65395	8.61 0.37626	14.09 0.07945	18.63 0.01697	34.09 0.00004	0.1546
200	0.45 0.99991	2.07 0.97879	3.94 0.8625	5.94 0.65395	8.46 0.38987	13.11 0.10812	17.32 0.02694	31.16 0.00013	0.1608
300	0.52 0.99985	2.17 0.97535	3.82 0.87299	5.35 0.7196	7.37 0.49729	11.23 0.189	14.84 0.06233	23.18 0.00314	0.2534
400	0.72 0.99947	1.91 0.98365	3.23 0.91911	4.57 0.80239	6.25 0.61925	9.54 0.29881	12.6 0.12637	17.76 0.0231	0.3698
500	0.59 0.99975	1.64 0.99013	2.77 0.94795	3.87 0.86866	5.23 0.73273	7.72 0.46129	9.95 0.26855	14.87 0.06172	0.501

Number of pathogens identified is 8

Table 4.5e Quantiles of observed 5000 χ^2 values of seed variety IS-8080.

n	MIN	0.95-uq	0.75-uq	0.50-uq	0.25-uq	0.05-uq	0.01-uq	max	@KS-D
10	0.86 0.99998	0.86 0.99998	1.8 0.99908	3.1 0.98934	9.43 0.58227	40.22 0.00003	112.81 0	341.64 0	0.6082
20	0.62 1	1.03 0.99994	2.49 0.9959	4.84 0.93867	12.19 0.34953	53.43 0	66.24 0	133.7 0	0.4614
30	0.9 0.99997	1.41 0.99972	3.06 0.98991	5.74 0.89014	11.76 0.38195	40.11 0.00003	53.11 0	88.27 0	0.3925
40	1.24 0.99985	1.78 0.99913	3.73 0.97716	6.8 0.81504	12.19 0.34953	32.56 0.00062	50.81 0	79.25 0	0.3239
50	1.31 0.9998	2 0.9985	3.92 0.9722	7.06 0.7942	14.57 0.20304	29.56 0.00186	43.67 0.00001	94.6 0	0.2961
75	1.29 0.99982	2.6 0.99503	4.61 0.94858	7.42 0.76412	15.4 0.16491	25.41 0.00794	33.59 0.00042	48.31 0	0.2721
100	1.41 0.99972	2.83 0.99279	4.87 0.9373	8.37 0.67983	13.64 0.25356	22.1 0.02361	30.18 0.00148	59.04 0	0.2318
125	1.55 0.99955	2.97 0.99112	5.17 0.92265	9.29 0.59514	12.92 0.29859	20.4 0.04015	26.93 0.00471	38.44 0.00007	0.189
150	1.59 0.9995	3.1 0.98934	5.63 0.89688	8.89 0.63205	12.54 0.32445	19.2 *0.0576*	26.1 0.00627	44.77 0.00001	0.1511
175	1.47 0.99966	3.09 0.98948	5.86 0.88254	8.42 0.67526	11.78 0.38041	17.73 0.08806	22.86 0.0185	40.48 0.00003	0.1782
200	1.27 0.99983	3.31 0.98596	5.96 0.87602	8.34 0.68256	11.27 0.42093	16.61 0.11995	21.61 0.02757	34.2 0.00034	0.1893
300	1.49 0.99963	3.74 0.97691	5.79 0.887	7.5 0.75727	9.6 0.56669	13.64 0.25356	16.96 0.10906	30.13 0.00151	0.3169
400	1.94 0.9987	3.65 0.97906	5.13 0.92471	6.48 0.8395	8.09 0.70522	11.32 0.41686	14.09 0.22805	19.84 0.04758	0.4608
500	2.28 0.99725	3.23 0.98732	4.35 0.95854	5.47 0.90629	6.87 0.80951	9.44 0.58135	11.44 0.40717	15.05 0.18022	0.5889

Number of pathogens identified is 11

Bibliography

[1] Adcock, C.J. (1987), "A Bayesian Approach to Calculating Sample Sizes for Multinomial Sampling," The Statistician, 36, 155-159.

[2] Adcock, C.J. (1997), "Sample Size Determination: A Review," The Statistician, 46, No.2, 261-283.

[3] Anscombe, F.J. (1961), "Estimating a Mixed-Exponential Response Law," Journal of American Statistical Association, 56, 493-502.

[4] Anderson, T.W., and Herman Burstein (1967), "Approximating the Upper Binomial Confidence Limit," Journal of American Statistical Association, 62, 857-861.

[5] Blumenthal, S., and Dahiya, R.C. (1981), "Estimating the Binomial Parameter n," Journal of American Statistical Association, 76, 903-909.

[6] Blumenthal, S., and Marcus, R. (1975), "Estimating Population Size With Exponential Failure," Journal of American Statistical Association, 28, 913-922.

[7] Chandra, S., Huaman, Z., Hari Krishna S., and Ortiz, R. (2002), "Optimal Sampling Strategy and Core Collection Size of Andean Tetraploid Potato Based on Isozyme Data- A Simulation Study," Theoretical and Applied Genetics, 104, 1325-1334.

[8] Cochran, G. (1963), Sampling Techniques, 2nd edition, New York: Wiley.

[9] Desu, M.M, and Raghava Rao, D. (1990), Sample Size Methodology, London: Academic Press.

[10] Draper, N., and Guttmann, I. (1971), "Bayesian Estimation of the Binomial Parameter," Technometrics, 13, 667-673.

[11] Feldman, D., and Fox, M. (1968), "Estimation of the Parameter n in Binomial Distribution," Journal of American Statistical association, 63, 150-158.

[12] Gosh, M., and Meeden, G. (1975), "How Many Tosses Of a Coin?," Sānkhya-A, 37, 523-529.

[13] Herman Burstein. (1971), Attribute Sampling-Tables and Explanations, New York: McGraw- Hill.

[14] Jelinski, Z., and Moranda, P. (1972), In Statistical Computer Performance Evaluation, New York: Academic Press.

[15] Merran, E., Nicholas, H., and Peacock, B. (1993), Statistical Distributions, 2nd edition, New York: John Wiley & Sons.

[16] Olkin, I., John Petkau, A., and James, V. Zidek. (1981), "A Comparison of n Estimators for the Binomial Distribution," Journal of American Statistical association, 76, 637-642.

[17] Roger, J. Lewis. (2000), "Power Analysis and Sample Size Determination: Concepts and Software Tools," Annual Meeting of the Society for Academic Emergency Medicine (SAEM), San Francisco, California.

[18] Kotz, S., and Johnson, L. Norman. (1986), "Population or Sample Size Estimation," Encyclopedia of Statistical Sciences, Vol.7, 100-110.

[19] Sokal, R.R., and Rohlf, F.J. (1981), Biometry, New York: W.H. Freeman and Co.

[20] Sukhatme, P.V., and Sukhatme, B.V. (1970), Sampling Theory of Surveys, Bombay: Asia Publishing House.

[21] Wald, A. (1947), Sequential Analysis. New York: Wiley.

[22] Wittes, J.T. (1970), Estimation of Population Size: The Bernoulli Census. Unpublished Ph.D. thesis, Harvard University, Cambridge.