RATIO ESTIMATORS IN AGRICULTURAL RESEARCH

A thesis presented in partial fulfilment of the requirements

for the degree of

Master of Science

in

Statistics

at Massey University, Palmerston North, New Zealand

CHUN GUI QIAO

B. Agr. Sc., M. Agr. Sc., PhD (Agr. Sc.)

2002

ACKNOWLEDGEMENTS

I would like to express my heart-felt thanks to my Principal Supervisor Professor Graham R Wood and the Co-supervisor Associate Professor Chin Diew Lai. It has been the most productive period in my academic career to study for a Master degree with them, which I have enjoyed so much. It was through discussions with these supervisors that my original ideas were extended into the formal framework of the thesis. The feedback from them on my work was crucial in ensuring the quality of the thesis. I have learned so much from them in scientific methodology, such as composing, structuring, and presenting academic work that is valuable in pursuing a career associated with statistics. Hence, I am grateful to both of them for leading me in the right direction. Mrs Wendy Brown has been most supportive in every aspect throughout the study; I thank her for being such a nice colleague. Thanks also go to Dr Zhenzi Zhang for her assistance in Word Processing during compilation of the thesis.

Special thanks are given to Professor Jingyong Ma from Rice Research Institute at Jilin Agricultural University in China, for kindly providing rice breeding data for the analysis in Chapter 4 of the thesis.

I would also like to express my gratitude to my wife Hang Yu and daughter Yu (Alice) Qiao for their full support of my academic career and my Master of Science degree study.

ABSTRACT

This thesis addresses the problem of estimating the ratio of quantitative variables from several independent samples in agricultural research. The first part is concerned with estimating a binomial proportion, the ratio of discrete counts, from several independent samples under the assumption that there is a single underlying binomial proportion p in the population of interest. The distributions and properties of two linear estimators, a weighted average and an arithmetic average, are derived and merits of the approaches discussed. They are both unbiased estimators of the population proportion, with the weighted average having lower variability than the arithmetic average. These findings are obtained through a first principles analysis, with a geometrical interpretation presented. This variability result is also a consequence of the Rao-Blackwell theorem, a well-known result in the theory of statistical inference. Both estimators are used in the literature but we conclude that the weighted average estimate should always be used when the sample sizes are unequal. These results are illustrated by a simulation experiment and are validated using survey data in the study of lodging percentage of sunflower cultivar, *Improved Peredovic*, in Jilin Province, China in 1994.

The second part of the research addresses the problem of estimating the ratio μ_X / μ_Y of the means of continuous variables in agricultural research. The distributional properties of the ratio X/Y of independent normal variables are examined, both theoretically and using simulation. The results show that the moments of the ratio do not exist in general. The moments exist, however, for a punctured normal distribution of the denominator variable if we only sample points for which $|Y| > \varepsilon$, ε being a small positive quantity. We draw out the practical rule-of-thumb that the ratio of two independent normal variables can be used to estimate μ_X / μ_Y when the coefficient of variation of the denominator variable is sufficiently small (less than or equal to 0.2).

Lastly the thesis evaluates the relative merits of two common estimators of the ratio of the means of continuous variables in agricultural research, an arithmetic average and a weighted average, via simulation experiments using normal distributions. In the first

simulation, the ratio and common coefficient of variation are changed while the sample size is kept moderately large. In the second simulation, the ratio and sample size are changed while the coefficient of variation is held constant. Results show that the weighted average always provides a better estimate of the true ratio and has lower variability than the arithmetic average. It is recommended that the weighted average be used for estimating the ratio from several pairs of observations. These results are tested using research data from rice breeding multi-environment trials in Jilin Province, China in 1995 and 1996. These data are used to demonstrate the diagnostic approach developed for assessing the 'safety' use of the arithmetic and the weighted average methods for estimating the ratio of the means of independent normal variables.

TABLE OF CONTENTS

LIST OF PAPERS SUBMITTED TO ACADEMIC JOURNALS FOR PUBLICATION BY CANDIDATE RELEVANT TO THE THESIS

C. G. Qiao, G. R. Wood and C. D. Lai (2002) Estimating a binomial proportion from several independent samples in agricultural research.

C. G. Qiao, G. R. Wood, C. D. Lai, D. W. Luo and J. Y. Ma (2002) Comparison of two common estimators of the ratio of the means of continuous measurements in agricultural research.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 RATIOS OF QUANTITATIVE VARIABLES

A ratio of quantitative variables, characterised by two variables, the numerator and denominator, is often used in scientific research, survey studies and daily life. Some of these are concerned with the degree of success, such as the success rate of a particular practice. Others are related to the comparison of a novel method relative to an existing one, such as an innovative technique relative to a standard approach. Depending on the composition variables, ratios can generally be expressed or classified into four major categories:

1) ratios of non-negative counting variables to positive counting variables (proportions or percentages);
2) ratios of continuous variables (proportions);
3) ratios of continuous to positive discrete counting variables; and
4) ratios of discrete counting variables to continuous variables.

In the first category, where the numerator and denominator are both discrete counting variables, the ratio can be further classified into two subcategories. In the first subcategory the denominator variable represents the total number of counts, while the numerator variable denotes the number of "successes" out of this total number of counts. Hence, this special kind of ratio is referred to as proportion or percentage, for example, the proportion or percentage of plants infected by a particular disease over the total number of plants sampled. In the second subcategory, the numerator and denominator variables denote counts of different nature and hence the ratio is known as the average number of occurrence for the numerator variable per unit count of the denominator variable. An example of this is the average number of cars possessed per family in a given community, which can be sensibly estimated as the ratio of the total number of cars over the total number of families surveyed in the population. In the second major category, where the numerator and denominator variables are both continuous, the ratio can also take several different forms.

The numerator and denominator of a ratio may measure the same quantitative attribute of two contrasting methodologies such as

i) ratio of grain yields of two crop varieties (a new variety and a local control);

ii) a fraction over the total amount, such as harvest index, which is the ratio of the economic yield over the total biological yield of field crops, or

iii) the average amount of one attribute per unit amount of another attribute, such as grain yield in kilograms per hectare for a particular crop species.

Examples of the third category are the grain yield (weight) per plant and the average weight per person within a community. Examples of the fourth category are the average number of people inhabited on a unit area of land (for example, per square metre) or the average number of insect pests parasitising plants per unit area of farmland.

Some ratios are easily identified as belonging to one of the four categories, such as the exchange rate of two currencies or the relative cost of groceries in Australia and New Zealand (expressed as price ratio of the specific grocery items between the two countries in local dollars). There are other types of ratio the numerator and/or denominator of which involve some kind of computation from the raw data. Examples of these include (1) the linear correlation coefficient, which is the ratio of the sum of cross-products between two variables over the square root of the product of the sums of squares for the two variables; (2) heritability, which is the ratio of the genetic variance to the total phenotypic variance; (3) the linear regression coefficient, which is the ratio of the sum of cross-products between the two variables to the sum of squares of the independent variable; (4) mid-parent heterosis, whose denominator is the mean of the two parents for a particular attribute such as grain yield, while the numerator is the difference between the hybrid and the mean of the two parents for the same character. These were termed naive estimators of ratio by Frankel (1971) and Rao and Kuzik (1975), who even regarded partial and multiple regression coefficients as belonging to this type of ratio. All of them can also be characterised as matching one of the four major groups. The appropriate estimation of ratio is hence of practical importance.

1.2 RATIO ESTIMATORS

If only one pair of measurements are made and collected for the numerator and denominator, the estimation of the ratio is straightforwardly carried out by division. In situations where a series of such ratios, proportions or percentages need to be pooled or averaged, however, a serious question will arise: which way of averaging should be used? There have been confusions especially in estimating ratio of continuous and of discrete counting variables in agricultural research. Unfortunately, this has not been well documented in the literature.

Estimators of ratios of discrete counts

Let x_i (i=1,2,...,K) be one of the K independent binomial samples with size n_i. There are two common methods for estimating the ratio of such count data p from several samples, arithmetic average and the weighted. The arithmetic average method estimates the ratio via dividing the sum of the individual ratio estimates of these samples by the number of samples, using the formula $\bar{p}_A = \left(\sum \frac{x_i}{n_i} \right) / K$. The weighted average, a contrasting approach, calculates the ratio via dividing the sum of the numerators by the sum of the denominators of a series of ratio estimates, employs the mathematical expression $\bar{p}_W = \sum w_i \frac{x_i}{n_i} = \left(\frac{n_1}{\sum n_i} \frac{x_1}{n_1} + \frac{n_2}{\sum n_i} \frac{x_2}{n_2} + ... + \frac{n_K}{\sum n_i} \frac{x_K}{n_K} \right) = \frac{\sum x_i}{\sum n_i}$. Due to the nature of the numerator and denominator, these estimate a binomial proportion, and will be thus referred to throughout the thesis. These two approaches have been widely used in studies of proportion data in applied research, especially agricultural sciences. Although some textbooks have advocated that the weighted average be adopted against the arithmetic average, the theoretical distributions and justifications of such an approach have not been provided. There have been no reports on the evaluations of and comparison between these two contrasting methods.

Estimators of a ratio of continuous variables

Let (x_i, y_i), $i = 1,2,...,n$, be a random sample of observations from a bivariate population, such as normal population $N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, and for each observation a ratio is calculated as x_i / y_i. There are two popular ways in agricultural research to estimate the ratio of two population means μ_x / μ_y, the arithmetic average approach, with $\overline{R}_A = \left(\sum \frac{x_i}{y_i} \right) / n$, and the weighted average approach, with

$$\overline{R}_W = \sum w_i \frac{x_i}{y_i} = \left[\left(\frac{y_1}{\sum y_i} \right) \left(\frac{x_1}{y_1} \right) + \left(\frac{y_2}{\sum y_i} \right) \left(\frac{x_2}{y_2} \right) + ... + \left(\frac{y_n}{\sum y_i} \right) \left(\frac{x_n}{y_n} \right) \right] = \frac{\sum x_i}{\sum y_i} = \frac{\bar{x}}{\bar{y}}.$$

There are several other estimators of the ratio of continuous variables for estimating the ratio of two population means (Hartley and Ross 1954; Quenouille 1956; Mickey 1959; Durbin 1959; Pascual 1961; Kokan 1963; Tukey 1958; Tin 1965). They are functions of the weighted and/or the arithmetic average ratio estimators. They have not, however, attracted attention from agricultural scientists. Only the weighted and the arithmetic average ratio estimators have gained popularity, with the latter being more favoured; the remaining estimators appear only in the sampling survey areas. Again, the relative merits of both methods have not been compared and theoretical justifications for using either of them have not been explored.

It is intuitively obvious that the weighted average method should be used in estimating the ratio of either discrete counting or continuous variables. Our intuition, however, often fails in practice for various reasons. Statistically, it is a matter of whether to adopt the idea of weighted averaging (and hence the weighted average method) or not (and hence the arithmetic average method). In essence, these two methods relate to averaging the series of ratio estimates before or after division. These options outline the skeleton of the thesis, as is shown in Figure 1.1.

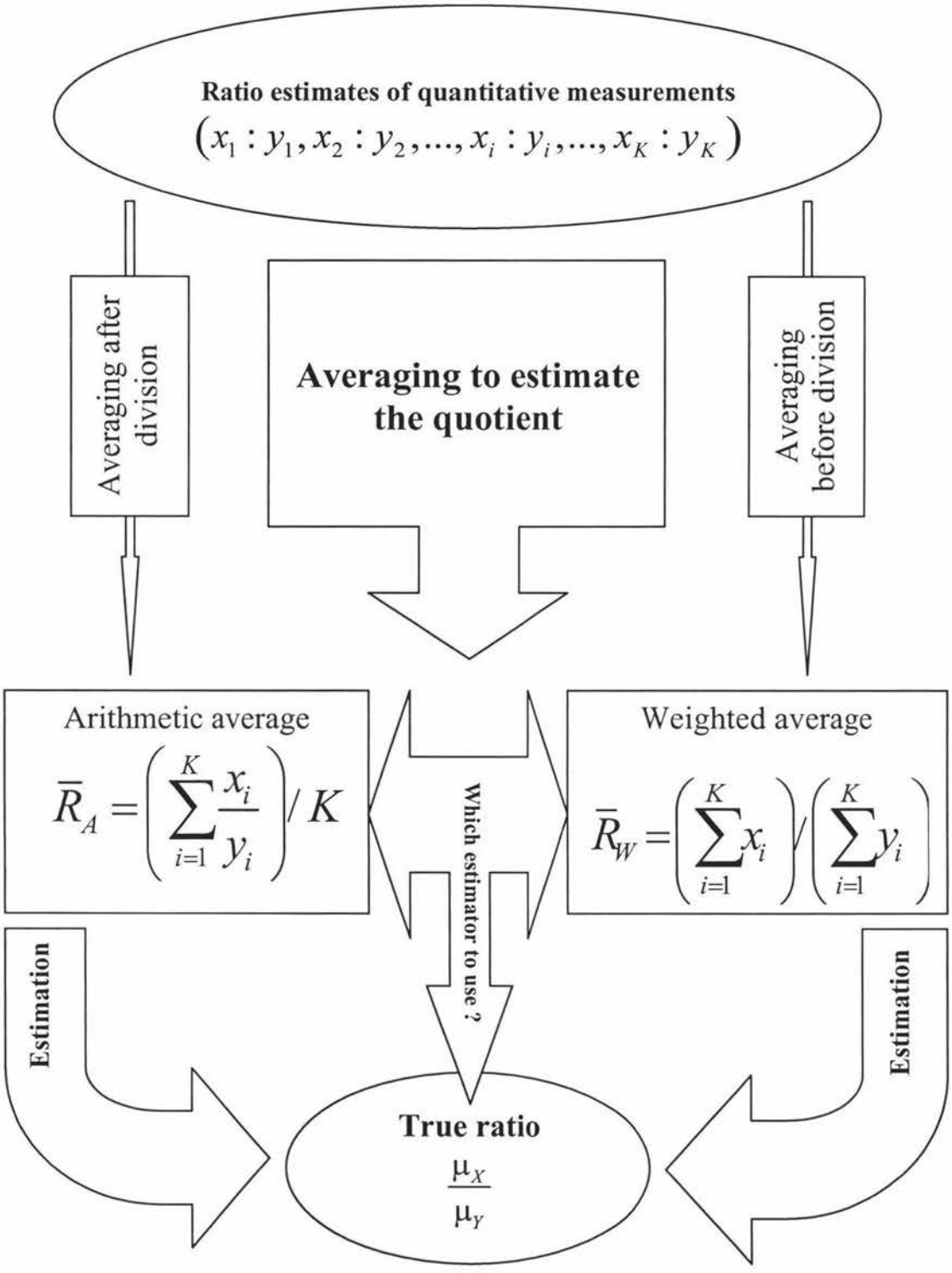

Figure 1.1 A diagram outlining the skeleton of the research comparing two common estimators of ratio, where X and Y can be either discrete counting variable or continuous variable. In the case of a ratio of discrete counting variables, $\overline{R}_A$ and $\overline{R}_W$ are replaced by $\overline{P}_A$ and $\overline{P}_W$, respectively.

1.3 MOTIVATION OF THE RESEARCH

The motivation of this research originated from the author's investigation of sunflower lodging percentages in China in 1994 (Qiao *et al.* 1994) and the study of relative performance of rice varieties in grain yield, also in China in 1995 and 1996 (Jingyong Ma 1996, personal communication). In both cases a series of ratio estimates needed to be pooled or averaged over different environments. The first belonged to the ratio of discrete counting variables, while the second belonged to the ratio of continuous variables. For the sunflower study, a survey was conducted in 1994 in the Western Region of Jilin Province, China to investigate the lodging percentage of a commercial sunflower cultivar *Improved Peredovic* in five sites (locations). The technicians at each site were asked to take a random sample of at least 500 plants for the measurement, but were encouraged to take larger samples if possible. The results are listed in Table 1.1, with number of lodged plants and total number of plants specified in each location. The aim was to estimate the average or pooled lodging proportion of the cultivar in the whole region. For the rice breeding multi-environment experiments conducted over eleven locations in Jilin Province, China in 1995 and 1996, the grain yield data were analysed to quantify the increase in grain yield of each variety over the control variety. In the regional testing program, the grain yield is customarily expressed as the percentage of each test variety relative to the control variety. The results are listed in Table 1.2; the aim is to estimate the mean percent yield increase of each of the test varieties over the control variety.

Table 1.1 Raw data from field inspection of lodging for sunflower cultivar *Improved Peredovic* at five locations (counties) in the Western Region of Jilin Province, China in 1994.

Location (county)	Number of plants lodged	Number of plants sampled	Percentage of lodging
Baicheng	265	1560	17.0
Zhenlai	250	1840	13.6
Da'an	462	2413	19.1
Changling	518	3627	14.3
Nongan	464	3027	15.3
The arithmetic average binomial proportion estimator ($\overline{P}_A$)			15.9
The weighted average binomial proportion estimator ($\overline{P}_W$)			15.7

Table 1.2 Grain yield performance of six rice varieties and the estimates of ratio between each of the test varieties and the control by the arithmetic and weighted average in a multi-environment trial during 1995 and 1996.

Location	Grain yield (kg/ha)	Percentage of control (%)	Grain yield (kg/ha)	Percentage of control (%)	Grain yield (kg/ha)	Percentage of control (%)	Grain yield of control (kg/ha)
1995	Jiu 9214		Chang 90-40		Ji K911		Control
Changchun	7083	104.4	7358	108.5	7068	104.2	6783
Dongfeng	5733	117.8	5934	121.9	3867	79.4	4868
Gongzhuling	8604	100.8	8664	101.5			8535
Jilin	7800	107.7	7290	100.6			7245
Lishu	8168	112.0	8100	111.1	7650	104.9	7292
Tonghua	7955	101.8					7815
Yanbian	8532	105.2	8652	106.7	8283	102.2	8106
Yushu	10082	105.0	9963	103.8	9638	100.4	9600
Arithmetic average ratio estimate $\overline{R}_A$		106.8		107.7		98.2	
Weighted average ratio estimate $\overline{R}_W$		106.2		106.7		99.6	
1996			Jiu 9421		Jiuhua 2		Control
Chanhchun			7041	103.5	6879	101.1	6804
Dongfeng					11801	122.9	9600
Gongzhuling			8381	102.1			8210
Jilin			8815	103.7			8501
Jilin Agricultural University			8095	94.4	7212	84.1	8571
Lishu			8151	94.2			8651
Shulan			7701	101.3	8100	106.6	7601
Tonghua			8358	105.2	8508	107.1	7945
Yanbian			7982	90.4			8834
Yongji			8271	101.6	7445	91.5	8138
Arithmetic average ratio estimate $\overline{R}_A$				99.6		102.2	
Weighted average ratio estimate $\overline{R}_W$				99.4		102.6	

In both circumstances, the pooled ratio estimate could differ with the way of averaging (which amount to averaging the series of ratio estimates before or after division). This forms the drive for investigations on the theoretical foundation of the difference between the two methods (the arithmetic versus the weighted average) and for evaluation of them in a more general sense in agricultural research. Therefore, this project will concentrate on the study of ratios of discrete counting variables and ratios of continuous variables in agricultural research. It is hoped that the findings and implications of the research will be directly relevant and applicable to estimations of the other two types of ratio.

1.4 OBJECTIVES OF THE THESIS

The aims of this project were to compare the relative merits of the different estimators by their theoretical distributions and simulations. Their practical implications in agricultural research will be addressed, with examples illustrated. A generalisation principle for the choice of a suitable ratio estimator with associated rule-of-thumb will be presented, emphasising the practical applicability of the research project.

1.4.1 Research Methodology

The behaviour of estimators of a ratio of quantitative variables may be investigated in a variety of ways, including the following, as is outlined by McCarthy (1969): (1) Exact analytic, in which the functional form of a distribution or a joint distribution is assumed; (2) approximate analytic, in which Taylor series approximations are used; (3) empirical studies, in which the data from actual surveys or experiments are used; and (4) simulation, which is also referred to as Monte Carlo sampling from synthetic populations.

The exact analytic approach should be sought whenever possible, since it is the starting point for theoretical study. The empirical approach, employing actual survey or experiment data, permits the use of complex designs, and the properties of estimators of many parameters could be investigated with the help of a computer and the relevant packages. An obvious limitation of the empirical approach is that the results are strictly applicable only to the particular population(s) considered. However, the empirical studies are extremely valuable in providing guidelines on the performances of various methods of estimation. The

simulation methodology, on the other hand, enables the researchers to mimic all sorts of populations under diverse environmental conditions. Therefore, the results or findings from simulation experiments can be applicable over a wide range of situations, with generalisation justified. For the purposes of the present research, a combination of all the above methods will be adopted, with each employed wherever possible and appropriate.

1.4.2 Work Included in this Thesis

Because of the diverse nature of the relevant literature in ratio of discrete counting variables and continuous variables, it was considered appropriate to present separate reviews of literature when these topics are discussed in the respective chapters. From Chapter 2 to Chapter 4, theoretical studies will be based on mathematical derivations and enhanced by simulated data using Minitab 13 for Windows software. The results or findings will then be validated using real data in agricultural research.

In the second chapter, we will investigate how the ratio of a discrete counting variable to another positive discrete counting variable can be used to estimate the unknown binomial proportion, as is often referred to in the literature. The theoretical distributions of two popular estimators of a binomial proportion will be evaluated and relative merits of the two methods, the weighted average and arithmetic average compared. In Chapter 3, the distributional properties of the ratio of independent normal variables will be explored both theoretically and using simulation. A practical rule-of-thumb will be drawn for using the ratio of independent normal variables to estimate the ratio of the means of continuous variables. In Chapter 4, the theoretical justifications will be pinpointed for the appropriateness of two common estimators of the ratio of the means of continuous variables, the weighted average and arithmetic average methods, based on the findings of Chapter 3. The relative merits of the two estimators will be evaluated using simulation experiments. Recommendations will be provided for practical diagnosis in evaluating the suitability of the use of ratio estimators of continuous variables in agricultural research. The final chapter, Chapter 5, summarises the findings of the research project and pinpoints areas of further research.

CHAPTER 2
ESTIMATING A BINOMIAL PROPORTION FROM SEVERAL INDEPENDENT SAMPLES

2.1 INTRODUCTION

Many areas of agricultural research use count data, modelled using a binomial distribution, to estimate the proportionate occurrence of a certain event relative to the total number of possible outcomes. Examples include the proportion of plants affected by the incidence of a particular disease or insect pest, the survival percentage of insect pests after application of certain chemicals, the germination percentage of seeds, and the proportion of environments in which a new variety outperforms the local control. When count data from a series of such samples (always of different sizes) have been recorded and pooled for estimating the true proportion, the question arises as to how this proportion should be estimated.

Two methods are commonly considered: arithmetic averaging, which divides the sum of all these sample proportions by the total number of samples, and weighted averaging, which estimates the proportion via dividing the total count from all the samples by the total sample size. This issue has been noticed and addressed in some statistical textbooks by advocating weighted averaging (Ott 1993; Ott and Mendenhall 1994). Both versions, however, have been used in the agricultural research literature. For example, some researchers use arithmetic averaging (Narayanan *et al.* 1999; Casler and Santen 2000; Ismail *et al.* 2000), while others adopt the weighted method (Chen *et al.* 2000; Choi *et al.* 2000; Paderson and Brink 2000). Pitt (1994) recommends that the weighted approach be used and provides procedures for such a practice in the quality control area; it is further argued that there is no sense in using the arithmetic average approach. Kim *et al.* (2000) applies a severity index in the study of disease reaction of soybean cultivars, which in essence adopts the idea of the weighted average. No theoretical reasons, however, are given for the choice of the weighted approach over the other option. Although it is clear from general statistical inference (see Section 2.3 for details) that the weighted average is the "best" over all other unbiased estimators, we have not found any report on a direct

comparison between these two methods and the consequences of inappropriate use of them.

This chapter examines the theoretical inter-relations between these two estimators of the binomial proportion and provides a simulation study to illustrate the findings. It is assumed in this study that different samples from the population of interest has a constant binomial proportion; the weighted average is unreservedly recommended under this assumption.

2.2 THEORETICAL INTER-RELATIONS

Suppose K independent samples have been taken from a binomial distribution and let x_i (i=1,2,...,K) represent the number of occurrences of a particular event in the ith sample, with n_i the corresponding sample size. The true underlying proportion of occurrence p for this event is estimated by $y_i = x_i / n_i$. Evidently x_i follows a binomial distribution, or $x_i \sim B(n_i, p)$. This represents a situation of repeated sampling from a single distribution, which can be one biological environment in the context of agricultural research. In each of these samples, the occurrence of the event for a particular individual is not influenced by those of others in the same sample.

A simple way to estimate the population proportion p is to use the formula $\bar{p}_A = \sum y_i / K$, referred to as the arithmetic average approach in this context. A contrasting approach, which is termed the weighted average method in this study, employs the mathematical formula $\bar{p}_W = \sum w_i y_i = \left(\frac{n_1}{\sum n_i} \frac{x_1}{n_1} + \frac{n_2}{\sum n_i} \frac{x_2}{n_2} + ... + \frac{n_K}{\sum n_i} \frac{x_K}{n_K} \right) = \frac{\sum x_i}{\sum n_i}$. We now compare the expected values (population means) and the variances of these two estimates.

2.2.1 Distribution of $\bar{p}_A$

The mean and variance of $\bar{p}_A$ are derived as follows. Since $y_i = x_i / n_i$, where $x_i \square B(n_i, p)$, we have $E(y_i) = p$ and

$$Var(y_i) = \frac{1}{n_i^2} Var(x_i) = \frac{1}{n_i^2} n_i p(1-p) = \frac{1}{n_i} p(1-p)$$

Under certain conditions $\bar{p}_A$ then approximately follows a normal distribution with mean

and variance given by

$$E(\overline{p}_A) = E\left(\frac{\sum y_i}{K}\right) = \frac{\sum E(y_i)}{K} = \frac{Kp}{K} = p \text{ and}$$

$$Var(\overline{p}_A) = Var\left(\frac{\sum y_i}{K}\right) = \frac{\sum Var(y_i)}{K^2} = \frac{\sum p(1-p)/n_i}{K^2} = \frac{p(1-p)}{K^2}\sum \frac{1}{n_i} \qquad (2.1)$$

2.2.2 Distribution of $\overline{p}_W$

The mean and variance of $\overline{p}_W$ are derived as follows.

Since $x_i \sim B(n_i, p)$, we have $\sum x_i \sim B\left(\sum n_i, p\right)$, whence

$$E\left(\sum x_i\right) = p\sum n_i \text{ and } Var\left(\sum x_i\right) = \left(\sum n_i\right)p(1-p).$$

Then $\overline{p}_W$ will approximately follow a normal distribution with mean and variance

$$E(\overline{p}_W) = E\left(\frac{\sum x_i}{\sum n_i}\right) = \frac{E\left(\sum x_i\right)}{\sum n_i} = \frac{p\sum n_i}{\sum n_i} = p \text{ and}$$

$$Var(\overline{p}_W) = Var\left(\frac{\sum x_i}{\sum n_i}\right) = \frac{Var\left(\sum x_i\right)}{\left(\sum n_i\right)^2} = \frac{\left(\sum n_i\right)p(1-p)}{\left(\sum n_i\right)^2} = \frac{p(1-p)}{\sum n_i} \qquad (2.2)$$

That is, $\overline{p}_W \sim N\left(p, \frac{p(1-p)}{\sum n_i}\right)$, if $n = \sum n_i$ is large and $np \geq 5$, $n(1-p) \geq 5$.

2.2.3 Variance Ratio R

We note that $\overline{p}_W$ is the maximum likelihood estimator of p (Johnson *et al.* 1992). In order to compare $\overline{p}_A$ and $\overline{p}_W$ the conditions for normality are not required. It can be seen from the above derivations that $E(\overline{p}_W) = E(\overline{p}_A) = p$, thus both are unbiased estimators of the population proportion. However, the variances of $\overline{p}_W$ and $\overline{p}_A$ are different and both are functions of the sample sizes n_i. The relative merit of $\overline{p}_W$ and $\overline{p}_A$ can thus be measured by the ratio of these two variances, R. This is calculated as:

$$R=\frac{Var(\overline{p}_W)}{Var(\overline{p}_A)}=\left\{\frac{p(1-p)}{\sum n_i}\right\}/\left\{\frac{p(1-p)}{K^2}\sum\frac{1}{n_i}\right\}=K^2/\left\{\left(\sum n_i\right)\left(\sum\frac{1}{n_i}\right)\right\} \qquad (2.3)$$

If $R<1$, the variance of $\overline{p}_W$ is smaller than that of $\overline{p}_A$; if $R=1$, the two variances are equal. We now show that $R\le 1$ always and hence that $\overline{p}_W$ is the better estimator.

2.2.4 Mathematical Analysis

Consider vectors x and y, where

$$\mathbf{x}=\left(n_1^{\frac{1}{2}},n_2^{\frac{1}{2}},...,n_K^{\frac{1}{2}}\right) \text{ and } \mathbf{y}=\left(\frac{1}{n_1^{\frac{1}{2}}},\frac{1}{n_2^{\frac{1}{2}}},...,\frac{1}{n_K^{\frac{1}{2}}}\right)$$

Let **<x, y>** denote the inner product of **x** and **y,** and $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ denote the norms of **x** and **y,** respectively. It follows from the Cauchy-Schwarz inequality that the relationship between the modulus of the inner product and the norms is

$|<\mathbf{x},\mathbf{y}>|\le\|\mathbf{x}\|\ \|\mathbf{y}\|$, where

$$|<\mathbf{x},\mathbf{y}>|=\left(n_1^{\frac{1}{2}}\frac{1}{n_1^{\frac{1}{2}}}+n_2^{\frac{1}{2}}\frac{1}{n_2^{\frac{1}{2}}}+...+n_K^{\frac{1}{2}}\frac{1}{n_K^{\frac{1}{2}}}\right)=(1+1+...+1)=\sum 1=K,$$

$$\|\mathbf{x}\|=\left\{\left(n_1^{\frac{1}{2}}\right)^2+\left(n_2^{\frac{1}{2}}\right)^2+...+\left(n_K^{\frac{1}{2}}\right)^2\right\}^{\frac{1}{2}}=\left(\sum n_i\right)^{\frac{1}{2}} \text{ and}$$

$$\|\mathbf{y}\|=\left\{\left(\frac{1}{n_1^{\frac{1}{2}}}\right)^2+\left(\frac{1}{n_2^{\frac{1}{2}}}\right)^2+...+\left(\frac{1}{n_K^{\frac{1}{2}}}\right)^2\right\}^{\frac{1}{2}}=\left(\sum\frac{1}{n_i}\right)^{\frac{1}{2}}$$

Therefore, $K\le\left(\sum n_i\right)^{\frac{1}{2}}\left(\sum\frac{1}{n_i}\right)^{\frac{1}{2}}$. Since $K>0$, this is equivalent to

$$K^2\le\left(\sum n_i\right)\left(\sum\frac{1}{n_i}\right) \text{ or } K^2/\left\{\left(\sum n_i\right)\left(\sum\frac{1}{n_i}\right)\right\}\le 1.$$

Following (2.3),

$$R=\frac{Var(\overline{p}_W)}{Var(\overline{p}_A)}=K^2/\left\{\left(\sum n_i\right)\left(\sum\frac{1}{n_i}\right)\right\}\le 1$$

Note that $R=1$ when there is equality in the Cauchy-Schwarz inequality. This is equivalent

to having $\mathbf{x}=c\mathbf{y}$ for some real number c or $n_i^{\frac{1}{2}} = c\dfrac{1}{n_i^{\frac{1}{2}}}$ for all i, or $n_i=c$ for all i. Thus $\overline{p}_A$ is as good as $\overline{p}_W$ only when all samples have the same size.

2.2.5 Geometrical Interpretation

If V is a real inner product space, then the angle θ between the nonzero vectors $\mathbf{x},\mathbf{y} \in \mathbf{V}$ (Figure 2.1) is defined by

$$\cos\theta = \frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\|\|\mathbf{y}\|}, \tag{2.4}$$

(Lipschutz 1968). Hence, $<\mathbf{x},\mathbf{y}>=\|\mathbf{x}\|(\|\mathbf{y}\|\cos\theta)$, indicating that the inner product is the product of the length of $\mathbf{x}$ ($\|\mathbf{x}\|$) and the length of the projection of $\mathbf{y}$ on $\mathbf{x}$ ($\|\mathbf{y}\|\cos\theta$). The vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal or perpendicular if $\theta = \dfrac{\pi}{2}$ or $\cos\theta = 0$. They are collinear if $\cos\theta = 1$ whence $\theta = 0$ (Lipschutz 1968).

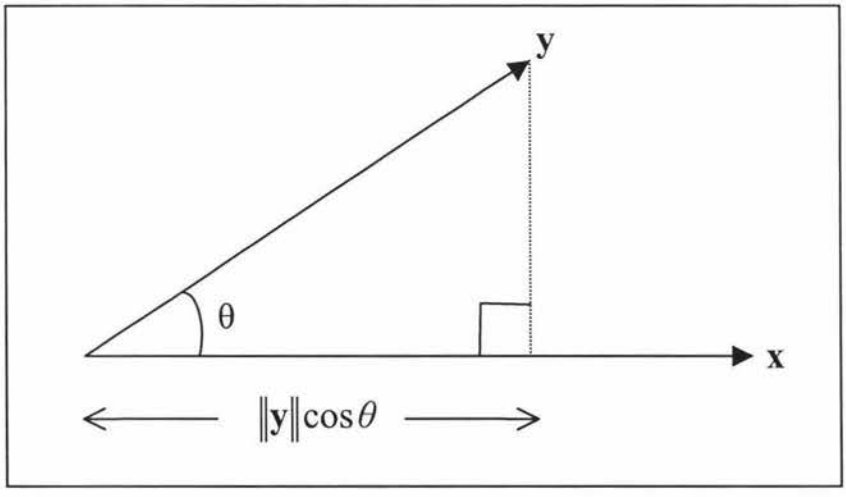

Figure 2.1 For a fixed K^2, vectors $\mathbf{x}=\left(n_1^{\frac{1}{2}},n_2^{\frac{1}{2}},...,n_K^{\frac{1}{2}}\right)$ and $\mathbf{y}=\left(\dfrac{1}{n_1^{\frac{1}{2}}},\dfrac{1}{n_2^{\frac{1}{2}}},...,\dfrac{1}{n_K^{\frac{1}{2}}}\right)$ are shown, where the ratio of the variances $Var(\overline{p}_W)/Var(\overline{p}_A)$ is equal to $\cos^2\theta$. Thus $\overline{p}_W$ becomes progressively better than $\overline{p}_A$ as the angle θ increases.

Here in our study of the variances of the estimates of the two ways of averaging the sample proportions, $< \mathbf{x},\mathbf{y} >= K > 0$, with $\|\mathbf{x}\| = \left(\sum n_i\right)^{\frac{1}{2}}$ and $\|\mathbf{y}\| = \left(\sum \frac{1}{n_i}\right)^{\frac{1}{2}}$. From (2.4) we have $\frac{|< \mathbf{x},\mathbf{y} >|^2}{\|\mathbf{x}\|^2 \|\mathbf{y}\|^2} = \cos^2\theta$, whence

$$R = \frac{Var(\overline{p}_W)}{Var(\overline{p}_A)} = \frac{K^2}{\left(\sum n_i\right)\left(\sum \frac{1}{n_i}\right)} = \frac{|< \mathbf{x},\mathbf{y} >|^2}{\|\mathbf{x}\|^2 \|\mathbf{y}\|^2} = \cos^2\theta .$$

This shows that the ratio R of the variances of the two averaging methods depends on the angle θ between the vectors $\mathbf{x}$ and $\mathbf{y}$. When $\theta = 0$, the two vectors are collinear, giving $\cos^2\theta = 1$. In this case, all the n_i's are equal and hence the variances of the two estimates are the same (R=1). When θ approaches $\frac{\pi}{2}$, or the two vectors are almost perpendicular, the value of $\cos^2\theta$ will approach 0 and thus R will be arbitrarily small. Since K>0, however, it follows that $\cos^2\theta > 0$ and therefore the two vectors are never perpendicular. Within this range, the value of $\cos^2\theta$ will decrease as the angle becomes larger and in turn the weighted average will have smaller variance than the arithmetic average.

2.2.6 Effect of Sample Size Difference

We now consider the two-sample case (K=2) with population proportion p kept constant, assuming without loss of generality that $n_1 < n_2$, in order to examine the relationship of (2.3). When the total sample size $n = n_1 + n_2$ is fixed the variance of $\overline{p}_W$ will remain constant irrespective of the relative values of n_1 and n_2, since $Var(\overline{p}_W) = \frac{p(1-p)}{n_1 + n_2}$. The variance of $\overline{p}_A$, however, is influenced by the relative size of n_1 and n_2, since $Var(\overline{p}_A) = \frac{p(1-p)}{2^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. The reduction of variance in the larger sample, caused by the inequality of n_1 and n_2, does not compensate for the greater increase of variance in the smaller sample. Hence, the variance of the weighted average estimate is always smaller

than that of the arithmetic average estimate, even if both sample sizes are large. The ratio of the two variances is thus

$$R = 2^2 / \left\{ (n_1 + n_2) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\} = \frac{\left(\sqrt{n_1 n_2} \right)^2}{\{(n_1 + n_2)/2\}^2} < 1,$$

since the arithmetic mean is always greater than the geometric mean when $n_1 \neq n_2$.

The magnitude of the difference depends on how much the two sample sizes n_1 and n_2 differ; for a fixed $n = n_1 + n_2$, the difference between the two variances, $Var(\bar{p}_A) - Var(\bar{p}_W)$, is

$$p(1-p) \left\{ \left(\frac{1}{2^2} \right) \left(\frac{n_1 + n_2}{n_1 n_2} \right) - \frac{1}{n_1 + n_2} \right\} = \left\{ \frac{p(1-p)}{4} \right\} \left\{ \frac{1}{(n_1 n_2) n} \right\} (n_1 - n_2)^2 \qquad (2.5)$$

The first term on the right hand side of (2.5) is a constant $\frac{p(1-p)}{4}$. The other two terms of $\frac{1}{(n_1 n_2) n}$ and $(n_2 - n_1)^2$ can be easily shown to reach a maximum when $n_2 - n_1$ reaches a maximum at $n_2 = n - n_1$, $n_1 = 1$. Hence, the difference between the two variances is maximised when the magnitude of the difference between the two sample sizes $n_2 - n_1$ reaches a maximum.

The effect of the difference in sample sizes $n_2 - n_1$ on the estimates of variances for both averaging methods, and on the ratio between the two variances, was examined at total sample size n of 20, 100, 200, 2000 and 20,000. Since similar results were obtained for each of these total sample sizes, only the results for total sample size of 20 are shown in Figure 2.2 below. As the difference between the two sample sizes increased, the variance estimate for arithmetic averaging increased, while that for weighted averaging remained constant (Figure 2.2). The variance ratio between the two averaging methods was a function of the difference between the sample sizes, being smaller for greater sample size differences and larger for smaller sample size differences (Figure 2.2).

It is evident that the difference in sample sizes determines the relative merit of the

averaging methods when the total sample size and the number of samples are fixed. The sample size difference affects the variance estimate of the arithmetic average, but has no effect on that of the weighted average. For cases of more than two samples of unequal size, the relationships between difference in sample sizes and the variance ratio become complex, but $R \leq 1$ always holds from Cauchy-Schwarz inequality.

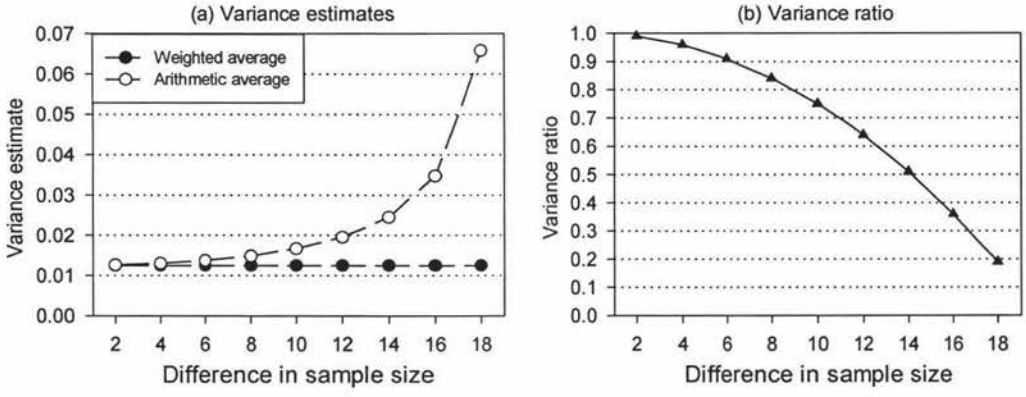

Figure 2.2 The influence of sample size difference on (a) the variance estimates of the weighted and arithmetic averaging approaches and (b) the variance ratio, where total sample size and population proportion are kept at $n = \sum n_1 + n_2 = 20$ and $p = 0.5$.

2.2.7 Implications and Recommendations

The variance ratio $R = K^2 / \left\{ \left(\sum n_i \right) \left(\sum \frac{1}{n_i} \right) \right\}$ is an indicator of the relative merit of the weighted and arithmetic averages. It is influenced by three factors, K^2, $n = \sum n_i$ and $\sum \frac{1}{n_i}$. The ratio R will become larger as we increase K (hence the first factor K^2) provided that the other two factors $n = \sum n_i$ and $\sum \frac{1}{n_i}$ can be kept approximately constant. The ratio decreases as the total sample size $n = \sum n_i$ increases provided that K^2 and $\sum \frac{1}{n_i}$ can be kept approximately constant. For a fixed K^2 and $n = \sum n_i$, R will decrease as the

sample size differences $n_{i+1} - n_i$ increase; these influence the third factor $\sum \frac{1}{n_i}$. The advantage of weighted over arithmetic averaging depends on the differences in sample sizes, the number of samples and the total sample size. The interrelationships between these factors, however, tend to become more complicated when the number of samples or the total sample size gets larger; this requires further investigation.

In agricultural research or surveys, it is recommended that the weighted averaging approach be employed when proportion data from a series of samples needs to be pooled or averaged in a single environment in which the measure of interest occurs with constant proportion p. The measure of interest in practical situations, however, can vary substantially from one site to another, which makes it possible for the overall measure to follow a mixture of binomial distributions, instead of one single binomial. In this case, the modelling approach recommended by Wood (1999) should be considered and further study comparing the two estimators would be worthwhile. The Restricted Maximum Likelihood (REML) methodology may also be considered. Another situation in which $\bar{p}_A$ has to be used is when n_i and x_i are not known individually but only the $y_i = x_i / n_i$ are available. For example, when the data collector only presents the percentages for different samples, we have no choice but to use the arithmetic averaging approach as a substitute for weighted averaging.

2.3 STATISTICAL INFERENCE

Alternatively, we can think of the collection of all K samples as a large binomial sample of size $\sum n_i$, that is, the experiment may be considered as having observed $\sum x_i$ "successes" from a binomial sample of size $n = \sum n_i$ and hence the best estimator of p is naturally $\bar{p}_W = \frac{\sum x_i}{\sum n_i}$, not $\bar{p}_A = \left(\sum \frac{x_i}{n_i} \right) / K$. Based on the concept and the definition of a "best" estimator described by Guenther (1973), $\bar{p}_W$ is considered the "best" estimator for p in the sense of being unbiased and having the smallest variance. This property also follows from the fact that $\bar{p}_W$ is a sufficient statistic for p (Johnson *et al.* 1992; Guenther 1973). Thus, it

is a uniformly minimum variance unbiased estimator (UMVUE) by the Rao-Blackwell theorem (Mood *et al.* 1974).

2.4 DIFFERENCE BETWEEN $\overline{p}_A$ AND $\overline{p}_W$

The difference between $\overline{p}_A$ and $\overline{p}_W$ is influenced by two factors, the individual binomial proportion estimates $y_i = x_i / n_i$ and the sample size differences $n_i - n_j$. It is easy to prove that $\overline{p}_A$ and $\overline{p}_W$ will be similar if the $y_i = x_i / n_i$ are similar or the $n_i - n_j$ are small. The proof is demonstrated as follows.

(1) If the individual binomial proportion estimates $y_i = x_i / n_i$ are similar, then $\frac{x_i}{n_i} \approx \frac{x_j}{n_j} = \alpha$, for all i, j. Thus $\overline{p}_A = \left(\sum \frac{x_i}{n_i} \right) / K \approx K\alpha / K = \alpha$, while $\overline{p}_W = \frac{\sum x_i}{\sum n_i} \approx \frac{\sum \alpha n_i}{\sum n_i} = \alpha$, leading to $\overline{p}_A \approx \overline{p}_W$. (2) If the sample size differences $n_i - n_j$ are small, then $n_i \approx n_j = n$, for all i, j. Thus $\overline{p}_A = \left(\sum \frac{x_i}{n_i} \right) / K \approx \left(\sum \frac{x_i}{n} \right) / K = \frac{\sum x_i}{nK}$, while $\overline{p}_W = \frac{\sum x_i}{\sum n_i} \approx \frac{\sum x_i}{\sum n} = \frac{\sum x_i}{nK}$, leading to $\overline{p}_A \approx \overline{p}_W$.

Therefore, $\overline{p}_A$ and $\overline{p}_W$ will be different only if both the individual binomial proportion estimates and the sample size differences are large. Thus, the interrelationship between $\overline{p}_A$ and $\overline{p}_W$ can be summarized using the four possible cases shown in Table 2.1.

Table 2.1 A summary of the relationship between the binomial proportion estimators $\overline{p}_A$ and $\overline{p}_W$ under the assumption that p is constant for the population.

Sample sizes (n_i, n_j)	Individual proportion estimates ($y_i = x_i / n_i$)	
	Similar	Different
Similar	Case 1: $\overline{p}_A \approx \overline{p}_W$	Case 2: $\overline{p}_A \approx \overline{p}_W$
Different	Case 3: $\overline{p}_A \approx \overline{p}_W$	Case 4: $\overline{p}_A \neq \overline{p}_W$

In the context of this present study, in which a single binomially distributed population with constant p is assumed, it is clear that $\overline{p}_A$ and $\overline{p}_W$ will be similar for most situations, since conditions for Case 4 are not easily met. That is probably one of the important reasons why little attention has been made to differentiate between $\overline{p}_A$ and $\overline{p}_W$ in practical agricultural research. If the differences between the individual proportion estimates are too large, the assumption that there is a single underlying binomial proportion p may not hold. In such situations it may be appropriate to consider the data as a mixture of binomial distributions. Estimation of the mixing distribution is studied in Wood (1999). Similarly, if the differences in sample sizes are too large, it may cause some concern to pool all the samples for a single estimate of the binomial proportion. These are the considerations in modelling the binomial proportion in applied research.

2.5 SIMULATION STUDY

A simulation experiment was run to compare the two averaging methods via examining their means and variances at varying population proportions. At each of the nine equally-spaced population proportions p ranging from 0.1 to 0.9, a set of 400 random binomial samples ($K = 400$), comprising 100 each for sizes (n_i) of 10, 50, 100 and 200, respectively, was generated (hence n=36,000 and sample size differences were kept constant, estimated as $Var(n_i) = 71.152$, $\overline{n} = 90$, $CV_n = 0.791$) using Minitab 13 software. Estimates of $\overline{p}_A$ and $\overline{p}_W$ were obtained from the 400 samples over the range of population proportions. This process was repeated 200 times, resulting in 200 pairs of proportion estimates ($\overline{p}_A$ and $\overline{p}_W$), which then formed the data points for calculation of the observed means and variances of these two methods. Estimates of $Var(\overline{p}_A)$ and $Var(\overline{p}_W)$ as well as their ratio R are then presented, together with the theoretical values based on (2.1), (2.2) and (2.3), in Figure 2.3.

The results showed that the proportion estimates of the two averaging methods $\overline{p}_A$ and $\overline{p}_W$, when averaged over the 200 repetitions, were almost identical for each of the population proportion values (results not listed). This illustrated the theoretical derivation

in the previous section that both $\overline{p}_A$ and $\overline{p}_W$ are unbiased estimates of the population proportion for binomial data. The variances of $\overline{p}_A$ and $\overline{p}_W$, however, altered with p (Figure 2.3a). The weighted average approach consistently provided a more reliable estimate of the true proportion than the arithmetic average method, giving stable and low variance estimates over the range of population proportions (Figure 2.3a). The observed variance estimates generally agreed with the theoretical predictions for both methods over the range of population proportions, although for the arithmetic average there was a possible deviation from the predicted values at p=0.4 and p=0.5 (Figure 2.3a). The variance ratios R for these methods were stable across all nine population proportions, with deviations from the theoretical expectation being small for the observed ratio. The population proportion did not influence the magnitude of R, which was always smaller than one (Figure 2.3b). These results illustrate the theoretical interrelationships between the two methods in (2.1), (2.2) and (2.3).

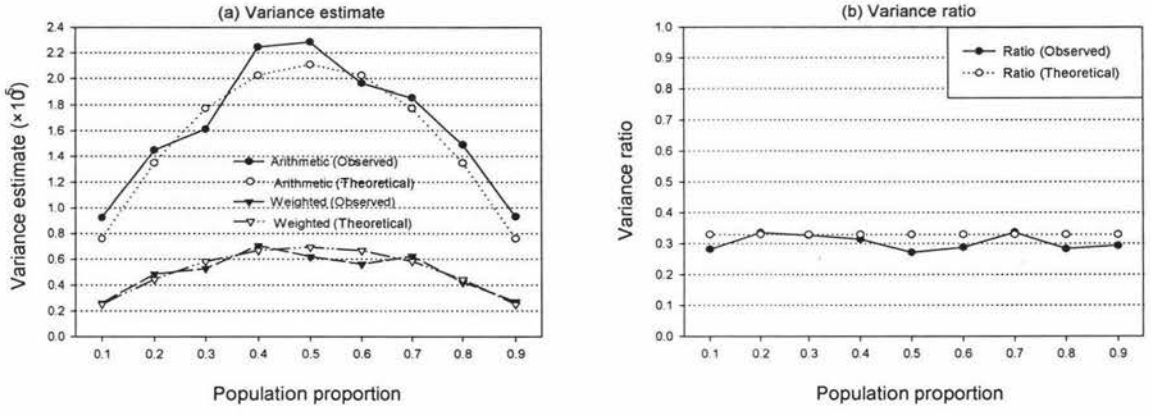

Figure 2.3 Showing a) variance estimates for weighted and arithmetic averages and b) their ratios from the 200 sets of 400 binomial samples from the simulation study. Also shown are the theoretical values for 400 samples of total size 36,000 with the specified sample sizes.

2.6 AN APPLICATION IN AGRICULTURAL RESEARCH

This concerns the author's investigation of sunflower lodging percentages in China in 1994, where a series of ratio estimates needed to be pooled or averaged over different environments (Qiao *et al.* 1994). 'Lodging' is a terminology used in agronomy to describe

field plants' falling onto the ground, thus always causing a substantial amount of grain yield loss in agricultural production. A survey was conducted in 1994 in the Western Region of Jilin Province, China to investigate a sunflower cultivar, *Improved Peredovic*, for lodging percentage in five locations where lodging was a severe problem. The results have already been listed in Table 1.1, but are repeated in Table 2.2 for convenience of the reader, where the number of lodged plants and the total number of plants sampled in each location are shown. The aim was to estimate the average or pooled lodging proportion of the sunflower cultivar in the whole region. It was assumed that all the locations were of equal importance in terms of commercial production for this cultivar. The whole survey was regarded as a multi-environment trial (MET) comprising five independent samples (locations) each estimating the same binomial proportion of lodging of *Improved Peredovic*. A pooled estimate of the proportion using data from all five samples was considered appropriate, since the proportion estimates of all five locations are similar.

Table 2.2 Estimates of lodging percentages for all five locations (counties) and a comparison of two lodging proportion estimators for sunflower cultivar *Improved Peredovic* in the Western Region of Jilin Province, China in 1994.

Location	Number of plants lodged x_i	Number of plants sampled n_i	Percentage of lodging $\hat{p}_i$
Baicheng	265	1560	17.0
Zhenlai	250	1840	13.6
Da'an	462	2413	19.1
Changling	518	3627	14.3
Nongan	464	3027	15.3
$\overline{p}_A$			15.9
$\overline{p}_W$			15.7
Variance estimate for the arithmetic average $Var(\overline{p}_A)$			0.000012
Variance estimate for the weighted average $Var(\overline{p}_W)$			0.000011
Variance ratio of the two estimates $R = Var(\overline{p}_W)/Var(\overline{p}_A)$			0.902

The first part of Table 2.2 details the data collected in each location (considered as a sample in this context). The second part of the table gives estimates (using our two estimators) of the population binomial proportion, the variances of the two estimators and their ratio

$R = Var(\overline{p}_W)/Var(\overline{p}_A)$. Using data from all the five locations, the average lodging percentage of sunflower cultivar *Improved Peredovic* is $\overline{p}_A$= 15.9% if the arithmetic average method is adopted, in comparison to a slightly lower estimate of $\overline{p}_W$ = 15.7% with the weighted average method (Table 2.2). The difference in the overall lodging percentage estimates seems to be small. The variance of $\overline{p}_A$ (0.00012), however, estimated using (2.1) in which p is replaced by $\overline{p}_A$, is larger than that of $\overline{P}_W$ (0.00011), estimated using (2.2) in which p is replaced by $\overline{p}_W$, with $R = Var(\overline{p}_W)/Var(\overline{p}_A) = 0.902$. These results show that even though the difference between $\overline{p}_A$ and $\overline{p}_W$ is small, $\overline{p}_W$ is still more appropriate than $\overline{p}_A$ in that the former has a smaller variance than the latter.

For this set of data, the differences between sample sizes are moderately large while the differences in the individual proportion estimates across the five locations are relatively small. This corresponds to Case 3 of Table 2.1 and hence the $\overline{p}_A$ and $\overline{p}_W$ are similar. This validates the finding of the earlier section that $\overline{p}_A$ and $\overline{p}_W$ will not differ noticeably if the differences in both sample sizes and the individual proportion estimates are not sufficiently large. Similarly, only a small difference is found between the calculated values of $\overline{p}_A$ and $\overline{p}_W$ for published data found from other available sources (Chen *et al.* 2000; Choi *et al.* 2000; Paderson and Brink 2000). This is because for most of the published data the differences between sample sizes n_i and between the separate sample proportion estimates are small, so that differences between $\overline{p}_A$ and $\overline{p}_W$ are also small. Data of this type can be found in Levine *et al.* (2000, p 131) and Kempthorne (1957, pp 152-154), where the $\overline{p}_A$ and $\overline{p}_W$ estimates were 73.3% and 72.9% for the former, and 52.0% and 51.5% for the latter. Examples of this type may explain why the issue of choice between the arithmetic and weighted average methods has not drawn sufficient attention from applied agricultural scientists, and why many keep using the arithmetic average approach. Alternatively, even if sample size differences are large for some samples, they get averaged out when the number of samples is large with many samples of similar sizes, leading to similar $\overline{p}_A$ and $\overline{p}_W$ estimates.

The differences in sample sizes and in individual proportion estimates will affect the difference in the two similar $\overline{p}_A$ and $\overline{p}_W$. Only if both of them are sufficiently large, can we expect a substantial difference between $\overline{p}_A$ and $\overline{p}_W$. It should be noted that this constant p model for estimating the binomial proportion was adopted in accordance with the local breeding practice in which the Western Region sampled by these five locations represents a similar environmental challenge (semi-dry and semi-arid). In situations where these locations were sampled from a mixture of binomial distributions with different p value at each location, other more appropriate alternative methods should be sought.

2.7 CONCLUSIONS

Two methods for estimating a binomial proportion p from several independent samples have been investigated. The first is the arithmetic average $\overline{p}_A$ and the second is the weighted average $\overline{p}_W$. Each method was shown to provide an unbiased estimate of p, but the weighted estimator always has the same or lower variance. The advantage of $\overline{p}_W$ becomes more evident as the differences between sample sizes and differences between individual estimates grow.

It is therefore recommended that the weighted average approach be used for averaging a series of proportions in agricultural research if the assumption holds that there is a single binomial population p. However, $\overline{p}_A$ has to be used when n_i and x_i are not known but only the y_i are available, for example, when the data collector only presents the percentages for different samples. If the assumption of a constant p does not hold (combining data from several distinct environments, for instance), the estimation of the mixing binomial distributions by Wood (1999) or the REML methodology should be considered.

CHAPTER 3
THE RATIO OF INDEPENDENT NORMALLY DISTRIBUTED VARIABLES

3.1 INTRODUCTION

In many practical applications, one wishes to estimate the ratio of the means of independent normal variables X and Y. Consider a pair of independent random variables $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. Given a sample (X_i, Y_i), $i = 1,...,n$, from the joint distribution, two estimators for μ_X / μ_Y are often used in the agriculture literature,

$$\overline{R}_A = \left(\sum_{i=1}^{n} \frac{X_i}{Y_i} \right) / n \text{ and } \overline{R}_W = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\overline{X}_n}{\overline{Y}_n}$$

We refer to $\overline{R}_A$ and $\overline{R}_W$ as the arithmetic average and the weighted average ratio estimators, respectively. Intuition suggests that $\overline{R}_A = \left(\sum \frac{X_i}{Y_i} \right) / n$ is a poor estimator of μ_X / μ_Y. This is because Y_i can be small and positive, leading to large and positive X_i / Y_i, thus biasing the final average upwards. It "averages after division". In contrast, $\overline{R}_W = \frac{\overline{X}_n}{\overline{Y}_n}$ should be a better estimator of μ_X / μ_Y as very small $\overline{Y}_n$ values are less likely to occur, thus lessening the upward bias. It "averages before division". Hence, $\overline{R}_W$ appears generally superior to $\overline{R}_A$.

It is clear from these formulae that both statistics involve ratios of independent normal random variables. This observation motivates us to take a close look at the distributional properties of X / Y. These properties will provide us with insight into the behaviour of the two estimators, to be addressed in Section 4.5. As will be discussed in the next chapter, many ratio-type estimators for μ_X / μ_Y have been proposed in the statistical literature, as reviewed by Tin (1965); however, we limit the scope of our investigation to the above two estimators. We note in passing that in the context of sample surveys, estimation of the ratio of two means is often required. The aims of this chapter are:

1) To present theoretical background on the distribution of the ratio of two independent normal variables X and Y;

2) To focus our attention on the critical quantity, the coefficient of variation of the denominator;

3) To provide the practical rule-of-thumb which indicates when the ratio of the sample means of two independent normal variables can be used satisfactorily.

3.2 THE RATIO OF INDEPENDENT NORMAL VARIABLES

We begin by considering the probability density function of the ratio $R = X/Y$, where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. Both of the ratio estimators $\overline{R}_A$ and $\overline{R}_W$ under consideration are directly related to R (as will be shown explicitly in the next chapter). It is well known that when $\mu_X = \mu_Y = 0$, R has a Cauchy distribution (Lukacs 1975, p43; Lukacs and Laha 1964, p56-57), the distribution of the central t with one degree of freedom. The Cauchy distribution does not possess finite moments of any order, in particular, the mean and variance do not exist (Johnson *et al.* 1994, p301). Springer (1979, p139) states "It is not surprising, therefore, that the distribution of the ratio of nonstandardized normal independent random variables also has this property (no absolute finite moments), since the standardized normal random variable is a special case of a nonstandardised normal random variable with mean zero and variance one." This suggests the non-existence of the moments of R in general. We will later demonstrate that the moments can exist when the denominator variable Y is bounded away from zero, through an examination of the probability density function of R. We now describe two approaches in the literature to the derivation of the probability density function of $R = X/Y$.

3.2.1 Springer's Approach

In general, the distribution of the quotient of two independent random variables can be obtained through Mellin convolutions and Mellin integral transforms (see, for example, Chapter 4 of Springer 1979). Springer (1979, p139-148) found the probability density function of $W = \dfrac{X/\sigma_X}{Y/\sigma_Y}$ and then $R = X/Y$ through the use of the simple transformation $R = (\sigma_X/\sigma_Y)W$. The density function of W has different expressions over four sections:

$$h(w)=\begin{cases} h_2^-(w), & -\infty < w < -1 \\ h_1^-(w), & -1 < w < 0 \\ h_1^+(w), & 0 < w < 1 \\ h_2^+(w), & 1 < w < \infty \end{cases}$$

The probability density function has removable discontinuities at $w=-1$, 0 and 1. Each of these component functions consists of four complicated infinite series involving gamma functions. Springer (1979, p148) gives a probability density function plot when $(\mu_X/\sigma_X)/(\mu_Y/\sigma_Y)=0.5$. It is apparent that the result given by Springer (1979) is rather unwieldy for computational purposes.

3.2.2 Kamerud's Approach

Kamerud (1978) gave the probability density function of $R=X/Y$ explicitly. There is an error in her derivation of the density function of W that we rectify in the following, making necessary subsequent adjustment to the density function.

Define $U=X/\sigma_X$, $V=Y/\sigma_Y$ and thus $U\sim N(\frac{\mu_X}{\sigma_X},1)$, $V\sim N(\frac{\mu_Y}{\sigma_Y},1)$. Set $W=U/V$ and let g be its density function. Then the density function of R can be obtained through a simple transformation $R=X/Y=(\sigma_X/\sigma_Y)W$. Replacing μ_1 and μ_2 in Kamerud (1978) by μ_X/σ_X and μ_Y/σ_Y, respectively, we have

$$g(w)=(2\pi)^{-1}Q\exp(M),\quad W=U/V$$

where $M=-\frac{1}{2}(\frac{\mu_Y}{\sigma_Y}w-\frac{\mu_X}{\sigma_X})^2 s^2$, $Q=ks(2\pi)^{1/2}[1-2\Phi(-k/s)]+2s^2\exp(-k^2/2s^2)$, $s=(w^2+1)^{-1/2}$, $k=(\frac{\mu_X}{\sigma_X}w+\frac{\mu_Y}{\sigma_Y})s^2$ and Φ is the standard normal cumulative distribution function. The probability density of R is then given by $f(r)=\frac{\sigma_Y}{\sigma_X}g\left(\frac{\sigma_Y}{\sigma_X}r\right)$.

In contrast to the method given in Springer (1979), Kamerud's expressions are easy to compute numerically. Hence, Kamerud's approach will be used to evaluate the distributional properties of the ratio of two normal variables in the next section.

3.3 DISTRIBUTIONAL PROPERTIES OF THE RATIO OF NORMAL VARIABLES

Though $R = X/Y$, $\mu_X, \mu_Y \neq 0$ is related to the Cauchy, this relationship becomes weaker as the coefficient of variation for denominator variable decreases. This will be demonstrated by the simulation on $\overline{R}_A$ and $\overline{R}_W$ conducted in Section 4.4. Alternatively, if the standard deviation σ_Y is sufficiently reduced while the two means are held constant, R behaves as if it has a distribution having a finite mean and standard deviation. The following theorem in stochastic convergence (see the next subsection), a graphical demonstration of the density function by Kamerud's approach and simulation studies will provide us some insights into the distributional properties of R.

3.3.1 Stochastic Convergence of the Ratio of Normal Variables

We consider two forms of stochastic convergence:

(1) $X_n \xrightarrow{P} X$, convergence in probability, if for every $\epsilon > 0$, $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.

(2) $X_n \xrightarrow{D} X$, convergence in distribution, if $F_n(x) \to F(x)$ as $n \to \infty$ at every continuity point x of F, where F_n and F are the cumulative density functions of X_n and X, respectively.

We note that $X_n \xrightarrow{P} X$ implies $X_n \xrightarrow{D} X$. If $X = c$, c being a constant, then $X_n \xrightarrow{P} c$ if and only if $X_n \xrightarrow{D} c$. We now present the following relevant results.

Lemma (Lukacs 1975, Corollary to Theorem 2.3.3)

Let $g(x, y)$ be a continuous function of the real variables x and y. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $g(X_n, Y_n) \xrightarrow{P} g(X, Y)$, as $n \to \infty$.

Theorem: Let $\overline{X}_n$ and $\overline{Y}_n$ be means of samples of size n, drawn independently from normal populations, with $\mu_X \neq 0$ and $\mu_Y \neq 0$. Then $\overline{X}_n / \overline{Y}_n \xrightarrow{P} \mu_X / \mu_Y$.

Proof: From the weak law of large numbers, $\overline{X}_n \xrightarrow{P} \mu_X$ and $\overline{Y}_n \xrightarrow{P} \mu_Y$. Take $g(x,y) = x/y$, $X_n = \overline{X}_n$, $Y_n = \overline{Y}_n$, $X = \mu_X$ and $Y = \mu_Y$ in the above lemma and the theorem follows immediately.

Thus $\overline{X}_n / \overline{Y}_n$ converges to μ_X / μ_Y in distribution. However, the graphs to be presented will show evidence of the existence of a mean of the ratio of independent variables under certain conditions, to be discussed in the next section (Figures 3.1-3.7). We note that if $X_1,...,X_n \sim N(\mu_X,\ \sigma_X)$ and $Y_1,...,Y_n \sim N(\mu_Y,\ \sigma_Y)$, then X_i / Y_i and $\overline{X}_n / \overline{Y}_n$ are ratios of two normal random variables. The only difference between them is that the coefficients of variation of the numerator and denominator of the latter are reduced by a factor of $\sqrt{n}$, i.e., $CV_{\overline{X}_n} = CV_{X_i} / \sqrt{n}$ and $CV_{\overline{Y}_n} = CV_{Y_i} / \sqrt{n}$. Thus the CV for both the numerator and denominator is influenced by the sample size n when the weighted average is adopted. The sample size affects the arithmetic average ratio estimator in a possibly less favourable manner, as we shall illustrate in Chapter 4.

3.3.2 Plots of the Density Function of the Ratio by Kamerud's Approach

TSome typical plots (Figures 3.1-3.7) are drawn using the density function of Kamerud, with varying population coefficients of variation for both the numerator and denominator variables. Note that a long tail or multi-peak is an indication that the moments of R may not exist. In Figures 3.1-3.3, the CV of Y is small (0.1) but the CV of X ranges from small (0.1), moderately large (0.5) to extremely large (5.0). All these three density functions are fairly symmetric around $\mu_X / \mu_Y = 1$, having a bell-shape like that of a normal distribution.

The long tail in Figure 3.4 and multi-peak in Figure 3.5, where the CV is small for the numerator but large for the denominator, clearly indicate that the moments, especially the mean of the distribution, may not exist. A moderately large CV creates a long tail in the distribution (Figure 3.6), while an extremely large CV creates a mean close to zero with symmetrical distribution (Figure 3.7), for both the numerator and denominator variables. In both cases there is evidence that the moments may not exist.

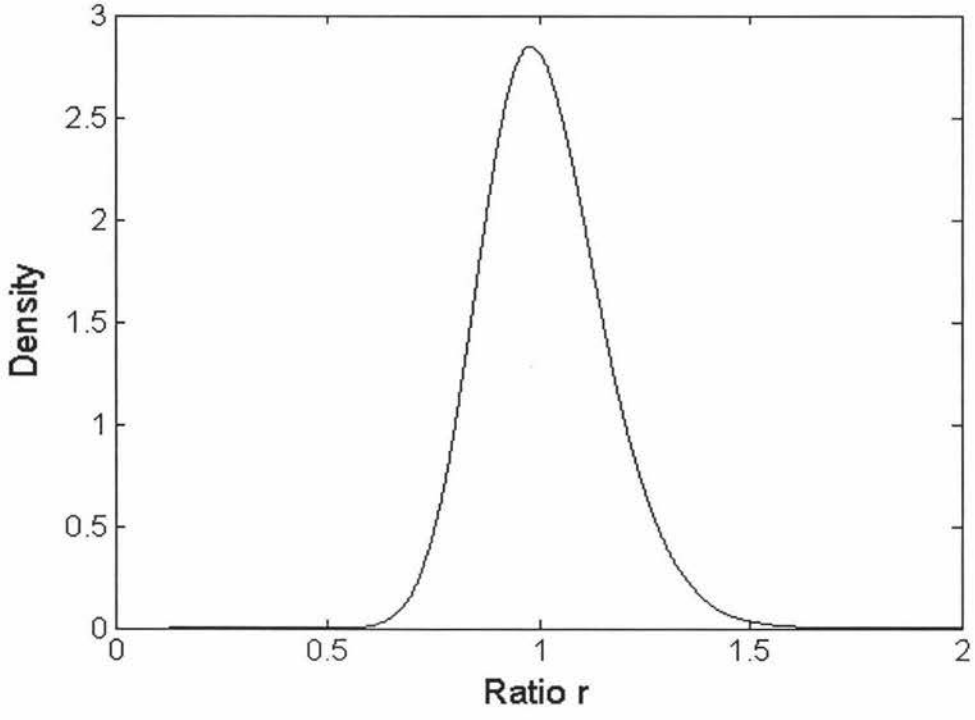

Figure 3.1 Density function for the ratio of two normal variables $X \sim N(100,10)$ and $Y \sim N(100,10)$. Here $\mu_X / \mu_Y = 1$, $CV_X = 0.1$, $CV_Y = 0.1$

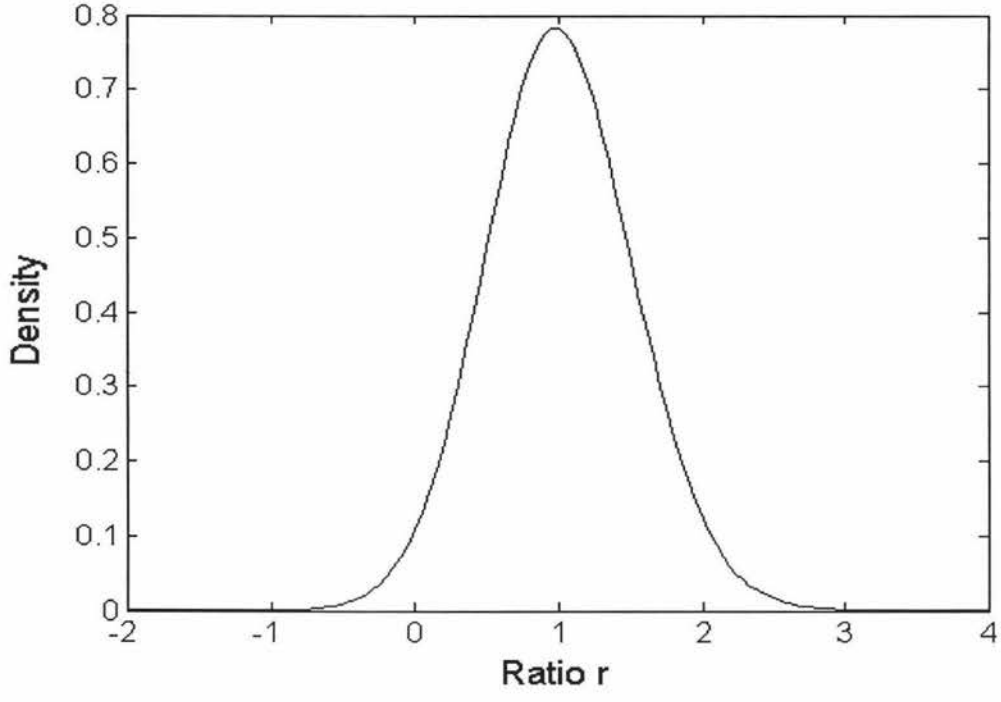

Figure 3.2 Density function for the ratio of two normal variables $X \sim N(100,50)$ and $Y \sim N(100,10)$. Here $\mu_X / \mu_Y = 1$, $CV_X = 0.5$, $CV_Y = 0.1$

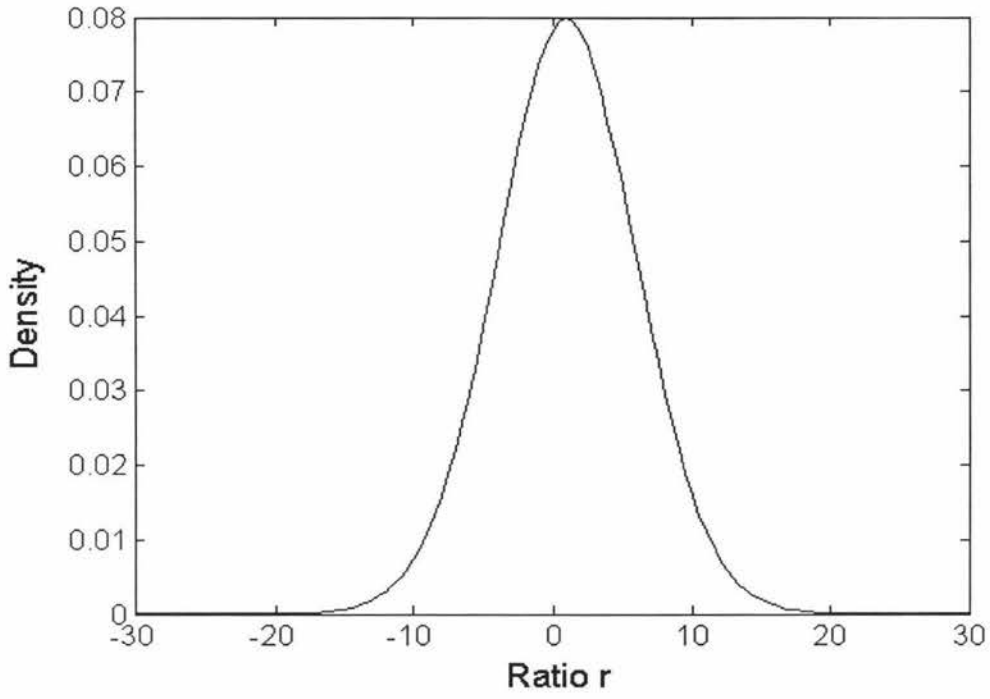

Figure 3.3 Density function for the ratio of two normal variables $X \sim N(100,500)$ and $Y \sim N(100,10)$. Here μ_X / μ_Y=1, $CV_X = 5.0$, $CV_Y = 0.1$

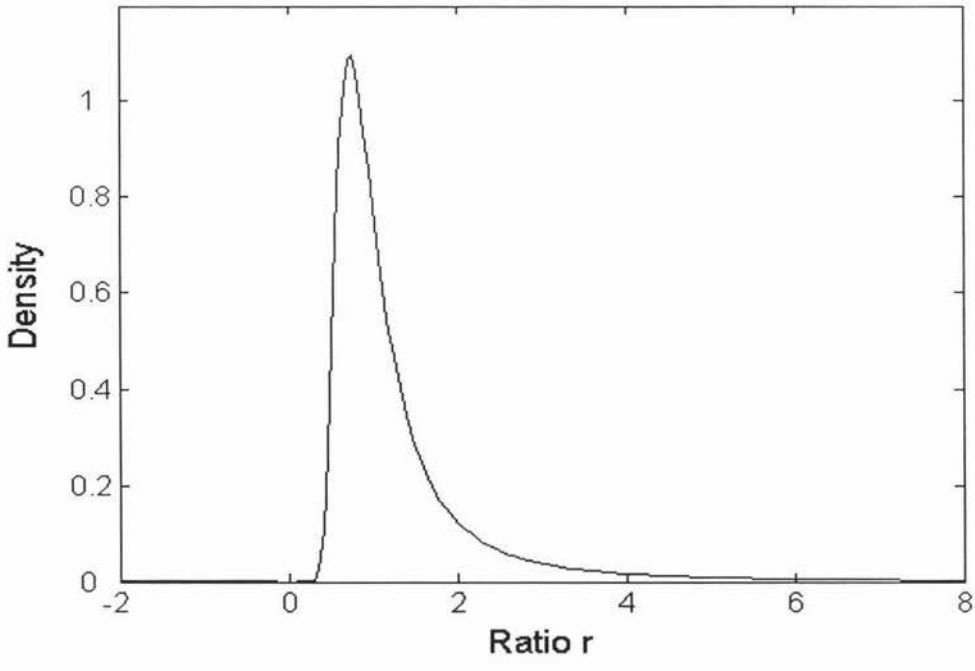

Figure 3.4 Density function for the ratio of two normal variables $X \sim N(100,10)$ and $Y \sim N(100,50)$. Here $\mu_X / \mu_Y = 1$, $CV_X = 0.1$, $CV_Y = 0.5$

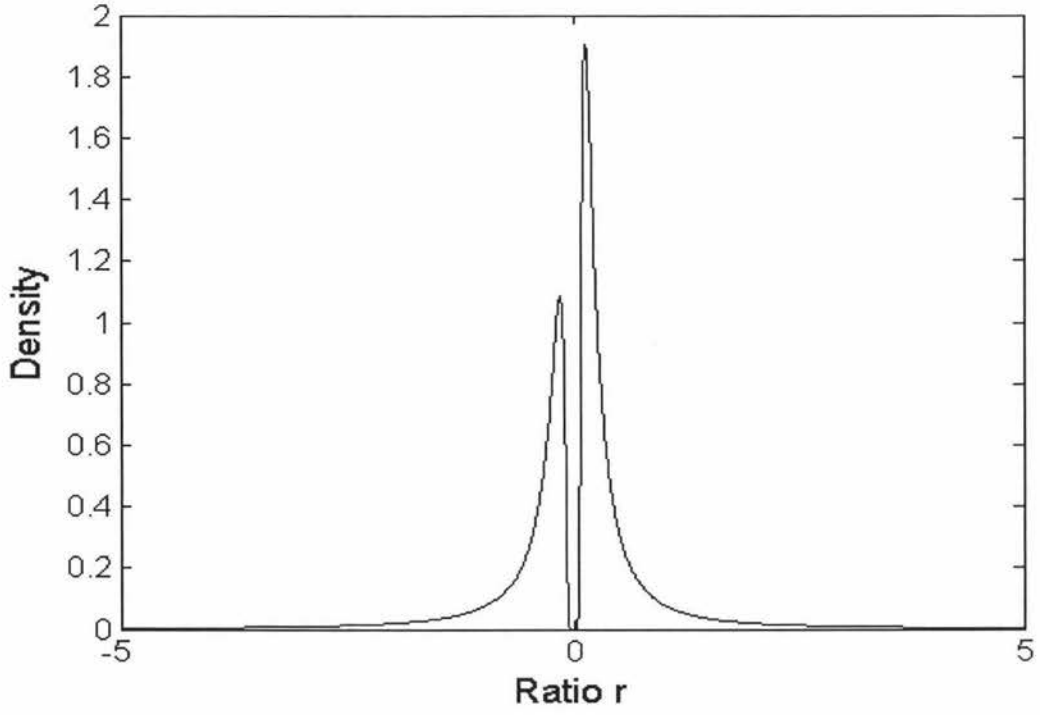

Figure 3.5 Density function for the ratio of two normal variables $X \sim N(100,10)$ and $Y \sim N(100,500)$. Here $\mu_X / \mu_Y = 1$, $CV_X = 0.1$, $CV_Y = 5.0$

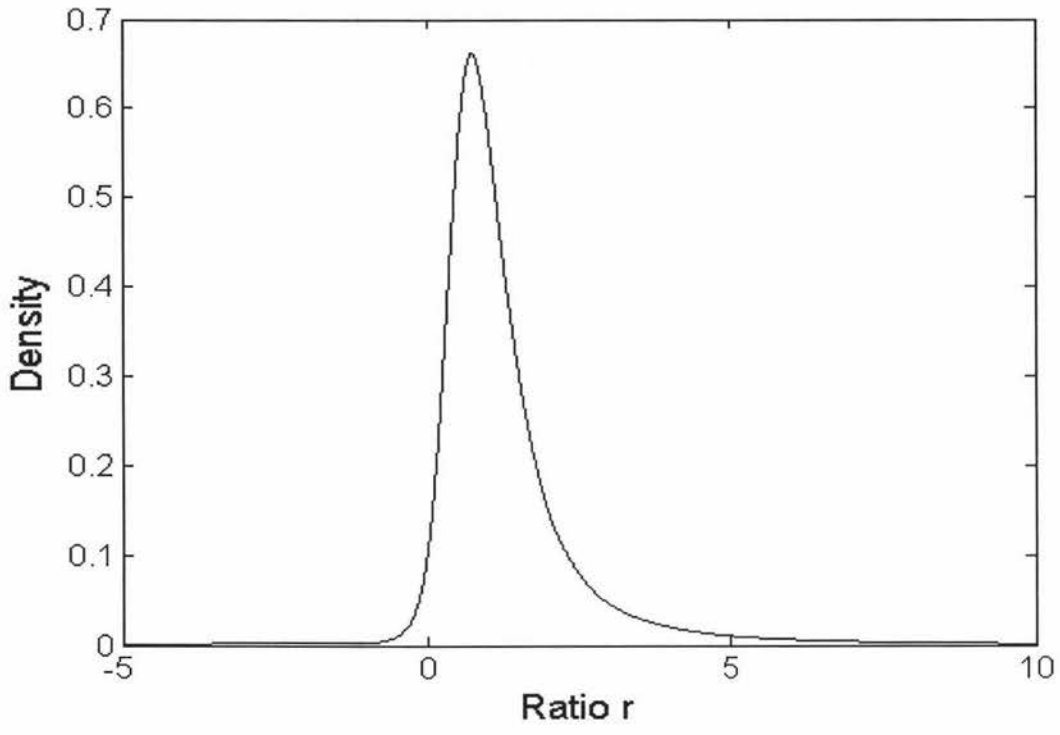

Figure 3.6 Density function for the ratio of two normal variables $X \sim N(100,50)$ and $Y \sim N(100,50)$. Here $\mu_X / \mu_Y = 1$, $CV_X = 0.5$, $CV_Y = 0.5$

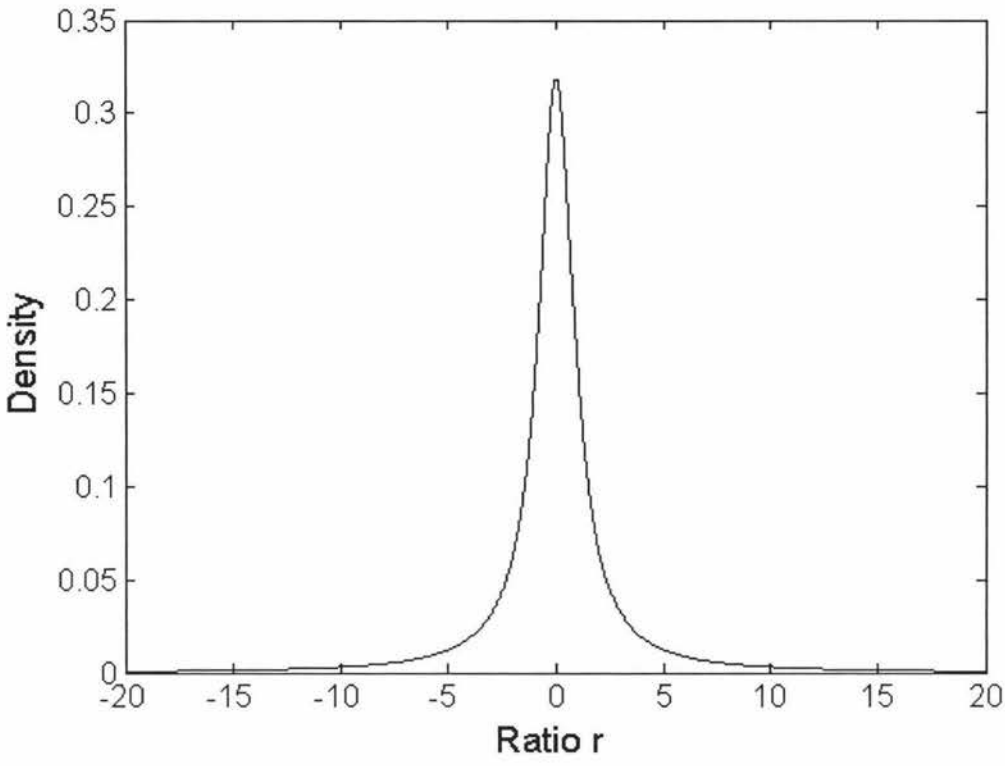

Figure 3.7 Density function for the ratio of two normal variables $X \sim N(100,500)$ and $Y \sim N(100,500)$. Here $\mu_X / \mu_Y = 1$, $CV_X = 5.0$, $CV_Y = 5.0$

For small CV of Y, the moments of the ratio of independent normal variables 'appears' toexcist. This is due to the fact that we did not sample very small Y values in the above graphical presentation and hence we were effectively sampling from $X/Y \| Y| > \varepsilon$, a punctured normal for the denominator variable. The moment of X/Y exists in this situation. Both the arithmetic and the weighted average methods take the form of a ratio of normal variables. We will demonstrate later that, as far as estimation of μ_X / μ_Y is concerned, both the arithmetic and the weighted average methods can be used when the CV of the denominator variable is sufficiently small. The circumstances beyond which the ratio of two normal variables cannot be used safely are now investigated using simulation.

3.3.3 Simulation of the Distribution of the Ratio of Normal Random Variables

Minitab 13 for Windows was used to simulate the distributional properties of the ratio $R = X/Y$ of two normal variables $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. The population means of both variables were fixed at 100 and hence $\mu_X / \mu_Y = 1$. The population

standard deviations of both variables took the values 10, 20,..., 100, 200 and 500, leading to both CV_X and CV_Y taking the values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0 and 5.0. For each of these 144 combinations, 500,000 pairs of (X_i, Y_i) were sampled; the mean, median, standard deviation and interquartile range of the ratios $R_i = X_i / Y_i$ were examined.

The simulation results, listed in Table 3.1, indicate that the sample mean, median, standard deviation and interquartile range of the ratio are all strongly influenced by the CV of the denominator variable. The effect of the CV, however, showed a different pattern for the sample moments, the mean and standard deviation, than for the median and interquartile range. When $CV_Y \leq 0.2$, the mean of R remains close to $\mu_X / \mu_Y = 1$, while the standard deviation of R increases approximately linearly as CV_X increases (Table 3.1). It appears that the variation of R is almost purely determined by the variation in the numerator variable when the CV of the denominator is small. It seems that $CV_Y = 0.2$ is an appropriate CV cut-off point for the denominator; for larger CV_Y values the mean deviates substantially from $\mu_X / \mu_Y = 1$ and the standard deviation increases accordingly. In contrast, the CV of the numerator seems to have no influence on the mean of R, and a relatively small influence on the standard deviation. Hence, the effect of increasing the CV of the denominator is much stronger than that for the numerator.

The sample mean of the ratios fails to estimate μ_X / μ_Y when $CV_Y > 0.4$, while the standard deviation is extremely large, with erratic behaviour, when $CV_Y > 0.3$. For the sample means to serve as reasonable estimators of μ_X / μ_Y for this sample size (500,000), CV_Y apparently has to be kept sufficiently small ($CV_Y \leq 0.2$ appears to suffice). In practical applied research, it is rare for the CV of a normal variable to be larger than 5.0. Thus, as long as $CV_Y \leq 0.2$, it makes empirical sense to use the ratio estimator X / Y.

The influence of the CV on the median of the ratios is similar to that which it has on the mean, except that the median deviates significantly from the true ratio only when

Table 3.1 Simulation of the ratio distribution: mean, median, standard deviation and interquartile range for 500,000 pairs of observations X_i / Y_i, where $X_i \sim N(\mu_X, \sigma_X)$ and $Y_i \sim N(\mu_Y, \sigma_Y)$, under varying coefficients of variation (CV), with $\mu_X / \mu_Y = 100/100 = 1$.

CV_X	CV_Y											
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2	5
						Mean						
0.1	1.011	1.047	1.129	1.487	4.260	1.544	-5.481	1.781	0.707	-0.055	0.022	-0.917
0.2	1.010	1.047	1.131	1.403	6.020	1.550	-7.294	1.948	0.600	0.077	0.081	-0.956
0.3	1.010	1.047	1.128	1.679	5.069	1.561	-3.521	1.655	0.686	-0.008	-0.091	-0.668
0.4	1.010	1.047	1.143	1.535	7.106	1.559	-2.912	1.746	0.637	-0.123	0.120	-1.112
0.5	1.011	1.047	1.142	1.226	5.337	1.892	-4.569	1.781	0.855	0.527	-0.683	-1.381
0.6	1.009	1.045	1.125	1.504	1.215	1.244	-6.581	1.585	0.591	-0.303	0.143	-0.703
0.7	1.009	1.046	1.136	1.432	8.460	1.567	-4.507	1.088	0.636	0.218	0.331	-0.867
0.8	1.010	1.047	1.149	1.335	8.038	1.147	-1.548	1.823	0.321	-0.393	0.423	-0.679
0.9	1.008	1.045	1.110	1.656	4.122	0.363	3.341	1.400	1.047	-0.431	0.189	-1.737
1.0	1.011	1.048	1.144	1.561	8.005	0.490	-11.990	1.694	1.171	-0.020	-1.204	-0.249
2.0	1.008	1.045	1.065	1.437	-7.078	2.247	-28.964	3.249	0.229	0.834	1.548	-0.311
5.0	1.011	1.044	1.186	0.415	15.293	0.490	28.645	4.248	1.806	0.340	-4.750	-3.286
						Median						
0.1	1.000	1.000	1.000	0.993	0.972	0.932	0.880	0.823	0.764	0.709	0.362	0.123
0.2	1.000	1.000	1.000	0.993	0.971	0.929	0.874	0.817	0.757	0.702	0.355	0.119
0.3	1.000	1.000	0.999	0.993	0.968	0.924	0.869	0.808	0.749	0.693	0.344	0.112
0.4	1.000	1.000	1.001	0.991	0.966	0.919	0.861	0.800	0.738	0.679	0.329	0.101
0.5	1.001	1.001	1.001	0.991	0.962	0.914	0.853	0.789	0.724	0.666	0.312	0.089
0.6	0.999	0.999	0.998	0.988	0.959	0.905	0.844	0.778	0.713	0.651	0.295	0.077
0.7	1.000	0.999	0.999	0.989	0.957	0.901	0.838	0.769	0.702	0.642	0.282	0.068
0.8	0.999	1.000	0.999	0.990	0.954	0.898	0.834	0.765	0.695	0.632	0.270	0.061
0.9	0.997	0.997	0.995	0.987	0.952	0.895	0.825	0.756	0.688	0.624	0.259	0.055
1.0	1.000	1.000	0.999	0.988	0.952	0.894	0.824	0.753	0.684	0.618	0.253	0.054
2.0	1.001	0.999	0.997	0.985	0.947	0.887	0.811	0.735	0.658	0.597	0.222	0.044
5.0	0.998	0.998	0.995	0.988	0.945	0.886	0.815	0.732	0.662	0.592	0.208	0.038
						Standard deviation						
0.1	0.102	0.531	12.554	169.4	2341.9	330.7	5124.0	557.7	462.1	329.3	371.3	422.2
0.2	0.213	0.629	10.723	130.6	3661.0	365.2	6744.2	662.4	453.5	275.2	434.4	423.5
0.3	0.309	0.532	12.206	267.2	2782.7	377.5	3412.4	549.3	460.9	319.2	513.2	282.8
0.4	0.421	0.601	4.410	223.8	4247.5	329.3	3334.7	658.6	491.8	343.4	485.5	539.4
0.5	0.534	0.855	6.951	94.3	3177.6	424.0	4680.2	505.3	484.9	237.0	532.9	693.2
0.6	0.672	0.760	14.109	273.5	711.9	330.5	6742.4	452.3	506.4	467.9	434.1	337.3
0.7	0.735	0.914	4.984	97.9	5225.1	391.5	4432.0	312.0	598.2	216.4	553.7	437.6
0.8	0.847	1.049	5.326	141.3	4677.4	355.9	3339.3	976.3	587.5	411.0	491.9	398.3
0.9	0.966	1.218	23.278	358.0	2113.1	324.9	1358.5	475.8	731.5	522.5	496.3	737.9
1.0	1.420	1.306	4.920	218.8	4331.1	531.1	9374.2	768.1	574.6	616.8	672.5	177.8
2.0	2.144	2.378	53.307	147.6	5112.0	553.1	20843.0	1827.4	528.8	505.3	697.1	340.5
5.0	5.135	5.857	30.811	963.1	10735.2	1100.5	18988.5	2293.6	1578.8	1649.8	1949.9	2840.4
						Interquartile range						
0.1	0.192	0.307	0.445	0.589	0.719	0.812	0.866	0.891	0.894	0.877	1.238	0.577
0.2	0.302	0.386	0.504	0.635	0.758	0.844	0.891	0.919	0.918	0.901	1.226	0.573
0.3	0.427	0.491	0.591	0.707	0.818	0.898	0.940	0.964	0.959	0.938	1.205	0.565
0.4	0.557	0.608	0.694	0.798	0.893	0.969	1.004	1.020	1.017	1.001	1.172	0.557
0.5	0.691	0.735	0.810	0.901	0.992	1.057	1.088	1.097	1.091	1.070	1.141	0.549
0.6	0.822	0.860	0.930	1.015	1.093	1.153	1.177	1.185	1.180	1.154	1.128	0.544
0.7	0.954	0.988	1.052	1.130	1.203	1.255	1.283	1.283	1.274	1.251	1.130	0.548
0.8	1.088	1.123	1.179	1.253	1.321	1.371	1.390	1.390	1.381	1.351	1.152	0.555
0.9	1.229	1.258	1.314	1.382	1.451	1.494	1.502	1.508	1.489	1.460	1.188	0.569
1.0	1.360	1.391	1.441	1.513	1.570	1.615	1.628	1.623	1.603	1.573	1.239	0.587
2.0	2.702	2.730	2.783	2.844	2.903	2.934	2.931	2.897	2.849	2.768	1.951	0.878
5.0	6.775	6.829	6.914	7.024	7.125	7.158	7.123	7.023	6.870	6.659	4.510	2.006

$CV_Y > 0.5$, a larger critical value. As CV_Y increases it has a less damaging influence on the estimate of the interquartile range than it does on the estimate of the standard deviation.

This simulation was repeated first with $\mu_X / \mu_Y = 10/100 = 0.1$, then with $\mu_X / \mu_Y = 100/10 = 10$. Tables 3.2 and 3.3 show the results. The mean, median, standard deviation and interquartile range of the ratios behave similarly to the case where $\mu_X / \mu_Y = 1$. This provides circumstantial evidence that the magnitude of μ_X / μ_Y does not influence the manner in which the sample mean and median estimate μ_X / μ_Y.

3.4 MOMENTS OF THE RATIO AND IMPLICATIONS

3.4.1 Moments of the Ratio of Normal Variables

It is well known that moments do not exist for the ratio of two standardised normal variables in the literature (Lucacs and Laha 1964; Lucacs 1975). Springer (1979) generalises the result to the ratio of two non-standardised normal variables. The latter work even provides mathematical proof for the non-existence of moments of the ratio of normal variables, but the derivation is somewhat complicated. Rao (1952) and Frishman (1975), however, argue for the existence of moments of the ratio of non-standardised normal variables, under what they described as 'mildly restrictive conditions'. These conditions are $0 < Y < 2\mu_Y$ for $\mu_Y > 0$, and $-2\mu_Y < Y < 0$ for $\mu_Y < 0$, $\mu_X, \mu_Y \neq 0$. Exact expressions are given for the mean and variance of the ratio of normal variables appropriate for these situations. Moses (1962) uses order statistics to estimate the ratio of two independent continuous positive random variables (known as Wilcoxon test theory) using simulation, but has a similar restriction problem. Rao and Beegle (1967) state that almost all the ratio estimators examined in their study involve linear functions of ratios of two correlated normal random variables and that the exact distributions of them are difficult to find.

We now present a simple proof of the non-existence of moments of the ratio of normal variables, with the denominator variable standardised. The conclusion can be easily extended to the case where both numerator and denominator variables are non-standardised normal variables.

Table 3.2 Simulation of the ratio distribution: mean, median, standard deviation and interquartile range for 500,000 pairs of observations X_i / Y_i, where $X_i \sim N(\mu_X, \sigma_X)$ and $Y_i \sim N(\mu_Y, \sigma_Y)$, under varying coefficients of variation (CV), with $\mu_X / \mu_Y = 10/100 = 0.1$.

CV_X	CV_Y											
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	2.0	5.0
						Mean						
0.1	0.101	0.105	0.114	0.123	0.163	0.238	0.157	0.322	-0.115	0.345	0.069	-0.049
0.2	0.101	0.105	0.115	0.124	0.159	0.234	0.155	0.278	-0.122	0.388	0.073	-0.045
0.3	0.101	0.105	0.114	0.130	0.151	0.223	0.159	0.354	-0.015	0.381	0.091	-0.089
0.4	0.101	0.105	0.116	0.119	0.152	0.356	0.159	0.353	-0.096	0.434	0.070	-0.124
0.5	0.101	0.105	0.114	0.133	0.158	0.313	0.135	0.422	-0.089	0.124	0.143	-0.018
0.6	0.101	0.105	0.112	0.142	0.177	0.312	0.185	0.165	-0.209	-0.080	0.052	-0.037
0.7	0.101	0.105	0.114	0.117	0.164	0.269	0.154	0.296	-0.163	0.357	0.098	0.041
0.8	0.101	0.104	0.116	0.110	0.152	0.307	0.141	0.132	0.012	0.281	0.164	-0.080
0.9	0.101	0.105	0.112	0.122	0.176	0.227	0.226	0.347	-0.314	0.596	0.173	-0.010
1.0	0.101	0.105	0.115	0.097	0.175	0.358	0.163	0.280	0.169	0.809	0.075	0.002
2.0	0.101	0.105	0.110	0.131	0.298	0.384	0.216	0.677	-0.337	1.769	0.059	-0.041
5.0	0.101	0.105	0.122	0.144	-0.008	0.065	0.284	1.697	0.306	-0.871	0.263	-0.126
						Median						
0.1	0.100	0.100	0.100	0.099	0.097	0.093	0.088	0.082	0.076	0.071	0.036	0.012
0.2	0.100	0.100	0.100	0.099	0.097	0.093	0.088	0.082	0.076	0.070	0.036	0.012
0.3	0.100	0.100	0.100	0.099	0.097	0.093	0.087	0.081	0.075	0.069	0.035	0.011
0.4	0.100	0.100	0.100	0.099	0.096	0.092	0.086	0.080	0.074	0.068	0.033	0.010
0.5	0.100	0.100	0.100	0.099	0.096	0.091	0.085	0.079	0.073	0.066	0.031	0.009
0.6	0.100	0.100	0.100	0.099	0.096	0.091	0.085	0.078	0.071	0.065	0.030	0.008
0.7	0.100	0.100	0.100	0.099	0.096	0.090	0.084	0.077	0.071	0.064	0.028	0.007
0.8	0.100	0.100	0.100	0.099	0.095	0.090	0.083	0.076	0.070	0.063	0.027	0.006
0.9	0.100	0.100	0.100	0.099	0.095	0.090	0.083	0.076	0.069	0.062	0.026	0.006
1.0	0.100	0.100	0.100	0.099	0.095	0.090	0.083	0.075	0.068	0.062	0.025	0.005
2.0	0.100	0.100	0.100	0.098	0.094	0.088	0.081	0.073	0.066	0.059	0.022	0.004
5.0	0.100	0.100	0.100	0.098	0.095	0.088	0.082	0.074	0.066	0.060	0.022	0.004
						Standard deviation						
0.1	0.015	0.027	1.729	12.062	24.622	74.680	20.090	132.284	110.835	178.748	52.971	43.921
0.2	0.023	0.033	1.712	10.313	23.268	74.447	22.138	101.829	111.236	207.974	58.041	41.235
0.3	0.032	0.041	1.915	16.175	22.526	63.122	18.615	140.737	49.251	221.688	67.039	67.660
0.4	0.042	0.050	2.246	5.390	30.073	155.84	18.629	154.039	111.595	222.907	66.641	89.348
0.5	0.052	0.059	2.000	16.707	27.427	120.73	14.768	176.467	96.660	48.214	95.498	26.681
0.6	0.062	0.069	2.156	19.025	26.651	137.86	22.886	49.580	192.556	104.768	29.967	30.854
0.7	0.072	0.079	2.607	5.238	29.076	103.95	23.912	189.196	159.395	170.832	84.168	19.081
0.8	0.082	0.090	1.861	4.589	26.437	126.98	21.198	45.853	52.381	128.222	102.073	58.453
0.9	0.092	0.100	1.645	9.744	46.162	74.601	34.863	154.652	191.999	361.484	148.660	15.828
1.0	0.102	0.111	1.440	12.816	31.276	151.06	23.839	118.560	134.042	512.798	41.014	33.141
2.0	0.203	0.217	2.441	6.898	54.132	163.05	78.355	324.828	246.458	1192.91	47.787	57.508
5.0	0.507	0.538	11.342	39.494	106.436	109.97	113.483	836.927	340.641	747.149	309.250	111.677
						Interquartile range						
0.1	0.019	0.031	0.044	0.059	0.072	0.081	0.087	0.089	0.089	0.088	0.123	0.058
0.2	0.030	0.039	0.050	0.064	0.075	0.085	0.090	0.092	0.091	0.090	0.122	0.057
0.3	0.043	0.049	0.059	0.071	0.082	0.090	0.095	0.096	0.096	0.094	0.120	0.056
0.4	0.056	0.061	0.069	0.080	0.090	0.097	0.101	0.102	0.101	0.100	0.117	0.056
0.5	0.069	0.073	0.081	0.090	0.099	0.106	0.109	0.110	0.109	0.107	0.114	0.055
0.6	0.082	0.086	0.093	0.101	0.109	0.115	0.118	0.118	0.118	0.116	0.112	0.055
0.7	0.096	0.099	0.105	0.113	0.121	0.126	0.129	0.129	0.127	0.125	0.113	0.055
0.8	0.109	0.112	0.118	0.125	0.132	0.137	0.140	0.139	0.137	0.135	0.115	0.056
0.9	0.123	0.126	0.131	0.138	0.144	0.150	0.151	0.150	0.148	0.146	0.118	0.057
1.0	0.136	0.139	0.144	0.150	0.157	0.162	0.164	0.162	0.160	0.157	0.124	0.059
2.0	0.271	0.274	0.279	0.285	0.291	0.295	0.294	0.291	0.285	0.277	0.194	0.088
5.0	0.675	0.681	0.690	0.700	0.710	0.714	0.712	0.701	0.684	0.663	0.451	0.199

Table 3.3 Simulation of the ratio distribution: mean, median, standard deviation and interquartile range for 500,000 pairs of observations X_i / Y_i, where $X_i \sim N(\mu_X, \sigma_X)$ and $Y_i \sim N(\mu_Y, \sigma_Y)$, under varying coefficients of variation (CV), with $\mu_X / \mu_Y = 100/10 = 10$.

CV_X	CV_Y											
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2	5
						Mean						
0.1	10.100	10.465	11.284	15.989	10.988	4.054	16.340	26.283	9.724	11.057	5.299	0.812
0.2	10.104	10.468	11.280	16.249	10.831	4.568	15.347	25.864	8.591	7.460	4.643	1.253
0.3	10.101	10.469	11.221	17.916	10.565	2.448	19.678	20.277	7.144	11.582	2.661	0.645
0.4	10.088	10.452	11.309	15.420	12.385	3.550	16.206	23.086	1.370	11.999	5.693	1.056
0.5	10.093	10.461	11.136	20.126	12.231	5.116	6.082	20.347	2.420	2.072	5.596	0.411
0.6	10.121	10.483	11.262	16.967	9.316	5.705	10.165	24.524	-4.077	10.534	0.947	0.845
0.7	10.110	10.481	11.311	17.965	10.173	7.782	13.993	23.548	-7.372	-1.464	-2.376	1.364
0.8	10.099	10.469	11.312	15.948	11.120	3.496	25.136	31.595	13.340	12.825	12.931	0.913
0.9	10.110	10.476	11.549	14.351	10.175	5.371	-7.253	29.567	14.069	-3.172	-5.250	0.500
1.0	10.136	10.498	11.488	20.622	13.746	5.802	23.218	31.051	-14.000	13.625	9.095	0.731
2.0	10.087	10.435	10.716	22.070	5.573	-1.373	-5.028	5.297	19.288	34.986	1.231	-1.021
5.0	10.082	10.465	11.606	-2.937	13.747	7.392	-10.072	-4.896	105.103	21.729	-1.531	5.674
						Median						
0.1	9.995	10.001	9.989	9.949	9.717	9.317	8.791	8.216	7.648	7.078	3.629	1.224
0.2	10.002	10.008	9.999	9.953	9.711	9.296	8.767	8.167	7.601	7.020	3.570	1.182
0.3	9.992	10.001	9.985	9.943	9.681	9.250	8.696	8.083	7.488	6.923	3.457	1.108
0.4	9.984	9.990	9.970	9.904	9.638	9.183	8.604	7.976	7.374	6.788	3.300	1.000
0.5	9.994	9.998	9.985	9.900	9.604	9.118	8.519	7.871	7.250	6.656	3.134	0.874
0.6	10.018	10.027	10.000	9.921	9.610	9.101	8.466	7.789	7.178	6.544	2.976	0.768
0.7	10.015	10.023	10.011	9.929	9.572	9.040	8.409	7.702	7.051	6.437	2.830	0.672
0.8	10.004	10.001	10.000	9.896	9.543	8.992	8.364	7.636	6.975	6.315	2.698	0.602
0.9	9.998	10.012	9.985	9.918	9.531	8.964	8.308	7.577	6.901	6.241	2.608	0.551
1.0	10.038	10.041	10.018	9.940	9.570	8.997	8.327	7.545	6.876	6.211	2.542	0.523
2.0	9.968	9.983	9.966	9.864	9.496	8.867	8.126	7.336	6.618	5.976	2.251	0.400
5.0	9.980	10.035	9.989	9.831	9.534	8.725	8.075	7.333	6.542	5.850	2.123	0.401
						Standard deviation						
0.1	1.521	2.772	102.7	2646.4	1622.5	3154.1	8410.7	7482.3	8189.5	7103.0	3914.4	505.2
0.2	2.364	3.369	100.0	3313.1	1473.5	2815.2	10582.5	7541.6	9925.0	5380.9	3789.1	526.3
0.3	3.209	4.118	113.1	3406.2	1574.8	4130.8	10149.1	6631.4	9338.7	7672.6	3764.8	563.8
0.4	4.203	5.088	80.7	3147.5	1271.5	3733.2	9306.7	6163.6	10652.8	7817.8	4635.4	540.2
0.5	5.227	5.924	171.0	4322.7	1473.2	3712.1	7955.8	5908.3	11372.3	5508.6	3427.7	517.9
0.6	6.233	6.931	143.7	3302.9	2177.1	3213.4	5801.7	8414.5	13135.2	6287.7	2653.8	605.7
0.7	7.281	7.947	92.6	4296.5	2518.6	2968.1	12075.4	7998.9	11770.5	5537.3	3073.5	745.1
0.8	8.263	9.106	108.7	2373.8	1002.6	4330.4	11380.2	8813.7	12924.8	4620.3	7722.5	689.6
0.9	9.246	10.052	120.2	4335.0	2436.5	2721.5	20303.4	10480.3	15518.6	4865.7	3900.3	760.6
1.0	10.257	11.876	127.8	4876.4	2097.8	2902.9	13801.6	7912.3	23560.8	11680.6	7562.5	625.4
2.0	20.305	21.608	268.2	5176.5	2192.7	5685.7	30904.2	8492.6	12070.0	14693.0	4667.9	1195.6
5.0	50.867	53.820	356.0	9897.7	7196.8	8457.6	35447.5	19662.5	44975.9	24106.7	10272.4	2496.9
						Interquartile range						
0.1	1.915	3.074	4.436	5.892	7.193	8.129	8.655	8.874	8.899	8.813	12.375	5.771
0.2	3.024	3.871	5.041	6.351	7.600	8.436	8.948	9.158	9.129	9.030	12.260	5.732
0.3	4.271	4.927	5.901	7.101	8.196	8.975	9.453	9.579	9.563	9.437	12.019	5.667
0.4	5.579	6.101	6.929	7.986	8.952	9.685	10.065	10.175	10.117	10.002	11.690	5.569
0.5	6.890	7.340	8.067	9.020	9.921	10.505	10.865	10.929	10.871	10.695	11.385	5.495
0.6	8.206	8.615	9.278	10.124	10.975	11.527	11.817	11.868	11.771	11.605	11.287	5.474
0.7	9.557	9.940	10.512	11.332	12.063	12.585	12.868	12.870	12.728	12.529	11.313	5.494
0.8	10.928	11.277	11.814	12.582	13.244	13.718	13.957	13.887	13.795	13.615	11.573	5.579
0.9	12.241	12.576	13.089	13.818	14.501	14.926	15.104	15.014	14.877	14.614	11.930	5.702
1.0	13.608	13.913	14.439	15.159	15.737	16.141	16.267	16.197	16.012	15.748	12.447	5.914
2.0	27.031	27.312	27.796	28.499	29.094	29.360	29.342	29.027	28.453	27.762	19.450	8.792
5.0	67.672	68.199	68.990	70.298	71.083	71.534	71.107	70.074	68.591	66.432	45.140	20.042

Theorem: $E(X/Y)$ does not exist, where $X \sim N(\mu_X, 1)$ and $Y \sim N(0, 1)$.

Proof:

$$E\left(\frac{X}{Y}\right) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{0-} \frac{x}{y} f_x(x) f_y(y) dx + \int_{0+}^{\infty} \frac{x}{y} f_x(x) f_y(y) dx \right\} dy$$

$$= \int_{-\infty}^{\infty} x f_x(x) \left\{ \int_{-\infty}^{0-} \frac{f_y(y)}{y} dy + \int_{0+}^{\infty} \frac{f_y(y)}{y} dy \right\} dx$$

$$= \frac{1}{\sqrt{2\pi}} \mu_X \left\{ \int_{-\infty}^{0-} \frac{e^{-\frac{1}{2}y^2}}{y} dy + \int_{0+}^{\infty} \frac{e^{-\frac{1}{2}y^2}}{y} dy \right\}$$

Now $\left\{ \int_{-\infty}^{0-} \frac{e^{-\frac{1}{2}x^2}}{x} dx + \int_{0+}^{\infty} \frac{e^{-\frac{1}{2}x^2}}{x} dx \right\} = \lim_{\varepsilon \downarrow 0} \int_{-\infty}^{-\varepsilon} \frac{e^{-\frac{1}{2}x^2}}{x} dx + \lim_{\varepsilon \downarrow 0} \int_{\alpha\varepsilon}^{\infty} \frac{e^{-\frac{1}{2}x^2}}{x} dx$, where $\alpha > 0$.

Let $u = \frac{1}{2}x^2$, $du = x\,dx$; Let $v = \frac{1}{\varepsilon^2}u$, $dv = \frac{2}{\varepsilon^2}du$, $\varepsilon > 0$.

Also $\int_{\varepsilon}^{\infty} \frac{e^{-\frac{1}{2}x^2}}{x} dx = \frac{1}{2}\int_{\frac{\varepsilon^2}{2}}^{\infty} \frac{e^{-u^2}}{x}\, du = \frac{1}{2}\int_{1}^{\infty} \frac{e^{-\frac{\varepsilon^2}{2}v}}{v}\, dv = \frac{1}{2}E_1\left(\frac{\varepsilon^2}{2}\right)$, from Abramowitz and Stegun (1964, 5.1.4 on p. 228)

Similarly $\int_{-\infty}^{\varepsilon} \frac{e^{-\frac{1}{2}x^2}}{x} dx = \frac{1}{2}E_1\left(\frac{\varepsilon^2}{2}\right)$, where $E_1\left(\frac{\varepsilon^2}{2}\right)$ is the exponential integral.

Thus, $\left\{ \int_{-\infty}^{0-} \frac{e^{-\frac{1}{2}x^2}}{x} dx + \int_{0+}^{\infty} \frac{e^{-\frac{1}{2}x^2}}{x} dx \right\} = \lim_{\varepsilon \downarrow 0} \frac{1}{2}\left[E_1\left(\frac{\alpha^2\varepsilon^2}{2}\right) - E_1\left(\frac{\varepsilon^2}{2}\right) \right]$

From Abramowitz and Stegun (1964, 5.1.11 on p. 229) we have

$$E_1(z) = -\alpha - \ln z - \sum_{n=1}^{\infty} \frac{(-1)^n z^n}{(n\,n!)}$$

So $E_1\left(\frac{\alpha^2\varepsilon^2}{2}\right) = E_1\left(\frac{\varepsilon^2}{2}\right) = -2\ln\alpha - \sum_{n=1}^{\infty}(-1)^n \left(\frac{\alpha\varepsilon^2}{2}\right)^n /(n\,n!) + \sum_{n=1}^{\infty}(-1)^n \left(\frac{\varepsilon^2}{2}\right)^n /(n\,n!)$

As $\varepsilon \to 0$, the right hand side converges to $-2\ln\alpha$. Since the limit depends on α, the first moment of X/Y, that is, $E(X/Y)$ does not exist. This can be extended to higher moments, which do not exist for a similar reason.

This theorem can be extended to the general case as follows. Let $X \sim N(\mu_X, \sigma_X)$ and

$Y \sim N(\mu_Y, \sigma_Y)$ be two independent normal variables. Clearly $E(X/Y) = E(X)E(1/Y)$. Thus $E(X/Y)$ does not exist if $E(1/Y)$ does not exist. Piegorsch and Casella (1985) noted in their Example 2.2 that if $f(y)$ is the normal density with mean μ and standard deviation σ, such that

$$f(y) = \left[1/(2\pi\sigma^2)^{1/2}\right] \exp\left[-1/2(y-\mu)^2/\sigma^2\right], \quad -\infty < y < \infty,$$

then $f(y) > 0$ and $E(Y^{-1})$ does not exist. Thus, $E(X/Y)$ does not exist.

3.4.2 Implications in Applied Research

The non-existence of moments of the ratio of normal variables presents a problem. In practical applications, as long as we avoid sampling in an interval around $Y = 0$, moments of X/Y will exist. If we let ε be a sufficiently small positive quantity, then X_i/Y_i can be used to estimate the ratio of μ_X/μ_Y, provided $|Y_i| > \varepsilon$. Hall (1979) showed that if a positive random variable Y has a singly truncated normal distribution from below, denoted by $N_a(\mu, \sigma)$, where $Y > a > 0$, then the inverse moments $E(Y^{-1})$ and $E(Y^{-2})$ can be approximated very accurately by expressions involving Dawson's integral, which are independent of the truncation point a, provided that $\left(\frac{\sigma}{\mu}\right)^2 \le \frac{a}{\mu} \le \frac{1}{25}$. This will ensure the existence of the expectation of the ratio of two independent normal variables $E(X/Y)$ when $\left(\frac{\sigma}{\mu}\right)^2 \le \frac{1}{25}$, or $CV_Y = \frac{\sigma}{\mu} \le \frac{1}{5} = 0.2$. Nahmias and Wang (1978) approximated $E(Y^{-n}; Y > t)$ for $t > 0$ and n a positive integer when Y is a normal variate having the property that the mean is substantially larger than the standard deviation, that is, $\mu >> \sigma$. According to the study, for any $n \ge 1$, this computation will require using numerical integration methods. The method developed is based on approximating the normal density by a gamma density, simplifying, and then re-approximating the resulting gamma density by the normal. They further validated the findings by comparing the estimated moments from the approximation and the exact moments calculated directly by numerical integration, where μ/σ is said to be relatively large if it is six or larger. This is almost equivalent to the constraint set by our investigations on the inverse of μ/σ, CV_Y, which should be held smaller than or equal

to 0.2 for the apparent existence of moments. Their results show that as μ/σ increases, the agreement between the approximation and the numerical integrals became closer across a range of values of μ, σ, t and n. The central idea behind all these investigations is similar, namely to make the denominator variable non-zero, a condition easily met in practical research.

The findings also suggest that if we want to use the sample mean of ratios X_i/Y_i to estimate μ_X/μ_Y then the larger the sample we use, the smaller the CV_Y we will need to avoid sample points getting close to zero in the denominator. When CV_Y is sufficiently small, there is almost no chance for a value of Y very close to zero being sampled, thus ensuring the existence of sample moments.

When CV_Y is very small, Y behaves as $Y \| Y| \geq \varepsilon$ for some $\varepsilon > 0$ (a punctured normal), thus the moments of $1/Y$ can be accurately approximated (Hall 1979; Nahmias and Wang 1978). This leads to "apparent" existence of the sample moments of X/Y. From our simulations and the results of Hall (1979), $CV_Y \leq 0.2$ can be used as a condition which determines the usefulness of $\overline{R}_A$ and $\overline{R}_W$. These imply that $\overline{R}_A$ can be used when $CV_Y \leq 0.2$ and $\overline{R}_W$ can be used when $CV_{\overline{Y}} \leq 0.2$.

3.5 CONCLUSIONS

The moments (mean and variance) of the ratio X/Y of two independent normal variables do not exist. The moments exist, however, if we avoid sampling points for which $|Y| \leq \varepsilon$, ε being a small positive quantity. This avoidance results in a punctured normal. The practical rule-of-thumb is that the ratio of two independent normal variables can be used to estimate μ_X/μ_Y when the coefficient of variation of the denominator variable is sufficiently small (smaller than or equal to 0.2).

This applies to the 'safe' use of the arithmetic and weighted average ratio estimators; $\overline{R}_A$ can be used when $CV_Y \leq 0.2$ and $\overline{R}_W$ can be used when $CV_{\overline{Y}} \leq 0.2$. Further research is needed to prove the existence of the moments of $X/Y \| Y| > \varepsilon$.

CHAPTER 4
COMPARISON OF TWO COMMON ESTIMATORS OF THE RATIO OF THE MEANS OF CONTINUOUS VARIABLES

4.1 INTRODUCTION

A ratio of continuos attributes (typically with normal distributions) is commonly used to assess the relative merits of two contrasting treatments, practices or methodologies in agricultural research. Examples include heterosis in plant breeding, the ratio of performance of a hybrid relative to its parents in a trait such as yield; the ratio of grain yield of a new crop variety to that of the commercial control across a range of environments; harvest index, the ratio between economic and biological yields of plants; and relative efficiency, the ratio between errors of two biological models in agricultural research. It is important to know how the true ratio of the two population means should be estimated when several such ratios are available.

Two methods are widely used for averaging different ratio estimates in agricultural research. The first is the arithmetic average approach, which divides the sum of all the ratio estimates by the total number of estimates (Kaeppler *et al.* 2000; Moreau *et al.* 1999; Qiao *et al.* 2000). The second is known as the weighted average approach, which estimates the true ratio via dividing the sum of all the numerators by the sum of all the denominators of the individual ratio estimates (Robinson *et al.* 1988; Haque *et al.* 1997; Witcombe *et al.* 1999). When used on the same set of data to estimate the ratio of two population means, these two approaches may give different results or even reach contradictory conclusions in some circumstances. It is a matter of practical importance to examine the relative merits of these two methods.

It should be noted that several alternative ratio estimators have been developed in other areas of research for estimating the ratio of two population means (Hartley and Ross 1954; Quenouille 1956; Mickey 1959; Durbin 1959; Pascual 1961; Kokan 1963; Tukey 1958; Tin 1965). They have not, however, attracted attention from agricultural scientists. Lahiri (1951) shows that the weighted average is an unbiased estimate of the ratio of the population means if the sample is drawn with probability proportional to the sum of the denominators. Cochran (1977) also defines the sample estimate of the true

ratio by the weighted average approach. In a Monte Carlo simulation study of various ratio estimators, Rao and Beegle (1967) find that the weighted average performs reasonably well in general relative to the other estimators under investigation, while the arithmetic average is not mentioned. We have not, however, found any report in the literature comparing the arithmetic and weighted average methods. The present study evaluates these two estimators of a ratio of normal means through simulation and provides some practical recommendations concerning their use in agricultural research. We conclude this section by remarking that related research was conducted in Chapter 2 where the corresponding estimators of a binomial proportion using several independent samples in agricultural research were investigated. That work provided the impetus for the current chapter.

4.2 LITERATURE REVIEW

In many practical applications, one wishes to estimate the ratio of the means of independent normal variables X and Y. Geary (1930) gave the distribution of $W = X/Y$ when $\mu_X = \mu_Y = 0$, where $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$ with correlation coefficient ρ. Fieller (1932) and Marsaglia (1956) considered the general problem with non-zero population means and studied the ratio in the form of

$$Z = \frac{a+X}{b+Y},$$

where $X \sim N(0,1)$ and $Y \sim N(0,1)$ and a and b are constants. Hinkley (1969) stated that Z has no great advantage over W, since the distributions of both involved the bivariate normal distributions. Hinkley further derived the general distribution of ratios of this type and compared it with the approximation obtained by assuming the denominator variable to be of constant sign. This may reflect many of the practical situations, for example the specifications of the US pharmacopoeia (Roberts 1969), where the denominator variables of a ratio are measurements of positive signs, but is obviously inappropriate for generalisation.

The ratio of continuous variables is often employed in sample surveys for estimating the population mean μ_Y of a characteristic of interest Y or the population ratio utilising a supplementary variate x that is positively correlated with y. In the area of sampling techniques the weighted average ratio estimator $\overline{R}_W = \overline{x}/\overline{y}$ is often referred to as the

classical ratio estimator (Cochran 1977). It is well known that the classical ratio estimator is biased and often, in practice, the bias may be negligible compared to the standard error and can be neglected (Rao and Beegle 1967). Kokan (1963) investigated large-sample stabilities of the variance and proposed the unbiased variance estimator of $\overline{R}_W$. Rao and Beegle (1967) presented a formula for estimating the variance of $\overline{R}_W$ and considered that the bias of the variance was of order $1/n$. However, no rule-of-thumb has been provided in deciding when the bias is negligible and hence $\overline{R}_W = \overline{x}/\overline{y}$ can be suitably used. On the other hand, the bias may become considerable in surveys with many strata of small or moderate size samples within strata if it is considered appropriate to use "separate ratio estimators" (Rao and Beegle 1967). In these situations, the use of unbiased or approximately unbiased (i.e., estimators with a smaller bias than the classical ratio estimator) ratio estimators may be of great advantage. Hence, considerable attention has been given to the development of unbiased or approximately unbiased ratio estimators. Several alternative ratio estimators have been developed for estimating the ratio of two population means (Hartley and Ross 1954; Quenouille 1956; Mickey 1959; Durbin 1959; Pascual 1961; Kokan 1963; Tukey 1958; Tin 1965).

Hartley and Ross (1954) gave an exact upper bound for the bias of $\overline{R}_W$ and proposed an unbiased ratio estimator of μ_X/μ_Y as

$$t_1 = \overline{R}_A + \frac{n}{(n-1)\mu_X}\left(\overline{y} - \overline{R}_A\overline{x}\right).$$

Goodman and Hartley (1958) showed that the variance of t_1 will always be larger than that of $\overline{R}_W$, for large n. Later on other types of ratio estimators were developed, based on dividing the sample at random into g groups, each of size m, as $n = mg$. Following Mickey (1959), another unbiased ratio estimator was given by

$$t_2 = \overline{r}_g + \frac{g}{\mu_X}\left(\overline{y} - \overline{r}_g\overline{x}\right),$$

where $\overline{r}_g = \sum_1^g \overline{R}_W'/g$ and $\overline{R}_W'$ is the classical ratio estimator computed from the sample after omitting the j-th group, that is, $\overline{R}_W' = (n\overline{y} - m\overline{y}_j)/(n\overline{x} - m\overline{x}_j)$, where $\overline{y}_j$ and $\overline{x}_j$ are the sample means computed from the j-th group. It is clear that t_2 reduced to t_1 for the

case of $n=2$.

Quenouille (1956) proposed a method of reducing estimation bias from order $1/n$ to $1/n^2$, based on random division of the sample into groups. Durbin (1959) applied this method to ratio estimators and proposed the following estimator $t_3 = g\overline{R}_W - (g-1)\overline{r}_g = \frac{1}{g}\sum_1^g r_{Qj}$, where $r_{Qj} = g\overline{R}_W - (g-1)\overline{R}_W'$ is called pseudo-values by Tukey and has bias of order n^{-2} at most. He showed that if the regression of Y on X is linear and X normally distributed, t_3 with $g=2$ has a smaller asymptotic variance than $\overline{R}_W$. Rao (1965) demonstrated that for the above models both asymptotic bias and asymptotic variance of t_3 are decreasing functions of g, so that $g=n$ would be the optimum choice for large or moderately large *n*. Durbin (1959) also considered the case where the regression of Y on X is linear, but X has a gamma distribution. He showed that although the variance of t_3 with $g=2$ is slightly increased compared to that of $\overline{R}_W$, the reduction in bias is such that the mean square error of t_3 is reduced. Rao and Webster (1966) showed that both bias and variance of t_3 are decreasing functions of g if X follows a gamma distribution, so that $g=n$ would be the optimum choice. For $g>2$, t_3 has a smaller variance than $\overline{R}_W$. Following Tukey (1958), Rao and Beegle (1967) presented the simple estimator

$$v(t_3) = g^{-1}(g-1)^{-1}\sum_1^g (r_{Qj} - t_3)^2$$

as the variance estimator of t_3, since the g estimators r_{Qj} may be treated as approximately independent and $t_3 = \sum_1^g r_{Qj}/g$. Tin (1965) investigated the large-sample bias, variance and approach to normality of the ratio estimators, $\overline{R}_W$, t_3 with $g=2$, and Beale's (1962) estimator

$$t_4 = \overline{R}_W \left(1 + \frac{1}{n}\frac{s_{xy}}{\overline{x}\,\overline{y}}\right) \Big/ \left(1 + \frac{1}{n}\frac{s_x^2}{\overline{x}^2}\right), \text{ and the modified estimator}$$

$$t_5 = \overline{R}_W \left[1 + \frac{1}{n}\left(\frac{s_{xy}}{\overline{x}\,\overline{y}} - \frac{s_x^2}{\overline{x}^2}\right)\right]$$

The results indicated that t_5 was slightly better than t_4 which in turn was better than t_3

with $g=2$. The comparison was validated using Monte Carlo study for large and moderately large samples. Rao and Webster (1966) made an exact comparison between t_5 and t_3 $g=2$ and found that their precisions were approximately the same. Pascual (1961) and Sastry (1965) proposed another ratio estimator with $g=n$ as

$$t_6=\frac{g}{g-1}\bar{R}_W-\frac{g}{g(g-1)}\sum_1^g\left(\bar{y}_j/\bar{x}_j\right),$$

which is identical to t_3 when $g=2$. Murthy and Nanjamma (1959) used t_6 when g independent and interpenetrating sub-samples each of size m were available. Pascual (1961) proposed the following estimator obtained by estimating the bias of $\bar{R}_W$ approximately

$$t_7=\bar{R}_W+\frac{1}{(n-1)\bar{X}}\left(\bar{y}-\bar{R}_A\bar{x}\right).$$

Pascual (1961) and Sastry (1965) investigated the properties of t_6 with $g=n$ and t_7, but Rao and Beegle (1967) described the investigations as "not very satisfactory", since some of the assumptions were set too stringent. Royall and Eberhardt (1975) compared four estimators for the variance of the ratio estimator under various linear prediction (super-population) models, for the estimation of the numerator variable of the ratio. These are the conventional statistic, the jack-knife estimator, a weighted least-square estimator from linear prediction theory and a new estimator from adjusting the conventional estimator in ways suggested by linear prediction theory. They chose to use the weighted average ratio estimator $\bar{R}_W$ for the estimation of the ratio of population numerator variable to denominator variable.

These ratio estimators have not, however, attracted attention from agricultural scientists. In contrast, the weighted average and the arithmetic average ratio estimators are widely used in applied research (Robinson *et al.* 1988; Haque *et al.* 1997; Witcombe *et al.* 1999; Kaeppler *et al.* 2000; Moreau *et al.* 1999; Qiao *et al.* 2000). Using the same set of data, these two approaches may give different results, or even provide contradictory conclusions in some circumstances. Hence it is of practical importance to examine the relative merits of these two methods.

4.3 TWO RATIO ESTIMATORS

4.3.1 Definitions of the Two Estimators of a Ratio

Suppose a sample of observations (X_i / Y_i), $i = 1, 2, ..., n$ is taken from a bivariate normal population $N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ and for each observation a ratio is calculated as X_i / Y_i. There are two popular ways in agricultural research to estimate the ratio of two population means μ_X / μ_Y, the arithmetic average approach, with $\overline{R}_A = \left(\sum \frac{X_i}{Y_i} \right) / n$, and the weighted average approach, with

$$\overline{R}_W = \sum W_i \frac{X_i}{Y_i} = \left[\left(\frac{Y_1}{\sum Y_i} \right) \left(\frac{X_1}{Y_1} \right) + \left(\frac{Y_2}{\sum Y_i} \right) \left(\frac{X_2}{Y_2} \right) + ... + \left(\frac{Y_n}{\sum Y_i} \right) \left(\frac{X_n}{Y_n} \right) \right] = \frac{\sum X_i}{\sum Y_i} = \frac{\overline{X}_n}{\overline{Y}_n}$$

If Y_i is the same for every X_i / Y_i in the sample, then $\overline{R}_A$ and $\overline{R}_W$ will be equal. When they are different, which is generally the case for real data in agricultural research, there can be a noticeable difference between the two estimators. We now show that the weighted average is superior to the arithmetic average. For many practical applications in agricultural research, the correlation coefficient between the two variables under study (X_i and Y_i) can be assumed to be zero, that is, X_i and Y_i are independent. Hence, we will only consider the cases in which X_i and Y_i are assumed to be independent.

4.3.2 Distributions of the Two Ratio Estimators

It was indicated in the preceding chapter that the population moments of the ratio of two normal variables X / Y do not exist unless we impose a condition that $|Y| > \varepsilon$. The rule-of-thumb is also defined as $CV_Y \leq 0.2$ for the 'safe' use of X / Y in agricultural research. When $CV_Y \leq 0.2$, the probability of getting values in $|Y| \leq \varepsilon$ is very small, where ε is a sufficiently small quantity, and hence X / Y behaves very much like $X / Y \| Y| > \varepsilon$. Consequently, a comparison of the variances of the two ratio estimators is meaningless if $X / Y \| Y| > \varepsilon$ cannot be assumed, since the standard deviations of $\overline{R}_A$ and $\overline{R}_W$ may not exist. We therefore adopt the arithmetic mean and the median as measures of central tendency, and the range and the interquartile range as measures of variation, in order to compare the two estimators, using simulation.

4.4 SIMULATION STUDIES

Random samples were generated, using Minitab 13 for Windows software, from two independent normal distributions, $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, to evaluate the relative merits of the two ratio estimators. The coefficients of variation of the two populations were assumed equal, or $\sigma_X / \mu_X = \sigma_Y / \mu_Y = CV$. A preliminary simulation was conducted to compare the distributions of the two estimators graphically. Systematic simulations were then conducted for a more in-depth evaluation of the distributions using parameter values typically found in agricultural studies. Central tendency, variation and skewness were examined for each of the two estimators.

4.4.1 A Preliminary Simulation

The distributions of $\overline{R}_A$ and $\overline{R}_W$ were simulated for the particular case where $\mu_X = \sigma_X = 200$ and $\mu_Y = \sigma_Y = 100$, hence $\mu_X / \mu_Y = 2.0$ and there is moderately large population variation ($CV = 1.0$). Two hundred samples, each of size $n = 300$, were drawn from each of the numerator and denominator populations. The distributions of $\overline{R}_A$ and $\overline{R}_W$ are graphically compared in Figure 4.1.

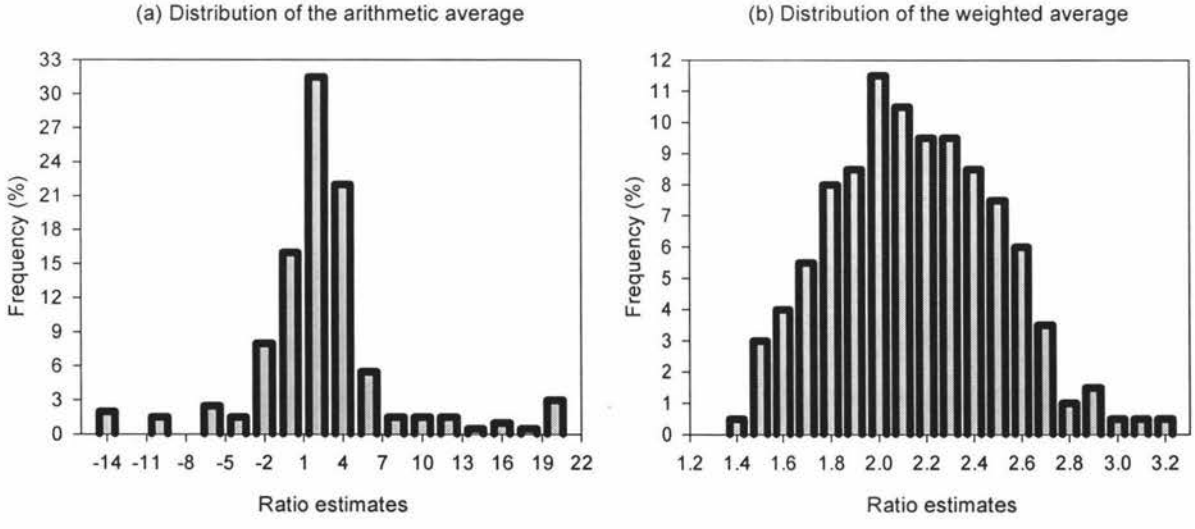

Figure 4.1 A comparison of $\overline{R}_A$ and $\overline{R}_W$ distributions using 200 ratio estimates. Each ratio was generated using a sample of size 300 from each of two normal populations, with $\mu_X = \sigma_X = 200$ and $\mu_Y = \sigma_Y = 100$, and hence μ_X / μ_Y=2.0 and $CV = 1.0$. Note the different scales used in (a) and (b). The means of $\overline{R}_A$ and $\overline{R}_W$ over the 200 estimates are -3.43 and 2.01, respectively.

Estimator $\overline{R}_W$ is far more concentrated around the true ratio (2.0) than is $\overline{R}_A$, and the variation of the former is much smaller than that of the latter. The distribution of $\overline{R}_A$ is also more right-skewed than that of $\overline{R}_W$, indicating that $\overline{R}_A$ gives some unusually large values while $\overline{R}_W$ is concentrated near the true ratio. In this fairly typical example it is evident that the weighted average is better at estimating the ratio of the two population means. The reason for this contrast is apparent and is explained as follows. For ratio estimator $\overline{R}_A$, since $X \sim N(200, 200)$ and $Y \sim N(100, 100)$, $CV_Y = 100/100 = 1.0 > 0.2$, as defined for the populations. Thus the $\overline{R}_A$ estimates are meaningless, since the expected value of X/Y and hence the expected value of $\overline{R}_A = \left(\sum_{i=1}^{300} \frac{X_i}{Y_i} \right) / 300$ does not exist, leading to the widely varying and biased $\overline{R}_A$ values. For ratio estimator $\overline{R}_W$, in contrast, $\overline{X}_{300} \sim N\left(200, 200/\sqrt{300}\right)$, $\overline{Y}_{300} \sim N\left(100, 100/\sqrt{300}\right)$, thus $CV_{\overline{Y}_{300}} = 100/\sqrt{300} = 0.0577 < 0.2$. Hence, the expected value of $\overline{R}_W = \overline{X}_{300} / \overline{Y}_{300}$ exists, leading to useful $\overline{R}_W$ values. In conclusion, $\overline{R}_W$ here is a better estimator of μ_X / μ_Y than $\overline{R}_A$.

4.4.2 Simulation 1: Comparison as μ_X / μ_Y and *CV* Change, with *n* Fixed

This simulation compared the two estimators as the ratio of population means (μ_X / μ_Y) and common coefficient of variation change, with $n = 300$ fixed throughout. Random samples were generated from two independent normal populations $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$. The means of the two populations were set at (1) $\mu_X = 500$, $\mu_Y = 100$; (2) $\mu_X = 200$, $\mu_Y = 100$; (3) $\mu_X = 100$, $\mu_Y = 100$; (4) $\mu_X = 100$, $\mu_Y = 200$; (5) $\mu_X = 100$, $\mu_Y = 500$. The *CV* was set at the values of 0.1, 0.2, 0.5, 1.0, 2.0 and 5.0. For each of the population mean and coefficient of variation combinations, 200 pairs of normal random samples of size $n = 300$ were taken. The sample size of $n = 300$ was considered to be large enough to capture the properties of the two ratio estimators in cases where resources are not limited.

Results (Table 4.1) showed that for the large sample size of $n = 300$ chosen, the arithmetic mean and median of $\overline{R}_A$ are very close to μ_X / μ_Y for low levels of

Table 4.1 A comparison of the centre, spread and skewness of the arithmetic average and the weighted average ratio estimators, for a range of true ratio and coefficient of variation (CV) values. Each result is based on 200 random samples of size 300 from normal populations.

CV	Arithmetic average					Weighted average				
	μ_X=500 μ_Y=100 μ_X/μ_Y=5.0	μ_X=200 μ_Y=100 μ_X/μ_Y=2.0	μ_X=100 μ_Y=100 μ_X/μ_Y=1.0	μ_X=100 μ_Y=200 μ_X/μ_Y=0.5	μ_X=100 μ_Y=500 μ_X/μ_Y=0.2	μ_X=500 μ_Y=100 μ_X/μ_Y=5.0	μ_X=200 μ_Y=100 μ_X/μ_Y=2.0	μ_X=100 μ_Y=100 μ_X/μ_Y=1.0	μ_X=100 μ_Y=200 μ_X/μ_Y=0.5	μ_X=100 μ_Y=500 μ_X/μ_Y=0.2
	Arithmetic mean									
0.1	5.05	2.02	1.01	0.51	0.20	5.00	2.00	1.00	0.50	0.20
0.2	5.23	2.09	1.05	0.52	0.21	5.01	2.00	1.00	0.50	0.20
0.5	9.52	0.73	1.72	0.69	-0.03	5.00	2.00	1.00	0.50	0.20
1.0	-0.09	-3.43	0.81	-0.01	-0.16	5.03	2.01	1.01	0.50	0.20
2.0	0.79	2.24	0.42	0.71	0.13	5.21	2.05	1.03	0.51	0.20
5.0	6.67	-2.61	1.13	-0.56	0.09	5.57	2.21	1.21	0.55	0.22
	Median									
0.1	5.05	2.02	1.01	0.51	0.20	5.00	2.00	1.00	0.50	0.20
0.2	5.24	2.09	1.05	0.52	0.21	5.00	2.00	1.00	0.50	0.20
0.5	6.46	2.42	1.30	0.63	0.26	5.00	2.00	1.01	0.50	0.20
1.0	4.58	1.36	0.91	0.36	0.17	5.03	2.02	1.01	0.49	0.20
2.0	0.76	0.27	0.39	0.15	0.07	5.13	2.04	1.00	0.51	0.20
5.0	-0.35	-0.04	0.07	-0.05	0.05	4.90	2.09	0.96	0.51	0.21
	Interquartile range									
0.1	0.05	0.02	0.01	0.01	0.00	0.05	0.02	0.01	0.00	0.00
0.2	0.14	0.05	0.03	0.01	0.00	0.11	0.04	0.02	0.02	0.00
0.5	4.12	1.13	0.71	0.40	0.13	0.25	0.10	0.06	0.03	0.01
1.0	9.31	3.13	1.63	0.93	0.36	0.51	0.25	0.10	0.06	0.02
2.0	10.86	3.41	2.05	0.90	0.39	1.17	0.48	0.23	0.12	0.05
5.0	12.55	4.53	2.01	0.98	0.41	2.49	1.19	0.53	0.29	0.13
	Range									
0.1	0.20	0.09	0.04	0.02	0.01	0.18	0.09	0.04	0.02	0.00
0.2	0.98	0.21	0.11	0.06	0.02	0.45	0.18	0.11	0.05	0.02
0.5	273.01	237.67	58.37	53.48	66.26	1.29	0.44	0.23	0.11	0.04
1.0	1388.17	856.90	56.50	132.34	77.33	2.26	0.87	0.42	0.21	0.11
2.0	1788.08	333.96	105.81	182.05	21.73	4.45	1.75	1.06	0.57	0.19
5.0	901.10	610.32	223.26	63.05	7.82	25.10	5.95	13.24	1.41	0.68
	Skewness									
0.1	-0.16	-0.12	-0.20	-0.07	3.93	-0.07	-0.13	-0.24	0.06	0.08
0.2	-1.24	0.08	0.10	-0.03	-0.42	0.08	-0.02	0.00	0.07	0.22
0.5	7.43	-8.02	8.79	1.09	-13.88	-0.11	0.36	0.16	0.23	-0.17
1.0	-9.72	-11.43	-2.05	-4.30	-13.51	0.26	0.02	0.44	0.25	0.45
2.0	1.67	8.13	1.54	6.09	1.05	0.45	0.24	0.98	0.66	0.37
5.0	8.99	-6.60	8.19	-3.98	2.13	3.44	1.07	7.24	1.11	1.13

population variation ($CV = 0.1$ and $CV = 0.2$). Under large population variation ($CV \geq 0.5$), these central tendency measures for $\overline{R}_A$ begin to deviate markedly from the corresponding true ratio, even producing changes of sign (Table 4.1). Thus, the performance of $\overline{R}_A$ is adversely affected even by a moderate amount of population variation. In contrast, the arithmetic mean and median of $\overline{R}_W$ remain stable and close to μ_X / μ_Y for the whole range of population CVs. The influence of the true ratio on the comparison between the two estimators becomes more marked as μ_X / μ_Y increases, when the arithmetic mean and median of $\overline{R}_A$ tend to deviate much more from the true ratio than do those of $\overline{R}_W$.

The range and interquartile range of $\overline{R}_A$ are generally small for low levels of population variation ($CV = 0.1$ and $CV = 0.2$), but become extremely large when the CV is 0.5 or bigger (Table 4.1). The larger the μ_X / μ_Y, the more the range and interquartile range of $\overline{R}_A$ deviate from the true ratios. In contrast, the range and interquartile range of $\overline{R}_W$ are generally small and unaffected by both population CV and μ_X / μ_Y, except for $CV = 5.0$ and $\mu_X / \mu_Y = 5.0$, when there is a large increase in both measures of variation. This shows that even with large sample size, the weighted average estimate can still deviate from the true ratio when the coefficient of variation or the mean ratio is extremely large. The effect of μ_X / μ_Y on $\overline{R}_W$, however, is generally much smaller than its effect on $\overline{R}_A$.

The reason why $\overline{R}_W$ performs better than $\overline{R}_A$ over the range of population CV values is discussed in the preceding section. Adopting $\overline{R}_W$ reduces the CV of the denominator and hence meets the minimum requirement of $CV_{\bar{Y}} \leq 0.2$ for most of the cases. In contrast, adopting $\overline{R}_A$ produces no change in the population CV and does not meet the minimum requirement of $CV_Y \leq 0.2$ for most of the situations listed in Table 4.1. For example, with the weighted average method, the denominator CV or $CV_{\bar{Y}}$ is estimated as 0.006, 0.012, 0.029, 0.058, 0.115 and 0.289, respectively, with only the last one (0.289) not satisfying the

$CV_{\bar{Y}} \leq 0.2$ condition. In comparison, for the arithmetic average method, the denominator CV or CV_Y is 0.1, 0.2, 0.5, 1.0, 2.0 and 5.0, respectively, as originally defined, with only the first and second values meeting the $CV_Y \leq 0.2$ condition for the moments to exist. Therefore, $\overline{R}_W$ is always better than $\overline{R}_A$.

We now turn to the skewness of the estimates. A near-zero skewness indicates a broadly symmetrical distribution. A large positive skewness indicates a strongly right-skewed distribution in which the mean is increased by some unusually high values, while a large negative skewness indicates a strongly left-skewed distribution in which the mean is decreased by some extremely low values (Levine *et al.* 2000). The distribution of $\overline{R}_W$ is generally more symmetrical, with fewer extreme values than are seen in the distribution of $\overline{R}_A$, as indicated by the skewness measure in Table 4.1.

To summarise, this simulation provides evidence that $\overline{R}_W$ is a better estimator of μ_X / μ_Y than $\overline{R}_A$ when the sample size is large.

4.4.3 Simulation 2: Comparison as *n* and μ_X / μ_Y Change, with *CV* Fixed

This simulation was conducted to investigate the effect of sample size n on the comparison between the two ratio estimators, with random samples drawn at varying sample sizes from two independent normal populations $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$. The means of the two populations were set at (1) $\mu_X = 2$, $\mu_Y = 1$; (2) $\mu_X = 1$, $\mu_Y = 1$; (3) $\mu_X = 1$, $\mu_Y = 2$, while the population coefficient of variation was held constant ($CV = 1.0$). For each of these three cases, 50 random samples of 5, 10, 20, 50, 100, 200, 500, 1000, and 10,000 pairs of normal observations were generated.

Results showed that when the population variation is set at a reasonably high level (CV=1.0), the arithmetic mean and median of $\overline{R}_A$ never improve and show no tendency to approach the true ratio, as sample size increases (Table 4.2). The corresponding measures of $\overline{R}_W$ lie consistently closer to the true ratio than those of $\overline{R}_A$ for each of the mean ratios

Table 4.2 A comparison of the centre, spread and skewness of the arithmetic average and the weighted average ratio estimators for a range of true ratio and sample sizes. The *CV* is held constant at one. Each result is based on 50 random samples of specific size, from normal populations.

Sample size	5	10	20	50	100	200	500	1000	10000
Arithmetic average (μ_X=2, μ_Y=1, μ_X/μ_Y=2.0)									
Mean	4.94	3.30	1.26	3.10	3.69	3.67	-3.21	4.29	-0.54
Median	1.82	1.23	1.71	1.51	1.47	2.27	1.98	1.91	1.23
Interquartile range	3.48	5.94	4.32	3.45	4.06	4.00	3.52	4.51	5.18
Range	1971.68	104.21	47.95	426.77	95.00	58.72	355.56	68.01	55.21
Skewness	-0.54	2.46	-1.30	-1.05	4.56	1.40	-5.56	2.84	-0.77
Arithmetic average (μ_X=1, μ_Y=1, μ_X/μ_Y=1.0)									
Mean	0.77	-1.00	2.69	1.99	-26.41	0.97	0.58	2.59	2.10
Median	0.45	0.34	0.77	0.61	0.18	0.80	0.79	1.02	0.95
Interquartile range	1.25	2.72	1.73	1.39	2.53	1.58	2.15	2.05	1.46
Range	38.95	63.66	63.37	77.10	1332.98	28.67	43.69	106.83	37.78
Skewness	4.97	-5.61	4.69	6.62	-7.07	3.68	1.19	0.29	4.38
Arithmetic average (μ_X=1, μ_Y=2, μ_X/μ_Y=0.5)									
Mean	-0.06	0.20	0.34	-0.16	0.46	0.48	-0.35	0.00	0.57
Median	0.40	0.41	0.20	0.26	0.50	0.39	0.49	0.10	0.65
Interquartile range	0.97	0.65	1.05	0.69	0.98	1.04	1.00	1.09	0.39
Range	47.63	12.18	11.52	9.44	21.39	15.74	31.25	26.65	6.44
Skewness	0.37	-2.73	1.05	-1.95	-2.00	0.05	-3.96	-0.96	-3.19
Weighted average (μ_X=2, μ_Y=1, μ_X/μ_Y=2.0)									
Mean	2.77	2.18	2.18	2.11	2.06	2.00	2.00	2.00	2.00
Median	1.89	2.12	2.15	2.14	2.06	2.00	2.00	2.00	2.00
Interquartile range	1.21	1.38	0.77	0.50	0.35	0.22	0.19	0.09	0.03
Range	22.49	5.77	5.01	2.15	1.40	0.99	0.58	0.34	0.12
Skewness	4.26	1.18	1.98	-0.15	0.05	0.02	0.11	-0.15	0.61
Weighted average (μ_X=1, μ_Y=1, μ_X/μ_Y=1.0)									
Mean	1.93	1.17	1.04	1.05	1.02	1.00	1.00	0.99	1.00
Median	1.03	1.14	0.92	1.02	1.00	0.99	1.00	0.99	1.00
Interquartile range	1.57	0.51	0.26	0.20	0.23	0.13	0.07	0.06	0.02
Range	30.97	2.59	2.47	1.15	0.70	0.54	0.25	0.23	0.06
Skewness	6.55	1.38	2.66	1.28	0.54	-0.08	-0.34	0.87	-0.34
Weighted average (μ_X=1, μ_Y=2, μ_X/μ_Y=0.5)									
Mean	0.21	0.54	0.48	0.53	0.50	0.51	0.50	0.50	0.50
Median	0.43	0.46	0.46	0.54	0.49	0.51	0.50	0.50	0.50
Interquartile range	0.40	0.29	0.18	0.16	0.11	0.06	0.05	0.03	0.00
Range	18.80	1.87	0.81	0.46	0.37	0.25	0.14	0.12	0.02
Skewness	-6.81	2.71	0.94	-0.46	0.70	0.73	0.13	0.52	0.05

and sample sizes. As sample size increases, both the mean and median of $\overline{R}_W$ approach the true ratio.

The reason for the weighted average method to outperform the arithmetic average method over the range of sample sizes from 5 to 10,000 is also directly related to the CV of the denominator populations. For the former, the denominator population $CV_{\overline{Y}}$, being influenced by sample sizes, is 0.447, 0.316, 0.224, 0.141, 0.100, 0.071, 0.045, 0.032, and 0.010, respectively over the range of sample sizes specified in Table 4.2. Thus, as sample size reaches 50, $CV_{\overline{Y}}$ falls under 0.2 for $\overline{R}_W$. For $\overline{R}_A$, in contrast, CV_Y remains the same (1.0) over all sample sizes, never meeting the minimum requirement. Hence, $\overline{R}_W$ is always better than $\overline{R}_A$.

The range and interquartile range are both large for $\overline{R}_A$ under varying sample sizes and ratios of population means, being always larger than those for $\overline{R}_W$. As sample size increases, $\overline{R}_A$ never stabilises when CV=1.0, while $\overline{R}_W$ stabilises around the true ratio for this CV (Table 4.2). The ratio μ_X / μ_Y does not appear to influence the comparison between the two estimators. The distribution of $\overline{R}_A$ is generally more skewed than $\overline{R}_W$, as evidenced by the skewness measures in Table 4.2.

In summary, $\overline{R}_W$ remains a universally better estimator of the true ratio than $\overline{R}_A$ over a range of sample sizes and mean ratios, under a reasonably high level of population variation.

4.5 THEORETICAL CONSIDERATIONS OF TWO RATIO ESTIMATORS

From the previous section it is evident that X_i / Y_i is a reasonable estimator of μ_X / μ_Y provided $CV_Y \leq 0.2$. This observation will provide the reason why $\overline{R}_W$ improves as the sample size n increases, while for $\overline{R}_A$ this is not the case; hence $\overline{R}_W$ will be regarded as a

superior estimator. We now examine $\overline{R}_A$ and $\overline{R}_W$ separately and conclude that $\overline{R}_W$ can be used if $CV_{\overline{Y}_n} \leq 0.2$, while $\overline{R}_A$ can be adopted if $CV_Y \leq 0.2$.

4.5.1 Weighted Average Ratio Estimator

Recall, $\overline{R}_W$ is called the weighted average estimator, so named because it can be written as

$$\overline{R}_W = \frac{\overline{X}_n}{\overline{Y}_n} = \left(\frac{Y_1}{\sum Y_i}\right)\left(\frac{X_1}{Y_1}\right) + \left(\frac{Y_2}{\sum Y_i}\right)\left(\frac{X_2}{Y_2}\right) + \ldots + \left(\frac{Y_n}{\sum Y_i}\right)\left(\frac{X_n}{Y_n}\right).$$

Since $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$, it follows that $\overline{X}_n \sim N\left(\mu_X, \sigma_X/\sqrt{n}\right)$ and $\overline{Y}_n \sim N\left(\mu_Y, \sigma_Y/\sqrt{n}\right)$. Hence, $CV_{\overline{X}_n} = \frac{\sigma_X/\sqrt{n}}{\mu_X}$ and $CV_{\overline{Y}_n} = \frac{\sigma_Y/\sqrt{n}}{\mu_Y}$. If $\mu_X, \mu_Y \neq 0$ and $\frac{\sigma_Y/\sqrt{n}}{\mu_Y} \leq 0.2$, then our simulations demonstrate that $\overline{R}_W = \overline{X}_n / \overline{Y}_n$ is an acceptable estimator of μ_X/μ_Y. Thus, for practical purposes, we recommend that $\overline{R}_W$ be used to estimate μ_X/μ_Y, since taking a sample of sufficiently large size n will reduce the coefficient of variation of $\overline{Y}_n$.

In designing a research experiment or survey, the sample size n needed to make a reasonably good estimate of μ_X/μ_Y can be determined in the following way. Take a sample of size n from $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. In order for $\overline{R}_W = \overline{X}_n / \overline{Y}_n$ to estimate μ_X/μ_Y, we require $\frac{\sigma_Y/\sqrt{n}}{\mu_Y} \leq 0.2$, or $n \geq 25\frac{\sigma_Y^2}{\mu_Y^2}$. In practical situations, the population means and standard deviations of interest are rarely known, but they can be estimated by the relevant sample means and standard deviations. Hence, the above inequality can be approximated by $n \geq 25\frac{s_Y^2}{\overline{Y}_n^2}$.

Realistically speaking, the sample size n is always either predetermined or data are analysed with known sample size. Thus, sample results can be examined to see if they

satisfy the requirement $\frac{s_Y/\sqrt{n}}{\bar{Y}_n} \leq 0.2$. This will provide a general guideline for evaluating the suitability of the weighted average method in estimating the ratio of the means of two normal variables.

4.5.2 Arithmetic Average Ratio Estimator

Estimator $\bar{R}_A = \sum_{i=1}^{n}(X_i/Y_i)/n$ is an equally weighted average of n ratios X_i/Y_i. We can adopt the same methodology used in evaluating the weighted average method to assess the suitability of $\bar{R}_A$. The coefficient of variation of Y_i, however, is $\frac{\sigma_Y}{\mu_Y}$ in this case, instead of $\frac{\sigma_Y/\sqrt{n}}{\mu_Y}$. If $\sigma_Y/\mu_Y > 0.2$, for example, then X_i/Y_i is a poor estimator of μ_X/μ_Y. Taking a larger sample size n is of little use. Naturally the sample value of $\frac{s_Y}{\bar{Y}_n}$ can be used as a diagnostic tool for evaluation of the appropriateness of $\bar{R}_A$. Thus, we recommend use of $\bar{R}_A$ only if the coefficient of variation of Y_i is sufficiently small, that is, $CV_Y = \frac{s_Y}{\bar{Y}_n} \leq 0.2$. The simulation results in Chapter 3 support our recommendation.

4.6 APPLICATION OF THE TWO ESTIMATORS IN RICE TRIALS

A rice breeding multi-environment trial (MET) was conducted in eleven locations in Jilin Province, China during 1995 and 1996 for an evaluation of grain yield performance of the new varieties developed by different plant breeding institutes (Jingyong Ma 1996, personal communication). Seven of the locations appeared in both years, while each of the other four locations was used in only one of the two years (Table 4.3). In such studies a subset of rice varieties are added in or dropped out from the regional variety testing program every year, based on their overall performance (mainly grain yield) relative to the control variety. This makes the field evaluation of rice varieties progress in a roll-over pattern.

Table 4.3 Grain yield performance of six rice varieties and the estimates of percent grain yield of each of the test varieties relative to the control by the arithmetic and weighted average in a multi-environment trial (MET) during 1995 and 1996.

Location	Grain yield (kg/ha)	Percentage of control (%)	Grain yield (kg/ha)	Percentage of control (%)	Grain yield (kg/ha)	Percentage of control (%)	Grain yield of control (kg/ha)
1995	Jiu 9214		Chang 90-40		Ji K911		Control[¶]
Changchun	7083	104.4	7358	108.5	7068	104.2	6783
Dongfeng	5733	117.8	5934	121.9	3867	79.4	4868
Gongzhuling	8604	100.8	8664	101.5			8535
Jilin	7800	107.7	7290	100.6			7245
Lishu	8168	112.0	8100	111.1	7650	104.9	7292
Tonghua	7955	101.8					7815
Yanbian	8532	105.2	8652	106.7	8283	102.2	8106
Yushu	10082	105.0	9963	103.8	9638	100.4	9600
Mean	7994	7530*	7994	7490*	7301	7330*	
Standard deviation	1255	1388†	1285	1494†	2144	1741†	
Coefficient of variation	0.157	0.184‡	0.161	0.200‡	0.294	0.237‡	
$\overline{R}_A$	106.8		107.7		98.2		
$\overline{R}_W$	106.2	$CV_{\bar{Y}} = 0.065$	106.7	$CV_{\bar{Y}} = 0.076$	99.6	$CV_{\bar{Y}} = 0.106$	
1996			Jiu 9421		Jiuhua 2		Control[¶]
Changchun			7041	103.5	6879	101.1	6804
Dongfeng					11801	122.9	9600
Gongzhuling			8381	102.1			8210
Jilin			8815	103.7			8501
Jilin Agricultural University			8095	94.4	7212	84.1	8571
Lishu			8151	94.2			8651
Shulan			7701	101.3	8100	106.6	7601
Tonghua			8358	105.2	8508	107.1	7945
Yanbian			7982	90.4			8834
Yongji			8271	101.6	7445	91.5	8138
Mean			8088	8139*	8324	8110*	
Standard deviation			498	630†	1804	941†	
Coefficient of variation			0.062	0.077‡	0.217	0.116‡	
$\overline{R}_A$			99.6		102.2		
$\overline{R}_W$			99.4	$CV_{\bar{Y}} = 0.026$	102.6	$CV_{\bar{Y}} = 0.047$	

*, † and ‡ denote mean, standard deviation and coefficient of variation, respectively, of the control variety across all the locations in which the grain yields of both the test variety and the control are available for that year. ¶ The control variety is Jiyin 12 for both 1995 and 1996.

The percent grain yield of each test variety relative to the control variety (Jiyin 12) is used to assess the yield improvement of the new variety at the locations in both years. The percent grain yield of each variety relative to the control, listed in Table 1.2, is presented again here in Table 4.3, for convenience of the reader. This is a typical application of a ratio of continuous variables (grain yield in this instance), where the grain yields of the test varieties and the control are independent of each other and both follow a normal distribution. Thus, the overall percent grain yield performance of each test variety over the control, averaged or pooled over all trials in the MET, is estimated using both the arithmetic and weighted average methods.

4.6.1 Estimation of the Pooled Percent Yield Improvement against Control

The results show that $\overline{R}_A$ and $\overline{R}_W$ are similar for two test varieties (Jiu 9421 and Jiuhua 2) in 1996, while they demonstrate some difference for the other three test varieties in 1995 (Table 4.3). The difference between $\overline{R}_A$ and $\overline{R}_W$ thus appears environment-dependent, being relatively large in 1995, but small in 1996. There is a degree of variation in the difference between $\overline{R}_A$ and $\overline{R}_W$, ranging from 0.2% to 1.4% (Table 4.3). From the plant breeding point of view, there is reason (to be discussed in the following section) to believe that the difference between $\overline{R}_A$ and $\overline{R}_W$ for rice varieties Chang 90-40 and Ji K911 is large enough to draw attention, while that for variety Jiu 9421 is negligible. The difference between the two estimators for the two remaining varieties (Jiu 9214 and Jiuhua 2) lies somewhere in between the above categories. It hence requires further investigation.

It is regulated by the Jilin Provincial Crop Variety Evaluation Committee (1995) that a new variety has to exceed the control in grain yield by at least 5% for self-pollinated crop species such as rice and wheat, while by 10% for cross-pollinated crop species such as corn or sunflower, before it can be considered for release and commercialisation. The regulation imposed by the Committee is most stringent, and it is usually difficult for a test variety to increase grain yield by an extra 1% against the control variety. Thus a 1% difference between the two ways of estimating the pooled ratio of the two rice varieties under comparison can make a real difference in deciding whether a particular variety should be

released. Therefore, based on the difference of 1% and 1.4% between $\overline{R}_A$ and $\overline{R}_W$ for Chang 90-40 and Ji K911, it is evident that the two ways of estimating the ratio of continuous variables can influence the decision of plant breeding in terms of recommendation for release and commercialisation. These findings indicate that the weighted average ratio estimator $\overline{R}_W$ should be used in practical agricultural research.

4.6.2 Application of the Diagnostic Approach in Rice Trials

The difference between these two estimators ranges from 0.2% to 1.4%, depending on the coefficient of variation for the denominator variable, the grain yield of the control. When the CV of the control is equal to or larger than 0.2, as in the case of Ji K911 and Chang 90-40, the two estimators differ by a reasonably large amount, being 1.4% and 1.0%, respectively. At the other extreme, in the case of Jiu 9421, the CV of the control is only 0.077 and hence there is almost no difference between $\overline{R}_W$ and $\overline{R}_A$. Between these two cases, the magnitude of the control CV almost determines the difference between $\overline{R}_W$ and $\overline{R}_A$. In general, if the CV of the grain yields for the control is smaller than or equal to 0.2 over the range of environments in which the test variety is being compared with the control, the difference between $\overline{R}_W$ and $\overline{R}_A$ will be small. On the other hand, if the CV is larger than 0.2, a noticeable difference can be expected between the two estimators. Furthermore, $CV_{\overline{Y}}$, the CV of the denominator of estimator $\overline{R}_W$, is smaller than 0.2 and is much smaller than the CV of the denominator of estimator $\overline{R}_A$ for each of the five comparisons between the test varieties and the control (Table 4.3). It is clear from the preceding sections that the weighted average method should be adopted in these situations. The weighted average ratio estimator is thus shown to be superior to the arithmetic average ratio estimator in this example.

It should be noted that the above analysis only represents one of the most popular practices in plant breeding and is used for illustration of the arithmetic and weighted average ratio estimators. Other important agronomic traits such as resistance to environmental stresses have to be considered before final decision is made. Alternatively, researchers may choose to employ combined analysis of the MET using either conventional analysis of variance

model or REML methodology (Qiao 2001). The approach used by Scobie and Saville (2000) in wool research may also be considered by plotting a graph of variety yield against control yield to correlate $\overline{R}_A$ with other useful information of the trials and to assist in identifying unusual locations. These are, however, beyond the scope of the present study and require further investigations in plant breeding experiments.

4.7 RECOMMENDATIONS AND DISCUSSION

4.7.1 General Recommendation for the Choice of the Ratio Estimator

Of the two ratio estimators studied in this experiment, $\overline{R}_A$ has been favoured in the agricultural research literature (Moreau *et al.* 1999; Schittenhelm 1999; Ismail and Hall 1999; Gumber *et al.* 1999; Qiao *et al.* 2000; Kaeppler *et al.* 2000; Ittu *et al.* 2000; Paderson and Brink 2000; Bromley *et al.* 2000). The main reason for this appears to be that our intuition fails us when it comes to choice of the estimator of a ratio of quantitative measurements.

In the situation where we are estimating a proportion using samples x_i / n_i, $i = 1,...,K$ with x_i being a discrete count of successes in n_i trials, it is more intuitively evident that we should use $\sum x_i / \sum n_i$ rather than $\sum (x_i / n_i) / K$. This discrete problem has already been addressed in Qiao *et al.* (2001). We need to transfer this intuition to the case where there are continuous quantitative variables in the numerator and denominator of the estimator. This, combined with the simulation evidence, indicates strongly that the weighted average should be used in preference to the arithmetic average.

4.7.2 Single versus Multi-environment Analysis

Suppose a plant breeder examines a new wheat variety and the commercial control variety in a regional variety test conducted at n = 24 environments, the aim being to compare the relative yield performance of both varieties for each of these environments and for all 24 environments in general. With this multi-environment test for specific and broad adaptation of the cultivars, the plant breeder's analysis might run as follows.

Single-environment analysis

Following collection of the yield data the plant breeder would first calculate the ratio of grain yields for the new (X_i) and control variety (Y_i) as X_i / Y_i ($i = 1,...,24$) to examine how much the former is better than the latter at each environment. This common practice provides a test for specific adaptation and gives the impression that each ratio X_i / Y_i is a statistic from one sample or environment.

Multi-environment analysis

The next natural step would be to average all 24 ratios giving $\overline{R}_A = \left(\sum_i^{24} \frac{X_i}{Y_i} \right) / 24$, as a test for broad adaptation. The belief is that the arithmetic average of the sample ratios is better than a single ratio for estimation of μ_X / μ_Y, the ratio of the population means for the two varieties.

We consider that the single trial analysis is correct, while the multi-environment analysis is poor. In quantitative genetics and plant breeding, it is more appropriate to consider the whole set of 24 environments as a sample from the target population of environments (Podlich *et al.* 1999; Cooper and Podlich 1999; Qiao 2001). The observations X_i / Y_i, which correspond to environments, have been mistaken as independent "samples" from a single distribution. Based on the distributional properties of the two ratio estimators, the separate X_i / Y_i should not be weighted equally, thus giving $\overline{R}_A$, in the overall analysis. The observation X_i / Y_i should be given a weight of $Y_i / \left(\sum Y_i \right)$ based on the magnitude of Y_i, thus giving $\overline{R}_W$.

4.7.3 Situations where the Arithmetic Average Approach should be Used

Ratios of different attributes

When ratio estimates for a series of attributes at a single environment are to be pooled or averaged to provide an overall assessment of the relative merits of two contrasting treatments, it is inappropriate to average the original measurements to give $\overline{X}_n$ and $\overline{Y}_n$ and

then calculate $\overline{R}_W = \overline{X}_n / \overline{Y}_n$. In an experiment with corn hybrids, for example, Bromley *et al.* (2000) first estimate the ratios between two biological models for seedling vigor, brittle snap, plant height, ear length, yield, days to pollen shed, root lodging and stalk lodging, respectively. The arithmetic average is then used to estimate the pooled ratio. It did not make sense to use the weighted average in this circumstance, while the arithmetic average seems to be more appropriate.

Unavailability of original measurements

The arithmetic average has to be used as a substitute for the weighted average in cases where the separate X_i and Y_i values are not available for various reasons yet the X_i / Y_i values remain. Care must be taken, however, to check for possible effects from extreme single ratio estimates in the series. In general, the arithmetic average can be a good estimate if there is little fluctuation in the components.

Small population variation and large sample size

When sample size is extremely large and the numerator and denominator population variation is very small relative to the population mean, the two estimators will be similar. In such cases, the arithmetic average can be used without major problems. In reality, small variation is sometimes encountered, but sample size will most likely be limited by the availability of resources. Hence the relative merits of the two methods need to be considered even in these situations.

Same or similar denominators of the ratio estimates

If the denominators of the series of ratio estimates are the same, the two estimators will agree, irrespective of the mean ratio, the population coefficients of variation and the sample size. The arithmetic average method can also be considered in situations where the denominators of a series of ratio estimates in a sample are similar, since the results will be comparable to those using the weighted average approach.

4.8 CONCLUSIONS

Differences in the denominators should be considered when averaging several estimates of a ratio of continuous variables; the weighted average method accommodates this variation and is better than the arithmetic average method. Based on our simulation studies, we recommend use of the weighted average approach for estimating the true ratio from a series of ratio estimates in agricultural research. The arithmetic average approach, however, has to be adopted when only the individual ratios are recorded.

According to the distributional properties of the two estimators examined in the previous chapter, the empirically determined critical coefficient of variation value (0.2) for the denominator of the ratio of continuous variables can be used to evaluate the suitability of both estimators. This provides a rational basis for favouring the weighted average ratio estimator in applied research. A practical diagnostic formula has been proposed to assess the reliability of the weighted average ratio estimator, namely that the coefficient of variation for the denominator mean $\overline{Y}_n$ is no larger than 0.2, or $CV_{\overline{Y}_n} = \frac{\sigma_Y}{\mu_Y \sqrt{n}} \leq 0.2$. The arithmetic average ratio estimator is of less use and should be employed only when the coefficient of variation for the denominator Y_i is no larger than 0.2, or $CV_Y = \frac{\sigma_Y}{\mu_Y} \leq 0.2$.

CHAPTER 5
GENERAL CONCLUSIONS AND FURTHER RESEARCH

5.1 GENERAL CONCLUSIONS

Two common estimators, the weighted average and the arithmetic average methods, have been systematically studied for estimating the ratio of a discrete counting variable to a positive discrete counting variable and the ratio of two independent continuous normal variables, both widely used in agricultural research. The research includes theoretical derivations, simulation studies and validation by practical research applications in agriculture. Results showed that the weighted average should be adopted in pooling or averaging a series of ratio data estimates in scientific research or survey studies. Hence, provided the assumptions underlying the analysis hold, the weighted average is unreservedly recommended over the arithmetic average methods. The findings of this thesis can be extended to other applied areas such as economics, sociology or environmental science.

The weighted average estimator always has smaller variance than the arithmetic average estimator for the ratio of a discrete counting variable to a positive discrete counting variable. Hence, the former should be used whenever possible. In situations in which the ratio estimates are available and the sample sizes and/or numerators are not available, the arithmetic average method has to be used. These findings, however, only apply to situations in which there is a constant binomial proportion p for the population of interest. Further studies are needed to examine whether these relationships still hold when the assumption of a constant binomial proportion does not hold.

Although the moments of the ratio of normal variables do not exist in general, the moments may exist if we avoid sample points at an interval for the denominator to approach zero, which is a punctured normal. In practical applications involving the ratio of two continuous normal variables, sample data can be used for assessment of the suitability of the weighted average and the arithmetic average methods in the estimation of the ratio. In both cases, the

coefficient of variation of the denominator variable should be kept smaller than or equal to 0.2 if the ratio is to be appropriately estimated. Based on this diagnostic approach, use of the weighted average ratio estimator is always justified in comparison to the arithmetic average method, since the coefficient of variation of the denominator for the former is always smaller.

5.2 FURTHER RESEARCH

Among the four major categories of ratio outlined in the first chapter, only the first two are studied and conclusions drawn for their practical application in agricultural research or surveys. Within these two categories, however, only situations in which the numerator and denominator are measuring the same attribute have been thoroughly investigated. We may expect that similar conclusions will be reached for situations in which the numerator and denominator are measuring different attributes of either discrete counting variables or continuous variables, although this may need further validation.

The weighted average and the arithmetic average estimators of ratio may also be applicable to the third and the fourth categories, involving the ratio of a continuous variable to a discrete positive counting variable and the ratio of a discrete counting variable to a continuous variable. However, the applicability of these two estimators remains to be empirically compared and theoretically evaluated in this context. The same research methodologies may be adopted for further exploring the theoretical properties and the suitability of the ratio estimators of these two categories in practical situations. When all these have been achieved, a similar conclusion regarding the comparison between these two ratio estimators may be reached and their practical implications may be examined and validated using both simulation method and real data.

Factors influencing the difference between the weighted average and the arithmetic average methods for binomial proportion data have been identified as difference in sample sizes, difference in individual proportion estimates of the samples, number of samples, total sample size and sum of the reciprocals of individual sample sizes. It is not clear which of these factors plays the most important role in deciding the difference between the two

methods in estimating the binomial proportion of the population. Situations need to be identified when the weighted average will be superior to the arithmetic average and when the latter has advantages over the former.

When there are several binomial samples, each having a different binomial proportion, we then need to consider a mixture of binomial distributions of the form

$$P(X=x)=\binom{n}{x}\sum_{j=1}^{K} w_j p_j^x (1-p_j)^{n-x},$$

where $n=n_1+...+n_K$, $w_j>0$ and $\sum_{j=1}^{K} w_j=1$.

The problem of estimating the mixing distribution (that is, estimating the $w_j=1,...,K$) and p_j is a difficult one. This issue is important in that when one has access to various binomial samples, he/she tends to assume that the individual proportion estimates come from a common binomial population and hence it may be simplistic just to pool these estimates for a combined estimate of the true population proportion. The question is, however, whether these samples come from the same binomial populations. If the answer is yes, then the weighted averaging method can be adopted for an estimate of the pooled binomial proportion of the unknown population. If the answer is no, on the other hand, the method of estimation recommended by Wood (1999) can then be used to estimate the parameters of this mixing distribution. The Restricted Maximum Likelihood (REML) methodology may also be considered for dealing with non-constant binomial proportions. This will be another interesting area that deserves further investigation and validation using real data from applied research.

The study of ratio estimators of continuous variables in this thesis was based on the assumption that X_i and Y_i are independent normal random variables. Under certain conditions, however, the numerator and denominator normal variables of the ratio may be correlated. This situation remains to be investigated. Furthermore, the present research has not considered situations where the normal assumption is not met or in doubt. Non-normal data occurs frequently, especially for data with smaller sample size. Thanks to the central

limit theorem, the normal assumption is both unnecessary and too restrictive. It is of practical importance to find out whether the relationship between the weighted average and arithmetic average ratio estimators still holds when one or both of the numerator and denominator variables of the ratio do not follow a normal distribution. It is desirable to find a robust estimator to estimate the ratio of the means of two continuous variables irrespective of the distribution. With experience gained from the current study, these goals should be within reach.

The other area of research interest is to compare other types of ratio estimators in the literature with the two estimators discussed in this thesis, that is, the $\overline{R}_W$ and $\overline{R}_A$ commonly used in agricultural research. Although preliminary work has been done in the literature in theoretical terms, the interrelationships between them have not been experimentally evaluated. It is therefore of practical importance to find out how these ratio estimators are ranked in their suitability for estimation of ratio from the agricultural point of view.

The ratio of continuous variables was investigated using simulated and experimental data of paired independent normal distributions in the present study. It is of practical importance to understand whether the same conclusions hold for situations where the numerator and denominator variables are correlated, which do occur in many agricultural applications. Both simulations and experiments are required to answer this question.

REFERENCES

Abramowitz, M. and Stegun I. A. (1964) *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York, USA.

Beale, E. M. L. (1962) Some uses of computers in operational research. *Industrielle Organisation* **31**: 51-52.

Bromley, C. M., Van Vleck, L. D., Johnson, B. E. and Smith, O. S. (2000) Estimation of genetic variance in corn from F_1 performance with and without pedigree relationships among inbred lines. *Crop Science* **40**: 651-655.

Casler, M. D. and Santen, E. V. (2000) Patterns of variation in a collection of meadow fescue accessions. *Crop Science* **40**: 248-255.

Chen, S., Lin, X. H., Xu, C. G. and Zhang, Q. (2000) Improvement of bacterial blight resistance of 'Minghui 63', an elite restorer line of hybrid rice, by molecular marker-assisted selection. *Crop Science* **40**: 239-244.

Choi, H. W., Lemaux, P. G. and Cho, M. J. (2000) Increased chromosomal variation in transgenic versus nontransgenic barley (*Hordeum vulgare* L.) plants. *Crop Science* **40**: 524-533.

Cochran, W. G. (1977) *Sampling techniques* (3rd Ed). John Wiley and Sons, Inc., New York, USA.

Cooper, M. and Podlich, D. W. (1999) Genotype × environment interactions, selection response and heterosis. In *The genetics and exploitation of heterosis in crops.* (Eds. J. G. Coors and S. Pandey) ASA-CSSA-SSSA, Madison, Wisconsin, USA (Madison; N).

Durbin, J. (1959) A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* **46**: 477-480.

Fieller, E. C. (1932) The distribution of the index in a normal bivariate population. *Biometrika* **24** (Part III and IV): 428-440.

Frankel, M. R. (1971) *An empirical investigation of some properties of multivariate statistical estimates from complex samples*. Ph.D. Thesis, University of Michigan, Michigan, USA.

Frishman, F. (1975) On the arithmetic means and variances of ratios and products of two random variables. *A modern course on statistical distributions in scientific work* **1**: 401-

406.

Geary, R.C. (1930) The frequency distribution of the quotient of two normal variates. Journal of Royal Statistical Society **93**: 442-446.

Guenther, W. C. (1973) *Concepts of statistical inference* (2dn Ed). McGraw-Hill Kogakusha, Ltd. Tokyo, Japan. pp 127-131.

Gumber, R. K., Schill, B., Link, W., Kittlitz, E. V. and Melchinger, A. E. (1999) Mean, genetic variance, and usefulness of selfing progenies from intra- and inter-pool crosses in faba beans (*Vicia Faba* L.) and their prediction from parental parameters. *Theoretical and Applied Genetics* **98**: 569-580.

Hall, R. L. (1979) Inverse moments for a class of truncated normal distributions. *The Indian Journal of Statistics* **41** (*Series B*): 66-76.

Haque, A. K. M. A., Choudhury, N. H., Quasem, M. A. and Arboleda, J. R. (1997) Rice post-harvest practices and loss estimates in Bangladesh – Part III: parboiling to milling. *Agricultural Mechanisation in Asia, Africa and Latin America* **28**: 51-55.

Hartley, H. O. and Ross, A. (1954) Unbiased ratio estimations. *Nature* **174**: 270-271.

Hinkley, D. V. (1969) On the ratio of correlated normal random variables. *Biometrika* **56**: 635-639.

Ismail, A. M. and Hall, A. E. (1999) Reproductive-stage heat tolerance, leaf membrane thermostability and plant morphology in cowpea. *Crop Science* **39**: 1762-1768.

Ismail, A. M., Hall, A. E. and Ehlers, J. D. (2000) Delayed-leaf-senescence and heat-tolerance trait mainly are independently expressed in cowpea. *Crop Science* **40**: 1049-1055.

Ittu, M., Saulescu, N. N., Hagima, I., Ittu, G. and Mustatea, P. (2000) Association of Fusarium head blight resistance with Gliadin loci in a winter wheat cross. *Crop Science* **40**: 62-67.

Jilin Provincial Crop Variety Evaluation Committee (1995) *Crop Variety Evaluation Regulations of Jilin Province*. Scientific and technological Press of Jilin Provincce, Changchun, China. pp 65-74.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate distributions*, Vol 1 (2nd Edition), John Wiley and Sons, Inc., New York, USA.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate discrete distribution* (2dn

Ed). Wiley Interscience, New York, USA. pp. 125-127.

Kaeppler, S. M., Parke, J. L., Mueller, S. M., Senior, L., Stuber, C. and Tracy, W. F. (2000) Variation among maize inbred lines and detection of quantitative trait loci for growth at low phosphorus and responsiveness to arbuscular mycorrhisal fungi. *Crop Science* **40**: 358-364.

Kamerud, D. B. (1978) Solution to Problem 6104: The random variable X/Y, X, Y normal. *American Mathematical Monthly* **85**: 206-207.

Kempthorne, O. (1957) An introduction to genetic statistics. John Wiley & Sons, Inc., New York, USA.

Kim, H. S., Hartman, G. L., Manandhar, J. B., Graef, G. L., Steadman, J. R. and Diers, B. W. (2000) Reaction of soybean cultivars to sclerotinia stem rot in field, greenhouse, and laboratory evaluations. *Crop Science* **40**: 655-669.

Kokan, A. R. (1963) A note on the stability of the estimates of standard errors of the ordinary mean estimate and the ratio estimate. *Calcutta Statistical Association Bulletin* **12**: 149-158.

Lahiri, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Bulletin of International Statistical Institution* **33**: 133-140.

Levine, D. M., Krehbiel, T. C. and Berenson, M. L. (2000) *Business statistics: a first course* (2nd Ed). Prentice Hall, Upper Saddle River, New Jersey, USA.

Lipschutz, S. (1968) *Schaum's outline of theory and problems of linear algebra*. Schaum's Outline Series, McGraw Book Company, New York, USA. pp. 279-306.

Lukacs, E. (1975) *Stochastic convergence*. 2nd Edition. Academic Press, New York, USA

Lukacs, E. and Laha, R. G. (1964) *Applications of Characteristic Functions*. Hafner Pub. Co., Griffin, London, UK.

McCarthy, P. (1969) Pseudoreplication: Further evaluation and application of the balanced half-sample technique, National Center for Health Statistics, Washington, D.C., Series 2, No. 31.

Mickey, M. R. (1959) Some finite population unbiased ratio and regression estimators. *Journal of American Statistical Association* **54**: 594-612.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974) *Introduction to the theory of statistics* (3rd Ed). McGraw-Hill Book Company, Tokyo, Japan. pp. 321-324.

Moreau, L., Monod, H., Charcosset, A. and Gallais, A. (1999) Marker-assisted selection with spatial analysis of unreplicated field trials. *Theoretical and Applied Genetics* **98**: 234-242.

Moses, L. E. (1962) Use of Wilcoxon test theory in estimating the distribution of a ratio by Monte Carlo Methods. *Annals of Mathematical Statistics* **33**: 1194-1196.

Nahmias, S. and Wang, S. S. (1978) Approximating partial inverse moments for certain normal variates with an application to decaying inventories. *Naval Research Logistics Quarterly* **25**: 405-413.

Narayanan, R. A., Atz, R., Nenny, R., Young, N. D. and Somers, D. A. (1999) Expression of soybean cyst nematode resistance in transgenic hairy roots of soybean. *Crop Science* **39**: 1680-1686.

Ott, R. L. (1993) *An introduction to statistical methods and data analysis* (4th Ed). Wadsworth, Inc., Belmont, California, USA. pp. 380-384.

Ott, R. L. and Mendenhall, W. (1994) *Understanding statistics* (6th Ed). Wadsworth, Inc., Belmont, California, USA. pp. 406-410.

Paderson, G. A. and Brink, G. E. (2000) Seed production of white clover cultivars and naturalised populations when grown in a pasture. *Crop Science* **40**: 1109-1114.

Pascual, J. N. (1961) Unbiased ratio estimators in stratified sampling. *Journal of American Statistical Association* **56**: 70-87.

Piegorsch, W. W. and Casella, G. (1985) The existence of the first negative moment. *The American Statistician* **39**: 60-62.

Pitt, H. (1994) *SPC for the rest of us - a personal path to statistical process control.* Addison-Wesley Publishing Company, Inc, New York, USA. pp. 259-276.

Podlich, D. W., Cooper, M. and Basford, K. E. (1999) Computer simulation of a selection strategy to accommodate genotype-environment interactions in a wheat recurrent selection programme. *Plant Breeding* **118**: 17-28.

Qiao, C. G. (2001) *Evaluation of correlated genetic advance theory as a framework for analysing genotype by environment interactions in wheat breeding multi-environment trials.* PhD Thesis, the University of Queensland, Brisbane, Australia. July 2001.

Qiao, C. G., Basford, K. E., DeLacy, I. H. and Cooper, M. (2000) Evaluation of experimental designs and spatial analyses in wheat breeding trials. *Theoretical and*

Applied Genetics **100**: 9-16.

Qiao, C. G., Shan, L. M., Yang, F., Zhu, Y. and Kong, F. J. (1994). *A report on field examinations of sunflower cultivar Improved Peredovic in disease and lodging resistance, and other agronomic performance in the Western Region of Jilin Province.* Research Report of Jilin Agricultural University, Changchun, China, October 1994.

Quenouille, M. H. (1956) Notes on bias in estimation. *Biometrika* **43**: 353-360.

Rao, C. R. (1952) *Advanced statistical methods in biometrics research.* John Wiley and Sons, Inc., New York, USA.

Rao, J. N. K. and Beegle, L. D. (1967) A Monte Carlo study of some ratio estimators. *The Indian Journal of Statistics* **29**(*Series B*): 47-56.

Rao, J. N. K. and Kuzik, R. A. (1974) Sampling errors in ratio estimation. *The Indian Journal of Statistics* **36**(*Series C, Pt* 1): 43-58.

Rao, J. N. K. and Webster, J. T. (1966) On two methods of bias reduction in the estimation of ratios. *Biometrika* **53**: 571-577.

Roberts, C. (1969) Fill weight variation release and control of capsules, tablets, and sterile solids. *Technometrics* **11:** 161-175.

Robinson, D. L., Kershaw, C. D. and Ellis, R. P. (1988) An investigation of two dimensional yield variability in breeders' small plot barley trials. *Journal of Agricultural Science, Cambridge* **111**: 419-426.

Royall, R. M. and Eberhardt, K. R. (1975) Variance estimators for the ratio estimators. *The Indian Journal of Statistics* **37**(*Series C, Pt* 1): 43-52.

Sastry, K. V. R. (1965) Unbiased ratio estimators. *Journal of Indian Society of Agricultural Statistics* **17**: 19-29.

Schittenhelm, S. (1999) Agronomic performance of root chicory, Jerusalem artichoke, and sugarbeet in stress and nonstress environments. *Crop Science* **39**: 1815-1823.

Scobie, D. R. and Saville, D. J. (2000) The misuse of ratios and percentages in wool research. *Asian-Australian Journal of Animal Science* **13 Supplement**: 461-464.

Springer, M. D. (1979) The algebra of random variables. John Wiley and Sons, Inc., New York, USA.

Tin, M. (1965) Comparison of some ratio estimators. *Journal of the American Statistical Association* **60**: 294-307.

Tukey, J. W. (1958) Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics* (Abstract) **29**: 614.

Witcombe, J. R., Petre, R., Jones, S. and Joshi, A. (1999) Farmer participatory crop improvement. IV. The spread and impact of a rice variety identified by participatory varietal selection. *Experimental Agriculture, Cambridge* **35**: 471-487.

Wood, G. R. (1999) Binomial mixtures: geometric estimation of the mixing distribution. *Annals of Statistics* **27**: 1706-1721.