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 $\overline{p}_A = \left(\sum \frac{x_i}{n_i}\right) / K$. 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 $\overline{p}_W = \sum w_i \frac{x_i}{n_i} = \left(\frac{n_1}{\sum n_i}\frac{x_1}{n_1} + \frac{n_2}{\sum n_i}\frac{x_2}{n_2} + \ldots + \frac{n_K}{\sum n_i}\frac{x_K}{n_K}\right) = \frac{\sum x_i}{\sum n_i}$. 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

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

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.

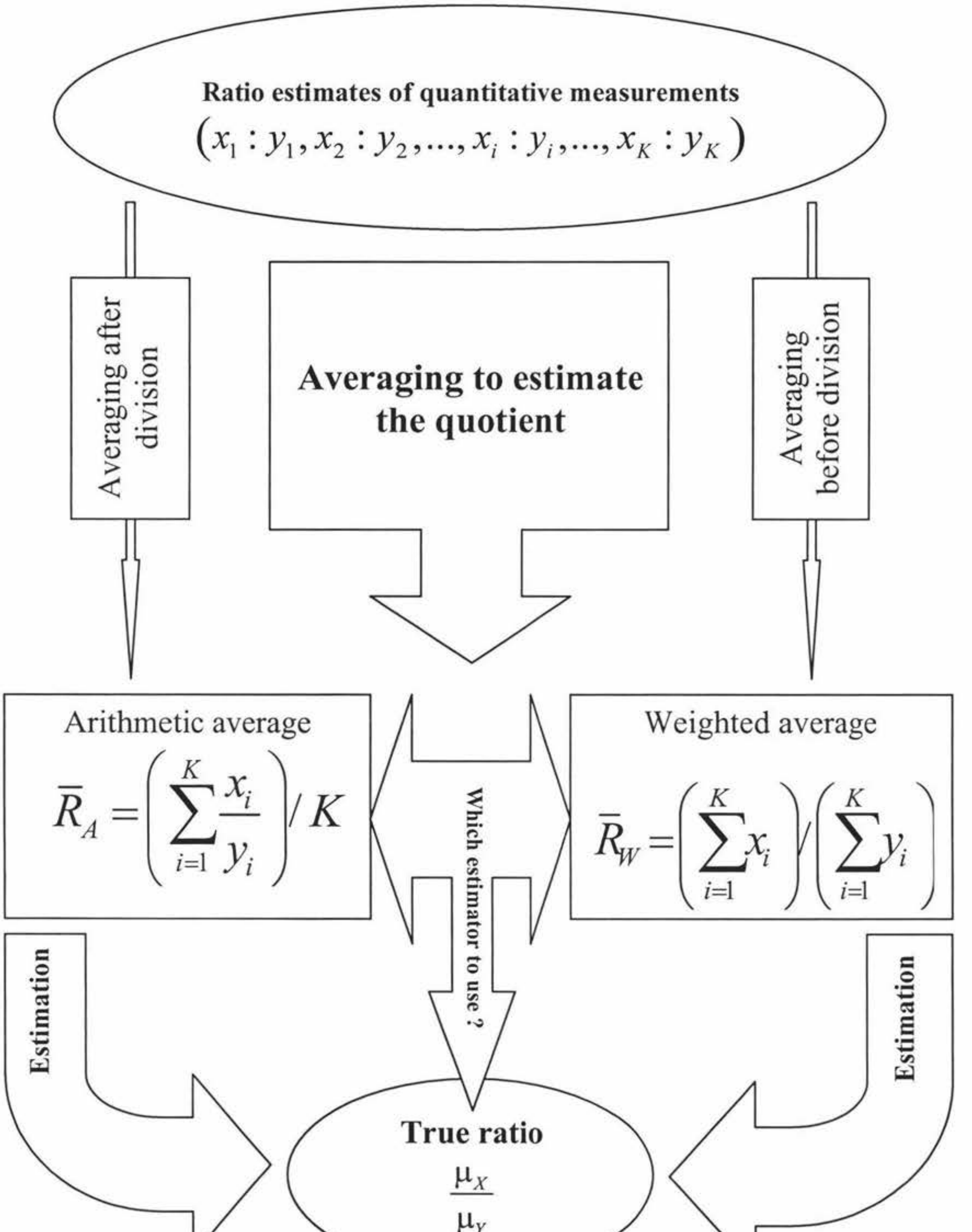

Figure 1.1 A diagram outlining the skeleton of the research comparing two common estimators of ratio, where X and Y can be either discrete counting variable or continuous variable. In the case of a ratio of discrete counting variables, $\bar{R}_A$ and $\bar{R}_W$ are replaced by $\bar{P}_A$ and $\bar{P}_W$,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