Spring 2001

Statistical modeling and inference regarding risk in case control studies

Deborah Kay Shepherd

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

STATISTICAL MODELING AND INFERENCE REGARDING RISK IN CASE CONTROL STUDIES

by

Deborah Kay Shepherd, M.S.

A Dissertation Presented in Parial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

COLLEGE OF ENGINEERING AND SCIENCE
LOUISIANA TECH UNIVERSITY

May 2001

UMI Number: 3000447

UMI Microform 3000447

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

LOUISIANA TECH UNIVERSITY

THE GRADUATE SCHOOL

March 22, 2001
Date

We hereby recommend that the dissertation prepared under our supervision by Deborah Kay Shepherd

entitled Statistical Modeling and Inference Regarding Risk in Case Control Studies

be accepted in partial fulfillment of the requirements for the Degree of Ph.D. in Computational Analysis and Modeling

Raja Nassar
Supervisor of Dissertation Research

Raja Nassar
Head of Department

Department

Recommendation concurred in:

Raja Nassar

Advisory Committee

Approved:
Director of Graduate Studies

Approved:
Director of the Graduate School

Dean of the College

GS Form 13
2/97

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It is understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author Deborah K Shepherd

Date April 11, 2001

GS Form 14
2/97

ABSTRACT

Some of the common measures of risk used in epidemiology today are the relative risk, the odds ratio, the attributable risk, and the chi-square goodness of fit test. All of these measures have their shortcomings. A new approach to measuring risk in case-control studies is to use the unitless measure of the coefficient of variation of incidence of disease over the risk categories, $\hat{k}^2$, first proposed by Begg et al. (1998). Begg et al. (1998), also showed that the product of multiple risk factors may be compared to an overall measure of the square of the coefficient of variation of the incidence of disease over all risk categories known and unknown, S, the standardized incidence ratio. It is shown that $S = k_t^2 + 1$, where k_t^2 represents the square of the coefficient of variation of the incidence of disease over all risks. If the risks are independent, then an estimate of S may be calculated from a case-control study as $\hat{S} = \prod_{i=1}^{r}\left(\hat{k}_i^2 + 1\right)$ and $\ln\hat{S} = \ln\left(\prod_{i=1}^{r}\left(\hat{k}_i^2 + 1\right)\right) = \sum_{i=1}^{r}\ln\left(\hat{k}_i^2 + 1\right)$. The parameter S may be available from a source such as a cancer registery. If $\hat{S} = \prod_{i=1}^{r}\ln\left(\hat{k}_i^2 + 1\right)$ is much smaller than S then it may be that not all risks have been considered.

The distribution and statistical properties of $\hat{k}^2$, $\ln\left(\hat{k}_i^2 + 1\right)$, and $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2 + 1\right)$ have not been investigated. In this study, it is shown that the distribution of $\hat{k}^2$ for one risk factor with multiple levels, is $Gamma\left(\frac{c-1}{2}, 2\frac{n-m}{n*m}\right)$. A simulation study was conducted to investigate the power of this statistic for testing $H_o : \hat{k}^2 = 0$ vs $H_a : \hat{k}^2 \neq 0$ for one risk factor with multiple levels. The simulation study confirmed the power of the test statistic to be very good as long as the sample size was at least 200 for both cases and controls.

The measure $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$ is of interest because it may be used to compare the sum of the logarithms of risk factors used in a study to the natural log of the overall square of the coefficient of variation of the incidence of disease over all risks known and unknown, $\ln S = \sum_{i=1}^{t} \ln(k_i^2 + 1)$. Also, this study investigates the distribution of the statistic $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$ and the power of this statistic when used to test $H_o : \sum_{i=1}^{r} \ln(k_i^2 + 1) = 0$ vs. $H_a : \sum_{i=1}^{r} \ln(k_i^2 + 1) > 0$ and $H_o : \sum_{i=1}^{r} \ln(k_i^2 + 1) = \ln S$ vs. $H_a : \sum_{i=1}^{r} \ln(k_i^2 + 1) \neq \ln S$.

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDEMENTS

I wish to express my sincere appreciation to my advisor Dr. Raja Nassar for his expert guidance throughout the course of this dissertation. I also wish to thank Dr. Gu and Dr. Alexander for their helpful suggestions while serving on the committee.

Also, I would like to thank Dr. Richard Greechie for helping me to obtain funding so that I might pursue this degree.

I am blessed with a wonderful family that has provided me with love and encouragement. Special thanks to my husband for his patience during this effort.

CHAPTER 1

INTRODUCTION

An important issue in our society today is health care. Epidemiologists strive to find what factors are associated with certain diseases. Last, (1988) has defined epidemiology as "the study of the distribution and determinants of health related states or events in specified populations, and the application of this study to control health problems." Because these measures of risk are used to set health policy and control disease, it is important to understand their properties and limitations.

Some of the most prevalent measures of risk used by epidemiologists today include the relative risk, odds ratio, chi-square goodness of fit test, and the attributable risk. There are advantages and disadvantages associated with each of these measures. For example, a disadvantage associated with the relative risk and the odds ratio is that the prevalence rate of the risk factor is not accounted for in the target population (Whittmore 1983). Levin (1953) proposed the attributable risk which was the first measure of risk that took the prevalence rate of the risk factor in the target population into consideration. Although the attributable risk considers more information, it is not without shortcomings. The attributable risk is dependent on the definition of the base one category of the risk factor; therefore, different researchers can compute different values of attributable risk for the same data (Begg et al. 1998).

Another measure of association between risk and disease was proposed by Begg, et al. (1998). He proposed a statistic in which the measure of risk was not dependent on the base line category, but computed by calculating the square of the coefficient of variation of the incidence rate over the risk categories. Begg showed that this statistic may be compared to the

standardized incidence ratio of second primaries of the disease which is shown to be the square of the coefficient of variation of the incidence rate over the risk categories for the entire population and all risk factors, known and unknown. Begg developed a nonparametric estimator for the square of the coefficient of variation of the incidence rate over the risk categories, $\hat{k}_b^2$, for a retrospective model. There has not been any investigation into hypothesis testing using the $\hat{k}_b^2$ statistic. A related measure to Begg's statistic is the square of the sample coefficient of variation of the incidence rate over the risk categories, $\hat{k}^2$.

A point of interest would be to test whether the square of the coefficient of variation of the incidence rate over the risk categories is zero. Another point of interest may be whether the square of the coefficient of variation of the incidence rate over the categories for two independent risk factors is the same. If there is more than one risk factor, then a test of the sum of the log of the squares of the coefficient of variation of the incidence rate over the risk categories is of interest. This sum could be tested against the log of the standardized incidence ratio of second primaries of the disease.

This study investigates the asymptotic distribution and properties of the square of the sample coefficient of variation of incidence rate over the risk categories, $\hat{k}^2$, for one risk factor with variable levels. Also, the distribution of $\ln(\hat{k}^2+1)$, $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$, and the difference, $\hat{k}_1^2-\hat{k}_2^2$, is investigated. A simulation to calculate the size and power of the square of the sample coefficient of variation of the incidence rate over the risk categories, $\hat{k}^2$, for testing a factor as a significant risk is investigated. The size and power associated with this statistic is also compared to the well known chi-square test. The size and power of the test statistics $\ln(\hat{k}^2+1)$ and $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$ for testing $Ho : \ln(k^2+1) = 0$ vs. $Ha : \ln(k^2+1) \neq 0$ and $Ho : \sum_{i=1}^{r}\ln(k_i^2+1) = 0$ vs. $Ha : \sum_{i=1}^{r}\ln(k_i^2+1) \neq 0$, respectively, are investigated. The

statistic $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$ is of importance because it may be compared with the standardized incidence ratio of second primaries, S. Therefore, the size and power of this statistic for testing $Ho : \sum_{i=1}^{r} \ln(k_i^2 + 1) = \ln S$ vs. $Ha : \sum_{i=1}^{r} \ln(k_i^2 + 1) \neq \ln S$, is also investigated. The size and power of the difference, $\hat{k}_1^2 - \hat{k}_2^2$, is investigated for testing $Ho : \hat{k}_1^2 - \hat{k}_2^2 = 0$ vs. $Ha : \hat{k}_1^2 - \hat{k}_2^2 \neq 0$.

1.1 Related Research Concerning Study Designs

One of the earliest methods of studying the effect of a risk factor on a disease outcome is using a 2×2 table where both the risk factor and the disease outcome are dichotomous. The data for a 2×2 table can be displayed in different ways, depending on the sampling scheme used. There are three study designs used most often in the literature with respect to the 2×2 table. They are the case-control (retrospective), prospective, and cross-sectional study designs. The study designs differ in the way the population is sampled. For clarity, the following notations will be used to represent groups of individuals in the tables:

N = number of people in the study population
n_{10} = number of cases with no risk factor present
n_{11} = number of cases with the risk factor present
n_{00} = number of controls with no risk factor present
n_{01} = number of controls with the risk factor present
m = total number of cases in the study population
n = total number of controls in the study population
q_0 = conditional probability of no risk given the disease is present
q_1 = conditional probability of risk given the disease is present
p_0 = conditional probability of no risk given no disease
p_1 = conditional probability of risk given no disease
e = total number of people in the study population with the risk factor

ne = total number of people in the study population without the risk factor

d_0 = conditional probability of disease given no risk factor present

d_1 = conditional probability of disease given the risk factor is present

nd_0 = conditional probability of no disease given no risk factor present

nd_1 = conditional probability of no disease given the risk factor is present

DR = probability of disease with risk factor present

DNR = probability of disease with no risk factor present

NDR = probability of no disease with risk factor present

$NDNR$ = probability of no disease with no risk factor present.

Here, controls refer to the individuals in the study without the disease and cases refer to the individuals in the study with the disease.

1.1.1 Prospective Study

A prospective study design resembles an experiment, therefore making it useful if a causal inference is desired (Kleinbaum, Kupper, and Morgenstern 1982). This type of study requires a cohort of individuals to be followed, *before* the onset of disease, for a set time period during which the onset of the disease is recorded. One of the major advantages of this type of study is the ability to calculate the incidence rate of the disease. Using the notation from table 1.1, this may be calculated by $\frac{n_{11}+n_{10}}{N}$. The study population may be sampled as stratified on the risk factor or unstratified. The sample, if unstratified, is grouped into exposed and nonexposed to the risk factor (Walter 1976). This type of study may be very costly because of the time required for the onset of disease and the large amount of individuals that need to be included in the study if the disease is rare. The following 2×2 table reflects the sample frequency distribution and cell probabilities of such a design using the notations listed above.

Table 1.1 2×2 Table for a Prospective Study Design

	Disease		Total	Disease		Total
Risk Factor	Yes	No		Yes	No	
Yes	n_{11}	n_{01}	e	$d_1 = \frac{n_{11}}{e}$	$nd_1 = \frac{n_{01}}{e}$	1
No	n_{10}	n_{00}	ne	$d_0 = \frac{n_{10}}{ne}$	$nd_0 = \frac{n_{00}}{ne}$	1
Total	$n_{11}+n_{10}$	$n_{01}+n_{00}$	N			

Here, the number of individuals with and without the risk factor are fixed. The variables n_{11} and n_{10} are considered to be independent binomial variables with parameters (e, d_1) and (ne, d_0), respectively. In this model there does not exist a method to calculate the prevalence of risk in the target population. The relative risk, odds ratio, and attributable risk may be calculated from this type of study design.

1.1.2 Cross-Sectional Study

In a cross-sectional study, an unstratified sample of size N is collected from the target population. This sample is then grouped into four categories, which are risk and disease, risk and no disease, no risk and disease, and no risk and no disease. The 2×2 table that would represent this type of study is given below.

Table 1.2 2×2 Table for a Cross-Sectional Study Design

	Disease		Total	Disease		Total
Risk Factor	Yes	No		Yes	No	
Yes	n_{11}	n_{01}	$n_{11}+n_{01}$	$DR = \frac{n_{11}}{N}$	$NDR = \frac{n_{01}}{N}$	
No	n_{10}	n_{00}	$n_{10}+n_{00}$	$DNR = \frac{n_{10}}{N}$	$NDNR = \frac{n_{00}}{N}$	
Total	$n_{11}+n_{10}$	$n_{01}+n_{00}$	N			1

In this type of study the variables n_{11}, n_{01}, n_{10}, and n_{00} are considered to be multinomial with parameters $(N; DR, NDR, DNR, NDNR)$.

This type of study design is easy and economical to conduct. The prevalence of disease may be measured from this type of study. Using the notation from table 1.2, the prevalence of disease may be calculated by $\frac{n_{11}+n_{10}}{N}$. A disadvantage of this type of study is that it is not possible to tell which occurred first, the risk or the disease. Therefore, the reason for the association between risk and disease is not easy to assess. However, this type of study may be useful for investigating factors that are fixed characteristics of individuals, such as race (Beaglehole, Bonita, and Kjellstrom 1993).

1.1.3 Case-Control Study

In a case-control study (also called a retrospective study) two samples are drawn, one from a population of cases and another from a population of controls. In this type of study, the fixed variables will be the total number of controls and the total number of cases, that is n and m in the notation given above. Because this type of study is like a snapshot in time, there is not a follow-up period, making the case-control study less expensive and less time consuming than the prospective study. A disadvantage of the case-control study is that risk factor data are collected from the individual after his or her disease status is known. Consequently, the accuracy of this data is heavily reliant on the individual's memory or perception. The following table represents the frequency distribution and cell probabilites of a case-control study.

Table 1.3 2 × 2 Table for a Case-Control Study

	Disease		Total	Disease	
Risk Factor	Yes	No		Yes	No
Yes	n_{11}	n_{01}	$n_{11}+n_{01}$	$q_1 = \frac{n_{11}}{n_{11}-n_{10}}$	$p_1 = \frac{n_{01}}{n_{01}-n_{00}}$
No	n_{10}	n_{00}	$n_{10}+n_{00}$	$q_0 = \frac{n_{10}}{n_{11}-n_{10}}$	$p_0 = \frac{n_{00}}{n_{01}-n_{00}}$
Total	$n_{11}+n_{10}$	$n_{01}+n_{00}$	N	1	1

Here, n_{11} and n_{10} are considered to be multinomial random variables with parameters

$(n_{11}+n_{10};q_1,q_0)$ and n_{01} and n_{00} are considered to be multinomial random variables with parameters $(n_{01}+n_{00};p_1,p_0)$. The cases and controls are drawn from two populations, therefore, making the two multinomial distributions independent of one another. The square of the coefficient of variation of the incidence rate over the risk categories presented in this study assumes this case-control design.

1.2 Related Research Concerning Measures of Risk

Four common measures of risk used in the literature today include the odds ratio, the relative risk, the attributable risk, and the chi-square goodness of fit test. All of these measures are not without their shortcomings. The relative risk and the odds ratio do not take into account the prevalence rate of the risk factor. Therefore, if the risk factor is very influential on the disease, but very rare in the target population, then it may not pose a major health problem. On the other hand, the attributable risk includes the prevalence rate of the risk factor but is highly dependent on the base-line category of risk, i.e., the relative risk is one or is minimum (Begg et al. 1998). Different researchers may perform the same experiment with the exception of the definition of a base-line category of risk and come up with two different measures for the attributable risk. For this reason, Begg et al. (1998) proposed an alternative method to calculate a measure of risk. He proposed using the square of the coefficient of variation of the incidence rate over the risk categories to measure the degree of a risk factor. Section 1.2.4 will address this approach.

1.2.1 Relative Risk

The relative risk is an incidence ratio that compares two risks. Using the notation in Table 1.3 the relative risk of developing a disease, given the risk factor of interest is present, would be

$$\textit{relative risk} = RR$$
$$= \frac{\text{Incidence of disease in a group with the risk factor}}{\text{Incidence of disease in a group without the risk factor}}$$
$$= \frac{\frac{n_{11}}{n_{11}+n_{01}}}{\frac{n_{10}}{n_{10}+n_{00}}}. \qquad (1.1)$$

The relative risk can only be calculated from a prospective study. Miettenan (1972) developed a procedure to deal with the confounding of factors if the relative risk is calculated using more than one risk factor. Cornfield (1951) showed that the relative risk can be estimated by the odds ratio in a retrospective study.

1.2.2 Odds Ratio

The odds of an event can be described by $\frac{P(\text{event A occurs})}{P(\text{event A does not occur})}$. In terms of our discussion of risk factor vs. disease, the odds an individual has a risk factor given the disease is

$$\frac{P(\text{individual has the factor} \mid \text{disease})}{P(\text{individual does not have the factor} \mid \text{disease})}.$$

The odds an individual has the risk factor given no disease may be written as

$$\frac{P(\text{individual has the factor} \mid \text{no disease})}{P(\text{individual does not have the factor} \mid \text{no disease})}.$$

The ratio of the odds of an individual having a risk factor given the disease to the odds of an individual having a risk factor given no disease is called the odds ratio and written as

$$\textit{odds ratio} = OR = \frac{\frac{P(\text{individual has the factor} \mid \text{disease})}{P(\text{individual does not have the factor} \mid \text{disease})}}{\frac{P(\text{individual has the factor} \mid \text{no disease})}{P(\text{individual does not have the factor} \mid \text{no disease})}}$$

$$= \frac{\frac{n_{11}}{n_{11}+n_{10}}}{\frac{n_{10}}{n_{11}+n_{10}}} \Big/ \frac{\frac{n_{01}}{n_{01}+n_{00}}}{\frac{n_{00}}{n_{01}+n_{00}}}$$

$$= \frac{\frac{n_{11}}{n_{10}}}{\frac{n_{01}}{n_{00}}}$$

$$= \frac{n_{11}n_{00}}{n_{10}n_{01}} \tag{1.2}$$

If the disease is rare, then the odds ratio can be used to estimate the relative risk in a retrospective study. For simplicity let D = *disease*, F = *factor*, ND = *no disease*, and NF = *no factor*. From the above discussion, it can be shown, that

$$RR = \frac{P(D|F)}{P(D|NF)} \tag{1.3}$$

$$OR = \frac{\frac{P(F|D)}{P(NF|D)}}{\frac{P(F|ND)}{P(NF|ND)}} = \frac{P(F|D)P(NF|ND)}{P(F|ND)P(NF|D)} \tag{1.4}$$

Using Bayes' rule, we may rewrite $P(F|D)$ and $P(F|ND)$as

$$P(F|D) = \frac{P(D|F)P(F)}{P(D|F)P(F) + P(D|NF)P(NF)}$$

$$P(F|ND) = \frac{P(ND|F)P(F)}{P(ND|F)P(F) + P(ND|NF)P(NF)}$$

and

$$\frac{P(F|D)}{P(F|ND)} = \frac{\frac{P(D|F)}{P(D|F)P(F) + P(D|NF)P(NF)}}{\frac{P(ND|F)}{P(ND|F)P(F) + P(ND|NF)P(NF)}}$$

In like manner,

$$\frac{P(NF|ND)}{P(NF|D)} = \frac{\dfrac{P(ND|NF)}{P(ND|NF)P(NF)+P(ND|F)P(F)}}{\dfrac{P(D|NF)}{P(D|NF)P(NF)+P(D|F)P(F)}}$$

Replacing the terms in the odds ratio with the expressions above, we may rewrite the odds ratio as,

$$OR = \frac{P(F|D)P(NF|ND)}{P(F|ND)P(NF|D)}$$

$$= \frac{\dfrac{P(D|F)}{P(D|F)P(F)+P(D|NF)P(NF)}}{\dfrac{P(ND|F)}{P(ND|F)P(F)+P(ND|NF)P(NF)}} \times \frac{\dfrac{P(ND|NF)}{P(ND|NF)P(NF)+P(ND|F)P(F)}}{\dfrac{P(D|NF)}{P(D|NF)P(NF)+P(D|F)P(F)}}$$

$$= \frac{P(D|F)P(ND|NF)}{P(ND|F)P(D|NF)} \tag{1.5}$$

If the disease is rare then

$$P(ND|NF) \cong 1$$

and

$$P(ND|F) \cong 1$$

consequently,

$$OR \cong \frac{P(D|F)}{P(D|NF)} = RR \tag{1.6}$$

1.2.3 Attributable Risk

Attributable risk was first proposed by Levin in 1953. He defined the attributable risk as a "measure of the proportion of the disease in the population which can be attributed to the factor" (Levin 1953). The attributable risk has also been called the etiologic fraction (Miettenan 1974) and is described as the proportion of cases that are attributed to the risk factor. The etiologic fraction is calculated using incidence rates and only deals with positive

risk factors, i.e. $I_0 \leq I$. Let I = incidence rate of disease in the population, and I_0 = incidence rate of the disease in the population *without* the risk factor, then

$$AR = \frac{I - I_0}{I}. \tag{1.7}$$

Because incidence rates are not always available, the above can be rewritten as

$$\begin{aligned}
AR &= \frac{I - I_0}{I} \\
&= \frac{P(D) - P(D|NF)}{P(D)} \\
&= \frac{P(D|F)P(F) + P(D|NF)P(NF) - P(D|NF)}{P(D|F)P(F) + P(D|NF)P(NF)} \\
&= \frac{P(D|F)P(F) + P(D|NF)(P(NF) - 1)}{P(D|F)P(F) + P(D|NF)P(NF)} \\
&= \frac{P(D|F)P(F) - P(D|NF)(1 - P(NF))}{P(D|F)P(F) + P(D|NF)P(NF)} \\
&= \frac{P(D|F)P(F) - P(D|NF)P(F)}{P(D|F)P(F) + P(D|NF)P(NF)} \\
&= \frac{P(F)(P(D|F) - P(D|NF))}{P(D|F)P(F) + P(D|NF)P(NF)} \\
&= \frac{P(F)\left(\frac{P(D|F)}{P(D|NF)} - 1\right)}{\frac{P(D|F)P(F)}{P(D|NF)} + \frac{P(D|NF)P(NF)}{P(D|NF)}}.
\end{aligned}$$

Recall that the relative risk is $RR = \frac{P(D|F)}{P(D|NF)}$, consequently,

$$\begin{aligned}
AR &= \frac{P(F)(RR - 1)}{P(F)RR + 1 - P(F)} \\
&= \frac{P(F)(RR - 1)}{P(F)(RR - 1) + 1}.
\end{aligned} \tag{1.8}$$

Written this way, the attributable risk can be calculated from a retrospective study by estimating the RR with the odds ratio, that is,

$$AR \simeq \frac{P(F)(OR - 1)}{P(F)(OR - 1) + 1}. \tag{1.9}$$

The distribution of $1 - AR$ was derived by Walters (1985). Walters (1976) considered the effects of a nuisance factor when calculating the attributable risk. He cites an example of three

risk factors, hyperlipoproteinaemia, smoking, and high diastolic blood pressure, for ischaemic heart disease. In this example, age is the nuisance factor. A nuisance factor is one that affects the disease but is not of interest to the experimenter (Dean and Voss 1999). Walters proposed a weighted average of the attributable risk over the levels of the nuisance factor, age, where the weights are calculated as the proportion of cases in each age group i, i.e.

$$AR_{average} = \frac{\sum_{i=1}^{c} w_i AR_i}{\sum_{i=1}^{c} w_i}. \tag{1.10}$$

In the same paper, Walters suggested a way to deal with the confounding of multiple factors which was first used by Meittinen (1972) with respect to the relative risk. Interaction of multiple factors is addressed by Walters (1983) and by Bruzzi et al. (1985).

1.2.4 Chi-Square Goodness of Fit Test

Pearson (1900) was the first to propose the chi-square goodness of fit test. If in the case of a case-control study there is a factor with c levels, then the distribution of the cases, X_i (where X_i is the number of cases in the i^{th} category) and the controls, Y_i (where Y_i is the number of controls in the i^{th} category) is multinomial with parameters $m, q_0, q_1, \ldots, q_c$ and $n, p_0, p_1, \ldots, p_c$, respectively. We can use the chi-square goodness of fit test to test the null hypothesis $q_0 = p_0$, $q_1 = p_1, \ldots, q_c = p_c$. The test statistic used for this test is given by

$$\sum_{i=1}^{c} \frac{(X_i - mq_i)^2}{mq_i} + \sum_{i=1}^{c} \frac{(Y_i - np_i)^2}{np_i} \tag{1.11}$$

The distribution of this statistic is $\chi^2(2c-2)$. From the case-control study, we can estimate the parameters p_i and q_i. Assuming the null hypothesis, H_o, is true, p_i and q_i can be estimated from the observed frequencies X_i and Y_i as $\frac{X_i + Y_i}{n+m}$, so the test statistic may be written as

$$\sum_{i=1}^{c} \frac{\left(X_i - m\left(\frac{X_i - Y_i}{n+m}\right)\right)^2}{m\left(\frac{X_i - Y_i}{n+m}\right)} + \sum_{i=1}^{c} \frac{\left(Y_i - n\left(\frac{X_i - Y_i}{n+m}\right)\right)^2}{n\left(\frac{X_i - Y_i}{n+m}\right)}. \quad (1.12)$$

The statistic in Eq. (1.12) is $\chi^2(c-1)$, $c-1$ degrees of freedom are lost because estimates replaced the actual parameters. Chase (1972) has considered the situation where the estimates for the parameters are estimated independently of the sample. If H_o is accepted, then the risk factor is not a significant risk for the disease because the distribution of the cases and controls across each level of the risk factor is the same. Begg et al. (1998) has proposed a statistic which measures the square of the coefficient of variation of the incidence rates over the risk categories in a case-control study. This statistic measures the degree of the risk for a given factor and is similar to the idea of the chi-square test mentioned above. In other words, it is a measure of the similarity or dissimilarity of the two multinomial distributions of the cases and controls.

1.2.5 Begg's Estimate of k^2

A new approach to measure the degree of risk associated with a given factor was proposed by Begg et al. (1998). He suggested using the square of the coefficient of variation of the incidence of disease over the categories of risk, k^2. The square of the coefficient of variation is a unitless measure of relative variability (Shafer and Sulivan 1986). It is defined as the ratio of the standard deviation to the mean of a random variable X, that is, $\frac{\sqrt{V(X)}}{E(X)}$. To demonstrate Begg's proposed estimate, hypothetical data will be used. The data displayed below is a hypothetical population in which incidence of disease may be calculated over the categories of risk, unlike a case-control study.

Table 1.4 Incidence of Disease Over Risk Categories

Risk Factor Category i	Cases, X_i	Controls, Y_i	$q_i = \frac{X_i}{240}$	$p_i = \frac{Y_i}{240}$	Incidence of disease
Category 1	40	80	0.17	0.33	$0.33 = \frac{X_i}{X_i - Y_i}$
Category 2	65	40	0.27	0.17	0.62
Category 3	100	60	0.42	0.25	0.63
Category 4	20	40	0.08	0.17	0.33
Category 5	15	20	0.06	0.08	0.43
Total	240	240	1.00	1.00	

Calculating k^2, the square of the coefficient of variation of the incidence of disease over the risk factor categories, gives

$$k^2 = \frac{\frac{\sum_{i=1}^{5}(I_i - \overline{I})^2}{5}}{\overline{I}^2} \qquad (1.13)$$
$$= 0.279$$

where $\overline{I}$ is the overall incidence of the disease and I_i is the incidence of the disease in the i^{th} category. If on the other hand, there is a population where the incidence of the disease is more or less evenly distributed over the risk categories, the hypothetical data may look as follows.

Table 1.5 Incidence of Disease Over Risk Categories

Risk Factor Category i	Cases, X_i	Controls, Y_i	$q_i = \frac{X_i}{240}$	$p_i = \frac{Y_i}{240}$	Incidence of disease
Category 1	75	80	0.31	0.33	$0.48 = \frac{X_i}{X_i - Y_i}$
Category 2	44	40	0.18	0.17	0.52
Category 3	58	60	0.24	0.25	0.49
Category 4	45	40	0.19	0.17	0.53
Category 5	18	20	0.08	0.08	0.47
Total	240	240	1.00	1.00	

The square of the coefficient of variation of the incidence of disease over the risk categories, k^2, in this hypothetical population would be 0.04. As can be seen from the two examples above, the more spread the incidence of disease over the risk categories, the larger is the square of the

coefficient of varation of the incidence of disease over the risk categories. The larger k^2, the more evidence there is that the risk factor under consideration is truly a risk. Begg has suggested a non-parametric approach to estimate the square of the coefficient of variation of the incidence of disease over the risk categories from a case-control study. His estimate or statistic will be denoted by $\hat{k}_B^2$. He also proposes a way to compare this statistic to the square of the overall variation of the incidence of disease over all risk categories, known or unknown, for the entire population. Begg did not investigate the distribution of $\hat{k}_B^2$ nor did he propose a test statistic for drawing inferences about k^2. In the subsequent sections, the distribution and properties of $\hat{k}^2$ are investigated. The first part of the next chapter will provide background on Begg's estimate of k^2.

CHAPTER 2

COEFFICIENT OF VARIATION OF THE INCIDENCE OF DISEASE OVER THE RISK CATEGORIES

In the retrospective model, the sampling procedure was discussed in Chapter 1 but will be summarized here for continuity. A random sample of size m is taken from a population of cases. Another random sample of size n is taken from a population of controls. These samples are then stratified into k risk categories which are determined by the experimenter. The sample of cases and controls are independent with a multinomial distribution of $M(m, q_1, q_2, \ldots, q_c)$ and $M(n, qcon_1, qcon_2, \ldots qcon_c)$, respectively, where $q_i = P(category\ i|disease)$ and $qcon_i = P(category\ i|\text{no } disease)$. The accuracy required of the study results usually determine the sample sizes drawn from the populations above. It is recognized also that cost, logistics, and amount of disease in the population are also of considerable importance when selecting a sample. There are no definitive rules in the literature for selecting sample sizes for a case-control study. From the literature, the range of the proportion of cases to controls is from 1:1 to 1:4 (Beaglehole, Bonita, and Kjelistrom 1993). For the simple dichotomous case, factor or no factor, the data would be displayed in a 2 × 2 table as follows:

Table 2.1 Data for a Case-Control Study Displayed in a 2 × 2 Table

	Disease	No Disease
Factor	x_1	y_1
No Factor	x_2	y_2
	$m = x_1 + x_2$	$n = y_1 + y_2$

where x_1 is the number of people with the disease that have the factor, y_1 is the number of people without the disease that have the factor, x_2 is the number of people with the disease that do not have the factor, and y_2 is the number of people without the disease and without the factor.

If the factor can be divided into more than two categories, then the general case of c categories or c levels of risk can be displayed in a $c \times 2$ table as follows:

Table 2.2 Data for a Case-Control Study Displayed in a $c \times 2$ Table

	Disease	No Disease
Factor Level 1	x_1	y_1
Factor Level 2	x_2	y_2
⋮	⋮	⋮
Factor Level c	x_c	y_c
Total	$m = \sum_{i=1}^{c} x_i$	$n = \sum_{i=1}^{c} y_i$

Of interest is to compute a statistic $u(x_1, x_2, \cdots x_c, y_1, y_2, \cdots y_c)$ in order to estimate the square of the coefficient of variation of the incidence of disease over the risk categories, that is, $k^2 = \sum_{i=1}^{c} \frac{q_i^2}{p_i} - 1$, from the independent samples. Using the maximum likelihood estimate of k^2 is the most desirable approach because of the desirable properties the estimate possesses (Craig and Hogg 1978).

2.1 The Square of the Coefficient of Variation of the Incidence of Disease over the Risk Categories Estimated from a Case-Control Study

For clarity, a list of the variables that will be used throughout the rest of the study is given.

D = disease

R_i = risk category i

n = number of controls

m = number of cases

y_i = number of observed controls in category i

x_i = number of observed cases in category i

c = number of categories

$q_i = P(R_i|D)$

$qcon_i = P(R_i|\text{no } D)$

$p_i = P(R_i)$

I_i = incidence of disease in category i

$= P(D|R_i)$

μ_I = mean incidence of disease in the population

$$= \sum_{i=1}^{c} I_i p_i = P(D)$$

The expected value of a discrete random variable X, is given by $\sum_x xP(X = x)$ and the expected value of X^2, is given by $\sum_x x^2 P(X = x)$. The variance of X is by definition $E(X^2) - E(X)^2$. Therefore, in this case, the variance of the incidence of disease over the risk categories can be written as

$$\sum_{i=1}^{c} I_i^2 p_i - \left(\sum_{i=1}^{c} I_i p_i \right)^2$$

$$= \sum_{i=1}^{c} I_i^2 p_i - \mu_I^2. \qquad (2.1)$$

If both terms above are divided by μ_I^2, then the square of the coefficient of variation of the incidence of disease over the risk factors is obtained as

$$\frac{\sum_{i=1}^{c} I_i^2 p_i - \mu_I^2}{\mu_I^2}$$

$$= \frac{\sum_{i=1}^{c} I_i^2 p_i - \mu_I^2}{\mu_I^2}$$

$$= \frac{\sum_{i=1}^{c} I_i^2 p_i}{\mu_I^2} - 1.$$

Taking the square root of the above expression gives the coefficient of variation of the incidence of disease over the risk categories, $k = \sqrt{\frac{\sum_{i=1}^{c} I_i^2 p_i}{\mu_I^2} - 1}$.

To estimate k^2 from a case-control study, the following observations were made by Begg et al. (1998).

$$\begin{aligned} q_i &= P(R_i|D) \\ &= \frac{P(R_i)P(D|R_i)}{P(D)} \\ &= \frac{p_i I_i}{\sum_{i=1}^{c} p_i I_i} \end{aligned} \tag{2.2}$$

Hence,

$$\begin{aligned} \hat{q}_i &= \frac{\frac{y_i + x_i}{n+m}\frac{x_i}{n_i + m_i}}{\sum_{i=1}^{c} \frac{y_i + x_i}{n+m}\frac{x_i}{y_i + x_i}} \\ &= \frac{\frac{x_i}{n+m}}{\sum_{i=1}^{c} \frac{x_i}{n+m}} \\ &= \frac{x_i}{m} \end{aligned} \tag{2.3}$$

Also, the incidence of disease can be written as

$$\begin{aligned} I_i &= P(D|R_i) \\ &= \frac{P(D)P(R_i|D)}{P(R_i)} \\ &= \frac{\mu_I q_i}{p_i} \end{aligned} \tag{2.4}$$

If there are multiple risk factors the above may be written as

$$
\begin{aligned}
I(\bar{r}) &= P(D|\bar{r}) \\
&= \frac{P(D)P(\bar{r}|D)}{P(\bar{r})} \\
&= \frac{\mu_I q_1 (q_2|q_1)(q_3|q_2 q_1)\cdots(q_r|q_{r-1}q_{r-2}\cdots q_1)}{p_1 (p_2|p_1)(p_3|p_2 p_1)\cdots(p_r|p_{r-1}p_{r-2}\cdots p_1)},
\end{aligned} \tag{2.5}
$$

where $\bar{r}$ represents an array of risks. This study will concentrate on the case of only one risk factor or more than one independent risk factors. The square of the coefficient of variation of the incidence of disease over the risk categories for one risk category may be written as

$$
k^2 = \frac{\sum_{i=1}^{c} I_i^2 p_i}{\mu_I^2} - 1
$$

replacing I_i with the above expression in Eq. (2.4) gives,

$$
\begin{aligned}
k^2 &= \frac{\sum_{i=1}^{c} \left(\frac{\mu_I q_i}{p_i}\right)^2 p_i}{\mu_I^2} - 1 \\
&= \frac{\sum_{i=1}^{c} \frac{\mu_I^2 q_i^2}{p_i^2} p_i}{\mu_I^2} - 1 \\
&= \sum_{i=1}^{c} \frac{q_i^2}{p_i} - 1
\end{aligned} \tag{2.6}
$$

An assumption that will be made is that the disease is rare in the population; therefore, $p_i \cong qcon_i$. This assumption can be seen from the following argument.

$$
\begin{aligned}
& P(R_i) \\
&= P(R_i|D)P(D) + P\left(R_i|\text{no } D\right)P\left(\text{no } D\right) \\
&\cong P(R_i|D) \times 0 + P\left(R_i|\text{no } D\right) \times 1 \\
&= P\left(R_i|\text{no } D\right) \\
&= qcon_i
\end{aligned}
$$

The theory from this point on will assume that $p_i \cong qcon_i$ unless otherwise stated. Begg et al. (1998) proposes a method to calculate k^2 for the entire population by using the standardized incidence ratio, S, which is the overall incidence rate of second occurences of the disease. In

the notation above, this would be calculated as

$$
\begin{aligned}
S &= \frac{\sum_{i=1}^{t} q_i I_i}{\sum_{i=1}^{t} p_i I_i} \\
&= \frac{\sum_{i=1}^{t} P(R_i|D)P(D|R_i)}{P(D)} \\
&= \frac{\sum_{i=1}^{t} \frac{P(R_i|D)P(R_i \cap D)}{P(R_i)}}{P(D)} \\
&= \frac{\sum_{i=1}^{t} \frac{P(R_i|D)P(R_i|D)P(D)}{P(R_i)}}{P(D)} \\
&= \sum_{i=1}^{t} \frac{P(R_i|D)P(R_i|D)}{P(R_i)} \\
&= \sum_{i=1}^{t} \frac{q_i^2}{p_i} \qquad (2.7)
\end{aligned}
$$

so it can be seen that there is a relationship between S and k^2, that is

$$S = k^2 + 1. \qquad (2.8)$$

S is an estimate that may be available from data bases that record such data for different diseases such as cancer registries. Using this approach, k^2 for the entire population with respect to all risk factors, known and unknown, can be calculated. If an estimate of k^2 is calculated from a case-control study in which there are several risk factors, it may be compared to the overall $k^2 = S - 1$ obtained from an appropriate registry. If the estimated k^2 is smaller than the overall k^2 this may be an indication that there are more risk factors than the case-control study considered.

The following assumptions are made in deriving k^2:

1. The second occurrence of the cancer is distinguishable from a metastatic spread of the first cancer under study.
2. The fundamental risk status of a patient does not change due to the diagnosis of

the first primary.

3. The factors that affect risk of cancer incidence do not affect subsequent survival.

2.2 Begg's Nonparametric Estimate of k^2

Begg et al. (1998) argues that the statistic, $\hat{k}^2$, is biased and is an inflated estimate of k^2. This is shown to be true from the simulation study provided in this study. In order to estimate k^2, Begg derives an expression for k^2 by first simplifying the problem with the assumption that cases and controls are distribution free. He divides a continuous risk factor into risk categories by ranking c controls into $c+1$ risk categories according to the rank of the control's level of the risk factor. He makes the assumption that the risk categories are uniformly distributed with respect to the risk factor. Therefore, the probability of an individual in risk category i, is $P(\text{risk category } i) = p_i = \frac{1}{c+1}$. To see this in a more formal way, it can be shown that the probability of each risk category can be represented as a length, $E(Z_{i-1} - Z_i)$ on a real line between 0 and 1, where Z_i, $i = 1,2,\ldots,c$, is a set of ordered statistics from a uniform distribution on the interval (0,1). The probability density function of the i^{th} ordered statistic, Z_i, $i = 1,2,\ldots,c$, is

$$\begin{aligned} g(z_i) &= \frac{c!}{(i-1)!(c-i)!}[F(z_i)]^{i-1}[1-F(z_i)]^{c-i}f(z_i) \\ &= \frac{c!}{(i-1)!(c-i)!}[z_i]^{i-1}[1-z_i]^{c-i} \qquad 0 < z_i < 1 \qquad (2.9) \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

The expected value of Z_i may be readily calculated as given below.

$$E(Z_i) = \int_0^1 \frac{z_i c!}{(i-1)!(c-i)!}[z_i]^{i-1}[1-z_i]^{c-i}dz_i = \frac{i}{c+1}$$

As can be seen, $E(Z_{i-1} - Z_i) = \frac{i-1}{c+1} - \frac{i}{c+1} = \frac{1}{c+1}$. Also, $E\left[(Z_{i-1} - Z_i)^2\right]$ will be used in the derivation of the estimate of k^2 and is given below.

$$
\begin{aligned}
& E[(Z_{i+1} - Z_i)^2] \\
&= V(Z_{i+1} - Z_i) + [E(Z_{i+1} - Z_i)]^2 \\
&= V(Z_{i+1}) + V(Z_i) - 2COV(Z_{i+1}Z_i) + [E(Z_{i+1} - Z_i)]^2
\end{aligned}
$$

Notice that the variance of Z_i can be estimated as,

$$
\begin{aligned}
& V(Z_i) \\
&= E(Z_i^2) - E(Z_i)^2 \\
&= \int_0^1 \frac{z_i^2 c!}{(i-1)!(c-i)!}[z_i]^{i-1}[1-z_i]^{c-i}dz_i - \int_0^1 \frac{z_i c!}{(i-1)!(c-i)!}[z_i]^{i-1}[1-z_i]^{c-i}dz_i
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{i(i+1)}{(c+1)(c+2)} - \frac{i^2}{(c+1)^2} \\
&= \frac{i(c-i+1)}{(c+1)^2(c+2)}
\end{aligned}
$$

and the $COV(Z_{i+1}Z_i)$ is

$$
\begin{aligned}
& COV(Z_{i+1}Z_i) \\
&= E(Z_{i+1}Z_i) - E(Z_{i+1})E(Z_i) \\
&= \int_0^1 \int_0^1 \frac{z_{i+1} z_i c! z_i^{i-1}(1-z_{i+1})^{c-i-1}}{(i-1)!(c-i-1)!} dz_{i+1}dz_i - \frac{i(i+1)}{(c+1)^2} \\
&= \frac{i(c-i)}{(c+1)^2(c+2)}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& E(Z_{i+1} - Z_i)^2 \\
&= V(Z_{i+1}) + V(Z_i) - 2COV(Z_{i+1}Z_i) + [E(Z_{i+1} - Z_i)]^2 \\
&= \frac{(i+1)(c-i)}{(c+1)^2(c+2)} + \frac{i(c-i+1)}{(c+1)^2(c+2)} - 2\frac{i(c-i)}{(c+1)^2(c+2)} + \frac{1}{(c+1)^2} \\
&= \frac{2}{(c+1)(c+2)}.
\end{aligned} \tag{2.10}
$$

From the above discussion, the average "length" of a risk category corresponds to the expected value of an individual belonging to category i, that is $E(p_i) = \frac{1}{c+1}$ and $E(p_i^2) = \frac{2}{(c+1)(c+2)}$. Begg makes the assumption that within each risk category the ratio of q_i to p_i remains constant, that is, $r_i = \frac{q_i}{p_i}$. The distribution of the cases are assumed to be multinomial, with parameters

$M(m; q_1, q_2, \ldots, q_c)$. Similar to a Baysian approach, Begg treats q_i, $i = 1, 2, \ldots, c$, the parameters of the multinomial distribution, as a random variable. The first two moments of q_i are given below.

$$\begin{aligned} & E(q_i) \\ &= E(r_i p_i) \\ &= r_i E(p_i) \\ &= r_i \frac{1}{c+1} \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} & E(q_i^2) \\ &= E(r_i^2 p_i^2) \\ &= r_i^2 E(p_i^2) \\ &= r_i^2 \frac{2}{(c+1)(c+2)}. \end{aligned} \tag{2.12}$$

Again with a Baysian approach, the distribution of the cases is now dependent on the distribution of the parameters q_i. To calculate the expected value of the cases, X_i, the expectation must be conditioned on the parameters, that is,

$$\begin{aligned} & E(X_i) \\ &= E[E(X_i|q_i)] \\ &= E(mq_i) \\ &= mE(q_i) \\ &= mr_i \frac{1}{c+1} \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} & E(X_i^2) \\ &= E[E(X_i^2|q_i)] \\ &= E\left[V(X_i|q_i) + [E(X_i|q_i)]^2\right] \\ &= E[mq_i(1-q_i) + m^2 q_i^2] \\ &= mE(q_i) - mE(q_i^2) + m^2 E(q_i^2) \\ &= mE(q_i) + mE(q_i^2)(m-1) \\ &= mr_i \frac{1}{c+1} + m(m-1) r_i^2 \frac{2}{(c+1)(c+2)}. \end{aligned} \tag{2.14}$$

Notice, that

$$\begin{aligned} & k^2 \\ &= \sum_{i=1}^{c} \frac{q_i^2}{p_i} - 1 \\ &= \sum_{i=1}^{c} p_i r_i^2 - 1 \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} & \sum_{i=1}^{c} p_i r_i \\ &= \sum_{i=1}^{c} p_i \frac{q_i}{p_i} \\ &= \sum_{i=1}^{c} q_i \\ &= 1. \end{aligned}$$

Solving Eq. (2.14) for r_i^2 and substituting it in Eq. (2.15) gives,

$$r_i^2 = \frac{(c+1)(c+2)}{2m(m-1)} E(X_i^2) - \frac{(c+2)}{2(m-1)} r_i \tag{2.16}$$

and

$$\begin{aligned} k^2 &= \sum_{i=1}^{c} p_i r_i^2 - 1 \\ &= \sum_{i=1}^{c} p_i \left\{ \frac{(c+1)(c+2)}{2m(m-1)} E(X_i^2) - \frac{(c+2)}{2(m-1)} r_i \right\} - 1 \\ &= \sum_{i=1}^{c} p_i \frac{(c+1)(c+2)}{2m(m-1)} E(X_i^2) - \sum_{i=1}^{c} p_i r_i \frac{(c+2)}{2(m-1)} - 1 \\ &= \frac{(c+1)(c+2)}{2m(m-1)} \sum_{i=1}^{c} p_i E(X_i^2) - \sum_{i=1}^{c} \frac{x_i}{m} \frac{(c+2)}{2(m-1)} - 1. \end{aligned} \tag{2.17}$$

Begg replaces $E(X_i^2)$ with x_i^2 in order to apply a "shrinkage" factor to the estimate. Also, Begg replaces p_i with its empirical estimator to give the statistic,

$$\hat{k}_b^2 = \frac{(c+1)(c+2)}{2m(m-1)} \sum_{i=1}^{c} \frac{1}{c+1} x_i^2 - \sum_{i=1}^{c} \frac{x_i}{m} \frac{(c+2)}{2(m-1)} - 1$$

$$= \frac{(c+2)}{2m(m-1)} \sum_{i=1}^{c} x_i(x_i - 1) - 1. \tag{2.18}$$

If there is more than one control per risk category, which is the case in an actual study, then the controls *within* each risk category are assumed to be uniformly distributed. Again, within each risk category, the controls are assumed to be a set of ordered statistics, Z_{ij}, $j = 1, 2, \ldots, y_i$, $i = 1, 2, \ldots c$. The same argument as above may be used here to derive the estimate for $E(p_i)$. Now the average "length" of the i^{th} interval will be as large as the number of controls in that interval, or the expected value of the y_i^{th} ordered statistic in the i^{th} category, that is $E(p_i) = \frac{y_i}{n+1}$ and $E(p_i^2) = \frac{y_i(1+y_i)}{(n+1)(n+2)}$. Now we have for the expected value of q_i, $E(q_i) = E(r_i p_i) = r_i \frac{y_i}{n+1}$ and $E(q_i^2) = E(r_i^2 p_i^2) = r_i^2 \frac{y_i(1+y_i)}{(n+1)(n+2)}$. Replacing these expected values in the expected value for X_i^2, gives

$$E(X_i^2) = mE(q_i) + m(m-1)E(q_i^2)$$

$$= mr_i \frac{y_i}{n+1} + m(m-1) r_i^2 \frac{y_i(1+y_i)}{(n+1)(n+2)} \tag{2.19}$$

and solving this equation for r_i^2 gives,

$$r_i^2 = \frac{(n+1)(n+2)}{y_i(1+y_i)m(m-1)} \left[E(X_i^2) - mr_i \frac{y_i}{n+1} \right]. \tag{2.20}$$

Begg's estimate for k^2 then becomes

$$\hat{k}_b^2 = \sum_{i=1}^{c} p_i \left\{ \frac{(n+1)(n+2)}{y_i(1+y_i)m(m-1)} \left[x_i^2 - m r_i \frac{y_i}{n+1} \right] \right\} - 1$$

$$= \frac{(n+1)(n+2)}{m(m-1)} \sum_{i=1}^{c} \frac{p_i}{y_i(1+y_i)} x_i^2 - \sum_{i=1}^{c} \frac{x_i}{m} \frac{y_i(n+2)}{y_i(1+y_i)(m-1)} - 1$$

$$= \frac{(n+1)(n+2)}{m(m-1)} \sum_{i=1}^{c} \frac{y_i}{y_i(1+y_i)(n+1)} x_i^2 - \sum_{i=1}^{c} \frac{x_i}{m} \frac{y_i(n+2)}{y_i(1+y_i)(m-1)} - 1$$

$$= \frac{(n+2)}{m(m-1)} \sum_{i=1}^{c} \frac{x_i(x_i-1)}{1+y_i} - 1. \qquad (2.21)$$

2.3 Maximum Likelihood Estimate of k^2, $\hat{k}^2$

The maximum likelihood estimate of the parameter k^2 can be calculated from the likelihood function associated with $\hat{k}^2$. The probability distribution of the cases is, as stated above, multinomial $M(m; q_1, q_2, \ldots, q_c)$ i.e,.

$$P[X_1 = x_1, \ldots, X_c = x_c] = \frac{m!}{x_1! x_2! \ldots x_c!} q_1^{x_1} q_2^{x_2} \ldots (1 - q_1 - q_2 \cdots - q_{c-1})^{x_c}$$

and the probability distribution of the controls is also multinomial $M(n; p_1, p_2, \ldots, p_c)$, i.e,

$$P[Y_1 = y_1, \ldots, Y_c = y_c] = \frac{n!}{y_1! y_2! \ldots y_c!} p_1^{y_1} p_2^{y_2} \ldots (1 - p_1 - p_2 \cdots - p_{c-1})^{y_c}.$$

The likelihood function is then

$$L(q_1, q_2, \ldots, q_c, p_1, p_2, \ldots, p_c | x_1, x_2, \ldots, x_c, y_1, y_2, \ldots, y_c)$$

$$= \frac{m!}{x_1! x_2! \ldots x_c!} q_1^{x_1} q_2^{x_2} \ldots (1 - q_1 - q_2 \cdots - q_{c-1})^{m - x_1 - x_2 \cdots - x_{c-1}} \times$$

$$\frac{n!}{y_1! y_2! \ldots y_c!} p_1^{y_1} p_2^{y_2} \ldots (1 - p_1 - p_2 \cdots - p_{c-1})^{n - y_1 - y_2 \cdots - y_{c-1}}$$

and the log likelihood is

$$\ln L(q_1, q_2, \ldots, q_c, p_1, p_2, \ldots, p_c | x_1, x_2, \ldots, x_c, y_1, y_2, \ldots, y_c)$$

$$= \ln \frac{m!}{x_1! x_2! \ldots x_c!} + x_1 \ln q_1 + \cdots + (m - x_1 - x_2 \cdots - x_{c-1}) \ln(1 - q_1 - q_2 \cdots - q_{c-1}) +$$

$$\ln \frac{n!}{y_1! y_2! \ldots y_c!} + y_1 \ln p_1 + \cdots + (n - y_1 - y_2 \cdots - y_{c-1}) \ln(1 - p_1 - p_2 \cdots - p_{c-1}).$$

Taking derivatives of the above expression with respect to p_i and q_i for $i = 1, 2 \ldots, c$, and

setting them equal to zero, in order to solve for the maximum likelihood estimators, gives the following,

$$\begin{aligned}&\frac{\partial}{\partial q_1}\ln L(q_1,q_2,\ldots,q_c,p_1,p_2,\ldots,p_c|x_1,x_2,\ldots,x_c,y_1,y_2,\ldots,y_c)\\ &=\frac{x_1}{q_1}+\frac{m-x_1-x_2\cdots-x_{c-1}}{1-q_1-q_2\cdots-q_{c-1}}(-1)\\ &=0\end{aligned}$$

and

$$\begin{aligned}\frac{x_1}{q_1}&=\frac{m-x_1-x_2\cdots-x_{c-1}}{1-q_1-q_2\cdots-q_{c-1}}\\ q_1&=\frac{x_1(1-q_1-q_2\cdots-q_{c-1})}{m-x_1-x_2\cdots-x_{c-1}}.\end{aligned}$$

Likewise,

$$\begin{aligned}&\frac{\partial}{\partial q_2}\ln L(q_1,q_2,\ldots,q_c,p_1,p_2,\ldots,p_c|x_1,x_2,\ldots,x_c,y_1,y_2,\ldots,y_c)=0\\ &\frac{\partial}{\partial q_2}\ln L(q_1,q_2,\ldots,q_c,p_1,p_2,\ldots,p_c|x_1,x_2,\ldots,x_c,y_1,y_2,\ldots,y_c)=0\\ &\qquad\vdots\\ &\frac{\partial}{\partial q_{c-1}}\ln L(q_1,q_2,\ldots,q_c,p_1,p_2,\ldots,p_c|x_1,x_2,\ldots,x_c,y_1,y_2,\ldots,y_c)=0\\ &\frac{\partial}{\partial p_1}\ln L(q_1,q_2,\ldots,q_c,p_1,p_2,\ldots,p_c|x_1,x_2,\ldots,x_c,y_1,y_2,\ldots,y_c)=0\\ &\qquad\vdots\\ &\frac{\partial}{\partial p_{c-1}}\ln L(q_1,q_2,\ldots,q_c,p_1,p_2,\ldots,p_c|x_1,x_2,\ldots,x_c,y_1,y_2,\ldots,y_c)=0\end{aligned}$$

and

$$\begin{aligned}q_2&=\frac{x_2(1-q_1-q_3\cdots-q_{c-1})}{m-x_1-x_3\cdots-x_{c-1}}\\ &\vdots\\ q_{c-1}&=\frac{x_2(1-q_1-q_2\cdots-q_{c-2})}{m-x_1-x_2\cdots-x_{c-2}}\\ p_1&=\frac{y_1(1-p_2-p_3\cdots-p_{c-1})}{n-y_2-y_3\cdots-y_{c-1}}\\ &\vdots\\ p_{c-1}&=\frac{y_1(1-p_1-p_2\cdots-p_{c-2})}{n-y_1-y_2\cdots-y_{c-2}}.\end{aligned}$$

Now there are $c-1$ equations and $c-1$ unknowns. Solving for the unknowns,

$q_1, q_2, \ldots, q_{c-1}, p_1, p_2 \ldots, p_{c-1}$ we get the following,

$$\begin{aligned} \hat{q}_1 &= \frac{x_1}{m} \\ \hat{q}_2 &= \frac{x_2}{m} \\ &\vdots \\ \hat{q}_{c-1} &= \frac{x_{c-1}}{m} \end{aligned}$$

$$\begin{aligned} \hat{p}_1 &= \frac{y_1}{n} \\ \hat{p}_2 &= \frac{y_2}{n} \\ &\vdots \\ \hat{p}_{c-1} &= \frac{y_{c-1}}{n}. \end{aligned}$$

Therefore, the maximum likelihood estimates are the sample estimates. The maximum likelihood estimate for k^2 is

$$\hat{k}^2 = \sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} - 1. \tag{2.22}$$

Another important aspect about our estimate is to determine if it is unbiased. If it is an unbiased estimator, then we know it has minimum variance for all unbiased estimators. The asymptotic expectation of $\hat{k}^2 | q_1, q_2, \ldots, q_c, p_1, p_2 \ldots, p_c$ will be derived in the next section.

2.4 Expected Value of $\hat{k}^2$

Begg's estimate, $\hat{k}_b^2$, relies on assumptions that may not be realistic. Also, the statistical properties of $\hat{k}_b^2$ are not known. A simpler and more useful approach is to use the maximum likelihood estimator $\hat{k}^2 = \sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i^2} - 1$ where $\hat{q}_i$ and $\hat{p}_i$ are the maximum likelihood estimates of q_i and p_i. In this section, the expected value of the maximum likelihood estimator

$$E(\hat{k}^2) = E\left(\sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} - 1 \right)$$

will be derived for the asymptotic case (i.e., assuming that n and m are large).

Let $\hat{v}_i = \frac{\hat{q}_i^2}{\hat{p}_i}$, then $E(\hat{v}_i) = E\left(\frac{\hat{q}_i^2}{\hat{p}_i}\right) = E(\hat{q}_i^2)E\left(\frac{1}{\hat{p}_i}\right)$ because $\hat{q}_i$ and $\hat{p}_i$ are independent. The exact distribution of $E\left(\frac{1}{\hat{p}_i}\right)$ is given by Stephen (1945), but an asymptotic approach similar to that used by Gupta (1975) will be used here.

$$E\left(\frac{1}{\hat{p}_i}\right)$$
$$= E\left(\frac{1}{1-\hat{p}_j}\right)$$

where, $\hat{p}_j = \hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_{i-1} + \hat{p}_{i+1} + \cdots + \hat{p}_c = 1 - \hat{p}_i$.

$$E\left(\frac{1}{1-\hat{p}_j}\right)$$
$$= \sum_{y_j=0}^{n} P(Y_j = y_j)\left(\frac{1}{1-\hat{p}_j}\right)$$

$$= \sum_{y_j=0}^{n} P(Y_j = y_j)\left(\frac{1}{1-\frac{y_j}{n}}\right)$$

Since there is a positive probability that $\frac{y_j}{n}$ can equal 1, therefore, making the denominator zero, a small arbitrary constant e will be added to the total number of controls, n. This gives

$$E\left(\frac{1}{1-\hat{p}_j}\right) \cong \sum_{y_j=0}^{n} P(Y_j = y_j)\left(\frac{1}{1-\frac{y_j}{n+e}}\right). \tag{2.23}$$

Expanding the denominator in Eq. (2.23), the expression may be rewritten as

$$\sum_{y_j=0}^{n} P(Y_j = y_j)\frac{1}{1-\frac{y_j}{n+e}}$$

$$= \sum_{y_j=0}^{n} P(Y_j = y_j)\left(1 + \frac{y_j}{(n+e)} + \frac{y_j^2}{(n+e)^2} + \frac{y_j^3}{(n+e)^3} + \cdots + \frac{y_j^t}{(n+e)^t} + \cdots\right)$$

$$= \sum_{y_j=0}^{n} P(Y_j = y_j) + \sum_{y_j=0}^{n} P(Y_j = y_j)\frac{y_j}{(n+e)} + \sum_{y_j=0}^{n} P(Y_j = y_j)\frac{y_j^2}{(n+e)^2} +$$

$$\cdots + \frac{y_j^t}{(n+e)^t} + \cdots$$

$$= 1 + E\left(\frac{Y_j}{n+e}\right) + E\left(\frac{Y_j^2}{(n+e)^2}\right) + \cdots + E\left(\frac{Y_j^t}{(n+e)^t}\right) + \cdots$$

$$= 1 + \frac{E(Y_j)}{(n+e)} + \frac{E(Y_j^2)}{(n+e)^2} + \cdots + \frac{E(Y_j^t)}{(n+e)^t} + \cdots \quad .$$

If $n \geq t$, the moments of Y_j , are given in general by,

$$E(Y_j^t) = (n+e)(n+e-1)\cdots(n+e-t+1)p_j^t +$$
$$(1+2+\cdots+t-1)(n+e)(n+e-1)\cdots(n+e-t+2)p_j^{t-1} + O((n+e)^{t-2}).$$

Dividing the above expression by $(n+e)^t$ gives

$$\frac{E(Y_j^t)}{(n+e)^t} = \frac{(n+e)(n+e-1)\cdots(n+e-t+1)}{(n+e)^t}p_j^t +$$
$$\frac{(1+2+\cdots+t-1)(n+e)(n+e-1)\cdots(n+e-t+2)}{(n+e)^t}p_j^{t-1} + O\left((n+e)^{-2}\right)$$
$$= \frac{(n+e)}{(n+e)}\frac{(n+e-1)}{(n+e)}\cdots\frac{(n+e-(t-1))}{(n+e)}p_j^t +$$
$$\frac{t(t-1)}{2(n+e)}\frac{(n+e)}{(n+e)}\frac{(n+e-1)}{(n+e)}\cdots\frac{(n+e-(t-2))}{(n+e)}p_j^{t-1} + O\left((n+e)^{-2}\right)$$
$$= \left(1-\frac{1}{(n+e)}\right)\left(1-\frac{2}{(n+e)}\right)\cdots\left(1-\frac{t-1}{(n+e)}\right)p_j^t +$$
$$\frac{t(t-1)}{2(n+e)}\left(1-\frac{1}{(n+e)}\right)\left(1-\frac{2}{(n+e)}\right)\cdots\left(1-\frac{t-2}{(n+e)}\right)p_j^{t-1} + O\left((n+e)^{-2}\right)$$

neglecting the terms which have a power of n greater than two in the denominator gives,

$$\frac{E(Y_j^t)}{(n+e)^t} \cong p_j^t - \left(\frac{1}{n+e} + \frac{2}{n+e} + \cdots + \frac{t-1}{n+e}\right)p_j^t +$$
$$\frac{t(t-1)}{2(n+e)}p_j^{t-1}$$
$$= p_j^t - \frac{t(t-1)}{2(n+e)}p_j^t + \frac{t(t-1)}{2(n+e)}p_j^{t-1}. \qquad (2.24)$$

If $n < t$, all of the terms are excluded because they are either zero or have a term with a power of n greater than 2 in the denominator. The expected value, $E\left(\frac{1}{1-\hat{p}_j}\right)$, can be written as

$$\begin{aligned}
&E\left(\frac{1}{1-\hat{p}_j}\right) \\
&= 1 + \frac{E(Y_j)}{n+e} + \frac{E(Y_j^2)}{(n+e)^2} + \cdots + \frac{E(Y_j^t)}{(n+e)^t} + \cdots \\
&= 1 + p_j + p_j^2 - \left(\frac{2}{2(n+e)}\right)p_j^2 + \left(\frac{2}{2(n+e)}\right)p_j \\
&\quad + \cdots + p_j^t - \left(\frac{t(t-1)}{2(n+e)}\right)p_j^t + \left(\frac{t(t-1)}{2(n+e)}\right)p_j^{t-1} + \cdots \\
&\cong \sum_{t=0}^{\infty} p_j^t - \left(\frac{t(t-1)}{2(n+e)}\right)p_j^t + \left(\frac{t(t-1)}{2(n+e)}\right)p_j^{t-1}. \qquad (2.25)
\end{aligned}$$

An approximation sign is used in Eq. (2.25) because the terms in which the denominator has a power of n that is greater than or equal to two are assumed to be close enough to zero to be ignored. Calculating the sum above, one obtains

$$\begin{aligned}
&\sum_{t=0}^{\infty} p_j^t - \left(\frac{t(t-1)}{2(n+e)}\right)p_j^t + \left(\frac{t(t-1)}{2(n+e)}\right)p_j^{t-1} \\
&= \sum_{t=0}^{\infty} p_j^t - \sum_{t=0}^{\infty}\left(\frac{t(t-1)}{2(n+e)}\right)p_j^t + \sum_{t=0}^{\infty}\left(\frac{t(t-1)}{2(n+e)}\right)p_j^{t-1} \\
&= \frac{1}{1-p_j} - \sum_{t=0}^{\infty}\left(\frac{p_j^2}{2(n+e)}\frac{\partial^2}{\partial p_j^2}p_j^t\right) + \sum_{t=0}^{\infty}\left(\frac{p_j}{2(n+e)}\frac{\partial^2}{\partial p_j^2}p_j^t\right) \\
&= \frac{1}{1-p_j} - p_j^2\frac{\partial^2}{\partial p_j^2}\sum_{t=0}^{\infty}\left(\frac{1}{2(n+e)}p_j^t\right) + p_j\frac{\partial^2}{\partial p_j^2}\sum_{t=0}^{\infty}\left(\frac{1}{2(n+e)}p_j^t\right) \\
&= \frac{1}{1-p_j} - p_j^2\frac{\partial^2}{\partial p_j^2}\left(\frac{1}{2(n+e)(1-p_j)}\right) + p_j\frac{\partial^2}{\partial p_j^2}\left(\frac{1}{2(n+e)(1-p_j)}\right) \\
&= \frac{1}{1-p_j} - \frac{p_j^2}{(n+e)(1-p_j)^3} + \frac{p_j}{(n+e)(1-p_j)^3} \\
&= \frac{1}{1-p_j} + \frac{p_j}{(n+e)(1-p_j)^2}. \qquad (2.26)
\end{aligned}$$

Since $p_i = 1 - p_j$, Eq. 2.26 may be rewritten as

$$\frac{1}{p_i} + \frac{(1-p_i)}{(n+e)p_i^2}.$$

For large n, the small constant e may be ignored. From the above, it is evident that the maximum likelihood estimate is bias, that is,

$$\begin{aligned}
&E\left(\frac{\hat{q}_i^2}{\hat{p}_i}\right) \\
&= E(\hat{q}_i^2)E\left(\frac{1}{\hat{p}_i}\right) \\
&= E(\hat{q}_i^2)E\left(\frac{1}{1-\hat{p}_j}\right) \\
&= \left(V(\hat{q}_i) + (E(\hat{q}_i))^2\right)E\left(\frac{1}{1-\hat{p}_j}\right) \\
&= \left(\frac{q_i(1-q_i)}{m} + q_i^2\right)\left(\frac{1}{p_i} + \frac{(1-p_i)}{np_i^2}\right) \\
&= \frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{mnp_i^2} + \frac{q_i^2}{p_i} + \frac{q_i^2(1-p_i)}{np_i^2}.
\end{aligned}$$

Therefore, the asymptotic expectation of $\hat{k}^2$ is

$$\begin{aligned}
E(\hat{k}^2) &= E\left(\sum_{i=1}^{c}\frac{\hat{q}_i^2}{\hat{p}_i} - 1\right) \\
&= E\left(\sum_{i=1}^{c}\frac{\hat{q}_i^2}{(1-\hat{p}_j)} - 1\right) \\
&= \sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{mnp_i^2} + \frac{q_i^2}{p_i} + \frac{q_i^2(1-p_i)}{np_i^2}\right) - 1. \qquad (2.27)
\end{aligned}$$

If this estimate were unbiased, then the expectation of the estimator would equal the parameter it estimates. In the case of $\hat{k}^2$ this would be $E(\hat{k}^2) = \left(\sum\frac{q_i^2}{p_i} - 1\right)$. From the above expression it is possible to show that the bias is

$$\sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{mnp_i^2} + \frac{q_i^2(1-p_i)}{np_i^2}\right) \qquad (2.27a)$$

and an estimate of the bias is

$$\sum_{i=1}^{c}\left(\frac{\hat{q}_i(1-\hat{q}_i)}{m\hat{p}_i} + \frac{\hat{q}_i(1-\hat{q}_i)(1-\hat{p}_i)}{mn\hat{p}_i^2} + \frac{\hat{q}_i^2(1-\hat{p}_i)}{n\hat{p}_i^2}\right).$$

The simulation study shows that this is a very good approximation of the bias of $\hat{k}^2$.

The asymptotic variance is developed in Section 2.5 using both the 2-D Taylor series approximation and the delta method. The next section investigates the asymptotic expectation of Begg's estimate.

2.5 Expected Value of $\hat{k}_b^2$

Begg did not develop an expected value for his statistic, so an approximate expectation is derived below. Begg's statistic is,

$$\hat{k}_b^2 = \frac{n+2}{m(m-1)} \sum_{i=1}^{c} \frac{x_i(x_i-1)}{1+y_i} - 1 \tag{2.28}$$

Therefore, the expected value of $\hat{k}_b^2$ may be written as,

$$\begin{aligned} E(\hat{k}_b^2) &= E\left(\frac{n+2}{m(m-1)} \sum_{i=1}^{c} \frac{X_i(X_i-1)}{1+Y_i} - 1 \right) \\ &= \frac{n+2}{m(m-1)} E\left(\sum_{i=1}^{c} \frac{X_i(X_i-1)}{1+Y_i} - 1 \right) \\ &= \frac{n+2}{m(m-1)} \sum_{i=1}^{c} E(X_i(X_i-1)) E\left(\frac{1}{1+Y_i} \right) - 1, \end{aligned} \tag{2.29}$$

where X_i and Y_i are independent binomially distributed random variables with parameters m, q_i and n, p_i, respectively. The last line in Eq. (2.29) may be written in such a way because of the independence of X_i and Y_i. The following is an approximate expression for $E\left(\frac{1}{1+Y_i} \right)$.

$$E\left(\frac{1}{1+Y_i} \right) = \sum_{y_i=0}^{\infty} P(Y_i = y_i) \frac{1}{1+y_i}.$$

Expanding the term, $\frac{1}{1+y_i}$, the above can be rewritten as,

$$
\begin{aligned}
&E\left(\frac{1}{1+Y_i}\right)\\
&=\sum_{y_i=0}^{\infty} P(Y_i=y_i)\frac{1}{1+y_i}\\
&=\sum_{y_i=0}^{\infty} P(Y_i=y_i)(1-y_i+y_i^2-y_i^3+\cdots)\\
&=\sum_{y_i=0}^{\infty} P(Y_i=y_i)-\sum_{y_i=0}^{\infty} y_i P(Y_i=y_i)+\sum_{y_i=0}^{\infty} y_i^2 P(Y_i=y_i)-\cdots\\
&=1-E(Y_i)+E(Y_i^2)-\cdots.
\end{aligned}
$$

Since, Y_i has a binomial distribution, with paramters n, p_i, the moments may be written in general as,

$$
\begin{aligned}
E(Y_i) &= np_i\\
E(Y_i^t) &= n(n-1)\cdots(n-t+1)p_i^t\\
&\quad +(1+2+\cdots+t-1)n(n-1)\cdots(n-t+2)p_i^{t-1}+O(n^{t-2}).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&E\left(\frac{1}{1+Y_i}\right)\\
&=1-E(Y_i)+E(Y_i^2)-E(Y_i^3)\cdots\\
&=1-np_i+n(n-1)p_i^2+np_i-(n(n-1)(n-2)p_i^3+(1+2)n(n-1)p_i^2)+\cdots
\end{aligned}
$$

After some cancellation of terms and ignoring any term of order (n^{t-2}), the above can be rewritten as,

$$
\begin{aligned}
&E\left(\frac{1}{1+Y_i}\right)\\
&=1-2n(n-1)p_i^2+5n(n-1)(n-2)p_i^3-9n(n-1)(n-2)(n-3)p_i^4\cdots\\
&=1+\sum_{t=2}^{\infty}\left\{\left(\frac{t(t+1)}{2}-1\right)(-1)^{t-1}(n(n-1)\cdots(n-t+1))p_i^t\right\}. \qquad (2.30)
\end{aligned}
$$

If the numerator and denominator are multiplied by n^t then Eq. (2.30) can be simplified further. As such,

$$
\begin{aligned}
&E\left(\frac{1}{1+Y_i}\right)\\
&= 1+\sum_{t=2}^{\infty}\left(\frac{t(t+1)}{2}-1\right)(-1)^{t+1}(n(n-1)\cdots(n-t+1))\frac{n^t}{n^t}p_i^t\\
&= 1+\sum_{t=2}^{\infty}\left(\frac{t(t+1)}{2}-1\right)(-1)^{t+1}\frac{(n(n-1)\cdots(n-t+1))}{n^t}(np_i)^t\\
&= 1+\sum_{t=2}^{\infty}\left(\frac{t(t+1)}{2}-1\right)(-1)^{t+1}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-t+1}{n}(np_i)^t\\
&= 1+\sum_{t=2}^{\infty}\left(\frac{t(t+1)}{2}-1\right)(-1)^{t+1}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{(t-1)}{n}\right)(np_i)^t.
\end{aligned}
$$

The above can be written as

$$
\begin{aligned}
&E\left(\frac{1}{1+Y_i}\right)\\
&= 1+\sum_{t=2}^{\infty}\left(\frac{t(t+1)}{2}-1\right)(-1)^{t+1}\left(1-\frac{1}{n}-\frac{2}{n}-\cdots-\frac{t-1}{n}\right)(np_i)^t\\
&= 1+\sum_{t=2}^{\infty}\left(\frac{t(t+1)}{2}-1\right)(-1)^{t+1}\left(1-\frac{t(t-1)}{2n}\right)(np_i)^t\\
&= 1+\frac{-1}{(1+np_i)^5}(-2np_i^2+2n^2p_i^2+5n^2p_i^3+5n^3p_i^3+n^3p_i^4+4n^4p_i^4+n^5p_i^5).
\end{aligned}
$$

Therefore, the approximate expected value for the statistic proposed by Begg is,

$$
\begin{aligned}
&E\left(\hat{k}_b^2\right)\\
&= \frac{n+2}{m(m-1)}\sum_{i=1}^{c}E(X_i(X_i-1))E\left(\frac{1}{1+Y_i}\right)-1\\
&= \frac{n+2}{m(m-1)}\sum_{i=1}^{c}\{(V(X_i)+(E(X_i))^2-E(X_i))\\
&\quad\left(1+\frac{-1}{(1+np_i)^5}(-2np_i^2+2n^2p_i^2+5n^2p_i^3+5n^3p_i^3+n^3p_i^4+4n^4p_i^4+n^5p_i^5)\right)\}-1\\
&= \frac{n+2}{m(m-1)}\sum_{i=1}^{c}\{(mq_i(1-q_i)+m^2q_i^2-mq_i)\\
&\quad\left(1+\frac{-1}{(1+np_i)^5}(-2np_i^2+2n^2p_i^2+5n^2p_i^3+5n^3p_i^3+n^3p_i^4+4n^4p_i^4+n^5p_i^5)\right)\}-1. \qquad (2.31)
\end{aligned}
$$

2.6 Asymptotic Variance of $\hat{k}^2$

In order to develop the two dimensional Taylor series approximation for the variance of $\hat{k}^2$, each term in the expression for $\hat{k}^2$ will be approximated separately, that is, let $g_i = \frac{\hat{q}_i^2}{\hat{p}_i}$ then,

$$g_i \cong g_i(\mu) + \frac{\partial g_i(\mu)}{\partial \hat{q}_i}(\hat{q}_i - q_i) + \frac{\partial g_i(\mu)}{\partial \hat{p}_i}(\hat{p}_i - p_i)$$

where $g(\mu) = \frac{q_i^2}{p_i}$. If we take the variance of the above expression, we obtain

$$\begin{aligned}
V(g_i) &\cong V\left(g_i(\mu) + \frac{\partial g_i(\mu)}{\partial \hat{q}_i}(\hat{q}_i - q_i) + \frac{\partial g_i(\mu)}{\partial \hat{p}_i}(\hat{p}_i - p_i)\right) \\
&= V(g_i(\mu)) + V\left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}(\hat{q}_i - q_i)\right) + V\left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}(\hat{p}_i - p_i)\right) \\
&\quad + 2\left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)\left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right) COV(\hat{q}_i, \hat{p}_i) \\
&= \left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)^2 V(\hat{q}_i - q_i) + \left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right)^2 V(\hat{p}_i - p_i) \\
&\quad + 2\left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)\left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right) COV(\hat{q}_i, \hat{p}_i) \\
&= \left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)^2 V(\hat{q}_i) + \left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right)^2 V(\hat{p}_i) \\
&\quad + 2\left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)\left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right) COV(\hat{q}_i, \hat{p}_i).
\end{aligned}$$

Since $\hat{q}_i$ and $\hat{p}_i$ are independent, the last term is equal to zero. Hence,

$$V(g_i) = \left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)^2 V(\hat{q}_i) + \left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right)^2 V(\hat{p}_i). \tag{2.32}$$

Each term in the above expression can now be calculated.

$$V(\hat{p}_i) = \frac{p_i(1-p_i)}{n}$$
$$V(\hat{q}_i) = \frac{q_i(1-q_i)}{m}$$

$$\frac{\partial g_i}{\partial \hat{q}_i} = \frac{2\hat{q}_i}{\hat{p}_i}$$
$$\frac{\partial g_i}{\partial \hat{q}_i} = \frac{-\hat{q}_i^2}{\hat{p}_i}.$$

Therefore,

$$V\left(\sum_{i=1}^{c} g_i\right) = \sum_{i=1}^{c} V(g_i) + 2\sum_{j=1}^{c-1}\sum_{i=j+1}^{c} COV(g_i g_j),$$

where,

$$\begin{aligned} COV(g_i g_j) &= E(g_i g_j) - E(g_i)E(g_j) \\ &\cong E\left\{\left(g_i(\mu) + \frac{\partial g_i(\mu)(\hat{q}_i - q_i)}{\partial \hat{q}_i} + \frac{\partial g_i(\mu)(\hat{p}_i - p_i)}{\partial \hat{p}_i}\right) \times \right. \\ &\quad \left.\left(g_j(\mu) + \frac{\partial g_j(\mu)(\hat{q}_j - q_j)}{\partial \hat{q}_j} + \frac{\partial g_j(\mu)(\hat{p}_j - p_j)}{\partial \hat{p}_j}\right)\right\} - E(g_i)E(g_j) \\ &= \left(\frac{\partial g_i(\mu)}{\partial \hat{q}_i}\right)\left(\frac{\partial g_j(\mu)}{\partial \hat{q}_j}\right) COV(\hat{q}_i, \hat{q}_j) + \left(\frac{\partial g_i(\mu)}{\partial \hat{p}_i}\right)\left(\frac{\partial g_j(\mu)}{\partial \hat{p}_j}\right) COV(\hat{p}_i, \hat{p}_j). \end{aligned}$$

Since,

$$COV(\hat{q}_i \hat{q}_j) = -\frac{q_i q_j}{m}$$

and

$$COV(\hat{p}_i \hat{p}_j) = -\frac{p_i p_j}{n},$$

then

$$\begin{aligned} V(\hat{k}^2) = V\left(\sum_{i=1}^{c} g_i\right) &\cong \sum_{i=1}^{c}\left(\frac{4q_i^2(q_i(1-q_i))}{p_i^2 m} + \frac{q_i^4(p_i(1-p_i))}{p_i^4 n}\right) \\ &\quad + 2\sum_{j=1}^{c-1}\sum_{i=j+1}^{c}\left(\frac{4q_i q_j(-q_i q_j)}{p_i p_j m} + \frac{4q_i^2 q_j^2(-p_i p_j)}{p_i^2 p_j^2 n}\right). \end{aligned} \tag{2.33}$$

2.7 Asymptotic Variance of $\hat{k}_b^2$

An asymptotic variance for Begg's estimate may be derived in a similar way. Recall Begg's estimate is

$$\hat{k}_b^2 = \frac{n+2}{m(m-1)}\sum_{i=1}^{c}\frac{X_i(X_i-1)}{1+Y_i} - 1.$$

Let $g_i = \frac{X_i(X_i-1)}{Y_i+1}$, then

$$
\begin{aligned}
V(X_i) &= m(q_i(1-q_i)) \\
V(Y_i) &= n(p_i(1-p_i)) \\
COV(X_iX_j) &= -m(q_iq_j) \\
COV(Y_iY_j) &= -n(p_ip_j)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial g_i}{\partial X_i} &= \frac{2X_i - 1}{Y_i + 1} \\
\frac{\partial g_i}{\partial Y_i} &= -\frac{X_i^2 - X_i}{(Y_i + 1)^2}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
V\left(\sum_{i=1}^{c} g_i\right) \cong & \sum_{i=1}^{c}\left(\frac{2X_i - 1}{Y_i + 1}\right)^2 m(q_i(1-q_i)) + \\
& \sum_{i=1}^{c}\left(-\frac{X_i^2 - X_i}{(Y_i + 1)^2}\right)^2 n(p_i(1-p_i)) + \\
& 2\sum_{j=1}^{c-1}\sum_{i=j+1}^{c}\left(\left(\frac{2X_i - 1}{Y_i + 1}\right)\left(\frac{2X_j - 1}{Y_j + 1}\right)(-m(q_iq_j))\right) + \\
& 2\sum_{j=1}^{c-1}\sum_{i=j+1}^{c}\left(\left(-\frac{X_i^2 - X_i}{(Y_i + 1)^2}\right)\left(-\frac{X_j^2 - X_j}{(Y_i + 1)^2}\right)(-n(p_ip_j))\right).
\end{aligned} \tag{2.34}
$$

Therefore,

$$
\begin{aligned}
V(\hat{k}_b^2) &= V\left(\frac{n+2}{m(m-1)}\sum_{i=1}^{c}\frac{X_i(X_i-1)}{1+Y_i} - 1\right) \\
&= \left(\frac{n+2}{m(m-1)}\right)^2 V\left(\sum_{i=1}^{c}\frac{X_i(X_i-1)}{1+Y_i}\right) \\
&= \left(\frac{n+2}{m(m-1)}\right)^2 V\left(\sum_{i=1}^{c} g_i\right)
\end{aligned} \tag{2.35}
$$

where $V(g_i)$ is given in Eq. (2.34).

2.8 Variance of $\hat{k}^2$ Using the Delta Method

The delta method, as outlined by Bishop, (1975) can be used to develop an expression

for the variance of $\hat{k}^2$. The method uses the first two terms of the Taylor series expansion to approximate the function $f(X)$ of the random variable X with mean μ, that is,

$$f(X) \cong f(\mu) + \frac{\partial f(\mu)}{\partial (X)}(X - \mu).$$

Taking the variance of the above expansion gives an approximate expression for the variance of $f(X)$,

$$\begin{aligned} V(f(X)) &\cong V\left(f(\mu) + \frac{\partial f(\mu)}{\partial (X)}(X - \mu)\right) \\ &= \left(\frac{\partial f(\mu)}{\partial (X)}\right)^2 V(X). \end{aligned}$$

This idea can be expanded to a function h with multiple random variables in the following way.

$$V(h(X,Y,Z)) = \begin{bmatrix} \frac{\partial f(\mu)}{\partial (X)} & \frac{\partial f(\mu)}{\partial (Y)} & \frac{\partial f(\mu)}{\partial (Z)} \end{bmatrix} \begin{bmatrix} V(X) & COV(XY) & COV(XZ) \\ COV(YX) & V(Y) & COV(YZ) \\ COV(ZX) & COV(ZY) & V(Z) \end{bmatrix} \begin{bmatrix} \frac{\partial f(\mu)}{\partial (X)} \\ \frac{\partial f(\mu)}{\partial (Y)} \\ \frac{\partial f(\mu)}{\partial (Z)} \end{bmatrix}$$

For $\hat{k}^2$ there are c categories, so there are $2c$ random variables. The asymptotic variance can then be calculated by multiplying the appropriate matrices, that is

$$V(\hat{k}_2) = \begin{bmatrix} \frac{\partial \hat{k}^2}{\partial(\hat{q}_1)} & \frac{\partial \hat{k}^2}{\partial(\hat{q}_2)} & \cdots & \frac{\partial \hat{k}^2}{\partial(\hat{q}_c)} & \frac{\partial \hat{k}^2}{\partial(\hat{p}_1)} & \frac{\partial \hat{k}^2}{\partial(\hat{p}_2)} & \cdots & \frac{\partial \hat{k}^2}{\partial(\hat{p}_c)} \end{bmatrix} \times$$

$$\begin{bmatrix} V(\hat{q}_1) & COV(\hat{q}_1\hat{q}_2) & \cdots & COV(\hat{q}_1\hat{q}_c) & 0 & 0 & 0 & 0 \\ COV(\hat{q}_2\hat{q}_1) & V(\hat{q}_2) & \cdots & COV(\hat{q}_2\hat{q}_c) & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ COV(\hat{q}_c\hat{q}_1) & COV(\hat{q}_c\hat{q}_2) & \cdots & V(\hat{q}_c) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & V(\hat{p}_1) & COV(\hat{p}_1\hat{p}_2) & \cdots & COV(\hat{p}_1\hat{p}_c) \\ 0 & 0 & 0 & 0 & COV(\hat{p}_2\hat{p}_1) & V(\hat{p}_2) & \cdots & COV(\hat{p}_2\hat{p}_c) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & COV(\hat{p}_c\hat{p}_1) & COV(\hat{p}_c\hat{p}_2) & \cdots & V(\hat{p}_c) \end{bmatrix} \times$$

$$\begin{bmatrix} \frac{\partial \hat{k}^2}{\partial(\hat{q}_1)} \\ \frac{\partial \hat{k}^2}{\partial(\hat{q}_2)} \\ \vdots \\ \frac{\partial \hat{k}^2}{\partial(\hat{q}_c)} \\ \frac{\partial \hat{k}^2}{\partial(\hat{p}_1)} \\ \frac{\partial \hat{k}^2}{\partial(\hat{p}_2)} \\ \vdots \\ \frac{\partial \hat{k}^2}{\partial(\hat{p}_c)} \end{bmatrix}$$

Substituting the appropriate expressions in the matrix above gives

$$V(\hat{k}_2) = \begin{bmatrix} \frac{2\hat{q}_1}{\hat{p}_1} & \frac{2\hat{q}_2}{\hat{p}_2} & \cdots & \frac{2\hat{q}_c}{\hat{p}_c} & \frac{-\hat{q}_1^2}{\hat{p}_1} & \frac{-\hat{q}_2^2}{\hat{p}_2} & \cdots & \frac{-\hat{q}_c^2}{\hat{p}_c} \end{bmatrix} \times$$

$$\begin{bmatrix} \frac{\hat{q}_1(1-\hat{q}_1)}{m} & \frac{-\hat{q}_1\hat{q}_2}{m} & \cdots & \frac{-\hat{q}_1\hat{q}_c}{m} & 0 & 0 & 0 & 0 \\ \frac{-\hat{q}_2\hat{q}_1}{m} & \frac{\hat{q}_2(1-\hat{q}_2)}{m} & \cdots & \frac{-\hat{q}_2\hat{q}_c}{m} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{-\hat{q}_c\hat{q}_1}{m} & \frac{-\hat{q}_c\hat{q}_2}{m} & \cdots & \frac{\hat{q}_c(1-\hat{q}_c)}{m} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\hat{p}_1(1-\hat{p}_1)}{n} & \frac{-\hat{p}_1\hat{p}_2}{n} & \cdots & \frac{-\hat{p}_1\hat{p}_c}{n} \\ 0 & 0 & 0 & 0 & \frac{-\hat{p}_2\hat{p}_1}{n} & \frac{\hat{p}_2(1-\hat{p}_2)}{n} & \cdots & \frac{-\hat{p}_c\hat{p}_1}{n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \frac{-\hat{p}_c\hat{p}_1}{n} & \frac{-\hat{p}_c\hat{p}_2}{n} & \cdots & \frac{\hat{p}_c(1-\hat{p}_c)}{n} \end{bmatrix} \times$$

$$\begin{bmatrix} \frac{2\hat{q}_1}{\hat{p}_1} \\ \frac{2\hat{q}_2}{\hat{p}_2} \\ \vdots \\ \frac{2\hat{q}_c}{\hat{p}_c} \\ \frac{-\hat{q}_1^2}{\hat{p}_1} \\ \frac{-\hat{q}_2^2}{\hat{p}_2} \\ \vdots \\ \frac{-\hat{q}_c^2}{\hat{p}_c} \end{bmatrix}$$

and

$$V(\hat{k}^2) = \sum_{i=1}^{c} \left\{ 2\left(2\frac{q_i^2}{p_i}\frac{1-q_i}{m} \right)\frac{q_i}{p_i} + \frac{q_i^4(1-p_i)}{p_i^4 n} \right\} + 2\sum_{j=1}^{c-1}\sum_{i=j+1}^{c}\left(4\frac{-q_i^2 q_j^2}{p_i p_j m} + \frac{-q_i^2 q_j^2}{p_i p_j n} \right)$$

which is the same as the expression for the variance in Eq. (2.33).

2.9 Expected Value and Asymptotic Variance of $\ln(\hat{k}^2 + 1)$

The expected value of $\ln(\hat{k}^2 + 1)$ may be approximated by a Taylor series expansion about the mean, k_0^2.

$$f(\hat{k}^2) = \ln(\hat{k}^2 + 1)$$
$$\cong \ln(k_0^2 + 1) + \frac{\partial f(k_0^2)}{\partial \hat{k}^2}(\hat{k}^2 - k_0^2) + \frac{\partial^2 f(k_0^2)}{2(\partial \hat{k}^2)^2}(\hat{k}^2 - k_0^2)^2. \quad (2.36)$$

Taking the expectation of both sides of Eq. (2.36), gives the expected value of $\ln(\hat{k}^2 + 1)$ as follows,

$$E(\ln(\hat{k}^2 + 1)) \cong E\left(\ln(k_0^2 + 1) + \frac{\partial f(k_0^2)}{\partial \hat{k}^2}(\hat{k}^2 - k_0^2) + \frac{\partial^2 f(k_0^2)}{2(\partial \hat{k}^2)^2}(\hat{k}^2 - k_0^2)^2 \right)$$
$$= \ln(k_0^2 + 1) + \frac{1}{k_0^2 + 1}E(\hat{k}^2 - k_0^2) - \frac{1}{2(k_0^2 + 1)^2}E\left((\hat{k}^2 - k_0^2)^2\right)$$
$$= \ln(k_0^2 + 1) - \frac{1}{2(k_0^2 + 1)^2}V(\hat{k}^2). \quad (2.37)$$

The asymptotic variance of $\ln(\hat{k}^2 + 1)$ may also be approximated by the first two terms in a Taylor series expansion about the mean, k_0^2.

$$f(\hat{k}^2) = \ln(\hat{k}^2 + 1)$$
$$\cong \ln(k_0^2 + 1) + \frac{\partial f(k_0^2)}{\partial \hat{k}^2}(\hat{k}^2 - k_0^2). \quad (2.38)$$

Taking the variance of both sides of Eq. (2.38) gives

$$V(\ln(\hat{k}^2 + 1)) \cong V\left(\ln(k_0^2 + 1) + \frac{\partial f(k_0^2)}{\partial \hat{k}^2}(\hat{k}^2 - k_0^2) \right)$$
$$= \frac{1}{(k_0^2 + 1)^2}V(\hat{k}^2). \quad (2.39)$$

2.10 Asymptotic Expectation and Variance of $\sum_{i=1}^{r} \ln(\hat{k}_i^2 + 1)$

The underlying assumption in the sum, $\sum_{i=1}^{r} \ln(\hat{k}_i^2 + 1)$, is that the $\hat{k}_i^2$, $i = 1, 2, \ldots, c$ are independent of one another. The expectation of $\sum_{i=1}^{r} \ln(\hat{k}_i^2 + 1)$ is the sum of the expectations of $\ln(\hat{k}_i^2 + 1)$ for $i = 1, 2, \ldots, r$, that is,

$$E\left(\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)\right)=\sum_{i=1}^{r}E\left(\ln\left(\hat{k}_i^2+1\right)\right)$$

$$\cong\sum_{i=1}^{r}\left(\ln(k_{i0}^2+1)-\frac{1}{2(k_{i0}^2+1)^2}V\left(\hat{k}_i^2\right)\right). \tag{2.40}$$

Similarly, the variance of $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)$ is the sum of the variances of $\ln\left(\hat{k}_i^2+1\right)$ for $i=1,2,\ldots,r$,

$$V\left(\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)\right)=\sum_{i=1}^{r}V\left(\ln\left(\hat{k}_i^2+1\right)\right)$$

$$\cong\sum_{i=1}^{r}\left(\frac{1}{\left(\hat{k}_i^2+1\right)^2}V\left(\hat{k}_i^2\right)\right). \tag{2.41}$$

CHAPTER 3

ASYMPTOTIC DISTRIBUTION OF $\hat{k}^2$

The distribution of $\hat{k}^2$ is different depending on whether or not the cases and controls are distributed the same throughout the categories of the risk factor. There are two possible situations. First, the risk factor under consideration is not truly a risk, that is, the cases and controls have the same distribution and second, the risk factor under consideration is a risk, meaning the cases and controls are distributed differently. These two cases will be treated separately in the following sections.

3.1 Asymptotic Distribution of $\hat{k}^2$ given the Factor under Consideration is not a Risk Factor

The distribution of $\hat{k}^2$, if the risk factor under consideration is truly not a risk, can be shown to be asymptotically $Gamma\left(\frac{c-1}{2}, \frac{2(m+n)}{mn}\right)$, where c is the number of categories in the case-control study.

A case-control study is represented as in the table below, where x_i, y_i, q_i, and p_i represent the cases, controls, percent of cases, and percent of controls in category i respectively.

Table 3.1 Data Representation of a Case-Control Study

Category, i	Cases, x_i	Controls, y_i	Percent Cases, $\hat{q}_i$	Percent Controls, $\hat{p}_i$
1	x_1	y_1	$\hat{q}_1 = \frac{x_1}{m}$	$\hat{p}_1 = \frac{y_i}{n}$
2	x_2	y_2	$\hat{q}_2 = \frac{x_2}{m}$	$\hat{p}_2 = \frac{y_1}{n}$
⋮	⋮	⋮	⋮	⋮
c	x_c	y_c	$\hat{q}_c = \frac{x_c}{m}$	$\hat{p}_c = \frac{y_c}{n}$
Total	m	n	1	1

Using the notation from the table, notice that $\hat{k}^2$ may be written in the following way,

$$\begin{aligned}
\hat{k}^2 &= \sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} - 1 \\
&= \sum_{i=1}^{c} \frac{(\hat{q}_i - \hat{p}_i)^2}{\hat{p}_i} \\
&= \sum_{i=1}^{c} \frac{m^2(\hat{q}_i - \hat{p}_i)^2}{m_i^2 \hat{p}} \\
&= \sum_{i=1}^{c} \frac{(m\hat{q}_i - m\hat{p}_i)^2}{m_i^2 \hat{p}} \\
&= \frac{1}{m} \sum_{i=1}^{c} \frac{(x_i - m\hat{p}_i)^2}{m\hat{p}_i}.
\end{aligned} \tag{3.1}$$

Theorem 3.1 $\hat{k}^2$ has a $Gamma\left(\frac{c-1}{2}, \frac{2(m+n)}{mn}\right)$ distribution with shape parameter $\frac{c-1}{2}$ and scale parameter $\frac{2(m+n)}{mn}$.

Proof:

From statistical theory, it is well known that for large m and n, $\sum_{i=1}^{c} \frac{(x_i - mq_i)^2}{mq_i} \sim \chi^2(c-1)$, $\sum_{i=1}^{c} \frac{(y_i - np_i)^2}{np_i} \sim \chi^2(c-1)$, and $\sum_{i=1}^{c} \frac{(x_i - mq_i)^2}{mq_i} + \sum_{i=1}^{c} \frac{(y_i - np_i)^2}{np_i} \sim \chi^2(2c-2)$ (Bishop 1975). If q_i and p_i are estimated by $\hat{q}_i = \hat{p}_i = \frac{x_i + y_i}{m+n}$, $i = 1, 2, \ldots, c$, then,

$$\sum_{i=1}^{c}\frac{(x_i-m\hat{q}_i)^2}{m\hat{q}_i}+\sum_{i=1}^{c}\frac{(y_i-n\hat{p}_i)^2}{n\hat{p}_i}$$

$$=\sum_{i=1}^{c}\frac{\left(x_i-m\frac{x_i+y_i}{m+n}\right)^2}{m\frac{x_i+y_i}{m+n}}+\sum_{i=1}^{c}\frac{\left(y_i-n\frac{x_i+y_i}{m+n}\right)^2}{n\frac{x_i+y_i}{m+n}}\sim\chi^2(c-1). \qquad (3.2)$$

Now it can be shown that Eq. (3.2) above is equal to $(\frac{nm}{n+m})\hat{k}^2$, implying that $(\frac{nm}{n+m})\hat{k}^2$ has a chi-square distribution with $c-1$ degrees of freedom , multiplied by $(\frac{n+m}{nm})$, which is equivalent to a $Gamma\left(\frac{c-1}{2},\frac{2(n+m)}{nm}\right)$. From Eq. 3.2 it is seen that

$$\sum_{i=1}^{c}\frac{\left(x_i-m\frac{x_i+y_i}{m+n}\right)^2}{m\frac{x_i+y_i}{m+n}}+\sum_{i=1}^{c}\frac{\left(y_i-n\frac{x_i+y_i}{m+n}\right)^2}{n\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{\left(\frac{x_i(m+n)-m(x_i+y_i)}{m+n}\right)^2}{m\frac{x_i+y_i}{m+n}}+\sum_{i=1}^{c}\frac{\left(\frac{y_i(m+n)-n(x_i+y_i)}{m+n}\right)^2}{n\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{\left(\frac{x_in-my_i}{m+n}\right)^2}{m\frac{x_i+y_i}{m+n}}+\sum_{i=1}^{c}\frac{\left(\frac{y_im-nx_i}{m+n}\right)^2}{n\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{n^2\left(\frac{x_i-\frac{m}{n}y_i}{m+n}\right)^2}{m\frac{x_i+y_i}{m+n}}+\sum_{i=1}^{c}\frac{n^2\left(\frac{x_i-\frac{m}{n}y_i}{m+n}\right)^2}{n\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{nn^2\left(\frac{x_i-m\hat{p}_i}{m+n}\right)^2+mn^2\left(\frac{x_i-m\hat{p}_i}{m+n}\right)^2}{nm\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{(n+m)n^2\left(\frac{x_i-m\hat{p}_i}{m+n}\right)^2}{nm\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{\frac{n^2}{m+n}(x_i-m\hat{p}_i)^2}{nm\frac{x_i+y_i}{m+n}}$$

$$=\sum_{i=1}^{c}\frac{n(x_i-m\hat{p}_i)^2}{m(x_i+y_i)}. \qquad (3.3)$$

Under the assumption that the cases and controls have the same distribution, $x_i=\frac{m}{n}y_i$, and Eq. (3.3) can be expressed as

$$\sum_{i=1}^{c} \frac{n(x_i - m\hat{p}_i)^2}{m(x_i + y_i)}$$

$$= \sum_{i=1}^{c} \frac{n(x_i - m\hat{p}_i)^2}{m\left(\frac{m}{n}y_i + y_i\right)}$$

$$= \sum_{i=1}^{c} \frac{n(x_i - m\hat{p}_i)^2}{m\left(\frac{m}{n}y_i + y_i\right)}$$

$$= \sum_{i=1}^{c} \frac{n(x_i - m\hat{p}_i)^2}{my_i\left(\frac{m+n}{n}\right)}$$

$$= \sum_{i=1}^{c} \frac{n(x_i - m\hat{p}_i)^2}{m(m+n)\hat{p}_i}$$

$$= \frac{n}{(m+n)} \sum_{i=1}^{c} \frac{(x_i - m\hat{p}_i)^2}{m\hat{p}_i}$$

$$= \frac{n}{(m+n)} m\hat{k}^2. \qquad (3.4)$$

Therefore, $\hat{k}^2$ is $Gamma\left(\frac{c-1}{2}, \frac{2(m-n)}{mn}\right)$.

To show that $\frac{(n-m)}{nm}\chi^2(c-1)$ is a $Gamma\left(\frac{c-1}{2}, \frac{2(n-m)}{nm}\right)$, the transformation of variable technique may be used. Let T be a chi-square with $c-1$ degrees of freedom, that is, $T \sim \chi^2(c-1)$. The probability distribution of T is given by

$$f(t) = \frac{t^{\alpha-1}e^{-\frac{t}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}dt$$

where, $\alpha = \frac{c-1}{2}$ and $\beta = 2$. Let $t = \frac{mn}{(m+n)}\hat{k}^2$ and $dt = \frac{mn}{(m-n)}d\hat{k}^2$, so the probability distribution of $\hat{k}^2$ is

$$f(\hat{k}^2) = \frac{\left(\frac{mn}{(m+n)}\hat{k}^2\right)^{\alpha-1} e^{-\frac{\frac{mn}{(m+n)}\hat{k}^2}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \frac{mn}{(m+n)}d\hat{k}^2$$

$$= \frac{(\hat{k}^2)^{\alpha-1} e^{-\frac{\frac{mn}{(m+n)}\hat{k}^2}{\beta}}}{\Gamma(\alpha)\left(\beta\frac{(m+n)}{mn}\right)^{\alpha}}d\hat{k}^2.$$

This distribution is $Gamma(\alpha, \beta')$, where $\alpha = \frac{c-1}{2}$ and $\beta' = \frac{(m-n)\beta}{mn}$. ■

The simulation study confirms the above theorem. Table 3.2 presents comparisons of the mean and variance of $\hat{k}^2$ from theory and simulation. A full explanation of the simulation study is given in Chapter 4.

Table 3.2 Comparison between Theory and Simulation Concerning the Mean and Variance of $\hat{k}^2$

$k^2 = 0$		Simulation	Simulation	Theoretical	Theoretical
Sample	Sample	Sample	Sample	Average	Variance
Size m	Size n	Average	Variance		
100	100	0.08203	0.003963	0.08	0.0032
500	500	0.01574	0.000123	0.016	0.000128
1000	1000	0.00855	0.000037	0.008	0.000032
3000	3000	0.00257	0.000003	0.0026	0.000003
5000	5000	0.00156	0.000001	0.0016	0.000001
50	100	0.12351	0.008968	0.120	0.007200
300	500	0.02152	0.000225	0.021	0.000227
500	800	0.01351	0.000101	0.013	0.00008
1000	3000	0.00532	0.000014	0.0053	0.000014
5000	7000	0.00137	0.000001	0.0013	0.000001

It is seen from the table that there is good agreement between theory and simulation. As the sample size increases, the difference between the theoretical and simulated values decreases.

3.2 Asymptotic Distribution of $\hat{k}^2$ given the Factor under Consideration is a Risk Factor

In the case where a risk factor is considered a risk, the distribution of $\hat{k}^2 = \frac{1}{m}\sum_{i=1}^{c}\frac{(x_i - m\hat{p}_i)^2}{m\hat{p}_i}$ is different because in the chi-square statistic, $\sum_{i=1}^{c}\frac{(x_i - m\hat{q}_i)^2}{m\hat{q}_i}$, the expected value of the cases in each category is not equal to the observed value, (i.e. since the cases and controls are distributed differently, $m\hat{p}$ is different than $m\hat{q}$). The distribution of $\hat{k}^2$

is dependent on the degree of this difference. If the difference between observed and expected is very small, then $\frac{mn}{(m+n)}\hat{k}^2$ can be seen to have a noncentral chi-square limiting distribution. However, if the difference is not very small, then the limiting distribution is $N(\mu_{\hat{k}^2}^2, \sigma_{\hat{k}^2}^2)$ (Bishop 1975). Considering the first case, q_i is close to p_i, a derivation of the noncentral chi-square can be found as follows: It is seen from Eq. (3.4), (assuming $x_i \cong \frac{m}{n}y_i$ or $q_i \cong p_i$) that

$$\hat{k}^2 = \frac{nm}{(m+n)}\sum_{i=1}^{c}\frac{(x_i - m\hat{p}_i)^2}{m\hat{p}_i}$$
$$= \frac{nm}{(m+n)}\sum_{i=1}^{c}\frac{(x_i - m\hat{q}_i + m\hat{q}_i - m\hat{p}_i)^2}{m\hat{p}_i}.$$

Now $\hat{k}^2$ has a noncentral chi-square distribution multiplied by $\frac{(n+m)}{mn}$ with a noncentrality parameter of $\frac{mn}{(m+n)}\sum_{i=1}^{c}\frac{(m\hat{q}_i - m\hat{p}_i)^2}{m\hat{p}_i}$, and $c-1$ degrees of freedom,

$$\hat{k}^2 \sim \chi^2\left(c-1, \frac{mn}{(m+n)}\sum_{i=1}^{c}\frac{(mq_i - mp_i)^2}{mp_i}\right). \qquad (3.5)$$

The difference between the p_i, q_i $i = 1,2,\ldots,c$, must be small enough to keep the power of the test statistic $\hat{k}^2 \sim \chi^2\left(c-1, \frac{mn}{(m+n)}\sum_{i=1}^{c}\frac{(mq_i - mp_i)^2}{mp_i}\right)$ bounded away from one as n increases (Kendall and Stuart 1979). The simulation study shows if there is an absolute difference, $|p_i - q_i|$, greater than 0.025 for any one category of risk, then the power is not bounded away from one. This corresponds to a value for k^2 of approximately 0.006. Table 3.3 compares the power of the test statistic, $\hat{k}^2 \sim \chi^2\left(c-1, \frac{mn}{(m+n)}\sum_{i=1}^{c}\frac{(mq_i - mp_i)^2}{mp_i}\right)$, simulated from the indicated populations of k^2, with five categories of risk when the size of the test is $\alpha = 0.05$. Setting the noncentrality parameter equal to zero gives the size of the test. The table gives the maximum difference between the parameters p_i and q_i for $i = 1,2,\ldots,c$.

Table 3.3 Comparison of the Power of the Test Statistic $\hat{k}^2$ for Various Populations with Different Sample Sizes from Simulation

		$n = m = 5000$	$n = m = 7000$	$n = m = 9000$
k^2	Absolute Value of the Maximum Difference	$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.05$
	$\lvert p_i - q_i \rvert$			
0.00000	0.000	0.047	0.051	0.046
0.00100	0.010	0.200	0.270	0.340
0.00196	0.014	0.393	0.524	0.624
0.00324	0.018	0.622	0.776	0.856
0.00400	0.020	0.736	0.870	0.936
0.00506	0.023	0.814	0.945	0.969
0.00625	0.025	0.912	0.982	0.995
0.00702	0.027	0.937	0.986	0.999
0.00812	0.029	0.969	0.994	1.000
0.01600	0.040	1.000	1.000	1.000
0.05000	0.050	1.000	1.000	1.000

From Table 3.3, it is apparent that the difference between p_i and q_i, $i = 1, 2, \ldots, c$, must be very small in order to keep the power bounded away from one. Therefore, unless $\hat{k}^2$ calculated from the sample is very small, it may be best to consider $\hat{k}^2$ to be asymptotically normally distributed. If the number of risk categories, c, is increased given the same value of k^2, the power is reduced (Kendall and Stuart 1976). Table 3.4 gives a brief summary comparing the sample mean and variance from the simulation study to the mean and variance of the noncentral chi-square for the indicated populations with five categories of risk, sample size $m = n = 5000$ and 1000 replications.

Table 3.4 Comparisons of Mean and Variance from the Noncentral Chi-Square in Eq. (3.5) and from Simulation

$m = n = 5000$	Simulation	Simulation	Theoretical	Theoretical
	Sample	Sample	Average	Variance
k^2	Average	Variance		
0.00025	0.0018	0.000002	0.00185	0.000002
0.00100	0.0026	0.000003	0.00260	0.000003
0.00196	0.0035	0.000004	0.00356	0.000004
0.00324	0.0048	0.000006	0.00484	0.000006
0.00400	0.0056	0.000007	0.00560	0.000007
0.00506	0.0066	0.000009	0.00666	0.000009
0.00625	0.0079	0.000011	0.00785	0.000011
0.00702	0.0085	0.000012	0.00862	0.000013
0.00812	0.0099	0.000014	0.00972	0.000014
0.01600	0.0174	0.000026	0.01760	0.000027
0.05000	0.0505	0.000100	0.05065	0.0000797

As can be seen from Table 3.4, the simulated variance starts to deviate from the variance of the theoretical noncentral chi-square distribution when k^2 is as large as 0.05.

When $p_i \neq q_i$ $i = 1, 2, \ldots, c$, and the difference is not very small, the distribution of $\hat{k}^2$ has a normal limiting probability distribution, $N\left(\mu_{\hat{k}^2}, \sigma^2_{\hat{k}^2}\right)$. Here, $\mu_{\hat{k}^2}$ is the expected value of $\hat{k}^2$ that was derived in section 2.4, Eq. (2.27), that is,

$$E\left(\hat{k}^2\right) \cong \sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{m(n+e)p_i^2} + \frac{q_i^2}{p_i} + \frac{q_i^2(1-p_i)}{(n+e)(p_i)^2}\right) - 1,$$

and $\sigma^2_{\hat{k}^2}$ is given in Eq. (2.33) by

$$V\left(\hat{k}^2\right) \cong \sum_{i=1}^{c}\left(\frac{4q_i^2(q_i(1-q_i))}{p_i^2 Ncases} + \frac{q_i^4(p_i(1-p_i))}{p_i^4 Ncontrols}\right) + 2\sum_{j=1}^{c-1}\sum_{i=j-1}^{c}\left(\frac{4q_iq_j(-q_iq_j)}{p_ip_j Ncases} + \frac{4q_i^2q_j^2(-p_ip_j)}{p_i^2p_j^2 Ncontrols}\right).$$

From the simulation study, a brief summary is given below using the parameters

$\mathbf{p} = (0.2, 0.2, 0.2, 0.2, 0.2)$ for the controls, and $\mathbf{q} = (0.5, 0.2, 0.15, 0.1, 0.05)$ for the cases, with $k^2 = 0.625$. Again, there are five risk categories and 1000 replications.

Table 3.5 Comparisons of Mean and Variance from the $N(\mu_{k^2}^2, \sigma_{k^2}^2)$ Distribution and from Simulation for Different Sample Sizes

	Simulation	Simulation	Theoretical	Theoretical
Sample	Sample	Sample	Average	Variance
Size, $m = n$	Average	Variance		
100	0.742	0.1283	0.725	0.1258
500	0.647	0.0197	0.645	0.0185
1000	0.634	0.0089	0.635	0.0089
3000	0.629	0.0029	0.628	0.0029
5000	0.626	0.0017	0.627	0.0017

The following table uses the same parameters as those for Table 3.5 with the exception of the sample size of the cases and controls.

Table 3.6 Comparisons of Mean and Variance from the $N(\mu_{k^2}^2, \sigma_{k^2}^2)$ Distribution and from Simulation for Different Sample Sizes

		Simulation	Simulation	Theoretical	Theoretical
Sample	Sample	Sample	Sample	Average	Variance
Size m	Size n	Average	Variance		
100	200	0.69442	0.0688	0.6952	0.0696
300	500	0.6546	0.0237	0.6498	0.0232
1000	3000	0.6306	0.0051	0.6309	0.0051
3000	5000	0.6258	0.0020	0.6274	0.0021
5000	7000	0.6269	0.0015	0.6266	0.0014

3.3 Asymptotic Distribution of $\ln(\hat{k}^2+1)$ and $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$

The distribution of $\ln(\hat{k}^2+1)$ is of importance if more than one risk factor is investigated. As Begg et al. (1998) pointed out, the standardized incidence ratio, $S = k^2 + 1$, is a measure of the square of the overall coefficient of variation of the incidence of disease over all risk categories known and unknown for the entire population. If k_i^2 is the square of the coefficient of variation of the incidence of disease over risk i, for $i = 1,2\ldots,t$, where t represents all of the risks and they are all independent of one another, then $S = \prod_{i=1}^{t}(k_i^2+1)$. Taking the log of both sides gives

$$\ln S = \ln\left(\prod_{i=1}^{t}(k_i^2+1)\right)$$
$$= \sum_{i=1}^{t}\ln(k_i^2+1).$$

Notice, that the value of t is unknown in practice, but an estimate of $\sum_{i=1}^{t}\ln(k_i^2+1)$ may be calculated from a case-control study by

$$\sum_{i=1}^{r}\ln(\hat{k}_i^2+1), \tag{3.6}$$

where r is the number of independent risk factors included in the study. If the null hypothesis is rejected in a test such as $H_o : \sum_{i=1}^{r}\ln(\hat{k}_i^2+1) = \ln S$ vs. $H_a : \sum_{i=1}^{r}\ln(\hat{k}_i^2+1) \neq \ln S$, then there is evidence that not all of the risk factors associated with the disease are included in the case-control study. In order to conduct such a test, the distribution of $\ln(\hat{k}_i^2+1)$ and $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$ must be known. Here, two cases will be considered. The first case will address the distribution of $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$ under the null hypothesis that all the risk factors are not risks,

that is, $H_o : \sum_{i=1}^{r} \ln(k_i^2 + 1) = 0$ vs. $H_a : \sum_{i=1}^{r} \ln(k_i^2 + 1) \neq 0$. The second case will address the distribution of $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$ under the null hypothesis that the sum of all the risk factors is equal to the parameter, $\ln S$, calculated from an appropriate registry, that is, $H_o : \sum_{i=1}^{r} \ln(k_i^2 + 1) = \ln S$ vs. $H_a : \sum_{i=1}^{r} \ln(k_i^2 + 1) \neq \ln S$. For the first case, the distribution of $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$, under the assumption that the null hypothesis is true, (i.e., k_i^2 is not a risk for all $i = 1,2,\ldots,r$ or alternatively, $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right) = 0$) may be found by making the following transformation of variables. As stated in section 3.1, $\hat{k}_i^2$ has the following distribution under the null hypothesis of no risk,

$$\hat{k}_i^2 \sim Gamma\left(\frac{c-1}{2}, \frac{2(m+n)}{mn}\right).$$

If $Y = \hat{k}_i^2 + 1$, then $\hat{k}_i^2 = Y - 1$ and $d\hat{k}_i^2 = dy$. Let the distribution of $\hat{k}_i^2$ under the null hypothesis be represented by

$$f\left(\hat{k}_i^2\right) = \frac{\left(\hat{k}_i^2\right)^{\alpha-1} e^{-\frac{\hat{k}_i^2}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}} d\hat{k}_i^2, \quad \hat{k}_i^2 > 0.$$

where $\beta' = 2\frac{(m+n)}{mn}$. Making the transformation of $Y = \hat{k}_i^2 + 1$, gives

$$f(y) = \frac{(y-1)^{\alpha-1} e^{-\frac{(y-1)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}} dy, \quad y > 1.$$

Now, let $Z = \ln Y = \ln\left(\hat{k}^2 + 1\right)$, so that $Y = e^z$ and $dy = e^z dz$. Making another transformation of variables gives

$$f(z) = \frac{(e^z - 1)^{\alpha-1} e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}} e^z dz, \quad z > 0. \tag{3.7}$$

Since we are considering the null hypothesis of no risk, $\hat{k}_i^2$ will be small and $Z = \ln\left(\hat{k}_i^2 + 1\right)$ will also be small. Consequently, we may approximate e^{-z} with the first two terms of the Taylor series about $Z = 0$, that is, $e^{-z} \cong 1 + Z$. Substituting this in Eq. (3.7) gives

$$
\begin{aligned}
f(z) &\cong \frac{(z)^{\alpha-1} e^{-\frac{(z)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}}(z+1)dz, \quad z > 0 \\
&= \frac{(z)^{\alpha} e^{-\frac{(z)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}}dz + \frac{(z)^{\alpha-1} e^{-\frac{(z)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}}dz \\
&= \beta' \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\left(\frac{(z)^{\alpha} e^{-\frac{z}{\beta'}}}{\Gamma(\alpha+1)(\beta')^{\alpha-1}}dz \right) + \left(\frac{(z)^{\alpha-1} e^{-\frac{z}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}}dz \right).
\end{aligned}
$$

From the above it can be seen that $f(z)$ is approximately the sum of a $Gamma\left(\alpha+1, (\beta')^2 \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}\right)$ and a $Gamma(\alpha, \beta')$. The goal is to sum the r terms that make up $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$. If the risks all have the same number of risk categories, then an approximate distribution of $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$ is given by

$$
Gamma\left(r(\alpha+1), (\beta')^2 \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\right) + Gamma(r\alpha, \beta'). \tag{3.8}
$$

If the risk categories differ for each risk factor, then the approximate distribution is given by

$$
Gamma\left(\sum_{i=1}^{r}(\alpha_i + 1), (\beta')^2 \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\right) + Gamma\left(\sum_{i=1}^{r} \alpha_i, \beta'\right). \tag{3.9}
$$

Another approach is to use the moment generating function to derive the distribution of $\sum_{i=1}^{r} \ln\left(\hat{k}_i^2 + 1\right)$ under the assumption of no risk. The moment generating function for the random variable $Z = \ln Y = \ln\left(\hat{k}_i^2 + 1\right)$ is given, by definition, to be

$$M_Z(t) = M_{\ln(\hat{k}^2+1)}(t) = E(e^{Zt}) = \int_0^\infty \frac{e^{zt}(e^z-1)^{\alpha-1}e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma(\alpha)(\beta')^\alpha}e^z dz$$

$$= \int_0^\infty \frac{e^{zt+z}(e^z-1)^{\alpha-1}e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma(\alpha)(\beta')^\alpha}dz.$$

Since the null hypothesis of no risk is assumed to be true, $\ln(\hat{k}^2+1)$ may be approximated with the first two terms of the Taylor series about zero, that is, $\ln(\hat{k}^2+1) \cong \hat{k}^2$ and

$$M_{\ln(\hat{k}^2+1)}(t) \cong M_{\hat{k}^2}(t) = E(e^{\hat{k}^2 t}) = \int_0^\infty \frac{e^{\hat{k}^2(t-1)}(q)^{\alpha-1}e^{-\frac{(\hat{k}^2)}{\beta'}}}{\Gamma(\alpha)(\beta')^\alpha}\frac{1}{\hat{k}^2}d\hat{k}^2$$

$$= \int_0^\infty \frac{(\hat{k}^2)^{\alpha-1}e^{-\frac{(\hat{k}^2)}{\beta'}-\hat{k}^2 t}}{\Gamma(\alpha)(\beta')^\alpha}d\hat{k}^2$$

$$= \int_0^\infty \frac{(\hat{k}^2)^{\alpha-1}e^{-\hat{k}^2\frac{1-\beta' t}{\beta'}}}{\Gamma(\alpha)(\beta')^\alpha}d\hat{k}^2$$

$$= \left(\frac{1}{1-\beta' t}\right)^\alpha \int_0^\infty \frac{(\hat{k}^2)^{\alpha-1}e^{-\hat{k}^2\frac{1-\beta' t}{\beta'}}}{\Gamma(\alpha)\left(\frac{\beta'}{1-\beta' t}\right)^\alpha}d\hat{k}^2$$

$$= \left(\frac{1}{1-\beta' t}\right)^\alpha. \tag{3.10}$$

This is the moment generating function of that of a $Gamma(\alpha, \beta')$. To obtain the moment generating function of $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$, the moment generating function for each $\hat{k}_i^2$, $i = 1, 2, \ldots, r$, may be multiplied together since each of the $\hat{k}_i^2$, $i = 1, 2, \ldots, r$, are independent. Assuming each risk has the same number of risk categories,

$$M_{\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)}(t) = \left(\frac{1}{1-\beta' t}\right)^{r\alpha}.$$

Alternately, if each of the risk factors has a different number of risk categories, say α_i, $i = 1,2,\ldots,r$, then

$$M_{\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)}(t) = \left(\frac{1}{1-\beta' t}\right)^{\sum_{i=1}^{r}\alpha_i}.$$

Therefore, if $\hat{k}_i^2$, $i = 1,2,\ldots,r$, all contain the same number of controls, n, and the same number of cases, m, then the only change necessary is to sum the r parameters, α_i, $i = 1,2,\ldots,r$. Even though an approximation of $\ln(\hat{k}^2+1)$ was made in order to get a closed form expression for the moment generating function of $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$, the exact distribution may be used with parameters $\alpha = \sum_{i=1}^{r}\alpha_i$ and β',

$$f\left(\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)\right) = f(z) = \frac{(e^z-1)^{\sum_{i=1}^{r}\alpha_i-1}\, e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma\left(\sum_{i=1}^{r}\alpha_i\right)(\beta')^{\sum_{i=1}^{r}\alpha_i}} e^z dz. \qquad (3.11)$$

The simulation study shows Eq. (3.11) to be a very good approximation of the distribution of $\sum_{i=1}^{r}\ln(\hat{k}_i^2+1)$. Table 3.7 below is a comparison of the theoretical mean and variance to the simulated mean and variance of $\ln(\hat{k}_i^2+1)$. Tables 3.8 and 3.9 compare the theoretical mean and variance to the simulated mean and variance of $\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$ and $\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)$, respectively.

Table 3.7 Comparison between Theory and Simulation Concerning the Mean and Variance of $\ln(\hat{k}^2 + 1)$

$k^2 = 0$	Simulation	Simulation	Theoretical	Theoretical
Sample	Sample	Sample	Average	Variance
Size, $m = n$	Average	Variance		
100	0.0773	0.003033	0.0757	0.002573
500	0.0156	0.000118	0.0158	0.001221
1000	0.0084	0.000036	0.0080	0.000031
3000	0.0026	0.000003	0.0026	0.000003
5000	0.0016	0.000001	0.0016	0.000001

Table 3.8 Comparison between Theory and Simulation Concerning the Mean and Variance of $\sum_{i=1}^{2} \ln(\hat{k}^2 + 1)$

$k_1^2 = k_2^2 = 0$	Simulation	Simulation	Theoretical	Theoretical
Sample	Sample	Sample	Average	Variance
Size, $m = n$	Average	Variance		
100	0.1611	0.008029	0.1461	0.004502
500	0.0314	0.000266	0.0314	0.000236
1000	0.0159	0.000067	0.0158	0.000062
3000	0.0053	0.000007	0.0053	0.000007
5000	0.0031	0.000002	0.0031	0.000002

Table 3.9 Comparison between Theory and Simulation Concerning the Mean and Variance of $\sum_{i=1}^{3} \ln(\hat{k}^2 + 1)$

$k_1^2 = k_2^2 = k_3^2 = 0$	Simulation	Simulation	Theoretical	Theoretical
Sample	Sample	Sample	Average	Variance
Size, $m = n$	Average	Variance		
100	0.2404	0.012319	0.2121	0.005955
500	0.0478	0.000407	0.0467	0.000345
1000	0.0248	0.000107	0.0237	0.000091
3000	0.0080	0.000011	0.0079	0.000010
5000	0.0048	0.000004	0.0048	0.000004

From the tables it can be seen that the agreement is stronger the larger the sample size, as would be expected.

Under the null hypothesis, $\sum_{i=1}^{r} \ln(k_i^2 + 1) = 0$, the distribution in Eq. (3.8), (3.9), or (3.11) may be used to determine the critical region of rejection for the test statistic $\sum_{i=1}^{r} \ln(\hat{k}_i^2 + 1)$. The simulation study shows Eq. (3.11) to be the best choice. If the null hypothesis is rejected, then the next step would be to perform multiple comparisons among the risk factors. The next section derives the asymptotic distribution of the difference of two independent risk factors, $\hat{k}_1^2 - \hat{k}_2^2$.

Under the null hypothesis, $H_o : \sum_{i=1}^{r} \ln(k_i^2 + 1) = \ln S$, the simulation study shows that $\ln(\hat{k}_i^2 + 1)$ has an asymptotic normal distribution with mean

$$\ln(\mu_i + 1) - \frac{1}{2(\mu_i + 1)^2}\sigma_i^2 \tag{3.12}$$

and variance

$$\frac{\sigma_i^2}{(\mu_i + 1)^2}. \tag{3.12a}$$

Here, μ_i is the mean associated with $\hat{k}_i^2$ and given by Eq. (2.27). Likewise, σ_i^2 is the variance associated with $\hat{k}_i^2$ and is given by Eq. (2.33). Eq. (3.12) and (3.13) may be derived as follows. Using the Taylor series expansion about μ_i, $\ln\left(\hat{k}_i^2+1\right)$ may be approximated as

$$\ln\left(\hat{k}_i^2+1\right) \cong \ln(\mu_i+1)+\frac{1}{\mu_i+1}\left(\hat{k}_i^2-\mu_i\right)-\frac{1}{2(\mu_i+1)^2}\left(\hat{k}_i^2-\mu_i\right)^2.$$

The mean of this may be calculated as

$$\begin{aligned} E\left(\ln\left(\hat{k}_i^2+1\right)\right) &= E\left(\ln(\mu_i+1)+\frac{1}{\mu_i+1}\left(\hat{k}_i^2-\mu_i\right)-\frac{1}{2(\mu_i+1)^2}\left(\hat{k}_i^2-\mu_i\right)^2\right) \\ &= \ln(\mu_i+1)-\frac{1}{2(\mu_i+1)^2}E\left(\left(\hat{k}_i^2-\mu_i\right)^2\right) \\ &= \ln(\mu_i+1)-\frac{1}{2(\mu_i+1)^2}V\left(\hat{k}_i^2\right) \\ &= \ln(\mu_i+1)-\frac{1}{2(\mu_i+1)^2}\sigma_i^2 \end{aligned}$$

and the variance as

$$\begin{aligned} V\left(\ln\left(\hat{k}_i^2+1\right)\right) &= V\left(\ln(\mu_i+1)+\frac{1}{\mu_i+1}\left(\hat{k}_i^2-\mu_i\right)\right) \\ &= \frac{1}{(\mu_i+1)^2}V\left(\hat{k}_i^2\right) \\ &= \frac{1}{(\mu_i+1)^2}\sigma_i^2. \end{aligned}$$

The distribution of $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)$ is the sum of r asymptotically normal random variables and, therefore, is asymptotically normal with mean

$$\sum_{i=1}^{r}\left(\ln(\mu_i+1)-\frac{1}{2(\mu_i+1)^2}\sigma_i^2\right) \tag{3.13}$$

and variance

$$\sum_{i=1}^{r}\frac{\sigma_i^2}{(\mu_i+1)^2}. \tag{3.13a}$$

The test statistic for the hypothesis test $Ho: \ln(k_i^2+1)=\ln(\mu_i+1)=\ln S$ vs. $Ha: \ln(k_i^2+1)\neq\ln(\mu_i+1)=\ln S$ is

$$\frac{\ln\left(\hat{k}_i^2+1\right)-\left(\ln S-\frac{1}{2(\mu_i-1)^2}\sigma_i^2\right)}{\frac{\sigma_i}{(\mu_i+1)}}$$

and has a standard normal distribution. In most situations, μ_i and σ_i^2 will not be available, only the parameter S is obtainable from an appropriate registry. Therefore, an alternative to this test statistic for large sample sizes is to estimate μ_i and σ_i^2 from the sample to give

$$\frac{\ln\left(\hat{k}_i^2+1\right)-\left(\ln S-\frac{\hat{\sigma}_i^2}{2\left(\hat{k}_i^2-1\right)^2}\right)}{\frac{\hat{\sigma}_i}{\left(\hat{k}_i^2+1\right)}}. \tag{3.14}$$

Table 3.10 compares the average over 1000 replications of the theoretical mean and the average over 1000 replications of the theoretical variance given in Eq. (3.12) and Eq. (3.12a), respectively, to the simulated sample mean and sample variance of $\ln\left(\hat{k}_i^2+1\right)$ under the null hypothesis, $\ln(k_i^2+1)=\ln(\mu_i+1)=\ln S$. Here, the parameter value is $k_i^2=0.625$ and $m=n=5000$ with five categories of risk. It can be seen from the table that there is good agreement between the simulated and theoretical values.

Table 3.10 Comparisons of the Mean and Variance from the Simulation to that of Eq. (3.12) and Eq. (3.12a), Respectively

$n = 5000, m = 5000$	Simulation Sample Average	Simulation Sample Variance	Average of the Sample Mean from Eq. (3.12) over 1000 Replications	Average of the Sample Variance from Eq. (3.12a) over 1000 Replications
k_i^2				
0.000	0.0016	0.000001	0.0016	0.000002
0.109	0.1053	0.000189	0.1053	0.000195
0.201	0.1852	0.000147	0.1852	0.00153
0.308	0.2710	0.000413	0.2708	0.000448
0.399	0.3375	0.000577	0.3372	0.000559
0.504	0.4088	0.000565	0.4085	0.000610
0.625	0.4873	0.000624	0.4869	0.000655
0.745	0.5883	0.000503	0.5881	0.000503
0.900	0.6424	0.000984	0.6419	0.000995

Table 3.10a gives the power of this test statistic at the α level of 0.05. Although the distribution is a bit skewed, it is conservative.

Table 3.10a. Power of the Test Statistic $\frac{\ln(\hat{k}_i^2+1)-\left(\ln S-\frac{\hat{\sigma}_i^2}{2\cdot\hat{k}_i^2+1^2}\right)}{\frac{\hat{\sigma}_i}{(\hat{k}_i^2+1)}}$, where $\mu_i = 0.625$, $\ln S = \ln 1.625 = 0.485508$ and $m = n = 5000$ with Five Categories of Risk

$n = 5000, m = 5000$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_i^2		
0.000	1.000	0.000
0.109	1.000	0.000
0.201	1.000	0.000
0.308	1.000	0000
0.399	1.000	0.000
0.504	0.871	0.000
0.625	0.016	0.024
0.799	0.000	0.998
0.900	0.000	1.000

The test statistic for the hypothesis test, $Ho : \sum_{i=1}^{r} \ln(k_i^2+1) = \ln S$ vs. $Ha : \sum_{i=1}^{r} \ln(k_i^2+1) \neq \ln S$ is given by

$$\frac{\sum_{i=1}^{r} \ln(\hat{k}_i^2+1) - \sum_{i=1}^{r}\left(\ln(\mu_i+1) - \frac{1}{2(\mu_i+1)^2}\sigma_i^2\right)}{\sqrt{\sum_{i=1}^{r} \frac{\sigma_i^2}{(\mu_i+1)^2}}}$$

$$= \frac{\sum_{i=1}^{r} \ln(\hat{k}_i^2+1) - \ln S + \sum_{i=1}^{r} \frac{1}{2(\mu_i+1)^2}\sigma_i^2}{\sqrt{\sum_{i=1}^{r} \frac{\sigma_i^2}{(\mu_i+1)^2}}}.$$

Again, in most situations, μ_i and σ_i^2 will not be available, only the parameter S is obtainable from an appropriate registry. Therefore, an alternative to this test statistic is to estimate μ_i and σ_i^2 from the sample to give

$$\frac{\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)-\ln S+\sum_{i=1}^{r}\frac{1}{2\left(\hat{k}_i^2-1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{r}\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}. \tag{3.14a}$$

Table 3.11 compares the average over 1000 replications of the theoretical mean and the average over 1000 replications of the theoretical variance given in Eq. (3.13) and Eq. (3.13a), respectively, to the simulated sample mean and sample variance of $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$ under the null hypothesis, $\sum_{i=1}^{2}\ln(k_i^2+1)=\ln S$. Here, the parameter values are $k_1^2=\mu_1=0.625$, $k_2^2=\mu_2=0.799$, and $\ln S=\sum_{i=1}^{2}\ln(k_i^2+1)=1.0727$ with sample sizes of $m=n=5000$ and five categories of risk. Again, it can be seen from the table that there is good agreement between the simulated and theoretical values.

Table 3.11 Comparisons of the Mean and Variance from the Simulation to that of Eq. Eq. (3.13) and Eq. (3.13a), Respectively

$n = 5000$ $m = 5000$		Simulation Sample Average	Simulation Sample Variance	Average of the Sample Mean from Eq. (3.13) over 1000 Replications	Average of the Sample Variance from Eq. (3.13a) over 1000 Replications
k_1^2	k_2^2				
0.000	0.625	0.48	0.00646	0.48	0.00656
0.109	0.625	0.59	0.00082	0.59	0.00084
0.201	0.625	0.67	0.00077	0.67	0.00080
0.308	0.625	0.75	0.00108	0.75	0.00100
0.399	0.625	0.82	0.00121	0.82	0.00121
0.504	0.625	0.89	0.00122	0.89	0.00126
0.625	0.625	0.97	0.00128	0.97	0.00129
0.799	0.625	1.07	0.00114	1.07	0.00115
0.900	0.625	1.12	0.00154	1.12	0.00160

Tables 3.11a and 3.11b give the power of this test statistic at the α level of 0.05 for the indicated parameters.

Table 3.11a. Power of the Test Statistic $\dfrac{\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)-\ln S-\sum_{i=1}^{2}\frac{1}{2(\hat{k}_i^2+1)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{2}\frac{\hat{\sigma}_i^2}{(\hat{k}_i^2+1)^2}}}$, where $k_1^2 = 0.625$,

$k_2^2 = 0.799$, $\ln S = \sum_{i=1}^{2}\ln(k_i^2+1) = 1.0727$ and $m = n = 5000$

with Five Categories of Risk

$n = 5000, m = 5000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.050	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	1.000	0.000
0.399	0.625	1.000	0.000
0.504	0.625	1.000	0.000
0.625	0.625	0.778	0.000
0.799	0.625	0.024	0.026
0.900	0.625	0.001	0.310

Table 3.11b. Power of the Test Statistic $\dfrac{\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)-\ln S-\sum_{i=1}^{2}\frac{1}{2(\hat{k}_i^2+1)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{2}\frac{\hat{\sigma}_i^2}{(\hat{k}_i^2+1)^2}}}$, where $k_1^2 = 0.625$,

$k_2^2 = 0.000$, $\ln S = \sum_{i=1}^{2}\ln(k_i^2+1)$ =0.485508 and $m = n = 5000$

with Five Categories of Risk

$n = 5000, m = 5000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	Power	Power
0.000	0.625	0.017	0.026
0.109	0.625	0.000	0.969
0.201	0.625	0.000	1.000
0.308	0.625	0.000	1.000
0.399	0.625	0.000	1.000
0.504	0.625	0.000	1.000
0.625	0.625	0.000	1.000
0.799	0.625	0.000	1.000
0.900	0.625	0.000	1.000

It can be seen from Table 3.11b, that the distribution becomes a bit skewed if one of the parameters, k^2, in the sum is zero. However, this is not serious, especially in the case of a two-tailed test.

Tables 3.12, 3.12a, and 3.12b are similar to 3.11, 3.11a, and 3.11b for the test statistic $\dfrac{\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)-\ln S-\sum_{i=1}^{3}\frac{1}{2(\hat{k}_i^2+1)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3}\frac{\hat{\sigma}_i^2}{(\hat{k}_i^2+1)^2}}}$ with the indicated parameters.

Table 3.12 Comparisons of the Mean and Variance from the Simulation to that of Eq. (3.13) and Eq. (3.13a), Respectively

$n = 5000$ $m = 5000$			Simulation Sample Average	Simulation Sample Variance	Average of the Sample Mean from Eq. (3.13) over 1000 Replications	Average of the Sample Variance from Eq. (3.13a) over 1000 Replications
k_1^2	k_2^2	k_3^2				
0.000	0.625	0.799	1.08	0.00114	1.08	0.00115
0.109	0.625	0.799	1.18	0.00126	1.18	0.00130
0.201	0.625	0.799	1.26	0.00127	1.26	0.00131
0.308	0.625	0.799	1.35	0.00150	1.35	0.00160
0.399	0.625	0.799	1.41	0.00165	1.41	0.00170
0.504	0.625	0.799	1.48	0.00171	1.48	0.00175
0.625	0.625	0.799	1.56	0.00150	1.56	0.00161
0.799	0.625	0.799	1.66	0.00163	1.66	0.00166
0.900	0.625	0.799	1.72	0.00206	1.72	0.00210

Tables 3.12a and 3.12b give the power of this test statistic at the α level of 0.05 for the indicated parameters.

Table 3.12a. Power of the Test Statistic $\dfrac{\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)-\ln S-\sum_{i=1}^{3}\frac{1}{2(\hat{k}_i^2+1)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3}\frac{\hat{\sigma}_i^2}{(\hat{k}_i^2+1)^2}}}$, where $k_1^2 = 0.625$,

$k_2^2 = 0.799$, $k_3^2 = 0.201$, $\ln S = \sum_{i=1}^{3}\ln(k_i^2+1)$ =1.25589, and $m = n = 5000$

with Five Categories of Risk

$n = 5000, m = 5000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.998	0.000
0.109	0.625	0.799	0.538	0.000
0.201	0.625	0.799	0.021	0.024
0.308	0.625	0.799	0.000	0.627
0.399	0.625	0.799	0.000	1.000
0.504	0.625	0.799	0.000	1.000
0.625	0.625	0.799	0.000	1.000
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

Table 3.12b. Power of the Test Statistic $\dfrac{\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)-\ln S+\sum_{i=1}^{3}\frac{1}{2(\hat{k}_i^2+1)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3}\frac{\hat{\sigma}_i^2}{(\hat{k}_i^2+1)^2}}}$, where $k_1^2 = 0.625$, $k_2^2 = 0.000$, $k_3^2 = 0.799$, $\ln S$ =1.07273, and $m = n = 5000$ with Five Categories of Risk

$n = 5000, m = 5000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2	Power	Power
0.000	0.625	0.799	0.021	0.028
0.109	0.625	0.799	0.000	0.849
0.201	0.625	0.799	0.000	1.000
0.308	0.625	0.799	0.000	1.000
0.399	0.625	0.799	0.000	1.000
0.504	0.625	0.799	0.000	1.000
0.625	0.625	0.799	0.000	1.000
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

3.4 Asymptotic Distribution of $D = \hat{k}_1^2 - \hat{k}_2^2$ Assuming the Risk Factors are Equal

The distribution of $\hat{k}_1^2 - \hat{k}_2^2$ is of interest in order to compare the degree of risk of different risk factors. The hypothesis test to be conducted is

$$H_o : k_1^2 - k_2^2 = 0$$
$$vs.$$
$$H_a : k_1^2 - k_2^2 \neq 0.$$

A test statistic for this hypothesis test may be given by

$$\frac{\hat{k}_1^2 - \hat{k}_2^2 - \mu_{\hat{k}_1^2 - \hat{k}_2^2}}{\sqrt{\sigma^2_{\hat{k}_1^2 - \hat{k}_2^2}}} = \frac{\hat{k}_1^2 - \hat{k}_2^2 - \left(\mu_{\hat{k}_1^2} - \mu_{\hat{k}_2^2}\right)}{\sqrt{\sigma^2_{\hat{k}_1^2 - \hat{k}_2^2}}}, \tag{3.15}$$

where $\mu_{\hat{k}_i^2}$ is the mean associated with $\hat{k}_i^2$ and given by Eq. (2.27). It may be worth noting here that the mean, $\mu_{\hat{k}_1^2} - \mu_{\hat{k}_2^2}$, under the null hypothesis may be rewritten as

$$\begin{aligned} &\mu_{\hat{k}_1^2 - \hat{k}_2^2} \\ &= \mu_{\hat{k}_1^2} - \mu_{\hat{k}_2^2} = (k_1^2 + bias_1) - (k_2^2 + bias_2) \\ &= bias_1 - bias_2. \end{aligned}$$

Notice that the expected value, $\mu_{\hat{k}_i^2}$ of $\hat{k}_i^2$ given by Eq. (2.27) is dependent on p_r, q_r, $r = 1, 2, \ldots, c$. There may be two distinct sets of p_r, q_r, $r = 1, 2, \ldots, c$, that give the same value of the parameter k_i^2. For a simple example, consider the following,

$$\begin{aligned} k_1^2 = 0.17522 \text{ with } q_1 &= 0.147826, q_2 = 0.252174, q_3 = 0.6 \\ p_1 &= 0.1, p_2 = 0.13, p_3 = 0.77 \end{aligned}$$

and

$$\begin{aligned} k_2^2 = 0.17522 \text{ with } q_1 &= 0.2, q_2 = 0.2, q_3 = 0.6 \\ p_1 &= 0.1, p_2 = 0.13, p_3 = 0.77. \end{aligned}$$

Even though the parameter values of k_1^2 and k_2^2 are the same, the associated expected value given by Eq. (2.27) of each of these parameters is different. The difference is due to the bias, which is also dependent on p_r, q_r, $r = 1, 2, \ldots, c$. For this simple illustration with $m = n = 5000$,

$$\begin{aligned} bias_1 &= 0.00168126 \\ bias_2 &= 0.00178917. \end{aligned}$$

In most situations, $bias_1 - bias_2$ will not be available from the sample but for large sample sizes may be estimated by

$$\hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}=\sum_{i=1}^{c}\left(\frac{\hat{q}_{1i}(1-\hat{q}_{1i})}{m\hat{p}_{1i}}+\frac{\hat{q}_{1i}(1-\hat{q}_{1i})(1-\hat{p}_{1i})}{m(n)\hat{p}_{1i}^2}+\frac{\hat{q}_{1i}^2(1-\hat{p}_{1i})}{(n)(\hat{p}_{1i})^2}\right)-$$
$$\sum_{i=1}^{c}\left(\frac{\hat{q}_{2i}(1-\hat{q}_{2i})}{m\hat{p}_{2i}}+\frac{\hat{q}_{2i}(1-\hat{q}_{2i})(1-\hat{p}_{2i})}{m(n)\hat{p}_{2i}^2}+\frac{\hat{q}_{2i}^2(1-\hat{p}_{2i})}{(n)(\hat{p}_{2i})^2}\right), \quad (3.16)$$

which is the difference between two estimates of the bias given by Eq. (2.27a).

An estimate of $\sigma^2_{\hat{k}_1^2-\hat{k}_2^2}$ in Eq.(3.15), may be obtained by a weighted average of the estimates of the variance (given by Eq. (2.33)) from both samples, that is,

$$\begin{aligned}\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2} &= V\left\{2\left(\hat{\omega}_1\hat{k}_1^2+(1-\hat{\omega}_1)\hat{k}_2^2\right)\right\}\\ &= 4\left(\hat{\omega}_1^2V\left(\hat{k}_1^2\right)+\hat{\omega}_2^2V\left(\hat{k}_2^2\right)\right),\end{aligned} \quad (3.17)$$

where $\hat{\omega}_1$ is given by

$$\hat{\omega}_1=\frac{\frac{1}{V\left(\hat{k}_1^2\right)}}{\frac{1}{V\left(\hat{k}_1^2\right)}+\frac{1}{V\left(\hat{k}_2^2\right)}},$$

and $V(k_i^2)$, $i = 1,2$, is estimated from the sample, that is,

$$\hat{V}\left(\hat{k}^2\right)\cong\sum_{i=1}^{c}\left(\frac{4\hat{q}_i^2(\hat{q}_i(1-\hat{q}_i))}{\hat{p}_i^2m}+\frac{\hat{q}_i^4(\hat{p}_i(1-\hat{p}_i))}{\hat{p}_i^4n}\right)$$
$$+2\sum_{j=1}^{c-1}\sum_{i=j-1}^{c}\left(\frac{4\hat{q}_i\hat{q}_j(-\hat{q}_i\hat{q}_j)}{\hat{p}_i\hat{p}_jm}+\frac{4\hat{q}_i^2\hat{q}_j^2(-\hat{p}_i\hat{p}_j)}{\hat{p}_i^2\hat{p}_j^2n}\right).$$

The test statistic in Eq. (3.15) may be estimated by,

$$\frac{\hat{k}_1^2-\hat{k}_2^2-\hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2}}}, \quad (3.18)$$

where $\hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}$ and $\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2}$ are given by Eq. (3.16) and (3.17), respectively. The simulation study shows that the test statistic in Eq. (3.18) has an asymptotic standard normal distribution.

Tables 3.13 and 3.13a give the power of this test statistic at the α level of 0.05 for the

indicated parameters.

Table 3.13 Power of the Test Statistic $\frac{\hat{k}_1^2 - \hat{k}_2^2 - \hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2}}}$, where $k_1^2 = 0.625$, $k_2^2 = 0.625$ and $m = n = 5000$ with Five Categories of Risk

$n = 5000, m = 5000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	Power	Power
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	1.000	0.000
0.399	0.625	0.992	0.000
0.504	0.625	0.597	0.000
0.625	0.625	0.022	0.023
0.799	0.625	0.000	0.849
0.900	0.625	0.000	0.973

Table 3.13a. Power of the Test Statistic $\frac{\hat{k}_1^2 - \hat{k}_2^2 - \hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2}}}$, where $k_1^2 = 0.000$, $k_2^2 = 0.000$ and $m = n = 5000$ with Five Categories of Risk

$n = 5000, m = 5000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	Power	Power
0.000	0.000	0.021	0.022
0.109	0.000	0.000	0.000
0.201	0.000	0.000	1.000
0.308	0.000	0.000	1.000
0.399	0.000	0.000	1.000
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.900	0.000	0.000	1.000

CHAPTER 4

SIMULATION

4.1 Simulation Procedure

For studying the size and power of the test statistic $\hat{k}^2$, 1000 random samples from different populations were generated using the RNMTN Fortran routine in IMSL (IMSL 1987). Each sample constituted two independent multinomials $M(m;q_1,q_2,\ldots,q_c)$ and $M(n;p_1,p_2,\ldots,p_c)$ for cases and controls, respectively. Two, four, five, six, and eight categories of risk, c, were simulated. For the two, four, six, and eight category cases, four different populations with four sample sizes were simulated, and from the 5 category cases, 11 different populations and 26 sample sizes were simulated. A summary of the simulation parameters used are given below.

For the 5 category case, different sample sizes of (n,m), as shown in tables 4.6-4.19, were used in the simulation with population parameters given in Table 4.1.

Table 4.1 Population Parameters used for Simulating Five Categories of Risk

Population	**q** (Cases)	**p** (Controls)
$k^2 = 0$	0.2, 0.2, 0.2, 0.2, 0.2	0.2, 0.2, 0.2, 0.2, 0.2
$k^2 = 0.049$	0.1, 0.16, 0.31, 0.27, 0.16	0.07, 0.11, 0.33, 0.28, 0.21
$k^2 = 0.109$	0.21, 0.16, 0.17, 0.34, 0.12	0.22, 0.22, 0.22, 0.21, 0.13
$k^2 = 0.201$	0.062 ,0.222, 0.242, 0.352, 0.122	0.23, 0.23, 0.22, 0.23, 0.09
$k^2 = 0.308$	0.14, 0.21, 0.29, 0.16, 0.2	0.2, 0.3, 0.15, 0.25, 0.1
$k^2 = 0.399$	0.27, 0.18, 0.35, 0.15, 0.05	0.5, 0.2, 0.15, 0.1, 0.05
$k^2 = 0.504$	0.2, 0.3, 0.25, 0.15, 0.1	0.4, 0.15, 0.1, 0.2, 0.15
$k^2 = 0.625$	0.5, 0.2, 0.15, 0.1, 0.05	0.2, 0.2, 0.2, 0.2, 0.2
$k^2 = 0.746$	0.2, 0.2, 0.2, 0.2, 0.2	0.5, 0.2, 0.15, 0.1, 0.05
$k^2 = 0.799$	0.02, 0.13, 0.5, 0.05, 0.3	0.2, 0.2, 0.2, 0.2, 0.2
$k^2 = 0.901$	0.13, 0.22, 0.3, 0.15, 0.2	0.5, 0.2, 0.15, 0.1, 0.05

The parameters were chosen to reflect data similar to that found in the literature. From each of the 1000 samples within a given population, the following statistics were computed:

a. $\hat{k}^2 = \sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} - 1$

b. $\hat{k}_b^2 = \frac{n+2}{m(m-1)} \sum_{i=1}^{c} \frac{x_i(x_i-1)}{1+y_i} - 1$

c. $bias_{\hat{k}^2} = \sum_{i=1}^{c} \left(\frac{\hat{q}_i(1-\hat{q}_i)}{m\hat{p}_i} + \frac{\hat{q}_i(1-\hat{q}_i)(1-\hat{p}_i)}{m(n)\hat{p}_i^2} + \frac{\hat{q}_i^2(1-\hat{p}_i)}{(n)(\hat{p}_i)^2} \right)$

d. $\hat{V}(\hat{k}^2)$, the estimate of $V(\hat{k}^2)$ defined in Section 2.6, Eq. (2.33).

e. $\hat{V}(\hat{k}_b^2)$, the estimate of $V(\hat{k}_b^2)$ defined in Section 2.7, Eq. (2.35).

f. $\ln(\hat{k}^2 + 1) = \ln\left(\sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} \right)$

g. $\hat{V}(\ln(\hat{k}^2 + 1))$, the estimate of $V(\ln(\hat{k}^2 + 1))$ defined in Section 2.9, Eq. (2.39).

h. $\sum_{i=1}^{2}(\ln(\hat{k}_i^2 + 1))$ and $\sum_{i=1}^{3}(\ln(\hat{k}_i^2 + 1))$.

i. $\hat{V}\left(\sum_{i=1}^{2}\left(\ln\left(\hat{k}_i^2+1\right)\right)\right)$ and $\hat{V}\left(\sum_{i=1}^{3}\left(\ln\left(\hat{k}_i^2+1\right)\right)\right)$, the corresponding estimates of $V\left(\sum_{i=1}^{2}\left(\ln\left(\hat{k}_i^2+1\right)\right)\right)$ and $V\left(\sum_{i=1}^{3}\left(\ln\left(\hat{k}_i^2+1\right)\right)\right)$ given in Eq. (2.41), where k_i^2 is independent of k_j^2 for all i,j $i \neq j$.

j. $\hat{k}_2^2 - \hat{k}_1^2$, where $\hat{k}_2^2$ is independent of $\hat{k}_1^2$.

k. $\hat{V}\left(\hat{k}_2^2 - \hat{k}_1^2\right)$, the estimate of $V\left(\hat{k}_2^2 - \hat{k}_1^2\right)$ defined by Eq. (3.14).

Additionally, the statistics computed over the 1000 replications are

a. $\hat{\sigma}^2_{\hat{k}^2}$, the sample variance of $\hat{k}^2$ from the 1000 replications or samples.

b. $\hat{\sigma}^2_{\hat{k}_b^2}$, the sample variance of $\hat{k}_b^2$ from the 1000 replications or samples.

c. $\hat{\sigma}^2_{\ln\left(\hat{k}^2+1\right)}$, the sample variance of $\left(\ln\left(\hat{k}^2+1\right)\right)$ from the 1000 replications or samples.

d. the sample variance from the 1000 replications of $\sum_{i=1}^{2}\left(\ln\left(\hat{k}_i^2+1\right)\right)$, and the sample variance from the 1000 replications of $\sum_{i=1}^{3}\left(\ln\left(\hat{k}_i^2+1\right)\right)$.

e. $\hat{\sigma}^2_{\left(\hat{k}_2^2-\hat{k}_1^2\right)}$, the sample variance of $\hat{k}_2^2 - \hat{k}_1^2$ from the 1000 replications or samples.

The average of $\hat{k}^2$ over the 1000 replications was calculated and compared to

1. its parameter value $k^2 = \sum_{i=1}^{c}\frac{q_i^2}{p_i} - 1$.

2. the expected value defined in Section 2.4, Eq. (2.27), that is,

$$E\left(\hat{k}^2\right) = \sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{m(n)p_i^2} + \frac{q_i^2}{p_i} + \frac{q_i^2(1-p_i)}{(n)(p_i)^2}\right) - 1.$$

3. the theoretical expected value based on the $Gamma\left(\frac{c-1}{2}, \left(\frac{m-n}{mn}\beta\right)\right)$ distribution (or equivalently, $\frac{m-n}{mn}\chi^2(c-1)$ distribution if the parameter value for k^2 is zero, or to the noncentral chi-square distribution $\frac{m-n}{mn}\chi^2\left(c-1, \frac{mn}{m-n}\sum\frac{(q_i-p_i)^2}{p_i}\right)$ if $k^2 > 0$).

The average of $\hat{k}_b^2$ over 1000 replications is calculated and compared with its parameter value of $k^2 = \sum_{i=1}^{c}\frac{q_i^2}{p_i} - 1$ and its expected value given in Section 2.5, Eq. (2.31). The average of the

1000 replications of $bias_{\hat{k}^2}$ is compared to the parameter value of

$$bias_{k^2} = \sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{m(n+e)p_i^2} + \frac{q_i^2(1-p_i)}{(n+e)(p_i)^2} \right).$$

The averages of $V(\hat{k}^2)$ and $V(\hat{k}_b^2)$ over the 1000 replications are compared to their corresponding parameter values of $V(\hat{k}^2)$ and $V(\hat{k}_b^2)$ defined in Section 2.6, Eq. (2.33) and Section 2.7, Eq. (2.35), respectively. These averages are also compared to their corresponding variances, $\hat{\sigma}^2_{\hat{k}^2}$ and $\hat{\sigma}^2_{\hat{k}_b^2}$. Additionally, the average of $V(\hat{k}^2)$ and the variance $(\hat{\sigma}^2_{\hat{k}^2})$ are compared to the theoretical variance based on the associated $Gamma\left(\frac{c-1}{2}, \left(\frac{m-n}{mn}\beta\right)\right)$ (or equivalently, $\frac{m-n}{mn}\chi^2(c-1)$ distribution if the parameter value for k^2 is zero, or to the noncentral chi-square distribution $\frac{m-n}{mn}\chi^2(c-1, \frac{mn}{m+n}\sum\frac{(q_i-p_i)^2}{p_i})$ if $k^2 > 0$).

The average of $\ln(\hat{k}^2+1)$ is calculated over the 1000 replications and compared with the corresponding expected value given in Section 2.9, Eq. (2.37),

$$E(\ln(\hat{k}^2+1)) \cong \ln(E(\hat{k}^2)+1) - \frac{1}{2(E(\hat{k}^2)+1)^2}V(\hat{k}^2).$$

The average of $V(\ln(\hat{k}^2+1))$ over the 1000 replications is calculated and compared to its parameter value $V(\ln(\hat{k}^2+1))$ defined in Section 2.9, Eq. (2.39),

$$V(\ln(\hat{k}^2+1)) \cong \frac{1}{(E(\hat{k}^2)+1)^2}V(\hat{k}^2)$$

and to its corresponding variance, $\hat{\sigma}^2_{\ln(\hat{k}^2+1)}$. Similar comparisons are made for $\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$, $\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)$, $V\left(\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)\right)$, and $V\left(\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)\right)$; that is, the average of $\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$ and the average of $\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)$ over 1000 replications are compared with the corresponding expected values given in Section 2.10, Eq. (2.40). The averages of

$V\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right)$ and $V\left(\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)\right)$ over the 1000 replications are compared with the respective parameter values given by Eq. (2.41) in Section 2.10 and compared with their respective sample variances.

The average of $\hat{k}_2^2-\hat{k}_1^2$, where $\hat{k}_2^2$ and $\hat{k}_1^2$ are simulated from two independent populations and, therefore, are independent random variables. The average of $\hat{k}_2^2-\hat{k}_1^2$ is compared with its corresponding expected value given by Eq. (3.13a). Additionally, the average of $V\left(\hat{k}_2^2-\hat{k}_1^2\right)$ over the 1000 replications is compared with Eq. (3.14).

The power associated with a test of the null hypothesis, *Ho*: the factor is not a risk, vs. the alternative hypothesis, *Ha*: the factor is a risk, is also considered in the simulation study. The null hypothesis for no risk is $k^2 = 0$ and the alternative hypothesis is $k^2 \neq 0$. The power of the test [calculated as the percent of the 1000 $\hat{k}^2$'s that exceeded the critical value of rejection for a given α obtained from the $Gamma\left(\frac{c-1}{2},\left(\frac{m+n}{mn}\beta\right)\right)$ distribution (or equivalently, $\frac{m+n}{mn}\chi^2(c-1)$)] is then reported for a particular population and sample size. The power is calculated at α=0.01, 0.025, 0.05 and 0.10 levels.

For the two, four, six, and eight category cases, sample sizes of $(n,m) = (50,50)$, $(100,100)$, $(500,500)$, $(1000,1000)$, $(3000,3000)$, and $(5000,5000)$ were used in the simulation with parameters given in tables 4.2-4.5.

Table 4.2 Population Parameters used for Simulating Two Categories of Risk

Two Category		
Population	**q** (Cases)	**p** (Controls)
$k^2 = 0.000$	0.500, 0.500	0.500, 0.500
$k^2 = 0.048$	0.375, 0.625	0.485, 0.515
$k^2 = 0.100$	0.555, 0.445	0.400, 0.600
$k^2 = 0.496$	0.745, 0.255	0.400, 0.600

Table 4.3 Population Parameters used for Simulating Four Categories of Risk

Four Category		
Population	**q** (Cases)	**p** (Controls)
$k^2 = 0.000$	0.250, 0.250, 0.250, 0.250	0.250, 0.250, 0.250, 0.250
$k^2 = 0.050$	0.289, 0.211, 0.319, 0.181	0.25, 0.25, 0.25, 0.25
$k^2 = 0.102$	0.310, 0.190, 0.310, 0.190	0.200,0.300,0.300,0.200
$k^2 = 0.503$	0.200, 0.200, 0.480, 0.120	0.250, 0.300, 0.200, 0.250

Table 4.4 Population Parameters used for Simulating Six Categories of Risk

Six Category		
Population	**q** (Cases)	**p** (Controls)
$k^2 = 0.000$	0.200, 0.100, 0.200, 0.200, 0.200, 0.100	0.200, 0.100, 0.200, 0.200, 0.200, 0.100
$k^2 = 0.048$	0.200,0.150,0.210,0.210,0.130,0.100	0.200, 0.100, 0.200, 0.200, 0.200, 0.100
$k^2 = 0.100$	0.300, 0.100, 0.210, 0.210, 0.130, 0.050	0.200, 0.100, 0.200, 0.200, 0.200, 0.100
$k^2 = 0.496$	0.210, 0.300, 0.100, 0.100, 0.220, 0.070	0.100, 0.200, 0.200, 0.200, 0.100, 0.200

Table 4.5 Population Parameters used for Simulating Cases for Eight Categories of Risk

Eight Category	
Population	**q** (Cases)
$k^2 = 0.000$	0.100, 0.100, 0.200, 0.100, 0.100, 0.100, 0.200, 0.100
$k^2 = 0.050$	0.100,0.127,0.195,0.105,0.073,0.148,0.152,0.100
$k^2 = 0.104$	0.115, 0.095, 0.175, 0.110, 0.095, 0.155, 0.155, 0.100
$k^2 = 0.502$	0.05, 0.100, 0.100, 0.070, 0.200, 0.150, 0.105, 0.225

Table 4.5a. Population Parameters used for Simulating Controls forEight Categories of Risk

Eight Category	
Population	**p** (Controls)
$k^2 = 0.000$	0.100, 0.100, 0.200, 0.100, 0.100, 0.100, 0.200, 0.100
$k^2 = 0.050$	0.100, 0.100, 0.200, 0.100, 0.100, 0.100, 0.200, 0.100
$k^2 = 0.104$	0.150, 0.050, 0.150, 0.150, 0.100, 0.100, 0.200, 0.100
$k^2 = 0.502$	0.150, 0.050, 0.150, 0.150, 0.100, 0.100, 0.200, 0.100

Parameters were chosen to reflect data similar to that found in the literature. From each of the 1000 samples within a given population, the following statistics were computed:

a. $\hat{k}^2 = \sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} - 1$

b. $bias_{\hat{k}^2} = \sum_{i=1}^{c} \left(\frac{\hat{q}_i(1-\hat{q}_i)}{m\hat{p}_i} + \frac{\hat{q}_i(1-\hat{q}_i)(1-\hat{p}_i)}{m(n)\hat{p}_i^2} + \frac{\hat{q}_i^2(1-\hat{p}_i)}{(n)(\hat{p}_i)^2} \right)$

c. $\hat{V}(\hat{k}^2)$, an estimate of $V(\hat{k}^2)$ defined in Section 2.6, Eq. (2.33).

Similar comparisons were made as that for the five category simulation.

4.2 Simulation Results

Convergence of the estimates to their expected values begins to occur at a sample size of $(n,m) = (500, 500)$ but for brevity, only the comparison at a sample size of $n = m = 5000$ for the simulation study with five categories of risk will be shown.

4.2.1 Results Regarding Measures of Mean and Variance for $\hat{k}^2$ and $\hat{k}_b^2$

Table 4.6 shows agreement between the theoretical expected value of the chi-square distribution given in Chapter 3, $E(\hat{k}^2)$ in Eq. (2.27), and the simulation average. Table 4.6a compares the simulation average of $\hat{k}_b^2$ to the expected value, $E(\hat{k}_b^2)$, given by Eq. (2.31).

Table 4.6 Comparison of Theory and Simulation Concerning the Mean of $\hat{k}^2$

$n = 5000$ $m = 5000$					
Risk Categories 5	Average of Bias over 1000 Replications	Bias Eq. (2.27a)	$E(\hat{k}^2)$ Eq. (2.27)	Average from Simulation	Chi-Square Mean Central ($k^2 = 0$) Noncental ($k^2 > 0$)
	$bias_{\hat{k}^2}$	$bias_{k^2}$			value
k^2					
0.000	0.001604	0.001600	0.001600	0.001555	0.00160
0.049	0.002004	0.002001	0.051001	0.050507	0.05173
0.109	0.001661	0.001658	0.110660	0.111196	0.11119
0.201	0.001849	0.001843	0.202840	0.203571	0.20418
0.308	0.002511	0.002498	0.310500	0.311583	0.31165
0.399	0.002665	0.002655	0.401660	0.401869	0.40257
0.504	0.002989	0.002984	0.506980	0.505457	0.50713
0.625	0.001978	0.001975	0.626980	0.626550	0.62715
0.746	0.005678	0.005638	0.751640	0.751454	0.75256
0.799	0.002085	0.002079	0.801080	0.801460	0.80190
0.901	0.005736	0.005720	0.906720	0.901945	0.90674

Table 4.6a. Comparison of Theory and Simulation Concerning the Mean of $\hat{k}_b^2$

$n = 5000, m = 5000$		
Risk Categories = 5	$E(\hat{k}_b^2)$ Eq.(2.31)	Average from Simulation
k^2		
0.000	0.00020	0.000153
0.049	0.04926	0.048717
0.109	0.10965	0.109759
0.201	0.20152	0.201966
0.308	0.30832	0.309339
0.399	0.39974	0.399490
0.504	0.50448	0.502774
0.625	0.62532	0.624899
0.746	0.74699	0.746145
0.799	0.79936	0.799737
0.901	0.90115	0.896608

Table 4.7 shows agreement between the variance in Eq. (2.33), the theoretical variance from the chi-square distribution, and the sample variance. As expected, the noncentral chi-square variance is only a good estimate if p_i $i = 1, 2, \ldots, c$, and q_i $i = 1, 2, \ldots, c$ are "not very different" from each other. Table 3.4 in Section 3.2 compares the mean and variance of $\hat{k}^2$ when $|p_i - q_i|$ for all $i = 1, 2, \ldots, c$, is very small. In the later case, there is very good agreement between the theoretical and simulated variances.

Table 4.7 Comparison of Theory and Simulation Concerning the Variance of $\hat{k}^2$

$n = 5000, m = 5000$			Chi-Square	
Risk Categories 5		Average of 1000 Sample Estimates from Simulation	Central ($k^2 = 0$) Noncental ($k^2 > 0$)	Variance over 1000 Samples from Simulation
k^2	$V(\hat{k}^2)$	$\hat{V}(\hat{k}^2)$	Variance	Variance
	Eq. (2.33)			
0.000	0.000002	0.000000	0.000001	0.000001
0.049	0.000102	0.0000983	0.000079	0.000100
0.109	0.000242	0.0002379	0.000180	0.000235
0.201	0.000222	0.0002168	0.000323	0.000214
0.308	0.000772	0.0007572	0.000494	0.000712
0.399	0.001103	0.0010908	0.000810	0.001135
0.504	0.001387	0.0013753	0.001523	0.001717
0.625	0.001735	0.0017271	0.001000	0.001717
0.746	0.003731	0.0036365	0.001195	0.003778
0.799	0.001638	0.0016225	0.001279	0.001639
0.901	0.003647	0.0035972	0.001443	0.003575

4.2.2 Results for $\ln(\hat{k}^2 + 1)$ Regarding Mean and Variance with Five Categories of Risk

Table 4.8 shows good agreement among $E(\ln(\hat{k}^2 + 1))$ from Eq. (2.37), the mean of the theoretical distribution of $\ln(\hat{k}^2 + 1)$ from Eq. (3.7), and the average or mean from simulation under the null hypothesis $H_o : \ln(k^2 + 1) = 0$. The mean of the theoretical distribution only applies under the null hypothesis and is calculated as

$$E(\ln(\hat{k}^2 + 1)) = E(Z) \tag{4.1}$$

$$= \int_0^\infty z f(z) dz, \tag{4.2}$$

where $f(z)$ from Eq. (3.7) is given as

$$f(z) = \frac{(e^z - 1)^{\alpha-1} e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}} e^z dz, \quad z > 0.$$

Table 4.8 Comparison of Theory and Simulation Concerning the Mean of $\ln(\hat{k}^2 + 1)$

$n = 5000, m = 5000$				
Risk Categories = 5		Average from Simulation		Expected Value for Null Hypothesis
			Eq. (2.37)	Eq. (4.2)
k^2	$\ln(k^2 + 1)$		$E(\ln(\hat{k}^2 + 1))$	$E(\ln(\hat{k}^2 + 1))$
0.000	0.000000	0.001553	0.001552	0.001598
0.049	0.047893	0.049228	0.049186	–
0.109	0.103844	0.105343	0.105264	–
0.201	0.183389	0.185220	0.185167	–
0.308	0.268550	0.271029	0.270898	–
0.399	0.336091	0.337518	0.337375	–
0.504	0.408239	0.408814	0.408679	–
0.625	0.485508	0.486138	0.486014	–
0.746	0.557709	0.559835	0.559638	–
0.799	0.587231	0.588346	0.588268	–
0.901	0.642275	0.642385	0.642246	–

Table 4.9 presents the different measures of the variance of the data when $\ln(\hat{k}^2 + 1)$ was simulated for a sample size of $(m, n) = (5000, 5000)$ and five categories of risk. Results show good agreement between the variance over 1000 replications, $V(\ln(\hat{k}^2 + 1))$ from Eq. (2.39), the average of the estimates of $V(\ln(\hat{k}^2 + 1))$ from the 1000 replications, and the variance of the theoretical distribution of $\ln(\hat{k}^2 + 1)$ from Eq. (3.7) under the null hypothesis

$H_o : \ln\left(\hat{k}^2 + 1\right) = 0$. The variance of the theoretical distribution only applies under the null hypothesis and is calculated as

$$V\left(\ln\left(\hat{k}^2 + 1\right)\right) = V(Z)$$
$$= \int_0^\infty z^2 f(z)dz - (E(Z))^2. \tag{4.3}$$

Table 4.9 Comparison of Theory and Simulation Concerning the Variance of $\ln\left(\hat{k}^2+1\right)$

n = 5000 m = 5000					
Risk Categories=5		Variance over 1000 Samples from Simulation		Average of 1000 Sample Estimates from Simulation	Variance for Null Hypothesis $\ln\left(\hat{k}^2+1\right) = 0$
			Eq. (2.39)		Eq. (4.3)
k^2	$\ln(k^2+1)$		$V\left(\ln\left(\hat{k}^2+1\right)\right)$	$V\left(\ln\left(\hat{k}^2+1\right)\right)$	$V\left(\ln\left(\hat{k}^2+1\right)\right)$
0.000	0.000000	0.000001	0.000000	0.000002	0.00000127
0.049	0.047893	0.000090	0.000089	0.000092	–
0.109	0.103844	0.000189	0.000193	0.000195	–
0.201	0.183389	0.000147	0.000150	0.000153	–
0.308	0.268550	0.000413	0.000443	0.000448	–
0.399	0.336091	0.000577	0.000557	0.000559	–
0.504	0.408239	0.000565	0.000608	0.000610	–
0.625	0.485508	0.000647	0.000654	0.000653	–
0.746	0.557709	0.001219	0.001192	0.001204	–
0.799	0.587231	0.000503	0.000501	0.000503	–
0.901	0.642275	0.000984	0.000996	0.001000	–

4.2.3 Results Regarding Mean and Variance of $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$ and $\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)$

Table 4.10 shows good agreement between different measures of the mean of

$\sum_{i=1}^{2} \ln\left(\hat{k}_i^2 + 1\right)$ for a sample size of $(m,n) = (5000, 5000)$ and five categories of risk under the null hypothesis $H_o : \sum_{i=1}^{2} \ln\left(\hat{k}^2 + 1\right) = 0$. Parameters used for each of k_1^2 and k_2^2 are as follows.

For $k_1^2 = 0$,

$$\mathbf{q} = \mathbf{p} = (0.2, 0.2, 0.2, 0.2, 0.2),$$

and for $k_2^2 = 0$,

$$\mathbf{q} = \mathbf{p} = (0.5, 0.2, 0.15, 0.1, 0.05).$$

Parameters for $k_i^2 \neq 0$ are the same as those in Table 4.1. The sample average, the expected value, $E\left(\sum_{i=1}^{2} \ln\left(\hat{k}_i^2 + 1\right)\right)$, given by Eq. (2.40), and the mean of the theoretical distribution of $\sum_{i=1}^{2} \ln\left(\hat{k}_i^2 + 1\right)$ under the null hypothesis, given by Eq. (3.11), are compared in Table 4.10. The theoretical mean is calculated as

$$E\left(\sum_{i=1}^{2} \ln\left(\hat{k}_i^2 + 1\right)\right) = E(Z) = \int_0^\infty z g(z) dz \tag{4.4}$$

where $g(z)$ is given by Eq. (3.11) as

$$g(z) = \frac{(e^z - 1)^{\sum_{i=1}^{r} \alpha_i - 1} e^{-\frac{(e^z - 1)}{\beta'}}}{\Gamma\left(\sum_{i=1}^{r} \alpha_i\right)(\beta')^{\alpha}} e^z dz.$$

Results in Table 4.10 show good agreement between theory and simulation.

Table 4.10 Comparison of Theory and Simulation Concerning the Mean of $\sum_{i=1}^{2}\ln\left(\hat{k}^2+1\right)$

n = 5000 m = 5000					
Risk Categories 5			Average from Simulation		Expected Value for Null Hypothesis
				Eq. (2.40)	Eq. (4.4)
k_1^2	k_2^2	$\sum_{i=1}^{2}\ln(k_i^2-1)$		$E\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2-1\right)\right)$	$E\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right)$
0.000	0.000	0.000000	0.003192	0.003206	0.0031936
0.000	0.103844	0.103844	0.106981	0.106867	–
0.000	0.183389	0.183389	0.186858	0.186770	–
0.000	0.268550	0.268550	0.272668	0.272502	–
0.000	0.336091	0.336091	0.339157	0.338978	–
0.000	0.408239	0.408239	0.410453	0.410282	–
0.000	0.485508	0.485508	0.487777	0.487617	–
0.000	0.557709	0.557709	0.561474	0.561241	–
0.000	0.587231	0.587231	0.589985	0.589871	–
0.000	0.642275	0.642275	0.644024	0.6438497	–

Table 4.11 presents different measures of the variance of $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$ for a sample size of $(m,n)=(5000,5000)$ and five categories of risk under the null hypothesis $H_o:\sum_{i=1}^{2}\ln\left(\hat{k}^2+1\right)=0$. Results show good agreement among the variance from simulation, $V\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right)$ from Eq. (2.41), average of $\hat{V}\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right)$ over 1000 replications and variance of the theoretical distribution of $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$ from Eq. (3.11). The variance of the theoretical distribution under the null hypothesis is calculated as

$$V\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right) = V(Z)$$

$$= \int_0^{\infty} z^2 g(z)dz - (E(Z))^2. \qquad (4.5)$$

Table 4.11 Comparison of Theory and Simulation Concerning the Variance of $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$

$n = 5000$ $m = 5000$						
Risk Categories 5			Variance over 1000 Samples from Simulation	Average of 1000 Sample Estimates from Simulation		Variance for Null Hypothesis
$\ln\left(\hat{k}_1^2+1\right)$	$\ln\left(\hat{k}_2^2+1\right)$	$\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$		$\hat{V}\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right)$	$V\left(\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)\right)$	$V\left(\sum_{i=1}^{2}\ln\left(\hat{k}^2+1\right)\right)$
					Eq. (2.41)	Eq. (4.5)
0.000	0.000000	0.000000	0.000002	0.000005	0.000000	0.000002
0.000	0.1038449	0.103844	0.000190	0.000198	0.000193	–
0.000	0.183389	0.183389	0.000149	0.000155	0.00015	–
0.000	0.268550	0.268550	0.000417	0.000450	0.000443	–
0.000	0.336091	0.336091	0.000578	0.000561	0.000557	–
0.000	0.408239	0.408239	0.000565	0.000613	0.000608	–
0.000	0.485508	0.485508	0.000647	0.000656	0.000654	–
0.000	0.557709	0.557709	0.001219	0.001207	0.001192	–
0.000	0.587231	0.587231	0.000504	0.000506	0.000501	–
0.000	0.642275	0.642275	0.000986	0.001003	0.000996	–

Similar comparisons are made for $\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)$ under null hypothesis $H_o : \sum_{i=1}^{3}\ln\left(\hat{k}^2+1\right) = 0$. The parameters used for each of k_1^2, k_2^2 and k_3^2 are as follows. For $k_1^2 = 0$,

$$\mathbf{q} = \mathbf{p} = (0.2, 0.2, 0.2, 0.2, 0.2),$$

for $k_2^2 = 0$,

$$\mathbf{q} = \mathbf{p} = (0.5, 0.2, 0.15, 0.1, 0.05)$$

and for $k_3^2 = 0$,

$$\mathbf{q} = \mathbf{p} = (0.102, 0.192, 0.338, 0.315, 0.053).$$

Parameters in the case of k_i^2, $i = 1,2,3$, in which $k_i^2 \neq 0$ are the same as those in Table 4.1. Comparisons concerning the mean of the data when $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$ is simulated are shown in Table 4.12, and comparisons concerning the variance are shown in Table 4.13. Results show good agreement between theory and simulation.

Table 4.12 Comparison of Theory and Simulation Concerning the Mean of $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$

$n = 5000$ $m = 5000$						
Risk Categories 5				Average from Simulation		Expected Value for Null Hypothesis
					Eq. (2.40)	Eq. (4.4)
$\ln(k_1^2 + 1)$	$\ln(k_2^2 + 1)$	$\ln(k_3^2 + 1)$	$\sum_{i=1}^{3} \ln(k_i^2 + 1)$		$E\left(\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)\right)$	$E\left(\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)\right)$
0.000	0.000	0.000	0.000000	0.004778	0.004810	0.004786
0.000	0.000	0.103844	0.103844	0.108567	0.108471	–
0.000	0.000	0.183389	0.183389	0.188444	0.188374	–
0.000	0.000	0.268550	0.268550	0.274253	0.274105	–
0.000	0.000	0.336091	0.336091	0.340743	0.340582	–
0.000	0.000	0.408239	0.408239	0.412038	0.411886	–
0.000	0.000	0.485508	0.485508	0.489362	0.489221	–
0.000	0.000	0.587231	0.587231	0.591571	0.591475	–
0.000	0.000	0.642275	0.642275	0.645609	0.645453	–

Table 4.13 Comparison of Theory and Simulation Concerning the Variance of $\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)$

$n = 5000$ $m = 5000$							
Risk Categories 5				Variance over 1000 Samples from Simulation	Average of 1000 Sample Estimates from Simulation		Varince for Null Hypothesis
$\ln(k_1^2+1)$	$\ln(k_2^2+1)$	$\ln(k_3^2+1)$	$\sum_{i=1}^{3}\ln\left(\hat{k}_i^2-1\right)$		$V\left(\sum_{i=1}^{3}\ln\left(\hat{k}_i^2-1\right)\right)$	$V\left(\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)\right)$	$V\left(\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)\right)$
						Eq. (2.41)	Eq. (4.5)
0.000	0.000	0.000000	0.000000	0.000004	0.000008	0.000000	0.000002
0.000	0.000	0.1038449	0.103844	0.000192	0.000200	0.000193	–
0.000	0.000	0.183389	0.183389	0.000150	0.000158	0.00015	–
0.000	0.000	0.268550	0.268550	0.000420	0.000453	0.000443	–
0.000	0.000	0.336091	0.336091	0.000580	0.000564	0.000557	–
0.000	0.000	0.408239	0.408239	0.000569	0.000615	0.000608	–
0.000	0.000	0.485508	0.485508	0.000650	0.000659	0.000654	–
0.000	0.000	0.587231	0.587231	0.000504	0.000509	0.000501	–
0.000	0.000	0.642275	0.642275	0.000983	0.001005	0.000996	–

The results for the null hypothesis $H_o : \sum_{i=1}^{2}\ln\left(\hat{k}^2+1\right) = \ln S$ were presented in Table 3.11 and the results for the null hypothesis $H_o : \sum_{i=1}^{3}\ln\left(\hat{k}^2+1\right) = \ln S$ were presented in Table 3.12

4.2.4 Results Regarding Mean and Variance of $\hat{k}_1^2 - \hat{k}_2^2$

Table 4.14 presents different measures of the mean of $\hat{k}_1^2 - \hat{k}_2^2$ for sample size of $(m,n) = (5000, 5000)$ and five categories of risk. The parameters used for each of the k_1^2 and k_2^2 are as follows. For $k_1^2 = k_2^2 = 0.625$,

$$\mathbf{q}=(0.50, 0.20, 0.15, 0.10, 0.05)$$
$$\mathbf{p}=(0.20, 0.20, 0.20, 0.20, 0.20).$$

Although, the random variables were generated from the same population, the simulation procedure used different seeds to randomly generate the samples; therefore, they are independent of one another. The parameters for the simulated k_i^2, $i = 1, 2$, in which

$k_i^2 \neq 0.625$ are the same as those in Table 4.1.

The sample average of $\hat{k}_1^2 - \hat{k}_2^2$ and the $E(\hat{k}_1^2 - \hat{k}_2^2) = E(\hat{k}_1^2) - E(\hat{k}_2^2)$, from Eq. (2.27) are compared in Table 4.14 as well as the simulated and theoretical bias for $\hat{k}_1^2$ and k_2^2. As can be seen, there is good agreement between the two measures of the mean.

Table 4.14 Comparison of Theory and Simulation Concerning the Mean of $\hat{k}_1^2$-$\hat{k}_2^2$

n = 5000 m = 5000							
Risk Categories 5		Average of Bias of $\hat{k}_1^2$ over 1000 Replications	Average of Bias of $\hat{k}_2^2$ over 1000 Replications	Average from Simulation	Theoritical Bias of k_1^2	Theoritical Bias of k_2^2	
k_1^2	k_2^2	$bias_{\hat{k}_1^2}$	$bias_{\hat{k}_2^2}$		$bias_{k_1^2}$	$bias_{k_2^2}$	$E(\hat{k}_1^2 - \hat{k}_1^2)$
					Eq. (2.27a)	Eq. (2.27a)	Eq. (2.27)
0.625	0.000	0.001993	0.001604	0.626694	0.001987	0.001600	0.6254
0.625	0.109	0.001993	0.001661	0.517053	0.001987	0.001658	0.5163
0.625	0.201	0.001993	0.001849	0.424678	0.001987	0.001843	0.4241
0.625	0.308	0.001993	0.002511	0.316666	0.001987	0.002498	0.3165
0.625	0.399	0.001993	0.002665	0.226380	0.001987	0.002656	0.225
0.625	0.504	0.001993	0.002989	0.122792	0.001987	0.002984	0.120
0.625	0.625	0.001993	0.001989	0.001699	0.001987	0.001987	0.000
0.746	0.625	0.005678	0.001993	0.123200	0.005639	0.001987	0.1246
0.799	0.625	0.002085	0.001993	0.173211	0.002080	0.001987	0.1740
0.901	0.625	0.005736	0.001993	0.273696	0.005721	0.001987	0.2797

4.2.5 Results Regarding 2, 4, 6, and 8 Categories of Risk

The average of $\hat{k}^2$ over 1000 replications was calculated and compared to its parameter value $k^2 = \sum_{i=1}^{c} \frac{q_i^2}{p_i} - 1$, to the expected value defined in Section 2.4, Eq. (2.27), that is,

$$E(\hat{k}^2) = \sum_{i=1}^{c} \left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{m(n)p_i^2} + \frac{q_i^2}{p_i} + \frac{q_i^2(1-p_i)}{(n)(p_i)^2} \right) - 1$$

and to the theoretical expected value based on the associated $Gamma\left(\frac{c-1}{2}, \left(\frac{m-n}{mn}\beta\right)\right)$ (or equivalently, $\frac{m-n}{mn}\chi^2(c-1)$ distribution if the parameter value for k^2 is zero, or to the

noncentral chi-square distribution $\frac{m+n}{mn}\chi^2(c-1, \frac{mn}{m+n}\sum\frac{(q_i-p_i)^2}{p_i})$ if $k^2 > 0$). The average of the 1000 replications of $bias_{\hat{k}^2}$ is compared to the parameter value of

$$bias_{k^2} = \sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{m(n+e)p_i^2} + \frac{q_i^2(1-p_i)}{(n+e)(p_i)^2}\right).$$

The average of $V(\hat{\hat{k}}^2)$ over the 1000 replications is compared to its corresponding parameter value of $V(\hat{k}^2)$ defined in Section 2.6, Eq. (2.33). The average of the estimate, $V(\hat{\hat{k}}^2)$, is also compared to its corresponding sample variance of $\hat{\sigma}^2_{\hat{k}^2}$. Additionally, the average of $V(\hat{\hat{k}}^2)$ is compared to the theoretical variance based on the associated $Gamma\left(\frac{c-1}{2}, (\frac{m+n}{mn}\beta)\right)$ (or equivalently, $\frac{m+n}{mn}\chi^2(c-1)$ distribution if the parameter value for k^2 is zero, or to the noncentral chi-square distribution $\frac{m+n}{mn}\chi^2(c-1, \frac{mn}{m+n}\sum\frac{(q_i-p_i)^2}{p_i})$ if $k^2 > 0$). It is seen from Tables 4.15 to 4.22 that there is good agreement between theory and simulation with regard to the mean and variance of $\hat{k}^2$. The variance under the alternative hypothesis differs from the noncentral chi-square variance because in the populations simulated, the p_i and q_i, $i = 1, 2, \ldots, c$, are too far apart to be distributed as noncentral chi-square; rather, they are better modeled as asymptotically normal.

Table 4.15 Comparison of Theory and Simulation Concerning the Mean of $\hat{k}^2$ with Two Categories

$n = 5000$ $m = 5000$					
	Average of Bias over 1000 Replications	Bias Eq. (2.27a)		Average from Simulation	Chi-Square Mean Central ($k^2 = 0$) Noncental ($k^2 > 0$)
k^2	$bias_{\hat{k}^2}$	$bias_{k^2}$	$E(\hat{k}^2)$		
0.000	0.000400	0.000400	0.000400	0.000375	0.000400
0.048	0.000392	0.000392	0.048835	0.048785	0.048843
0.100	0.000481	0.000481	0.100585	0.099740	0.100504
0.495	0.000589	0.000589	0.496527	0.495018	0.496337

Table 4.16 Comparison of Theory and Simulation Concerning the Variance of $\hat{k}^2$ with Two Categories

$n = 5000$ $m = 5000$				
		Average of 1000 Sample Estimates from Simulation	Chi-Square Variance Central ($k^2 = 0$) Noncental ($k^2 > 0$)	Variance over 1000 Samples from Simulation
	Eq. (2.33)			
k^2	$V(\hat{k}^2)$	$V(\hat{\hat{k}}^2)$		
0.000	0.000000	0.000001	0.000000	0.000000
0.048	0.000075	0.000070	0.000077	0.000075
0.100	0.000173	0.000173	0.000160	0.000173
0.495	0.000833	0.000814	0.000793	0.000832

Table 4.17 Comparison of Theory and Simulation Concerning the Mean of $\hat{k}^2$ with Four Categories

n = 5000 m = 5000					
	Average of Bias over 1000 Replications	Bias Eq. (2.27a)		Average from Simulation	Chi-Square Mean Central ($k^2 = 0$) Noncental ($k^2 > 0$)
k^2	$bias_{\hat{k}^2}$	$bias_{k^2}$	$E(\hat{k}^2)$		
0.000	0.001200	0.001200	0.001200	0.001215	0.001200
0.050	0.001220	0.001220	0.051476	0.051101	0.051456
0.102	0.001346	0.001347	0.103015	0.103001	0.102867
0.502	0.001683	0.001683	0.504617	0.504485	0.504133

Table 4.18 Comparison of Theory and Simulation Concerning the Variance of $\hat{k}^2$ with Four Categories

n = 5000 m = 5000				
		Average of 1000 Sample Estimates from Simulation	Chi-Square Variance Central ($k^2 = 0$) Noncental ($k^2 > 0$)	Variance over 1000 Samples from Simulation
k^2	$V(\hat{k}^2)$	$V(\hat{\hat{k}}^2)$		
	Eq. (2.33)			
0.000	0.000000	0.000002	0.000001	0.000001
0.048	0.000079	0.000080	0.000081	0.000072
0.100	0.000187	0.000178	0.000177	0.000175
0.495	0.001511	0.001519	0.000805	0.001444

Table 4.19 Comparison of Theory and Simulation Concerning the Mean of $\hat{k}^2$ with Six Categories

$n = 5000$ $m = 5000$					
	Average of Bias over 1000 Replications	Bias Eq. (2.27a)		Average from Simulation	Chi-Square Mean Central ($k^2 = 0$) Noncental ($k^2 > 0$)
k^2	$bias_{\hat{k}^2}$	$bias_{k^2}$	$E(\hat{k}^2)$	average	
0.000	0.002001	0.002001	0.002001	0.001969	0.002000
0.051	0.002206	0.002207	0.052706	0.051302	0.052500
0.101	0.001936	0.001936	0.102436	0.102666	0.102500
0.500	0.003255	0.003256	0.502756	0.502531	0.501500

Table 4.20 Comparison of Theory and Simulation Concerning the Variance of $\hat{k}^2$ with Six Categories

$n = 5000$ $m = 5000$				
		Average of 1000 Sample Estimates from Simulation	Chi-Square Variance Central ($k^2 = 0$) Noncental ($k^2 > 0$)	Variance over 1000 Samples from Simulation
	Eq. (2.33)			
k^2	$V(\hat{k}^2)$	$V(\hat{\hat{k}}^2)$		
0.000	0.000000	0.000003	0.000001	0.000001
0.048	0.000086	0.000094	0.000082	0.000092
0.100	0.000161	0.000166	0.000162	0.000166
0.495	0.000997	0.000954	0.000800	0.001012

Table 4.21 Comparison of Theory and Simulation Concerning the Mean of $\hat{k}^2$ with Eight Categories

$n = 5000$ $m = 5000$					
	Average of Bias over 1000 Replications	Bias Eq. (2.27a)		Average from Simulation	Chi-Square Mean Central ($k^2 = 0$) Noncental ($k^2 > 0$)
	$bias_{\hat{k}^2}$	$bias_{k^2}$	$E(\hat{k}^2)$	average	
k^2					
0.000	0.002801	0.002802	0.002802	0.002823	0.002800
0.050	0.003028	0.003028	0.052543	0.052164	0.052315
0.104	0.003529	0.003530	0.107655	0.106804	0.106925
0.502	0.004621	0.004623	0.506998	0.508477	0.505175

Table 4.22 Comparison of Theory and Simulation Concerning the Variance of $\hat{k}^2$ with Eight Categories

$n = 5000$ $m = 5000$				
		Average of 1000 Sample Estimates from Simulation	Chi-Square Variance Central ($k^2 = 0$) Noncental ($k^2 > 0$)	Variance over 1000 Samples from Simulation
	Eq. (2.33)			
k^2	$V(\hat{k}^2)$	$V(\hat{k}^2)$		
0.000	0.000000	0.000005	0.000002	0.000002
0.050	0.000091	0.000097	0.000081	0.000092
0.104	0.000239	0.000250	0.000168	0.000256
0.502	0.001064	0.001095	0.000806	0.001120

4.2.6 Simulation Results Regarding the Power of the Two Test Statistics $\hat{k}^2$ and χ^2 for Five Categories of Risk

The following tables present the powers of the test statistics, $\hat{k}^2 = \sum_{i=1}^{c} \frac{\hat{q}_i^2}{\hat{p}_i} - 1$ and χ^2, for the null hypothesis ($k^2 = 0$, or the factor is not a risk) vs. the alternative hypothesis ($k^2 \neq 0$) from simulation at the 0.01, 0.025, 0.050, and 0.100 α levels. The critical values for each α level were calculated for $\hat{k}^2$ from the $Gamma\left(\frac{c-1}{2}, \frac{2(m-n)}{mn}\right)$ distribution and for χ^2 from the chi-square distribution with $c-1$ degrees of freedom. As can be seen, the power is quite high for sample size 200 and over. For small k^2 values of 0.049 the power becomes good for sample size over 400. This is similar to the power found in the χ^2 test given in Eq. (1.12). As expected, the power of the test is better as the sample size gets larger. The power of the test for $\hat{k}^2$ is slightly better than that for the χ^2 statistic.

Table 4.23 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 50$, $n = 50$

$n = 50, m = 50$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.026	0.051	0.074	0.144	.009	.026	.047	.103
0.049	0.102	0.148	0.199	0.277	0.015	0.035	0.082	.154
0.109	0.127	0.186	0.246	0.339	0.062	0.125	0.194	.295
0.201	0.222	0.336	0.451	0.573	0.234	0.386	0.515	.649
0.308	0.364	0.458	0.554	0.660	0.203	0.320	.0436	.571
0.399	0.519	0.608	0.706	0.790	0.324	0.470	0.594	.711
0.504	0.641	0.731	0.802	0.873	0.456	0.606	0.722	.828
0.625	0.684	0.792	0.864	0.929	0.669	0.791	0.871	.938
0.746	0.800	0.867	0.904	0.945	0.604	0.729	0.828	.911
0.799	0.876	0.928	0.968	0.987	0.931	0.972	0.987	.997
0.901	0.904	0.946	0.961	0.974	0.851	0.915	0.945	.975

Table 4.24 Power of the Two Test Statistics $\hat{k}^2$ Eq.(2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 100$, $n = 100$

$n = 100, m = 100$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.015	0.027	0.051	0.099	0.011	0.021	0.033	0.085
0.049	0.130	0.191	0.276	0.367	0.042	0.086	0.172	0.274
0.109	0.240	0.330	0.406	0.532	0.180	0.269	0.372	0.496
0.201	0.533	0.685	0.786	0.881	0.678	0.782	0.856	0.924
0.308	0.715	0.793	0.852	0.905	0.605	0.725	0.809	0.882
0.399	0.850	0.896	0.935	0.956	0.778	0.853	0.913	0.951
0.504	0.929	0.960	0.982	0.990	0.888	0.938	0.970	0.985
0.625	0.978	0.989	0.997	0.999	0.973	0.993	0.997	0.999
0.746	0.981	0.993	0.997	0.999	0.964	0.985	0.995	0.998
0.799	0.999	0.999	0.999	1.000	0.999	1.000	1.000	1.000
0.901	0.999	1.000	1.000	1.000	0.999	1.000	1.000	1.000

Table 4.25 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 200$, $n = 200$

$n = 200, m = 200$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.015	0.037	0.059	0.098	0.011	0.027	0.052	0.091
0.049	0.241	0.342	0.440	0.539	0.149	0.239	0.333	0.473
0.109	0.510	0.633	0.728	0.814	0.449	0.587	0.694	0.798
0.201	0.945	0.983	0.992	0.999	0.983	0.992	0.997	1.000
0.308	0.971	0.990	0.994	0.998	0.962	0.981	0.992	1.000
0.399	0.996	0.996	0.999	1.000	0.992	0.996	0.999	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.26 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 300$, $n = 300$

$n = 300, m = 300$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.014	0.028	0.052	0.103	0.009	0.026	0.046	0.093
0.049	0.352	0.472	0.565	0.659	0.241	0.393	0.498	0.618
0.109	0.747	0.826	0.881	0.928	0.699	0.802	0.872	0.925
0.201	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.27 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 400$, $n = 400$

$n = 400, m = 400$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.010	0.024	0.048	0.091	0.008	0.021	0.038	0.088
0.049	0.490	0.611	0.707	0.801	0.408	0.529	0.646	0.767
0.109	0.903	0.941	0.961	0.981	0.882	0.936	0.958	0.979
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.28 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fft Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 500$, $n = 500$

$n = 500, m = 500$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.012	0.026	0.050	0.094	0.006	0.022	0.046	0.096
0.049	0.622	0.739	0.802	0.876	0.552	0.682	0.784	0.970
0.109	0.953	0.971	0.979	0.991	0.941	0.966	0.976	0.999
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.29 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 600$, $n = 600$

$n = 600, m = 600$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.013	0.027	0.049	0.131	0.120	0.240	0.050	0.094
0.049	0.729	0.825	0.875	0.939	0.673	0.785	0.854	0.912
0.109	0.988	0.993	0.996	0.999	0.986	0.992	0.995	0.999
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.30 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 800$, $n = 800$

$n = 800, m = 800$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.011	0.027	0.045	0.093	0.010	0.025	0.049	0.096
0.049	0.844	0.913	0.951	0.973	0.806	0.887	0.936	0.970
0.109	0.997	0.998	0.998	0.999	0.995	0.998	0.998	0.999
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.31 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 1000$, $n = 1000$

$n = 1000, m = 1000$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.015	0.037	0.068	0.116	0.015	0.034	0.062	0.113
0.049	0.945	0.974	0.990	0.993	0.927	0.969	0.985	0.993
0.109	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.32 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 3000$, $n = 3000$

$n = 3000, m = 3000$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.008	0.020	0.035	0.086	0.007	0.021	0.036	0.087
0.049	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.109	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.33 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 5000$, $n = 5000$

$n = 5000, m = 5000$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.008	0.016	0.035	0.080	0.008	0.017	0.032	0.081
0.049	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.109	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.34 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 7000$, $n = 7000$

$n = 7000, m = 7000$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.012	0.031	0.056	0.105	0.010	0.030	0.058	0.108
0.049	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.109	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.35 Power of the Two Test Statistics $\hat{k}^2$ Eq. (2.21a), and the χ^2 Goodness of Fit Test, Eq. (1.12), for Five Categories of Risk for Sample Sizes of $m = 9000$, $n = 9000$

$n = 9000, m = 9000$	Test Statistics							
	$\hat{k}^2$				χ^2			
k^2	**0.010**	**0.025**	**0.050**	**0.100**	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.011	0.026	0.046	0.095	0.012	0.027	0.045	0.095
0.049	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.109	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.201	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.308	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.399	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.504	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.625	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.746	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.799	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.901	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

4.2.7 Results Regarding the Power at the $\alpha = 0.05$ Level of the Test for $\ln(\hat{k}^2 + 1)$, $\sum_{i=1}^{2} \ln(\hat{k}_i^2 + 1)$, $\sum_{i=1}^{3} \ln(\hat{k}_i^2 + 1)$, and $\hat{k}_1^2 - \hat{k}_2^2$ for Five Categories of Risk

The power of the test statistic, $\ln(\hat{k}^2 + 1)$, for the null hypothesis $\ln(\hat{k}^2 + 1) = 0$, at the $\alpha = 0.05$ level was calculated for the sample sizes given in Tables 4.36 through 4.40. The parameters used to simulate the k^2 are the same as those specified in Section 4.1, and the parameters used to simulate the test statistic are indicated in the tables. The critical value of the test is determined by the distribution of the test statistic. For $\ln(\hat{k}^2 + 1)$, the distribution was derived in Section 3.3, and given by Eq. (3.7), that is,

$$f(z) = \frac{(e^z - 1)^{\alpha-1} e^{-\frac{(e^z - 1)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}} e^z dz, \quad z > 0.$$

The critical value Z_α for this test at the $\alpha = 0.05$ level is found by integrating over the distribution from zero to Z_α, where Z_α is the upper limit of the integral that gives an area under the curve of $1 - \alpha$,

$$\int_0^{Z_\alpha} f(Z)dZ = 0.95.$$

For $\sum_{i=1}^{2} \ln(\hat{k}_i^2 + 1)$ and $\sum_{i=1}^{2} \ln(\hat{k}_i^2 + 1)$, under the null hypothesis, $\sum_{i=1}^{2} \ln(\hat{k}_i^2 + 1) = 0$, the distribution is given by Eq. (3.11),

$$f(z) = \frac{(e^z - 1)^{\sum_{i=1}^{r} \alpha_i - 1} e^{-\frac{(e^z - 1)}{\beta'}}}{\Gamma\left(\sum_{i=1}^{r} \alpha_i\right)(\beta')^{\alpha}} e^z dz.$$

The critical value is again found by determining the value of the upper limit that makes the area under the curve equal to $1 - \alpha$. The tables show that the powers of the test statistics

$\ln(\hat{k}^2+1)$ and $\sum_{i=1}^{2}\ln(\hat{k}^2+1)$ are quite high for sample size 500 and larger.

Table 4.36 Power for $\ln(\hat{k}^2+1)$ and $\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$ at a Sample Size of $(m,n) = (100, 100)$ for Five Risk Categories

$n = 100, m = 100$		Critical Value	
$\alpha = 0.05$		0.17375	0.48261
k_1^2	k_2^2	$\ln(\hat{k}_1^2+1)$	$\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$
0.000	0.000	0.052	0.080
0.109	0.000	0.407	0.388
0.201	0.000	0.786	0.686
0.308	0.000	0.853	0.795
0.399	0.000	0.935	0.900
0.504	0.000	0.982	0.965
0.625	0.000	0.997	0.985
0.746	0.000	0.997	0.993
0.799	0.000	0.999	1.000
0.901	0.000	1.000	1.000

Table 4.37 Power for $\ln\left(\hat{k}^2+1\right)$ and $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$ at a Sample Size of $(m,n)=(500,500)$ for Five Risk Categories

$n=500, m=500$		Critical Value	
$\alpha=0.05$		0.03725	0.06019
k_1^2	k_2^2	$\ln\left(\hat{k}_1^2+1\right)$	$\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$
0.000	0.000	0.050	0.063
0.109	0.000	0.979	0.965
0.201	0.000	1.000	1.000
0.308	0.000	1.000	1.000
0.399	0.000	1.000	1.000
0.504	0.000	1.000	1.000
0.625	0.000	1.000	1.000
0.746	0.000	0.997	1.000
0.799	0.000	1.000	1.000
0.901	0.000	1.000	1.000

Table 4.38 Power for $\ln\left(\hat{k}^2+1\right)$ and $\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$ at a Sample Size of $(m,n) = (1000,1000)$ for Five Risk Categories

$n = 1000, m = 1000$		Critical Value	
$\alpha = 0.05$		0.018799	0.030545
k_1^2	k_2^2	$\ln\left(\hat{k}_1^2+1\right)$	$\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)$
0.000	0.000	0.060	0.062
0.109	0.000	1.000	1.000
0.201	0.000	1.000	1.000
0.308	0.000	1.000	1.000
0.399	0.000	1.000	1.000
0.504	0.000	1.000	1.000
0.625	0.000	1.000	1.000
0.746	0.000	0.997	1.000
0.799	0.000	1.000	1.000
0.901	0.000	1.000	1.000

Table 4.39 Power for $\ln\left(\hat{k}^2 + 1\right)$ and $\sum_{i=1}^{2} \ln\left(\hat{k}_i^2 + 1\right)$ at a Sample Size of $(m, n) = (3000, 3000)$ for Five Risk Categories

$n = 3000, m = 3000$		Critical Value	
$\alpha = 0.05$		0.006306	0.010285
k_1^2	k_2^2	$\ln\left(\hat{k}_1^2 + 1\right)$	$\sum_{i=1}^{2} \ln\left(\hat{k}_i^2 + 1\right)$
0.000	0.000	0.045	0.048
0.109	0.000	1.000	1.000
0.201	0.000	1.000	1.000
0.308	0.000	1.000	1.000
0.399	0.000	1.000	1.000
0.504	0.000	1.000	1.000
0.625	0.000	1.000	1.000
0.746	0.000	0.997	1.000
0.799	0.000	1.000	1.000
0.901	0.000	1.000	1.000

Table 4.40 Power for $\ln(\hat{k}^2+1)$ and $\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$ at a Sample Size of $(m,n) = (5000,5000)$ for Five Risk Categories

$n = 5000, m = 5000$		Critical Value	
$\alpha = 0.05$		0.0037885	0.0061845
k_1^2	k_2^2	$\ln(\hat{k}_1^2+1)$	$\sum_{i=1}^{2}\ln(\hat{k}_i^2+1)$
0.000	0.000	0.045	0.047
0.109	0.000	1.000	1.000
0.201	0.000	1.000	1.000
0.308	0.000	1.000	1.000
0.399	0.000	1.000	1.000
0.504	0.000	1.000	1.000
0.625	0.000	1.000	1.000
0.746	0.000	0.997	1.000
0.799	0.000	1.000	1.000
0.901	0.000	1.000	1.000

Tables 4.41 to 4.45 show the power of the test statistic $\sum_{i=1}^{3}\ln(\hat{k}_i^2+1)$ under the null hypothesis $\sum_{i=1}^{3}\ln(\hat{k}_i^2+1) = 0$, for different sample sizes. The parameters used in simulating the test statistic are indicated in the tables. As is seen from these tables, the power is high for sample size 500 or larger.

Table 4.41 Power for $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$ at a Sample Size of $(m,n) = (100, 100)$ for Five Risk Categories

$n = 100, m = 100$			Critical Value
$\alpha = 0.05$			0.61035
k_1^2	k_2^2	k_3^2	$\ln \sum_{i=1}^{3} \left(\hat{k}_i^2 + 1\right)$
0.000	0.000	0.000	0.085
0.109	0.000	0.000	0.397
0.201	0.000	0.000	0.651
0.308	0.000	0.000	0.771
0.399	0.000	0.000	0.870
0.504	0.000	0.000	0.952
0.625	0.000	0.000	0.979
0.799	0.000	0.000	1.000
0.901	0.000	0.000	1.000

Table 4.42 Power for $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$ at a Sample Size of $(m,n) = (500, 500)$ for Five Risk Categories

$n = 500, m = 500$			Critical Value
$\alpha = 0.05$			0.080760
k_1^2	k_2^2	k_3^2	$\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$
0.000	0.000	0.000	0.072
0.109	0.000	0.000	0.949
0.201	0.000	0.000	1.000
0.308	0.000	0.000	1.000
0.399	0.000	0.000	1.000
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.901	0.000	0.000	1.000

Table 4.43 Power for $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$ at a Sample Size of $(m,n) = (1000, 1000)$ for Five Risk Categories

$n = 1000, m = 1000$			Critical Value
$\alpha = 0.05$			0.041192
k_1^2	k_2^2	k_3^2	$\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$
0.000	0.000	0.000	0.072
0.109	0.000	0.000	1.000
0.201	0.000	0.000	1.000
0.308	0.000	0.000	1.000
0.399	0.000	0.000	1.000
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.901	0.000	0.000	1.000

Table 4.44 Power for $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$ at a Sample Size of $(m,n) = (3000, 3000)$ for Five Risk Categories

$n = 3000, m = 3000$			Critical Value
$\alpha = 0.05$			0.013921
k_1^2	k_2^2	k_3^2	$\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$
0.000	0.000	0.000	0.050
0.109	0.000	0.000	1.000
0.201	0.000	0.000	1.000
0.308	0.000	0.000	1.000
0.399	0.000	0.000	1.000
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.901	0.000	0.000	1.000

Table 4.45 Power for $\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)$ at a Sample Size of $(m,n) = (5000, 5000)$ for Five Risk Categories

$n = 5000, m = 5000$			Critical Value
$\alpha = 0.05$			0.008376
k_1^2	k_2^2	k_3^2	$\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)$
0.000	0.000	0.000	0.052
0.109	0.000	0.000	1.000
0.201	0.000	0.000	1.000
0.308	0.000	0.000	1.000
0.399	0.000	0.000	1.000
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.901	0.000	0.000	1.000

Tables 4.46 through 4.69 give the power of the test statistics, $\frac{\ln\left(\hat{k}_i^2+1\right)-\ln S-\frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}$, $\frac{\sum_{i=1}^{2}\ln\left(\hat{k}_i^2+1\right)-\ln S-\sum_{i=1}^{2}\frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{2}\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}$, and $\frac{\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)-\ln S-\sum_{i=1}^{3}\frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3}\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}$, derived in Section 3.3 for testing

$$H_o : \sum_{i=1}^{r}\ln(k_i^2+1) = \ln S$$

$$vs.$$

$$H_a : \sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right) \neq \ln S,$$

where $r = 1, 2,$ or 3. From simulation, the distribution for each of the three test statistics was shown to be approximately standard normal. As expected, this normal approximation becomes

better for large sample sizes. The power for the test statistics becomes good for all alternative hypotheses when the sample size is $m = n = 1000$ or greater. This may be seen from the tables below. The parameters used for each null hypothesis are given in the tables.

Table 4.46 Power for the Null Hypothesis $\ln(k^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (200, 200)$ and Five Categories of Risk

$k^2 = 0.625$		
$n = 200$ $m = 200$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k^2		
0.000	1.000	0.000
0.109	0.937	0.000
0.201	0.882	0.000
0.308	0.380	0.002
0.399	0.270	0.002
0.504	0.100	0.014
0.625	0.030	0.045
0.799	0.002	0.207
0.900	0.000	0.270

Table 4.47 Power for the Null Hypothesis $\ln(k^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (400, 400)$ and Five Categories of Risk

$k^2 = 0.625$		
$n = 400$ $m = 400$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k^2		
0.000	1.000	0.000
0.109	0.999	0.000
0.201	0.996	0.000
0.308	0.685	0.000
0.399	0.369	0.000
0.504	0.141	0.000
0.625	0.028	0.033
0.799	0.001	0.313
0.900	0.001	0.396

Table 4.48 Power for the Null Hypothesis $\ln(k^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (500, 500)$ and Five Categories of Risk

$k^2 = 0.625$		
$n = 500$ $m = 500$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k^2		
0.000	1.000	0.000
0.109	0.998	0.000
0.201	0.999	0.000
0.308	0.783	0.000
0.399	0.445	0.000
0.504	0.132	0.004
0.625	0.030	0.036
0.799	0.000	0.372
0.900	0.000	0.443

Table 4.49 Power for the Null Hypothesis $\ln(k^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (1000, 1000)$ and Five Categories of Risk

$k^2 = 0.625$		
$n = 1000$ $m = 1000$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k^2		
0.000	1.000	0.000
0.109	1.000	0.000
0.201	1.000	0.000
0.308	0.971	0.000
0.399	0.735	0.000
0.504	0.281	0.001
0.625	0.027	0.030
0.799	0.000	0.590
0.900	0.000	0.713

Table 4.50 Power for the Null Hypothesis $\ln(k^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (3000, 3000)$ and Five Categories of Risk

$k^2 = 0.625$		
$n = 3000$ $m = 3000$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2		
0.000	1.000	0.000
0.109	1.000	0.000
0.201	1.000	0.000
0.308	1.000	0.000
0.399	0.997	0.000
0.504	0.614	0.000
0.625	0.022	0.030
0.799	0.000	0.968
0.900	0.000	0.992

Table 4.51 Power for the Null Hypothesis $\ln(k^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (7000, 7000)$ and Five Categories of Risk

$k^2 = 0.625$		
$n = 7000$ $m = 7000$	$\alpha = 0.05$	
	$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2		
0.000	1.000	0.000
0.109	1.000	0.000
0.201	1.000	0.000
0.308	1.000	0.000
0.399	1.000	0.000
0.504	0.950	0.000
0.625	0.023	0.027
0.799	0.000	1.000
0.900	0.000	1.000

Table 4.52 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (200, 200)$ and Five Categories of Risk

$k_1^2 = 0.799, k_2^2 = 0.625$			
$n = 200, m = 200$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	0.922	0.000
0.109	0.625	0.744	0.000
0.201	0.625	0.609	0.000
0.308	0.625	0.337	0.000
0.399	0.625	0.203	0.001
0.504	0.625	0.131	0.005
0.625	0.625	0.071	0.010
0.799	0.625	0.019	0.040
0.900	0.625	0.100	0.049

Table 4.53 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (400, 400)$ and Five Categories of Risk

$k_1^2 = 0.799, k_2^2 = 0.625$			
$n = 400, m = 400$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	0.998	0.000
0.109	0.625	0.968	0.000
0.201	0.625	0.920	0.000
0.308	0.625	0.669	0.000
0.399	0.625	0.430	0.000
0.504	0.625	0.251	0.000
0.625	0.625	0.112	0.001
0.799	0.625	0.019	0.032
0.900	0.625	0.008	0.068

Table 4.54 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (500, 500)$ and Five Categories of Risk

$k_1^2 = 0.799, k_2^2 = 0.625$			
$n = 500, m = 500$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	0.987	0.000
0.201	0.625	0.966	0.000
0.308	0.625	0.754	0.000
0.399	0.625	0.524	0.000
0.504	0.625	0.278	0.000
0.625	0.625	0.113	0.002
0.799	0.625	0.022	0.032
0.900	0.625	0.006	0.065

Table 4.55 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (1000, 1000)$ and Five Categories of Risk

$k_1^2 = 0.799, k_2^2 = 0.625$			
$n = 1000, m = 1000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	0.967	0.000
0.399	0.625	0.827	0.000
0.504	0.625	0.533	0.000
0.625	0.625	0.204	0.001
0.799	0.625	0.019	0.037
0.900	0.625	0.007	0.098

Table 4.56 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (3000, 3000)$ and Five Categories of Risk

$k_1^2 = 0.799, k_2^2 = 0.625$			
$n = 3000, m = 3000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	1.000	0.000
0.399	0.625	0.999	0.000
0.504	0.625	0.954	0.000
0.625	0.625	0.559	0.000
0.799	0.625	0.021	0.025
0.900	0.625	0.000	0.200

Table 4.57 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (7000, 7000)$ and Five Categories of Risk

$k_1^2 = 0.799, k_2^2 = 0.625$			
$n = 7000, m = 7000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	1.000	0.000
0.399	0.625	1.000	0.000
0.504	0.625	1.000	0.000
0.625	0.625	0.896	0.000
0.799	0.625	0.027	0.025
0.900	0.625	0.000	0.355

Table 4.58 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (1000, 1000)$ and Five Categories of Risk

$k_1^2 = 0.000, k_2^2 = 0.625$			
$n = 1000, m = 1000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	0.013	0.036
0.109	0.625	0.000	0.412
0.201	0.625	0.000	0.897
0.308	0.625	0.000	0.985
0.399	0.625	0.000	1.000
0.504	0.625	0.000	1.000
0.625	0.625	0.000	1.000
0.799	0.625	0.000	1.000
0.900	0.625	0.000	1.000

Table 4.59 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (3000,3000)$ and Five Categories of Risk

$k_1^2 = 0.000, k_2^2 = 0.625$			
$n = 3000, m = 3000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	0.020	0.033
0.109	0.625	0.000	0.840
0.201	0.625	0.000	1.000
0.308	0.625	0.000	0.985
0.399	0.625	0.000	1.000
0.504	0.625	0.000	1.000
0.625	0.625	0.000	1.000
0.799	0.625	0.000	1.000
0.900	0.625	0.000	1.000

Table 4.60 Power for the Null Hypothesis $\sum_{i=1}^{2} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (7000,7000)$ and Five Categories of Risk

$k_1^2 = 0.000, k_2^2 = 0.625$			
$n = 7000, m = 7000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	0.017	0.026
0.109	0.625	0.000	0.969
0.201	0.625	0.000	1.000
0.308	0.625	0.000	1.000
0.399	0.625	0.000	1.000
0.504	0.625	0.000	1.000
0.625	0.625	0.000	1.000
0.799	0.625	0.000	1.000
0.900	0.625	0.000	1.000

Table 4.61 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (200,200)$ and Five Categories of Risk

$k_1^2 = 0.201$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 200, m = 200$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.100	0.012
0.109	0.625	0.799	0.032	0.026
0.201	0.625	0.799	0.005	0.053
0.308	0.625	0.799	0.003	0.123
0.399	0.625	0.799	0.001	0.187
0.504	0.625	0.799	0.000	0.290
0.625	0.625	0.799	0.000	0.437
0.799	0.625	0.799	0.000	0.658
0.900	0.625	0.799	0.000	0.692

Table 4.62 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (400, 400)$ and Five Categories of Risk

$k_1^2 = 0.201$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 400, m = 400$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.237	0.001
0.109	0.625	0.799	0.056	0.007
0.201	0.625	0.799	0.012	0.035
0.308	0.625	0.799	0.002	0.121
0.399	0.625	0.799	0.000	0.246
0.504	0.625	0.799	0.000	0.411
0.625	0.625	0.799	0.000	0.596
0.799	0.625	0.799	0.000	0.841
0.900	0.625	0.799	0.000	0.913

Table 4.63 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (500, 500)$ and Five Categories of Risk

$k_1^2 = 0.201$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 500, m = 500$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.313	0.000
0.109	0.625	0.799	0.072	0.007
0.201	0.625	0.799	0.015	0.043
0.308	0.625	0.799	0.000	0.160
0.399	0.625	0.799	0.000	0.302
0.504	0.625	0.799	0.000	0.509
0.625	0.625	0.799	0.000	0.658
0.799	0.625	0.799	0.000	0.898
0.900	0.625	0.799	0.000	0.959

Table 4.64 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m, n) = (1000, 1000)$ and Five Categories of Risk

$k_1^2 = 0.201$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 1000, m = 1000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.579	0.000
0.109	0.625	0.799	0.136	0.009
0.201	0.625	0.799	0.009	0.043
0.308	0.625	0.799	0.000	0.206
0.399	0.625	0.799	0.000	0.451
0.504	0.625	0.799	0.000	0.743
0.625	0.625	0.799	0.000	0.866
0.799	0.625	0.799	0.000	0.994
0.900	0.625	0.799	0.000	0.999

Table 4.65 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (3000, 3000)$ and Five Categories of Risk

$k_1^2 = 0.201$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 3000, m = 3000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.973	0.000
0.109	0.625	0.799	0.349	0.000
0.201	0.625	0.799	0.021	0.026
0.308	0.625	0.799	0.000	0.414
0.399	0.625	0.799	0.000	0.864
0.504	0.625	0.799	0.000	0.992
0.625	0.625	0.799	0.000	1.000
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

Table 4.66 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (7000, 7000)$ and Five Categories of Risk

$k_1^2 = 0.201$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 7000, m = 7000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	1.000	0.000
0.109	0.625	0.799	0.669	0.000
0.201	0.625	0.799	0.026	0.028
0.308	0.625	0.799	0.000	0.745
0.399	0.625	0.799	0.000	0.996
0.504	0.625	0.799	0.000	1.000
0.625	0.625	0.799	0.000	1.000
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

Table 4.67 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (1000, 1000)$ and Five Categories of Risk

$k_1^2 = 0.000$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 1000, m = 1000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.011	0.046
0.109	0.625	0.799	0.000	0.999
0.201	0.625	0.799	0.000	0.314
0.308	0.625	0.799	0.000	0.696
0.399	0.625	0.799	0.000	0.929
0.504	0.625	0.799	0.000	0.985
0.625	0.625	0.799	0.000	0.998
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

Table 4.68 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (3000,3000)$ and Five Categories of Risk

$k_1^2 = 0.000$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 3000, m = 3000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.018	0.032
0.109	0.625	0.799	0.000	0.651
0.201	0.625	0.799	0.000	0.988
0.308	0.625	0.799	0.000	1.000
0.399	0.625	0.799	0.000	1.000
0.504	0.625	0.799	0.000	1.000
0.625	0.625	0.799	0.000	1.000
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

Table 4.69 Power for the Null Hypothesis $\sum_{i=1}^{3} \ln(k_i^2 + 1) = \ln S$ at a Sample Size of $(m,n) = (7000, 7000)$ and Five Categories of Risk

$k_1^2 = 0.000$ $k_2^2 = 0.625$ $k_3^2 = 0.799$				
$n = 7000, m = 7000$			$\alpha = 0.05$	
			$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2	k_3^2		
0.000	0.625	0.799	0.024	0.030
0.109	0.625	0.799	0.000	0.940
0.201	0.625	0.799	0.000	1.000
0.308	0.625	0.799	0.000	1.000
0.399	0.625	0.799	0.000	1.000
0.504	0.625	0.799	0.000	1.000
0.625	0.625	0.799	0.000	1.000
0.799	0.625	0.799	0.000	1.000
0.900	0.625	0.799	0.000	1.000

Tables 4.70 through 4.78 give the power of the test statistics, $\frac{\hat{k}_1^2 - \hat{k}_2^2 - \hat{bias}_{\hat{k}_1^2 - \hat{k}_2^2}}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2 - \hat{k}_2^2}}}$, derived in Section 3.4 for testing

$$H_o : k_1 - k_2 = 0$$
$$vs.$$
$$H_a : k_1 - k_2 \neq 0.$$

The parameter values used to simulate the test statistics are indicated in the tables along with the sample size. The critical values are found by using the values of the standard normal distribution that gives an area under the curve equal to $1 - \frac{\alpha}{2}$.

For five categories of risk, the results show high power at sample sizes as small as $m = n = 200$ if both risk factors are not risks, that is, $k_1^2 = k_2^2 = 0$. If $k_1^2 = k_2^2 \neq 0$, the

sample sizes must be as large as $n = m = 1000$ before the test shows high power. The power of the test statistic is also dependent on the parameters p_i, q_i, $i = 1,2,\ldots,c$, that make up k^2. In fact, if p_i for some i is small and the corresponding q_i is large, then the variance of k^2 will be much larger than the variance of some other k^2 in which the p_i, $i = 1,2,\ldots c$, are equally distributed. For example, in the simulation study, test statistics are generated from two different populations with parameters $k^2 = 0.625$ and $k^2 = 0.746$ with corresponding values of **p** and **q** given by

$$\mathbf{q} = 0.5,\ 0.2,\ 0.15,\ 0.1,\ 0.05$$
$$\mathbf{p} = 0.2,\ 0.2,\ 0.2,\ 0.2,\ 0.2$$

and

$$\mathbf{q} = 0.2,\ 0.2,\ 0.2,\ 0.2,\ 0.2$$
$$\mathbf{p} = 0.5,\ 0.2,\ 0.15,\ 0.1,\ 0.05$$

respectively. The associated variance, $V(\hat{k}^2)$, of $k^2 = 0.625$ with $m = n = 5000$ (also given in Table 4.7) is 0.001735 and the $V(\hat{k}^2)$ of $k^2 = 0.746$ with $m = n = 5000$ is 0.003731. Notice that even though the variance in the test statistic, $\frac{\hat{k}_1^2 - \hat{k}_2^2 - \left(\hat{bias} - \hat{bias}_2\right)}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2 - \hat{k}_2^2}}}$ is pooled, the variance associated with the null hypothesisis of $k_1^2 = k_2^2 = 0.625$ is smaller than the variance associated with the alternative hypothesis, making the test statistic associated with the alternative hypothesis smaller than that of the null hypothesis. Therefore, the power for the alternative hypothesis drops below the power of the null hypothesis for small sample sizes. The recommendation is to use this test only when the $\hat{p}_i$ $i = 1,2,\ldots,c$, are fairly equally spaced for both $\hat{k}_1^2$ and $\hat{k}_2^2$, if both are assumed to be risks.

Table 4.70 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (200,200)$ and Five Categories of Risk

$k_1^2 = 0.625, k_2^2 = 0.625$			
$n = 200, m = 200$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	0.988	0.000
0.109	0.625	0.809	0.000
0.201	0.625	0.672	0.000
0.308	0.625	0.290	0.001
0.399	0.625	0.180	0.001
0.504	0.625	0.096	0.015
0.625	0.625	0.033	0.026
0.799	0.625	0.008	0.099
0.900	0.625	0.002	0.113

Table 4.71 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (400,400)$ and Five Categories of Risk

$k_1^2 = 0.625, k_2^2 = 0.625$			
$n = 400, m = 400$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	0.960	0.000
0.201	0.625	0.908	0.000
0.308	0.625	0.468	0.000
0.399	0.625	0.222	0.001
0.504	0.625	0.104	0.001
0.625	0.625	0.034	0.020
0.799	0.625	0.008	0.161
0.900	0.625	0.002	0.213

Table 4.72 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (500, 500)$ and Five Categories of Risk

$k_1^2 = 0.625, k_2^2 = 0.625$			
$n = 500, m = 500$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	0.976	0.000
0.201	0.625	0.962	0.000
0.308	0.625	0.557	0.000
0.399	0.625	0.291	0.001
0.504	0.625	0.109	0.004
0.625	0.625	0.033	0.032
0.799	0.625	0.002	0.179
0.900	0.625	0.002	0.250

Table 4.73 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (1000, 1000)$ and Five Categories of Risk

$k_1^2 = 0.625, k_2^2 = 0.625$			
$n = 1000, m = 1000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	0.829	0.000
0.399	0.625	0.497	0.000
0.504	0.625	0.178	0.003
0.625	0.625	0.032	0.034
0.799	0.625	0.000	0.276
0.900	0.625	0.000	0.402

Table 4.74 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (3000,3000)$ and Five Categories of Risk

$k_1^2 = 0.625, k_2^2 = 0.625$			
$n = 3000, m = 3000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	0.998	0.000
0.399	0.625	0.925	0.000
0.504	0.625	0.404	0.000
0.625	0.625	0.033	0.028
0.799	0.625	0.000	0.621
0.900	0.625	0.000	0.850

Table 4.75 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (7000,7000)$ and Five Categories of Risk

$k_1^2 = 0.625, k_2^2 = 0.625$			
$n = 7000, m = 7000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.625	1.000	0.000
0.109	0.625	1.000	0.000
0.201	0.625	1.000	0.000
0.308	0.625	1.000	0.000
0.399	0.625	1.000	0.000
0.504	0.625	0.737	0.000
0.625	0.625	0.027	0.025
0.799	0.625	0.000	0.948
0.900	0.625	0.000	0.996

Table 4.76 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (200,200)$ and Five Categories of Risk

$k_1^2 = 0.000, k_2^2 = 0.000$			
$n = 200, m = 200$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.000	0.024	0.023
0.109	0.000	0.001	0.383
0.201	0.000	0.000	0.751
0.308	0.000	0.000	0.858
0.399	0.000	0.000	0.928
0.504	0.000	0.000	0.970
0.625	0.000	0.000	0.990
0.799	0.000	0.000	1.000
0.900	0.000	0.000	1.000

Table 4.77 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (500,500)$ and Five Categories of Risk

$k_1^2 = 0.000, k_2^2 = 0.000$			
$n = 500, m = 500$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.000	0.020	0.023
0.109	0.000	0.000	0.835
0.201	0.000	0.000	0.997
0.308	0.000	0.000	0.998
0.399	0.000	0.000	0.999
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.900	0.000	0.000	1.000

Table 4.78 Power for the Null Hypothesis $k_1^2 - k_2^2 = 0$ at a Sample Size of $(m,n) = (1000, 1000)$ and Five Categories of Risk

$k_1^2 = 0.000, k_2^2 = 0.000$			
$n = 1000, m = 1000$		$\alpha = 0.05$	
		$-Z_{0.025} = -1.96$	$Z_{0.025} = 1.96$
k_1^2	k_2^2		
0.000	0.000	0.021	0.026
0.109	0.000	0.000	0.991
0.201	0.000	0.000	1.000
0.308	0.000	0.000	1.000
0.399	0.000	0.000	1.000
0.504	0.000	0.000	1.000
0.625	0.000	0.000	1.000
0.799	0.000	0.000	1.000
0.900	0.000	0.000	1.000

4.2.8 Results for the Power of the Test Statistic $\hat{k}^2$ with 2, 4, 6, and 8 Categories of Risk

Tables 4.79 through 4.76 show the power of the test for the null hypothesis, $k^2 = 0$, (the factor is not a risk) vs. the alternative hypothisis $k^2 \neq 0$ (the factor is a risk) at the 0.01, 0.025, 0.050, and 0.100 α levels. As is seen from the tables, the power of the test statistic increases as the risk categories decrease, so that when the risk factor has only two categories, the power of the test statistic is very good at a sample size $m = n = 500$. As can be seen, the test has high power for $\hat{k}^2 > 0.1$ for sample size of $m = n = 100$. The power of the test increases as the sample size increases, which is to be expected because the distribution of $\hat{k}^2$ is an asymptotic distribution.

Table 4.79 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Two Categories of Risk for Sample Sizes of $m = 100$, $n = 100$

$n = 100, m = 100$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.015	0.025	0.052	0.103
0.048	0.134	0.260	0.352	0.453
0.100	0.372	0.499	0.603	0.699
0.496	0.991	0.997	1.000	1.000

Table 4.80 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Two Categories of Risk for Sample Sizes of $m = 500$, $n = 500$

$n = 500, m = 500$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.008	0.020	0.043	0.094
0.048	0.824	0.883	0.943	0.971
0.100	0.986	0.991	0.994	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.81 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Two Categories of Risk for Sample Sizes of $m = 1000$, $n = 1000$

$n = 1000, m = 1000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.011	0.024	0.052	0.109
0.048	0.987	0.994	0.995	0.999
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.82 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Two Categories of Risk for Sample Sizes of $m = 5000$, $n = 5000$

$n = 5000, m = 5000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.012	0.022	0.044	0.093
0.048	1.000	1.000	1.000	1.000
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.83 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Four Categories of Risk for Sample Sizes of $m = 100$, $n = 100$

$n = 100, m = 100$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.012	0.030	0.055	0.108
0.048	0.136	0.201	0.273	0.382
0.100	0.250	0.350	0.453	0.568
0.496	0.940	0.899	0.945	0.977

Table 4.84 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Four Categories of Risk for Sample Sizes of $m = 500$, $n = 500$

$n = 500, m = 500$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.013	0.030	0.056	0.098
0.048	0.684	0.813	0.878	0.889
0.100	0.955	0.973	0.989	0.997
0.496	1.000	1.000	1.000	1.000

Table 4.85 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Four Categories of Risk for Sample Sizes of $m = 1000$, $n = 1000$

$n = 1000, m = 1000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.010	0.025	0.050	0.109
0.048	0.974	0.990	0.996	0.998
0.100	0.999	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.86 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Four Categories of Risk for Sample Sizes of $m = 5000$, $n = 5000$

$n = 5000, m = 5000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.008	0.024	0.052	0.096
0.048	1.000	1.000	1.000	1.000
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.87 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Six Categories of Risk for Sample Sizes of $m = 100$, $n = 100$

$n = 100, m = 100$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.019	0.044	0.070	0.116
0.048	0.107	0.169	0.230	0.330
0.100	0.166	0.262	0.361	0.482
0.496	0.936	0.962	0.978	0.988

Table 4.88 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Six Categories of Risk for Sample Sizes of $m = 500$, $n = 500$

$n = 500, m = 500$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.009	0.030	0.056	0.110
0.048	0.586	0.702	0.782	0.851
0.100	0.947	0.975	0.989	0.995
0.496	1.000	1.000	1.000	1.000

Table 4.89 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Six Categories of Risk for Sample Sizes of $m = 1000$, $n = 1000$

$n = 1000, m = 1000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.008	0.027	0.046	0.085
0.048	0.950	0.971	0.987	0.994
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.90 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Six Categories of Risk for Sample Sizes of $m = 5000$, $n = 5000$

$n = 5000, m = 5000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.007	0.023	0.045	0.095
0.048	1.000	1.000	1.000	1.000
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.91 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Eight Categories of Risk for Sample Sizes of $m = 100$, $n = 100$

$n = 100, m = 100$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.021	0.034	0.080	0.156
0.048	0.102	0.150	0.212	0.297
0.100	0.212	0.291	0.375	0.482
0.496	0.323	0.333	0.450	0.540

Table 4.92 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Eight Categories of Risk for Sample Sizes of $m = 500$, $n = 500$

$n = 500, m = 500$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.015	0.035	0.063	0.116
0.048	0.509	0.618	0.720	0.817
0.100	0.922	0.952	0.974	0.984
0.496	1.000	1.000	1.000	1.000

Table 4.93 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Eight Categories of Risk for Sample Sizes of $m = 1000$, $n = 1000$

$n = 1000, m = 1000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.008	0.025	0.060	0.121
0.048	0.904	0.939	0.968	0.985
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

Table 4.94 Power of the Test Statistic $\hat{k}^2$ for the Null Hypothesis $Ho : k^2 = 0$ for Eight Categories of Risk for Sample Sizes of $m = 5000$, $n = 5000$

$n = 5000, m = 5000$				
	$\hat{k}^2$			
k^2	**0.010**	**0.025**	**0.050**	**0.100**
0.000	0.012	0.030	0.047	0.102
0.048	1.000	1.000	1.000	1.000
0.100	1.000	1.000	1.000	1.000
0.496	1.000	1.000	1.000	1.000

CHAPTER 5

APPLICATIONS TO REAL DATA

The examples in this chapter are from data in the literature. The focus of this chapter is to apply the statistics, $\hat{k}^2$ and functions of $\hat{k}^2$, developed in this study, to real data. Also, a comparison of $\hat{k}^2$ and the odds ratio applied to data with two levels of risk is provided.

5.1 An Example Using $\hat{k}^2$ and the Odds Ratio to Test for an Association Between Risk and Disease

The example in this section uses data from a case-control study in which the objective is to determine if there is an association between cleft lip and or cleft palate in infants and first trimester maternal smoking (Christensen et al. 1999). The 2×2 table presenting data for a case-control study with two categories of risk was given in Chapter 1 but is repeated here for continuity.

Table 5.1 2×2 Table Representing Data in a Case-Control Study

	Disease		Total	Disease	
Risk Factor	Yes	No		Yes	No
Yes	n_{11}	n_{01}	$n_{11}+n_{01}=e$	q_1	p_1
No	n_{10}	n_{00}	$n_{10}+n_{00}=ne$	q_0	p_0
Total	$n_{11}+n_{10}=m$	$n_{01}+n_{00}=n$	N	1	1

The odds ratio calculated from this table is

$$OR = \frac{n_{11}n_{00}}{n_{01}n_{10}}.$$

The null hypothesis of no risk from a case-control study with two categories of risk using the odds ratio is

$$\begin{aligned} H_o &: OR \\ &= \frac{n_{11}n_{00}}{n_{01}n_{10}} \\ &= 1 \end{aligned}$$

vs. the alternative hypothesis of

$$\begin{aligned} H_a &: OR \\ &= \frac{n_{11}n_{00}}{n_{01}n_{10}} \\ &> 1. \end{aligned}$$

Another way to state the null hypothesis is

$$H_o : \frac{n_{11}}{m} = \frac{n_{01}}{n}.$$

Here, n_{11} and n_{01} are considered to be independent binomial random variables with parameters (m, q_1) and (n, p_1). For large sample size, the test statistic for testing the null hypothesis is given by

$$Z = \frac{\frac{n_{11}}{m} - \frac{n_{01}}{n} - 0}{\sqrt{\frac{q_1(1-q_1)}{m} + \frac{p_1(1-p_1)}{n}}}.$$

Under the null hypothesis, q_1 and p_1 may be estimated by $\frac{n_{11}+n_{01}}{m+n}$. So the test statistic becomes

$$\begin{aligned} Z &= \frac{\frac{n_{11}}{m} - \frac{n_{01}}{n} - 0}{\sqrt{\frac{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)}{m} + \frac{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)}{n}}} \\ &= \frac{\frac{n_{11}}{m} - \frac{n_{01}}{n} - 0}{\sqrt{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)}}. \end{aligned} \tag{5.1}$$

If both sides of Eq. (5.1) are squared, then the test statistic becomes that with a chi-square distribution with one degree of freedeom, that is,

$$
\begin{aligned}
Z^2 &= \frac{\left(\frac{n_{11}}{m}-\frac{n_{01}}{n}\right)^2}{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)} \\
&= \frac{\left(\frac{n_{11}}{n_{11}+n_{10}}-\frac{n_{01}}{n_{01}+n_{00}}\right)^2}{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)} \\
&= \frac{\left(\frac{n_{11}(n_{01}+n_{00})-n_{01}(n_{11}+n_{10})}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)} \\
&= \frac{\left(\frac{n_{11}n_{01}+n_{11}n_{00}-n_{01}n_{11}-n_{01}n_{10}}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)} \\
&= \frac{\left(\frac{n_{11}n_{00}-n_{01}n_{10}}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{\frac{n_{11}+n_{01}}{m+n}\left(1-\frac{n_{11}+n_{01}}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)} \\
&= \frac{\left(\frac{n_{11}n_{00}-n_{01}n_{10}}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{\frac{n_{11}+n_{01}}{n_{11}+n_{10}+n_{01}+n_{00}}\left(1-\frac{n_{11}+n_{01}}{n_{11}+n_{10}+n_{01}+n_{00}}\right)\left(\frac{1}{n_{11}+n_{10}}+\frac{1}{n_{01}+n_{00}}\right)} \\
&= \frac{\left(\frac{n_{11}n_{00}-n_{01}n_{10}}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{\frac{n_{11}+n_{01}}{N}\left(1-\frac{n_{11}+n_{01}}{N}\right)\left(\frac{N}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)} \\
&= \frac{\left(\frac{n_{11}n_{00}-n_{01}n_{10}}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{(n_{11}+n_{01})\left(\frac{N-(n_{11}+n_{01})}{N}\right)\left(\frac{1}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)} \\
&= \frac{\left(\frac{n_{11}n_{00}-n_{01}n_{10}}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)^2}{(n_{11}+n_{01})\left(\frac{(n_{00}+n_{10})}{N}\right)\left(\frac{1}{(n_{11}+n_{10})(n_{01}+n_{00})}\right)} \\
&= \frac{\frac{N}{(n_{11}+n_{10})(n_{01}+n_{00})}(n_{11}n_{00}-n_{01}n_{10})^2}{(n_{11}+n_{01})(n_{00}+n_{10})} \\
&= \frac{\frac{N}{(m)(n)}(n_{11}n_{00}-n_{01}n_{10})^2}{(e)(ne)} \\
&= \frac{N(n_{11}n_{00}-n_{01}n_{10})^2}{mn(ne)e} = \chi_1^2. \qquad (5.2)
\end{aligned}
$$

Here, the marginal totals are all considered to be "fixed." The reason this assumption may be used is that the marginals do not provide information about the association between the risk and the disease. The only information that may be obtained from the marginals in a 2 × 2 table is the amount of data for quantifying the association between the risk and disease. Therefore, no bias is introduced by treating all of the marginals as "fixed" (Kleinbaum, Kupper, and Morgenstern 1982).

For the data in table 5.2, the test statistic for the null hypothesis $H_o : OR = 1$ vs. $H_a : OR > 1$ can be calculated using Eq. (5.2) as

$$\begin{aligned} Z^2 &= \frac{N(n_{11}n_{00} - n_{01}n_{10})^2}{mn(ne)e} \\ &= \frac{474(75 \times 193 - 139 \times 67)^2}{142 \times 332 \times 260 \times 214} \\ &= 4.8150. \end{aligned}$$

Table 5.2 Table Representing Data from a Case-Control Study Investigating the Association between Cleft Lip/Palate and Maternal Smoking

	Cases	Controls	Total	Odds Ratio
Smoker	(Cleft lip and/or Cleft palate)			
yes	75	139	214	1.55
no	67	193	260	1.00
Total	142	332	474	

Since the test statistic has a chi-square distribution with one degree of freedom, it may be compared with the value from the chi-square distribution, χ^2_1, that gives an area under the curve of 0.95, providing a level of significance of 0.05. This value is $\chi^2_{1,0.05} = 3.841$. Therefore, at the $\alpha = 0.05$ level, the null hypothesis may be rejected, and smoking may be considered to be associated with cleft lip and/or palate. If a more conservative test were desired, then $\alpha = 0.01$ may be considered. In this case, $\chi^2_{1,0.01} = 6.635$ and the null hypothesis would not be rejected. If $\hat{k}^2$ is calculated from the data in Table 5.2, one obtains

$$\begin{aligned} \hat{k}^2 &= \frac{\left(\frac{75}{142}\right)^2}{\left(\frac{139}{332}\right)} + \frac{\left(\frac{67}{142}\right)^2}{\left(\frac{193}{332}\right)} - 1 \\ &= 0.0492. \end{aligned}$$

From Chapter 3, it is known that $\hat{k}^2$ has a $Gamma\left(\frac{1}{2}, 2\left(\frac{142-332}{142 \cdot 332}\right)\right)$ distribution under the null hypothesis $k^2 = 0$. The hypothesis test to be conducted is $H_o : \hat{k}^2 = 0$ vs. $H_a : \hat{k}^2 > 0$. The

value that gives an area under the curve of the $Gamma\left(\frac{1}{2}, 2\left(\frac{142-332}{142 \cdot 332}\right)\right)$ distribution of 0.95 is 0.0386. Therefore, the null hypothesis would be rejected at the $\alpha = 0.05$ level. Again, if a more conservative test were desired, then the value that gives an area under the curve of the $Gamma\left(\frac{1}{2}, 2\left(\frac{142-332}{142 \cdot 332}\right)\right)$ distribution equal to 0.99 is 0.0667, and like the situation with the odds ratio, the null hypothesis would not be rejected at the $\alpha = 0.01$ level. Table 5.3 gives data pertaining to the cases that only had an isolated cleft palate without the cleft lip.

Table 5.3 Data from a Case-Control Study Investigating the Association between Cleft Palate and Maternal Smoking

	Cases	Controls	Total	Odds Ratio
Smoker	(Cleft palate)			
yes	19	139	158	1.00
no	29	193	222	0.91
Total	48	332	380	

Conducting the hypothesis test, $H_o : OR = 1$ vs. $H_a : OR > 1$ at the $\alpha = 0.05$ level gives a test statistic of $Z^2 = 0.090$ and $\chi^2_{1,0.05} = 3.841$. Therefore, the null hypothesis may not be rejected at the $\alpha = 0.05$ level. If the hypothesis test, $H_o : \hat{k}^2 = 0$ vs. $H_a : \hat{k}^2 > 0$ is conducted at the $\alpha = 0.05$ level, the test statistic is $\hat{k}^2$ =0.002144 with a critical value for rejection, from the $Gamma\left(\frac{1}{2}, 2\left(\frac{48-332}{48 \cdot 332}\right)\right)$, of 0.0916. Therefore, in both tests the null hypothesis is not rejected.

5.2 An Example Using $\sum_{i=1}^{r}\left(\ln\left(\hat{k}_i^2 + 1\right)\right)$ as a Test Statistic for *r* Independent Risk Factors

The example in this section will use data from a case-control study in which the objective is to determine if there is an association between very preterm births and social differences (Ancel et al. 1999). Very preterm births are defined as birth before 22 to 32 weeks

of gestation. The factors considered are obstetric history, marital status, and maternal age. These factors are treated as independent factors. The statistic, $\sum_{i=1}^{3}\left(\ln\left(\hat{k}_i^2+1\right)\right)$, is calculated for the three factors and a hypothesis test of $H_o : \sum_{i=1}^{r}\left(\ln\left(\hat{k}_i^2+1\right)\right) = 0$ vs. $H_a : \sum_{i=1}^{r}\left(\ln\left(\hat{k}_i^2+1\right)\right) > 0$ is conducted. The data from the study are given in tables 5.4 through 5.6.

Table 5.4 Data from a Case-Control Study Investigating the Association between Obstetric History and Very Preterm Births

	Cases	Controls	Total	q_i	p_i
Obstetric history	(Very Preterm Birth)				
Primigravid women	562	2970	3532	0.350	0.382
Previous first-trimester abortion	375	1827	2202	0.234	0.235
Previous second-trimester abortion	100	233	333	0.062	0.030
Previous preterm birth	282	513	795	0.176	0.066
Multigravidae without any of the above outcomes	286	2231	2517	0.178	0.287
Total	1605	7774	9379	1	1

For this risk factor, $\hat{k}_1^2 = 0.261549$ and the distribution of $\hat{k}_1^2$ under the null hypothesis is the $Gamma\left(2, 2\frac{1605-7774}{1605\cdot7774}\right)$. The statistic $\ln\left(\hat{k}_1^2+1\right) = 0.23234$ and the distribution of $\ln\left(\hat{k}_1^2+1\right)$, under the null hypothesis that $\hat{k}_1^2$ is not a risk, is that derived in Section 3.3, Eq. (3.7), with $\alpha = 2$, $\beta' = 2\frac{1605-7774}{1605\cdot7774}$, that is,

$$f(Z) = \frac{(e^z-1)^{\alpha-1}e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma(\alpha)(\beta')^{\alpha}}e^z dz, \quad z > 0$$

$$= \frac{(e^z-1)^{2-1}e^{-\frac{(e^z-1)}{2\frac{1605-7774}{1605\cdot 7774}}}}{\Gamma(2)\left(2\frac{1605-7774}{1605\cdot 7774}\right)^2}e^z dz$$

$$= \frac{(e^z-1)^1 e^{-\frac{(e^z-1)}{0.0015}}}{(0.0015)^2}e^z dz.$$

Data for marital status and very preterm births is given in table 5.5.

Table 5.5 Data from a Case-Control Study Investigating the Association between Marital Status and Very Preterm Births

	Cases	Controls	Total	q_i	p_i
Marital Status	(Very Preterm Birth)				
Married	1149	6123	7272	0.733	0.797
Unmarried cohabiting	287	1168	1455	0.183	0.152
Unmarried, not cohabiting	132	392	524	0.084	0.051
Total	1568	7683	9251	1	1

For this risk factor, $\hat{k}_2^2 = 0.0328146$. The distribution for $\hat{k}_2^2$ under the null hypothesis of no risk is the $Gamma\left(1, 2\frac{1568-7683}{1568\cdot 7683}\right)$. The statistic $\ln\left(\hat{k}_2^2+1\right) = 0.032288$ and its distribution under the null hypothesis of no risk is again that derived in Eq. (3.7), with $\alpha = 1$, $\beta' = 2\frac{1568-7683}{1568\cdot 7683}$. Table 5.6 gives the data for the maternal age.

Table 5.6 Data from a Case-Control Study Investigating the Association between Marital Status and Very Preterm Births

	Cases	Controls	Total	q_i	p_i
Maternal age	(Very Preterm Birth)				
<20	99	350	449	0.061	0.045
20-24	298	1780	2078	0.184	0.229
25-29	486	2565	3051	0.300	0.333
30-34	407	2130	2537	0.251	0.274
35-39	254	785	1039	0.157	0.101
≥ 40	76	163	239	0.047	0.021
Total	1621	7773	9394	1	1

From the table 5.6, $\hat{k}_3^2$ is calculated to be 0.08243. The distribution of $\hat{k}_3^2$ under the null hypothesis of no risk is the $Gamma\left(\frac{5}{2}, 2\frac{1621+7773}{1621\cdot 7773}\right)$. The statistic $\ln\left(\hat{k}_3^2+1\right) = 0.079208$. Again, the distribution of $\ln\left(\hat{k}_3^2+1\right)$ under the null hypothesis is that given by Eq. (3.7) with $\alpha = \frac{5}{2}, \beta' = 2\frac{1621+7773}{1621\cdot 7773}$.

For the three independent risk factors,

$$\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right) = 0.23234 + 0.032288 + 0.079208$$
$$= 0.34384.$$

The distribution of $\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)$ under the null hypothesis that none of the k_i^2, $i = 1,2,3$ are risks was derived in Section 3.3 and is given by Eq. (3.11), that is,

$$f(z) = \frac{(e^z-1)^{\sum_{i=1}^{r}\alpha_i-1}\, e^{-\frac{(e^z-1)}{\beta'}}}{\Gamma\left(\sum_{i=1}^{r}\alpha_i\right)(\beta')^{\sum_{i=1}^{r}\alpha_i}}e^z dz$$

with $\sum_{i=1}^{3}\alpha_i = \frac{11}{2}$ and $\beta' = 0.0015$. Here, the sample sizes for the cases and controls are not

exactly the same, but they are approximately the same and $\beta' = 0.0015$ for all three risk factors.

From the distribution, $f(z)$, above, the critical value for rejecting $H_o : \sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right) = 0$ in favor of $H_a : \sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right) > 0$ at $\alpha = 0.05$ is 0.01465. Since 0.34384 is larger than 0.01465, the null hypothesis would then be rejected.

Now, if there were an available registry that kept data on very preterm births, a standardized incidence ratio, S, may be calculated and $\ln S$ may be compared to $\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right)$. For demonstration purposes, assume there were such a registry, and a value of S was found to be 1.78. Then the hypothesis test $H_o : \sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right) = \ln 1.78$ vs $H_a : \sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right) \neq \ln 1.78$ may be conducted, where the distribution of the test statistic,

$$\frac{\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right) - \ln S + \sum_{i=1}^{3} \frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3} \frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2 + 1\right)^2}}}$$

$$= \frac{\sum_{i=1}^{3} \ln\left(\hat{k}_i^2 + 1\right) - \ln 1.78 + \sum_{i=1}^{3} \frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3} \frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2 + 1\right)^2}}},$$

was discussed in Section 3.3 and has a standard normal distribution. Recall that if all of the risk factors have been considered in the study then $S = \prod_{i=1}^{t}(1 + k_i^2)$. For $\ln S = 1.78$, the value of the test statistic is

$$\frac{\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)-\ln 1.78+\sum_{i=1}^{3}\frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3}\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}$$

$$= \frac{0.343836-0.576613+0.0007618}{0.039034}$$

$$= -5.94384$$

and the null hypothesis would be rejected. This would indicate that not all of the risks have been included in the study and that more risk factors may be associated with the disease. On the other hand, if the standardized incidence ratio, S, were calculated to be 1.44, a value close to the actual sample estimate of $\left(\prod_{i=1}^{3}\left(1+\hat{k}_i^2\right)=1.41035\right)$, then our test statistic would be equal to

$$\frac{\sum_{i=1}^{3}\ln\left(\hat{k}_i^2+1\right)-\ln 1.44+\sum_{i=1}^{3}\frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{3}\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}$$

$$= \frac{0.343836-0.364643+0.0007618}{0.039034}$$

$$= -0.51353$$

and the null hypothesis would not be rejected.

5.3 An Example Using $D = \hat{k}_i^2 - \hat{k}_j^2$ as a Test Statistic for Comparison of Two Independent Risk Factors

The next step may be to decide which risk factors are different in the case-control. Using the same data as that in Section 5.2, hypothesis tests of $H_o : \hat{k}_i^2 - \hat{k}_j^2 = 0$ vs. $H_a : \hat{k}_i^2 - \hat{k}_j^2 \neq 0$, $i,j = 1,2,3$, $i \neq j$ may be conducted. The test statistic for this test was given in section 3.4 by Eq. (3.18), $\frac{\hat{k}_1^2 - \hat{k}_2^2 - \hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2}}}$ and has a standard normal distribution.

The following estimates may be made from the above data.

$$\hat{bias}_{\hat{k}_1^2-\hat{k}_2^2} = 0.0055720$$
$$\hat{\sigma}_{\hat{k}_1^2-\hat{k}_2^2} = 0.0230393$$
$$\hat{bias}_{\hat{k}_1^2-\hat{k}_3^2} = 0.0011214$$
$$\hat{\sigma}_{\hat{k}_1^2-\hat{k}_3^2} = 0.0361101$$
$$\hat{bias}_{\hat{k}_3^2-\hat{k}_2^2} = 0.0044506$$
$$\hat{\sigma}_{\hat{k}_3^2-\hat{k}_2^2} = 0.0206312.$$

The results of hypothesis tests, $H_o : k_i^2 - k_j^2 = 0$ vs $H_a : k_i^2 - k_j^2 \neq 0$ for $i,j = 1,2,3,\ i \neq j$, conducted at the $\alpha = 0.05$ level are summarized in table 5.7.

Table 5.7 Summary of Results from Hypothesis Test H_o: k_i^2–k_j^2= $bias_1$–$bias_2$ vs. H_a: k_i^2–k_j^2= $bias_1$–$bias_2$ for $i,j = 1,2,3,\ i \neq j$, Conducted at $\alpha = 0.05$

i	j	$\hat{k}_i^2$	$\hat{k}_j^2$	$\hat{k}_i^2 - \hat{k}_j^2$	Test Statistic $\dfrac{\hat{k}_i^2 - \hat{k}_j^2 - \left(\hat{bias}_i - \hat{bias}_j\right)}{\hat{\sigma}_{\hat{k}_i^2-\hat{k}_j^2}}$	Critical Values	Accept/ Reject H_o
1	2	0.261549	0.0328146	0.228734	9.68614	±1.96	reject
1	3	0.261549	0.0824296	0.179119	4.92930	±1.96	reject
3	2	0.082429	0.0328146	0.049615	2.18913	±1.96	reject

Therefore, none of the above risk factors may be considered to be the same level of risk.

CHAPTER 6

SUMMARY

The asymptotic distribution of $\hat{k}^2$ under the null hypothesis, $k^2 = 0$, for one risk factor with c levels, is $Gamma\left(\frac{c-1}{2}, 2\frac{n-m}{n*m}\right)$. Under the alternative hypothesis,. $k^2 \cong 0$, $\hat{k}^2$ has a noncentral $\chi^2\left(c-1, \frac{mn}{(m+n)}\sum_{i=1}^{c}\frac{(mq_i - mp_i)^2}{mp_i}\right)$ distribution If the parameters p_i, q_i, $i = 1,2,\ldots c$, are different, then $\hat{k}^2 \sim N\left(\mu_{\hat{k}^2}, \sigma^2_{\hat{k}^2}\right)$. Here, $\mu_{\hat{k}^2}$ is the expected value of $\hat{k}^2$ that was derived in section 2.4, Eq. (2.27) that is,

$$E\left(\hat{k}^2\right) \cong \sum_{i=1}^{c}\left(\frac{q_i(1-q_i)}{mp_i} + \frac{q_i(1-q_i)(1-p_i)}{m(n+e)p_i^2} + \frac{q_i^2}{p_i} + \frac{q_i^2(1-p_i)}{(n+e)(p_i)^2}\right) - 1,$$

and $\sigma^2_{\hat{k}^2}$ is given in Eq. (2.33) by

$$V\left(\hat{k}^2\right) \cong \sum_{i=1}^{c}\left(\frac{4q_i^2(q_i(1-q_i))}{p_i^2 Ncases} + \frac{q_i^4(p_i(1-p_i))}{p_i^4 Ncontrols}\right) + 2\sum_{j=1}^{c-1}\sum_{i=j-1}^{c}\left(\frac{4q_iq_j(-q_iq_j)}{p_ip_j Ncases} + \frac{4q_i^2q_j^2(-p_ip_j)}{p_i^2p_j^2 Ncontrols}\right).$$

The power of the test statistic under the null hypothesis is shown to be high for sample sizes of 200 and above.

The asymptotic distribution of $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2 + 1\right)$ under the null hypothesis, $\sum_{i=1}^{r}\ln(k_i^2 + 1) = 0$, $r \geq 1$, is shown to have a probability distribution function of $f(z) = \frac{\sum_{i=1}^{r}\alpha_i - 1 \; -\frac{(e^z-1)}{\beta'}}{(e^z-1)^{i=1} \quad e}{\Gamma\left(\sum_{i=1}^{r}\alpha_i\right)(\beta')^{\alpha}}e^z dz$ $z \geq 0, r \geq 1$. The simulation study shows this to be a very

good approximation of the distribution of $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)$, $r \geq 1$ under the null hypothesis $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)=0$, $r \geq 1$. The power of this test statistic under the null hypothesis $\sum_{i=1}^{r}\ln(k_i^2+1)=0$ is shown to be high for sample sizes of 500 and above.

Under the null hypothesis, $H_o : \sum_{i=1}^{r}\ln(k_i^2+1)=\ln S$ vs. $H_a : \sum_{i=1}^{r}\ln(k_i^2+1)\neq \ln S$, the test statistic $\dfrac{\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)-\ln S-\sum_{i=1}^{r}\frac{1}{2\left(\hat{k}_i^2+1\right)^2}\hat{\sigma}_i^2}{\sqrt{\sum_{i=1}^{r}\frac{\hat{\sigma}_i^2}{\left(\hat{k}_i^2+1\right)^2}}}$ has a standard normal distribution, where $\ln S$ is the standardized incidence ratio from an appropriate cancer registry and $\hat{k}_i^2$, $i=1,2,\ldots,r$, are independent. The simulation study again shows this to be a good approximation for the distribution of $\sum_{i=1}^{r}\ln\left(\hat{k}_i^2+1\right)$, $r \geq 1$ especially for large sample sizes.

A statistic to test the difference between two independent risk factors, $H_o : k^2-k_2^2=0$ vs. $H_a : k^2-k_2^2\neq 0$, is also developed. The test statistic in this case is

$$\frac{\hat{k}_1^2-\hat{k}_2^2-\hat{bias}_{\hat{k}_1^2-\hat{k}_2^2}}{\sqrt{\hat{\sigma}^2_{\hat{k}_1^2-\hat{k}_2^2}}}\sim N(0,1).$$

The simulation results show high power at sample sizes as small as $m=n=200$ if both risk factors are not risks, that is, $k_1^2=k_2^2=0$. If $k_1^2=k_2^2\neq 0$ the sample sizes must be as large as $n=m=1000$ before the test shows high power.

This study has assumed all risk factors are independent. Further research is needed to investigate a statistic to estimate S if the risk factors are dependent.

APPENDIX

SOURCE CODE FOR SIMULATION

```
//This program reads in and manipulates data from files produced
//by the program RNTMN. The program contains a class called appgi,
//which creates objects that contain the following attributes:
//      1.    the square of the sample coefficient of variation of
//            incidence of disease over the risk categories (CV) calculated
//            from a case-control study.
//      2.    the square of the coefficient of variation of
//            incidence of disease over the risk categories (CV) calculated
//            from the parameters that generated the sample.
//      3.    the asymptotic variance of the (CV) calculated from the sample.
//      4.    the estimated bias of the (CV) calculated from the sample.
//      5.    the parameter bias of the (CV) calculated from the parameters
//            that generated the sample.
//      6.    the natural log of the (CV) calculated from the sample.
//      7.    the natural log of the (CV) calculated from the parameters
//            that generated the sample.
//      8.    the asymptotic variance of the natural log of the (CV)
//            calculated from the sample.
//      9.    the asymptotic variance of the natural log of the (CV)
//            calculated from the parameters that generated the sample.
//      10.   the (CV) calculated from the sample using Begg's nonparmetric
//            estimate.
//      11.   the asymptotic variance of Begg's estimate calculated from
//            the sample.
//      12.   the chi-square test statistic.
//      13.   the total number of cases and controls in the sample
//      14.   the percent of cases and controls in each category in
//            the sample.
//Other member functions included in the class that
//are used to calculate the asymptotic variance are
//      covgigj
//      covpipj
//      covqiqj.

//Also, the class contains member functions called setup_files,
//setup_files1,...,setup_files10 that read from an
//external file called info.txt.

//Info.txt contains the name of all
//the files that the program needs to open and read. These
//are files that were created by the Fortran subroutine, RNTMN,
//and contain the random variates for the simulation.

//The member function called setup_ksqr reads from the external
//files created by RNTMN which contain the 1000 samples from each population.

//The main program calls a function called WriteTable that creates
//the heading of the tables, instantiates the objects from the
//class appgi, and declares and defines arrays to hold the object's
```

```
//attributes mentioned above.

//WriteTable calls a function called CalcStats that in turn
//calls the functions trial, powertable, and writetotable.

//The trial function calculates the sample
//averages for the statistics gathered from the sample.

//The powertable uses the appropriate critical value (defined at the
//beginning of the program with a pound define command) to calculate
//the percent of sample statistics that exceed the critical value.

//The writetotable writes the column heads of the table to the
//appropriate files and the statistics calculated by the method, trial.

#include <iostream>
#include <iomanip>
#include <stdlib.h>
#include <fstream.h>
#include <string.h>
#include <ctype.h>
#include <math.h>
#include <numeric>

//Below are the values that will remain constant in the program

#define cat 5          //number of levels in the case control study
#define reps 1000      //number of samples created from the population

#define ksqr0_valHo 0.0
#define ksqr0_lnvalHo 0.0
#define ksqr05_valHo 0.05
#define ksqr05_lnvalHo 0.049
#define ksqr1_valHo 0.1
#define ksqr1_lnvalHo 0.095
#define ksqr2_valHo 0.2
#define ksqr2_lnvalHo 0.182
#define ksqr3_valHo 0.3
#define ksqr3_lnvalHo 0.262
#define ksqr4_valHo 0.4
#define ksqr4_lnvalHo 0.336
#define ksqr5_valHo 0.5
#define ksqr5_lnvalHo 0.405
#define ksqr625_valHo 0.625
#define ksqr625_lnvalHo 0.486
#define ksqr8_valHo 0.8
#define ksqr8_lnvalHo 0.588
#define ksqr75_valHo 0.75
#define ksqr75_lnvalHo 0.560
#define ksqr9_valHo 0.9
```

```
#define ksqr9_InvalHo 0.641

//Below are the critical values used
//to determine the power of the test.
//These values are changed dependent on the
//test statistic being used.

#define ksqrho_50 0.739012
#define ksqrho_100 0.369506
#define ksqrho_200 0.0739012
#define ksqrho_300 0.0369506
#define ksqrho_400 0.00739012
#define ksqrho_500 0.0739012
#define ksqrho_600 0.0615844
#define ksqrho_700 0.0527866
#define ksqrho_800 0.0461883
#define ksqrho_900 0.0410562
#define ksqrho_1000 0.0369506
#define ksqrho_3000 0.0123169
#define ksqrho_5000 0.00739012
#define ksqrho_7000 0.00527866
#define ksqrho_9000 0.00410562

//Below are the commands to set up the files to read from and to

ofstream outfile;//("chi.txt",ios::out);
ofstream outf;//("normal.txt",ios::out);

//Here, there are 11 different sample sizes in each run of the program.
//There is an input file for the controls and one for the cases
//for each sample size.

ifstream controls;
ifstream cases;
ifstream controls1;
ifstream cases1;
ifstream controls2;
ifstream cases2;
ifstream controls3;
ifstream cases3;
ifstream controls4;
ifstream cases4;
ifstream controls5;
ifstream cases5;
ifstream controls6;
ifstream cases6;
ifstream controls7;
ifstream cases7;
ifstream controls8;
ifstream cases8;
```

```
ifstream controls9;
ifstream cases9;
ifstream controls10;
ifstream cases10;
ifstream info("info.txt",ios::in);


using namespace std;       //introduces namespace std


class appgi
{
public:
appgi();                                                    //constructor
appgi(string p);                                            //alternate constructor
double figurechisqr();                                      //method to calculate chi-square statistic
double figure_lnksqr();                                     //method to calculate ln ksqr,variance of lnksqr
double get_varlnksqr();                                     //method to return variance of lnksqr
double get_lnksqr();                                        //method to return theoretical ln ksqr
double get_avglnksqr();                                     //method to return sample ln ksqr
double covqiqj(int x,int y);                                //method to calculate covariance qi and qj
double covqconiqconj(int x,int y);                          //method to calculate covariance of qconi and
                                                            //qconj
double covpiqj(int x,int y);                                //method of calculate covariance pi and qj
double covpipj(int x,int y);                                //method of calculate covariance pi and pj
double covgigj(int x,int y);                                //method of calculate covariance gi and gj
void figure_ksqr();                                         //method to calculate ksqr from sample
double para_ksqr();                                         //method to calculate ksqr from parameters
double para_bias();                                         //method to calculate the bias from parameters
double calcksqrVAR();                                       //method to calculate the variance of ksqr
double calchisksqrVAR();                                    //method to calculate the variance of ksqrb
                                                            //(Begg's estimate)
void setup_ksqr();                                          //method to set up the files to read in the
                                                            //multinomial variate that were generated from
                                                            //RNTMN

void setup_ksqr1();
void setup_ksqr2();
void setup_ksqr3();
void setup_ksqr4();
void setup_ksqr5();
void setup_ksqr6();
void setup_ksqr7();
void setup_ksqr8();
void setup_ksqr9();
void setup_ksqr10();
void setup_files(int i);                                    //method to read in the appropriate files to read from
                                                            //for each population of ksqr

void setup_files1(int i);
void setup_files2(int i);
```

```
void setup_files3(int i);
void setup_files4(int i);
void setup_files5(int i);
void setup_files6(int i);
void setup_files7(int i);
void setup_files8(int i);
void setup_files9(int i);
void setup_files10(int i);
void setup_files11(int i);
double getTotalControls();          //method to return the total number of controls
double getTotalCases();             //method to return the total number of cases
void Bias();                        //method to calculate the bias
double get_Bias();                  //method to return the bias
double get_ksqr();                  //method to return ksqr
double get_hisksqr();               //method to return ksqrb
double get_ourksqr();


private:
double NumofCase[cat];              //this is an array of random variates
                                    //generated by RNTMN that will be
                                    //read in from a file
double NumofCon[cat];               //this is an array of random variates
                                    //generated by RNTMN that will be
                                    //read in from a file
double NumInCat[cat];               //this is calculated from the prior two
                                    //as NumofCase[i]+NumofCon[i]
double p[cat];                      //this is calculated by
                                    //NumInCat[i]/NumInStudy
double q[cat];                      //this is calculated by
                                    //NumofCase[i]/TotalCases
double qcon[cat];                   //this is calculated by
                                    //NumofCon[i]/TotalControl
double g[cat];                      //this is an array that keeps the
                                    //terms of ksqr before summing
double TotalNumInStudy;             //this is calculated as Total
                                    //Cases + TotalControls
double paraq[cat];                  //same as q[cat] but uses
                                    //parameters not sample
double parap[cat];                  //same as p[cat] but uses
                                    //parameters not sample
double para_q[cat];                 //same as q[cat] but uses
                                    //parameters not sample
double para_p[cat];                 //same as p[cat] but uses
                                    //parameters not sample
double sumofquotient;               //this is the sum of the terms in
                                    //ksqr from sample
double para_sumofquotient;          //this is the sum of the terms in ksqr
                                    //from parameters
```

```
double varq[cat];                          //the is array that holds the variance
                                                                    //of the cases

double varp[cat];
double varqcon[cat];                       //the is array that holds the variance
                                                                    //of the controls
double partialgiwrq[cat];                  //array of partial deriv w/r to q
double partialgiwrp[cat];                  //array of partial deriv w/r to p
double secpartialgiwrp[cat];               //array of partial deriv w/r to p for
                                                                    //ksqrb
double secpartialgiwrq[cat];               //array of partial deriv w/r to q for
                                                                    //ksqrb
double secpartialgiwrpq[cat];              //array of variance of ksqrb+1
double vargi[cat];                         //array of variance of ksqr+1
double nvarq[cat];                         //holds the varq for ksqrb
double nvarqcon[cat];                      //holds the varqcon for ksqrb
double ksqr;                               //holds the value for ksqr
double hisksqr_;                           //holds the value for ksqrb+1
double ourksqr_;                           //holds the value for ksqr+1
double hisksqr_sum;                        //holds the value for ksqrb+1
double ourksqr_sum;                        //holds the value for ksqr+1
double hisksqr[cat];                       //holds the value for ith term in
                                                                    //ksqrb+1
double ourksqr[cat];                       //holds the value for ith term in
                                                                    //ksqr+1
double covqp[cat];                         //holds the value for ith covariance
                                                                    //of p and q
double covqiqj_;                           //holds value for the covariance of
                                                                    //qi and qj
double covqconiqconj_;                     //holds value for the covariance of
                                                                    //qconi and qconj
double covpiqj_;                           //holds value for the covariance of
                                                                    //pi and qj
double covpipj_;                           //holds value for the covariance of
                                                                    //pi and pj
double covgigj_;                           //holds value for the covariance
                                                                    of terms in ksqr

double exvarksqr;
double exlnvarksqr;                        //holds the variance of ksqrb
double ex_lnksqr;                          //holds the value for theoretical ln
                                                                    //ksqr
double avg_lnksqr;                         //holds the value for sample ln
                                                                    //ksqr
double var_lnksqr;                         //holds the value for theoretical
                                                                    //variance of ln ksqr

double ksqrAVG;
char C[35];                                //character array for name of file
                                                                    //to read from
char D[35];                                //character array for name of file
                                                                    //to read from
```

```
double TotalControls;                          //total controls
double TotalCases;                             //total cases
double Controls;                               //controls
double Cases;                                  //cases
double bias[cat];                              //array for terms in the
                                               //calculation of the bias
double biassum;                                //holds the value for the sum of
                                               //the terms in the bias
double samplebias[cat];                        //this array holds the i^th category
                                               //computation in order to
                                               //calculate the sample bias
double samplebiassum;                          //this value holds the sample bias
double samplebiassum1;                         //this value holds the parameter
                                               //bias
};



appgi::appgi()                                 //constructor
{ for(int k=0;k<35;k++)
{C[k]='\0';                    //this is a character array that reads in the name of a file
D[k]='\0';                     //this is a character array that reads in the name of a file
}//end of for

outfile.precision(6);
outf.precision(6);
outfile.setf(ios::fixed,ios::floatfield);
outf.setf(ios::fixed,ios::floatfield);


TotalNumInStudy=0;
TotalControls=0;
TotalCases=0;
sumofquotient=0;
ksqr=0;
hisksqr_=0;
ourksqr_=0;
hisksqr_sum=0;
ourksqr_sum=0;
covqiqj_=0;
covqconiqconj_=0;
covpiqj_=0;
covpipj_=0;
covgigj_=0;
exvarksqr=0;
exlnvarksqr=0;
biassum=0;
samplebiassum=0;
```

```
samplebiassum1=0;
for(int j=0;j<cat;j++)
{NumofCase[j]=0;
NumofCon[j]=0;
NumInCat[j]=0;
p[j]=0;
q[j]=0;
qcon[j]=0;
g[j]=0;
para_p[j]=0;
para_q[j]=0;
bias[j]=0;
covqp[j]=0;
varq[j]=0;
varp[j]=0;
varqcon[j]=0;
partialgiwrq[j]=0;
partialgiwrp[j]=0;
secpartialgiwrp[j]=0;
secpartialgiwrq[j]=0;
secpartialgiwrpq[j]=0;
vargi[j]=0;
samplebias[j]=0;
hisksqr[j]=0;
ourksqr[j]=0;

}//end of for
}//end of constructor

double appgi::getTotalControls()
{return TotalControls;
}

double appgi::getTotalCases()
{return TotalCases;
}

double appgi::para_bias()
{double samplebiascat=0;
double samplebiascat1=0;
double firstterm=0;
double secterm=0;
double thirdterm=0;
for(int j=0;j<cat;j++)
      {

      firstterm=para_q[j]*(1-para_q[j])/(TotalCases*para_p[j]);

      secterm=para_q[j]*(1-para_q[j])*(1-para_p[j])/
```

```
    (TotalCases*(TotalControls)*(para_p[j]*para_p[j]));

    thirdterm=para_q[j]*para_q[j]*(1-para_p[j])/
    (TotalControls*(para_p[j])*(para_p[j]));

    samplebiassum1=samplebiassum1+firstterm+secterm+thirdterm;
    }//end of for

    return samplebiassum1;
}//end of para_bias()

double appgi::para_ksqr()                    //this method calculates the ksqr with the parameters
{double ksqr1=0;
for(int j=0;j<cat;j++)
{ksqr1=ksqr1+para_q[j]*para_q[j]/para_p[j];
}
ksqr1=ksqr1-1;
return ksqr1;
}//end para_ksqr()

void appgi::Bias()                                   //this method calculates the bias of ksqr
{
for(int j=0;j<cat;j++)
    {
    samplebias[j]=q[j]*(1-q[j])/(TotalCases*qcon[j])+
    q[j]*(1-q[j])*(1-qcon[j])/
    (TotalCases*(TotalControls)*(qcon[j]*qcon[j]))+
    q[j]*q[j]*(1-qcon[j])/(TotalControls*(qcon[j])*(qcon[j]));

    samplebiassum=samplebiassum+samplebias[j];
    }//end of for
}//end of Bias()

double appgi::get_Bias()                                  //this method returns the bias
{return samplebiassum;
}//end of getBias()

void appgi::setup_files(int b)                          //This method reads from a file
{                                                          //called info. The info file
char t;                                                    //lists the files that contain
int k=0;                                                //the random variates from RNTMN.
                                                           //The appropriate file is then
int samplenum;                                          //opened depending on the sample
samplenum=b;                                               //sample size. There are 11 of
                                                        //methods, one for each population
                                                        //of ksqr, but only two are listed for
                                                                                   //brevity.

info>>t;
while(isspace(t)==false)
```

```
{
C[k]=t;
t=info.get();
k=k+1;
}
k=0;
info>>t;
while(isspace(t)==false)
{
D[k]=t;
t=info.get();
k=k+1;
}
cases.open(C,ios::in);
controls.open(D,ios::in);
info>>Cases;
info>>Controls;
for(int i=0;i<cat;i++)
{
info>>paraq[i];
}
for(int i=0;i<cat;i++)
{
info>>parap[i];
}
}//end of setup files

void appgi::setup_files1(int b)
{
char t;
int k=0;
int samplenum;
samplenum=b;
info>>t;
while(isspace(t)==false)
{
C[k]=t;
t=info.get();
k=k+1;
}
k=0;
info>>t;
while(isspace(t)==false)
{
D[k]=t;
t=info.get();
k=k+1;
}
cases1.open(C,ios::in);
```

```
controls1.open(D,ios::in);
info>>Cases;
info>>Controls;
for(int i=0;i<cat;i++)
{
info>>paraq[i];
}
for(int i=0;i<cat;i++)
{
info>>parap[i];
}
}//end of setup1 files

void appgi::setup_ksqr()                    //This method reads in the multinomial
{                                           //variates that were generated by RNTMN
TotalCases=Cases;
TotalControls=Controls;
TotalNumInStudy=TotalCases+TotalControls;
     cases>>NumofCase[0];
     controls>>NumofCon[0];
     for(int j=1;j<cat;j++)
     {cases>>NumofCase[j];
     controls>>NumofCon[j];
     }

     for(int j=0;j<cat;j++)
     {para_p[j]=parap[j];
     para_q[j]=paraq[j];

     NumofCon[j]=NumofCon[j]+.5;

     }//end of for

     double sum=0;
     for(int j=0;j<cat;j++)
     {
     sum=sum+NumofCon[j];
     }//end of for

     TotalControls=sum;
}//end setup_ksqr

double appgi::figurechisqr()                //this method calculates the chi-square
{double firstterm=0;                                              //statistic
double firsttermsum=0;
double secondterm=0;
double secondtermsum=0;
double chisqr=0;
     for(int k=0;k<cat;k++)
```

```
        {
        firstterm=
        ((NumofCase[k]-TotalCases*(NumofCase[k]+NumofCon[k])/
        (TotalCases+TotalControls))*
        (NumofCase[k]-TotalCases*(NumofCase[k]+NumofCon[k])/
        (TotalCases+TotalControls)))/
        (TotalCases*(NumofCase[k]+NumofCon[k])/
        (TotalCases+TotalControls));

        firsttermsum=firsttermsum+firstterm;
        }//end of for

        for(int k=0;k<cat;k++)
        {
        secondterm=((NumofCon[k]-TotalControls*
        (NumofCase[k]+NumofCon[k])/
        (TotalCases+TotalControls))*
        (NumofCon[k]-TotalControls*(NumofCase[k]+NumofCon[k])/
        (TotalCases+TotalControls)))/
        (TotalControls*(NumofCase[k]+NumofCon[k])/
        (TotalCases+TotalControls));
        secondtermsum=secondtermsum+secondterm;
        }//end of for

chisqr=firsttermsum+secondtermsum;
return chisqr;
}//end of figurechisqr

void appgi::figure_ksqr()                    //This method calculates ksqr, the
{                                            //variance and covariances needed
                                                  //and the ksqr(Begg), called
                                                                  //hisksqr.
        for(int k=0;k<cat;k++)
        {
        q[k]=NumofCase[k]/TotalCases;
        qcon[k]=NumofCon[k]/TotalControls;

        NumInCat[k]=NumofCase[k]+NumofCon[k];
        p[k]=NumInCat[k]/TotalNumInStudy;

        g[k]=((q[k])*(q[k]))/(qcon[k]);
        sumofquotient=sumofquotient+g[k];

                        for(int i=0;i<cat;i++)
                {if(NumofCon[k]==0.5)
                {NumofCon[k]=0;}//end of if
                }//end of for
        hisksqr[k]=(TotalControls+2)*NumofCase[k]*(NumofCase[k]-1)/
        (TotalCases*(TotalCases-1)*(1+NumofCon[k]));
```

```
//put back the .5 num of Controls for rest
      for(int i=0;i<cat;i++)
      {if(NumofCon[k]==0)
      {NumofCon[k]=0.5;}//end of if
      }//end of for


      hisksqr_sum=hisksqr_sum+hisksqr[k];

      covqp[k]=(1/(TotalNumInStudy*TotalCases))*(
      TotalCases*(q[k])*(1-q[k])+(TotalCases*q[k])
      *(TotalCases*q[k])
      +(TotalCases*q[k])*
      TotalControls*qcon[k]-
      (TotalCases*q[k]+TotalControls*qcon[k])
      *TotalCases*q[k]);

                  varq[k]=(q[k]*(1-q[k])/TotalCases);

varp[k]=(1/pow(TotalNumInStudy,2))*(TotalControls*qcon[k]*(1-qcon[k])
        +TotalCases*q[k]*(1-q[k]));

varqcon[k]=(qcon[k]*(1-qcon[k])/TotalControls);

partialgiwrq[k]=2.0*q[k]/(qcon[k]);

partialgiwrp[k]=-1.0*q[k]*q[k]/(qcon[k]*qcon[k]);

                  //put back the 0 num of Controls for his
                  for(int i=0;i<cat;i++)
                  {if(NumofCon[k]==0.5)
                  {NumofCon[k]=0;}//end of if
                  }//end of for
secpartialgiwrp[k]=-1.0/(pow(1+NumofCon[k],2))
                  *(NumofCase[k]*NumofCase[k]-NumofCase[k]);

secpartialgiwrq[k]=1/(1+NumofCon[k])
                  *(2.0*NumofCase[k]-1);

                  //the second partial is the first partial of his stat

secpartialgiwrpq[k]=(secpartialgiwrq[k]*secpartialgiwrq[k])
                  *nvarq[k]+(secpartialgiwrp[k]*secpartialgiwrp[k])*nvarqcon[k];

                  //this thing above is the vargi for his stat

                  //put back the .5 num of Controls for rest
                  for(int i=0;i<cat;i++)
```

```
                    {if(NumofCon[k]==0)
                    {NumofCon[k]=0.5;}//end of if
                    }//end of for
vargi[k]=(partialgiwrq[k]*partialgiwrq[k])
          *varq[k]+(partialgiwrp[k]*partialgiwrp[k])*varqcon[k];

     }//end of for


hisksqr_=hisksqr_sum-1;
ksqr=sumofquotient-1;
}//end of figure_ksqr

double appgi::figure_lnksqr()                    //This method calculates the variance for ksqr
{
     ex_lnksqr=log(ksqr+1)-.5*(1/((ksqr+1)*(ksqr+1)))*calcksqrVAR();
     avg_lnksqr=log(ksqr+1);
     var_lnksqr=(1/(ksqr+1))*(1/(ksqr+1))*calcksqrVAR();
}//end of figure_lnksqr()

double appgi::get_lnksqr()                       //this method returns the theoretical ln ksqr
{return ex_lnksqr;
}

double appgi::get_avglnksqr()                    //this method returns the sample ln ksqr
{return avg_lnksqr;
}

double appgi::get_varlnksqr()                    //this method returns the
{return var_lnksqr;                              //theoretical variance ln ksqr
}

double appgi::get_ksqr()                         //this method returns ksqr
{return ksqr;
}//end of get_ksqr

double appgi::get_hisksqr()                      //this method returns ksqrb
{return hisksqr_;
}//end of get_ksqr

double appgi::get_ourksqr()
{return ourksqr_;
}//end of our_ksqr

double appgi::covqiqj(int x,int y)               //this method calculates covariance of qi and
                                                 //qj
{
covqiqj_=-q[x]*q[y]*TotalCases;
return covqiqj_;
}//end of covqiqj
```

```
double appgi::covqconiqconj(int x,int y)                    //this method calculates covariance
{                                                                                    //of qconi and qconj
covqconiqconj_=-qcon[x]*qcon[y]*TotalControls;
return covqconiqconj_;
}//end of covqconiqconj

double appgi::covpiqj(int x,int y)
{
     covpiqj_=(1/(TotalNumInStudy*TotalCases))*
     ((TotalCases*TotalCases)*(covqiqj(x,y)+q[x]*q[y])+TotalControls
     *qcon[x]*TotalCases*q[y]-
     (TotalCases*q[x]+TotalControls*qcon[x])*TotalCases*q[y]);
     return covpiqj_;
}//end of covpiqj

double appgi::covpipj(int x,int y)
{covpipj_=-1.0*secpartialgiwrq[x]*secpartialgiwrq[y]*(NumofCase[x]*NumofCase[y]/TotalCases
(-1.0)*secpartialgiwrp[x]*secpartialgiwrp[y]*(NumofCon[x]*NumofCon[y]/TotalControls);
//NOTE: the second partial is the first partial of his stat
return covpipj_;
}//end of covpipj

double appgi::covgigj(int x,int y)
{ covgigj_=-1.0*partialgiwrq[x]*partialgiwrq[y]*(q[x]*q[y]/TotalCases)+
(-1.0)*partialgiwrp[x]*partialgiwrp[y]*(qcon[x]*qcon[y]/TotalControls);
return covgigj_;
}//end of covgigj

double appgi::calcksqrVAR()                                  //this method calculates the variance of
                                                                                    //ksqr
{               double exlnvarfirst_term=0;
                double exlnvarsecond_term=0;
     for(int x=0;x<cat-1;x++)
     {

          for(int y=x+1;y<cat;y++)
          {exlnvarfirst_term=exlnvarfirst_term+2.0*covgigj(x,y);
          }//end of for

     }//end of outside for

     for(int x=0;x<cat;x++)
     {
          exlnvarsecond_term=exlnvarsecond_term+vargi[x];

     }//end of for
exlnvarksqr=exlnvarfirst_term+exlnvarsecond_term;
```

```
return exlnvarksqr;
}

double appgi::calchisksqrVAR()                    //this method calculates the variance of ksqrb
{                  double exlnvarfirst_term=0;
                  double exlnvarsecond_term=0;
     for(int x=0;x<cat-1;x++)
     {

          for(int y=x+1;y<cat;y++)
          {exlnvarfirst_term=exlnvarfirst_term+2.0*covpipj(x,y);
          }//end of for

     }//end of outside for

     for(int x=0;x<cat;x++)
     {
          exlnvarsecond_term=exlnvarsecond_term+secpartialgiwrpq[x];

     }//end of for
exlnvarksqr=exlnvarfirst_term+exlnvarsecond_term;
exlnvarksqr=pow((TotalControls+2)/(TotalCases*(TotalCases+1)),2)*exlnvarksqr;
return exlnvarksqr;
}//end of calchisksqrVAR()

//This is the start of the main program which calls the function write table.
//The function write table creates objects of appgi, the ksqr. The
//appropriate methods are called to calculate the attributes, then written
//to a table. The write table is called 14 times, once for each sample
//size.

void main()
{
void WriteTable(int b);
for(int i=1;i<15;i++)
     {
WriteTable(i);
}//end of for
}//end of main

//The function WriteTable calls the functions calc_stats. The calc_stats
//function calls the functions trial, powertable, and writetotable. The
//trial function calculates the sample averages and the sample variances
//for the 11 populations of ksqr and ln ksqr+1. The powertable reads the
//appropriate critical value from the defined values to calculate the
//power of the hypothesis test. The writetotable function formats the
//information calculated and prints it to the appropriate out file.

void WriteTable(int b);
```

```
void WriteTable(int b)
{
//this section defines the functions WriteTable will call

void calc_stats(double ksqrlist_0Bias[],double ksqrlist_0[],doubleksqrvar_0[],.double
&ksqrVAR,double &ksqrAVG,double &ksqrSTD,double &ksqrBiasAVG,do-uble
logksqrlist_0[],double &ksqrEXP,double &betachi_0,double &betahi_0,doublle
&betalo_0,double ksqrlnname1,double ksqrname1,double hisksqrlist_0[],doub-le
ksqrhisvar_0[],double logavgksqrlist_0[],double ksqrlogvar_0[],int q,double
ksqrvar_ho[],double ksqrlogvar_ho[],double ksqrhisvar_ho[],double ksqrlist_hoBias[],double
ksqrhislist_hoBias[],double ksqrloglist_hoBias[],double &ksqrBiasAVGho,doruble
&ksqrVARho,double &ksqrhisAVGho,double &ksqrhisVARho,double
&ksqrlogEXPho,double &ksqrlogVARho);

void trial(double ksqrlist_0Bias[],double ksqrlist_0[],double ksqrvar_0[],doublle
&ksqrVAR,double &ksqrAVG,double &ksqrSTD,double &ksqrBiasAVG,doruble
logksqrlist_0[],double &ksqrEXP,double &ksqrBiasAVGho,double &ksqrVA_Rho,double
ksqrhisAVGho,double ksqrhisVARho,double ksqrlogEXPho,double ksqrlogV_ARho,int q);

void powertable(double ksqrlist_0[],double ksqrvar_0[],double &betachi_0,do-uble
&betahi_0,double &betalo_0,double ksqrlist_hoBias[],double &ksqrAVGho,d-ouble
&ksqrVARho);

void writetotable(double &ksqrBiasAVG,double &ksqrname1,double &ksqrA_VG,double
&ksqrSTD,double &ksqrVAR,double &betachi_0,double &betahi_0,double &betalo_0);

//the following open the appropriate file depending on the sample size

int sampleno=b;
if(sampleno==1)
{outfile.open("chi_50.doc",ios::out);
outf.open("normal_50.doc",ios::out);
outfile2.open("chi2_50.doc",ios::out);
outf2.open("normal2_50.doc",ios::out);
}//end of if==1
if(sampleno==2)
{outfile.open("chi_100.doc",ios::out);
outf.open("normal_100.doc",ios::out);
outfile2.open("chi2_100.doc",ios::out);
outf2.open("normal2_100.doc",ios::out);
}//end of if==2
if(sampleno==3)
{outfile.open("chi_200.doc",ios::out);
outf.open("normal_200.doc",ios::out);
outfile2.open("chi2_200.doc",ios::out);
outf2.open("normal2_200.doc",ios::out);
}//end of if==3
if(sampleno==4)
{outfile.open("chi_300.doc",ios::out);
outf.open("normal_300.doc",ios::out);
```

```
outfile2.open("chi2_300.doc",ios::out);
outf2.open("normal2_300.doc",ios::out);
}//end of if==4
if(sampleno==5)
{outfile.open("chi_400.doc",ios::out);
outf.open("normal_400.doc",ios::out);
outfile2.open("chi2_400.doc",ios::out);
outf2.open("normal2_400.doc",ios::out);
}//end of if==5
if(sampleno==6)
{outfile.open("chi_500.doc",ios::out);
outf.open("normal_500.doc",ios::out);
outfile2.open("chi2_500.doc",ios::out);
outf2.open("normal2_500.doc",ios::out);
}//end of if==6
if(sampleno==7)
{outfile.open("chi_600.doc",ios::out);
outf.open("normal_600.doc",ios::out);
outfile2.open("chi2_600.doc",ios::out);
outf2.open("normal2_600.doc",ios::out);
}//end of if==7
if(sampleno==8)
{outfile.open("chi_800.doc",ios::out);
outf.open("normal_800.doc",ios::out);
outfile2.open("chi2_800.doc",ios::out);
outf2.open("normal2_800.doc",ios::out);
}//end of if==8
if(sampleno==9)
{outfile.open("chi_1000.doc",ios::out);
outf.open("normal_1000.doc",ios::out);
outfile2.open("chi2_1000.doc",ios::out);
outf2.open("normal2_1000.doc",ios::out);
}//end of if==9
if(sampleno==10)
{outfile.open("chi_3000.doc",ios::out);
outf.open("normal_3000.doc",ios::out);
outfile2.open("chi2_3000.doc",ios::out);
outf2.open("normal2_3000.doc",ios::out);
}//end of if==10
if(sampleno==11)
{outfile.open("chi_5000.doc",ios::out);
outf.open("normal_5000.doc",ios::out);
outfile2.open("chi2_5000.doc",ios::out);
outf2.open("normal2_5000.doc",ios::out);
}//end of if==11
if(sampleno==12)
{outfile.open("chi_7000.doc",ios::out);
outf.open("normal_7000.doc",ios::out);
outfile2.open("chi2_7000.doc",ios::out);
```

```
outf2.open("normal2_7000.doc",ios::out);
}//end of if==12
if(sampleno==13)
{outfile.open("chi_9000.doc",ios::out);
outf.open("normal_9000.doc",ios::out);
outfile2.open("chi2_9000.doc",ios::out);
outf2.open("normal2_9000.doc",ios::out);
}//end of if==13
if(sampleno==14)
{outfile.open("chi_10000.doc",ios::out);
outf.open("normal_10000.doc",ios::out);
outfile2.open("chi2_10000.doc",ios::out);
outf2.open("normal2_10000.doc",ios::out);
}//end of if==14

//the following declare the variables to be used throughout the program

double TotControls=0;
double TotCases=0;
    double betachi_0=0;
    double betahi_0=0;
    double betalo_0=0;

    double para_ksqr0=0;
    double para_ksqr05=0;
    double para_ksqr1=0;
    double para_ksqr2=0;
    double para_ksqr3=0;
    double para_ksqr4=0;
    double para_ksqr5=0;
    double para_ksqr625=0;
    double para_ksqr75=0;
    double para_ksqr8=0;
    double para_ksqr9=0;

//the following declare the arrays to hold the statistics calculated

double ksqrlist_0[reps];                          //array for the ksqr for each rep
double ksqrlist_0Bias[reps];                      //array for the bias of ksqr for each rep
double ksqrlist_hoBias[reps];                     //array for the ksqrb for each rep
double hisksqrlist_0[reps];                       //array for the ksqrb for each rep
double ksqrhislist_hoBias[reps];
double logksqrlist_0[reps];                       //array for the theoretical ln ksqr for each rep
double ksqrloglist_hoBias[reps];
double chisqr_0[reps];                            //array for the chi-square stat for each rep
double logavgksqrlist_0[reps];                    //array for the sample ln ksqr for each rep

//the above is repeated for each population of ksqr

double ksqrlist_05[reps];
```

```
double ksqrlist_05Bias[reps];
double hisksqrlist_05[reps];
double logksqrlist_05[reps];
double logavgksqrlist_05[reps];
double chisqr_05[reps];
double ksqrlist_1[reps];
double ksqrlist_1Bias[reps];
double hisksqrlist_1[reps];
double logksqrlist_1[reps];
double logavgksqrlist_1[reps];
double chisqr_1[reps];
double ksqrlist_2[reps];
double ksqrlist_2Bias[reps];
double hisksqrlist_2[reps];
double logksqrlist_2[reps];
double logavgksqrlist_2[reps];
double chisqr_2[reps];
double ksqrlist_3[reps];
double ksqrlist_3Bias[reps];
double hisksqrlist_3[reps];
double logksqrlist_3[reps];
double logavgksqrlist_3[reps];
double chisqr_3[reps];
double ksqrlist_4[reps];
double ksqrlist_4Bias[reps];
double hisksqrlist_4[reps];
double logksqrlist_4[reps];
double logavgksqrlist_4[reps];
double chisqr_4[reps];
double ksqrlist_5[reps];
double ksqrlist_5Bias[reps];
double hisksqrlist_5[reps];
double logksqrlist_5[reps];
double logavgksqrlist_5[reps];
double chisqr_5[reps];
double ksqrlist_625[reps];
double ksqrlist_625Bias[reps];
double hisksqrlist_625[reps];
double logksqrlist_625[reps];
double logavgksqrlist_625[reps];
double chisqr_625[reps];
double ksqrlist_8[reps];
double ksqrlist_8Bias[reps];
double hisksqrlist_8[reps];
double logksqrlist_8[reps];
double logavgksqrlist_8[reps];
double chisqr_8[reps];
double ksqrlist_75[reps];
double ksqrlist_75Bias[reps];
```

```
double hisksqrlist_75[reps];
double logksqrlist_75[reps];
double logavgksqrlist_75[reps];
double chisqr_75[reps];
double ksqrlist_9[reps];
double ksqrlist_9Bias[reps];
double hisksqrlist_9[reps];
double logksqrlist_9[reps];
double logavgksqrlist_9[reps];
double chisqr_9[reps];
double ksqrvar_0[reps];                          //array for variance of ksqr
double ksqrvar_ho[reps];
double ksqrlogvar_0[reps];                       //array for variance of ln ksqr
double ksqrlogvar_ho[reps];
double ksqrhisvar_0[reps];                       //array for variance of ksqrb
double ksqrhisvar_ho[reps];

//the above is repeated for the 11 populations of ksqr

double ksqrvar_05[reps];
double ksqrlogvar_05[reps];
double ksqrhisvar_05[reps];

double ksqrvar_1[reps]; //these are for the variance of each from
double ksqrlogvar_1[reps]; //each rep
double ksqrhisvar_1[reps];

double ksqrvar_2[reps]; //these are for the variance of each from
double ksqrlogvar_2[reps]; //each rep
double ksqrhisvar_2[reps];

double ksqrvar_3[reps]; //these are for the variance of each from
double ksqrlogvar_3[reps]; //each rep
double ksqrhisvar_3[reps];

double ksqrvar_4[reps]; //these are for the variance of each from
double ksqrlogvar_4[reps]; //each rep
double ksqrhisvar_4[reps];

double ksqrvar_5[reps]; //these are for the variance of each from
double ksqrlogvar_5[reps]; //each rep
double ksqrhisvar_5[reps];

double ksqrvar_625[reps]; //these are for the variance of each from
double ksqrlogvar_625[reps]; //each rep
double ksqrhisvar_625[reps];

double ksqrvar_8[reps]; //these are for the variance of each from
double ksqrlogvar_8[reps]; //each rep
```

```
double ksqrhisvar_8[reps];

double ksqrvar_75[reps]; //these are for the variance of each from
double ksqrlogvar_75[reps]; //each rep
double ksqrhisvar_75[reps];

double ksqrvar_9[reps]; //these are for the variance of each from
double ksqrlogvar_9[reps]; //each rep
double ksqrhisvar_9[reps];

double ksqrEXP=0;
double ksqrAVG=0;                                   //value to hold the average
double ksqrVAR=0;                                   //value to hold the theoretical variance
double ksqrSTD=0;                                   //value to hold the sample variance
    double ksqrBiasAVG=0;                           //value to hold the average bias
    double ksqrBiasAVGho=0;

//the following are the critical values needed for each sample
//size for the hypothesis testing

    if(sampleno==1)
    {ksqrBiasAVGho=ksqrho_50;}
    //end of if 1

    if(sampleno==2)
    {ksqrBiasAVGho=ksqrho_100;}
    //end of if 2

    if(sampleno==3)
    {ksqrBiasAVGho=ksqrho_200;}
    //end of if 3

    if(sampleno==4)
    {ksqrBiasAVGho=ksqrho_300;}
    //end of if 4

    if(sampleno==5)
    {ksqrBiasAVGho=ksqrho_400;}
    //end of if 5

    if(sampleno==6)
    {ksqrBiasAVGho=ksqrho_500;}
    //end of if 6

    if(sampleno==7)
    {ksqrBiasAVGho=ksqrho_600;}
    //end of if 7

    if(sampleno==8)
```

```
{ksqrBiasAVGho=ksqrho_800;}
//end of if 8

if(sampleno==9)
{ksqrBiasAVGho=ksqrho_1000;}
//end of if 9

if(sampleno==10)
{ksqrBiasAVGho=ksqrho_3000;}
//end of if 10

if(sampleno==11)
{ksqrBiasAVGho=ksqrho_5000;}
//end of if 11

if(sampleno==12)
{ksqrBiasAVGho=ksqrho_7000;}
//end of if 12
if(sampleno==13)
{ksqrBiasAVGho=ksqrho_9000;}
//end of if 13

double ksqrVARho=0;                 //value to hold the variance of critical value for ksqr
double ksqrhisAVGho=0;              //value to hold the mean of critical value for ksqr
double ksqrhisVARho=0;              //value to hold the variance of critical value for ksqrb
double ksqrlogEXPho=0;              //value to hold the mean of critical value for ln ksqr
double ksqrlogVARho=0;              //value to hold the variance of critical value for ln ksqr
double ksqrlist_0bias=0;                                    //holds value for the bias of ksqr
double ksqrlist_05bias=0;
double ksqrlist_1bias=0;
double ksqrlist_2bias=0;
double ksqrlist_3bias=0;
double ksqrlist_4bias=0;
double ksqrlist_5bias=0;
double ksqrlist_625bias=0;
double ksqrlist_75bias=0;
double ksqrlist_8bias=0;
double ksqrlist_9bias=0;

for(int j=0;j<reps;j++)
{
//The 1000 reps to calculate ksqr from 11 different populations
appgi ksqr_0;
appgi ksqr_05;
appgi ksqr_1;
appgi ksqr_2;
appgi ksqr_3;
appgi ksqr_4;
appgi ksqr_5;
```

```
appgi ksqr_625;
appgi ksqr_75;
appgi ksqr_8;
appgi ksqr_9;
    if(j==0)                                    //open files only once for each ksqr
    {

    ksqr_0.setup_files(b);
    ksqr_05.setup_files1(b);
    ksqr_1.setup_files2(b);
    ksqr_2.setup_files3(b);
    ksqr_3.setup_files4(b);
    ksqr_4.setup_files5(b);
    ksqr_5.setup_files6(b);
    ksqr_625.setup_files7(b);
    ksqr_75.setup_files8(b);
    ksqr_8.setup_files9(b);
    ksqr_9.setup_files10(b);

    }
ksqr_0.setup_ksqr();                        //sets up the ksqr using the appropriate
ksqr_05.setup_ksqr1();                                                      //file
ksqr_1.setup_ksqr2();
ksqr_2.setup_ksqr3();
ksqr_3.setup_ksqr4();
ksqr_4.setup_ksqr5();
ksqr_5.setup_ksqr6();
ksqr_625.setup_ksqr7();
ksqr_75.setup_ksqr8();
ksqr_8.setup_ksqr9();
ksqr_9.setup_ksqr10();
para_ksqr0=ksqr_0.para_ksqr();                  //this calculates each ksqr
para_ksqr05=ksqr_05.para_ksqr();
para_ksqr1=ksqr_1.para_ksqr();
para_ksqr2=ksqr_2.para_ksqr();
para_ksqr3=ksqr_3.para_ksqr();
para_ksqr4=ksqr_4.para_ksqr();
para_ksqr5=ksqr_5.para_ksqr();
para_ksqr625=ksqr_625.para_ksqr();
para_ksqr75=ksqr_75.para_ksqr();
para_ksqr8=ksqr_8.para_ksqr();
para_ksqr9=ksqr_9.para_ksqr();
ksqrlist_0bias=ksqr_0.para_bias();
ksqrlist_05bias=ksqr_05.para_bias();
ksqrlist_1bias=ksqr_1.para_bias();
ksqrlist_2bias=ksqr_2.para_bias();
ksqrlist_3bias=ksqr_3.para_bias();
ksqrlist_4bias=ksqr_4.para_bias();
ksqrlist_5bias=ksqr_5.para_bias();
```

```
ksqrlist_625bias=ksqr_625.para_bias();
ksqrlist_75bias=ksqr_75.para_bias();
ksqrlist_8bias=ksqr_8.para_bias();
ksqrlist_9bias=ksqr_9.para_bias();
ksqr_0.figure_ksqr();                                   //this calculates each ksqr
ksqr_05.figure_ksqr();
ksqr_1.figure_ksqr();
ksqr_2.figure_ksqr();
ksqr_3.figure_ksqr();
ksqr_4.figure_ksqr();
ksqr_5.figure_ksqr();
ksqr_625.figure_ksqr();
ksqr_75.figure_ksqr();
ksqr_8.figure_ksqr();
ksqr_9.figure_ksqr();
ksqr_0.figure_lnksqr();                                 //this calculates each lnksqr
ksqr_05.figure_lnksqr();
ksqr_1.figure_lnksqr();
ksqr_2.figure_lnksqr();
ksqr_3.figure_lnksqr();
ksqr_4.figure_lnksqr();
ksqr_5.figure_lnksqr();
ksqr_625.figure_lnksqr();
ksqr_75.figure_lnksqr();
ksqr_8.figure_lnksqr();
ksqr_9.figure_lnksqr();
ksqr_0.Bias();                                          //this calculates each bias
ksqr_05.Bias();
ksqr_1.Bias();
ksqr_2.Bias();
ksqr_3.Bias();
ksqr_4.Bias();
ksqr_5.Bias();
ksqr_625.Bias();
ksqr_75.Bias();
ksqr_8.Bias();
ksqr_9.Bias();
TotControls=ksqr_0.getTotalControls();
TotCases=ksqr_0.getTotalCases();
ksqrlist_0[j]=ksqr_0.get_ksqr();                        //this retrieves the ksqr's
ksqrlist_0Bias[j]=ksqr_0.get_Bias();
ksqrlist_hoBias[j]=ksqrlist_0Bias[j];
hisksqrlist_0[j]=ksqr_0.get_hisksqr();
ksqrhislist_hoBias[j]=0;
logksqrlist_0[j]=ksqr_0.get_lnksqr();                   //this is the Taylor avg
    ksqrloglist_hoBias[j]=logksqrlist_0[j];
    logavgksqrlist_0[j]=ksqr_0.get_avglnksqr();         //calc the sample avg
    chisqr_0[j]=ksqr_0.figurechisqr();
ksqrlist_05[j]=ksqr_05.get_ksqr();                      //this retrieves the ksqr's
```

```
ksqrlist_05Bias[j]=ksqr_05.get_Bias();
hisksqrlist_05[j]=ksqr_05.get_hisksqr();
logksqrlist_05[j]=ksqr_05.get_lnksqr();                    //this is the Taylor avg
logavgksqrlist_05[j]=ksqr_05.get_avglnksqr();               //calc the sample avg
chisqr_05[j]=ksqr_05.figurechisqr();
ksqrlist_1[j]=ksqr_1.get_ksqr();
ksqrlist_1Bias[j]=ksqr_1.get_Bias();
hisksqrlist_1[j]=ksqr_1.get_hisksqr();
logksqrlist_1[j]=ksqr_1.get_lnksqr();
logavgksqrlist_1[j]=ksqr_1.get_avglnksqr();
    chisqr_1[j]=ksqr_1.figurechisqr();
ksqrlist_2[j]=ksqr_2.get_ksqr();
ksqrlist_2Bias[j]=ksqr_2.get_Bias();
hisksqrlist_2[j]=ksqr_2.get_hisksqr();
logksqrlist_2[j]=ksqr_2.get_lnksqr();
logavgksqrlist_2[j]=ksqr_2.get_avglnksqr();
    chisqr_2[j]=ksqr_2.figurechisqr();
ksqrlist_3[j]=ksqr_3.get_ksqr();
ksqrlist_3Bias[j]=ksqr_3.get_Bias();
hisksqrlist_3[j]=ksqr_3.get_hisksqr();
logksqrlist_3[j]=ksqr_3.get_lnksqr();
logavgksqrlist_3[j]=ksqr_3.get_avglnksqr();
    chisqr_3[j]=ksqr_3.figurechisqr();
ksqrlist_4[j]=ksqr_4.get_ksqr();
ksqrlist_4Bias[j]=ksqr_4.get_Bias();
hisksqrlist_4[j]=ksqr_4.get_hisksqr();
logksqrlist_4[j]=ksqr_4.get_lnksqr();
logavgksqrlist_4[j]=ksqr_4.get_avglnksqr();
    chisqr_4[j]=ksqr_4.figurechisqr();
ksqrlist_5[j]=ksqr_5.get_ksqr();
ksqrlist_5Bias[j]=ksqr_5.get_Bias();
hisksqrlist_5[j]=ksqr_5.get_hisksqr();
logksqrlist_5[j]=ksqr_5.get_lnksqr();
logavgksqrlist_5[j]=ksqr_5.get_avglnksqr();
    chisqr_5[j]=ksqr_5.figurechisqr();
ksqrlist_625[j]=ksqr_625.get_ksqr();
ksqrlist_625Bias[j]=ksqr_625.get_Bias();
hisksqrlist_625[j]=ksqr_625.get_hisksqr();
logksqrlist_625[j]=ksqr_625.get_lnksqr();
logavgksqrlist_625[j]=ksqr_625.get_avglnksqr();
    chisqr_625[j]=ksqr_625.figurechisqr();
ksqrlist_8[j]=ksqr_8.get_ksqr();
ksqrlist_8Bias[j]=ksqr_8.get_Bias();
hisksqrlist_8[j]=ksqr_8.get_hisksqr();
logksqrlist_8[j]=ksqr_8.get_lnksqr();
logavgksqrlist_8[j]=ksqr_8.get_avglnksqr();
    chisqr_8[j]=ksqr_8.figurechisqr();
ksqrlist_75[j]=ksqr_75.get_ksqr();
ksqrlist_75Bias[j]=ksqr_75.get_Bias();
```

```
hisksqrlist_75[j]=ksqr_75.get_hisksqr();
logksqrlist_75[j]=ksqr_75.get_lnksqr();
logavgksqrlist_75[j]=ksqr_75.get_avglnksqr();
    chisqr_75[j]=ksqr_75.figurechisqr();
ksqrlist_9[j]=ksqr_9.get_ksqr();
ksqrlist_9Bias[j]=ksqr_9.get_Bias();
hisksqrlist_9[j]=ksqr_9.get_hisksqr();
logksqrlist_9[j]=ksqr_9.get_lnksqr();
logavgksqrlist_9[j]=ksqr_9.get_avglnksqr();
chisqr_9[j]=ksqr_9.figurechisqr();
ksqrvar_0[j]=ksqr_0.calcksqrVAR();                    //calc and return the var of ksqr
ksqrvar_ho[j]=ksqrvar_0[j];
ksqrhisvar_0[j]=ksqr_0.calchisksqrVAR();
    ksqrhisvar_ho[j]=ksqrhisvar_0[j];
    ksqrlogvar_0[j]=ksqr_0.get_varlnksqr();
    ksqrlogvar_ho[j]=ksqrlogvar_0[j];

ksqrvar_05[j]=ksqr_05.calcksqrVAR();                  //calc and return the var of ksqr
ksqrhisvar_05[j]=ksqr_05.calchisksqrVAR();
ksqrlogvar_05[j]=ksqr_05.get_varlnksqr();

ksqrvar_1[j]=ksqr_1.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_1[j]=ksqr_1.calchisksqrVAR();
ksqrlogvar_1[j]=ksqr_1.get_varlnksqr();

ksqrvar_2[j]=ksqr_2.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_2[j]=ksqr_2.calchisksqrVAR();
ksqrlogvar_2[j]=ksqr_2.get_varlnksqr();

ksqrvar_3[j]=ksqr_3.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_3[j]=ksqr_3.calchisksqrVAR();
ksqrlogvar_3[j]=ksqr_3.get_varlnksqr();

ksqrvar_4[j]=ksqr_4.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_4[j]=ksqr_4.calchisksqrVAR();
ksqrlogvar_4[j]=ksqr_4.get_varlnksqr();

ksqrvar_5[j]=ksqr_5.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_5[j]=ksqr_5.calchisksqrVAR();
ksqrlogvar_5[j]=ksqr_5.get_varlnksqr();

ksqrvar_625[j]=ksqr_625.calcksqrVAR();                //calc and return the var of ksqr
ksqrhisvar_625[j]=ksqr_625.calchisksqrVAR();
ksqrlogvar_625[j]=ksqr_625.get_varlnksqr();

ksqrvar_8[j]=ksqr_8.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_8[j]=ksqr_8.calchisksqrVAR();
ksqrlogvar_8[j]=ksqr_8.get_varlnksqr();
```

```
ksqrvar_75[j]=ksqr_75.calcksqrVAR();                    //calc and return the var of ksqr
ksqrhisvar_75[j]=ksqr_75.calchisksqrVAR();
ksqrlogvar_75[j]=ksqr_75.get_varlnksqr();

ksqrvar_9[j]=ksqr_9.calcksqrVAR();                      //calc and return the var of ksqr
ksqrhisvar_9[j]=ksqr_9.calchisksqrVAR();
ksqrlogvar_9[j]=ksqr_9.get_varlnksqr();

}//end of for

//this writes to the headings to the table
     outfile<<"THIS IS FOR ONE KSQR"<<endl<<endl;
     outfile<<"OURS–Assuming Chi-Square Distribution BOTH distributed
MULTINOMIAL"
<<endl;
outfile<<"HIS–Assuming Chi-Square Distribution CONTROL distributed
UNIFORM"<<endl;
outfile<<" trans/CASE MULTINOMIAL"<<endl;
outfile<<"LOG–Assuming Chi-Square LOG of KSQR"<<endl<<endl;
outfile<<"reps= "<<reps<<endl;
outfile<<" n = "<<TotControls<<endl;
outfile<<" m = "<<TotCases<<endl;
outfile<<"Ho: ksqr= "<<ksqr0_valHo<<endl;
outfile<<"Ho: lnksqr= "<<ksqr0_lnvalHo<<endl;
outfile<<endl;
outf<<"THIS IS FOR ONE KSQR"<<endl<<endl;
     outf<<"OURS–Assuming NORMAL Distribution BOTH distributed MULTINOMIAL"
<<endl;
outf<<"HIS–Assuming NORMAL Distribution CONTROL distributed UNIFORM"<<endl;
outf<<" trans/CASE MULTINOMIAL"<<endl;
outf<<"LOG–Assuming NORMAL– LOG of KSQR"<<endl<<endl;
outf<<"reps= "<<reps<<endl;
outf<<" n = "<<TotControls<<endl;
outf<<" m = "<<TotCases<<endl;
outf<<"Ho: ksqr= "<<ksqr0_valHo<<endl;
outf<<"Ho: lnksqr= "<<ksqr0_lnvalHo<<endl;
outf<<endl;

outfile<<endl;
outfile<<" OURS "<<endl;
outfile<<" sample "<<" sample "<<" taylor "<<endl;
//" "<<
//" sample "<<" sample "<<
//" taylor "<<" "<<endl;
outfile<<"ksqr bias "<<" avg "<<" var "<<" var "<<" power "<<endl<<endl;
//" avg "<<" var "<<
//" var "<<" power "<<endl<<endl;
outf<<endl;
outf<<" OURS "<<endl;
```

```
outf<<" sample "<<" sample "<<" taylor "<<
" below "<<" above "<<endl;//" sample "<<" sample "<<
outf<<"ksqr bias"<<" avg "<<" var "<<" var "
<<" power "
<<" power "<<endl;//
double ksqrlnname0=ksqr0_lnvalHo;
double ksqrname0=para_ksqr0;
double ksqrlnname05=ksqr05_lnvalHo;
double ksqrname05=para_ksqr05;
double ksqrlnname1=ksqr1_lnvalHo;
double ksqrname1=para_ksqr1;
double ksqrlnname2=ksqr2_lnvalHo;
double ksqrname2=para_ksqr2;
double ksqrlnname3=ksqr3_lnvalHo;
double ksqrname3=para_ksqr3;
double ksqrlnname4=ksqr4_lnvalHo;
double ksqrname4=para_ksqr4;
double ksqrlnname5=ksqr5_lnvalHo;
double ksqrname5=para_ksqr5;
double ksqrlnname8=ksqr8_lnvalHo;
double ksqrname8=para_ksqr8;
double ksqrlnname9=ksqr9_lnvalHo;
double ksqrname9=para_ksqr9;
double ksqrlnname625=ksqr625_lnvalHo;
double ksqrname625=para_ksqr625;
double ksqrlnname75=ksqr75_lnvalHo;
double ksqrname75=para_ksqr75;

int q=1;
for(q=1;q<25;)
{                                   //here the function calc_stats is called 11 times for each population
calc_stats(ksqrlist_0Bias,ksqrlist_0,ksqrvar_0,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_0bias,
logksqrlist_0,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname0,ksqrname0,
hisksqrlist_0,ksqrhisvar_0,
logavgksqrlist_0,ksqrlogvar_0,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_05Bias,ksqrlist_05,ksqrvar_05,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_05bias,
logksqrlist_05,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname05,ksqrname05,
```

```
hisksqrlist_05,ksqrhisvar_05,
logavgksqrlist_05,ksqrlogvar_05,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_1Bias,ksqrlist_1,ksqrvar_1,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_1bias,
logksqrlist_1,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname1,ksqrname1,
hisksqrlist_1,ksqrhisvar_1,
logavgksqrlist_1,ksqrlogvar_1,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_2Bias,ksqrlist_2,ksqrvar_2,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_2bias,
logksqrlist_2,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname2,ksqrname2,
hisksqrlist_2,ksqrhisvar_2,
logavgksqrlist_2,ksqrlogvar_2,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_3Bias,ksqrlist_3,ksqrvar_3,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_3bias,
logksqrlist_3,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname3,ksqrname3,
hisksqrlist_3,ksqrhisvar_3,
logavgksqrlist_3,ksqrlogvar_3,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_4Bias,ksqrlist_4,ksqrvar_4,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_4bias,
```

```
logksqrlist_4,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname4,ksqrname4,
hisksqrlist_4,ksqrhisvar_4,
logavgksqrlist_4,ksqrlogvar_4,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_5Bias,ksqrlist_5,ksqrvar_5,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_5bias,
logksqrlist_5,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname5,ksqrname5,
hisksqrlist_5,ksqrhisvar_5,
logavgksqrlist_5,ksqrlogvar_5,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_625Bias,ksqrlist_625,ksqrvar_625,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_625bias,
logksqrlist_625,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname625,ksqrname625,
hisksqrlist_625,ksqrhisvar_625,
logavgksqrlist_625,ksqrlogvar_625,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_75Bias,ksqrlist_75,ksqrvar_75,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_75bias,
logksqrlist_75,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname75,ksqrname75,
hisksqrlist_75,ksqrhisvar_75,
logavgksqrlist_75,ksqrlogvar_75,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
```

```
q=q+1;
calc_stats(ksqrlist_8Bias,ksqrlist_8,ksqrvar_8,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_8bias,
logksqrlist_8,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname8,ksqrname8,
hisksqrlist_8,ksqrhisvar_8,
logavgksqrlist_8,ksqrlogvar_8,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
calc_stats(ksqrlist_9Bias,ksqrlist_9,ksqrvar_9,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrlist_9bias,
logksqrlist_9,
ksqrEXP,betachi_0,
betahi_0,betalo_0,ksqrlnname9,ksqrname9,
hisksqrlist_9,ksqrhisvar_9,
logavgksqrlist_9,ksqrlogvar_9,q,ksqrvar_ho,ksqrlogvar_ho,
ksqrhisvar_ho,ksqrlist_hoBias,ksqrhislist_hoBias,ksqrloglist_hoBias,
ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,
ksqrlogVARho);
q=q+1;
}//end of for q<25
q=1;

//close the files that were open earlier in the program
outfile.close();
outf.close();
cases.close();
controls.close();
cases1.close();
controls1.close();
cases2.close();
controls2.close();
cases3.close();
controls3.close();
cases4.close();
controls4.close();
cases5.close();
controls5.close();
cases6.close();
controls6.close();
cases7.close();
controls7.close();
```

```
cases8.close();
controls8.close();
cases9.close();
controls9.close();
cases10.close();
controls10.close();
}//end of function

//this function calculates the sample average and variance

void trial(double ksqrlist_0Bias[],double ksqrlist_0[],double ksqrvar_0[],
double &ksqrVAR,double &ksqrAVG,double &ksqrSTD,double &ksqrBiasAVG,
double logksqrlist_0[],
double &ksqrEXP,double &ksqrBiasAVGho,double &ksqrhisAVGho,
double &ksqrVARho,double &ksqrhisVARho,
double &ksqrlogEXPho,double &ksqrlogVARho,int q);
void trial(double ksqrlist_0Bias[],double ksqrlist_0[],
double ksqrvar_0[],
double &ksqrVAR,double &ksqrAVG,double &ksqrSTD,
double &ksqrBiasAVG,
double logksqrlist_0[],
double &ksqrEXP,double &ksqrBiasAVGho,double &ksqrVARho,
double &ksqrhisAVGho,double &ksqrhisVARho,
double &ksqrlogEXPho,double &ksqrlogVARho,int q)
{
ksqrAVG=0;
ksqrSTD=0;
ksqrVAR=0;
ksqrEXP=0;
// ksqrBiasAVG=0;
for(int j=0;j<reps;j++)
{
ksqrVAR=ksqrVAR+ksqrvar_0[j];
}//end of for, list for variances
ksqrVAR=ksqrVAR/reps;
if(q==1||q==12)
{
ksqrVARho=ksqrVAR;
ksqrhisVARho=ksqrVAR;
ksqrlogVARho=ksqrVAR;
}
for(int j=0;j<reps;j++)
{
ksqrAVG=ksqrAVG+ksqrlist_0[j];
ksqrEXP=ksqrEXP+logksqrlist_0[j];
//ksqrBiasAVG=ksqrBiasAVG+ksqrlist_0Bias[j];
}//end of for,
ksqrAVG=ksqrAVG/reps;
ksqrEXP=ksqrEXP/reps;
```

```
//ksqrBiasAVG=ksqrBiasAVG/reps;
if(q==1||q==12)
{
ksqrhisAVGho=ksqrAVG;
ksqrlogEXPho=ksqrEXP;
}
for(int j=0;j<reps;j++)
{
ksqrSTD=ksqrSTD+(ksqrlist_0[j]-ksqrAVG)*(ksqrlist_0[j]-ksqrAVG);
}//end of for makes numerator for std
ksqrSTD=ksqrSTD/(reps-1);
}//end of trial functon

//this function calculates the power of the test using the
//appropriate critical value defined at the beginning of the
//program

void powertable(double ksqrlist_0[],
double ksqrvar_0[],double &betachi_0,
double &betahi_0,double &betalo_0,double ksqrlist_hoBias[],
double &ksqrAVGho,double &ksqrVARho);
void powertable(double ksqrlist_0[],
double ksqrvar_0[],double &betachi_0,
double &betahi_0,double &betalo_0,double ksqrlist_hoBias[],
double &ksqrAVGho,double &ksqrVARho)
{betachi_0=0;
betahi_0=0;
betalo_0=0;
for(int j=0;j<reps;j++)
{
if(j==0||j==500||j==999)
{//outd<<"ksqrAVGho/ksqrVARho "<<ksqrAVGho<<" "<<ksqrVARho<<endl;
}//end of j==0
if(ksqrlist_0[j]>ksqrAVGho)
{betachi_0=betachi_0+1;
}//end of if
if(ksqrlist_0[j]<-sqrt(ksqrAVGho))
{betalo_0=betalo_0+1;
}//end of if
if(ksqrlist_0[j]>sqrt(ksqrAVGho))
{betahi_0=betahi_0+1;
}//end of if
}//end of for
betachi_0=betachi_0/reps;
betahi_0=betahi_0/reps;
betalo_0=betalo_0/reps;
}//end of function powertable
```

```
//this function formats the data for ksqr and writes to the table

void writetotable(double &ksqrBiasAVG,double &ksqrname1,
double &ksqrAVG,double &ksqrSTD,double &ksqrVAR,
double &betachi_0,double &betahi_0,double &betalo_0);
void writetotable(double &ksqrBiasAVG,double &ksqrname1,
double &ksqrAVG,double &ksqrSTD,double &ksqrVAR,
double &betachi_0,double &betahi_0,double &betalo_0)
{
outtest<<"this is ksqrAVG during writtable "<<ksqrAVG<<endl;
outfile<<ksqrname1<<" "<<ksqrBiasAVG<<" "
<<ksqrAVG
<<setw(12)<<ksqrSTD<<setw(12)<<ksqrVAR<<setw(12)
<<betachi_0<<endl;
outf<<ksqrname1<<" "<<ksqrBiasAVG<<" "
<<ksqrAVG
<<setw(12)<<ksqrSTD<<setw(12)<<ksqrVAR<<setw(12)
<<betalo_0<<setw(10)<<betahi_0<<endl;
}//end of funtion writetotable

//this function calls all of the functions above for each population
//of ksqr

void calc_stats(double ksqrlist_0Bias[],double ksqrlist_0[],double ksqrvar_0[],
double &ksqrVAR,double &ksqrAVG,double &ksqrSTD,double &ksqrBiasAVG,
double logksqrlist_0[],
double &ksqrEXP,double &betachi_0,
double &betahi_0,double &betalo_0,
double ksqrlnname1,double ksqrname1,
double hisksqrlist_0[],double ksqrhisvar_0[],
double logavgksqrlist_0[],double ksqrlogvar_0[],int q,
double ksqrvar_ho[],double ksqrlogvar_ho[],
double ksqrhisvar_ho[],double ksqrlist_hoBias[],
double ksqrhislist_hoBias[],double ksqrloglist_hoBias[],
double &ksqrBiasAVGho,double &ksqrVARho,
double &ksqrhisAVGho,double &ksqrhisVARho,double &ksqrlogEXPho,
double &ksqrlogVARho);


void calc_stats(double ksqrlist_0Bias[],
double ksqrlist_0[],double ksqrvar_0[],
double &ksqrVAR,double &ksqrAVG,
double &ksqrSTD,double &ksqrBiasAVG,
double logksqrlist_0[],
double &ksqrEXP,double &betachi_0,
double &betahi_0,double &betalo_0,
double ksqrlnname1,
double ksqrname1,double hisksqrlist_0[],
double ksqrhisvar_0[],
```

```
double logavgksqrlist_0[],
double ksqrlogvar_0[],int q,
double ksqrvar_ho[],double ksqrlogvar_ho[],
double ksqrhisvar_ho[],double ksqrlist_hoBias[],
double ksqrhislist_hoBias[],
double ksqrloglist_hoBias[],
double &ksqrBiasAVGho,double &ksqrVARho,
double &ksqrhisAVGho,double &ksqrhisVARho,
double &ksqrlogEXPho,
double &ksqrlogVARho)
{
void trial(double ksqrlist_0Bias[],double ksqrlist_0[],double ksqrvar_0[],
double &ksqrVAR,double &ksqrAVG,double &ksqrSTD,double &ksqrBiasAVG,
double logksqrlist_0[],
double &ksqrEXP,double &ksqrBiasAVGho,double &ksqrVARho,
double &ksqrhisAVGho,double &ksqrhisVARho,
double &ksqrlogEXPho,double &ksqrlogVARho,int q);

void powertable(double ksqrlist_0[],
double ksqrvar_0[],double &betachi_0,
double &betahi_0,double &betalo_0,double ksqrlist_hoBias[],
double &ksqrAVGho,double &ksqrVARho);

void writetotable(double &ksqrBiasAVG,double &ksqrname1,
double &ksqrAVG,double &ksqrSTD,double &ksqrVAR,
double &betachi_0,double &betahi_0,double &betalo_0);
if(q<12)
{
trial(ksqrlist_0Bias,ksqrlist_0,ksqrvar_0,
ksqrVAR,ksqrAVG,ksqrSTD,ksqrBiasAVG,
logksqrlist_0,
ksqrEXP,ksqrBiasAVGho,ksqrVARho,
ksqrhisAVGho,ksqrhisVARho,
ksqrlogEXPho,ksqrlogVARho,q);

powertable(ksqrlist_0,
ksqrvar_ho,betachi_0,
betahi_0,betalo_0,ksqrlist_hoBias,ksqrBiasAVGho,ksqrVARho);

writetotable(ksqrBiasAVG,ksqrname1,
ksqrAVG,ksqrSTD,ksqrVAR,
betachi_0,betahi_0,betalo_0);

}//end of function calcstats

//This fortran program generates multinomial random variates using the
//parameter values designated in the program. The subroutine
//used is RNMTN.
// EXEC VSF2CLG,PARM.GO='NOXUFLOW'
//FORT.SYSIN DD *
```

```
INTEGER K, LDIR
PARAMETER (K=5, LDIR=1000)
INTEGER I, IR(LDIR,K), ISEED, J, N, NOUT, NR
REAL P(K)
EXTERNAL RNMTN, RNSET, UMACH
CALL UMACH (2, NOUT)
N = 50
P(1) =0.1
P(2) =0.16
P(3) =0.31
P(4) =0.27
P(5) =0.16
NR = 1000
ISEED = 234651
CALL RNSET (ISEED)
CALL RNMTN (NR, N, K, P, IR, LDIR)
WRITE (NOUT,10001) ((IR(I,J),J=1,K),I=1,NR)
10001 FORMAT (' 50 .05 : ', 5I7, /, (30X,5I7))
END
//LKED.SYSLIB DD DSN=DA.IMSL.LIBRARY.IMSL20,DISP=SHR
//
```

BIBLIOGRAPHY

Bain, Lee J., and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Boston: PWS Publishers, 1987.

Beaglehole, R., R. Bonita, and T. Kjellstrom. *Basic Epidemiology*. Geneva: World Health Organization, 1993.

Begg, Colin B., Jaya M. Satagopan, and Marianne Berwick. "A New Strategy for Evaluating the Impact of Epidemiologic Risk Factors for Cancer With Application to Melanoma." *Journal of the American Statistical Association* 93, no. 442 (1998): 415-426.

Bishop, Yvonne M. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press, 1975.

Bruzzi, P., S.B. Green, D.P. Byar, L.A. Brinton, and C. Schairer. "Estimating the Population Attributable Risk For Multiple Risk Factors Using Case-Control Data." *American Journal of Epidemiology* 122 (1985): 904-914.

Bury, Karl V. *Statistical Models in Applied Science*. New York: John Wiley and Sons, 1975.

Chase, Gerald R. "On the Chi-Square Test when the Parameters Are Estimated Independently of the Sample." *Journal of the American Statistical Association* 67, no. 339 (Sept 1972): 609-611.

Chernoff, Herman. "On the Distribution of the Likelihood Ratio." *Annuals of Mathematical Statistics* 25 (Sept 1954): 573-78.

Chernoff, Herman, and E. L. Lehmann. "The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit." *Annuals of Mathematical Statistics* 25 (Sept 1954): 579-86.

Cornfield, J. "A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast and cervix." *Journal of the National Cancer Institute* 11 (1951): 1269-1275.

Cramer, Harald. *Mathematical Methods of Statistics*. Princeton: Princeton University Press, 1945.

Dean, Angela, and Daniel Voss. *Design and Analysis of Experiments*. New York: Springer-Verlag, 1999.

Deitel, H. M. and P. J. Deitel. *C++ How to Program*, 2nd ed. Upper Saddle River: Prentice Hall, 1998.

Gupta, Prakash Chandra. "Estimation of Relative Risk and Attributable Risk from Epidemiological Studies." Ph.D. diss., Johns Hopkins University, 1975.

Hogg, Robert V., and Allen T. Craig. *Introduction to Mathematical Statistics*, 5th ed. Upper Saddle River: Prentice Hall, 1995.

IMSL. *The International Mathematical and Statistical Library Reference Manual.* IMSL Inc. Houston: (1987).

Johnson, N.L., and S. Kotz. *Continuous Univariate distributions*, Vol. 2. Boston: Houghton Mifflin Company, 1970.

Kendall, Maurice, and Alan Stuart. *The Advanced Theory of Statistics*, Vol. 2, 4th ed. New York: Macmillan Publishing Company, 1979.

Kleinbaum, David G., Lawrence L. Kupper, and Hal Morgenstern. *Epidemiologic Research*. New York: Van Nostrand Reinhold, 1982.

Kupper, L.L., J.M. Karon, D.G. Kleinbaum, H. Morgenstern, and D.K. Lewis. "Matching in Epidemiologic Studies: Validity and Efficiency Considerations." *Biometrics* 37 (1981): 271-291.

Last, J. M. *A Dictionary of Epidemiology*, 2nd ed. Oxford: Oxford University Press. 1988.

Levin, M. L. "The Occurrence of Lung Cancer in Man." *Acta Unio Internationalis Contra Cancrum* 19 (1953): 531-541.

Miettinen, O.S. "Components of the Crude Risk Ratio." *American Journal of Epidemiology*" 96 (1972): 168-172.

Miettinen, O.S. "Proportion of Disease Caused or Prevented by a Given Exposure, Trait or Intervention." *American Journal of Epidemiology* 99, no. 5 (1974): 325-332.

_____. "Confounding and Effect Modification." *American Journal of Epidemiology* 100, no. 5 (1974): 351-353.

_____. "Estimability and Estimation in Case-Referent Studies." *American Journal of Epidemiology* 103, no. 2 (1976): 226-236.

Mitra, Sujit Kumar. "On the Limiting Power Function of the Frequency Chi-Square Test." *Annuals of Statistics* (1958): 1221-1233.

Murthy, V. K., and A. V. Gafarian. "Limiting Distributions of Some Variations of the Chi-Square Statistic." *The Annuals of Mathematical Statistics* 41, no. 1 (1970): 188-202.

Patnaik, P. B. "The Power Function of the Test for the Difference between Two Proportions in a 2 × 2 Table." *Biometrica* 35 (1948): 157-175.

_____. "The Non-Central χ^2 and F distributions and their applications." *Biometrica* 36 (1949): 202-232.

Ross, Sheldon. *A First Course in Probability*. Upper Saddle River: Prentice Hall, Inc., 1998.

Rothman, K. J. *Modern Epidemiology*. Boston: Little, Brown, 1986.

Serfling, Robert J. *Approximation Theorems of Mathematical Statistics*. New York: Wiley and Sons, 1980.

Shafer, N. J. and J.A. Sullivan. "A Simulation Study of a Test for the Equality of the Coefficients of Variation." *Communications in Statistics-Simulation and Computation* 15, no. 3 (1986): 681-695.

Sillitto, G. P. "Note on Approximations to the Power Function of the 2 × 2 Comparative Trial." *Biometrika* 36 (1949): 347-352.

Stephan, Frederick F. "The Expected Value and Variance of the Reciprocal and other Negative Powers of a Positive Bernoullian Variate." *Mathematical Statistics* (June 1943): 50-61.

Stevens, W. L. "Mean and Variance of an Entry in a Contingency Table." *Biometrika* 38 (1951): 468-470.

Walter, S. D. "The distribution of Levin's measure of attributable risk." *Biometrika* 62, no. 2 (1975): 371-374.

_____. "Estimation and Interpretation of Attributable Risk in Health Research." *Biometrics* 32 (1976): 829-849.

_____. "Calculation of Attributable Risks From Epidemiological Data." *International Journal of Epidemiology* 7, no. 2 (June 1978): 175-82.

Whittemore, A. S. "Estimating attributable risk from case-control studies." *American Journal of Epidemiology* 117 (1983): 76-85.

_____. "Statistical Methods for Estimating Attributable Risk From Retrospective Data." *Statistics in Medicine* 1 (1982): 229-243.

VITA

Deborah Shepherd received a bachelor of music education from Northwestern State University in Natchitoches, Louisiana. She then joined the United States Air Force in order to perform in the Air Force band. While in the Air Force, Shepherd was stationed at Robins Air Force Base in Warner Robins, Georgia, then transferred to Clark Air Base in the Philippines.

After completing her enlistment in the Air Force, Shepherd went back to college to pursue a degree in mathematics from Southern Illinois University at Edwardsville, Illinois. From SIUE, she received another bachelor's degree in mathematics education and a master's degree in mathematics/statistics.

Shepherd worked as an actuarial assistant at General American Life Insurance Company in Saint Louis, Missouri, after receiving her master's degree. She moved back to Shreveport, Louisiana to be close to her family and taught high school math and physics. Shepherd returned to school at Louisiana Tech University in Ruston, Louisiana, to pursue a Ph.D. in computational analysis and modeling.