This is to certify that the

dissertation entitled

Factors influencing Pearson's chi-squared statistic's fit to its asymptotic distributions: Implications for sample size guidelines

presented by

Shelly Johann Naud

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Education

Major professor

Date 3/29/99

FACTORS INFLUENCING PEARSON'S CHI-SQUARED STATISTIC'S FIT TO ITS ASYMPTOTIC DISTRIBUTIONS: IMPLICATIONS FOR SAMPLE SIZE GUIDELINES

By

SHELLY JOHANN NAUD

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods
College of Education

1999

ABSTRACT

FACTORS INFLUENCING PEARSON'S CHI-SQUARED STATISTIC'S FIT TO ITS AYMPTOTIC DISTRIBUTIONS: IMPLICATIONS FOR SAMPLE SIZE GUIDELINES

By

SHELLY JOHAN NAUD

Recent sample size guidelines for Pearson's chi-squared statistic (X^2) have generally been based on simulation studies. These previous studies have mainly focused on the impact of small sample size on Type I error for a single test. A simulation study was carried out to evaluate the impact of small sample size on both Type I error and power approximation across four tests. It was found that power may be overestimated even though the sample size is large enough for the Type I error rate to be close to α. This problem is more serious for the test of independence than for the goodness of fit test.

A quantitative index, Pn, was proposed for contingency table tests. When sample size is larger than Pn, both Type I error and power of X^2 are fairly well approximated by the asymptotic distributions.

ACKNOWLEDGMENTS

I owe a debt of gratitude to the members of my dissertation committee: Dr. Alexander von Eye, committee chair, Dr. Betsy Jane Becker, advisor and guidance committee chair, Dr. Richard Houang, and Dr. Alka Indurkhya. I also wish to thank Dr. David Wagstaff for his extensive editorial comments on my first draft.

TABLE OF CONTENTS

LIST OF TABLES.. v

LIST OF FIGURES vi

LIST OF ABBREVIATIONS.. vii

INTRODUCTION 1

CHAPTER 1
THEORETICAL BACKGROUND 3

CHAPTER 2
SIMULATIONS.. 11

CHAPTER 3
THE FIRST QUESTION.. 16

CHAPTER 4
THE GOODNESS OF FIT TEST UNDER THE MULTINOMIAL SAMPLING MODEL 21

CHAPTER 5
THE GOODNESS OF FIT TEST UNDER THE PRODUCT MULTINOMIAL SAMPLING MODEL.. 33

CHAPTER 6
THE TEST OF INDEPENDENCE 40

CHAPTER 7
THE HOMOGENEITY TEST.. 51

CHAPTER 8
SUMMARY AND RECOMMENDATIONS 55

APPENDIX A 59

TABLES AND FIGURES... 70

BIBLIOGRAPHY 100

LIST OF TABLES

Table 4-1. Lower limits for the minimum expected cell frequency (e_{min}). 75
Table 4-2. Cell probabilities of tables generated for part 2 76

LIST OF FIGURES

Figure 2-1. Normal plots of the standardized residuals of the cell means.........70
Figure 3-1. Power plots for the test of independence.......................................71
Figure 3-2. Power plots for the homogeneity test...72
Figure 3-3. Power plots for the multinomial case of the goodness of fit test 73
Figure 3-4. Power plots for the product multinomial case of the goodness of fit test...74
Figure 4-1. Type I error rate versus sample size...78
Figure 4-2. Power plots: Rejection rate in percent versus effect size, w. 79
Figure 4-3. Rejection rates versus sample size for alternative hypotheses.......83
Figure 5-1. Differences in fit between the product multinomial and the multinomial cases of the goodness of fit test...84
Figure 5-2. Unequal sample sizes, 860 and 875 series 86
Figure 5-3. The number of small cell expectations within rows as a factor affecting power 88
Figure 5-4. Application problem...89
Figure 6-1. Type I error rate versus sample size, test of independence 90
Figure 6-2. Power plots, test of independence...91
Figure 6-3. Power plots for n = Pn...93
Figure 6-4. Differences between observed and expected power versus expected power...95
Figure 6-5. Application problem...96
Figure 7-1. Differences in fit between the homogeneity test and the test of independence...97

LIST OF ABBREVIATIONS

df Degrees of freedom.

e_i or e_{ij} Expected cell frequency (p. 2).

e_{min} A table's minimum expected cell frequency (p. 5).

ES Effect size (p. 5)

k Number of cells in a table.

H_0 Null hypothesis.

H_1 Alternate hypothesis.

n Sample size.

n5 The sample size where power is expected to be .5 for a large effect size (p. 5).

n8 The sample size where power is expected to be .8 for a moderate effect size (p. 5).

n(p) The number of cells with expected probabilities less than 1/k (p. 6).

o_i or o_{ij} Observed cell frequency (p. 2).

p_i or p_{ij} Expected cell probability (p. 2).

$p_{i.}$, $p_{.j}$ Marginal probabilities (p. 2).

p_{min} A table's minimum expected cell probability (p. 6).

Pn Sample size where the probability of getting a marginal total of zero is 1% (p. 43).

r Number of cells expectations less than 5 (p. 6).

R A global index evaluating the skewness of the distribution of cell expectations (p. 6).

w Cohen's index of effect size (p. 4).

X^2 Pearson's chi-squared statistic.

λ Noncentrality parameter (p. 3).

INTRODUCTION

Pearson's chi-squared statistic, X^2, first introduced in 1900, is currently widely known and used. Many of the researchers who used X^2 may not realize is that there is no consensus on sample size guidelines - available guidelines actually vary a great deal. Why is there such variablitiy? It is partly due to the different approaches for determining when an asymptotic distribution is a reasonable approximation. When sample sizes are small, the distribution of X^2 is a step function that cannot be well approximated by any continuous function. The older guidelines required that the distribution of X^2 be fairly smooth. To attain this criteria, sample sizes need to be large. Recent guidelines are generally based on simulation studies. As long as the actual Type I error rate is reasonably close to the nominal Type I error rate, α, the asymptotic distribution is considered adequate. The resulting sample size recommendations are considerably less stringent.

Though a considerable number of simulation studies have been done, the question of sample size has not been entirely resolved because additional factors complicate the problem. One such factor is the table's distribution of cell expectations. Tables where some of the expected cell frequencies are very small in comparison to the other cells apparently require different guidelines than table with uniform expectations.

This study proposes to address some of the gaps in the simulation research. One is related to the fact that a majority of the research has dealt with only one of the several tests that use X^2 as the test statistic. Although the asymptotic distributions of X^2 are the same across tests, the actual distribution of X^2 across tests is not necessarily similar when the sample size is small. This issue has not been studied systematically. A second issue addressed in this study is power. Although there have been studies on the impact of small sample sizes on power, these have had much less influence on sample size guidelines than the studies focusing on Type I error.

The comparison between tests is the focus of Chapter three. Each test is then considered individually in the following chapters. The first two chapters will cover theoretical and methodological issues.

In summary, this study will explore the behavior of Pearson's chi-squared statistic when the sample size is small and the table has a skewed distribution of expected cell frequencies. These are the conditions where the asymptotic distributions do not hold well. Both power and Type I error will be considered across different tests. Current recommendations for sample size will be evaluated based on these findings.

Chapter 1

THEORETICAL BACKGROUND

The first sections of this chapter will define the notation and terminology related to X^2, hypothesis testing and power, and some proposed indices. The sampling distributions and tests associated with categorical data are described in the last section.

Notation and formulas

The two-way frequency tables have I rows and J columns. The number of cells in the table is denoted by k with k = IJ. Marginal row and column probabilities, $p_{i.}$ and $p_{.j}$, are obtained by dividing the row and column totals, $n_{i.}$ and $n_{.j}$, by the total sample size, n (e.g., $p_{1.} = n_{1.}/n$). Depending on the sampling plan that is assumed to have generated the data, one or more of the marginal totals may be fixed or treated as constants. With such sampling plans, the marginal totals will used in some formulas instead of n.

The expected cell probabilities are denoted by p_{ij} with $p_{ij} = p_{i.}p_{.j}$. The expected cell frequencies (or expectations) are related to the cell probabilities: $e_{ij} = np_{ij}$. Each cell's count is referred to as the observed cell frequency (o_{ij}).

Pearson's chi-squared statistic provides a measure of the discrepancy between observed and expected cell frequencies:

$$X^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

If the expected values are close to the observed values, the value of X^2 is small; if the expected values are far from the observed values, the value of X^2 is large. Because the deviations are squared, the X^2 statistic gives more weight to observed cell frequencies that are much larger (or much smaller) than the expected cell frequencies.

When the null hypothesis is true, Pearson's X^2 is asymptotically distributed as the chi-square (χ^2). On the other hand, when the null hypothesis is false, the asymptotic distribution is the noncentral chi-square distribution. Both distributions have degrees of freedom (df) as a parameter. The noncentral chi-square distribution further depends on a second parameter

$$\lambda = n\sum\sum\frac{(p_{1ij} - p_{0ij})^2}{p_{0ij}}$$

where p_{0ij} refers to the cell probability under the null hypothesis (H_0) while p_{1ij} refers to the cell probability under the alternate hypothesis (H_1). Lambda increases in value as the two hypotheses become more discrepant, and it increases with the sample size. When λ is set to zero, the noncentral χ^2 is equivalent to the chi-square distribution.

Type I error, power, and Cohen's effect size index

When deciding whether to accept or reject the null hypothesis, researchers can make two types of error. They can reject the null hypothesis when it is true or they can fail to reject the null hypothesis when it is false. This former is referred to as Type I error and the probability of its occurrence is α. The second is a Type II error and its probability of occurrence is β. Power is the probability of rejecting H_0 when it is false; it is equal to $1 - \beta$.

Researchers have to balance the costs associated with the two errors. Choosing to make α very small decreases the risk of rejecting a null hypothesis that is true; however, power also decreases as a result. Choosing a relatively large α will result in a smaller β and, therefore, more power; however, this choice increases the risk of rejecting the null hypothesis when it is true. There is another alternative: Researchers can achieve an increase in power by increasing their sample size. In order to determine how large the sample should be, the researcher should have a reasonable estimate of the population effect size, ES. A small effect size indicates that the alternative hypothesis is not much different from the null hypothesis. Small differences are unlikely to be detected unless the sample size is large. On the other hand, a researcher can expect to detect large effect sizes with smaller samples. Estimates of effect size are determined from previous research or pilot studies whenever possible.

Cohen (1988, 1992) has defined a measure of effect size that is widely used. If previous findings are available, Cohen's index[1] can be calculated as follows: $w = \sqrt{\sum_i \sum_j \frac{(p_{1ij} - p_{0ij})^2}{p_{0ij}}}$. This index is closely related to the noncentrality parameter, $\lambda = nw^2$. Cohen provided the following guidelines for

[1] In his first edition of *Statistical Power Analysis for the Behavioral Sciences,* Cohen proposed a slightly different ES index: $e = \lambda/n = w^2$.

interpreting the values of w: 0.1 corresponds to a small effect size, 0.3 to a moderate effect size, while 0.5 is considered large. In practice, w is not likely to be greater than 0.9. Several effect size surveys have found the average w to be approximately 0.3, at least in the field of psychology (Haase et al., 1982; Cooper and Findley, 1982). Therefore, if a researcher lacks an empirically-based alternative hypothesis, setting w to 0.3 is a plausible alternative. Cohen suggests that power should be set at 0.8.

This study will evaluate the behavior of X^2 at two target sample sizes. The first, n5, is the sample size where the theoretical power is .5 for a large ES, i.e., n is determined after constraining the noncentrality parameter to be .5 and the noncentral χ^2 to be 0.5. If a researcher has a sample size equal to n5, he or she will have a 50-50 chance of detecting a large effect size. The second target sample size, n8, corresponds to a power of .8 for a medium ES (w = .3). N5 serves as the lower bound to sample sizes that may be considered by researchers while n8 reflects a reasonable goal for most research. The specific sample sizes that correspond to each target sample size are listed at the end of this chapter.

Measures of discrepancy

One problem that exists in the literature on categorical data is the lack of a quantitative index for describing tables where the expected cell frequencies are not all equal. Researchers often resort to qualitative descriptions such as "a highly skewed distribution of expected cell frequencies." Three quantitative indices are proposed here.

Many authors use the minimum expected cell frequency (e_{min}) as their criterion for indicating how discrepant the observed table is from a table where all cell frequencies are the same, i.e., the uniform table. Some researchers are also interested in the number of cells with small expectations (Cochran, 1952, 1954; Yarnold, 1970). In particular, Yarnold proposed r, the number of cells with an expected frequency of less than five. The disadvantage of using e_{min} or r is that they vary with the sample size. I propose two alternate indices that remain invariant: the minimum expected cell probability, p_{min}, and n(p), the number of cells that have probabilities less than 1/k, where k is the total number of cells. In a uniform table, $p_{min} = 1/k$ and $n(p) = 0$.

The third index used in the present study to indicate how discrepant an observed table is from the uniform table is a global index, $R = \Sigma\ 1/p_{ij}$. R is an element of three formulas for estimating the variance of X^2 (Pearson given in Lawal & Upton, 1980; Haldane given in Lawal, 1992; and Morris given in Koehler and Larntz, 1980). The use of R is of interest since it is a key component of the variance estimates, and the fit of X^2's distribution to its asymptotic distributions is thought to be related to the variance of X^2. When there are small cell expectations, the variance of X^2 can be much greater than the variance of χ^2 (Lawal, 1991).

Sampling distributions and tests for categorical data

There are two tests that are usually thought of whenever one deals with categorical data, namely the test of independence and the goodness of fit test. Researchers would conduct a test of independence to determine whether two variables are related, e.g., gender and level of job satisfaction. The degrees of freedom for X^2 is (I-1)(J-1). The expected cell frequencies are calculated from the marginal probabilities: $e_{ij} = np_{i.}p_{.j}$.

The procedure for the goodness of fit test differs from the test of independence in two ways. The expected cell probabilities are specified by the null hypothesis and the degrees of freedom is k-1. One application of this test, given by Pearson when he introduced his statistic in 1900 (cited in Agresti, 1990), is analyzing the outcomes from a roulette wheel. If the wheel shows no bias then each outcome has an equal probability of occurring, therefore, under the null hypothesis $e_i = n\ \pi = n\ (1/37)$. Only one subscript is used since these tables are one-dimensional.

In the two examples described above, the data are sampled from a single population. There are two possible sampling distributions, namely, the multinomial and the Poisson. The Poisson differs from the multinomial in that the sample size is not fixed; n itself has a Poisson distribution (Agresti, 1990).

It is possible to sample from more than one population. The relevant sampling distribution is the product multinomial. For the goodness of fit test, the degrees of freedom will be reduced by the number of groups sampled. Given i groups and J categories, df = I (J-1) = IJ - I = k - I. The corresponding contingency table test, i.e., the test of homogeneity, has the same degrees of freedom as the test of independence; both are constrained by the marginal totals.

This discussion of sampling distributions would not be complete without mentioning the hypergeometric sampling distribution. In this case, all of the marginal frequencies are fixed: $e_{ij} = n \pi_{i.} \pi_{j}$. Agresti (1990) maintains that the only appropriate test in this situation is Fisher's exact probability test, therefore it will not be considered in this study. Wickens (1989) presents other alternative that also are appropriate.

Agresti (1990, p. 39) uses an example to clarify the differences among the above sampling models. A two-way table is defined by seat-belt use (yes, no) and whether the driver survives the accident (yes, no). If the data include all reported accidents occurring on the Massachusetts turnpike in a year, then the cell frequencies are Poisson random variables. The cell observations have a multinomial distribution when a subset of the population is randomly chosen, say 100 accident reports. If the researcher decides to sample 50 drivers who didn't wear seat belts and 50 seat-belt users, then we have a product multinomial sampling distribution.

A different outcome needs to be chosen in order to illustrate the hypergeometric sampling distribution. Let's say the sample of accident reports (from the product multinomial case) are given to an expert who is asked to determine which 50 drivers were most likely to have worn seat belts. The expert's answers (likely, not likely) are compared to the actual data which were withheld from the expert (seat belt, no seat belt). The resulting 2 x 2 table will have marginal totals that are all fixed to equal 50 by design.

This study will focus on four tests for which Pearson's X^2 is appropriate. These tests are defined by the two dimensions that were described in this section, namely the sampling distribution and the method of calculativng the expected cell frequencies - the goodness of fit tests depend on H_0 while the contingency table tests depend on the marginal totals. The target sample sizes described earlier (n5 and n8) will vary depending on the tests as well as the table size. The degrees of freedom will vary also. These are listed in the following table.

	k	df	n5	n8
Goodness of fit tests				
Multinomial	4	3	24	122
	16	15	48	216
Product multinomial	4	2	20	108
	16	12	40	196
Contingency table tests				
Test of independence	4	1	16	88
(Multinomial)	16	9	40	176
Homogeneity test	4	1	16	88
(Product multinomial)	16	9	40	176

The distribution of X^2 for these four tests will be compared in Chapter 3 and considered separately in the following chapters.

Chapter 2

SIMULATIONS

The next five chapters will present the results from multiple simulations. The underlying procedures common to all are described in this chapter. Two related topics are treated in separate sections: confidence intervals and a description of the tables that are used in more than one chapter.

The simulation programs were written in UNIX SAS version 6.07 (SAS Institute Inc., 1990).

Four data sets were generated to assess the behavior of X^2 across tests. Using the same data eliminates variation in the generated data as a possible cause for any differences seen in the results.

The general strategy of the simulation programs was to partition the table by the predetermined cell probabilities (p_i). For example, the limits for cell 1 are $[0, p_1]$[2], the limits for cell 2 are $(p_1, p_2]$, and so on. Each generated random number, u, was then assigned to the cell for which $p_i \leq u < p_{i+1}$. The product multinomial case differed from the multinomial in that each row was treated as a separate table.

One critical aspect of simulations is the process used to generate the random numbers. The SAS uniform random number generator uses a prime

[2] The bracket is inclusive while the parenthesis is not: "[0" means "including zero" while "p_1)" means "up to, but not including, p_1."

modulus multiplicative generator with modulus 2^{31} - 1 and multiplier 397204094. This particular combination has been tested and found to be one of the better random number generators (Fishman and Moore, 1982). The programs were tested to see how well the generated data conformed to the target sampling distribution. The observed cell means were compared to their theoretical values: $d_{ij} = \bar{o}_{ij} - np_{ij}$. The standardized residuals, $r_{ij} = d_{ij} / \sigma_{ij}$[3], were plotted against the expected z scores, $z_{ij} = \Phi(\text{percentile rank}_{ij})$. These normal probability plots (Figure 2-1) are linear. The observed cell frequencies therefore follow the expected distribution.

The mean observed cell variances were compared to their theoretical values: $\sum\left(\frac{\overline{s_{ij}^2}}{\sigma_{ij}^2} - 1\right)$. The average deviation of the 96 cells from tables with a multinomial distribution is slightly below the expected value of 0 (-.0035) while it is slightly above 0 for the product multinomial case (.0046). The cell means and standard deviations are therefore both close to their expected values.

[3] The expected standard deviation depends on the sampling distribution and the number of replications. Given a multinomial sampling distribution and 1000 replications, $\sigma_{ij} = [np_{ij}(1-p_{ij})/1000]^{1/2}$ for the multinomial case. The standard deviation for the product multinomial case is $\sigma_{ij} = n_i p_{.j}(1-p_{.j})/1000]^{1/2}$.

Confidence intervals

Cochran (1952) and a number of other researchers have suggested the range .04 to .06 as acceptable lower and upper limits for observed Type I error rate when the nominal rate is .05. Other researchers have proposed a range that is more liberal (e.g., .03, .07; for example, Koehler and Larntz, 1980), and at least one researcher has proposed a range that is asymmetric, (.03, .06). Bradley et al. (1979) justified the latter by remarking that many researchers would accept a conservative bias. In actuality, these are tolerance limits and not confidence intervals since they were all set independently of the simulation. These ranges of varying widths do lead to differing interpretations of the behavior of X^2; a wider range obviously makes the X^2 appear to behave better than would a stringent one.

A 95% confidence interval is calculated based on the number of replications: $\theta \pm 1.96\ [(\alpha)(1 - \alpha)\ /\ (\text{number of replications})]^{1/2}$ where $\theta = \alpha$ for $w = 0$, and θ is the expected power for all other values of the effect size index. 1825 tables were generated for most of the simulations so that the resulting confidence interval would equal Cochran's limits (.4, .6).

Description of the tables

Two factors described in Chapter 1 were used to create the tables of set I, namely the global index, R, and the number of small cell expectations, n(p). The four values of R were chosen variance[4] of X^2 would range from being slightly discrepant from the theoretical variance to being two and a half time greater. The ratios of the Pearson estimate of the variance to the theoretical was kept the same for both table sizes. The following table lists the values of R and the corresponding estimates of the variance of X^2 for the two table sizes.

		Ratio of variance (X^2) to variance (χ^2)			
	k	7/6	1.5	2	2.5
R	4	32	52	82	112
Variance (X^2)	4	7	9	12	15
R	16	366	526	759	1006
Variance (X^2)	16	35	45	60	75

The tables of set II were created to evaluate the effect of changing marginal probabilities while holding p_{min} constant. Other tables were generated to deal with specific questions and are described in the appropriate chapters.

Detailed descriptions of set I and II tables are given in Appendix A. An alphanumeric code is used to identify these tables. The first symbol is a letter that represents the number of cells in the table: E, F, and S indicate that the table consists of 8, 4, and 16 cells respectively. Following the letter are the values of n(p) and the variance of X^2. For example, S435 refers to a sixteen-cell table with four small cells and the Pearson estimate variance of X^2 is 35.

[4] Pearson's estimate of the variance of X^2 is: $2(k - 1) + (R + k^2 - 2k - 2)/n$. I set n = 10 for four-cell tables and n = 16 for sixteen-cell tables. I prefer Pearson's formula to the others since it most closely matched the observed variance of X^2.

Because the tables in set II are variations of a specific table from set I, an additional letter is added to distinguish between the tables.

Chapter 3

THE FIRST QUESTION

Because the asymptotic distribution of Pearson's X^2 is the same across all tests whatever the underlying sampling model may be, there is a tendency to generalize the results from one test to all cases. Such generalizations may not be justified according to Cochran (1952, p. 326):

> Is the same X^2 test to be used for all cases [i.e., contingency tables with three underlying sampling distributions: multinomial, product multinomial, and hypergeometric]? In large samples there is no conflict, because X^2 has the same limiting distribution however the linear restrictions arise. This is not so in small samples, where the distribution of X^2 differs in the three cases.

A theoretical study of power substantiates this observation. Harkness and Katz (1964) found that both the hypergeometric and product multinomial cases of X^2 have more power than the multinomial (i.e., the test of independence). The hypergeometric case's superiority did not hold, though, when the marginal probabilities were skewed and n=20.

Few simulation studies have looked at more than one test. Roscoe and Byars (1971) considered the goodness of fit test and the homogeneity test (though the latter was referred to as the test of independence) and proposed sample size guidelines that are different for each test. Camilli and Hopkins (1978) considered both the homogeneity test and the test of independence and found the behavior of X^2 to be similar for both. However, neither of these studies used identical tables to assess the tests.

The above evidence suggest that the behavior of X^2 may indeed be different across tests when n is small but this issue has not been studied systematically. The differences in the behavior of X^2 across tests may not be a serious problem if the differences are small. Therefore the question is: How variable is the behavior of X^2 across tests when n is small?

Part 1. Type I error and power across tests when n=16

Methodology. The simulation programs and table specifications are described in Chapter 2. For this chapter, a subset of the set I tables were used, specifically the sixteen-cell tables with four small cell expectations (n(p) = 4). The sample size of 16 was determined based on the most liberal available guidelines, e.g., Koehler and Larntz (1980) for the goodness of fit test, Craddock and Flood, 1970, and Bradley et al., 1979, for the test of independence.

Results. The power plots for the test of independence (Figure 3-1, Panel a) and the homogeneity test (Figure 3-2, Panel a) appear very similar. The latter does have a smaller rejection rate when the effect size is small while power is greater for large ES but these differences are generally within the confidence limits (± .01) or not much larger. These two tests will be compared more extensively in Chapter 7. For now it suffices to say that the distribution of X^2 is similar for these two tests.

The multinomial and product multinomial cases of the goodness of fit test (Figures 3-3 and 3-4, Panels a) are also similar to each other. The observed power distributions are not as much alike as those of the contingency tables, but these two cases do not have the same expected distribution as they differ in the degrees of freedom. Therefore, a greater variability is to be expected. The product multinomial case does show a slightly more liberal trend. These two tests will be compared in chapter 5.

Marked differences are to be found between the contingency table tests and the goodness of fit tests. For the former, the Type I error rate (i.e., when w = 0) and power are both lower than predicted by the asymptotic distributions. The goodness of fit tests have a rejection rate that is greater than expected when the effect size is small while power tends to be overestimated for large ES. This overestimation, however, is not as dramatic as it is for the contingency table tests.

The simulations listed in the methodology section were focused on Type I error. The recommended sample size based on these studies is not sufficiently large for power to be well approximated by the noncentral χ^2 even for the table that is the least discrepant from the uniform (S435). The estimated power, however, is low - the maximum is .45 for w = .7. In other words, when the sample size is only 16, one is not likely to detect even a very large effect size. Therefore, from a practical point of view, these results are not of much interest. It may well be that power is well approximated by the noncentral χ^2 when n is large enough to detect a large or a moderate ES. If the asymptotic distributions have an acceptable fit to the actual distribution of X^2 when n is somewhat larger then we can ignore the erratic behaviors of X^2 noted in this section.

Part 2. Type I error and power across tests for larger n

Methodology. The same tables are used again although the sample sizes will correspond to n5 and n8. As defined in Chapter 2, n5 is the sample size that is large enough to detect a large effect size with a power of .5 while n8 is the sample size where X^2 is expected to detect a moderate ES with a power of .8. Again, $\alpha = .05$ and the 95% confidence limits are ± .01 of the theoretical values.

Results. The observed power of the test of independence is still seriously overestimated by the noncentral χ^2 distribution at n5 (Figure 3-3, Panel b). The actual power of X^2 is as low as 50% of the estimated power when the effect size is large. The fit between the actual and the theoretical distributions is much better at n8 (Figure 3-1, Panel c). At this sample size Pearson's X^2 has an observed power that is only .02 to .05 below that of the estimated power for the table with the most extreme cell probabilities (S475).

The power plots associated with the test of homogeneity follow the same trends (Figures 3-2, Panels b and c). The discrepancy between the observed and the theoretical power is actually less although the difference is too small to be discernible from the plots.

The power distributions of both goodness of fit tests are well approximated by the noncentral χ^2 distribution at n8 (Figures 3-3 and 3-4, Panels c). At n5 the Type I error rate is somewhat liberal and there is some overestimation for the larger ES, but both power plots show a reasonable fit to the noncentral χ^2 (Figures 3-3 and 3-4, Panels b).

To return to the question: How variable is the behavior of X^2 across tests when n is small? When n = 16 or n5, the lack of fit between the noncentral χ^2 and the actual power distributions of X^2 are most marked for the contingency table tests. The dissimilarity across tests appears to be minor when n = n8. When the sample size is large enough to detect a moderate effect size with adequate power, X^2 is well approximated by its asymptotic distributions for all four tests considered.

As the major differences found were between the goodness of fit cases on the one hand and the two contingency table tests on the other, This study will focus on the two multinomial cases.

Chapter 4

THE GOODNESS OF FIT TEST UNDER THE MULTINOMIAL SAMPLING MODEL

The earliest sample size recommendations for the goodness of fit test were based on the fact that Pearson used the multivariate normal distribution to approximate the multinomial distribution of the cells (Cochran, 1952). This approximation is valid only when expectations are large. It therefore became customary to recommend that all expected frequencies be at least 5 or even 10. Cochran proposed guidelines for assessing goodness of fit in the case of a unimodal distribution with only one or two small expectations. These guidelines were less stringent than those of his predecessors. He suggested that the minimum cell frequency could be as small as 0.5 when there was only one small e_i; that the minimum could be 1 when there were two cells with small e_i; and that all other cells should have frequencies of 5 or more. These guidelines are still cited although subsequent research, described below, has found them to be restrictive.

There is as yet no universally accepted set of guidelines, although a consensus has formed around the following findings (Roscoe and Byars, 1971; Moore, 1986; Read and Cressie, 1988):

1. X^2 has been found to be erratic when there is only one degree of freedom (Larntz, 1978). Roscoe and Byars recommend the exact binomial test in this situation.

2. When the expected cell probabilities are uniform, X^2 is robust for very small sample sizes (Wise, 1963). How small n can be is still disputed. Tate and Hyer (cited in Roscoe and Byars, 1971) suggest that e_i can be as small as 1. Koehler and Larntz (1980) suggest that the sample size, n, must be greater than $(10k)^{1/2}$ and no less than 10. The expected frequencies can become as small as .25 for large tables.

3. The distribution of X^2 is not well approximated by χ^2 when samples sizes are small and the expected cell frequencies are extremely different. Given an α of .05, Roscoe and Byars suggest all $e_i \geq 1$ when the departure from the uniform is moderate. For extreme departures, the minimum e_i should be 2. Koehler and Larntz (1980) suggest that their formula cited above can still be applied but the minimum n should be 15 when there is a departure from the uniform. They warn, however, that the Type I error rate will be inflated if there are many $e_i < 1$. Yarnold (1970) argues that when there are too many small cell expectations a distribution other than χ^2 should be used to approximate the distribution of X^2. He provided a lower bound for using χ^2: $e_{min} \geq 5r/k$ with $r = n(e_i < 5)$. This can be modified in order to calculated a sample size: $n \geq 5r/(kp_{min})$.

A few simulation studies have looked at the power of X^2. Hayman and Leone (1964), Slakter (1968), and Frosini (1978) showed that the power of X^2 is well approximated by the asymptotic distribution when the cell expectations are equal; but the approximation can be poor when there are some small e_i. Slakter recommended reducing the estimated power by 20% to get a better approximation of the actual power when n is less than 50.

Implications for researchers

Let's use an example to illustrate what happens when one applies the different guidelines given above. A statistician working for a state department wants to compare local statistics to the following national statistics for teachers' level of education.

Level of Education	Percentage
Less than Bachelor's	0.9
Bachelor's	51.3
Master's	44.9
Master's + 30 graduate credits	2.9

By Cochran's (1952, 1954) guidelines the sample size should be 112. Using Roscoe and Byars's (1971) recommendations for tables with an extreme departure n should be 223. The sample size is 15 by Koehler and Larntz's (1980) formula, but their caveat about too many small e_i probably applies to this case. Applying Yarnold's (1970) guidelines gives an n of 278. When power is the criteria for choosing the sample size, one finds that n5 is 24 while n8 is 122.

In summary, the various guidelines yield very different sample sizes; all but Cochran's are either smaller than n5 or are larger than n8! These guidelines will be compared empirically in the following simulations.

In summary, previous research suggests that the size of e_{min} depends on the size of the table (e_{min} decreases as k increases) and the number of small e_i (e_{min} increases as n(p) increases). X^2 apparently becomes unstable when there are both small and large e_i. The first two factors will be considered in part 1 and all three will be considered in part 2.

Part 1 Number and size of e_{min}

Methodology. Tables of dimensions 1 x k were generated, where k was equal to 4, 8, and 16. The number of cells with small expectations (n(p)) also varied for a total of 14 different tables (refer to Table 4-1). For each of these 14 tables, simulations were run for various sample sizes. Two minimum sample sizes, 10 and 16, were used for k = 4. The first minimum sample size is appropriate when the cell expectations are fairly uniform; the second minimum n is more appropriate when the cell expectations are skewed (Koehler and Larntz, 1980). The larger tables had the minimum sample size set to 16. The sample size was increased by increments of 0.5k until the maximum of 5k was reached.

The small cell probabilities of the tables were decreased until one or more Type I error rates fell out of the range (.036, .064). This range corresponds to the confidence limits when there are 1000 replications. For each table, k and n(p) remained fixed.

Several suggested minimum e_i are reported in table 4-1. Yarnold developed his index for $n \geq 5k$. His formula, $e_{min} \geq 5r/k$, was modified so that it could be used here. I substituted n(p) for his r. A trial index was created by combining Koehler and Larntz's formula for n with the modified form of Yarnold's index: n_{min} n(p) / k^2 with n_{min} = $(10\ k)^{1/2}$ or 10, whichever is larger.

Results. The smallest e_{min}s that had all Type I error rates falling within the confidence interval are listed in Table 4-1. As predicted by previous research, the minimum cell frequency, $e_{min,}$ increases as n(p) increases for any specific k; e_{min} is larger when there are fewer small cells in the table. The lower limits suggested for e_{min} by previous researchers are, in general, larger than the e_{min} observed by this simulation study. The e_{min} given by the modified version of Yarnold's formula is of particular interest since his values follow a similar pattern to that of the observed e_{min}. The trial index is closer to the observed e_{min} than any of the other guidelines but it falls below the observed e_{min} when k=16, therefore it may be too small for larger tables.

Part 2 Large e_i

Methodology. Five sets of tables (for a total of 13 tables) were generated where the size and number of p_{min} were held constant while the size and number of the maximum expected cell probabilities, p_{max}, were varied. The values of these cell probabilities are listed in Table 4-2. The ratios of pmax to p_{min} ranged from 10.8 to 78.5. All other cell expectations were set to n/k. The sample sizes were increased by increments of 0.5k. An arbitrary large n was chosen as the maximum.

Results. The Type I error rates are plotted against n in Figure 4-1. The plots show time-series type trends because the sample sizes are accumulative. The behavior of X^2 does not appear to be affected by the value of p_{max}. The tables with the largest ratios of p_{max} to p_{min} are not that much different from the tables with less extreme cell expectations. For example, table b in Panel B has a ratio of 73 but its distribution of Type I error rates is similar to that of table a with a ratio of 49. However, the other case with a very large ratio, namely table b in Panel E, does show a more liberal trend. The two tables in this panel have the largest discrepancy in their respective ratios: 18 for table a and 78 for table b. For such a very large discrepancy in ratios, the difference between the Type I error rates is hardly dramatic.

Controlling the number os small cells acted as a constraint to the size of p_{max}. After a point, the only way to increase the size of p_{max} is to increase the number of small cells. What appears to be true is that tables with the same number of small cells are similar, irrespective of the size of p_{max}, at least when k = 4 or 16. Panel E suggests that different results may be found for larger tables with many small cells. In these cases, n(p) would have less of a constraint on the relative size of p_{max}, resulting in much greater extremes in cell expectations.

The next section will look at the relationship between n(p), e_{min} and power approximations. Will these factors which were found to influence the Type I error rate also influence how well X^2's power is estimated by the noncentral χ^2 distribution?

Part 3 Power

Methodology. In Chapter 3, only one series of tables from set I was used for the comparison across tests. The power distributions of X^2 are presented here for all of the set I tables. The sample sizes correspond to n5 and n8. There are five different values of n(p) for k = 16: 1, 4, 8, 12, and 15. The four-cell tables have three possible values for n(p): 1, 2, and 3. At least one table in each series is expected to have Type I error rates close to α for very small sample sizes.

Results. The power plots for the sixteen-cell tables are presented in Figure 4-2. From Chapter 3 one would expect that the noncentral χ^2 would be a good approximation of actual power at n8 and less so at n5. The results provide a few surprises. In Panel f (k=4, n(p)=3) the fit is fairly good, as expected for n8. On the other hand, the four-cell tables with fewer small cells (Panels b and d) show a poorer fit for the same sample size.

Among the sixteen-cell tables, it is the table with only one small cell that shows a poor fit at n8 (Panel h). These results suggests that it may be the size of e_{min}, independent of the number of small cells, that affects the fit of the noncentral χ^2 to the observed power distribution of X^2. But this hypothesis is contradicted by the results for n5. Panels i and k (n(p) = 4 and 8 respectively) show a good fit although these tables have smaller cell probabilities than Panels m and o (n(p) = 12 and 15 respectively).

To further complicate matters, Panel c indicates that another factor is involved. There are three jumps in power: F212 jumps at w = .3, F209 jumps at w = .4, F201 jumps at w = .5. These jumps correspond to a change in the pattern of cell probabilities in H_1. The lower power corresponds to an H_1 where is greater when there is a trade-off between like cells, e.g., one small e_i

decreases by the amount that the other small e_i increases. A jump in power corresponds to an H_1 where a cell with a large expectation decreases while the other three cells increase. This latter pattern was used consistently for n(p) = 1 and 3. It thus appears that actual power was maximized (inadvertently!) by the H_1 used in these two series of simulations.

The above observation led to a question: Is degenerate power associated with an H_1 which posits that some small cell probabilities become even smaller? The tables with n(p) = 12 and 15 have a large number (6 to 12) of cells that are posited to have probabilities smaller than e_{min} under the null. This may explain in part the discrepancy seen between observed and expected power. This issue is explored in the next section.

Part IV Power for two different H_1

Methodology. Two sets of alternative hypotheses were created. One with a positive pattern, meaning that all cells with small possibilities were larger under H_1. Under the negative pattern, at least two-thirds of the small cells were set to .001. Simulations were run for four-cell tables and one sixteen-cell table, S1275. The effect size was set to w = .3 (moderate). The 95% confidence interval for the power distribution is $\theta \pm 1\%$.

Results. In Figure 4-3, the observed rejection rates are plotted against the sample size. Several of the observed power functions in Panels a and b do not increase smoothly as the sample size increases. In these extreme cases, the possible values for X^2 are restricted and the distribution for X^2 is a step function. The rejection rate decreases sharply when a specific set of observed values yields an X^2 that falls just below the critical value. For example, for the case k = 4, n(p) = 3, p_{min} = .01, under the negative pattern, the set of observed values (0, 0, 1, 9) occurs fairly frequently when n = 10. Its X^2 is larger than the critical value: $8.35 > \chi^2_{.05,3} = 7.815$. At n = 12 the similar set (0, 0, 1, 11) is no longer significant: $6.73 < \chi^2$. This results in the drop observed in the power function. In Panel b the plot appears to smooth out near n = 100 for n(p) = 1 (where e_{min} = 4) and at n = 120 for n(p) = 3 (where e_{min} = 4.8).

The power plots are comparatively smooth for the large table (Panel c), even though two-thirds of the cells have very small expectations: n(p) = 12, p_{min} = .012. The two extreme H_1s show that very different power plots can be created for the same table. At the maximum sample size e_{min} is 2.8.

In none of the plots do the two H_1s converge. In Panel a, there is a difference of 13% in the rejection rate between the two alternate hypotheses at n = 200 (e_{min} = .2). In Panel b, the disparity in the rejection rates between the two hypotheses is 6% for when e_{min} = 8. In Panel c, at the maximum sample size, e_{min} = 2.8 and the disparity in the rejection rates is nearly .10. The observed power plots are all outside the confidence interval of the asymptotic distribution - even when all cell expectations are greater than five (Panel b).

Discussion

Type I error was found to be sensitive to several factors: the size of the minimum expectations, the number of small expectations, and the size of the table. Power was found to be sensitive to an additional factor, namely the pattern of differences posited by the alternative hypothesis. Power plots where the small cells were larger under H_1 were quite different from those where a majority of the small cells were smaller.

The approximation of X^2's distribution by χ^2 does appear to be satisfactory for sample sizes smaller than those generally recommended. However, under the same conditions the power distributions of X^2 are not well approximated by the noncentral χ^2. As suggested by Figure 4-3 Panel b, power can be underestimated by the noncentral χ^2 even when the sample size is larger than that recommended by any of the present guidelines. Admittedly, the observed power is not greatly overestimated and the case used is extreme.

Any recommendations based on these limited number of cases would be premature. Further work controlling all four known factors is needed in order to develop reliable guidelines.

Application

This section is meant to illustrate how to apply the simulation results to a hypothetical example. A simulation was run to test the predictions made.[5]

An example was described in previously in the section "Implications for researchers." The four-cell table had two small cells. These cells represent the extremes on the spectrum of educational level. If local teachers are higher than the national average at one end of the educational spectrum, they are likely to be lower than the national average at the other. In other words, it's unlikely that a state having a higher percentage of teachers with advanced graduate degrees would also have more teachers who have not attained a bachelor's degree. Therefore, the alternate hypothesis is not likely to be an extreme case where both small cells are smaller than under H_0.

From the simulations in part 1, we can expect that the Type I error to be acceptable as long as $e_{min} \geq .96$. (Refer to Table 4-1, k = 4, n(p) = 2.) Given that p_{min} is .009 for this example, n should therefore be at least 107. The results of part 3 suggest that power is likely to be somewhat less than predicted by the noncentral χ^2 even when n = 122 (n8). (Refer to Figure 4-2 Panel d, case F215.) The actual power for this specific case was .03 less at n = n8 for the H_1 which

[5] The data presented in all four application sections are made up. The confirmatory simulation runs used data generated by *Numerical Recipes'* RAN2 (Press et al., 1992). This program uses a L'Ecuyer generator with a Bays-Durham shuffle.

posited that local teachers would have higher educational levels than the national average. Contrary to expectations, the other H_1 tested showed more power (+.02) than predicted by the asymptotic power distribution. The second H_1 posited that local teachers are less well educated than their national peers: P_{min} became larger under H_1.

The predictions based on the previous simulations were therefore not entirely misleading although the power trend for one of the alternative hypothesis was opposite of what was expected. Power cannot yet be accurately predicted by the results of this simulation study.

Chapter 5

THE GOODNESS OF FIT TEST UNDER THE PRODUCT MULTINOMIAL SAMPLING MODEL

It may be best to explain the product multinomial case of the goodness of fit test by contrasting it with the usual multinomial case. In the example used in the previous chapter, we were interested in teachers' level of education. Let's say that it is known that teachers' level of education is not homogeneous across all groups, specifically that high school teachers are more likely than any other group to have a graduate degree. If our sample has a higher percentage of high school teachers than in the national sample, this bias may cause us to erroneously reject the null hypothesis. One option for controlling this bias is to sample from each group and test against the expected proportions for each separate group. This, then, is the product multinomial version of the goodness of fit test.

The research question remains the same as for the multinomial case: Are local teachers comparable in level of education to the nation as a whole? The number of degrees of freedom, however, differs. For I groups and J categories, the correct degrees of freedom is I(J - 1) or k - I. Otherwise the goodness of fit test is carried out in the usual manner.

I have found no empirical studies for this version of the goodness of fit test. In Chapter 3, it was seen that the product multinomial case followed the same trends as its multinomial analog. In part 1, the extent of this similarity is evaluated by comparing the simulation results for the two tests. In part 2, the impact of varying the size of the samples is considered.

Part 1. Comparison to the multinomial case

Methodology. Set I tables with n(p) = 12 were used. These were chosen because the fit of the observed power distribution to the asymptotic was found to be poor. The differences in fit for the two sampling models had to be evaluated indirectly because of the discrepancy in the degrees of freedom: [Observed power (product multinomial case) - predicted power (df = 12)] - [Observed power (multinomial case) - predicted power (df = 15)].

Results. The differences in fit are plotted in Figure 5-1. At n = n5 the differences in fit are nearly all negative (Panel a). For the small effect sizes, where power is slightly underestimated for both sampling models, the negative differences mean that the multinomial case has a stronger liberal trend than the product multinomial case. The interpretation is different when the effect sizes are large. Power is overestimated in both cases, but more so for the product multinomial. These differences, however, are small with the largest (in absolute terms) being -.023.

At n = n8 the differences in fit are random - the product multinomial case does not show a consistent bias. The differences, again, are generally small. The two cases can therefore be considered as equivalent, at least when the group samples are all equal in size. This simulation is replicated in the next part with tables where the groups are not equal in size.

Part 2. Varying the size of the samples

Methodology. Set II tables are used, along with their set I counterparts, namely S860 and S875. These are the tables where the minimum expected cell frequencies are held constant while the marginal probabilities are varied. Since it was found in the previous chapter that the patterns of differences under H_1 affected power, this factor was controlled as much as possible. Specifically, I attempted to set the smallest frequencies equal across all tables for a given w. The table specifications can be found in Appendix A.

Results. Figure 5-2 presents the power plots. For both series, the best fit to the asymptotic power distribution occurs when the samples are equal (S860, S875). What is striking is the fact that both the 860 and 875 series have similar plots even though the minimum cell frequencies are smaller for the latter. The 875 series has only slightly less power (approximately -.02) than the 860 series when the effect size is large and n = n5. Both are reasonably well approximated by the noncentral χ^2 when n = n8.

The discrepancies seen in the power distributions at n = n5 (Panels a and c) cannot be explained by the factors that have been considered previously. E_{min} and H_1 patterns can be ruled out since these were held constant. Although the number of small cells do vary somewhat, discrepancies are seen between tables with the exact same n(p). For example, table c's observed power at w = .5 is .27 more than that of table a even though they both have n(p) = 6. Two other possible factors are marginal totals and the distribution of e_i within the rows.

Let's first consider marginal totals as a possible factor. There are two pairs of tables with the same fixed row totals (1: a and b; 2: c and d). Tables c and d do have similar Type I error rates and observed power distributions. The same cannot be said for tables a and b. They show a .17 disparity in power at w = .5. This finding seems to rule out marginal totals as a factor affecting the power of X^2.

The possibility that the distribution of e_i within each sample is the explanatory factor cannot be answered with the sample sizes used in this section. At n = 40, all of the e_{min} are below the minimum observed values found in Chapter 4 while they are all larger than the minimum values at n = 196. Other sample sizes are considered in the next simulation.

Part 3. Distribution of e_i within samples

Methodology. The same tables are used as in part 2. Fewer effect sizes were considered, namely w = .3 to .6. One sample size was chosen so that tables a and c would have e_{min} larger than the minimum observed value for e_{min} (as reported in Table 4-1) while tables b and d, with three small cells, will have an e_{min} below tha minimum observed value. This sample size is 96 for the 860 series and 128 for the 875 series. A second sample size was chosen near the minimum observed value for tables b and d.

Results. The power plots are presented in Figure 5-3. The distribution for S875d shows markedly less power. It is a case where n(p) = 3 therefore it and, to a lesser extent, S860d appear to confirm the expectation that power plots associated with tables having three small cells per group would have less power than the plots for tables with n(p) = 2 in each row. However, the other two tables with n(p) = 3, namely S860b and S875b, do not support this hypothesis. Their power plots are not consistently worse than those of other tables for the smaller sample size. Therefore, the number of small cells within each group does not appear to explain the discrepancies in the observed power distributions noted in part 2.

Discussion

When all samples are equal in size, the power distributions for the product multinomial case of the goodness of fit test are comparable to those for the multinomial case. When sample sizes are not equal, the fit of the observed power distributions to the asymptotic is not as good although this does not necessarily translate as loss of power. In the two series of tables with e_{min} held constant, three of the four tables with unequal samples had more power than the tables with equal sample sizes. I was not able to isolate what specific factor or, more likely, the combination of factors that could explain the discrepancies of the observed power from the asymptotic power distribution.

Application

The application problem will follow up on the example used at the beginning of this chapter. Let's say that the national survey of teachers' level of education yielded the following results when broken down into four groups. The total sample size is 13,060.

	< Bachelor's	Bachelor's	Master's	Master's + 30	Total
Primary					
N	64	2925	1577	4	4571
% of group	1.4	64.0	34.5	0.09	
% of all	0.49	22.4	12.1	0.03	
Upper Primary					
N	26	1698	1528	13	3265
% of group	0.8	52.0	46.8	0.4	
% of all	0.20	13.0	11.7	0.1	
Junior High					
N	13	654	706	65	1437
% of group	0.9	45.5	49.1	4.5	
% of all	0.1	5.0	5.4	0.5	
High School					
N	11	1428	2049	299	3787
% of group	0.3	37.7	54.1	7.9	
% of all	0.08	10.9	15.7	2.3	

From Table 4-1, we can expect that the Type I error rate will be acceptable if e_{min} is at least .44 (k = 16, n(p) = 8). As p_{min} is .0003, n should be 1437. They Type I error rate will be liberal for smaller sample sizes. The simulation results showed that X^2's power tends to be close to the power approximation. (Refer to Figure 4-2 Panel I.) However, the application table has cell expectations much smaller than any of the simulation tables, therefore power can be expected to be less.

The results from the confirmatory simulation run are presented in Figure 5-4. The group sizes are all equal. The four sample sizes considered correspond to expected powers of .80, .90, .95, and .99. The Type I error rates are all liberal, as predicted above. Observed power is considerably less than that of the noncentral χ^2 approximation for two of the alternative hypotheses.

The difference is more marked for the "Shift down" case where smaller cell frequencies were predicted for the Master's + 30 level. This result runs counter to the Chapter 4 application result where the "Shift down" H_1 showed more power! The hypothesis which posited no change for the small cells ("No extremes") had observed power close the nominal values.

In summary, the predicted trends were correct for both Type I error and power under the two hypotheses predicting differences for the small cells. Power, however, was much lower than I expected.

Chapter 6

THE TEST OF INDEPENDENCE

The test of independence differs from the goodness of fit test in that the expected cell probabilities are not predetermined but are calculated based on the marginal probabilities: $e_{ij} = n\ p_{i.}\ p_{.j}$. These expectations cannot be known precisely before collecting the data therefore determining sample size will be a process of guess-estimating. Some have suggested a multi-stage sampling procedure when there is very little information about the possible values of the marginal probabilities (e.g., Horn, 1977).

Simulation studies (Camilli and Hopkins, 1978; Craddock and Flood, 1970; Bradley et al., 1979) have consistently found that X^2 is robust as long as the marginal probabilities are not extremely skewed. For tables varying in size from 2x3 to 5x5 and with nearly equal expected frequencies, Craddock and Flood found that the χ^2 approximations of X^2 is accurate at the 90th, 95th and 98th percentiles for n as small as k. In their extensive simulation study, Bradley et al. found that Type I error rates will not exceed .06 unless both sets of marginal probabilities are extremely skewed. If one set of marginal probabilities is highly skewed while the other is nearly uniform, the Type I error rates are conservative. This conservative bias, as remarked by Bradley, appears to be tolerable to many researchers even though power may be adversely affected. Koehler (1986) and Agresti and Yang (as cited in Agresti, 1990) considered much larger tables. For 10x10 and 20x20 tables, e_{ij} can be as small as 0.5 when all the expected frequencies are equal. When both sets of marginal probabilities are highly skewed, Koehler found the χ^2 approximation to be poor for large, sparse tables. Agresti and Yang, on the other hand, found that the chi-square approximation is adequate given a large table (100 cells) and n = k for marginal probabilities as small as .05. Their tables were not as skewed as those in Koehler's study.

An empirical study on the power of Pearson's chi-squared test of independence for 2x2 tables was carried out by Bradley and Seely (1977). They found errors of approximation when n is small. These errors are most serious when a small n is combined with highly skewed marginal probabilities. For example, given n=20 and marginal probabilities of .1 and .9, the actual power is .8 whereas the power based on the noncentral χ^2 distribution is greater than .95.

In an earlier study Harkness and Katz (1964) compared power estimated by normal approximation methods developed by Patnaik and Sillitto with an exact test, the uniformly most powerful unbiased size α test (UMPUT), for three types of contingency tables. The power of all three tests was overestimated by the normal approximations though the discrepancies were not large. Only 2x2 tables and $n \leq 30$ were considered.

In summary, the simulation studies focusing on Type I error suggest that X^2 is robust when n is small unless the marginal probabilities are highly skewed. On the other hand, power simulations (i.e., Bradley and Seely, 1977) found that the noncentral χ^2 approximation is more sensitive to these factors, at least for 2x2 tables. The initial results presented in Chapter 3 bear this out: Power was found to be seriously overestimated for the generally recommended sample size, n = k or 16, and even for the larger sample size of 40 (n5).

Implications for researchers

Many different guidelines for sample size have been proposed. Cochran's (1952, 1954) guidelines are still frequently cited in textbooks. He suggested that at least 80% of cells should have $e_{ij} \geq 5$ while the remaining cells can have expected values as small as 1. As stringent as Cochran's guidelines are, there are researchers that have recommended even larger sample sizes. Hays (cited in Bradely et al., 1979) recommended that all $e_{ij} \geq 10$ when df=1 and a minimum of 5 for larger tables. Tate and Hyer (cited in Bradely et al., 1979) argued for a minimum e_{ij} of 20. Bradley et al. considered these recommendations as prohibitive and remarked that "traditional rules of thumb based on minimum expected frequency, without regard to the marginal distributions, do not provide selective protection against errors of approximation where such protection is needed most" (p. 1295).

Roscoe and Byars (1971)[6] suggested the following guidelines, given $\alpha = .05$: $n \geq 2k$ when the marginal probabilities are uniform; $n \geq 4k$ when the probabilities are moderately skewed; $n \geq 6k$ for tables with extremely skewed marginals.

[6] This study is cited frequently in the literature related to the test of independence although the actual sampling distribution used is the product multinomial. As the two sampling distributions were found to give similar results in Chapter 3, Roscoe and Byars's guidelines are included in this section.

A more recent set of guidelines based on simulation studies was offered by Wickens (1989, p. 30):

1. For tests with 1 degree of freedom, all the μ_{ij} [cell expectations] should exceed 2 or 3.
2. With more degrees of freedom, $\mu_{ij} \approx 1$in a few cells is tolerable.
3. In large tables up to 20% of the cells can have μ_{ij} appreciably less than 1.
4. The total sample should be at least 4 or 5 times the number of cells.
5. Samples should be appreciably larger when the marginal categories are not equally likely.

The main drawback to these guidelines is the vagueness of some of the terminology. When should one consider the marginal probabilities to be extremely rather than moderately skewed? How much is "substantially more?"

Obviously, these different guidelines lead to different sample sizes. To illustrate how different the sample sizes can be, ns are calculated for a few tables that will be used in the simulations.

Table	p_{min}	Cochra	Tate &	Roscoe	Wickens	Power	Power	Pn
		n Hays	Hyer	& Byars		n5	n8	
		$e_{min}=5$	$e_{min}=20$	n=6k	$e_{min}>1$			
S475	.0047	1064	4255	96	>213	40	176	241
S475b	.0047	1064	4255	96	>213	40	176	86
S875	.0085	589	2353	96	>118	40	176	153
S875b	.0085	589	2353	96	>118	40	176	75

Pn in the last column refers to an index that I wish to introduce here. When n is small, it is possible to end up with a marginal total of zero especially if the marginal probabilities are skewed. When that happens the expectations for that row's (or column's) cells are zero and it then becomes impossible to calculate X^2 for all cells. The probability of getting a marginal total of zero for a specific sample size can be calculated using $\sum_i (1-p_{i.})^n + \sum_j (1-p_{.j})^n$. This estimate is accurate for small probabilities (i.e., less than .05). Pn is the sample size where the probability of getting a marginal total of zero is .01. This index will be considered along with the other factors, namely e_{min} and R, in the following simulation. If any of these indices are useful in predicting when Type I error is close to α, we would then have a quantitative index that can be helpful in determining sample size.

Part 1. Type I error rate

Methodology. Set II tables were used where p_{min} was held constant within each series of tables while the marginal probabilities were manipulated. These tables are described in Chapter 2 and Appendix A. Sample sizes ranged from 16 to 1000. The 95% confidence interval for the Type I error rate is .4 to .6. Whenever a generated table did have a marginal total of 0, it was treated as a failure to reject H_0.

Results. The Type I error rates are plotted in Figure 6-1. The error rates substantiate Bradley et al.'s (1979) conclusion: When both sets of marginal probabilities are extremely skewed the error rates are higher than the nominal α; otherwise X^2 tends to be conservative. Apparently both sets of marginals need to have at least one probability less than .1 for the Type I error to become liberal (i.e., larger than expected).

For some of the tables, n must be quite large before Type I error falls within the confidence interval (notably S875b). If one sets wider tolerance limits, as did several of the researchers cited above, then these results do substantiate their conclusion that X^2 is fairly well approximated by χ^2 for the test of independence, even when the marginal probabilities are extremely skewed. The majority of tables have distributions that are within (.3, .7) for $n \geq 32$. There are exceptions, the more notable being S475, S860, S860c, and S875.

Neither p_{min} nor R appear to be useful for predicting how close the Type I error rate will be to the nominal, α. If p_{min} (or, alternatively, e_{min}) were the determining factor, then the error rates would be similar within each series. However, this is not the case. For example, S475a falls within the tolerance limits at $n = 40$, $e_{min} = .188$ while this doesn't happen for S475 until $n = 136$ and $e_{min} = .64$. There would also be noticeable differences across series. The 875 series should be worse than the 860 series ($p_{min} = .0085$ versus .0115 for the 860 series). The same argument can be made against R. The tables with the largest values are not necessarily the worse. By this criteria, all of the 475 tables should have poorer fit than the 875 tables (excepting S875 itself). Though the lowest R values (S875a, b, c and S860a) do tend to have good fits, this is not consistently true (S860).

The index based on the marginal totals, Pn, does show some usefulness in controlling Type I error. Sample sizes that are greater than Pn have error rates well within the tolerance limits.

Part 2. Power

Methodology. For comparative purposes, set I tables are presented here along with two tables from set II, namely S869b and S875b. These latter tables have Type I error rates that are higher than expected. Two sample sizes are considered for these sixteen-cell tables: n = 40 (n5) and 176 (n8).

Results. The power plots are presented in Figure 2. Power is well approximated by the noncentral χ^2 at n = n8. However, this is not the case when n = 40. For these tables with skewed marginals, power is fairly consistently overestimated by the noncentral chi-square distribution. This is true even for the tables associated with a liberal Type I error rate (Panel e, S860b and S875b). The observed power distributions for these tables are also overestimated in the range of interest, namely w = .5 to .7.

Four-cell tables with extremely skewed marginal probabilities have a particularity in that they have a restricted range for the effect size. If one column (or row) total is small relative to the other, there is an upper limit to the size of ES. In these trials, the largest effect size is w = .4. Power is overestimated for small n, but well approximated by the noncentral χ^2 at n = n8.

As was seen in Chapter 3, the overestimation of power is much greater for the test of independence than for the goodness of fit test. Given k = 16, when the number of small expectations was not very large (n(p) ≤ 8), the fit of the observed power distribution by the noncentral $\chi2$ was good for the latter test. For the test of independence, the observed power can be as little as half of that predicted by the noncentral χ^2. Another difference between the two tests is that the number of small cells does not seem to be a factor affecting power for the test of independence. The power plots are fairly similar across n(p) (i.e., compare Panels c, e, and g).

In summary, the power plots of S860b and S875b eliminate p_{min}/e_{min} and R as determining factors. If the first case were true, these plots would be similar to those of their respective set I counterparts, S860 and S875. The observed power for the former tables was greater for all effect sizes. If R was the determining factor, then their power plots would have showed less power than that of S860. However, this expectation is contradicted by the results.

In part 1, it was found that when $n \geq Pn$, Type I error was within the tolerance limits. Can the same be said for power? This question is the motivation for the next simulation.

Part 3. Pn and asymptotic fit

Methodology. The same set of sixteen-cell tables used in Part 2 are used here. The sample size was set to Pn rounded up to the nearest factor of .5k.

Results. The power plots are presented in Figures 3. The fit of the observed power distribution to the noncentral χ^2 is not ideal for all values of w. It seems worse when power is in the middle ranges. The difference between the observed and nominal powers are plotted against the nominal values in Figure 6-4. The relationship is parabolic for power estimates between .05 and .80. The maximum difference in fit is .09, corresponding to a 9% decrease in the rejection rate.

DISCUSSION

The above simulations confirm previous research: The chi-squared test of independence is quite robust as far as Type I error is concerned - as long as one accepts tolerance limits that are somewhat wider than the confidence interval. However, when marginal distributions are skewed and n is small, power can be seriously overestimated by the noncentral χ^2.

Most of the available guidelines for determining sample size recommend sample sizes that are much larger than needed. It was also found that the distribution of marginal probabilities is a better indicator of the Pearson statistic's fit to its asymptotic distribution than e_{min}.

One practical issue not raised in the literature on the test of independence is that small sample sizes may result in marginal totals of 0. A researcher can avoid this problem by calculating Pn, defined in this study as the sample size where the probability of getting a marginal total of zero is 1%. An easier method that yields a similar answer is to multiply the minimum estimated marginal probability by 5.5. This sample size is large enough for the Type I error to be reasonably close to α. Power, however, can be overestimated by as much as .09 when n = Pn. Some adjustment to power estimates is recommended.

An application

A professor is interested in knowing whether the level of exposure to advanced math courses is related to success in her introductory statistics course. Based on a survey she finds the following distribution for highest level of math course taken.

Factor 1: Highest level of math taken	Percentage
No college level math	10
College algebra	55
1 year of calculus	15
1 year or more beyond calculus	20

Based on previous experience, she expects the following distribution for grades.

Factor 2: Grade	Expected percentage
4.0	30
3.5	20
3	40
≤ 2.5	10

Her current enrollment is 40 students. Is the sample size large enough for a reasonable level of power?

To answer the question a plausible effect size must first be determined. One strategy is to calculate w for a possible set of data if a high (but not perfect) correlation exists. If the students are distributed as shown in the following table, w = .87, a considerably large ES. The expected power is better than .90 for w greater than .7.

	≤ 2.5	3.0	3.5	4.0
No college math	2	2	0	0
College algebra	2	14	4	2
1 year calculus	0	0	2	4
> 1 year calculus	0	0	2	6

It was found in this chapter's simulations that Type I error rates generally fell within the range .3 to .7 when the sample size was at least 32 for sixteen-cell table. (Refer to part 1.) The marginal totals of the application table are not extremely skewed - no proportion is expected to be less than .1 - therefore the trend of the Type I error should be conservative.

Marginal totals of zero are not a concern here but two marginal totals are less than five; a sample size of 40 is therefore less than Pn (which equals 51 for this example). Actual power can be expected to be overestimated by the noncentral χ^2. (Refer to Figure 6-2, Panels c and 3 for n(p) = 4 and 8 respectively. N(p) is 6 for the application table.) The overestimation will decrease as the effect size increases. (Refer to Figure 6-4). In spite of the overestimation, a sample size of 40 appears to be large enough for detecting a large effect size with a power greater than .80.

The confirmatory simulation run had a Type I error rate of 4.4% which does fall within the expected range. The power distribution is given in Figure 6-5. The discrepancy between observed and actual power does not consistently decrease as the effect size increased as was predicted above. The largest discrepancy, though, is for w = .5. Observed power is not too seriously overestimated, supporting the conclusion that the sample size is large enough.

Chapter 7

THE HOMOGENEITY TEST

The calculations for the test of homogeneity are carried out in the same manner as the test of independence. The difference is entirely in the sampling procedure. One set of marginal totals corresponds to the samples taken from the various populations. The objective is to determine whether the populations are similar on the characteristic measured. For example, one may ask whether career aspirations of medical students are similar across ethnic groups.

The homogeneity test has been studied less frequently than the test of independence. Camilli and Hopkins (1978) found the homogeneity test to be somewhat conservative when both sets of marginal probabilities were skewed ($e_{min} \leq 2$) but otherwise it was robust for 2 by 2 tables when the sample size was at least 20. A simulation study by Roscoe and Byars (1971) considered two equal groups and varying marginal probabilities on the second dimension (uniform, moderately, and extremely skewed). They reported X^2 to be "strikingly robust." At the .05 level, Type I error was conservative for the smallest sample sizes when the column totals were skewed. They also reported that when both sets of marginals were extremely skewed, the Type I errors were "a bit erratic (though generally conservative)." Garside and Mack (1976) calculated the exact Type I error rates for 2 x 2 tables. All but a very few error rates fell in the .04 to .06 range for $\alpha = .05$. Larntz (1978) tested a 2x3 table with two equal-sized groups. X^2 was close to nominal values for $n \geq 16$ and below nominal for smaller

n. These three studies are therefore consistent in finding that X^2 is robust and tends to be conservative when n is small, much like the results found for the test of independence.

I have not found any simulation studies on the power of the homogeneity test but there have been some theoretical work done. Meng and Chapman (1966) presents Neyman's proof that the optimum sample size for a 2x2 table is $n_1 = n_2 = N/2$. The test of independence has less power than a homogeneity test with equal group sizes. Harkness and Katz's (1964) theoretical study of exact power found that this superiority in power held for $n \leq 30$ and when the two groups were not equal in size. Although higher in power than the test of independence, the homogeneity test's power is still overestimated by the normal approximations developed by Patnaik.

Implications for researchers

Recommendations made for the test of independence appear appropriate for the homogeneity test. Ideally the all the samples would be equal in size as this would maximize power. When the marginal totals are skewed and/or n is small, the power of X^2 will not be closely approximated by the noncentral χ^2, but research suggest that the test of homogeneity is more robust then the test of independence. How much more robust is the question considered below.

Methodology. Set I tables with n(p) = 8 are used along with two tables from set II, namely S860b and S875b. The set I tables have equal sized groups while the set II tables have skewed marginals on both dimensions. Two sample sizes are used: n = n5 which is 40 for both tests, and n = Pn. The value of Pn will depend on the table.

Results. The power plots for the test of independence and the homogeneity test are presented in Figure 7-1, along with the differences found between the two tests' observed power. In Panel a one can see that the homogeneity test does tend to have more power for the larger effect sizes when n = 40 and its Type I error rate (w = 0) is slightly more conservative. This superiority does not hold when n increases (Panel d). The two tables with unequal sample sizes show the same pattern (Panel g): The homogeneity test's superiority in power appears to exist only for large effect sizes and small n. The maximum observed difference in power is .05 (Panel g) with nearly all other positive differences being less than .03.

Discussion

The homogeneity's test theoretical superiority in power over the test of independence was confirmed but found to be significant only for large ES and small n. Guidelines developed for the test of independence appear to be generalizable to the homogeneity test.

Application

From a ten-year old large-scale study, it was found that career aspirations among medical students differed across ethnic groups. A replication study is being considered. Previous data provide the following information. Sixty-five percent of medical students are white, 25% are black, 7% are Hispanic, and 3% are Asian. The breakdown for career aspirations is: Private practice, 54.0%;

Salaried positions, 12.9%; Faculty positions, 29.5%; the remaining 2.7% are lumped together as "Other." The effect size is expected to be moderate at best.

If the smallest group size is 10 to 24% of the overall sample size, Pn will be 175. Since only one set of the marginals will be extremely skewed, the Type I error rate can be expected to be conservative. A sample size of 176 is theoretically large enough to detect a moderate effect size with a power of .8. The simulation results suggest when the sample size is greater than Pn, Type I error will be reasonably close to α and the power approximation will also be close to the observed power. (Refer to Figure 7-1, Panel e, Table S875.)

The confirmatory simulation run with the smallest group making up 10% of the overall sample does substantiate the predictions: The Type I error rate was 5.8 and power was .82. The results were slightly better when all groups were set equal: The Type I error rate was 4.2 and power was .80.

Chapter 8

SUMMARY AND RECOMENDATIONS

Although the asymptotic distributions for Pearson's chi-squared statistic are the same across tests, it was shown here that X^2 behaves differently when n is small. The fit of X^2's observed distributions to the asymptotic is further worsened when the distribution of expected cell frequencies is not uniform. Under these conditions, the goodness of fit X^2 tends to have a liberal Type I error. In contrast, the test of independence is generally conservative unless both sets of marginal probabilities are extremely skewed. For both tests it was found that power estimation is more sensitive than Type I error. Overestimation of power is much more serious for the test of independence than the goodness of fit test. The product multinomial analogs of these tests have similar trends.

Several sample size guidelines were considered for each test. These yielded greatly divergent sample sizes. The objective of the earliest guidelines was to have a close approximation of X^2's Type I error rate by χ^2. These guidelines are stringent and their recommended sample sizes tend to be large. Later guidelines based on simulations considered a looser fit as acceptable, therefore these sample sizes are often considerably smaller. Though there have been empirical power studies, these haven't led to sample size guidelines. This study attempted to combine both perspectives for evaluating sample size guidelines.

A related problem is how best to describe tables with cell expectations that are not uniform. The minimum cell expectation is frequently the criteria used by sample size recommendations. It was found to not be a sufficient criteria for the goodness of fit test and it is not as useful as marginal totals for the test of independence. Several factors are involved in the former case: not only the size of the minimum cell expectation, but also the number of small expectations, the size of the table, and, for power, whether the small cells are smaller or larger under the alternative hypothesis. These factors cannot be all combined into a single index nor can a simple guideline be developed that would account for all of the factors.

The test of independence was easier to deal with. A quantitative index based on the marginal totals, Pn, was described. If the sample size is larger than Pn, a researcher can be confident that the actual distribution of X^2 is fairly well approximated by its asymptotic distributions.

Recommendations for future research

A tension exists between "good enough" for practical purposes and the theoretical perspective. Ideally the sample size should be large enough that the statistic's actual distribution will match its asymptotic distribution. Extreme cases, though, pose a dilemma for practitioners. Given a table with extremely small expectations, the sample size needs to be very large before one can expect a good approximation by the asymptotic distributions. This may neither be feasible nor even desirable. If the researcher is only interested in evaluating a moderate to large effect size but the recommended sample size is so large that it will detect a small to moderate effect size with better than .9 power, the researcher would be justified in thinking that some middle ground should be found! Guidelines that provide adjustments for less than ideal cases would help in this type of situation.

The tentative guidelines suggested here need to be refined and tested to other table sizes in order to make them more generalizable. Determining adjustments for less than ideal sample sizes would also require a large systematic simulation study. Extensions to smaller αs and multi-dimensional tables are two other areas where further research is needed.

Pearson's chi-squared statistic, in spite of its well-known shortcomings, is still the most used test for categorical data. With the growing emphasis on power issues, research on the factors influencing the power estimation of X^2 should become a greater priority.

APPENDICES

Table A-1. Marginal probabilities for tables

Set I.

Id	S135	S145	S160	S175	S435	S445	S460	S475
k	16	16	16	16	16	16	16	16
np	1	1	1	1	4	4	4	4
Var(X^2)	35	45	60	75	35	45	60	75
R	366	526	766	1006	366	526	766	1006
row 1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
row 2	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
row 3	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
row 4	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
column 1	0.195	0.191	0.1894	0.1888	0.080	0.044	0.026	0.019
column 2	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
column 3	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
column 4	0.305	0.309	0.3106	0.3112	0.420	0.456	0.474	0.481

Id	S835	S845	S860	S875	S1235	S1245	S1260	S1275
k	16	16	16	16	16	16	16	16
np	8	8	8	8	12	12	12	12
Var(X^2)	35	45	60	75	35	45	60	75
R	366	526	766	1006	366	526	766	1006
row 1	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
row 2	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
row 3	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
row 4	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
column 1	0.113	0.071	0.046	0.034	0.142	0.095	0.066	0.049
column 2	0.113	0.071	0.046	0.034	0.142	0.095	0.066	0.049
column 3	0.387	0.429	0.454	0.466	0.142	0.095	0.066	0.049
column 4	0.387	0.429	0.454	0.466	0.574	0.714	0.802	0.854

Id	S1525	S1545	S1560	S1575
k	16	16	16	16
np	15	15	15	15
Var(X^2)	35	45	60	75
R	366	526	766	1006
row 1	0.25	0.25	0.25	0.25
row 2	0.25	0.25	0.25	0.25
row 3	0.25	0.25	0.25	0.25
row 4	0.25	0.25	0.25	0.25
column 1	0.165	0.114	0.079	0.060
column 2	0.165	0.114	0.079	0.060
column 3	0.165	0.114	0.079	0.060
column 4	0.505	0.658	0.763	0.820

Table A-1 continued.

Set I.

Id	F107	F109	F112	F115	F207	F209	F212	F215
k	4	4	4	4	4	4	4	4
np	1	1	1	1	2	2	2	2
Var(X^2)	7	9	12	15	7	9	12	15
R	32	52	82	112	32	52	82	112
row 1	0.362	0.349	0.342	0.340	0.50	0.50	0.50	0.50
row 2	0.638	0.651	0.658	0.660	0.50	0.50	0.50	0.50
column 1	0.362	0.349	0.342	0.340	0.146	0.084	0.051	0.037
column 2	0.638	0.651	0.658	0.660	0.854	0.916	0.949	0.963

Id	F307	F309	F312	F315
k	4	4	4	4
np	3	3	3	3
Var(X^2)	7	9	12	15
R	32	52	82	112
row 1	0.196	0.118	0.074	0.054
row 2	0.804	0.882	0.926	0.946
column 1	0.196	0.118	0.074	0.054
column 2	0.804	0.882	0.926	0.946

Table A-1 continued.

Set II.

Id	S475a	S475b	S860a	S860b	S860c	S860d
k	16	16	16	16	16	16
np	4	4	8	8	8	8
Var(X^2)	75	75	60	60	60	60
R	1218	1137	682	870	847	913
row 1	0.125	0.063	0.125	0.125	0.063	0.063
row 2	0.125	0.063	0.125	0.125	0.063	0.063
row 3	0.375	0.438	0.125	0.125	0.063	0.063
row 4	0.375	0.438	0.625	0.625	0.813	0.813
column 1	0.038	0.076	0.092	0.092	0.184	0.184
column 2	0.038	0.076	0.092	0.092	0.184	0.184
column 3	0.462	0.424	0.408	0.092	0.316	0.184
column 4	0.462	0.424	0.4080	0.7240	0.316	0.448

Id	S875a	S875b	S875c	S875d
k	16	16	16	16
np	8	8	8	8
Var(X^2)	75	75	75	75
R	868	1157	992	1165
row 1	0.125	0.125	0.063	0.063
row 2	0.125	0.125	0.063	0.063
row 3	0.125	0.125	0.063	0.063
row 4	0.625	0.625	0.813	0.813
column 1	0.068	0.068	0.137	0.137
column 2	0.068	0.068	0.137	0.137
column 3	0.432	0.068	0.364	0.137
column 4	0.432	0.795	0.364	0.591

Table A-2. Description of cell probabilities

SET I k = 16

Id=S135		Id = S160	
w	Cell probabilities	w	Cell probabilities
0	.0075, 14*.06, .12	0	.0019, 14*.06, .12
0.1	.0127, 7*.06, 7*.07, .11	0.1	.0055, 14*.06, .12
0.2	.018, 7*.05, 7*.07, .11	0.2	.0091, 7*.06, 7*.07, .12
0.3	.0232, 7*.05, 7*.08, .1	0.3	.0127, 7*.05, 7*.07, .11
0.4	.0284, 7*.04, 7*.08, .1	0.4	.0164, 7*.05, 7*.08, .11
0.5	.0337, 7*.04, 8*.09	0.5	.02, 7*.04, 7*.08, .1
0.6	7*.0311, .04, 8*.09	0.6	.0236, 7*.04, 7*.08, .1
0.7	7*.0259, .04, .08, 7*.1	0.7	.0272, 7*.04, 7*.09, .1
Id=S145		Id = S175	
0	.0034, 14*.06, .12	0	.0013, 14*.06, .12
0.1	.0078, 7*.06, 7*.07, .12	0.1	.0045, 7*.06, 7*.07, .12
0.2	.0121, 7*.05, 7*.07, .11	0.2	.0076, 7*.06, 7*.07, .12
0.3	.0165, 7*.05, 7*.08, .11	0.3	.0108, 7*.05, 7*.07, .11
0.4	.0208, 7*.05, 7*.08, .1	0.4	.0139, 7*.05, 7*.08, .11
0.5	.0252, 7*.04, 7*.08, .1	0.5	.0171, 7*.05, 7*.08, .11
0.6	.0296, 7*.04, 7*.09, .1	0.6	.02, 7*.04, 7*.08, .1
0.7	8*.0339, 8*.09	0.7	.0234, 7*.04, 7*.08, .1

Note: All cell probabilities are rounded to two decimal places except for the minimum p.

Table A-2 continued

Id = S435	
w	Cell probabilities
0	4*.02, 8*.06, 4*.1
0.1	2*.015, 2*.026, 4*.06, 4*.07, 2*.1, 2*.11
0.2	2*.0095, 2*.03, 4*.05, 4*.07, 2*.09, 2*.09, 2*.12
0.3	2*.0043, 2*.04, 4*.05, 4*.08, 2*.09, 2*.12
0.4	2*.003, 2*.04, 4*.05, 2*.06, .07, 3*.1, .11, .14
0.5	2*.003, 4*.04, 5*.05, .08, 2*.12, .13 .16
0.6	2*.003, 2*.03, 3*.03, 4*.05, .07, 2*.13, .14, .17
0.7	2*.003, 2*.018, .03, 2*.04, 3*.05, .06, 2*.14, .15, .18
Id = S445	
0	4*.011, 8*.06, 4*.11
0.1	2*.007, 2*.015, 4*.06, 4*.07, 2*.11, 2*.12
0.2	2*.002, 2*.02, 4*.05, 4*.07, 2*.11, 2*.12
0.3	2*.001, 2*.02, 6*.05, .08, 2*.09, .1, .12, .14
0.4	2*.001, 2*.02, 2*.04, 4*.05, .07, .09, 2*.11, .14, .16
0.5	2*.001, 2*.02, 2*.03, 4*.05, .06, .08, 2*.12, .15, .17
0.6	2*.001, 2*.017, 2*.02, 5*.05, .07, 2*.13, .16, .08
0.7	2*.001, 2*.006, 2*.02, .04, 4*.05, .06, 2*.14, .17, .19

Id = S460	
w	Cell probabilities
0	4*.007, 8*.06, 4*.12
0.1	2*.003, 2*.01, 4*.06, 4*.07, 4*.12
0.2	2*.0006, 2*.013, 6*.06, 2*.08, .1, .11, .12, .14
0.3	2*.0006, 2*.013, 2*.04, 4*.06, 3*.09, .1, .14, .15
0.4	2*.0006,2*.013,2*.03,4*.06,.08,.09,2*.11,.15,.16
0.5	2*.0006, 2*.013, 2*.02, 5*.06, .08, 2*.12, .16, .17
0.6	2*.0006, 2*.01, 2*.013, .05, 5*.06, 2*.13, .17, .18
0.7	2*.0001, 2*.0006, 2*.013, .04, 5*.06, 2*.14, .18, .19
Id = S475	
0	4*.0047, 8*.06, 4*.12
0.1	4*.0047, 2*.05, 4*.06, 2*.07, 2*.11, 2*.13
0.2	4*.0047, 2*.04, 4*.06, 2*.08, 2*.1, 2*.14
0.3	4*.0047, 2*.03, 4*.06, 4*.09, 2*.15
0.4	4*.0047, 2*.02, 4*.06, 2*.08, 2*.1, 2*.16
0.5	4*.0047, 2*.012, 4*.06, 2*.07, 2*.11, 2*.17
0.6	2*.0017, 4*.0047, 6*.06, 2*.12, 2*.18
0.7	4*.0001, .0023, .016, .04, 5*.06, 2*.13, 2*.19

Table A-2 continued

Id = S835	
w	Cell probabilities
0	8*.028, 8*.1
0.1	4*.023, 4*.03, 4*.09, 4*.12
0.2	4*.018, 4*.04, 4*.09, 4*.11
0.3	4*.013, 4*.044, 4*.08, 4*.11
0.4	4*.007, 4*.05, 4*.08, 4*.12
0.5	4*.002, 4*.05, 4*.07, 4*.12
0.6	8*.028, 4*.03, 4*.16
0.7	4*.019, 8*.028, 4*.17

Id = S845	
0	8*.018, 8*11
0.1	4*.013, 4*.022, 4*.1, 4*.11
0.2	4*.009, 4*.03, 4*.1, 4*.12
0.3	4*.0046, 4*.03, 4*.09, 4*.12
0.4	4*.0003, 4*.04, 4*.09, 4*.12
0.5	8*.018, 4*.05, 4*.17
0.6	8*.018, 4*.04, 4*.18
0.7	8*.018, 4*.03, 4*.19

Id = S860	
w	Cell probabilities
0	8*.012. 8*.11
0.1	4*.008, 4*.015, 4*.11, 4*.12
0.2	4*.0043, 4*.019, 4*.11, 4*.12
0.3	8*.012, 4*.08, 4*.15
0.4	8*.012, 4*.07, 4*.16
0.5	8*.012, 4*.05, 4*.17
0.6	8*.012, 4*.04, 4*.18
0.7	8*.012, 4*.03, 4*.2

Id = S875	
0	8*.0085, 8*.12
0.1	4*.0053, 4*.012, 4*.11, 4*.12
0.2	4*.0022, 4*.015, 4*.11, 4*.12
0.3	8*.0085, 4*.08, 4*.15
0.4	8*.0085, 4*.07, 4*.16
0.5	8*.0085, 4*06, 4*.18
0.6	8*.0085, 4*.04, 4*.19
0.7	8*.0085, 4*.03, 4*.2

Table A-2 continued

	Id = S1235
w	Cell probabilities
0	12*.036, 4*.14
0.1	6*.03, 6*.04, 2*.14, 2*.15
0.2	6*.025, 6*.05, 2*.13, 2*.15
0.3	6*.02, 6*.05, 2*.13, 2*.016
0.4	6*.015, 6*.06, 2*.12, 2*.16
0.5	6*.0093, 6*.06, 2*.12, 2*.17
0.6	6*.0041, 6*.07, 2*.11, 2*.17
0.7	6*.0041, 2*.04, 2*.07, .08, 29.1, 2*.14, .2

	Id = S1245
0	12*.024, 4*.18
0.1	6*.019, 6.03, 2*.17, 2*.18
0.2	6*.015, 6*.03, 2*.17, 2*.19
0.3	6*.011, 6*.04, 2*.17, 2*.19
0.4	6*.0064, 6*.04, 2*.16, 2*.2
0.5	6*.002, 6*.05, 2*.16, 2*.2
0.6	6*.0013, 2*.02, 2*.05, 2*.07, .13, 29.18, .22
0.7	6*.0013, 2*.012, 2*.05, 2*.08, .12, .17, .19, .24

	Id = S1260
w	Cell probabilities
0	12*.017, 4*.02
0.1	6*.013, 6*.02, 4*.2,
0.2	6*.0092, 6*.02, 2*.19, 2*.21
0.3	6*.0055, 6*.03, 2*19, 2*.21
0.4	6*.0018, 6*.03, 2*.19, 2*.22
0.5	6*.0005, 2*.018, 2*.03, 2*.05, .17, 2*.2, .23
0.6	6*.0005, 2*.007, 2*.03, 2*.06, .16, .19, .21, .24
0.7	8*.0001, 2*.03, 29.07, .15, .18, .22, .25

	Id = S1275
0	12*.012, 4*.21
0.1	6*.009, 6*.015, 2*.21, 2*.22
0.2	6*.0059, 6*.018, 2*.21, 2*.22
0.3	6*.0027, 6*.02, 2*.2, 2*.22
0.4	6*.00006, 8*.02, 2*.23
0.5	6*.00006, 2*.007, 2*.02, 2*.04, .18, .21, .22, .24
0.6	8*.00006, 2*.02, 2*.05, .18, .2, .23, .25
0.7	9*.00006, .02, 3*.05, .15, .2, 2*.25

Table A-2 continued

Id = S1535	
w	Cell probabilities
0	15*.04, .38
0.1	8*.036, 7*.046, .39
0.2	8*.031, 7*.05, .39
0.3	8*.026, 7*.057, .4
0.4	8*.02, 7*.06, .4
0.5	8*.015, 7*.07, .4
0.6	8*.01, 7*.07, .41
0.7	8*.0047, 7*.08, .42
Id = S1545	
0	15*.029; .57
0.1	8*.024, 7*.03, .58
0.2	8*.02, 7*.04, .58
0.3	8*.016, 7*.04, .58
0.4	8*.011, 7*.05, .59
0.5	8*.0068, 7*.05, .59
0.6	8*.0024, 7*.05, .6
0.7	8*.0024, 4*.03, 3*.08, .62

Id = S1560	
w	Cell probabilities
0	15*.0198, .7
0.1	8*.016, 7*.024, .71
0.2	8*.013, 7*.03, .71
0.3	8*.009, 7*.03, .71
0.4	8*.0053, 7*.03, .72
0.5	8*.0016, 7*.04, .72
0.6	8*.0016, 4*.018, 3*.06, .74
0.7	4*.0094, 8*.0016, 3*.07, .75
Id = S1575	
0	15*.0149, .78
0.1	8*.012, 7*.018, .78
0.2	8*.0086, 7*.02, .78
0.3	8*.0054, 7*.02, .79
0.4	8*.0023, 7*.03, .79
0.5	8*.0023, 4*.012, 3*.04, .80
0.6	8*.0023, 4*.005, 3*.05, .81
0.7	8*.0004, 4*.0029, 3*.06, .82

Table A-2. Description of cell probabilities. Set II

SET II k = 16

	Id=S860a		Id = S860c
w	Cell probabilities	w	Cell probabilities
0	6*.0115, 6*.05, 2*.06, 2*.26	0	6*.0115, 6*.02, 2*.15, 2*.26
0.1	3*.0079, 3*.0151, 7*.05, .06, .25, .26	0.1	3*.0079,3*.015,3*.017,3*.02,2*.15,.25,.26
0.2	3*.0043, 3*.019, 3*.04, .05, 3*.06, .25, .26	0.2	3*.0043,3*.013,3*.019,3*.03,.14,.16,.25,.26
0.3	6*.0115, 3*.02, 2*.06, 3*.08, .23, .28	0.3	3*.00275, 6*.0115, 3*.04, 2*.15, .24, .27
0.4	6*.0115, 3*.015, 2*.06, 3*.09, .22, .29	0.4	6*.0115, 3*.012, 3*.03, *.18, 2*.23
0.5	3*.0064, 6*.0115, 2*.06, 2*.1, .21, .3	0.5	6*.0115, 3*.012, 3*.03, 2*.15, 2*.25
0.6	6*.0015, 3*.012, 3*.09, 2*.12, 2*.19	0.6	6*.0115, 3*.012, 3*.03, 2*.13, 2*.28
0.7	6*.0115, 3*.017, 3*.085, 2*.15, 2*.19	0.7	6*.0115, 3*.012, 3*.03, 2*.11, 2*.3
	Id=S860b		Id = S860d
0	9*.0115, 2*.06, 3*.09, .45	0	9*.0115, 3*.03, 3*.15, .36
0.1	4*.008, 5*.015, 3*.06, 3*.09, .45	0.1	4*.0081, 5*.015, 2*.2, .03, 3*.15, .36
0.2	4*.0045, 5*.0185, 3*.06, 2*.08, .1, .45	0.2	4*.0048, 5*.018, 2*.02, .03, 3*.15, .36
0.3	9*.0115, .017, 3*.09, .1, .41	0.3	9*.0115, 3*.03, .09, 2*.021, .3
0.4	9*.0115, .03, .07, 2*.09, 2*.1, .41	0.4	9*.0115, 3*.03, .07, 2*.23, .28
0.5	.0075, 8*.0115, .05, .06, 2*.09, 2*.11, .4	0.5	9*.0115, 3*.03, .04, 2*.25, .26
0.6	.0075, 8*.0115, .04, .06, 2*.09, 2*.11, .4	0.6	9*.0115, .02, 3*.03, .24, 2*.28
0.7	.0075, 8*.0115, .03, .07, 2*.09, .11, .4	0.7	.003, 9*.0115, 3*.03, .22, 2*.3

Note: All cell probabilities are rounded to two decimal places except for the minimum p.

Table A-2 continued

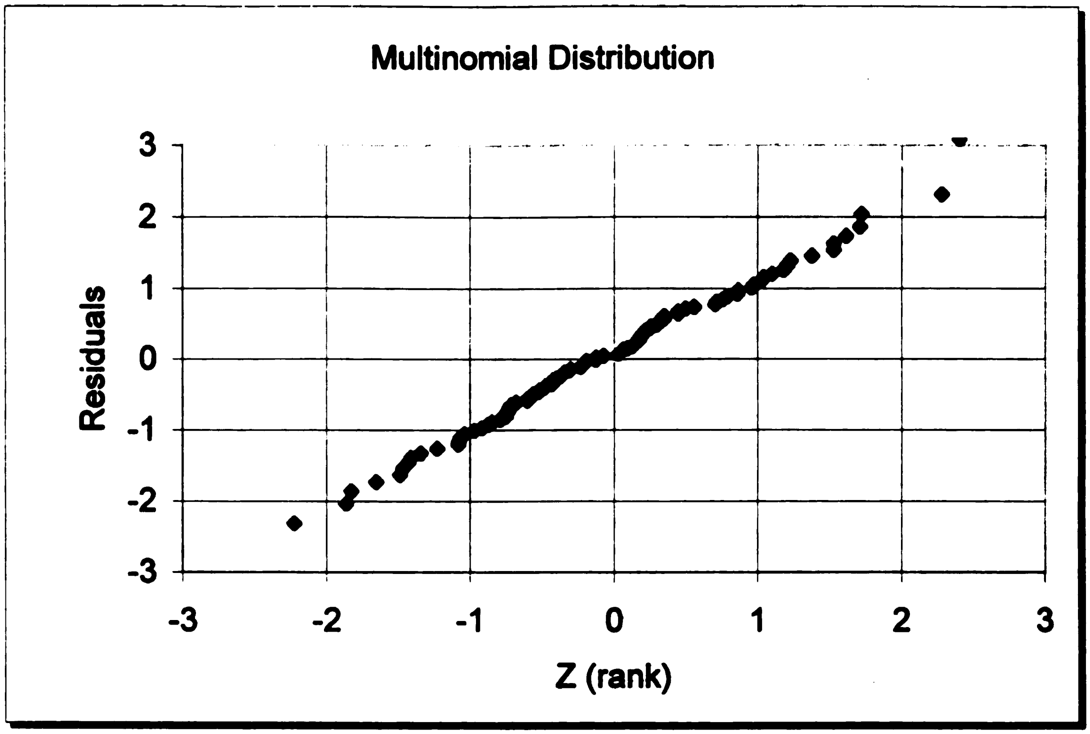
Id = S875a	
w	Cell probabilities
0	6*.0085, 2*.04, 6*.05, 2*.27
0.1	3*.0053, 3*.012, .04, 4*.05, 3*.06, 2*.27
0.2	3*.0022, 3*.015, 4*.04, .05, 3*.06, .26, .28
0.3	6*.0085, 3*.03, 2*.04, 3*.08, .24, .3
0.4	6*.0085, 3*.017, 2*.04, 3*.09, .23, .31
0.5	9*.0085, 2*.04, 3*.1, .22, .32
0.6	6*.0085, 3*.02, 3*.09, 2*.11, 2*.2
0.7	6*.0085, 3*.02, 3*.09, 2*.12, 2*.19

Id = S875b	
0	9*.0085, 3*.04, 3*.01, .5
0.1	4*.0054, 5*.012, 3*.04, 3*.1, .5
0.2	4*.0024, 5*.015, 3*.04, 2*.09, .11, .5
0.3	.0073, 9*.0085, 2*.08, 2*.1, .46
0.4	9*.0085, .03, .07, 2*.08, 2*.1, .46
0.5	9*.0085, 04, .06, 2*.08, 2*.1, .46
0.6	9*.0085, .05, .06, 2*.08, 2*.1, .46
0.7	9*.0085, .04, .07, 2*.08, 2*.1, .46

Id = S875c	
w	Cell probabilities
0	6*.0085, 6*.02, 2*.11, 2*.3
0.1	3*.0053, 3*.012, 3*.02, 3*.03, 2*.11, .29, .3
0.2	3*.0022,3*.015,3*.016,3*.03,.1,.12,.29,.3
0.3	3*.0045, 6*.0085, 3*.04, 2*.11, .28, .31
0.4	9*.0085, 3*.04, 2*.18, 2*.23
0.5	9*.0085, 3*.04, 2*.2, 2*.21
0.6	9*.0085, 3*.04, 2*.18, 2*.22
0.7	9*.0085, 3*.04, 2*.16, 2*.24

Id = S875d	
0	9*.0085, 3*.04, 3*.11, .48
0.1	4*.0055, 5*.012, 3*.04, 3*.11, .48
0.2	4*.0026, 5*.014, 2*.03, .04, 3*.11, .48
0.3	9*.0085, 3*.04, .06, 2*.17, .42
0.4	9*.0085, 4*.04, 2*.19, .41
0.5	9*.0085, .02, 2*.04, 2*.2, .39
0.6	.0059, 8*.0085, 2*.02, 2*.04, 2*.22, .37
0.7	.0027, .0059, 8*.0085, 3*.04, 2*.22, .37

TABLES AND FIGURES



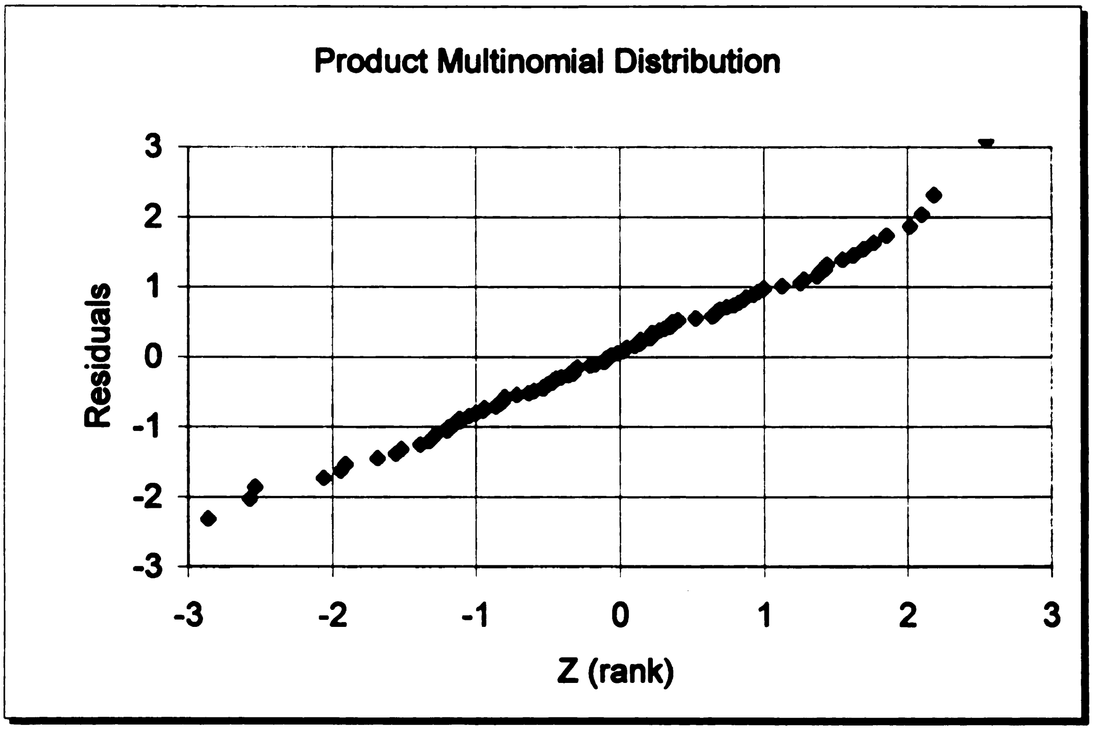


Figure 2-1. Normal plots of the standardized residuals of the cell means.

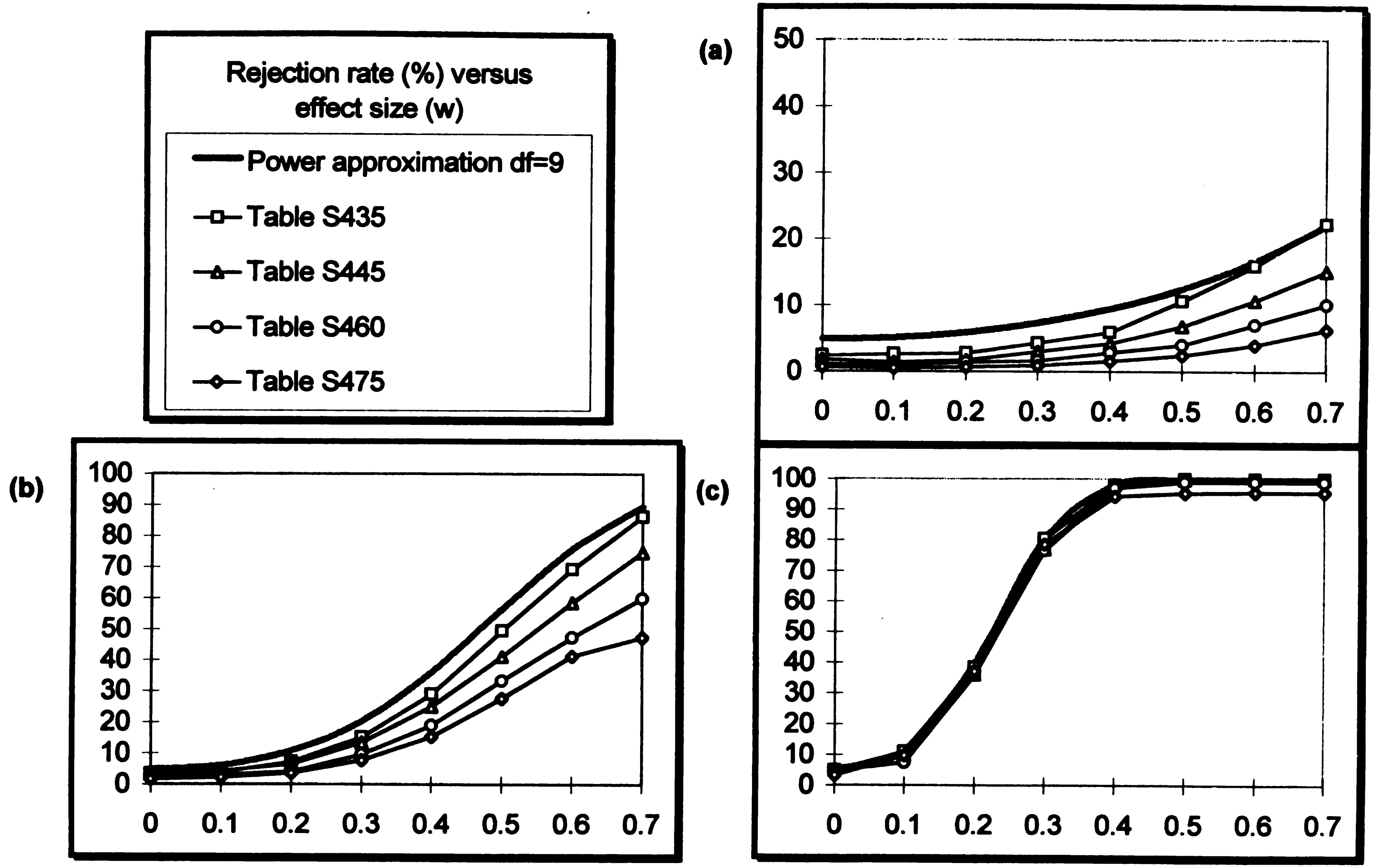


Figure 3-1. Power plots for the test of independence, 16-cell table, 4 small cell expectations, (a) n=16, (b) n=40,(c) n=176.

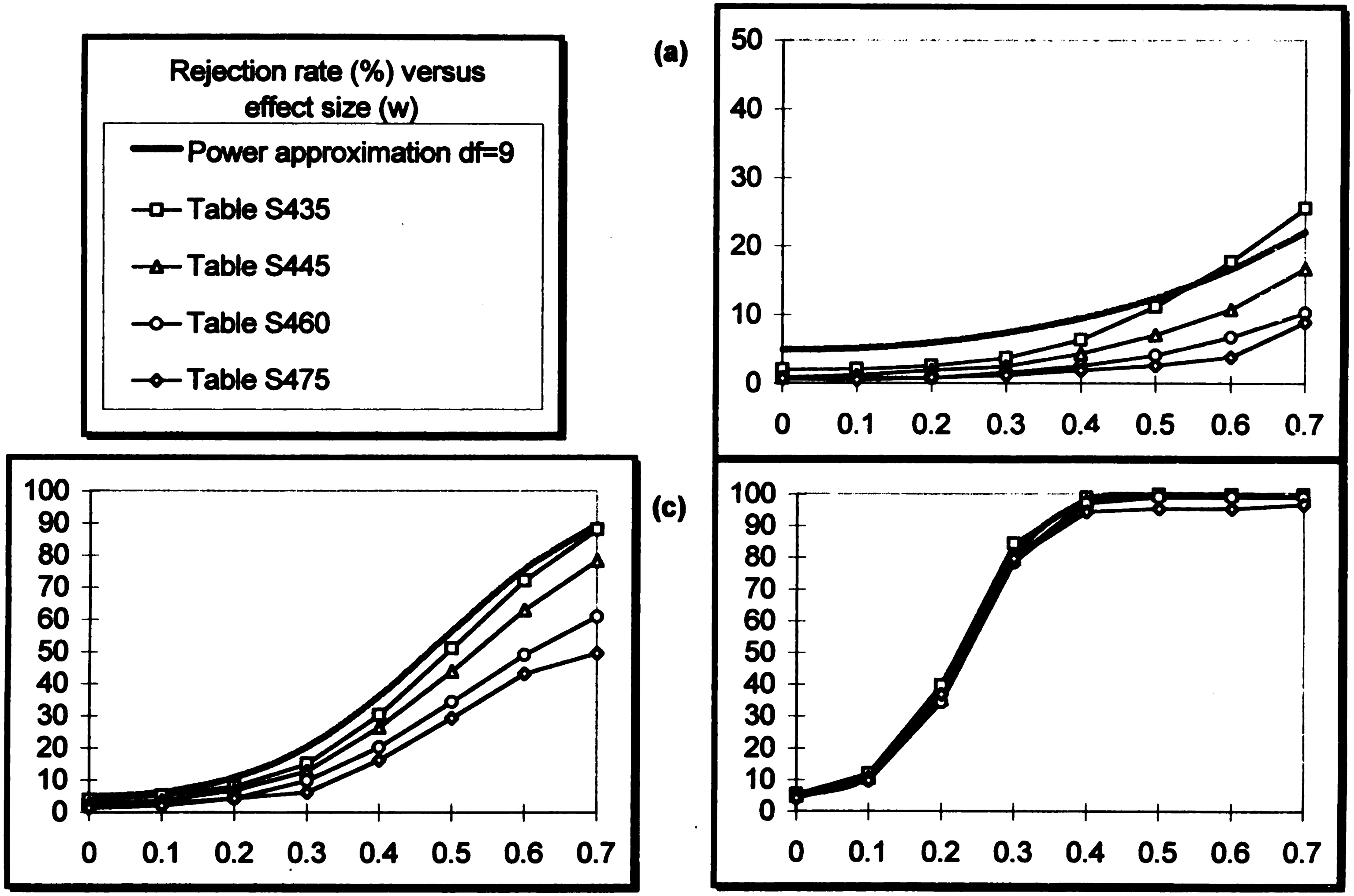


Figure 3-2. Power plots for the homogeneity test, 16-cell table, 4 small cell expectations, (a) n=16, (b) n=40,(c) n=176.

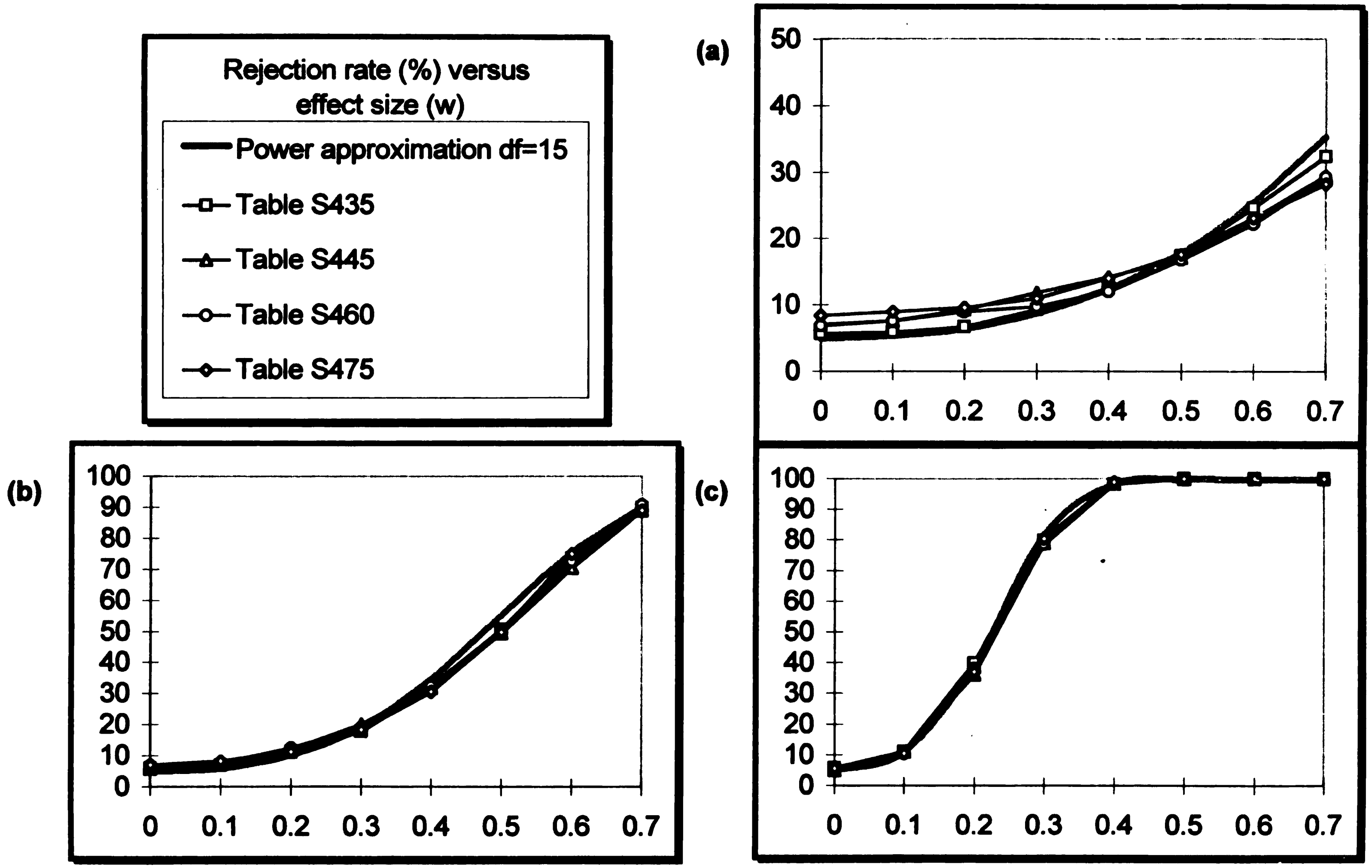


Figure 3-3. Power plots for the multinomial case of the goodness of fit test, 16-cell table, 4 small cell expectations, (a) n=16, (b) n=48,(c) n=216.

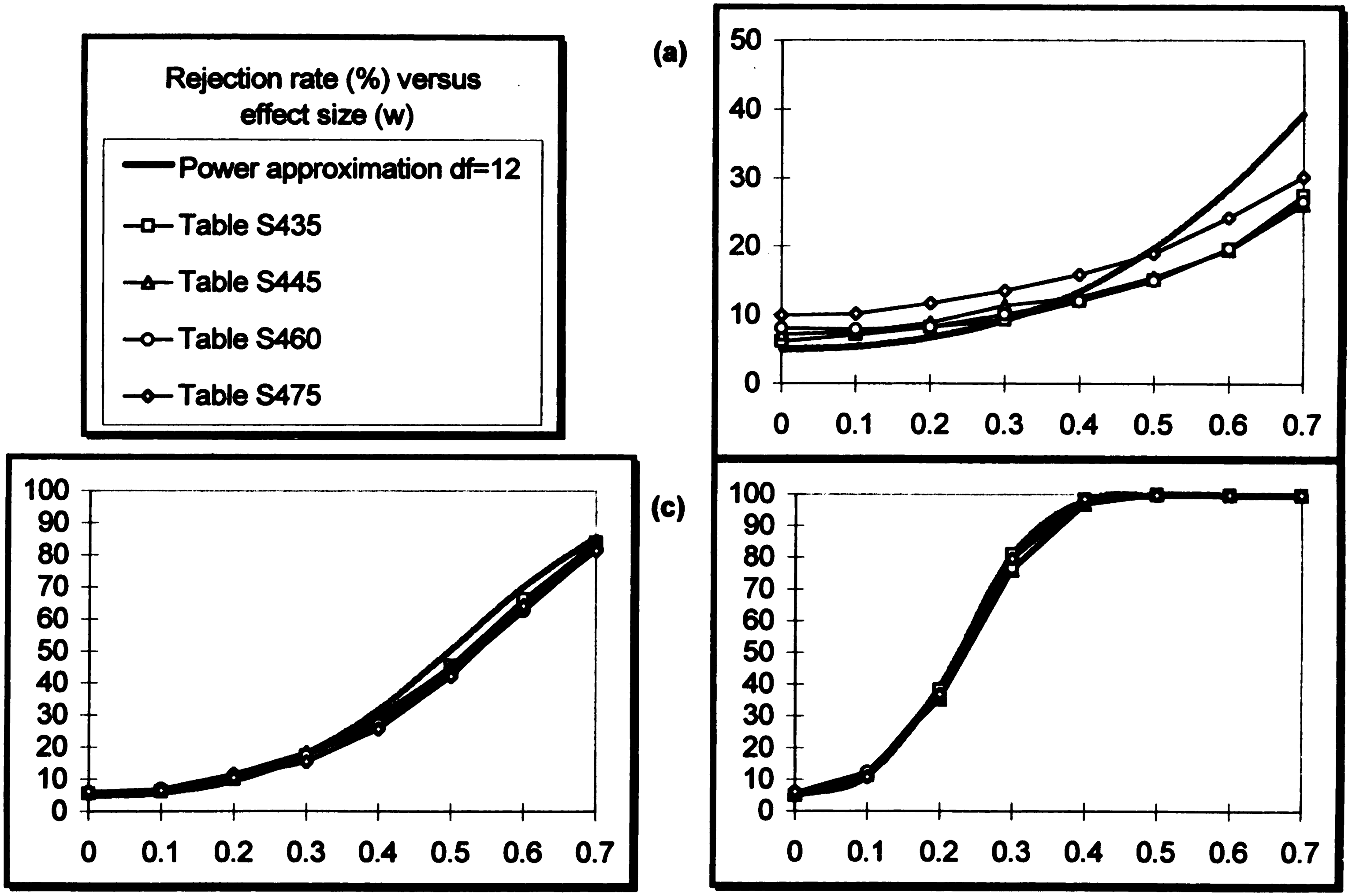


Figure 3-4. Power plots for the product multinomial case of the goodness of fit test, 16-cell table, 4 small cell expectations, (a) n=16, (b) n=40,(c) n=196.

Table 4-1. Lower limits for the minimum expected cell frequency (e_{min})

						Recommended minimums			
k	n(p)	R	p_{min}	n	Observed minimum e	Cochran	Yarnold (Modified)	Roscoe & Byars	Trial index
16	1	366	0.0075	16	0.12	0.5	0.31	1.00	0.05
16	2	370	0.0125	16	0.2	1	0.63	1.00	0.10
16	4	345	0.0225	16	0.36	5	1.25	1.00	0.20
16	8	373	0.0275	16	0.44	5	2.50	1.00	0.40
16	12	349	0.0375	16	0.6	5	3.75	1.00	0.59
16	15	319	0.0475	16	0.76	5	4.69	1.00	0.74
8	1	163	0.008	16	0.096	0.5	0.63	1.00	0.16
8	2	146	0.019	16	0.228	1	1.25	1.00	0.31
8	4	217	0.02	16	0.24	5	2.50	1.00	0.63
8	6	139	0.045	16	0.54	5	3.75	1.00	0.94
8	7	122	0.058	16	0.696	5	4.38	1.00	1.09
4	1	34	0.04	10	0.4	0.5	1.25	1.00	0.63
		49	0.025	16	0.4				
4	2	27	0.09	10	0.9	1	2.50	1.00	1.25
		38	0.06	16	0.96				
4	3	29	0.14	10	1.4	5	3.75	1.00	1.88
		33	0.095	16	1.52				

Table 4-2. Cell probabilities of tables generated for part 2.

							Ratio	
Set	k	n(p)	p_{min}	Subset	$n(p_{max})$	p_{max}	p_{max}/p_{min}	R
A	4	1	0.01	a	3	0.33	33	109
				b	2	0.37	37	109
				c	1	0.49	49	110
B	4	2	0.01	a	2	0.49	49	204
				b	1	0.73	73	205
C	16	2	0.0065	a	14	0.07	10.8	508
				b	8	0.08	11.8	504
				c	1	0.17	26.8	522
D	16	4	0.0065	a	12	0.08	12.5	763
				b	8	0.09	13.9	768
				c	1	0.29	44.1	795
E	16	8	0.0065	a	8	0.12	18.2	1298
				b	1	0.51	78.5	1345

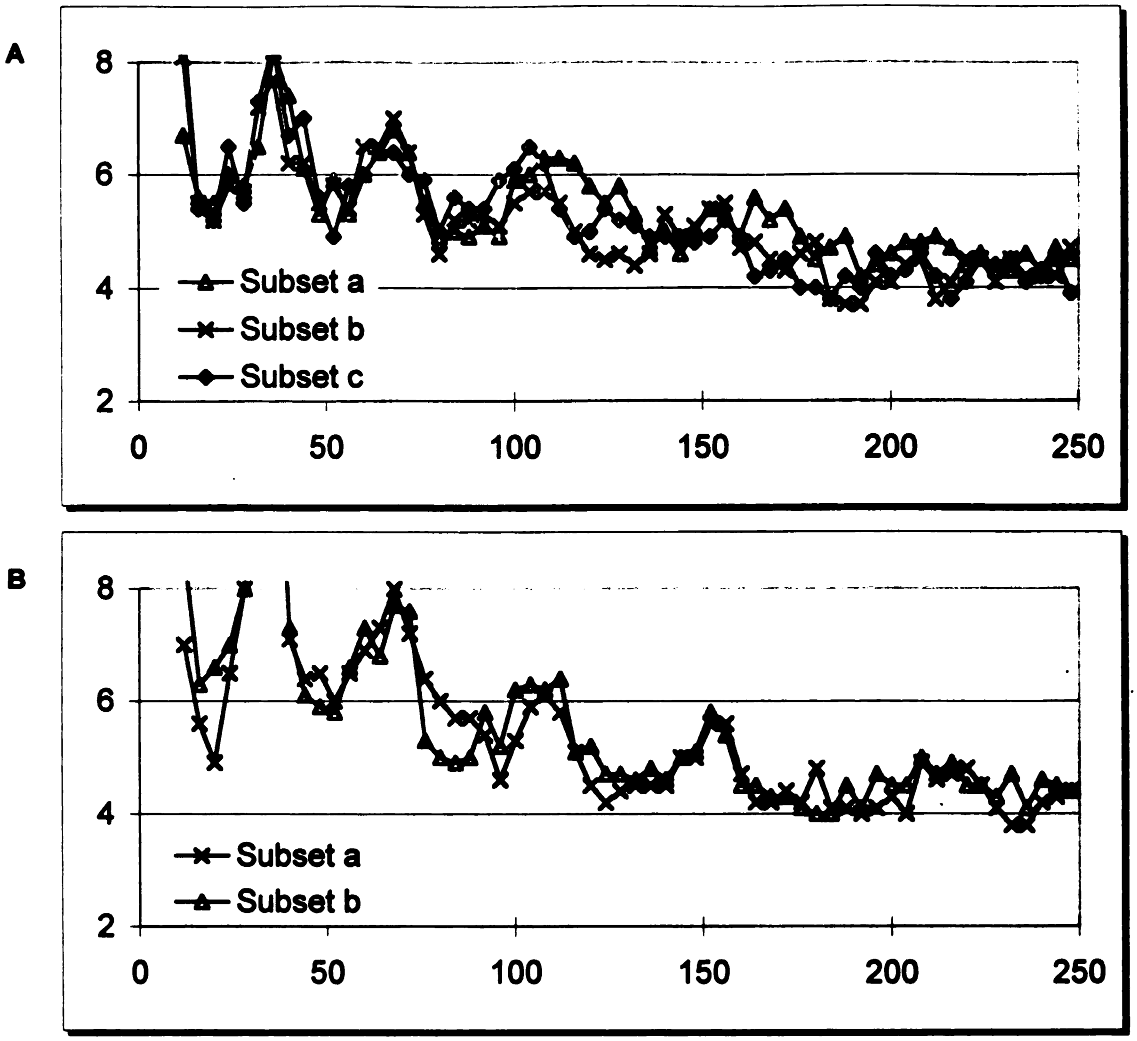


Figure 4-1. Type I error rate (in percent) versus sample size, 4-cell tables, (A) 1 and (B) 2 small cell expectations.

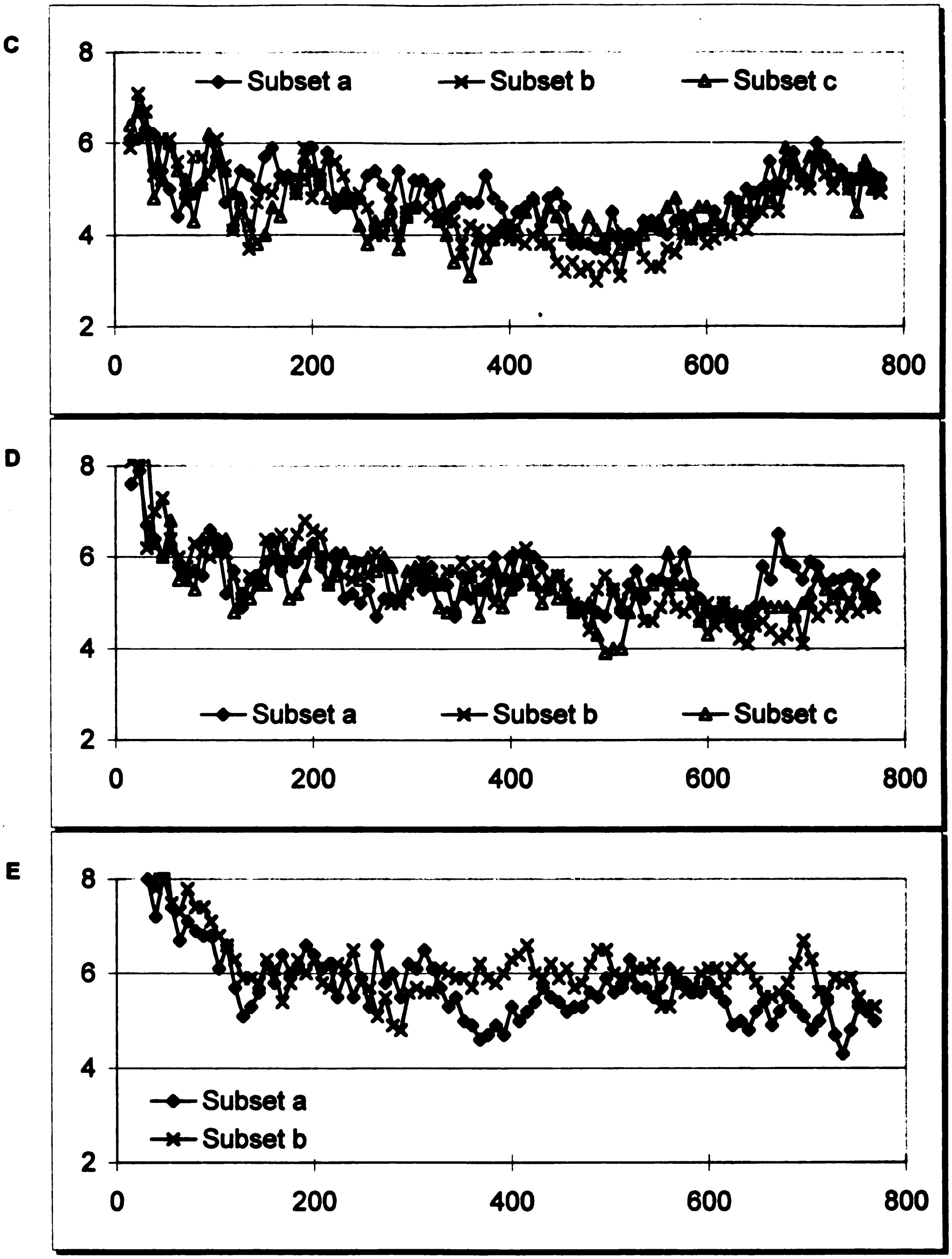


Figure 4-1 continued. 16-cell tables, (C) 2, (D) 4, and (E) 8 small cell expectations.

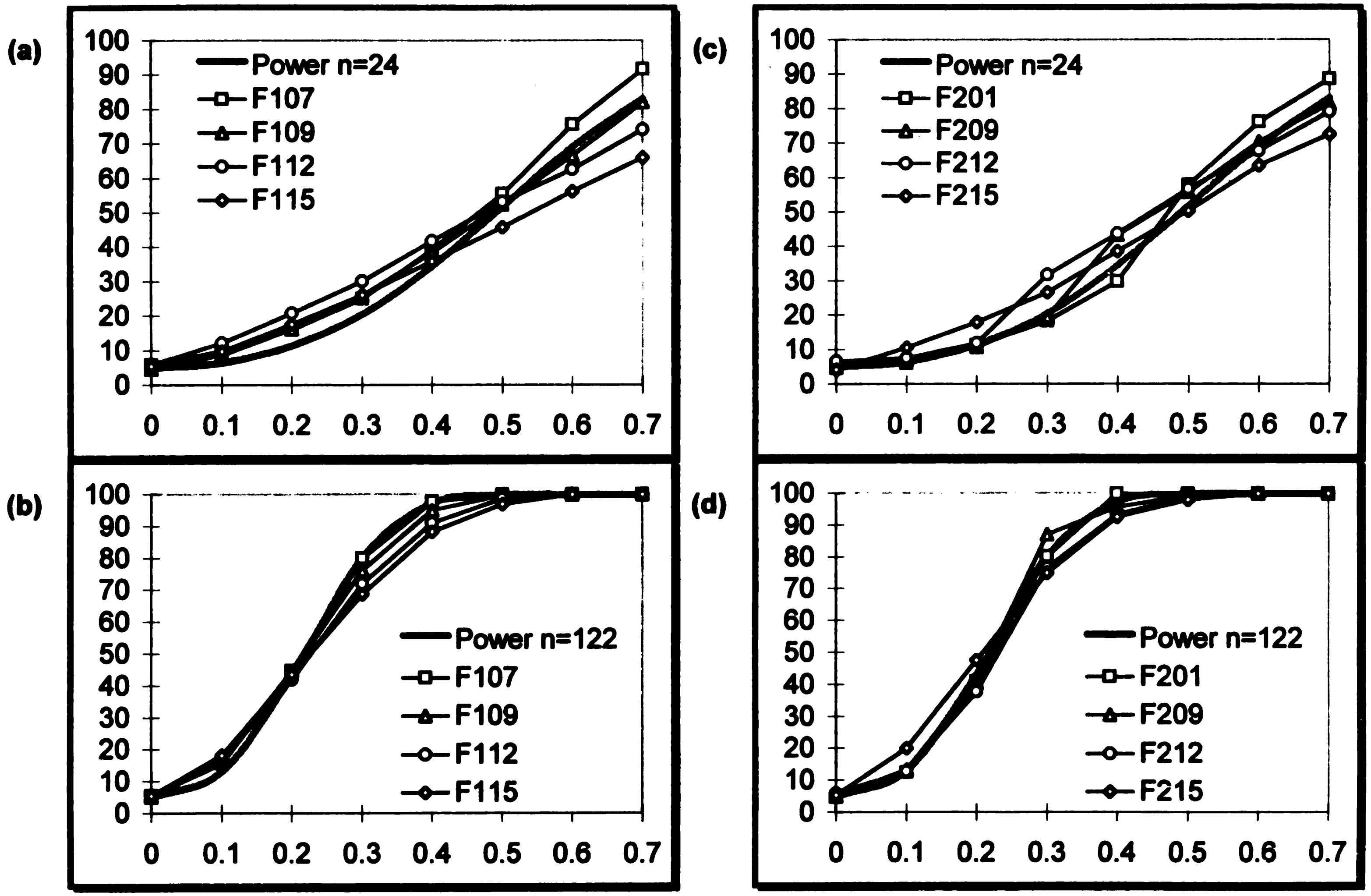


Figure 4-2. Power plots: rejection rates in percent versus effect size, w. (a) & (b) 4-cell tables with 1 small cell expectation. (c) & (d) 4-cell tables with 2 small cell expectations.

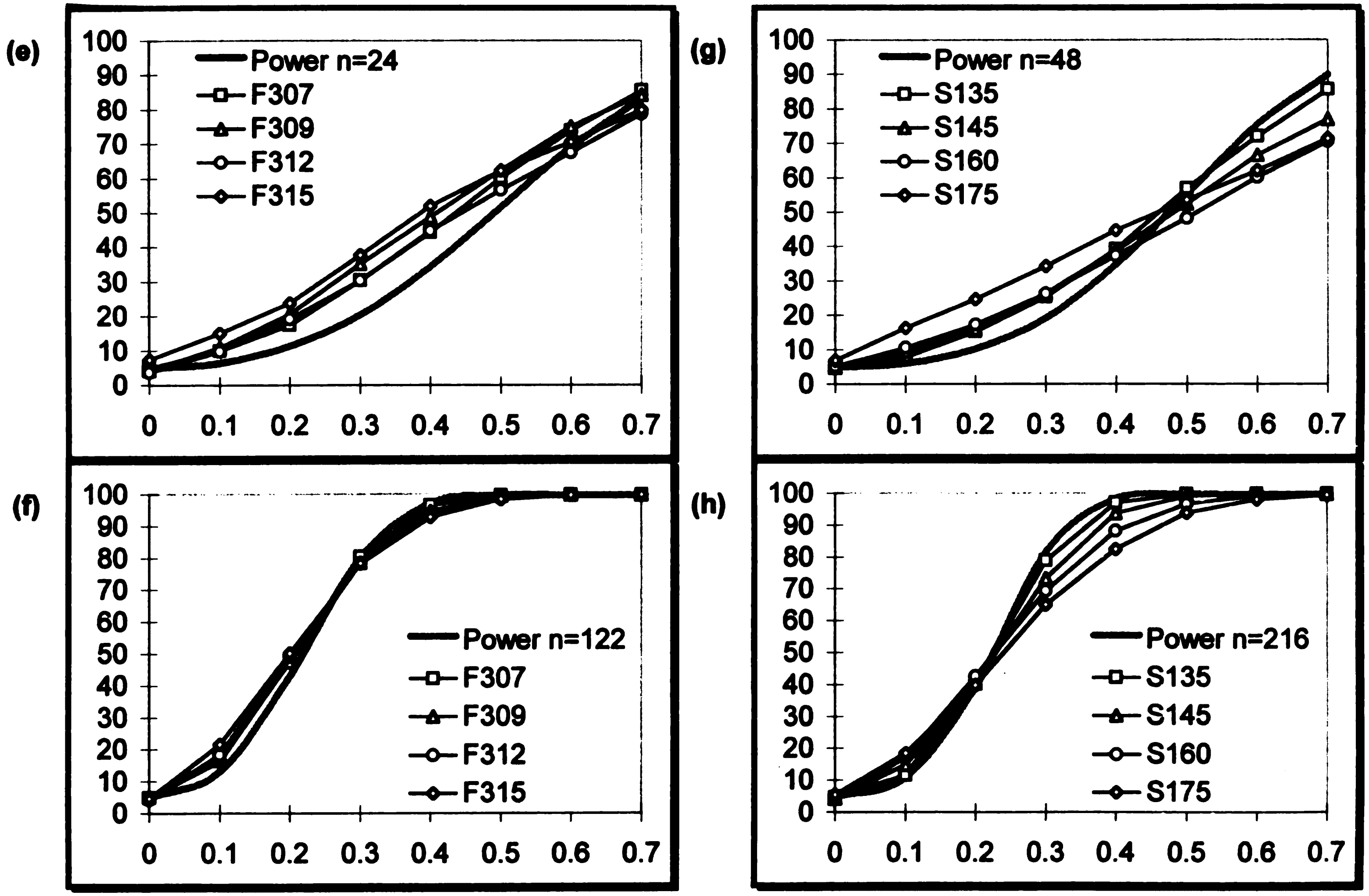


Figure 4-2 continued. (e) & (f) 4-cell tables with 3 small cell expectations, (g) & (h) 16-cell tables with 1 small cell expectation.

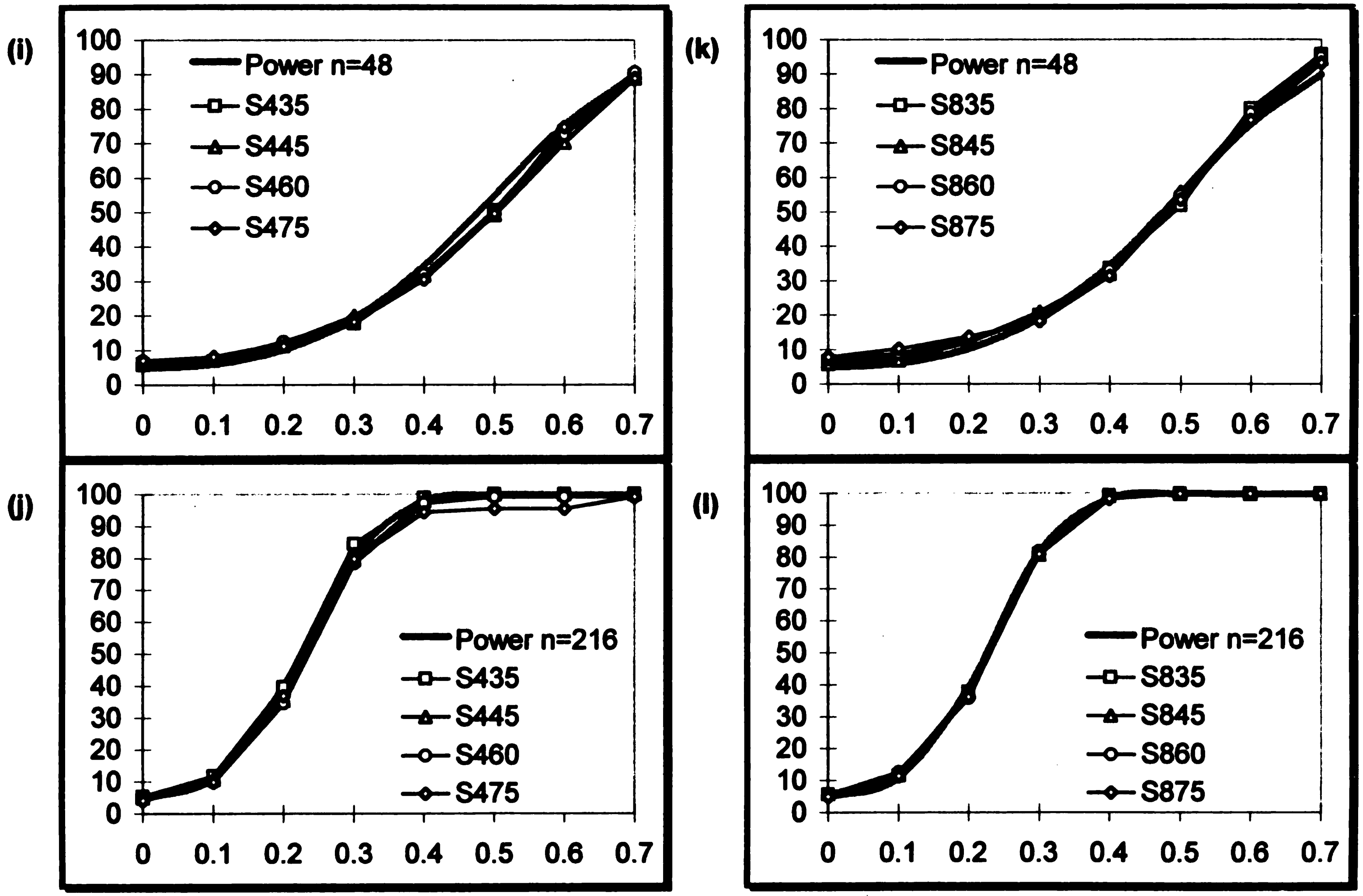


Figure 4-2 continued. (i) & (j) 16-cell tables with 4 small cell expectations, (k) & (l) 16-cell tables with 8 small cell expectations.

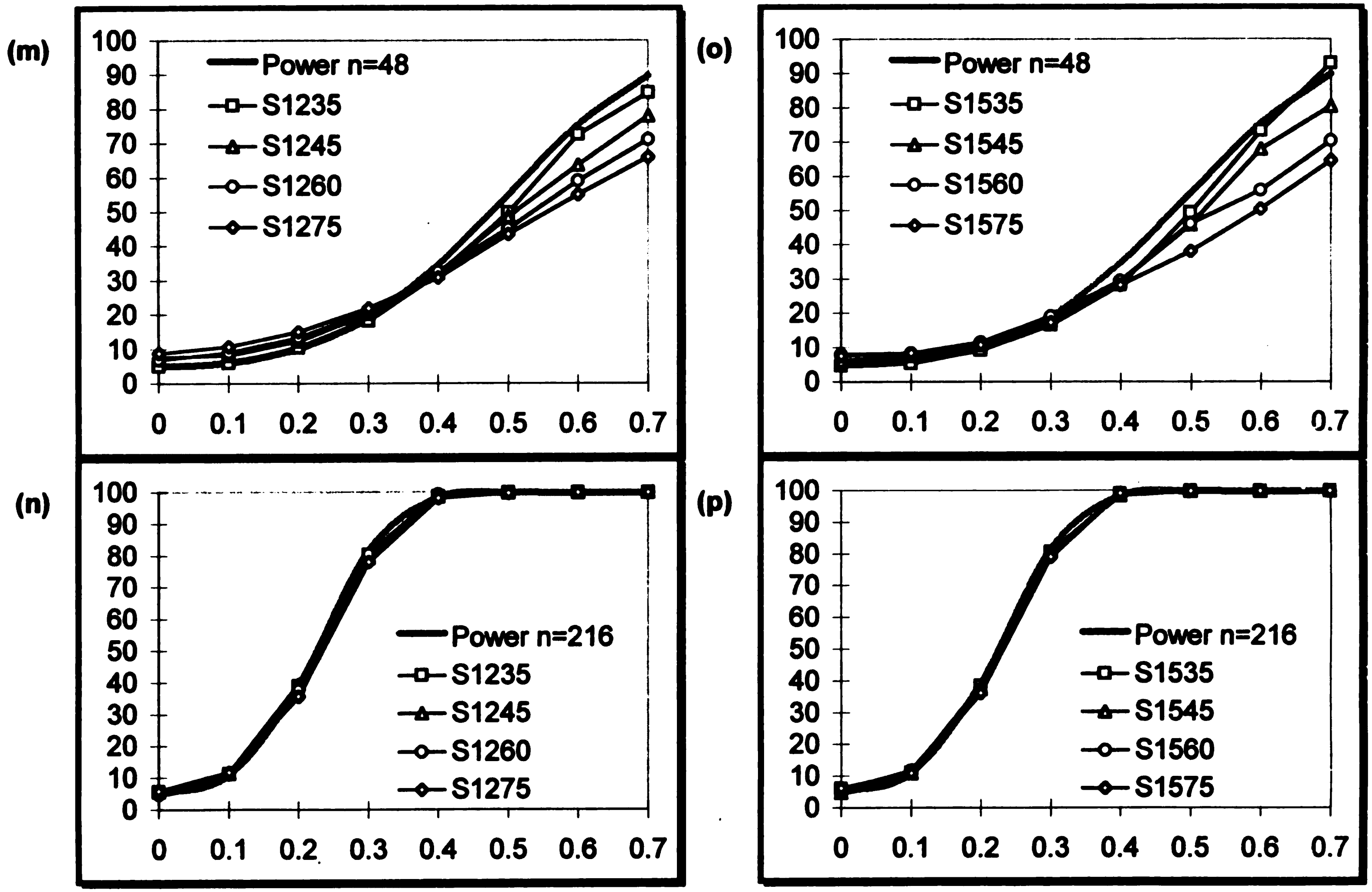


Figure 4-2 continued. (m) & (n) 16-cell tables with 12 small cell expectations, (o) & (p) 16-cell tables with 15 small cell expectations.

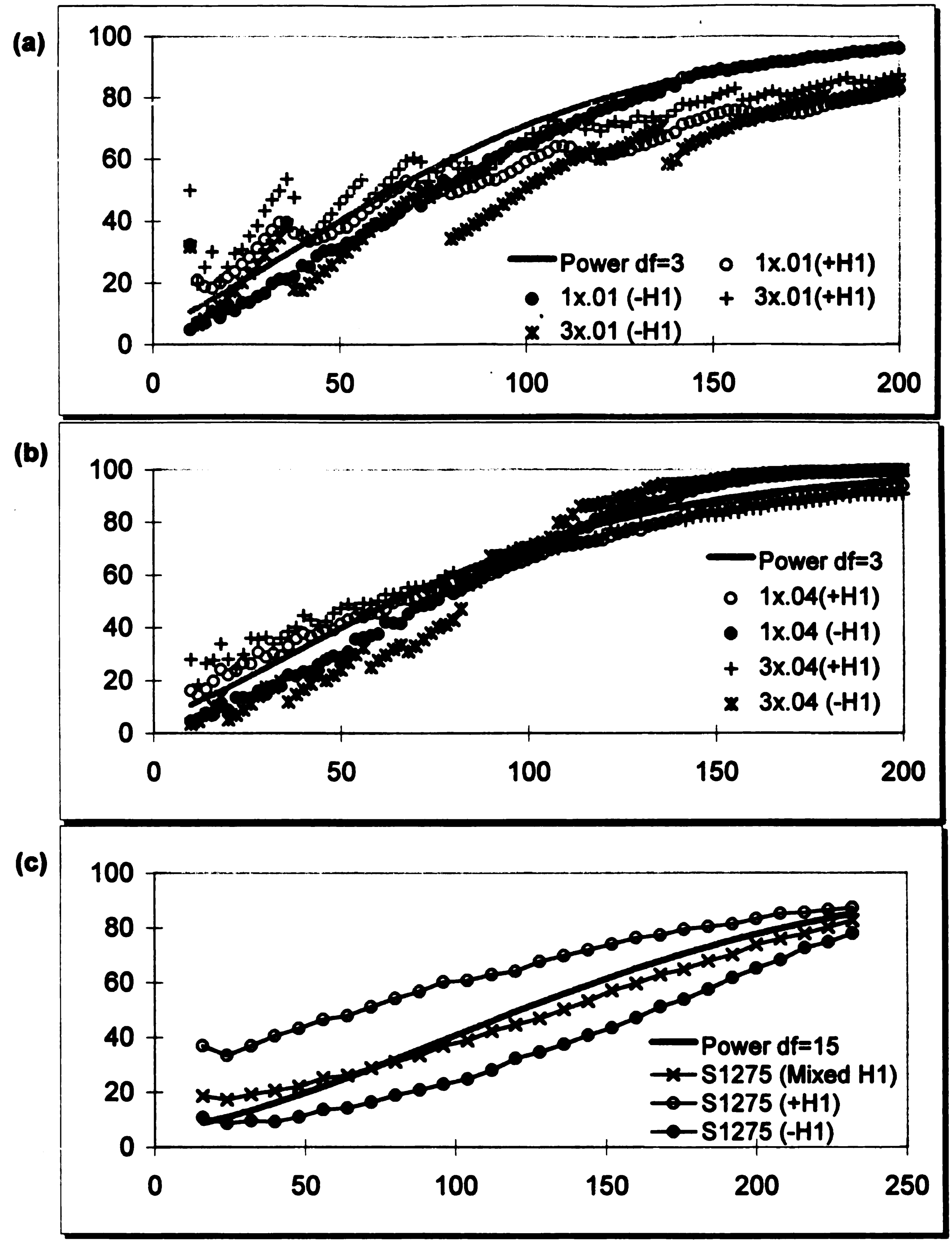


Figure 4-3. Rejection rates (%) versus sample size for alternative hypotheses: small cells increasing (+H1) or small cells decreasing (-H1), (a) and (b) 4-cell tables, n(p)=1 or 3, (c) 16-cell tables, n(p)=12.

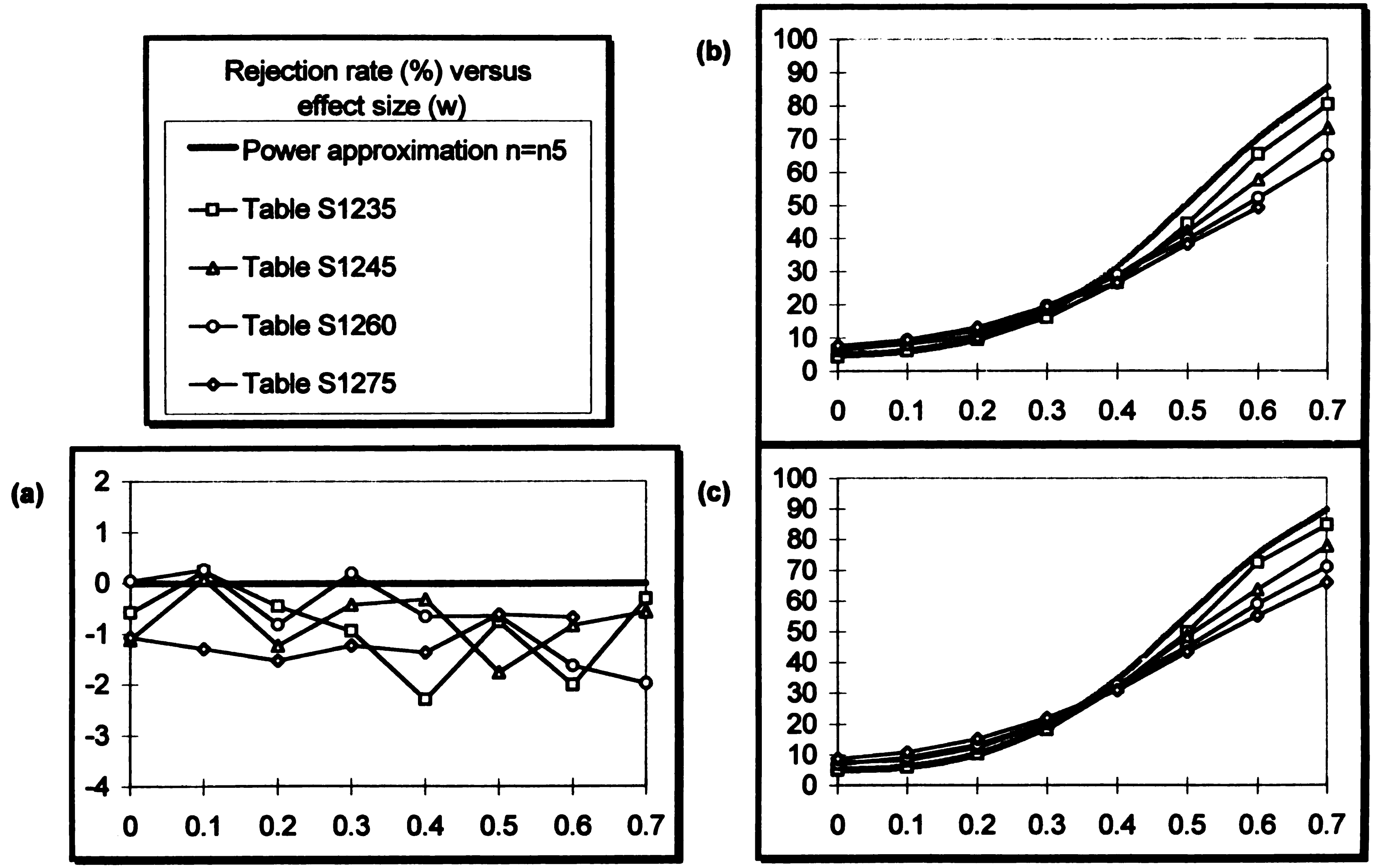


Figure 5-1. k=16, np=12, n=n5. (a) Difference in fit observed power to the asymptotic between (b) the product mutinomial case and (c) the multinomial case of the goodness of fit test.

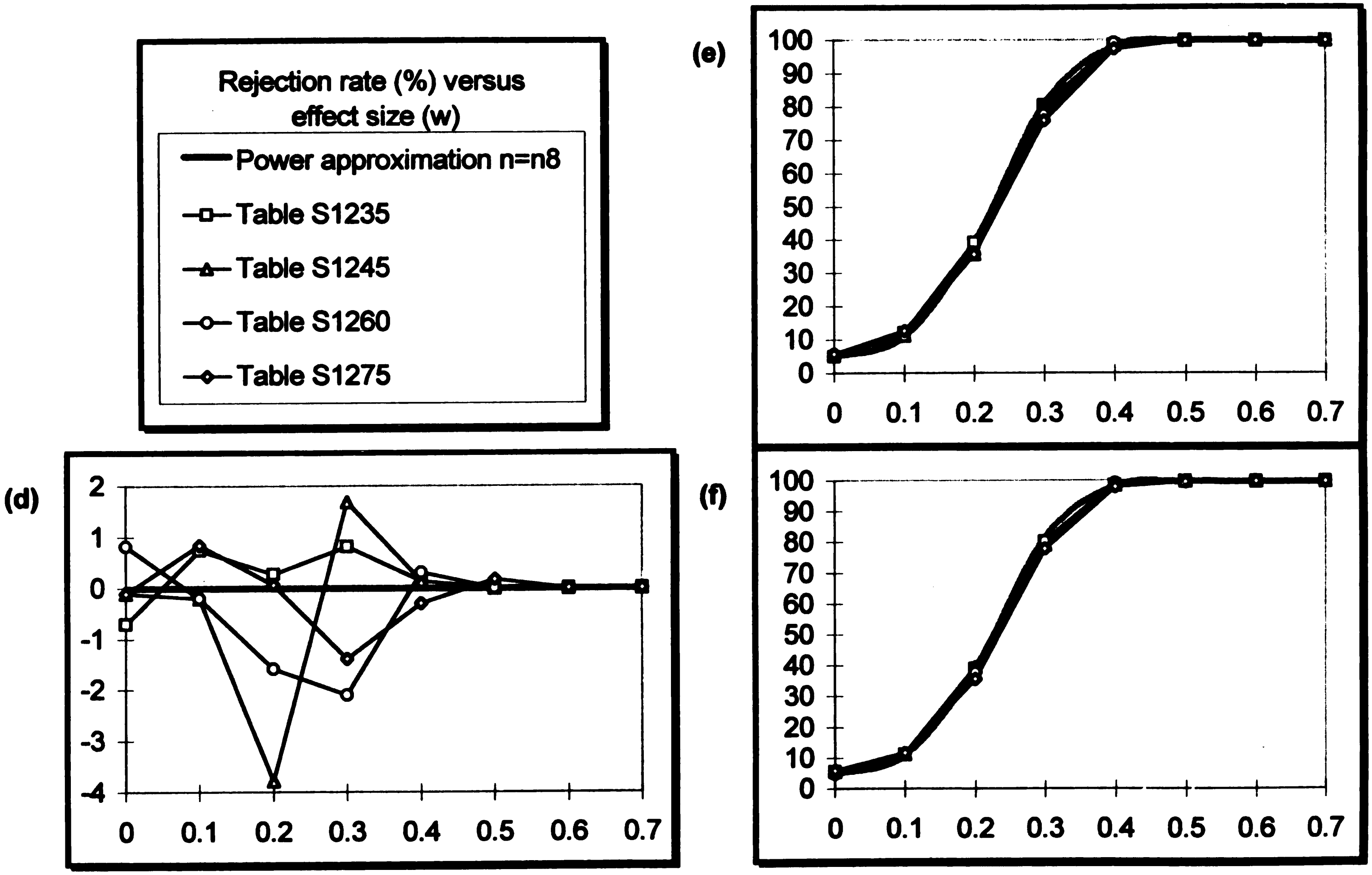


Figure 5-1 continued.

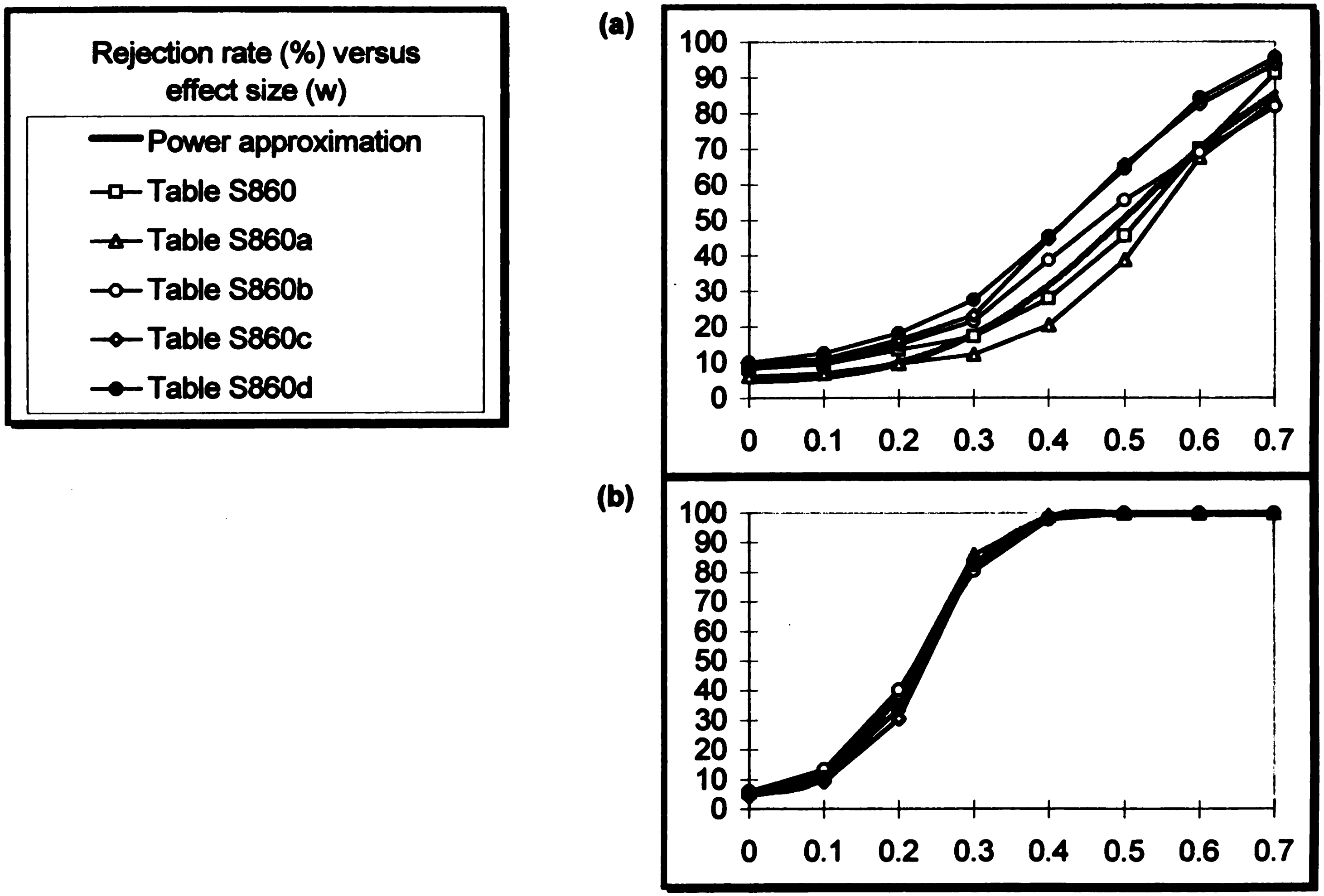


Figure 5-2. Unequal sample sizes, 860 series, (a) n5 = 40, (b) n8 = 196.

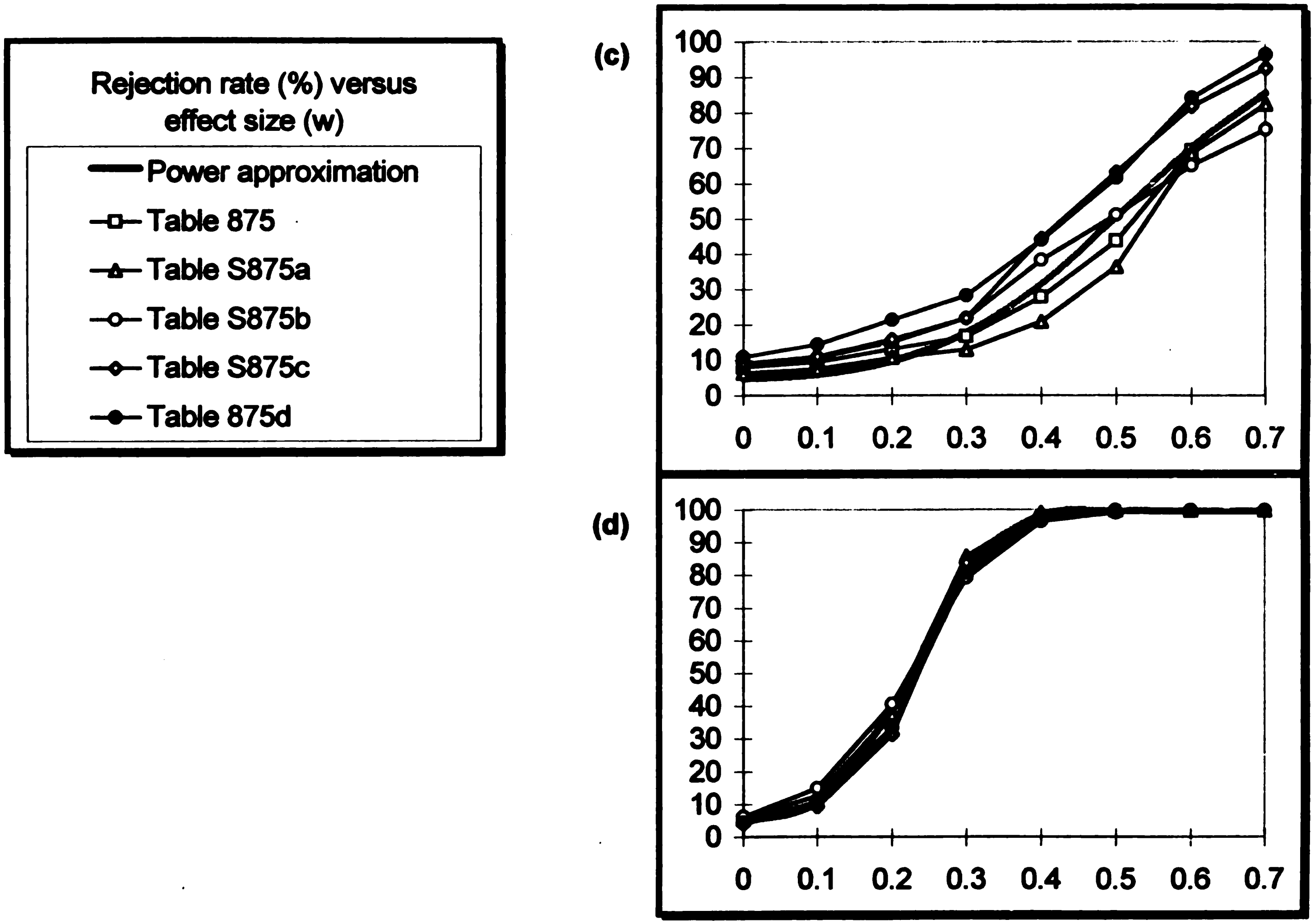


Figure 5-2 continued. 875 series (c) n5 = 40, (d) n8 = 196.

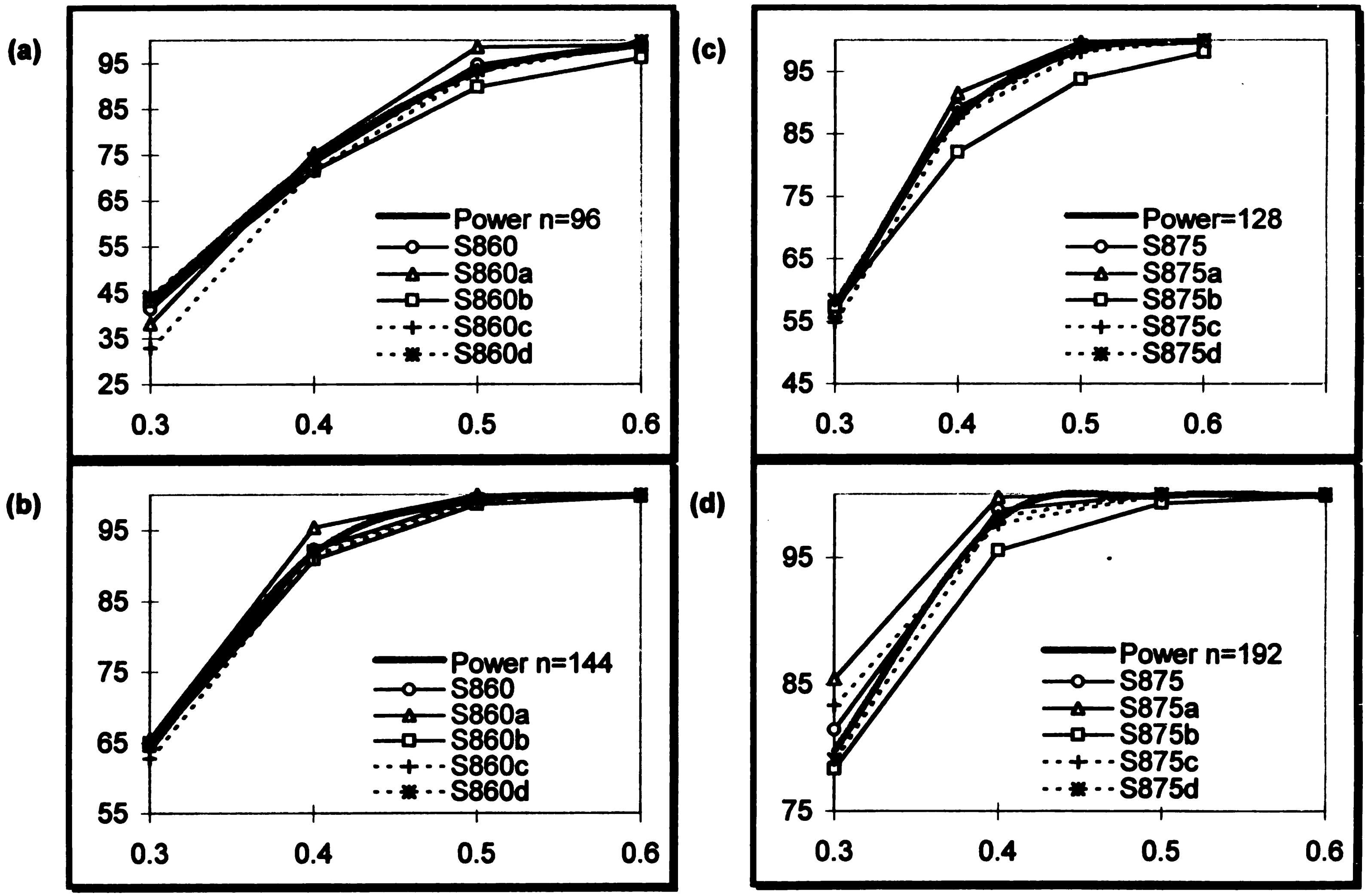


Figure 5-3. Distribution of e_{min} within rows. k=16, np=8, 860 and 875 series.

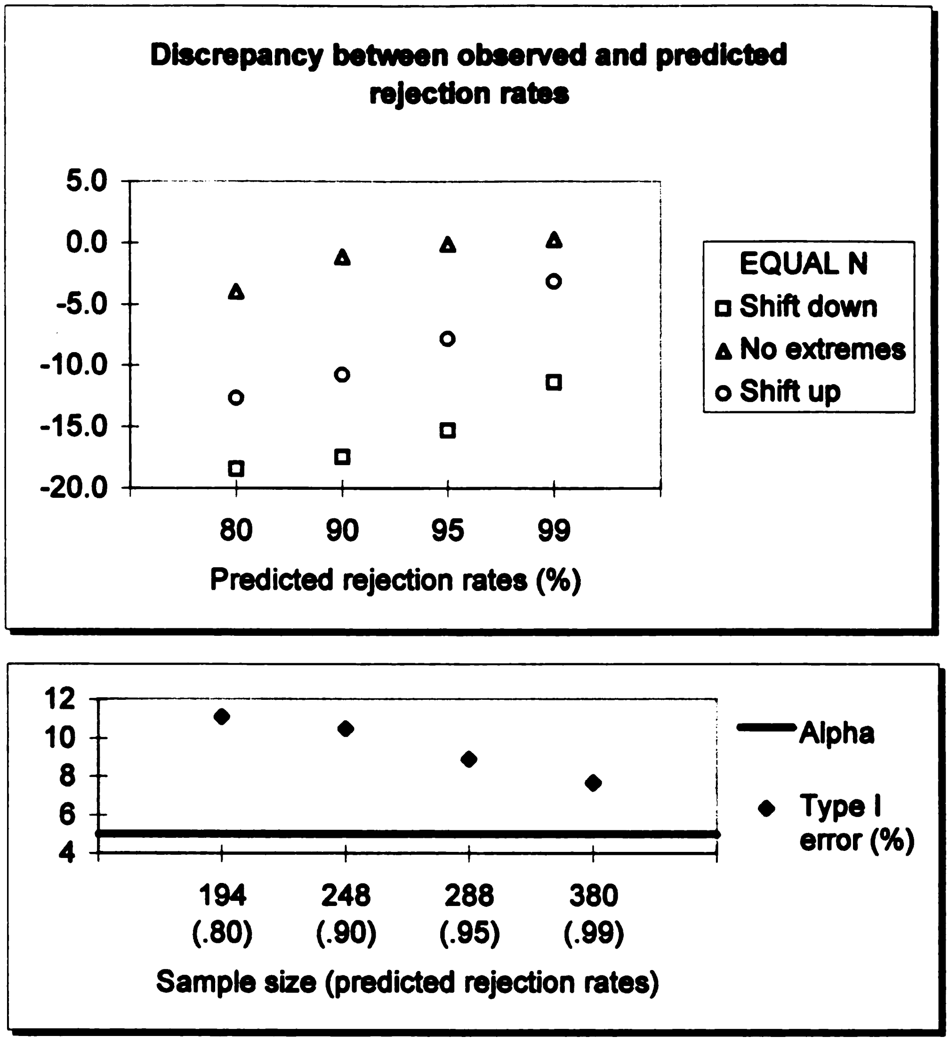


Figure 5-4. Application problem, confirmatory simulation results.

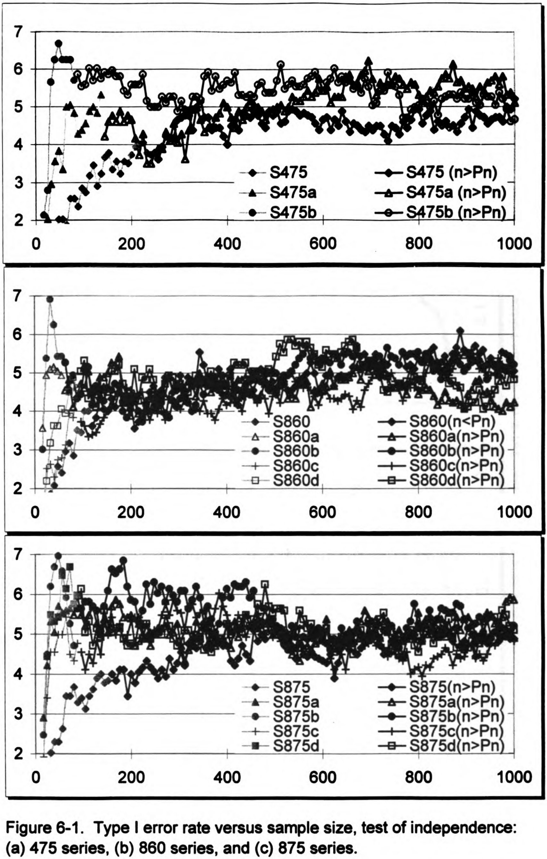


Figure 6-1. Type I error rate versus sample size, test of independence: (a) 475 series, (b) 860 series, and (c) 875 series.

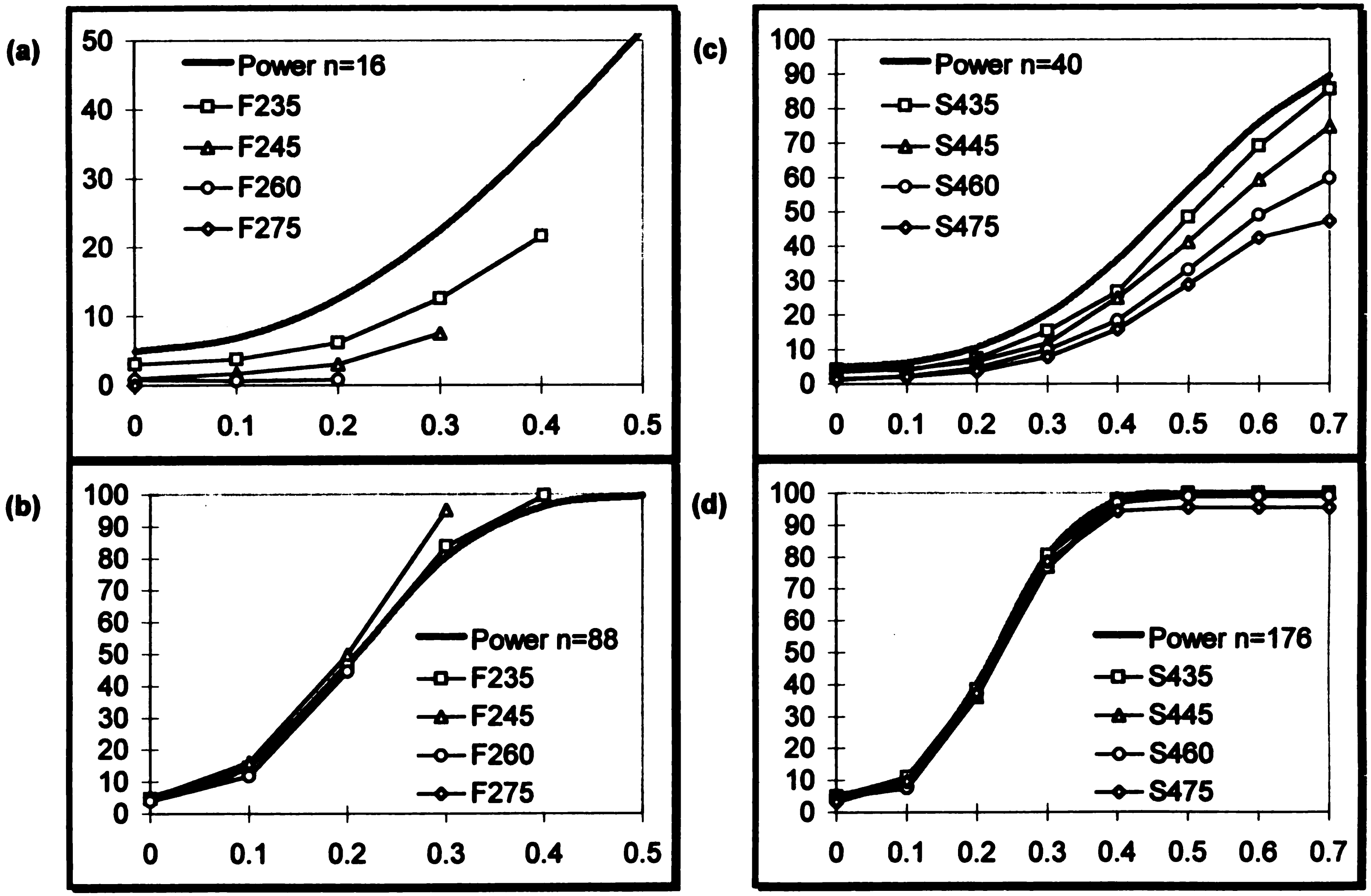


Figure 6-2. Power plots, test of independence: (a) & (b) 16-cell tables with 2 small cell expectations, (c) & (d) 16-cell tables with 4 small cell expectations.

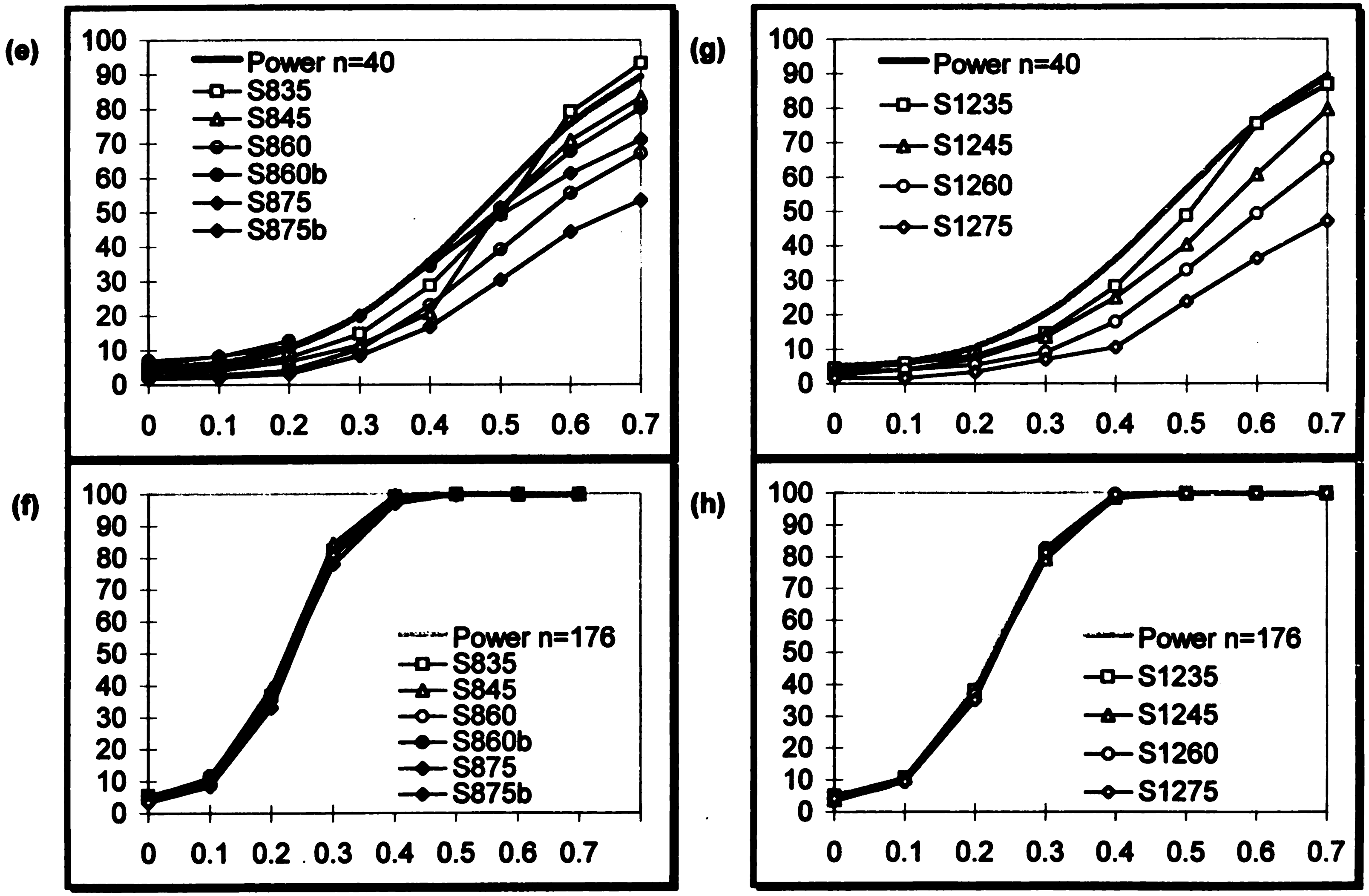


Figure 6-2 continued. (e) & (f) 16-cell tables with 8 small cell expectations, (g) & (h)16-cell tables with 12 small cell expectations.

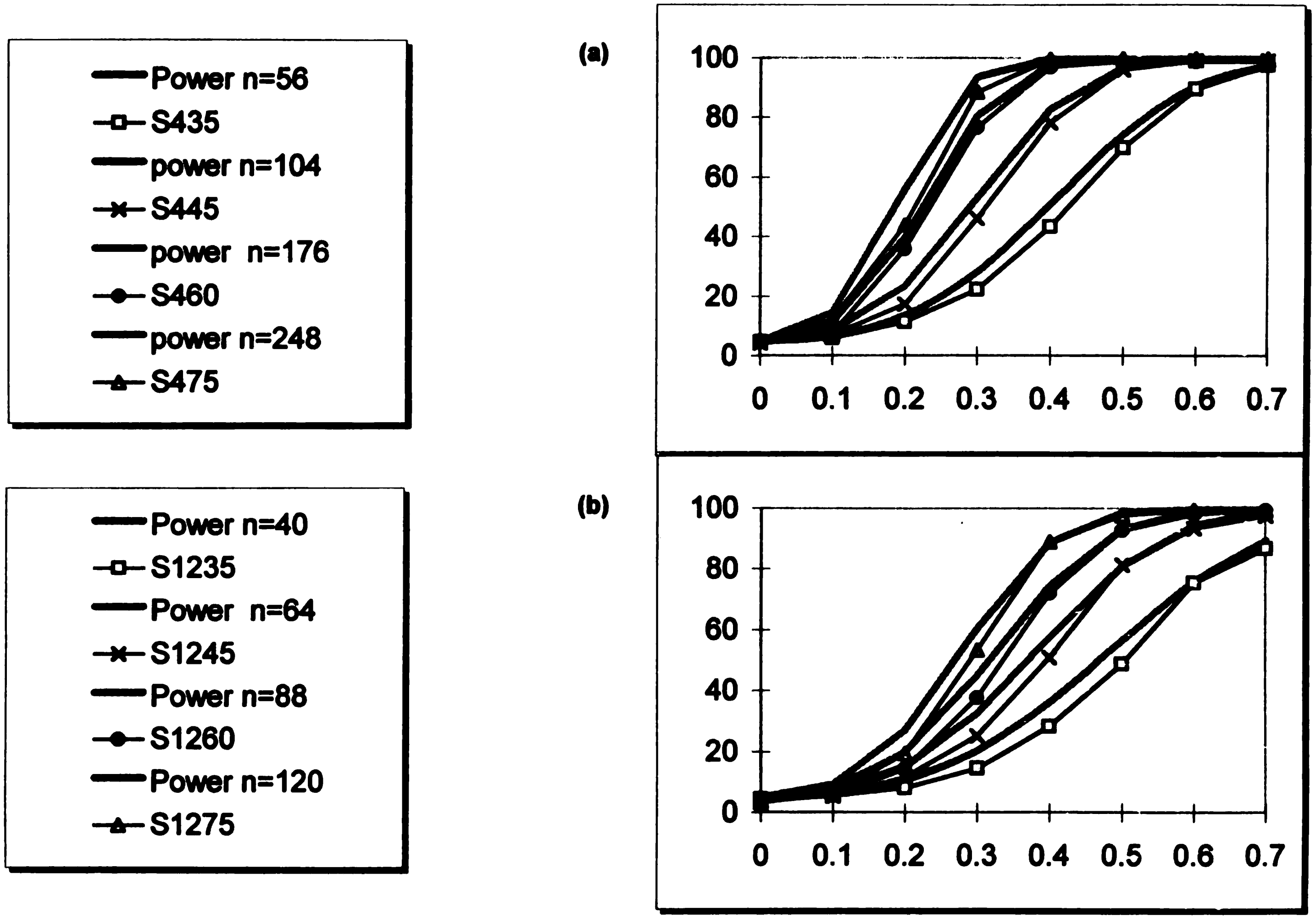


Figure 6-3. Power plots for n = Pn: (a) 4 small cell expectations (b) 12 small cell expectations.

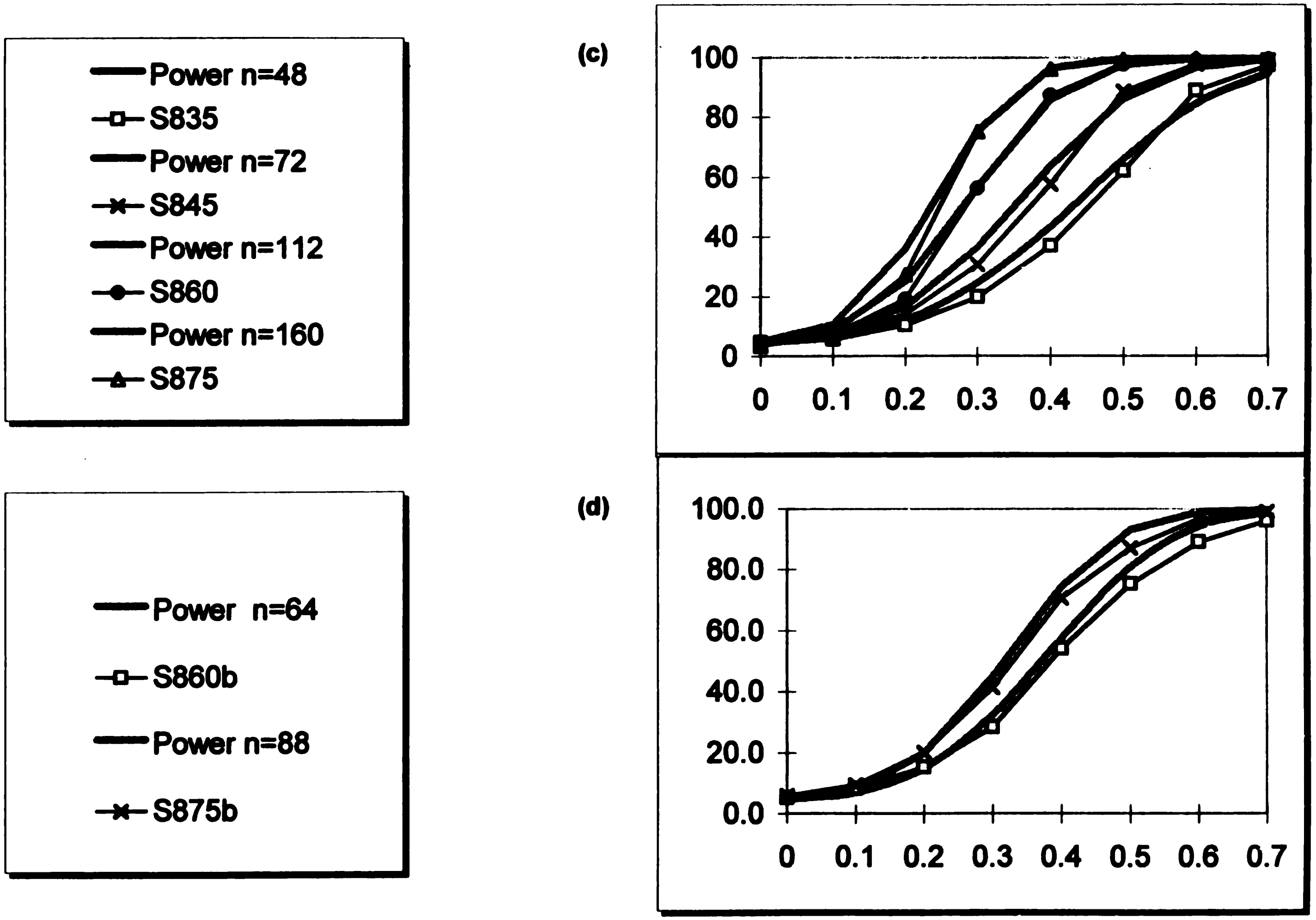


Figure 6-3 continued. (c) 8 small cell expectations, (d) 860 and 875 series.

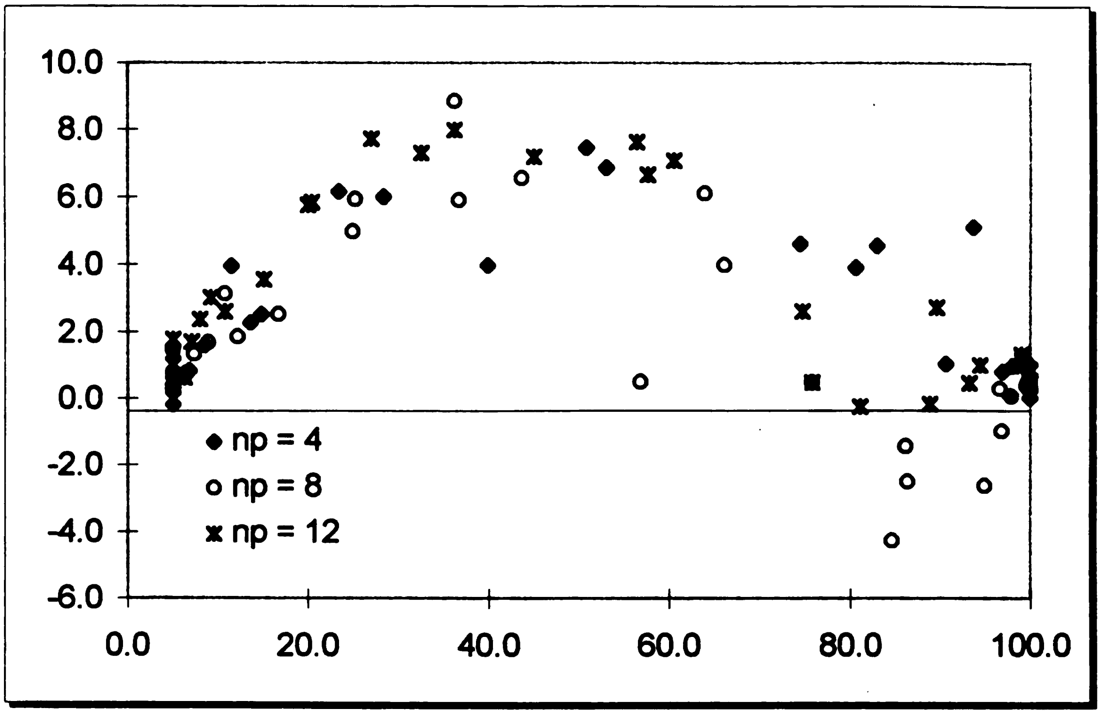


Figure 6-4. Differences between observed and expected power versus expected power.

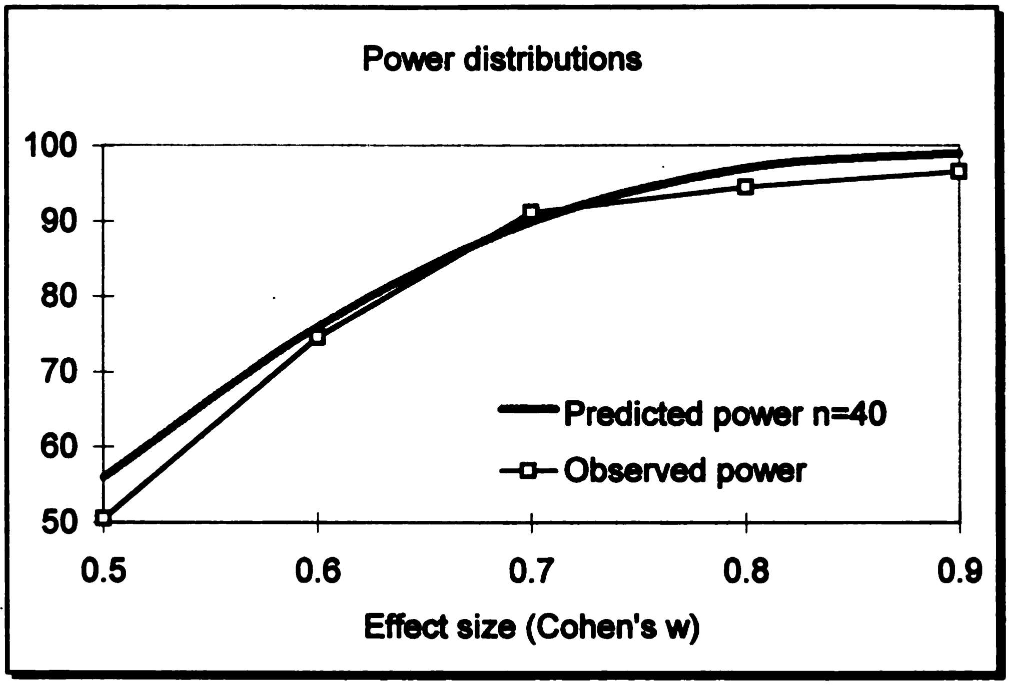


Figure 6-5. Application problem, confirmatory simulation results.

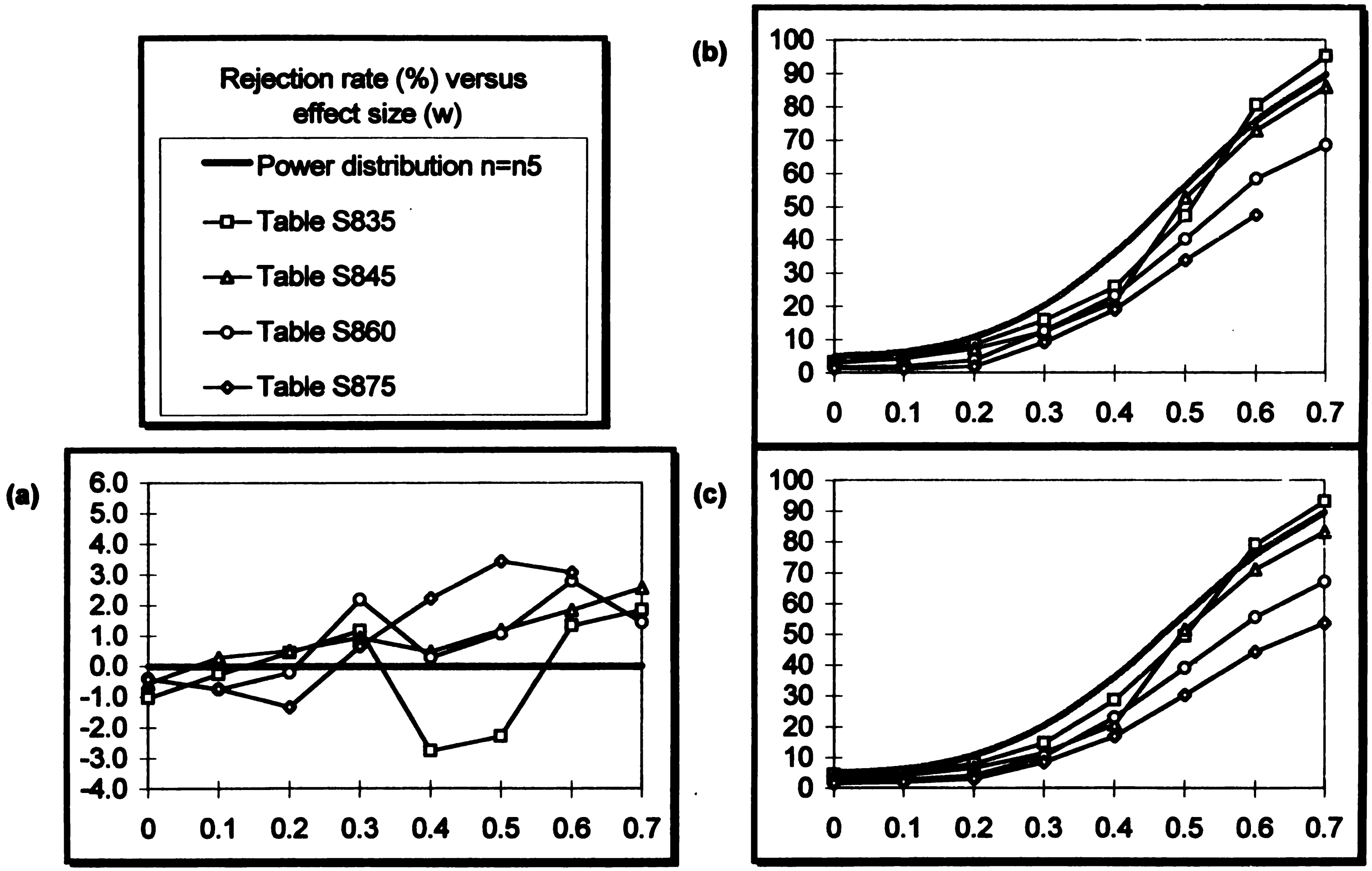


Figure 7-1. (a) Differences in fit (observed power - power approximation) between (b) the homogeneity test and (c) the test of independence. 16-cell tables with 8 small cell expectations.

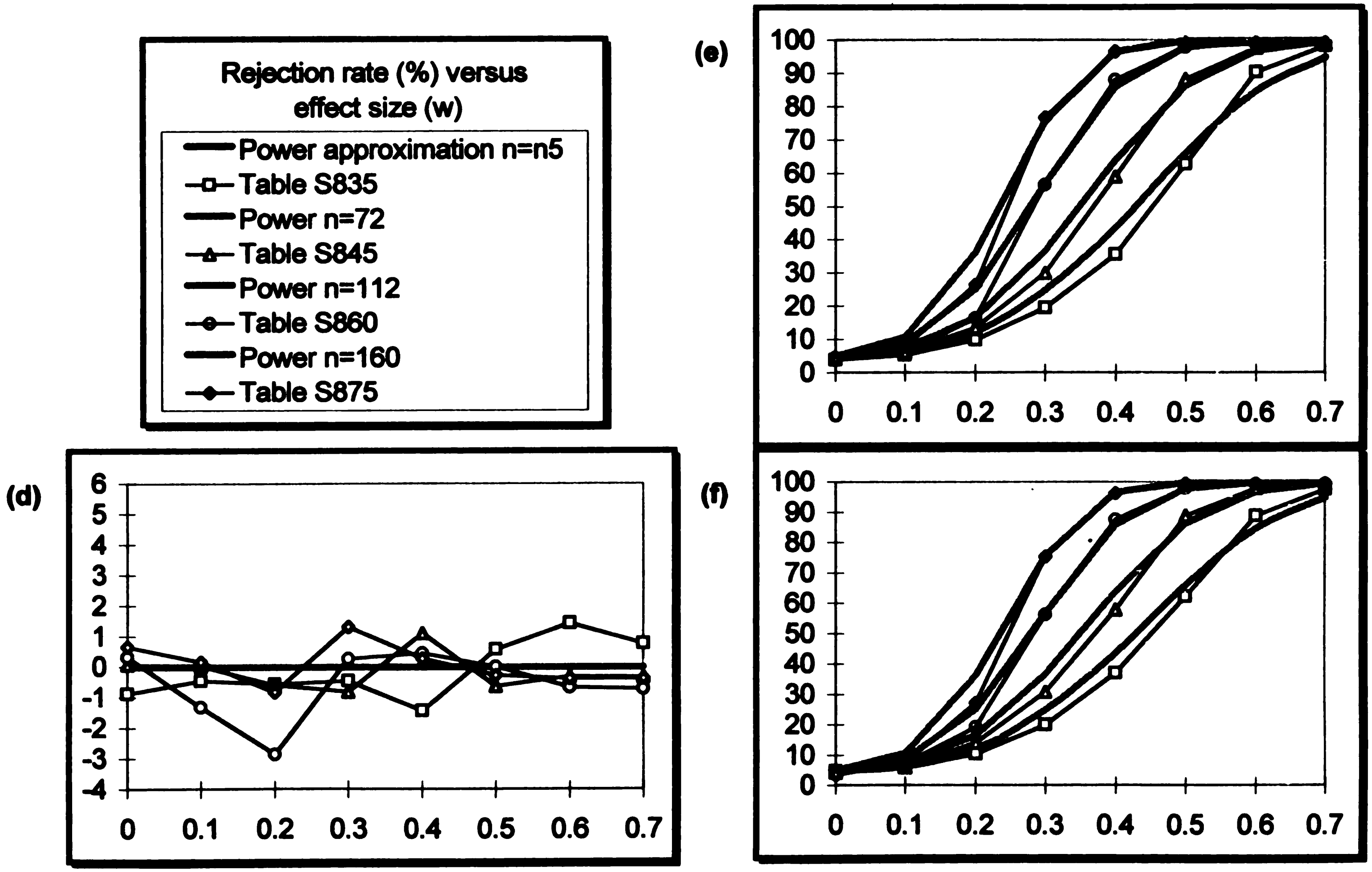


Figure 7-1 continued. The sample size is set to equal to Pn.

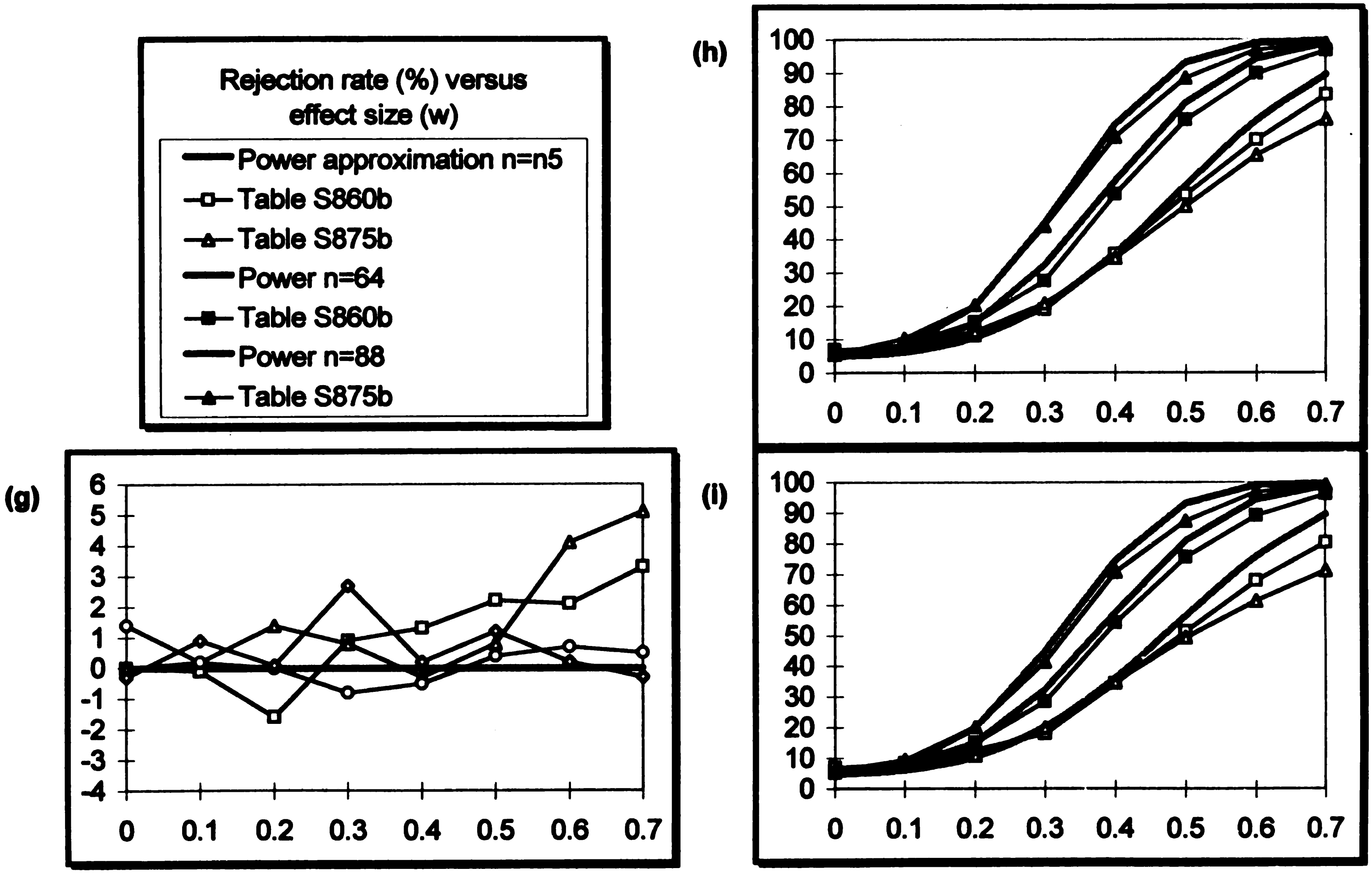


Figure 7-1 continued. 860 and 875 series.

BIBLIOGRAPHY

Bibliography

Agresti, A. (1990). *Categorical data analysis.* John Wiley & Sons.

Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979). Type I error rate of the chi-square test of independence in r x c tables that have small expected frequencies. *Psychological Bulletin, 86*(6), 1290-1297.

Bradley, D. R., & Seely, D. L. (1977). Empirical determination of the power of the chi-square test of independence in 2 x 2 tables. *Proceedings of the Statistical Computing Section of the American Statistical Association,* 138-144.

Camilli, G., & Hopkins, K. D. (1978). Applicability of chi-square to 2x2 contingency tables with small expected cell frequencies. *Psychological Bulletin, 85*(1), 163-167.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics, 23,* 315-345.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics, 10,* 417-451.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences , 2nd ed.* Hillsdate, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 9,* 168-173.

Craddock, J. M., & Flood, C. R. (1970). The distribution of the χ^2 statistic in small contingency tables. *Applied Statistics. Journal of the Royal Statistical Society, Series C, 19,* 173-181.

Fishman, G. S., & Moore, L. R. (1982). A statistical evaluation of multiplicative congruential random number generators with modulus 2^{31}- 1. *Journal of the American Statistical Association, 71,* 129-136.

Frosini, B. V. (1978). On the power function of the X^2 test. *Metron, 34,* 3-36.

Garside, G. R., & Mack, C. (1976). Actual Type I error probabilities for various tests in the homogeneity case of the 2x2 contingency table. *The American Statistician, 30*(1), 18-20.

Harkness, W. L. & Katz, L. (1964). Comparison of the power functions for the test of independence in 2x2 contingency tables. *Annals of Mathematical Statistics, 35,* 1115-1127.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29,* 58-65.

Horn, S. D. (1977). Goodness of fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics, 33,* 237-248.

Hayman, G. E. & Leone, F. C. (1964). Comparison of the power functions for the test of independence in 2x2 contingency tables. *Annals of Mathematical Statistics, 35,* 1115-1127.

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness of fit statistics for sparse multinomials. *Journal of the American Statistical Association, 75*(370), 336-344.

Koehler, K. J. (1986). Goodness of fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association, 81*(394), 483-493.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness of fit statistics. *Journal of the American Statistical Association, 73*(362), 253-263.

Lawal, H. B. (1992). A modified X^2 tests when some cells have small expectations in the multinomial distribution. *Journal of Statistical Computer Simulations, 40,* 15-27.

Lawal, H. B. & Upton, G. J. G. (1980). An approximation to the distribution of the X^2 goodness-of-fit statistic for use with small expectations. *Biometrika, 67* (2), 447-453.

Meng, R. C., and Chapman, D. G. (1966). *Journal of the American Statistical Association, 61,* 965-975.

Moore, D. S. (1986). Tests of chi-squared type. In R. B. D'Agostino and M. A. Stephens (Eds.), *Goodness-of-fit techniques.* NewYork: Marcel Dekker, Inc.

Ott, R. L. (1993). *An introduction to statistical methods and data analysis* (4th edition). Belmont, CA: Duxbury Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in Fortran, 2nd edition.* Cambridge University Press.

Read, T. T. C. & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer-Verlag.

Roscoe, J. T. & Byars, J. A. (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association, 66,* 336, 755-759.

SAS Institute Inc. (1990). *SAS language: Reference, version 6, first edition.* Cary, NC: SAS Institute, Inc.

Slakter, M. J. (1968). Accuracy of an approximation to the power of the chi-square goodness of fit test with small but equal expected frequencies. *Journal of the American Statistical Association, 63,* 912-924.

Von Eye, A. (1990). *Introduction to configural frequency analysis: The search for types and antitypes in cross-classifications.* Cambridge University Press.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Wise, M. E. (1963). Multinomial probabilities and the χ^2 and X^2 distributions. *Biometrika, 50,* 145-154.

Yarnold, J. K. (1970). The minimum expectation in X^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association, 65(330), 864-886.*