SCISPACE
formerly Typeset

Open access • Journal Article • DOI:10.1080/01621459.1980.10477473

An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials

— Source link

Kenneth J. Koehler, Kinley Larntz

Institutions: Iowa State University, University of Minnesota

Published on: 01 Jun 1980 - Journal of the American Statistical Association (Taylor & Francis Group)

Topics: Asymptotic theory (statistics), Goodness of fit, V-statistic, Asymptotic distribution and Pearson's chi-squared test

Related papers:

- Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics
- Goodness-of-fit tests for log-linear models in sparse contingency tables
- Multinomial goodness-of-fit tests
- The χ^2 Test of Goodness of Fit
- Central limit theorems for multinomial sums

Share this paper:

View more about this paper here: https://typeset.io/papers/an-empirical-investigation-of-goodness-of-fit-statistics-for-1hw2dnb6be

An Empirical Investigation of
Goodness-of-Fit Statistics
for Sparse Multinomials*

by

Kenneth J. Koehler[1] and Kinley Larntz[2]

Technical Report No. 327

September 1978

* This research was supported in part by Grant NIE-G76-0094 from the National Institute of Education, U.S. Department of Health, Education and Welfare.

1 Department of Statistics, Iowa State University

2 School of Statistics, University of Minnesota

ABSTRACT

Traditional discussions of goodness-of-fit tests for multinomial data consider asymptotic chi-squared properties under the assumption that all expected cell frequencies become large. However, this condition is not always satisfied and other asymptotic theories must be considered. For testing a specified simple hypothesis, Morris gave conditions for the asymptotic normality of the Pearson and likelihood ratio statistics when both the sample size and number of cells become large (even if the expected cell frequencies remain small). Monte Carlo techniques are used to examine the applicability of the normal approximations for moderate sample sizes with moderate numbers of cells.

KEY WORDS AND PHRASES: Asymptotic approximations, Likelihood ratio statistic, Pearson statistics, Chi-squared.

1. Introduction

The use of chi-squared tests for goodness of fit has become widespread since their introduction by Karl Pearson in 1900. Traditional consideration of large sample properties has depended upon the assumption that all expected cell frequencies become large. It is our contention that in many applications cell selection is dependent upon the sample size in such a way as to violate these traditional asymptotic assumptions. In this paper we explore the practical importance of recent results of Morris as they relate to this statistical question.

Let $(N_1, N_2, \ldots, N_k)$ be a multinomial random vector with probability parameter $\underset{\sim}{p} = (p_1, p_2, \ldots, p_k)$ such that $n = \sum_{i=1}^{k} N_i$ and $1 = \sum_{i=1}^{k} p_i$. Consider the problem of testing the null hypothesis H_0: $\underset{\sim}{p} = \underset{\sim}{q}$, for some completely specified probability vector $\underset{\sim}{q}$, against all possible alternatives. The test most frequently used is the one suggested by Karl Pearson (1900) which rejects H_0 for sufficiently large values of

$$X_k^2 = \sum_{i=1}^{k} (N_i - nq_i)^2/nq_i \quad . \tag{1.1}$$

This statistic will be referred to as the Pearson goodness-of-fit statistic.

The use of the likelihood ratio test statistic was proposed by J. Neyman and E. Pearson (1926). The likelihood ratio statistic rejects H_0 for large values of

$$G_k^2 = 2 \sum_{i=1}^{k} N_i \log(N_i/nq_i) \quad . \tag{1.2}$$

This statistic has become more popular as the availability of high-speed computers has increased.

When H_0 is true, both statistics are well known to have the same limiting central chi-squared distribution under the traditional limiting argument which requires that $\min_{1\leq i\leq k} np_i \to \infty$ as $n \to \infty$. Therefore, when all expected frequencies are large the chi-squared distribution can be used to establish approximate critical regions for each test statistic. However, it is not uncommon in practice to use the sample size to determine the number of cells. Then the number of cells is generally increased when the sample is increased. In that case, both X_k^2 and G_k^2 can be shown to have asymptotic normal distributions under conditions which allow both n and k to become large without necessarily requiring that $\min_{1\leq i\leq k} np_i \to \infty$. These conditions are reviewed in Section 2. Monte Carlo methods are used in Sections 4 and 5 to assess accuracy of the asymptotic normal and chi-squared approximations to the distributions of the Pearson and likelihood ratio test statistics for moderate numbers of cells and moderate sample sizes.

2. Asymptotic Normality

Several authors have demonstrated the asymptotic normality of certain goodness-of-fit statistics under conditions which do not require that all expected frequencies become large as the sample size increases.

The number of cells must increase with the sample size. A simple example is the test for a uniform distribution on a fixed interval where the interval is partitioned into a number of subintervals of equal length. If it is desired to achieve a specified expected frequency

λ for each subinterval, then k subintervals are used for a sample size of n, where k is selected to make n/k close to λ. If n is increased k would also be increased.

This leads to the consideration of the limiting distributions of goodness-of-fit statistics for sequences of multinomials of increasing dimension. Consider the sequence of multinomial random vectors

$$\{(N_{1,k(i)}, N_{2,k(i)}, \ldots, N_{k(i),k(i)})\}_{i=1}^{\infty}$$

where the i-th vector in the sequence has $k(i)$ cells, with sample size $n_k = \sum_{j=1}^{k} N_{j,k}$ and probability vector $(p_{1k}, p_{2k}, \ldots, p_{kk})$ with $1 = \sum_{j=1}^{k} p_{jk}$. (The underlying subscript i is hereafter suppressed to simplify notation.) We will require the sample size n_k to increase as k increases. Since the asymptotic moments for the statistics are derived from independent Poisson frequencies we need to define a corresponding sequence of Poisson random vectors. For each multinomial vector $(N_{1k}, N_{2k}, \ldots, N_{kk})$, let $(Y_{1k}, Y_{2k}, \ldots, Y_{kk})$ be a vector of independent Poisson random variables such that $E(Y_{ik}) = E(N_{ik})$.

Morris (1966, 1975) generalized a conditioning argument given by Steck (1957) to obtain a central limit theorem for sums of functions of multinomial frequencies. The method requires that the sum of functions of independent Poisson frequencies has a limiting normal distribution. Then under mild conditions the asymptotic normality of the sum under the multinomial distribution can be obtained by conditioning on the sum of the independent Poisson frequencies.

As special cases, Morris (1975) derived central limit theorems for the Pearson and likelihood ratio statistics. Although asymptotic normality is valid for certain classes of alternatives we only consider the case where the null hypothesis is true in this paper. In that case sufficient conditions for asymptotic normality as $k \to \infty$ are

(i) $\max_{1 \leq i \leq k} p_{ik} = o(1)$ as $k \to \infty$ and

(ii) $n_k p_{ik}$ is uniformly bounded below by some constant.

These conditions are not necessary and other sets of sufficient conditions have been given by Steck (1957) and Holst (1972, 1976).

When the null hypothesis is true the asymptotic mean and variance for the Pearson statistic are given by

$$\mu_{P,k} = k \tag{2.1}$$

and

$$\sigma^2_{P,k} = 2k + \sum_{j=1}^{k} (1 - k^{-1} p_{jk})/n_k p_{jk} \tag{2.2}$$

However, it can be shown that Morris's central limit theorem for the Pearson statistic is valid when $\mu_{P,k}$ and $\sigma^2_{P,k}$ are replaced by the corresponding exact moments. Exact moments for the Pearson statistic were derived by Haldane (1937) and it is easily seen that

$$E(X^2_k) = \mu_{P,k} - 1 \tag{2.3}$$

and

$$\mathrm{Var}(X^2_k) = \sigma^2_{P,k} - 2\left[1 + \frac{k-1}{n_k}\right] . \tag{2.4}$$

The effect on the accuracy of the normal approximation from replacing $\mu_{P,k}$ and $\sigma^2_{P,k}$ by the exact values is examined in Sections 4 and 5. Note that $\sigma^2_{P,k}$, and consequently $\mathrm{Var}(X^2_k)$, can be much larger than the chi-squared variance on $k - 1$ degrees of freedom when the expected frequencies are not all equal.

The asymptotic moments for the likelihood ratio statistic are also derived from independent Poisson random variables. Define the Poisson information kernel by

$$I(y,m) = \begin{cases} y \log(y/m) - y + m \,, & \text{if } y > 0 \\ m \,, & \text{if } y = 0 \end{cases} \quad .$$

Then the first two asymptotic moments are given by

$$\mu_{LR,k} = 2\sum_{j=1}^{k} E[I(Y_{jk},\ n_k p_{jk})] \tag{2.5}$$

and

$$\sigma^2_{LR,k} = 4\sum_{j=1}^{k} Var[I(Y_{jk},\ n_k p_{jk})] - n_k \gamma_k^2 \tag{2.6}$$

where

$$\gamma_k = \frac{2}{n_k} \sum_{j=1}^{k} Cov[I(Y_{jk},\ n_k p_{jk}),\ Y_{jk}] \quad .$$

An examination of these asymptotic moments is useful in determining when the asymptotic chi-squared approximation is appropriate. A graph of $E[I(Y,\ m)]$ is presented in Figure A for a Poisson random variable Y with mean m. The rapid decline of $E[I(Y,\ m)]$ as $m \to 0$ indicates that $\mu_{LR,k}$ can be much smaller than the chi-squared mean when many expected frequencies are smaller than one-half. However, the graph also shows that $\mu_{LR,k}$ is substantially larger than $k - 1$ when most expected frequencies are between one and five. The mean of the likelihood ratio statistic is close to $k - 1$ when almost all expected frequencies are large.

--- Insert Figure A about here ---

The graphs of [Var I(Y, m)] and Cov[I(Y, m), Y] presented in Figures B and C give a good indication of the behavior of $\sigma^2_{LR,k}$. The asymptotic variance can be much smaller than 2(k - 1) when most expected frequencies are smaller than one, but it is larger than 2(k - 1) when most expected frequencies are moderate. These figures indicate that the chi-squared approximation for the likelihood ratio statistic may give inflated critical levels when most expected frequencies are moderate and extremely conservative critical levels when most expected frequencies are smaller than one-half.

--- Insert Figures B and C about here ---

It is interesting to note that Pearson and likelihood ratio statistics have different limiting normal distributions as $k \to \infty$. The difference in behavior is largely due to the differing influence given to very small observed counts by the statistics. This effect was described by Larntz (1978) for expected frequencies in the range of 2.0-5.0. Here we examine the effect for smaller expected frequencies. Table 1 illustrates the general pattern. For a cell with an expected frequency larger than one an observed count of zero or one makes a larger minimum contribution to G^2_k than X^2_k. Consequently, when most expected cell frequencies are in the range of 1.0-5.0 the first two moments for G^2_k are larger than those for X^2_k. However, the contribution to X^2_k for a nonzero count can be quite large when the expected frequency is less than one, and the first two moments for X^2_k are larger than the corresponding moments for G^2_k when a sufficient number of expected frequencies are less than one.

--- Insert Table 1 about here ---

3. Monte Carlo Procedures

A Monte Carlo study was performed to assess the accuracy of the asymptotic chi-squared and normal approximations for moderate cell sizes when most expected frequencies do not exceed five. The objectives were (a) to determine when the normal approximation is sufficiently more accurate to justify the additional computation, (b) to examine how the accuracy of the asymptotic approximations is affected by departures from the conditions imposed by the central limit theorems, and (c) to determine when the use of exact means and variances provides better normal approximations.

In this study values of X_k^2 and G_k^2 were simulated for multinomials with 3, 4, 10, 40, 100, 400, and 1000 cells. For each cell size, sample sizes were selected such that $\lambda = n_k/k$ achieved the values 1/4, 1/2, 1, 2, 3, and 5 as closely as possible. Some cases with 400 and 1000 cells were omitted because of the extreme computational cost. For each of the nine null hypotheses selected and each combination of λ and k, 2500 multinomial random vectors were simulated. Each multinomial vector was used to produce a value for X_k^2 and G_k^2. Therefore, the X_k^2 and G_k^2 values are correlated.[1]

---

[1] Computations were performed using FORTRAN programs on a CDC 6600 computer. Multinomials were generated from uniform random numbers by classifying the uniforms into k categories [see Koehler (1977)]. The uniform random numbers were produced by a multiplicative congruential generator using modulus 2^{47} and multiplier 5^{17}.

Denote the probability simplex by

$$T_k = \{\underset{\sim}{p}_k \in R^k : \sum_{i=1}^{k} p_{ik} = 1 \text{ and } p_{ik} \geq 0 \text{ for all } 1 \leq i \leq k\} ,$$

and let

$$T_k^+ = \{\underset{\sim}{p}_k \in T_k : p_{1k} \geq p_{2k} \geq \cdots \geq p_{kk}\} .$$

Any point in T_k can be obtained from a permutation of the coordinates of some point in T_k^+; therefore only null hypotheses in T_k^+ need be considered. The nine null hypotheses examined in this study are labeled in Table 2. Null hypothesis 1 is the center of T_k and will be referred to as the hypothesis of symmetry. The other points were selected to cover a wide range of T_k^+. Hypothesis 5 is the center of gravity of T_k^+ when mass is uniformly distributed over T_k^+.

The accuracy of the asymptotic normal approximation was examined for three standardized versions of X_k^2 and G_k^2. The first two exact moments for X_k^2 were computed directly, but the first two exact moments for G_k^2 were estimated by a Monte Carlo procedure which uses X_k^2 as a control variate. The standardized statistics are denoted by the following symbols.

P_E --- X_k^2 standardized with $E(X_k^2)$ and $Var(X_k^2)$.

LR_E --- G_k^2 standardized with Monte Carlo estimates of the exact mean and standard deviation.

P_A --- X_k^2 standardized with $\mu_{P,k}$ and $\sigma^2_{P,k}$.

LR_A --- G_k^2 standardized with $\mu_{LR,k}$ and $\sigma^2_{LR,k}$.

P_C --- X_k^2 standardized with the mean and standard deviation of a chi-square random variable with

$k-1$ degrees of freedom.

LR_C --- G_k^2 standardized in the same manner as P_C .

Since a standardized chi-squared random variable converges in distribution to a standard normal random variable as the degrees of freedom increase, P_C and LR_C will be well approximated by the standard normal distribution for large k when the chi-squared distribution provides an adequate approximation for the distribution of X_k^2 and G_k^2 respectively.

Rejection levels and percentiles were simulated for all six of the standardized statistics for nominal levels, .001, .005, .01, .025, .05, .1(.1).9, .95, .975, .99, .995, .999. Complete tables for the .01 and .05 levels are available from the authors. Some special cases are presented in the next two sections to illustrate general trends.

4. The Symmetrical Case

Small sample properties of goodness-of-fit statistics have been most frequently studied for the null hypothesis of equal cell probabilities. One reason is that many goodness-of-fit problems can be transformed into the problem of assessing the goodness-of-fit of the uniform distribution on the unit interval and in that case it is reasonable to select cells of equal widths. Second, the computation, of exact probability levels is relatively simple since X_k^2 and G_k^2 are invariant under permutations of the observed frequencies when all cell probabilities are equal. In this section we examine the distributions of X_k^2 and G_k^2 under the null hypothesis of symmetry.

Exact probability levels for the Pearson statistic have been examined by Vessereau (1958), Nass (1958), Slakter (1966), Good, et. al (1970), Zahn and Roberts (1971), and Katti (1973) for small numbers of cells with equal expectations. Their consensus opinion is that the traditional chi-squared approximation does not introduce serious absolute errors at nominal levels .05 and .01 in the upper tail when $n \geq 10$. Zahn and Roberts recommend that $n \geq 25$ when the chi-squared approximation is used at similar nominal levels in the lower tail.

Citing Verrereau's work some authors have suggested that an appropriate rule for deciding when the chi-squared approximation for X_k^2 is adequate is to use the chi-squared approximation whenever $n \geq n_0$, where n_0 is a fixed positive integer. However, any fixed n_0 will be inadequate when k is sufficiently large. A more appropriate criterion is to require $n^2/k > c$ for some constant c. Unless n^2/k is sufficiently large, the Pearson statistic will have a high probability of assuming its minimum value and will not allow for an adequate continuous approximation. Our Monte Carlo results indicate that the chi-squared approximation is reasonably adequate for the symmetrical case when $k \geq 3$, $n \geq 10$, and $n^2/k \geq 10$.

It should be noted that $n^2/k \to \infty$ is a necessary condition for X_k^2 to have a limiting normal distribution as $k \to \infty$ under any null hypothesis. Therefore, the rule $n^2/k > c$ is also an appropriate guideline for the application of the normal approximation.

The distribution of the likelihood ratio statistic is generally not well approximated by the chi-squared distribution when $\lambda \leq 5$. Unlike the moments of X_k^2, the mean and variance of G_k^2 do not closely match the corresponding moments of the chi-squared distribution with $k - 1$ degrees of freedom. As noted in Section 2, the mean and variance of G_k^2

are smaller than the chi-squared moments when $\lambda < 0.5$ and larger when $\lambda > 1$. Hence the chi-squared approximation produces conservative critical levels in the first case and liberal critical levels in the latter case. The simulated critical levels for LR_C presented in Figures D, E, F, and G illustrate how extremely inaccurate the chi-squared approximation can be when the number of cells is moderately large. When $\lambda = 0.5$, $\mu_{LR,k} = (1.007)k$ and $\sigma^2_{LR,k} = (.56)k$ and the estimated critical level drops to 0.0024 at $k = 100$ for the .05 nominal level. When $\lambda = 1$, the estimated critical level is .9776 at $k = 1000$ for the .05 nominal level. Critical levels are most liberal when λ is close to 2, but even when $\lambda = 5$ the estimated critical level is 0.126 at $k = 100$ for the .05 nominal level.

The inadequacy of the chi-squared approximation was previously noticed by Good, et. al. (1970) who stated that "The distribution of the likelihood ratio statistic is by no means as well approximated by the chi-squared distribution as that of X^2 when $n/k < 1$." Larntz (1978) observed that the likelihood ratio statistic yields exact levels in excess of the nominal levels when the minimum expected frequencies are between 2 and 4.

Fortunately the standard normal distribution provides a good approximation for the right tail of the LR_A distribution. The normal approximation for LR_A is quite adequate at the .05 and .01 nominal levels when $k \geq 3$, $n \geq 15$, and $n^2/k > 10$. Figures D, E, F, and G show that the normal approximation is appreciably more accurate for LR_A than LR_E at the .01 nominal level. In addition, the normal approximation is generally more accurate for LR_A than either P_A or P_E for nominal levels smaller than .05 and moderate values of k. The estimated rejection levels for

P_A and P_E tend to be too large. The evaluation of the first four central moments indicates that the skewness converges to zero and the kurtosis converges to three faster for LR_A than for either P_A or P_E as $k \rightarrow \infty$.

--- Insert Figures D, E, F, and G about here ---

Monte Carlo power comparisons showed that for the null hypothesis of symmetry, X_k^2 is slightly more powerful for near alternatives. The Pearson test is decidedly dominant as the alternative moves toward a boundry of T_k which contains a high proportion of zeros and a few relatively large probabilities. The likelihood ratio test is dominant at alternatives which lie near boundries of T_k which contain a small proportion of near zero probabilities and have nearly equal probabilities in the remaining cells. This pattern agrees with observations made by West and Kempthorne (1971) from exact computations for 2, 3, and 4 cell examples. The boundries near which G_k^2 is dominant become close to the symmetrical null hypothesis in the Euclidean sense as k becomes large, but the areas where X_k^2 is more powerful may not. This indicates that X_k^2 is more powerful than G_k^2 for a very large portion of the simplex when k is moderately large.

5. SOME UNSYMMETRICAL CASES

General rules are more difficult to prescribe for unsymmetrical null hypotheses. In an extremely influential paper, Cochran (1954) gave a set of recommendations for the use of the chi-squared approximation for the Pearson statistic which generally require most expected frequencies to be at least five but allow a few to be between one and five. Vessereau found Cochran's recommendations to be stringent for the cases he considered, but he noticed that the chi-squared approximation for the Pearson statistic tends to produce inflated critical regions when small unequal expected frequencies are present. This phenomenon was partially explained in Section 2 where it was noted that under a null hypothesis with many small, unequal frequencies the variance of X_k^2 can be much larger than $2(k-1)$ but the mean is $k-1$.

Roscoe and Byars (1971) examined the chi-squared approximation for the Pearson statistic for the hypthesis of symmetry and two levels of skewness. Under their most extreme level of skewness they recommend that the chi-squared approximation be used at the .05 level only when $\lambda \geq 2$ and at the .01 level when $\lambda \geq 4$, where $\lambda = n/k$. The rule proposed by Roscoe and Byars works for their special cases, but in general no rule based solely on a minimum value of λ can hold under all unsymmetrical null hypotheses.

--- Insert Figures H and I about here ---

Monte Carlo rejection rates for all nine null hypothese are given in Figures H and I. In this study null hypothese 2, 3, and 4 all have at least one cell probability which does not become small for large values of k, and at least k-2 cells with small, equal probabilities. The chi-squared approximation for X_k^2 also gives very liberal critical regions under hypothesis 3, but they are not quite as bad as those under hypothesis 2. Hypothesis 4 is close enough to the center of the simplex so that the variance is not greatly inflated and, therefore, the chi-squared approximation is reasonably accurate for X_k^2. These observations are supported by the summary of Monte Carlo results given in Figures A and B for the nominal .05 and .01 levels. Figures J and K show that the chi-squared approximation for G_k^2 gives very conservative critical levels under hypothesis 2 when $\lambda \leq 5$. For this case the variance and mean of G_k^2 are smaller than the chi-squared moments. However, the general behavior of the chi-squared approximation for G_k^2 is exhibited under hypotheses 3 and 4. For those hypotheses the k-2 smallest expected frequencies are identical. The chi-squared approximation gives conservative critical levels when these expected frequencies do not exceed 0.5 and it gives quite liberal critical levels when these expected frequencies are between 1 and 5. The critical regions given by the chi-squared approximation are most liberal when these expected frequencies are near 2 and become increasingly conservative when the expected frequencies are made smaller

--- insert Figures J and K about here ---

Figures J and K indicate that the normal approximation is much more accurate than the chi-squared approximation under null hypotheses 2 when n^2/k is sufficiently large. Standardization by the asymptotic moments seems to be best. The normal approximations for LR_E and P_E tend to give critical levels which are too large at the .05 and .01 nominal levels. This result was also observed under hypotheses 3 and 4. It is interesting to note that the normal approximation is generally better for LR_A than P_A. The P_A critical levels tend to be too large, especially at the .01 nominal level. As in the symmetric case, the skewness and kurtosis tend to the normal values faster for LR_A than for P_A as k becomes large.

In general the normal approximation for LR_A is less affected by the presence of one or two large cell probabilities than the normal approximation for P_A. The Monte Carlo results for null hypotheses 2, 3, 4 suggest that the normal approximation for LR_A are not seriously misleading if $k \geq 10$, $n \geq 20$, and $n^2/k \geq 100$. These minimum values probably should be increased if a few cells contain more than ninety percent of the total probability.

Hypotheses 5 through 9 share the common property that no two cell probabilities are equal. For hypothesis 6 the behavior of the likelihood ratio and Pearson statistics is similar to their behavior under hypothesis 1 (symmetry) and hypothesis 4.

Null hypotheses 5, 7, 8, and 9 all have some very small expected frequencies. The presence of small expected frequencies has little effect on the normal approximation for LR_A, but the effect on the normal approximation for P_A varies with the number of small expected frequencies. In general, the normal approximation for LR_A is the most accurate at the .05 and .01 critical levels. As in previous cases, the chi-squared approximation for G_k^2 gives conservative critical levels when most expected frequencies are smaller than 0.5 and liberal levels when most expected frequencies are between 1 and 5. The chi-squared approximation for X_k^2 yields liberal critical levels when most expected frequencies are less than one.

Unlike the other cases, the presence of an extremely small expected frequency can cause the normal approximation for P_A and P_E to give very conservative critical regions. This is most dramatically illustrated by the estimated critical levels for hypothesis 9 presented in Figures L and and M. This hypothesis is close to hypothesis 1 in the sense that every cell but one has an expected frequency larger than $(0.9)\lambda$. The k-th cell has expected freqeuncy $(.1\lambda)/k$. Hence the conditional distribution of X_k^2 given that $N_{kk} = 0$ is well approximated by a chi-squared distribution with k-2 degrees of freedom. Furthermore, the probability that $N_{kk} = 0$ is $(1 - .1k^{-2})^k$, which converges to 1 as $k \to \infty$. Hence the distribution of X_k^2 severely deviates from the chi-squared distribution only in very extreme regions of the upper tail. However, the infrequent non-zero values of N_{kk}

$\sigma^2_{P,k}$ and $\mathrm{Var}(X^2_k)$ to be much larger than $2(k-2)$. Therefore, the normal approximate for P_A and P_E is conservative at commonly used critical levels, but the chi-squared approximations with $k-2$ degrees of freedom is quite adequate.

--- Insert Figures L and M about here ---

In general it was found that the normal approximation was more accurate for LR_A than for P_A. This is illustrated by the results in Figures H and I. In fact, the normal approximation for LR_A is even accurate when several very small expected frequencies are present.

Monte Carlo power computations show that for unsymmetrical null hypotheses, either test may be dominant. The area of dominance for the Pearson statistic is generally not nearly as broad as it is for the symmetrical null hypothesis case. In fact for some null hypotheses the likelihood ratio test completely dominates the Pearson test along specific directions.

As previously noted, Morris's central limit theorems are valid for a certain class of alternatives. Therefore, the normal approximations for X^2_k and G^2_k provide computationally inexpensive power approximations. However, Monte Carlo results indicate that it is not uncommon for these power approximations to be too large by as much as 20% for moderate power and moderate cell sizes. The discrepancy is generally smaller for G^2_k than for X^2_k.

6. SUMMARY AND RECOMMENDATIONS

Clearly for the null hypothesis of symmetry, the chi-squared approximation for the Pearson statistic is quite adequate at the .05 and .01 nominal levels for expected frequencies as low as .25 when $k \geq 3$,

$n \geq 10$, $n^2/k \geq 10$. The chi-squared approximation is generally easier to apply than the normal approximation since the former procedure does not require the calculation of a mean and standard deviation. Furthermore, the theoretical results of Holst, Morris and Stein and the numerical results summarized in Section 6 indicate that the Pearson test has some optimal local power properties in the symmetrical case when the number of cells is moderately large. Hence the Pearson goodness-of-fit test based on the traditional chi-squared approximation is preferred for the test of symmetry.

In general, the normal approximation for LR_A produces the most accurate critical regions for unsymmetrical hypotheses. The Monte Carlo results for null hypotheses 5, 6, 7, 8, and 9 suggest that the use of this approximation will not be seriously misleading for a wide range of null hypotheses in the interior of the simplex when $n \geq 15$, $n^2/k \geq 10$ and k is selected so that most expected frequencies are less than 5. Unlike the normal approximations for P_A and P_E, the accuracy of the normal approximations for LR_A is not seriously affected by the presence of a few extremely small expected frequencies. The chi-squared approximation for the Pearson statistic produces inflated rejection levels for unsymmetrical null hypotheses which contain many expected frequencies smaller than one.

The $C_{(m)}$ approximation for the Pearson statistic for the case of just a few small expected frequencies was proposed by Cochran (1946) and further studied by Yarnold (1970). The application of this approximation is limited to the cases for one and two small expected frequencies covered by the tables of percentage points given in Cochran's paper. Use of the normal approximation for LR_A eliminates the need for extensive tables of the $C_{(m)}$ approximation.

Modern computer programs which provide values of G_k^2 would have little trouble in providing values of LR_A. These values can be compared to readily available tables of the percentiles of the standard normal distribution.

FIGURE A

EXPECTED VALUE OF THE
POISSON INFORMATION KERNEL

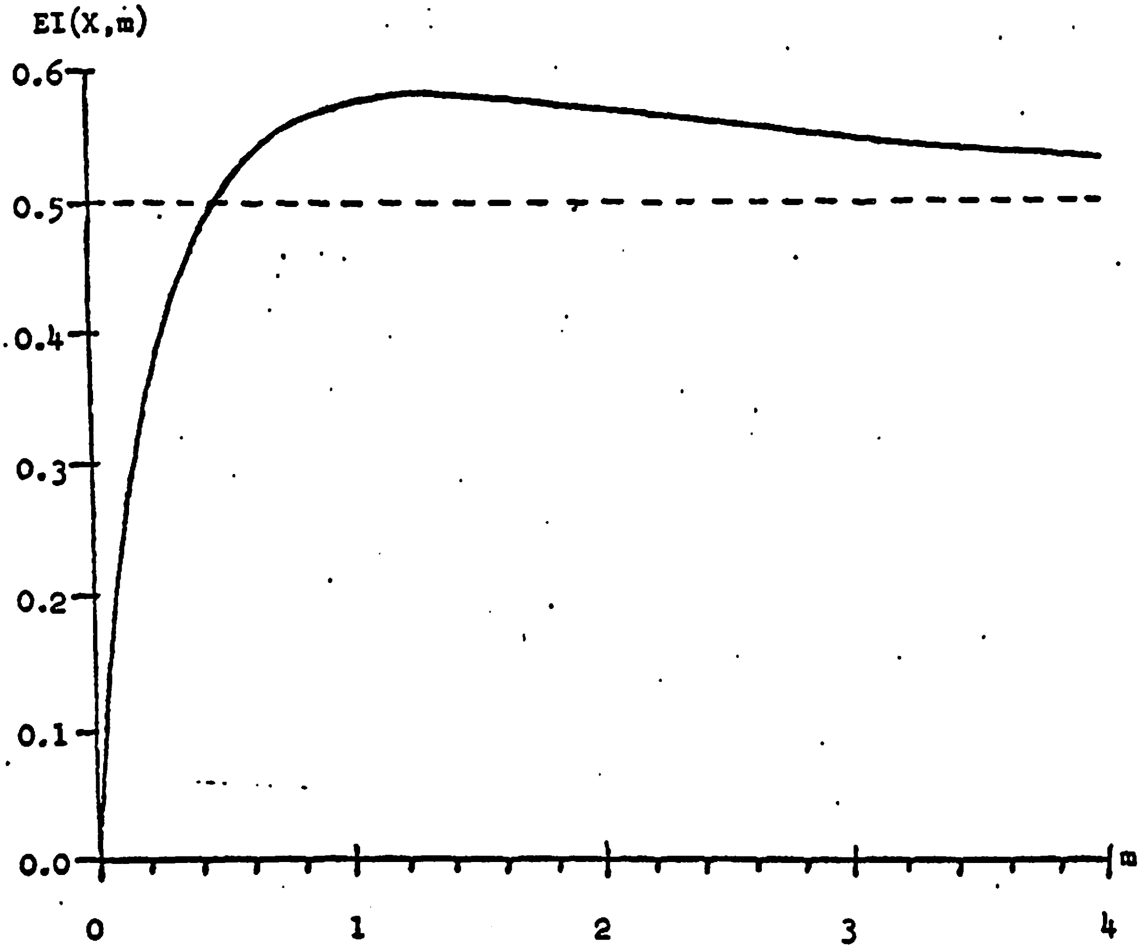


FIGURE B

VARIANCE OF THE

POISSON INFORMATION MODEL

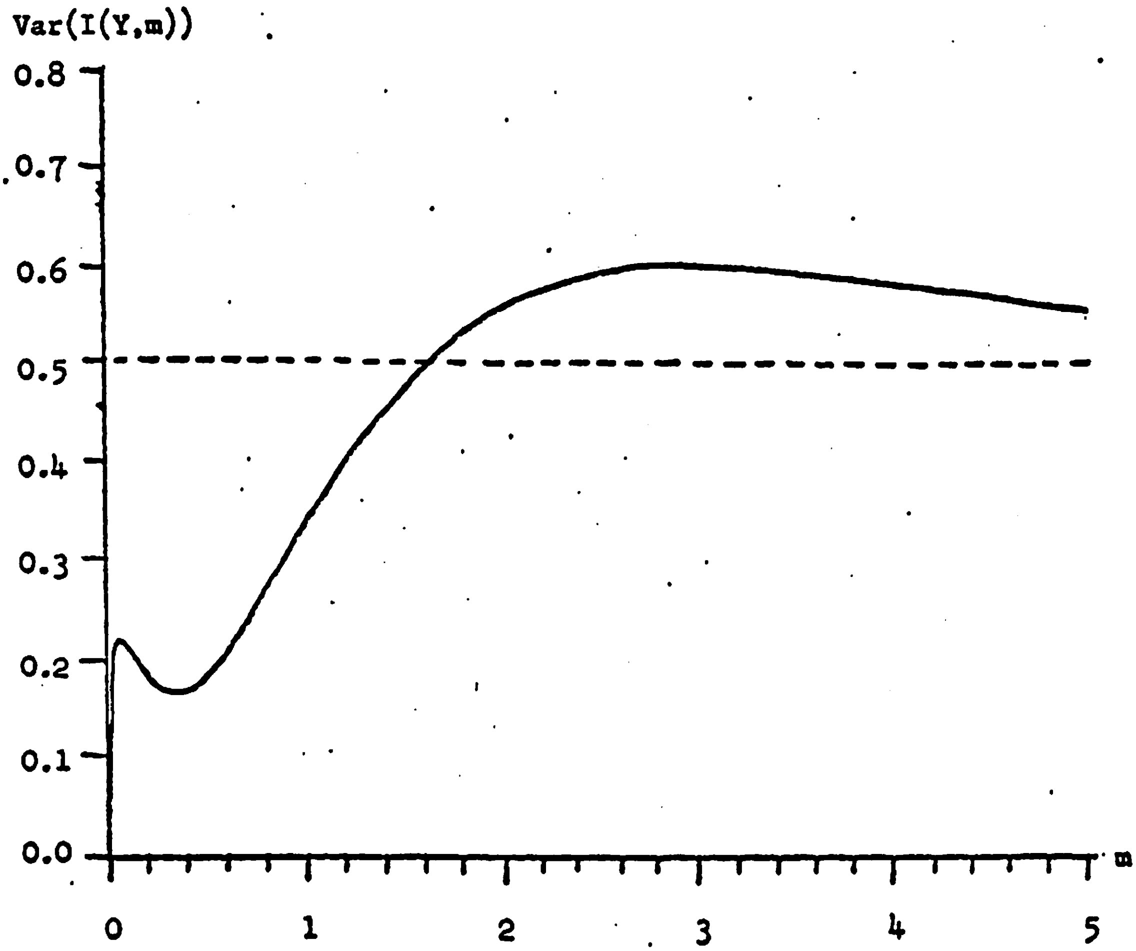


FIGURE C

COVARIANCE BETWEEN Y AND THE

POISSON INFORMATION KERNEL

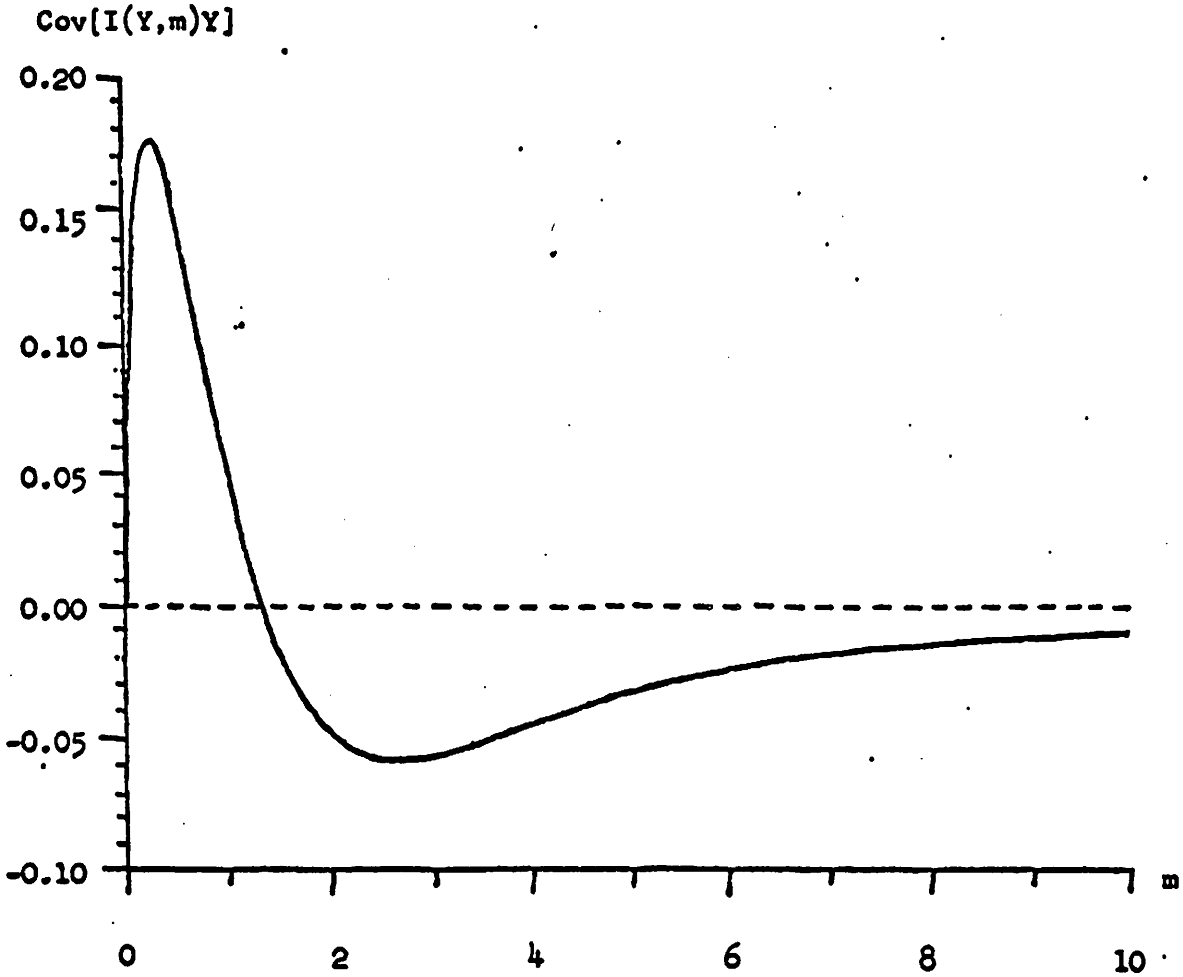


FIGURE D

Estimated Probability of Exceeding $Z_{.95} = 1.645$

Under Hypothesis 1 when n = .5k

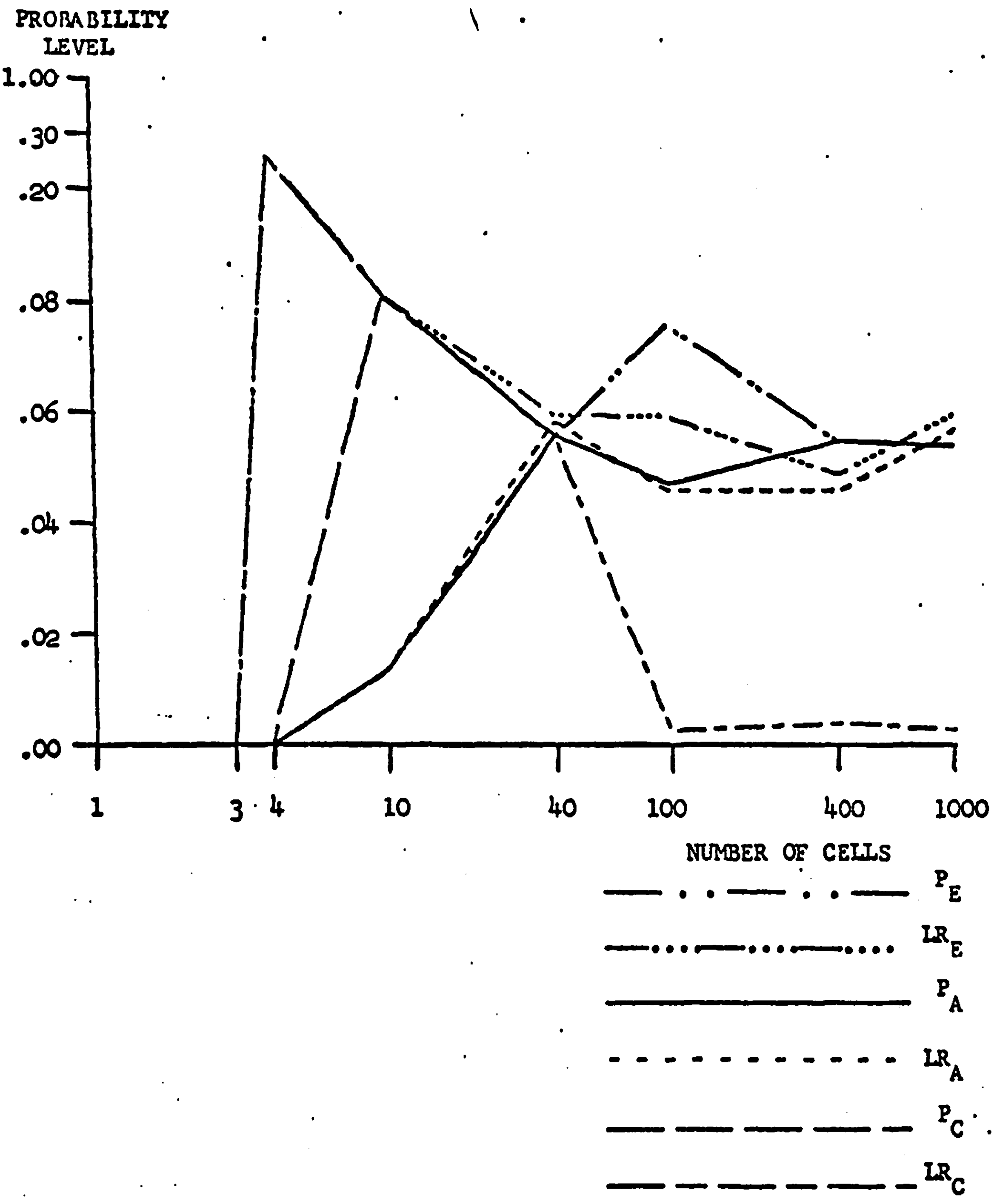


FIGURE E

Estimated Probability of Exceeding $Z_{.99} = 2.326$

Under Hypothesis 1 when $n = .5k$

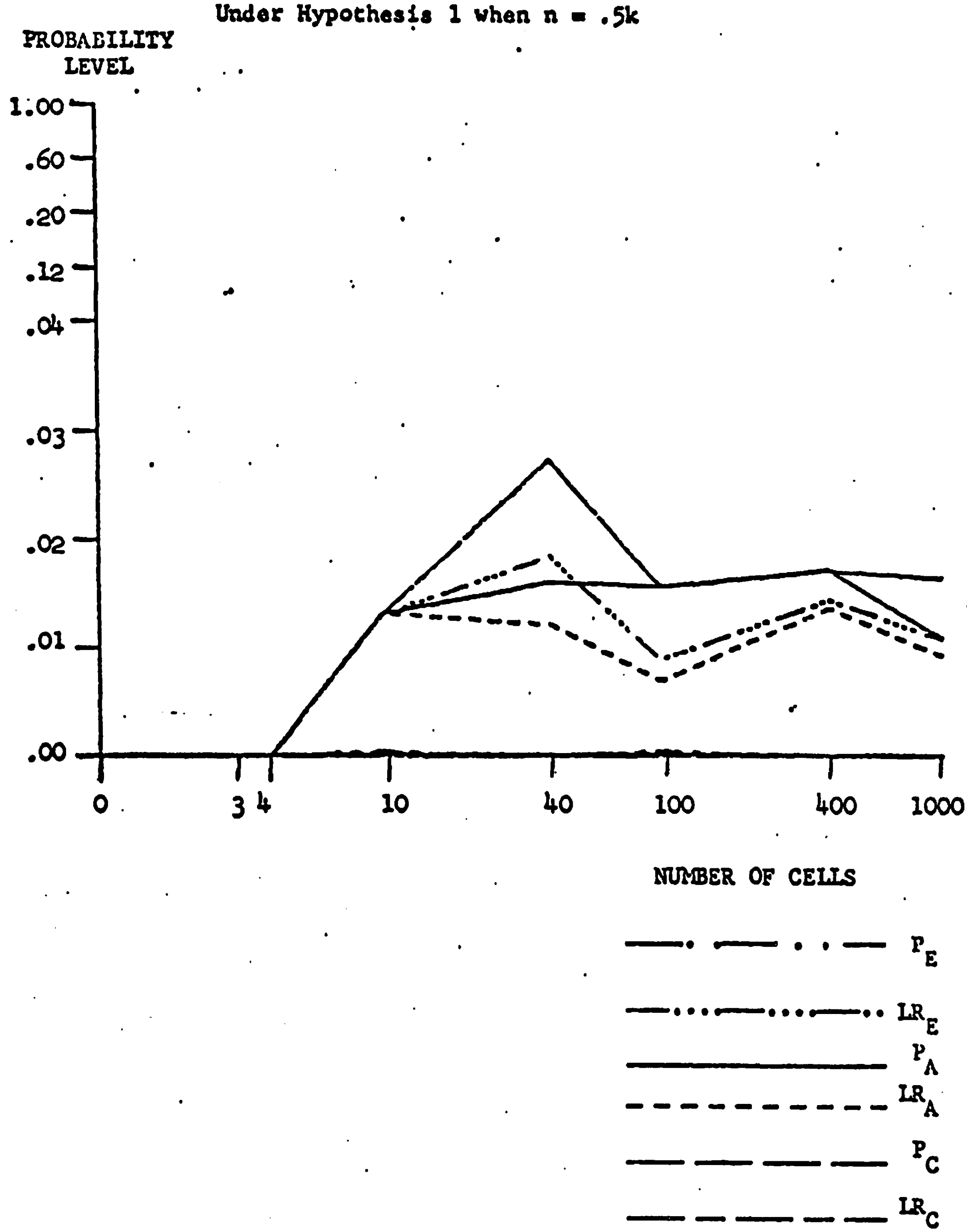


FIGURE F

Estimated Probability of Exceeding $Z_{.95} = 1.645$

Under Hypothesis 1 when $n = k$

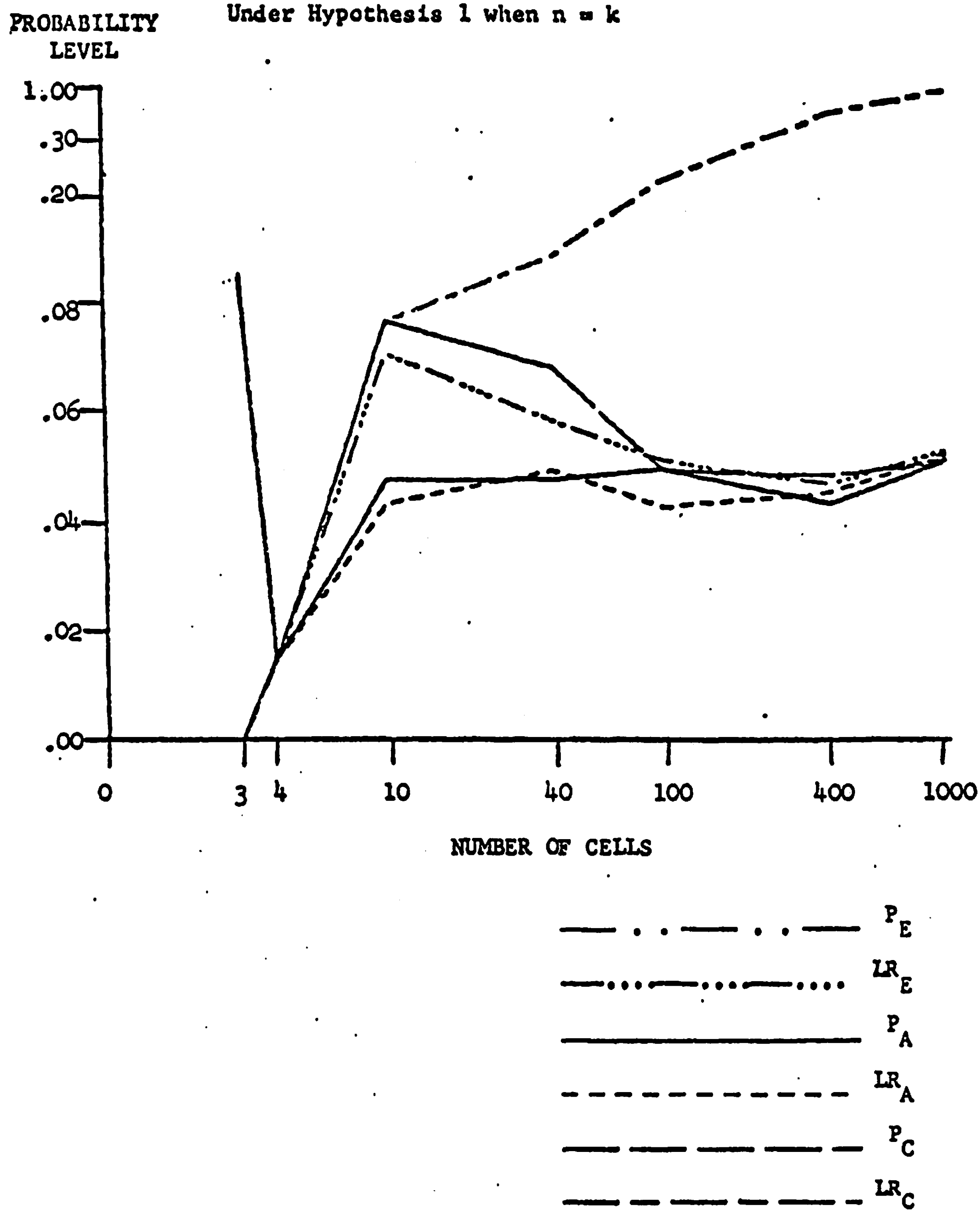
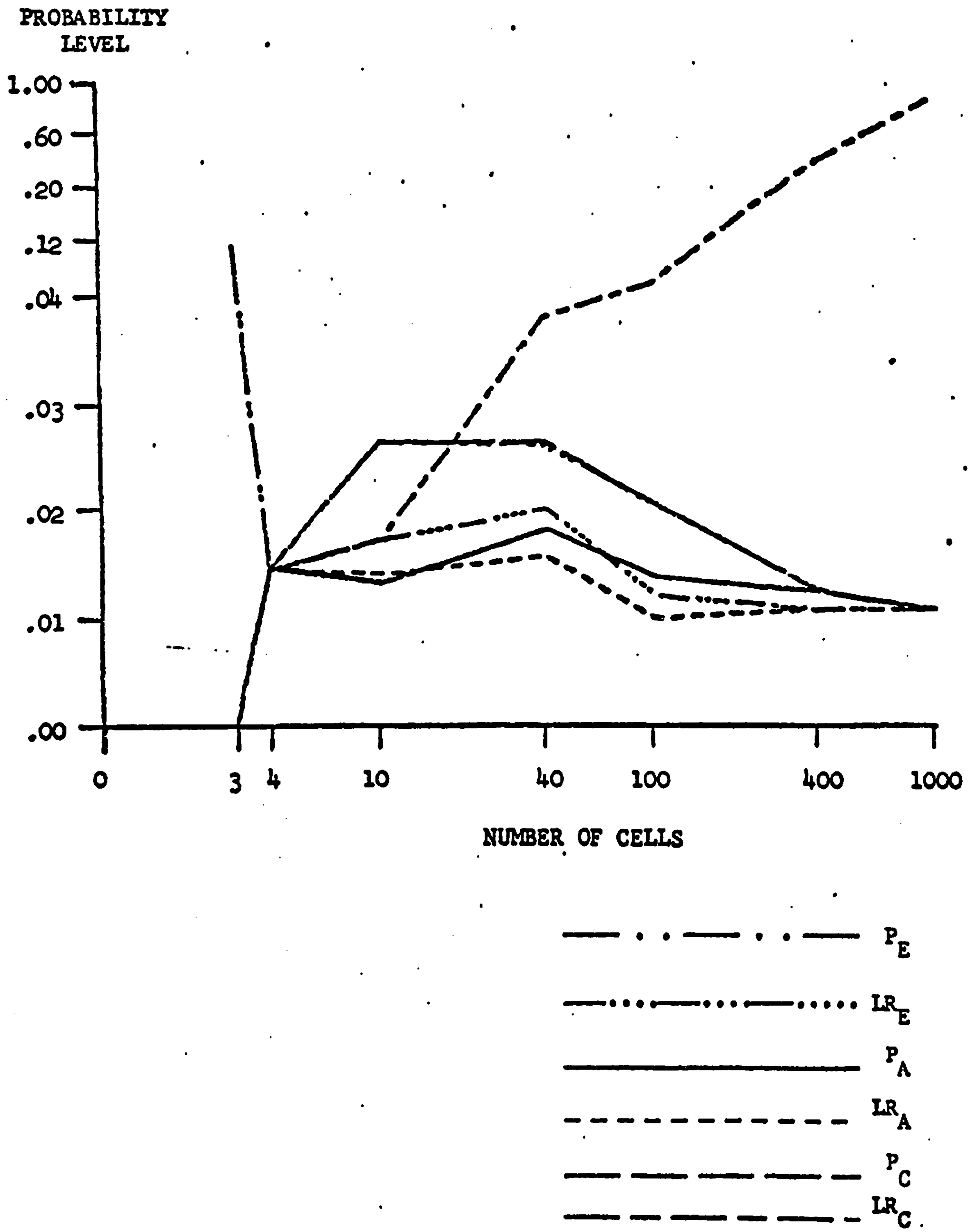


FIGURE G

Estimated Probability of Exceeding $Z_{.99} = 2.326$

Under Hypothesis 1, when n = k



H. Monte Carlo Rejection Levels for the Nominal .05 Level

Null Hypothesis	k	P_E 123456789	LR_E 123456789	P_A 123456789	LR_A 123456789	P_C 123456789	LR_C 123456789
λ=.25	3	x6⊗0⊗0 1	x6⊗0⊗0 1	060000 1	060000 1	0600⊗0 1	060000 1
	4	x700 0310	x700 0310	0700 0310	0700 0310	0700 0310	070000310
	10	026⊗ 10	028⊗ 6971	0261 0 20	0371 0371	0●⊗1⊗0⊗ 1	000000000
	40	8776 6330	7786 6	27 330	7	8●⊗ ⊗6●⊗6	000000000
	100	8 66 330	8866 3	330		8●⊗7⊗ ●⊗6	000000000
	400	6 36330	3 3	6 3 330		●⊗7⊗ ●●	000000000
λ=.50	3	0⊗2819782	0⊗⊗880782	0⊗2010782	00200222	0⊗2010782	002010222
	4	●⊗39 ⊗ 1	●⊗⊗9 ⊗8 1	0⊗3020 1	0030 0 1	0⊗30808 1	0030201 1
	10	86 77 30	8677 6762	163 6 3	16 3 2332	8●⊗ ⊗6⊗80	101000000
	40	766 7 0	66 6893	7		●⊗6⊗6●⊗	00000000
	100	7 6 6 30	777 9	6 3	66	●⊗7⊗6●●	000000000
	400	6 6 330		33		●⊗7⊗6●●	000000000
λ=1	3	⊗18⊗6⊗233	⊗⊗ ⊗8⊗⊗⊗3	012060233	012330231	⊗⊗8⊗6⊗⊗⊗3	⊗1 ⊗8⊗231
	4	1229 372	12 717⊗6	1211 1332	12 1 131	1●2 61⊗76	10 96 01
	10	7767 1	77666 86	32 31	1333	7●969 ⊗⊗	70273701
	40	6 7766 0	766 6	6 0	7	6●97⊗6●⊗	⊗01⊗1⊗00⊗
	100	66 0	66 77	6 0		⊗96⊗ ●●	●00⊗0●00●
	400	666 3 0	6 66 8	6 3 0	6	●97⊗ ●●	●00●0●00●
λ=2	3	2 776	68⊗7776	021133776	22123113	6 779	⊗26⊗⊗⊗116
	4	9 6 666	3 8 6 976	2 323166	3 3232113	3 66 9⊗8	918⊗6⊗117
	10	7 6766 2	6 6666666	3 3 2	323 3	7⊗778 ⊗⊗6	⊗07⊗6⊗027
	40	66 0	76 6 7 7	6 0		⊗8 ⊗ ⊗⊗	⊗0⊗⊗ ⊗00⊗
	100	66 6 0	6 76	6 6 0		⊗7 ⊗ ●⊗	●0●● ●00●
λ=3	3	766684 2	879899⊗⊗	272212 1	322 2 103	76 67⊗⊗6	829⊗9923
	4	796 67⊗87	896878876	3 3232 7	322333223	696 76⊗99	⊗2⊗⊗7⊗22
	10	8676 663	766667677	36 3	33 3	7⊗767 ⊗⊗6	⊗19⊗7⊗027
	40	66 6 61	6 6776	6 1		⊗7686⊗⊗	⊗0⊗⊗7⊗00⊗
	100	7 6 0	7 67	6 0		⊗769 ●⊗	⊗0●●6⊗00⊗
λ=5	3	92 767222	72677727	222232221	122333123	⊗2 766776	72987822
	4	6 6 66	6 666 ●8	2 3222222	233333133	6 668⊗⊗	9 ⊗787136
	10	777776666	777877666	7 6	3 3	7⊗7786⊗⊗9	92⊗99912
	40	6866 2	67 6	2		6⊗767 ⊗⊗	91⊗⊗79007
	100	66 1	6 6767	1		6⊗6 86⊗⊗	⊗1●⊗9⊗009

Code for Figure H.

Symbol	Range for Rejection Level
0	0 - 0.01
1	0.01 - 0.02
2	0.02 - 0.03
3	0.03 - 0.04
blank	0.04 - 0.06
6	0.06 - 0.07
7	0.07 - 0.08
8	0.08 - 0.09
9	0.09 - 0.10
◉	0.10 - 0.20
●	0.20 - 1.00
x	No result - statistic assumes only one value

I. Monte Carlo Rejection Levels for the Nominal .01 Level

Null Hypothesis	k	P_E 123456789	LR_E 123456789	P_A 123456789	LR_A 123456789	P_C 123456789	LR_C 123456789
λ=.25	3	x●000●0⊛⊛	x●00000⊛⊛	0●0 00⊛⊛	0●00000⊛⊛	0●000●0⊛⊛	0●00000⊛⊛
	4	x●000●0 6	x●000●0⊛6	0●0 ●0 6	0●00000⊛6	0●000●0⊛6	000000
	10	7● 9 861	76 6 96	79 9 861	76 6 66	●● ● ●⊛1	010000100
	40	877988661	868667776	877687661	6666 6 6	8●●9●8●●7	000000000
	100	6 6877670	9878 6 66	6 6777670	66 66	6●●9●7●●7	000000000
	400	676677 0	666 6 2	676676 0	6 6	6●●7●7●●	000000000
λ=.50	3	0 70 0777	0 70 0●●7	0 70 0777	0 70 0777	0●70 0●●7	0 70 0117
	4	0 ⊛070⊛⊛	0 ⊛070⊛⊛	0 ⊛020⊛⊛	0 ⊛07066	0●⊛⊛0⊛0⊛	0 002016
	10	6⊛⊛9792	⊛ 6 866	72792	6 66	●●8● ●●	002110010
	40	897788881	6776 6⊛9	696676881	766 76	8●●7●8●●6	000000000
	100	697887770	677 669	696786770	666 66	6●●8●7●●6	000100000
	400	66 7 660	6	66 7 660		6●●6● ●●6	000000000
λ=1	3	●679●●89⊛	0679●9896	0620⊛089⊛	062020226	0●79●0●●⊛	0679●●226
	4	766⊛●987	7⊛6⊛696●	6 ⊛6887	6 6 2	766⊛6●●⊛	667⊛2 8
	10	8⊛878⊛⊛9	668 77⊛76	⊛6 6699	6 7	8●968●●8	6 6 7026
	40	8⊛88978⊛0	6⊛866686	6⊛7686880	666 66	8●●8●7●●6	⊛01819007
	100	777676770	7 88	77 66770	6 6	7●⊛6●6●●	●00⊛1●00●
	400	66666660	6 6 8	6 666660	6	●⊛6●6●●	●00●●00●
λ=2	3	●87●⊛9●●●	●8⊛689●69	28 2⊛1668	2 221	●●76⊛9●●●	● ⊛⊛8● 9
	4	8⊛99⊛7●●⊛	9⊛9797977	6⊛7 6 6●⊛	6 6	8⊛97⊛6●●●	⊛1⊛⊛98 8
	10	7889⊛8987	686776997	66676887	6 6	7●⊛8●8●●⊛	●06⊛7⊛129
	40	78868697	76 867	687 7 97	6 6	7●⊛6⊛6●●7	●09● ●00⊛
	100	666686771	6 7 77	6 8 771	66	6●86●6●●	●0●● ●00●
λ=3	3	7●⊛8⊛⊛88	⊛⊛8⊛8⊛⊛⊛9	27 6		7⊛98⊛9●⊛9	● ●●⊛● 8
	4	⊛⊛9⊛99⊛89	8⊛7⊛⊛9⊛8⊛	76 6 876	6 66 6	6●98⊛8●●●	●69●⊛● 9
	10	7⊛9787⊛⊛8	8⊛6797⊛⊛8	6⊛6 7 ⊛⊛8	6 66	7⊛⊛76⊛●⊛⊛	⊛19⊛9⊛1 8
	40	68777789	687 67977	8767689	6 676	6⊛97⊛8●●7	●0●⊛7●00⊛
	100	686676891	86 6 68	76 7 881	6 6	6●86⊛6●●6	●0●⊛7●00●
λ=5	3	879⊛88787	87⊛98⊛78●	6 676	2 678	879⊛8878●	⊛6⊛⊛⊛⊛67⊛
	4	8⊛9789676	8⊛98887●⊛	⊛ 66	7 6 6	8●9788●●⊛	⊛ ⊛⊛⊛⊛ 68
	10	9899⊛898⊛	897888979	687676869	6 6666 6	8●⊛8⊛8⊛8●	⊛ ⊛⊛⊛⊛177
	40	6⊛677687	68666 66	7666 87	6 6 6	6⊛7796●⊛8	91⊛⊛89018
	100	67 6698	666 66767	7 6688	6	6⊛6 96●⊛6	⊛1⊛⊛8⊛009

Code for Figure I.

Symbol	Range for Rejection Level
0	0 - 0.0005
1	0.0005 - 0.0025
2	0.0025 - 0.0050
blank	0.0050 - 0.0150
6	0.0150 - 0.0200
7	0.0200 - 0.0250
8	0.0250 - 0.0300
9	0.0300 - 0.0350
⊗	0.0350 - 0.0500
●	0.0500 - 1.0000
x	No result - statistic assumes only one value

FIGURE J

Estimated Probability of Exceeding $Z_{.95} = 1.645$

Under Hypothesis 2, n = k

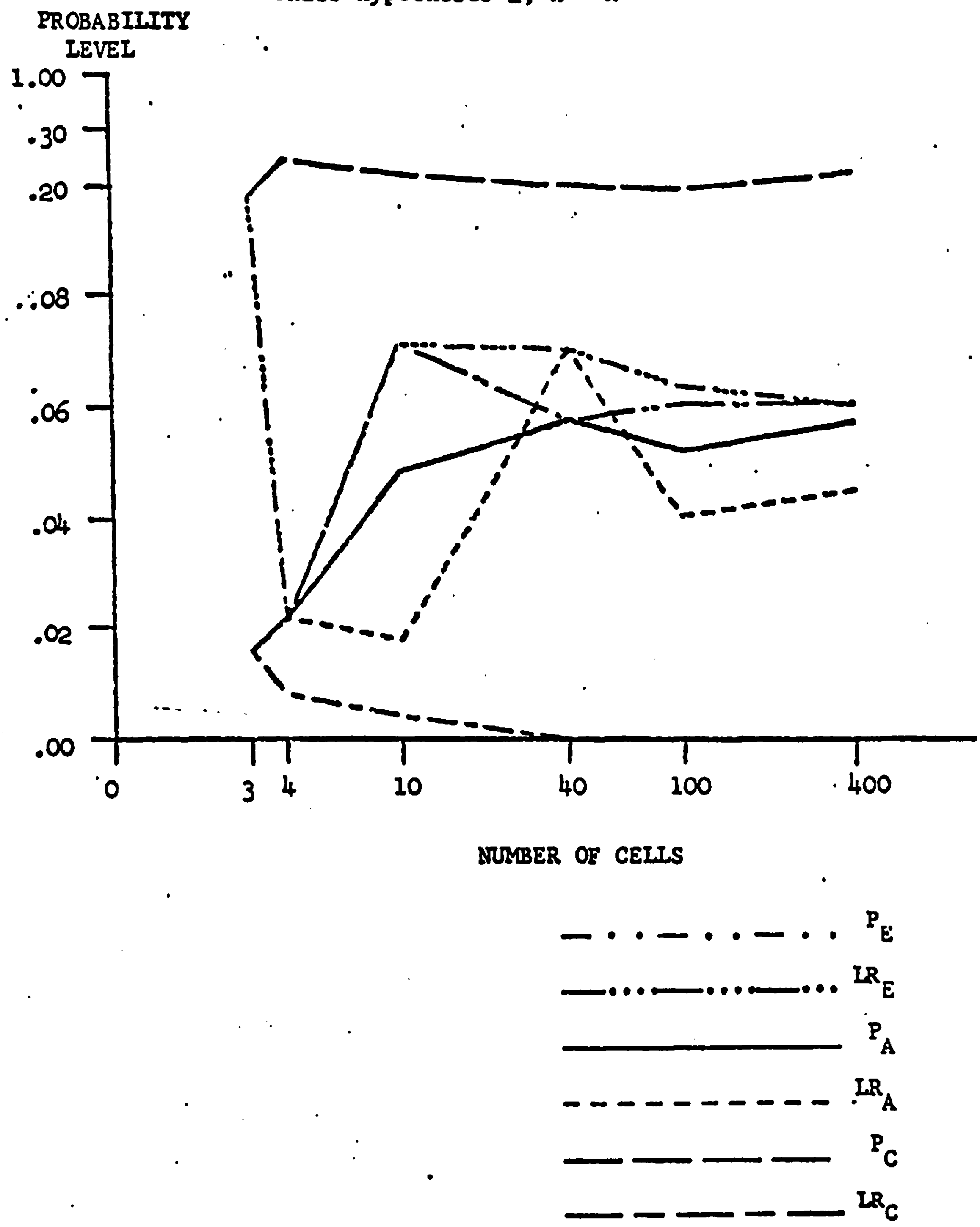


FIGURE K

Estimated Probability of Exceeding $Z_{.99} = 2.326$

Under Hypothesis 2 when $n = k$

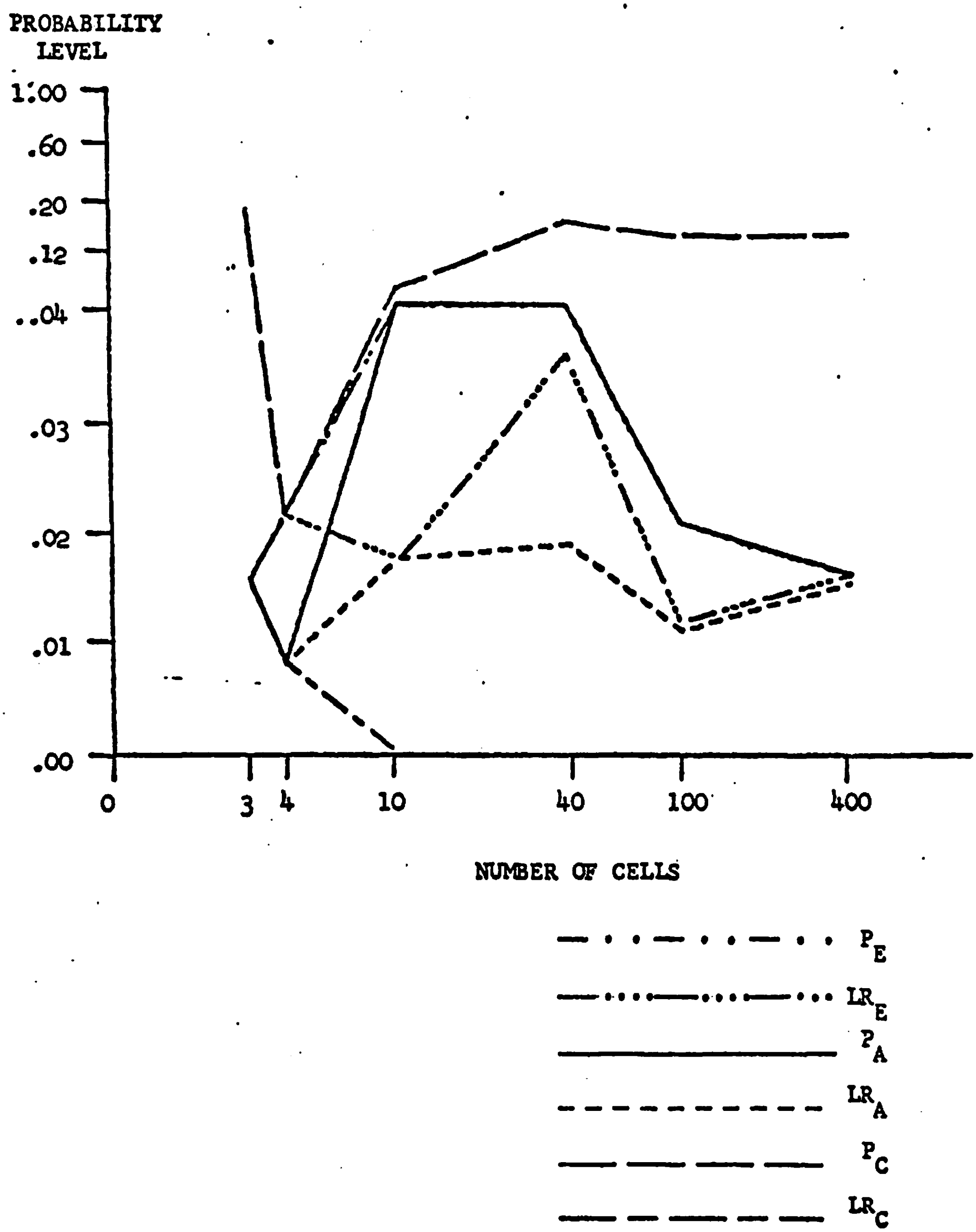


FIGURE L

Estimated Probability of Exceeding $Z_{.95} = 1.645$

Under Hypothesis 9 when $n = .5k$

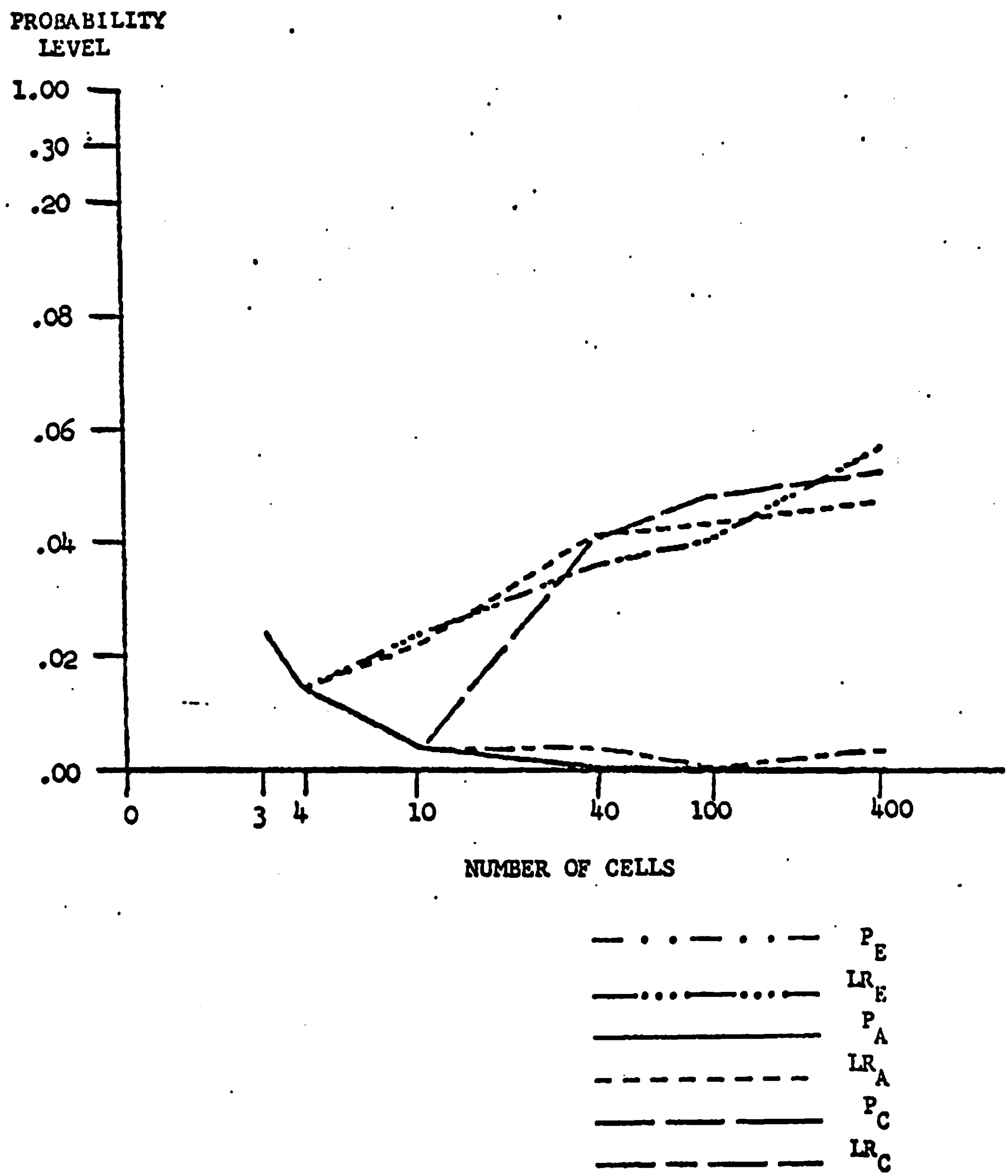


FIGURE M

Estimated Probability of Exceeding $Z_{.99} = 2.326$

Under Hypothesis 9 when $n = .5k$

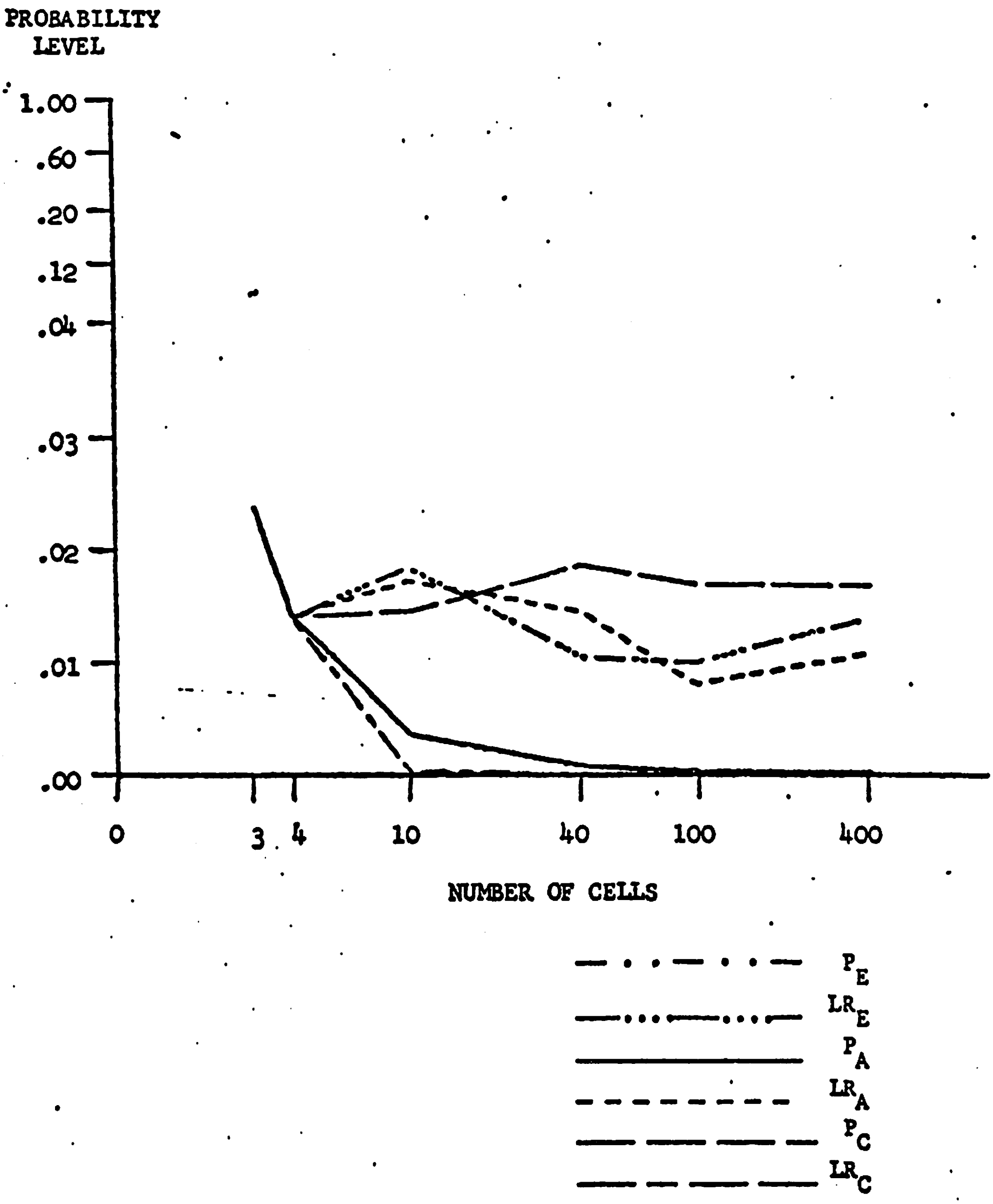


Table 1

MINIMUM CONTRIBUTIONS FOR OBSERVED
COUNTS OF ZERO AND ONE

Expected Frequency	Count of Zero X^2	Count of Zero G^2	Prob (Zero Count) Under Poisson	Count of One X^2	Count of One G^2	Prob (One Count) Under Poisson
5.00	5.00	10.00	.00674	3.200	4.781	.03369
3.00	3.00	6.00	.04979	1.333	1.803	.14936
2.00	2.00	4.00	.13533	0.500	0.614	.27067
1.50	1.50	3.00	.22313	0.167	0.189	.33470
1.00	1.00	2.00	.36788	0.000	0.000	.36788
0.75	0.75	1.50	.47237	0.083	0.074	.35427
0.50	0.50	1.00	.60653	0.500	0.386	.30326
0.25	0.25	0.50	.77880	2.250	1.273	.19470
0.10	0.10	0.20	.90483	8.100	2.806	.09048
0.05	0.05	0.10	.95123	18.050	4.091	.04756
0.01	0.01	0.02	.99004	98.010	7.230	.00990

NOTE: Minimum contribution for G^2 is $\lim_{n\to\infty} 2n\log(n/(n-np_j)) = 2np_j$ for a zero count in the j-th cell, and $\lim_{n\to\infty} 2\log(1/np_j) + 2(n-1)\log((n-1)/(n-np_j)) = -2\log(np_j) + 2(np_j-1)$ for a count of one. It is interesting to note the values for G^2 are limits of the OUTLIER values given by Gokhale and Kullback (1978, p. 64).

Table 2

NULL HYPOTHESES CONSIDERED IN THE STUDY

Label	Null Hypotheses
1	$(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$
2	$.1(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}) + .9(1, 0, \ldots, 0)$
3	$(\frac{3}{8} + \frac{1}{2k}, \frac{1}{8} + \frac{1}{2k}, \frac{1}{2k}, \ldots, \frac{1}{2k})$
4	$.9(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}) + .1(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$
5	$(c_1, c_2, \ldots, c_k)$, where $c_i = \frac{1}{k} \sum_{j=i}^{k} \frac{1}{j}$
6	$.1(c_1, c_2, \ldots, c_k) + .9(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$
7	$.1(c_1, c_2, \ldots, c_k) + .9(1, 0, \ldots, 0)$
8	$.1(c_1, c_2, \ldots, c_k) + .9(\frac{1}{2k}, \frac{1}{2k}, \ldots, \frac{1}{2k}, 0, \ldots, 0)$
9	$.1(c_1, c_2, \ldots, c_k) + .9(\frac{1}{k-1}, \ldots, \frac{1}{k-1}, 0)$

REFERENCES

Cochran, W.G. (1942), "The χ^2 correction for continuity," Iowa State College Journal of Science, 16, 421-436.

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

Gokhale, D.V. and Solomon Kullback (1978), The Information in Contingency Tables. New York, Marcel Dekker, Inc.

Good, I.J., I.N. Grover, and G.J. Mitchell (1970), "Exact distributions for X^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution," Journal of the American Statistical Association, 65, 267-283.

Haldane, J.B.S. (1937), "The exact value of moments of the distribution of χ^2, used as a test of goodness of fit when expectations are small," Biometrika, 29, 133-143.

Hoeffding, W. (1965), "Asymptotically optimal tests for multinomial distributions," Annals of Mathematical Statistics, 36, 369-401.

Holst, L. (1972), "Asymptotic normality and efficiency for certain goodness-of-fit tests," Biometrika, 59, 137-145.

Holst, L. (1976), "On multinomial sums," Mathematics Research Center Technical Summary Report No. 1629, University of Wisconsin-Madison.

Katti, S.J. (1973), "Exact distribution for the chi-square test in the one way table," Communications in Statistics, 2, 435-447.

Koehler, K.J. (1977), Goodness-of-fit statistics for large sparse multinomials, Unpublished Ph.D. dissertation, School of Statistics, University of Minnesota.

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

Morris, C. (1966), "Admissable Bayes procedures and classes of epsilon Bayes procedures for testing hypotheses in a multinomial distribution," Technical Report No. 55, Department of Statistics, Stanford University.

Morris, C. (1975), "Central limit theorems for multinomial sums," Annals of Statistics, 3, 165-188.

Nass, C.A.G. (1959), "The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism," Biometrika, 46, 365-385.

Neyman, J. and E.S. Pearson (1928), "On the use and interpretation of certain test criteria for purposes of statistical inference," Biometrika, 20-A, 175-247, 264-299.

Pearson, K. (1900), "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Phil. Mag., 50, 157-175.

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

Slakter, M.J. (1966), "Comparative validity of the chi-square and two modified chi-square goodness-of-fit tests for small but equal expected frequencies," Biometrika, 53, 619-622.

Steck, G.P. (1957), "Limit theorems for conditional distributions," University of California Publications in Statistics, Vol. 2, No. 12, 237-284.

Vessereau, A. (1958), "Sur les conditions d'application du criterium χ^2 de poisson," Bulletin International Institute of Statistics, 36, 87-101.

West, E.N. and O. Kempthorne (1971), "A comparison of the chi^2 and likelihood ratio tests for composite alternatives," Journal of Statistical Computation and Simulation, 1, 1-33.

Zahn, P.A. and G.C. Roberts (1971), "Exact χ^2 criterion tables with cell expectation one: An application to Coleman's measure of consenses," Journal of the American Statistical Association, 66, 145-148.