No. 44 (79-8)

The Number of Quadrats and the Goodness-of-Fit Test of the Quadrat Method for Testing Randomness in the Distribution of Points on a Plane

July 1979
(Revised in August, 1980)

Atsuyuki Okabe

ABSTRACT

This paper reexamines the conventional goodness-of-fit test used in the quadrat method and proposes an alternative goodness-of-fit test. It is first shown that the conventional test should be used when the number k of points is more than 4000 and the number n of quadrats should be around $n = .06k$. Second the conventional test is likely to misjudge a random point distribution as non-random when $k < 4000$ regardless of n. To avoid this misjudgment, last an alternative goodness-of-fit test is proposed.

Associate Professor,

Institute of Socio-Economic Planning

University of Tsukuba

Sakura, Ibaraki 305, Japan

(1) INTRODUCTION

The statistical analysis of a spatial distribution of point-like activity, such as a distribution of stores in an urban area or that of cities in a region, has been one of the major subjects of statistical geography. To deal with this subject, a variety of methods has been proposed, (for example, see the methods cited in Bartlett [1], Getis and Boots [5], King [6], Lewis [7], Rogers [9]). Among those methods, the goodness-of-fit test by quadrats or so-called the quadrat method is one of the most frequently used methods in the related literature. Conventionally the quadrat method takes the following procedure: first, to cover a study area, where a certain number, say k, of points are distributed, with an arbitrary number, say n, of equal quadrats; second, to count the number n_s of quadrats having s points, $s = 0, 1, \ldots, k$; third, to calculate the value of

$$X^2 = \sum_s \frac{(n_s - m_s)^2}{m_s} , \qquad (1\text{-}1)$$

where m_s is given by

$$m_s = n \frac{\lambda^s e^{-\lambda}}{s!} , \qquad s = 0, 1, \ldots, \qquad (1\text{-}2)$$

and $\lambda = k/n$; last, to test randomness by consulting the chi-square table.

Concerning the theoretical basis of this test, it appears to be accepted in the literature that the ordinary goodness-of-fit test developed by Pearson [8] is applicable to the quadrat method and hence the test statistic given by equation (1-1) follows the chi-square distribution if the number of quadrats is moderately large. This paper reexamins this theoretical basis. That is, the objectives of this paper are: first to show that the conventional method is likely to lead a biased conclusion; second to obtain the appropriate number of quadrats that reduces a bias to an allowable level; last to propose an alternative goodness-of-fit test.

To this end, Section 2 brings attention to von Mises [11] model which shows the exact distribution of n_s. (Note that this model is different from the multinominal distribution model considered in the ordinary goodness-of-fit test). With this exact distribution, the accuracy of approximation is analytically examined in Section 3 and numerically investigated in Section 4. Based upon these examinations, an appropriate number of quadrats which brings an allowable level of approximation is obtained. Since such a number does not always exist for a given number of points, an alternative goodness-of-fit test is proposed in Section 5. The paper ends in Section 6 summarizing the major results.

(2) THE EXACT AND ASYMPTOTIC DISTRIBUTION OF THE NUMBER OF QUADRATS HAVING S POINTS

It is well known that the model underlying the goodness-of-fit test is the multinominal distribution model (See Pearson [8], or Chapter 30 of Cramér [4], or Wise [12]). Stated explicitly, the model considers the situation in which n observations are randomly distributed into $k+1$ classes with probability p_s, $s = 0, 1, \ldots, k$. One may attempt to relate these n, k, s and p_s with those of the quadrat method. (p_s may correspond to m_s/n). This attempt, however, will not be successful to understand the quadrat method. Rather the model examined by von Mises [11], which is almost ignored in the literature of the quadrat method, is more relevant. Von Mises considers the model in which k balls are randomly placed in n boxies and obtains the distribution of boxies having s $(=0, 1, \ldots, k)$ balls. Since von Mises' result is indispensable for proceeding the following analysis, his result are briefly summarized here.

Let x_s be a random number of quadrats having s points when k points are randomly distributed over n quadrats. The probability $p\{x_s = m\}$ of the random variable x_s being equal to m is then given by

$$p\{x_s = m\} = \binom{n}{m} \sum_{i=0}^{n-m} \delta((m+i)s)(-1)^i \binom{n-m}{i} \times \frac{k!(n-(m+i))^{k-(m+i)s}}{(s!)^{m+i}(k-(m+i)s)!\, n^k} \tag{2-1}$$

$m = 0, 1, \ldots, k,$

where

$$\delta((m+i)s) = \begin{cases} 1 & \text{if } 0 \leq (m+i)s \leq k, \\ 0 & \text{if otherwise.} \end{cases} \tag{2-2}$$

The expected value $E(x_s)$ and variance $Var(x_s)$ of x_s are respectively given by

$$E(x_s) = m'_s = n \binom{k}{s} \left(\frac{1}{n}\right)^s \left(1 - \frac{1}{n}\right)^{k-s}, \tag{2-3}$$

$$Var(x_s) = \delta(2s) \frac{n(n-1)k!(n-2)^{k-2s}}{(s!)^2(k-2s)!\, n^k} - m'^2_s + m'_s . \tag{2-4}$$

The asymptotic distribution of equation (2-1) with respect to $n \to \infty$ provided that $k = \lambda n$ is written as

$$P\{x_s = m\} = \frac{m_s^m e^{-m_s}}{m!}, \qquad m = 0, 1, \ldots \tag{2-5}$$

where m_s is given by equation (1-2).[1] (Note that this Poisson distribution should not be confused with that of equation (1-2)).

(3) THE LIMITING ACCURACY OF APPROXIMATION WITH RESPECT TO THE NUMBER OF QUADRATS UNDER A FIXED NUMBER OF POINTS

When x is a random variable, the transformed variable $(x - E(x))/\sqrt{Var(x)}$ generally gives a crude approximation to the standard normal variable. In fact if a random variable follows the Poisson distribution, the variable $(x - E(x))/\sqrt{E(x)}$

approximately distributes standard-normally for a large $E(x)$. In the ordinary goodness-of-fit model, as is shown by Cochran (p.318 of [2]), a random variable x_s, (i.e., a random number of observations belonging to the s^{th} class), follows the Poisson distribution and hence the random variable $\Sigma(x_s - E(x_s))^2/E(x_s)$ almost shows the chi-square distribution. From this fact and equation (2-5) one may consider that the use of the test statistic given by equation (1-1) is also appropriate in the case of the quadrat method. It should be noted, however, that in the ordinary goodness-of-fit model, the random variable x_s exactly follows the Poisson distribution ([2]), while in the model of the quadrat method, x_s asymptotically follows the Poisson distribution as $n \to \infty$ keeping λ constant, which implies $k \to \infty$. In an actual situation, however, the number k of points is given, (i.e., fixed), although the number n of quadrats can freely be chosen. Therefore the asymptotic property obtained in equation (2-1) may practically be meaningless. Rather the asymptotic property with respect to $n \to \infty$ provided that k is fixed is of importance in an actual situation. In this section this asymptotic property will be examined and it will be shown that the ordinary goodness-of-fit test is not always applicable to the quadrat method.

Before setting into the analysis, it may be worthwhile to compare the exact distribution, i.e., equation (2-1), with the approximated distribution, i.e., equation (2-7), in a numerical example. The cases of $k = 25$; $n = 25, 50$; $s = 0, 3$ are depicted in figure 1a ∿ 1d. It may be read from this figure that the approximation is bad in the case of $s = 0$ but it is fairly good in the case of $s = 3$, and that in the case of $s = 0$ the approximation becomes worse as the number of quadrats increases, while it becomes better in the case of $s = 3$.

Figure 1 The exact and approximated distribution of the number x_s of quadrats having s points when the number k of points is 25: a) the number n of quadrats $= 25$, $s = 0$; b) $n = 25$, $s = 3$; c) $n = 50$, $s = 0$; d) $n = 50$, $s = 3$.

Now let us examine the above observation analytically. In the quadrat method, it is implicitly assumed that m'_s (given by equation (2-3)) can be approximated by m_s (given by equation (1-2)) and that $\mathrm{Var}(x_s)$ (given by equation (2-4)) can be approximated by m_s. First concerning the expected value, let us compare m_s with m'_s by

$$R_e\,(n \mid s,\ k) = \frac{m'_s}{m_s} = \left(1 - \frac{1}{n}\right)^{k-s} e^{\frac{k}{n}} \prod_{i=1}^{s-1} \left(1 - \frac{i}{k}\right). \qquad (3\text{-}1)$$

(Note that $\prod_{i=1}^{s-1} (1 - i/k) = 1$ for $s = 0, 1$). On the limit, the relation

$$\lim_{n\to\infty} R_e(h \mid s,\ k) = \prod_{i=1}^{s-1} \left(1 - \frac{i}{k}\right) \qquad (3\text{-}2)$$

holds. This implies that if a large number of quadrat is used, the approximated expected value m_s is equal to the exact expected value m'_s for $s = 0$ and 1, but m_s is not equal to m'_s for $s \geq 2$. Thus the error exists in the case of $s \geq 2$ even if the number of quadrats is large, but it will be shown later that this error is negligible in an actual situation.

Concerning the variance, the difference between $Var(x_s)$ and m_s becomes crucial. To see it, let

$$
\begin{aligned}
&R_v(n \mid s, k) \\
&= \frac{Var(x_s)}{m_s} \\
&= e^{\frac{k}{n}} \{ \frac{k!}{k^s s!\,(k-2s)!} \frac{1}{n^{s-1}} (1 - \frac{1}{n})(1 - \frac{2}{n})^{k-2s} \\
&\quad - \frac{k!^2}{k^s s!\,(k-s)!^2} \frac{1}{n^{s-1}} (1 - \frac{1}{n})^{2k-2s} \\
&\quad + \frac{k!}{k^s (k-s)!} (1 - \frac{1}{n})^{k-s} \}
\end{aligned}
\qquad (3\text{-}3)
$$

Since the analysis deffers according to $s = 0$, $s = 1$ and $s \geq 2$, these cases are separately examined. First in the case of $s = 0$, equation (3-3) becomes

$$
\begin{aligned}
R_v(n \mid 0, k) = e^{\frac{k}{n}} [n\{(1 - \frac{1}{n})(1 - \frac{2}{n})^k - (1 - \frac{1}{n})^{2k}\} \\
+ (1 - \frac{1}{n})^{k-2}].
\end{aligned}
\qquad (3\text{-}4)
$$

Furthermore, noticing that

$$(1 - \frac{1}{n})^{2k} = (1 - \frac{2}{n} + \frac{1}{n^2})^k = \sum_{i=0}^{k} \frac{1}{n^{2i}} (1 - \frac{2}{n})^{k-i},$$

equation (3-4) is written as

$$R_v(n \mid 0, k) = e^{\frac{k}{n}} \{ - (1 - \frac{2}{n})^k - \sum_{i=1}^{k} \frac{1}{n^{2i-1}} (1 - \frac{2}{n})^{k-i} + (1 - \frac{1}{n})^{k-2} \}. \qquad (3\text{-}5)$$

Taking the limit to $n = \infty$, equation (3-5) becomes

$$\lim_{n \to \infty} R_v(n \mid 0, k) = 0 . \qquad (3\text{-}6)$$

This equation shows that the difference between the approximated variance and the exact variance becomes great when a large number of quadrats is chosen. In the case of $s = 1$, equation (3-3) becomes

$$R_v(n \mid 1, k) = e^{\frac{k}{n}} \{(k - 1)(1 - \frac{1}{n})(1 - \frac{2}{n})^{k-2} - k(1 - \frac{1}{n})^{2k-2} + (1 - \frac{1}{n})^{k-1} \}. \qquad (3\text{-}7)$$

It is easily seen from this equation that

$$\lim_{n \to \infty} R_v(n \mid 1, k) = 0. \qquad (3\text{-}8)$$

Like the case of $s = 0$, this result shows that the approximated variance is extremely larger than the exact variance when the number of quadrats is large.

Last in the case of $s \geq 2$, it can be shown that

$$\lim_{n\to\infty} R_v(n \mid s \geq 2,\ k) = \prod_{i=1}^{s-1} (1 - \frac{i}{k}) . \qquad (3\text{-}9)$$

Thus on the limit of $n \to \infty$, the exact variance will be close to the approximated variance. (Note that $\lim_{n\to\infty} R_v = 1$ for $s = 0, 1$ and that $\lim_{n\to\infty,\ k\to\infty} R_v = 1$).

From the above examinations, (i.e., equations (3-6), (3-8) and (3-9)), one would notice a difficulty in the choice of the number of quadrats: if a large number of quadrats is used, $(n_i - m_i)^2/m_i$, $i = 0, 1$, in equation (1-1) (i.e., the conventional test statistic of the quadrat method) will provide an erroneous value of X^2; if the number of quadrat is not large, $(n_i - m_i)^2/m_i$, $i \geq 2$, in equation (1-1) will produce a bias. Therefore if the number of quadrats is arbitrary chosen, the conventional method is likely to lead a false conclusion. In the above analysis, however, the accuracy of approximation is not quantitatively examined. There might be a moderately large (or small) number of quadrats that will produce an allowable level of approximation. To cosider this problem a numerical analysis will be pursued in the next section.

(4) NUMERICAL EXAMINATION OF THE ACCURACY OF APPROXIMATION

In a practical situation, it may be assumed that the number of points and quadrats is between ten and ten thousands and that the number of classes used in the goodness-of-fit test is around five, i.e.,

$$n,\ k \in J = \{j \mid 10 \leq j \leq 10000\}; \quad s \in L = \{j \mid 0 \leq j \leq 4\}. \tag{4-1}$$

(Note that j is an integer). Under these assumptions, let us numerically examine the accuracy of approximation.

First concerning the expected value, suppose that 1% error is allowable and let S be the set of the number of quadrats and that of points (n, k) that produce less that 1% error, i.e.,

$$S = \{(n, k) \mid .99 \leq R_e(n, s, k) \leq 1.01 \quad \text{for all} \quad s \in L\}. \tag{4-2}$$

By numerical examination, S is obtained in figure 2. It can be concluded from this figure that if the number of points is larger than 114, the error in the expected value can be reduced to less than 1% by choosing an appropriate number of quadrats from S in figure 2.

Figure 2 The number of quadrats and that of points which produce less than 1% errors in the expected value

Concerning the variance, let us see, for example, the case of $k = 1000$ in figure 3 where the value of $R_v(n \mid s, 1000)$ is graphed with respect to $n \in J$ and $s \in L$. First figure 3 shows that R_v is less than one. This relation also holds in the others cases $(k \neq 1000)$, i.e.,

$$0 \leq R_v(n \mid s, k) < 1 \quad \text{for} \quad n, k \in J, \quad s \in L. \tag{4-3}$$

(The implication of this relation will be discussed in Section 5). Second figure 3 shows that the value $\overline{R}_v$ $(n \mid 1000)$ defined by

$$\overline{R}_v(n \mid k) = \min_s \{R_v(n \mid s, k)\}, \tag{4-4}$$

(i.e., the lower envelope of $R_v(n \mid k, s)$ with respect to $s \in L$), increases in the domain of $10 \leq n \leq 80$ and decreases in $n > 80$. This result numerically supports the analytical conclusion obtained in Section 3, that is, the conventional test statistic (given by equation (1-1)) will bring a false conclusion if a very large number of quadrats is chosen. In figure 3, the

Figure 3 The ratio of the exact variance to the approximated variance (the number of points is 1000)

best accuracy is read as $\overline{R}_v = .94$, which may not be within an allowable level. Like the case of the expected value, suppose that 1% error allowable. Then the numerical examination shows that there exists no n satisfying this level for all $k \in J$. Alternatively suppose that less than 2% (5%) error is allowable. Then there exists n satisfying this level. To be explicit, let $T_{.02}$ $(T_{.05})$ be the set of the number of quadrats and points that brings less than 2% (5%) errors, i.e.,

$$T_{.02} = \{(n, k) \mid .98 \leq \overline{R}_v(n \mid k) \leq 1.02\}, \tag{4-5}$$

$$T_{.05} = \{(n, k) \mid .95 \leq \overline{R}_v(n \mid k) \leq 1.05\}. \qquad (4\text{-}6)$$

From the numerical calculation $T_{.02}$ and $T_{.05}$ are obtained in figure 4. (Note that $T_{.02}, T_{.05} \subseteq S$). This figure says that if more than 2% (5%) error is not allowable and if the number of points is less than 4000 (1200), the conventional test statistic given by equation (1-1) should not be used. If the number of points is more than 4000 (1200), the number of quadrats should be chosen from $T_{.02}$ ($T_{.05}$) in figure 4. Roughly speaking the number of quadrats is 6% of the number of points. Obviously if the number of quadrats is not chosen from $T_{.02}$ ($T_{.05}$), the conventional test statistic is likely to bring a false conclusion. If should be noted that many empirical examples cited in the related literature do not satisfy this condition.

Figure 4 The number of quadrats and that of points which produce less than 2% or 5% errors in the variance

(5) AN ALTERNATIVE GOODNESS-OF-FIT

In an actual situation, the number k of points may be less than 4000. Actually the case of $k \geq 4000$ can hardly be found in the related literature. If $k < 4000$, as is shown in Section 4, the use of the conventional test statistic (equation (1-1)) is misleading. Then what test statistic should be used when $k < 4000$? To answer this question, let us examine a random variable defined by

$$\chi^2 = \sum_s \frac{(x_s - m_s')^2}{\mathrm{Var}(x_s)} \quad , \tag{5-1}$$

where m_s' and $\mathrm{Var}(x_s)$ are respectively given by equations (2-3) and (2-4). If χ^2 follows the chi-square distribution, randomness can be tested by

$$\chi^2 = \sum_s \frac{(n_s - m_s')^2}{\mathrm{Var}(x_s)} \quad . \tag{5-2}$$

To justify this alternative test statistic, it has to be shown that $(x_s - m_s')/\sqrt{\mathrm{Var}(x_s)}$ follows the standard normal distribution and that $(x_s - m_s')/\sqrt{\mathrm{Var}(x_s)}$, $s = 0, 1, \ldots$, are statistically independent each other. First, to examine the normality, a numerical method is used because the higher moments of x_s is analytically too complex to obtain. In the case of $s \geq 2$, the goodness-of-fit test shows that the exact distribution can practically be approximated by the Poisson distribution if the number of quadrats is moderately large. (Recall that this fact is analytically proved by equation (3-9). As an example, see figure 1b, d). When x_s follows the Poisson distribution, it is reported by Chochran [2], [3] that even if the expected value is moderately small, the random variable $(x_s - m_s)^2/m_s$ can practically be approximated by the chi-square distribution. (Also see Yarnold [13] and Roscoe and Byars [10]). In the case of $s = 0, 1$, the goodness-of-fit test shows that the fitness to the normal distribution is very good. An example is shown in figure 1a, c in which the normal distribution is depicted by a continuous line. If may hence be concluded that the normality condition is practically satisfied for all $s \in L$.

Second, to examine the statistical independence, let us see the correlation coefficient R_{st} of x_s and x_t. Von Mises [11][2] shows that the covariance $Cov(x_s, x_t)$ of x_s and x_t is given by

$$Cov(x_t, x_t) = \frac{n!}{(n-2)!} \frac{k!\,(n-2)^{k-s-t}}{s!\,t!\,(k-s-t)!\,n^k} - m'_s m'_t , \quad (5\text{-}3)$$

where m'_s and m'_t are given by equation (2-3). Hence the correlation coefficient is obtained from

$$R_{st} = \frac{Cov(x_s, x_t)}{\sqrt{Var(x_s)\,Var(x_t)}} , \quad (5\text{-}4)$$

where $Var(x_s)$ and $Var(x_t)$ are given by equation (2-4). From this equation it can be proved that

$$R_{st} \neq 0 \quad (5\text{-}5)$$

However the numerical examination shows that the order of R_{st} is smaller than 10^{-3}. Hence the random variables x_s and x_t $(s \neq t)$ can practically be regarded as the statistically independent variables.

Since the normality and independence conditions are practically satisfied, the random variable χ^2 given by equation (5-1) will approximately follow the chi-square distribution. This fact justifies the use of the alternative goodness-of-fit test given by equation (5-2). It should be noted that this alternative test can be used not only for the case of $k < 4000$ but also for the case of $k \geq 4000$. As is shown in Section 4, however, the conventional test statistic becomes almost the same as the alternative test statistic when $k \geq 4000$ and hence the conventional test statistics may be used in that case, for the computational work becomes easier.

Last let us remark a bias brought by the use of the conventional test. From equation (4-3) the relation $m_s > \mathrm{Var}(x_s)$ holds. From this result and the fact that m_s' can practically be approximated by m_s, (recall section 4), it follows that

$$\sum_s \frac{(n_s - m_s)^2}{m_s} < \sum_s \frac{(n_s - m_s)^2}{\mathrm{Var}(x_s)} . \tag{5-5}$$

Noticing that the left hand side is the conventional test statistic and the right hand side is the alternative test statistic, this relation shows that the conventional test statistic is likely to misjudge a non-random point distribution to be a random point distribution. This misjudgment will be crucial when the number of points is less that 4000.

(6) CONCLUSIONS

The major conclusions of this paper are summarized as follows. If less than 2% (5%) error is allowable, then

i) the conventional goodness-of-fit test of the quadrat method should be used when the number k of points in more than 4000 (1200). In this case the number n of quadrats should be chosen from $T_{.02}$ ($T_{.05}$) of figure 4. Roughly speaking, $n = .06k$.

ii) When the number of points is less than 4000 (1200), the alternative goodness-of-fit test given by equation (5-1) should be employed.

iii) The conventional goodness-of-test is likely to misjudge a non-random point distribution to be a random point distribution, especially when the number of points is less than 4000 (1200).

FOOTNOTE

(1) Equation (2-1), (2-3), (2-4) and (2-5) respectively correspond to equations (13), (4) (11) and (22) of von Mises [11].

(2) Equation (5-3) corresponds to equation (25) of von Mises [11].

ACKNOWLEDGEMENTS

I express my thanks to N. Matsubara who brought the related literature to my attention. I also thanks to S. Ishikawa for his computational works. Needless to say, I am solely responsible for any errors which may remain.

REFERENCES

1. Bartlett, M. S., (1960), Stochastic Population Models in Ecology and Epidemiology, London: Methuen.

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

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

4. Cramér, H., (1946), Mathematical Methods of Statistics, Princeton: Prinston University Press.

5. Getis, A., Boots, B., (1978), Models of Spatial Processes, Cambridge: Cambridge University Press.

6. King, L. J., (1969), Statistical Analysis in Geography, Englewood Cliffs, N. J.: Prentice-Hall.

7. Lewis, P., (1977), Maps and Statistics, London: Methuen.

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

9. Rogers, A., (1974), Statistical analysis of spatial dispersion, London: Pion.

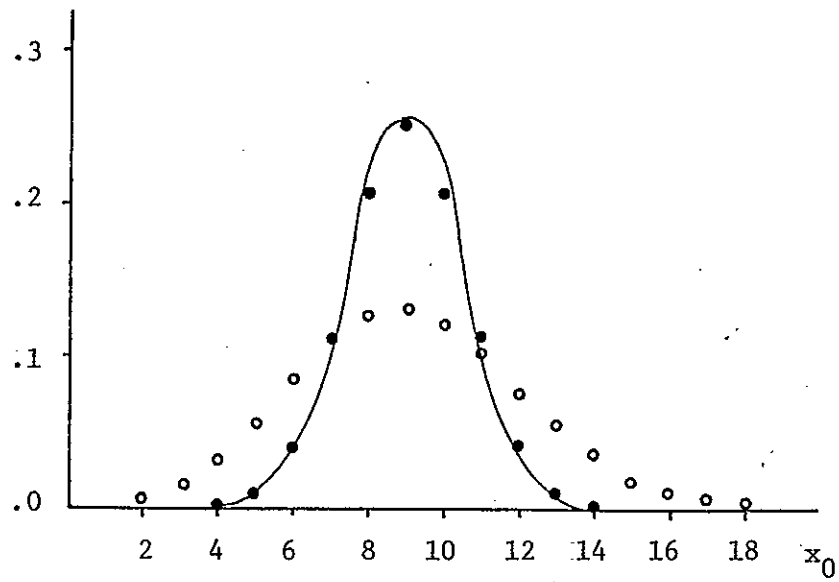
10. Roscoe, J. T., and Byars, J. A., (1971), An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistics, Journal of the American Statistical Association, 66, 755-759.

11. von Mises, R., (1939), Über Aufteilungs-und Besetzungs-Wahrscheinlich(k)eiten, Revue de la Faculté des Sciences de l'Université d'Istanbul, N. S., 4, 145-163.

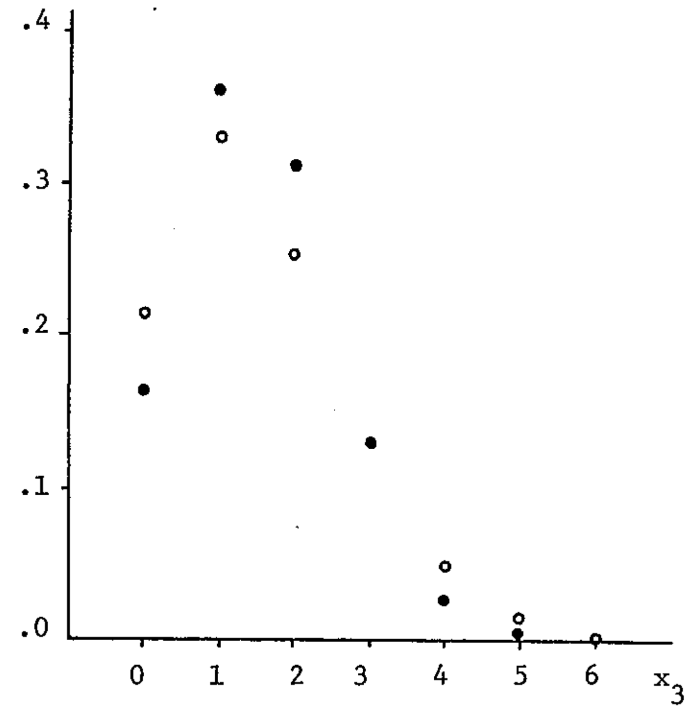
12. Wise, M. E., (1963), Multinomial probabilities and the X^2 and χ^2 distributions, Biometrika, 50, 145-154.

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

Figure 1 The exact and approximated distribution of the number x_s of quadrats having s points when the number k of points is 25 and that n of quadrats is 25 or 50

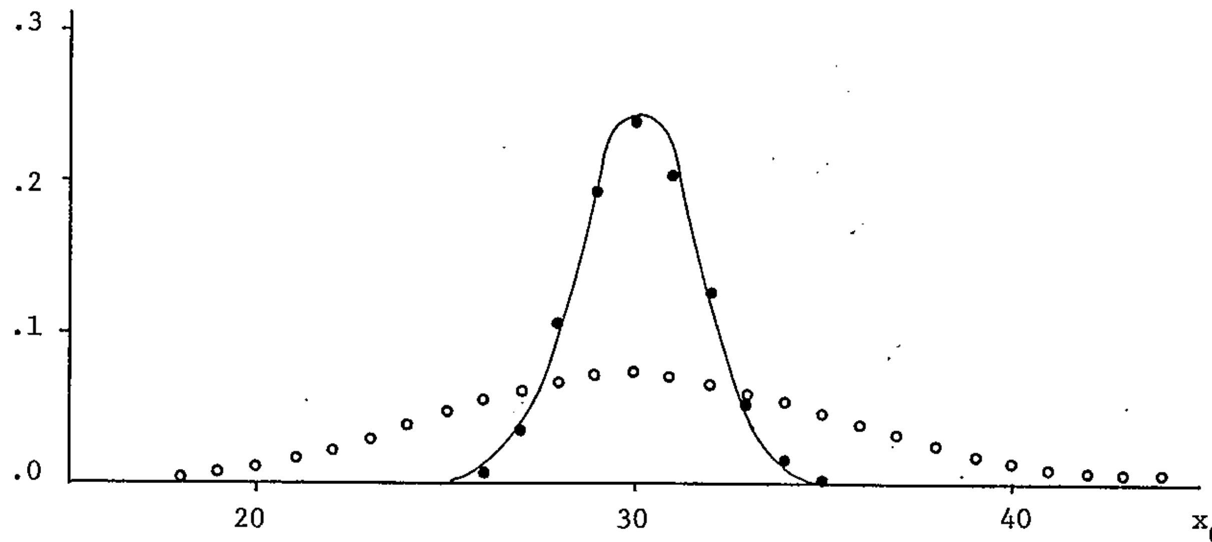


(a) k=n=25, s=0

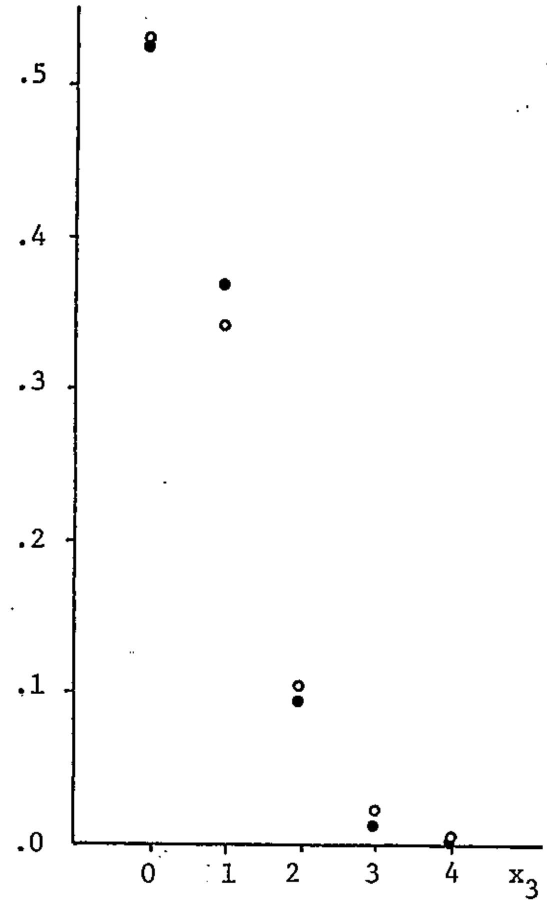


(b) k=n=25, s=3

Figure 1



(c) k=25, n=50, s=0



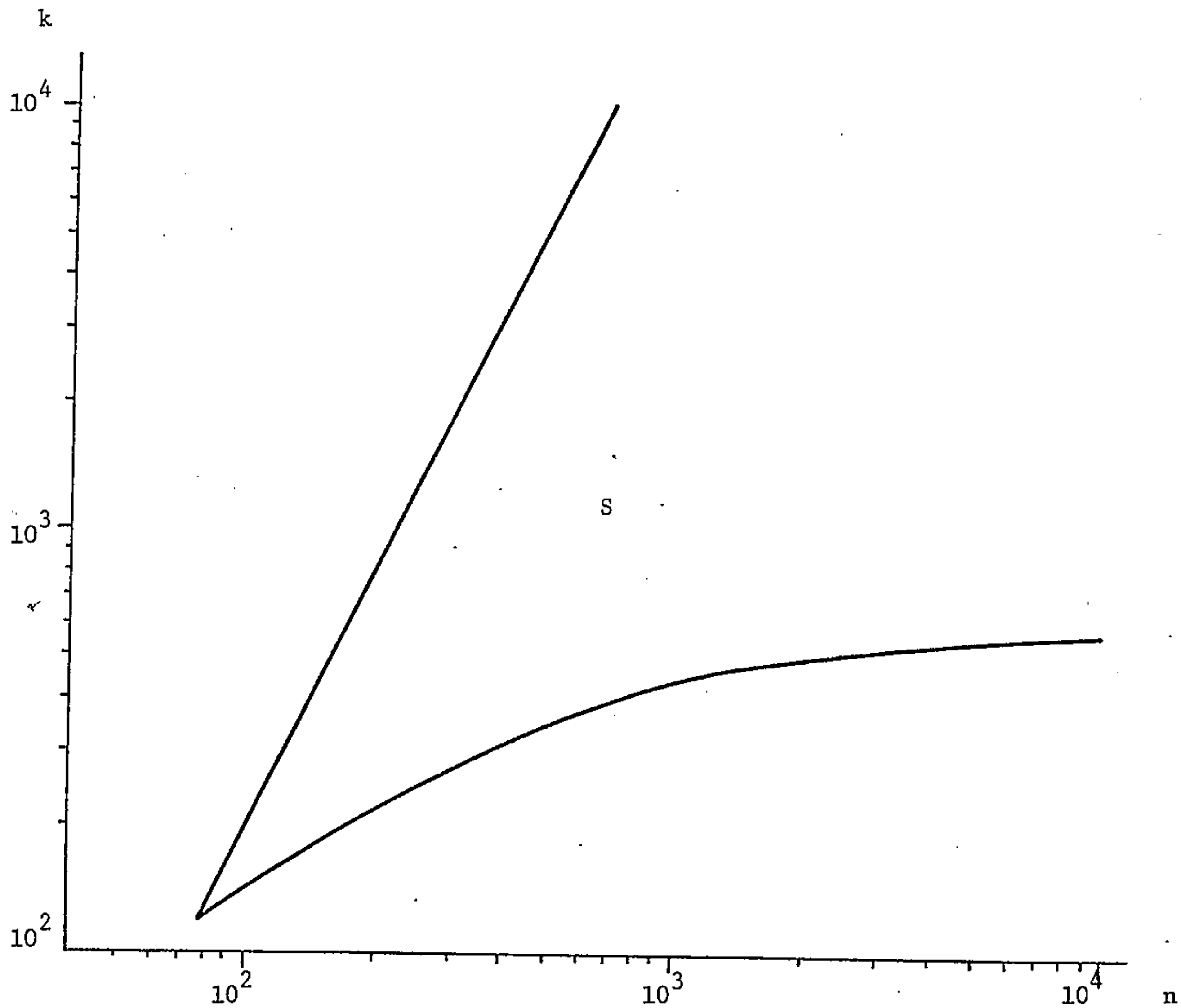
(d) k=25, n=50, s=3

● the exact distribution

○ the approximated distribution

—— the normal distribution

Figure 2 The number n of quadrats and that k of points which produce less than 1% errors in the expected value



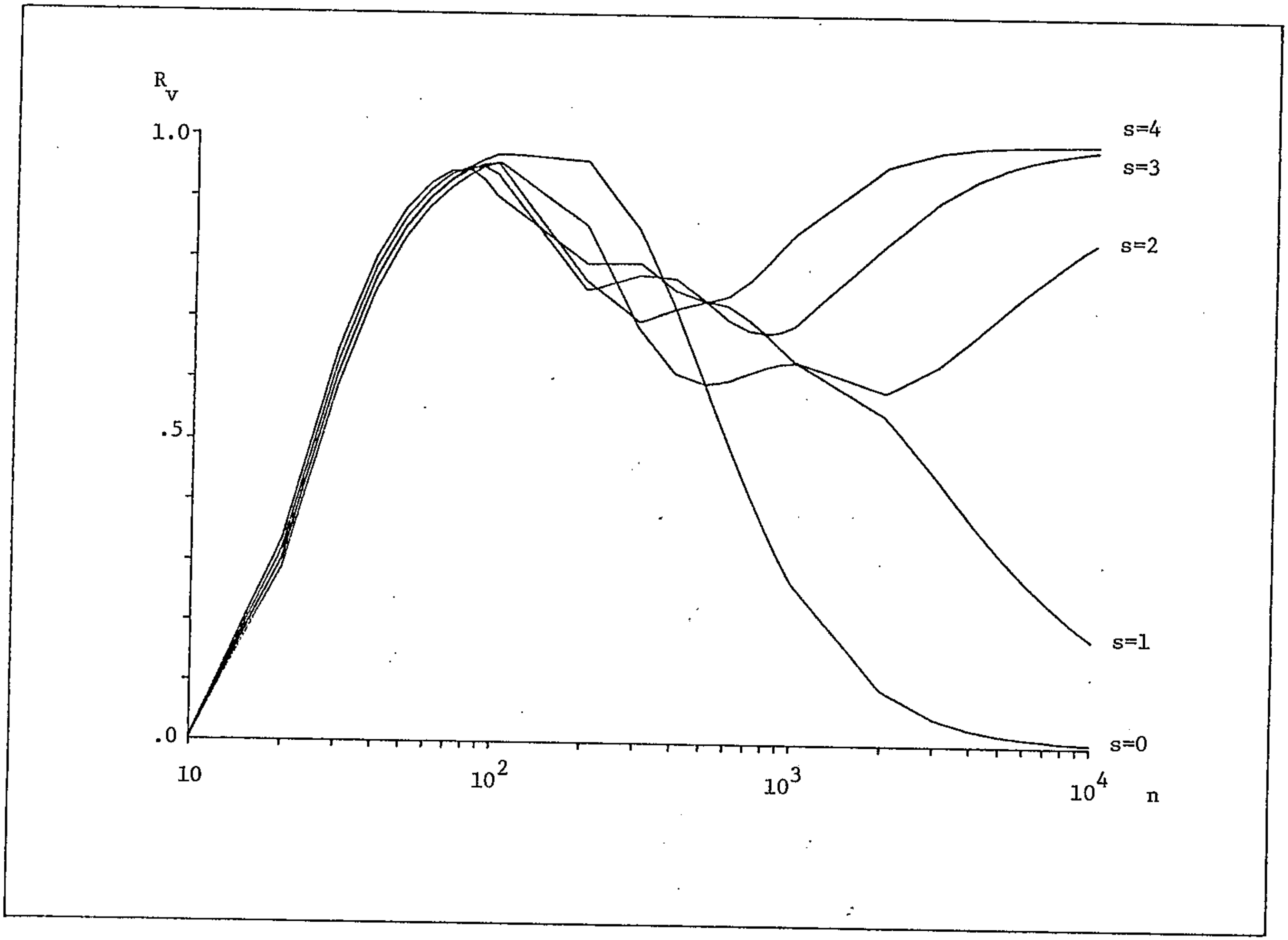


Figure 3 The ratio of the exact variance to the approximated variance, (the number of points is 1000 and that n of quadrats is $10 - 10^4$)

Figure 4 The number n of quadrats and that k of points which produce less than 2% ($T_{.02}$) and 5% ($T_{.05}$) errors in the variance