The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 265-275

An evaluation of the Mantel-Fleiss validity criterion for the Mantel-Haenszel statistic[1]

Younghae Chung[2] and Charles S. Davis[3]

Abstract

In testing the partial association between two variables after controlling for the S levels of a third factor, the Mantel and Haenszel (1959) statistic is often used. Since the statistic is based on the asymptotic distribution of the sum X of S hypergeometric variates, a guideline for the minimum requirements for the application of the statistic is useful. Mantel and Fleiss (1980) developed a criterion based on the guideline for the Pearson's X^2 statistic. The criterion requires the distance from the expected value to the closer bound of X to be at least five.

The Mantel-Fleiss (MF) criterion was studied through a simulation using the hypergeometric sampling scheme. The criterion is not satisfactory. The size of statistic exceeded nominal 0.05 level nearly 1/5 of the cases even when the criteion is met. However, the results show that the statistic is much more unstable and conservative when the criterion is not met.

1. Introduction

In many areas of health research, variables often contain only two categories and the relationship between two such variables is of interest. In such case, the resulting data can be summarized in a 2×2 table. The association between the two variables can then be tested using Pearson's chi-square statistic X^2 or the likelihood ratio statistic G^2, both of which have asymptotic chi-square distributions with one degree of freedom (χ_1^2) if the null hypothesis of no association is true.

Quite often, the relationship between the two variables is affected by other factors that are referred to as intervening factors. In the presence of such factors, the analysis carried out using the classic X^2 or G^2 statistics can be misleading. For example, consider a

[1] This research was supported in part by the National Cancer Institute grant CA39065.

[2] Lecturer, Dept of Computer Science and Statistics, Dongshin University, Naju-shi, Chonnam 520-714, Korea

[3] Professor, Dept of Preventive Medicine and Environmental Health, University of Iowa, Iowa City, Iowa 52242, U.S.A.

multicenter clinical trial where the effectiveness of a new treatment is compared with the standard treatment with respect to some outcome of interest that is measured on a dichotomous scale. Since the patient populations may be different among the centers, this difference needs to be taken into consideration in order to assess the true treatment-response relationship.

The resulting data can be summarized in S 2×2 tables, where S is the number of centers. For testing the average partial association between the treatment and response variables across the centers (strata), the Mantel-Haenszel (1959) statistic is commonly used. Under the hypothesis of no association between the treatment and the response in any of the strata, the statistic has an asymptotic χ_1^2 distribution.

Of the N_i subjects from the ith center, suppose n_{i1} subjects receive the new treatment and n_{i2} subjects receive the standard treatment. Let a_i and c_i denote the number of subjects with favorable response to the new and standard treatment, respectively (see Table 1). We assume the total sample size N_i and the row margins n_{i1} and n_{i2} are fixed.

Mantel and Haenszel (1959) proposed a statistic for testing the partial association using the hypergeometric probability model given fixed row margins and the number of favorable outcomes. Their test statistic is based on the distribution of $X=\sum_{i=1}^{S} a_i$, conditional on the row and column margins. Since each a_i has a central hypergeometric distribution under the null hypothesis, the expected value and the variance of a_i are

$$E(a_i)=\frac{n_{i1}m_{i1}}{N_i} \quad \text{and} \quad \mathrm{Var}(a_i)=\frac{n_{i1}n_{i2}m_{i1}m_{i2}}{N_i^2(N_i-1)},$$

and X has expected value $E(X)=\sum_{i=1}^{S} E(a_i)$ and variance $\mathrm{Var}(X)=\sum_{i=1}^{S}\mathrm{Var}(a_i)$. When

Table 1 Layout of the 2×2 table for the ith stratum

Treatment group	outcome Success	 Failure	Total
Treatment	a_i	b_i	n_{i1}
Control	c_i	d_i	n_{i2}
Total	m_{i1}	m_{i2}	N_i

N_i is large, a_i tends to normality. Even when N_i is as small as 2, the sum of the a_i values is asymptotically normally distributed from the central limit theorem for large S (Birch, 1964). The Mantel-Haenszel statistic, without continuity correction is

$$X^2_{\text{MH}} = \frac{(X - \text{E}(X))^2}{\text{Var}(X)} .$$

The statistic has an asymptotic χ^2_1 distribution when the null hypothesis of no association is true. When all N_i are large, the term $N_i - 1$ in the denominator of the variance can be replaced by N_i. In this case, X^2_{MH} is identical to the square of the statistic proposed by Cochran (1954). His test criterion, which is based on the assumption of independent binomial sampling, for testing the null hypothesis of no difference in success probabilities between the two binomial samples in any of the tables is the ratio of a weighted average difference of the proportions to the standard error. But only the MH statistic is appropriate when the N_i are small (Birch, 1964). When there are only 2 subjects in each stratum, as in a matched case-control study, the MH statistic is equivalent to the large sample test statistic proposed by McNemar (1947).

The validity of the chi-square approximation to the MH statistic has not been studied extensively. Even the results from the limited number of studies conducted on the small-sample properties of the MH test are conflicting. Considering several small tables (e.g., $N_i = 10$ and 15 for $S = 3$), Bennett and Kaneshiro (1974) reported from a simulation study that the MH statistic has significance levels close to the nominal .05 and .01 levels, and thus demonstrated the appropriateness of the normal approximation for the table configurations they considered. The MH test was also shown to maintain its size and power by O'Gorman, Woolson, Jones and Lemke (1988), where tables as small as $N_i = 4$, 8 and 16 were considered. The table configurations Bennett and Kaneshiro (1974) considered were further studied by Li, Simon and Gart (1979) who demonstrated instability when the success probabilities are close to zero or one. Results similar to Li et al (1979) were reported by Bennett and Underwood (1970) on the matched-pair design. However, no definitive guidelines are found from these studies.

Concerning the validity of the chi-square approximation to the distribution of the MH statistic, Mantel and Fleiss (1980) developed a criterion, based on the commonly followed guideline for the classic Pearson's X^2 statistic, that the smallest cell expectation should be at least 5. They noted that this guideline allows the expected value of the (1,1) cell entry in a 2×2 table to vary by at least 2 standard deviations in either direction. To ensure $\text{E}(X)$ to have enough variability, Mantel and Fleiss observed that the guideline be relative to

$X=\sum_{i=1}^{S} a_i$, not relative to the individual a_i. Let $(a_i)_L = \max(0,\ m_{i1} - n_{i2})$ and $(a_i)_U = \min(\ n_{i1},\ m_{i1})$ be the lower and upper bounds of the hypergeometric variate from table i, respectively. Then the lower and upper bounds of $X=\sum a_i$ are

$L = \sum (a_i)_L = \sum \max\ (0, m_{i1} - n_{i2}\ )$ and $U = \sum (a_i)_U = \sum \min\ (n_{i1}\ , m_{i1})$, respectively. The Mantel-Fleiss (MF) criterion for the chi-square approximation requires that $R = \min\ (\ \mathrm{E}(X) - L,\ U - \ \mathrm{E}(X))$ be at least 5.

For the special case where there are only two subjects in each table, as in matched pair studies, the MF criterion is much less strict than the usual criterion that is applied to McNemar's test. A commonly used guideline for McNemar's statistic is based on the criterion for the normal approximation to the binomial distribution (Rosner, 1986, p. 335). The normal approximation can be used when $npq \geq 5$ or $n \geq 20$ in this case, since $p = q = .5$. If we were to follow the criterion suggested by Mantel and Fleiss, $n = 10$ would be sufficient to give enough variation to the expected value of X. It is interesting to note that McNemar (1947) also suggested the use of the chi-square approximation when n, the sum of informative discordant cell frequencies, is at least 10.

2. Methods

A simulation study was carried out to investigate the appropriateness of the MF criterion for the validity of the asymptotic χ_1^2 distribution of the MH statistic. Since the asymptotic distribution of the X_{MH}^2 statistic was developed under the assumption of fixed row and column marginal distributions in every 2×2 table, the validity of the criterion was studied under the null hypothesis using a hypergeometric sampling model. Random variables a_i were generated from the hypergeometric distribution. A random number generated from a uniform (0, 1) distribution was transformed to a discrete uniform random integer in the range $[1,\ U - L + 1]$, where U and L are the maximum and minimum possible values of a_i for a given fixed total N_i, with fixed row margin n_{i1} and fixed column margin m_{i1}. This random variate was then converted to a hypergeometric variate using the alias method (Kronmal and Peterson, 1979). The remainder of the cell entries were determined from the marginal distributions.

It was possible to generate a complete set of S independent hypergeometric variables at the same time because the row and column marginal distributions remained constant in all S tables. The minimum and the maximum possible values as well as the expected values were

Table 2 First marginal totals (n_{i1} or m_{i1}) considered in the simulation study

	Number of Subjects in Each Stratum								
Margin type	2	3	4	5	6	8	10	15	20
B^*	1	1	2	2	3	4	5	7	10
U			1	1	1	2	2	3	5
VU						1	1	2	3

* B: Balanced margins: Integer of ($N_i/2$)
U: Unbalanced margins: Integer of ($N_i/4$)
VU: Very unbalanced margins: Integer of ($N_i/6$)

determined as soon as the margins were set. To ensure sufficient variability in the values of the statistics, only table dimensions with at least 10 possible values of the statistic X were considered for the simulation. This constraint was reflected in the table dimensions.

The table dimensions considered in our simulation study involved three factors:

(1) nine sample sizes for each of the S tables: N_i=2, 3, 4, 5, 6, 8, 10, 15 and 20.

(2) three row and column marginal distributions (n_{i1}/N_i and m_{i1}/N_i): 1/2, 1/4 and 1/6. These marginal distributions were chosen such that they would represent balanced, unbalanced, and very unbalanced marginal patterns. Even though the total sample size N_i cannot be assigned equally to the two row margins when it is odd, the row margins with $(N_i-1)/2$ were considered as balanced for the purposes of tabulation. When the sample size was small (N_i=2 or 3), only the balanced margins where $n_{i1}=1$ were considered. For medium sample sizes (N_i=4, 5 or 6), the balanced and the unbalanced margins (1/2 and 1/4) were considered. All three marginal distributions, balanced, unbalanced and very unbalanced margins, were considered for larger sample sizes (N_i= 8, 10, 15 or 20). When two marginal totals are the same, only one was considered as in N_i=6, where both 1/4 and 1/6 yield $n_{i1}=1$. Since the rows and the columns are interchangeable, only distinct combinations of marginal patterns were considered. For example, for N_i=6 a balanced row and unbalanced column marginal pattern was the same as an unbalanced row and balanced column marginal pattern. Thus only the balanced row and unbalanced column marginal pattern was considered. In this case, there were three types of marginal patterns. For $N_i \doteq 8$ and above, six types of marginal patterns were considered.

(3) ten possible values for the number of strata: $S=5$ for $N_i \geq 4$, $S=10$, 15, 20, 25, 30, 40, 50 for all N_i, and $S=75$ and 100 for $N_i \leq 4$.

The first marginal totals (n_{i1} or m_{i1}) considered for different margin types are shown in Table 2. For each combination of table configurations 10,000 replications were generated. The algorithm of Marsaglia, Zaman, and Tsang (1990) was used to generate the required uniform random numbers. Marsaglia et al reported that this generator had passed all of the stringent tests for randomness they considered and that it has a very long cycle length (2^{144}). This random number generator was chosen because it was the fastest among three random number generators we tested.

The size was determined by the proportion of the replications that exceeds the 95th percentile of the reference distribution, χ^2_1 in this case. Since the data are generated under the null hypothesis, we expect the size to be close to the nominal 0.05 level. Following Cochran's suggestion (see Upton, 1982), it is considered to be acceptable if the size of the test exceeds the nominal level by less than 20%. Thus, in this case, it is considered to be acceptable if the size reported is less than 0.06.

3. Results

The MF criterion for the chi-square approximation of the MH statistic requires the distance from the expected value to the lower and the upper bound of X to be at least five. Since the criterion only depends on the specifications of the table configurations, i.e., S, N_i, n_{i1} and m_{i1}, the MF criterion is either met or not met in all of the replications with the same table dimensions. As we limited our consideration to the table dimensions where there are at least 10 possible values of X, most of the balanced row and column margin cases met the MF criterion. The two that did not meet the MF criterion were ($S=10$, $N_i=3$) and ($S=5$, $N_i=5$), where the margins are not exactly $N_i/2$. When one of the margins is balanced, the tables meet the criterion except one, where the marginal total is not even, in both cases. The number of table configurations where the MF criterion is not met increases as the margins become more unbalanced. When both margins are very unbalanced, more than half of the table configurations considered did not meet the criterion. It is expected because when the margins are not balanced, the distribution is skewed and the distance from the expected value to the closer bound is short.

Figure 1 shows the plot of the size of the test against the distance R when the continuity correction is not employed. For plotting purpose, the logarithm of R is used. If the criterion were valid, most of the points are expected to fall below the reference bar (size=0.06, 20%

above the nominal level) when R is greater than 5, or $\log 5 > 1.609$. The plot shows, however, that many of the points fall above the reference bar even when the MF criterion is met (right side of the reference bar). Note that the statistic is more unstable when the criterion is not met — the minimum was 0.0164 and the maximum was 0.1033, while the size ranged from 0.0195 to 0.0875 when the criterion is met. Also, it appears to be more conservative — the average size was 0.0402 when the criterion is not met, compared to 0.0478 when it is met. It is only after R is as large as 80 ($\log R = 4.4$) when all of the sizes become smaller than 0.06. The criterion appears to be too liberal.

<Figure 1> about here.

Table 3 summarizes what is shown in figure 1. Whether the criterion is met or not, more than 17% of the table configurations have size above 0.06, i.e. unacceptable. The MF criterion does not seem to work adequately.

Interesting results were found when the continuity correction is used (Figure 2). All of the table dimensions give size below nominal 0.05 level even when the criterion is not met. In fact, the statistic is far too conservative when the criterion is not met.

<Figure 2> about here.

4. Conclusion and Discussions

The MF criterion for the chi-square approximation of the MH statistic was empirically evaluated. The criterion was not met in 41 of the 276 (14.9%) table configurations considered in the study. Whether the criterion was met or not, the proportion of the table configurations where the size exceeding 20% of the nominal 0.05 level (size > 0.06) was almost 20% — 17.1% when the criterion is not met and 19.1% when it is met. When the criterion is not met, however, the statistic was much more unstable and conservative.

Table 3 Proportion the size exceeding 20% of the nominal 0.05 level (uncorrected) by Mantel-Fleiss criterion

	Size of the test	
MF criterion	≤ 0.06	> 0.06
not met ($n=41$)	82.9	17.1
met ($n=235$)	80.9	19.1

The MH statistic with continuity correction is known to be conservative. It never gave size larger than 0.05 in this study. That means the criterion is not needed for the corrected statistic.

Regardless of the usage of the continuity-correction, the MF criterion does not seem to be discriminating the cases where the approximation is valid. This study leads to the conclusion that we need a better criterion.

There are some limitations in this study. We considered only table configurations where there were more than 10 possible values for X. Smaller tables need to be investigated for a better understanding. Also, we suggest another study based on the independent binomial sampling scheme since the MH statistic is the same as the square of Cochran's statistic.

References

[1] Bennett, B.M. and Kaneshiro, C. (1974). On the small-sample properties of the Mantel-Haenszel test for relative risk, *Biometrika*, Vol. 61, 233-236.

[2] Bennett, B.M. and Underwood, R.E. (1970). On McNemar's test for the 2×2 table and its power function, *Biometrics*, Vol. 26, 339-343.

[3] Birch, M.W. (1964). The detection of partial association, I: The 2×2 case, *Journal of Royal Statistical Society B.*, Vol. 26, 313-343.

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

[5] Kronmal, R.A. and Peterson, A.V. Jr. (1979). On the alias method for generating random variables from a discrete distribution, *American Statistician,* Vol. 33, 214-218.

[6] Li, S.H., Simon, R.M. and Gart, J.J. (1979). Small sample properties of the Mantel-Haenszel test, *Biometrika,* Vol. 66, 181-183.

[7] Mantel, N. and Fleiss, J.L. (1980). Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure, *American Journal of Epidemiology*, Vol. 112, 129-134.

[8] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of National Cancer Institute*, Vol. 22, 719-748.

[9] Marsaglia, G., Zaman, A. and Tsang, W.W. (1990). Toward a universal random number generator, *Statistics and Probability Letters,* Vol. 8. 35-39.

[10] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, Vol. 12, 153-157.

[11] O'Gorman, T.W., Woolson, R.F., Jones, M.P., and Lemke, J.H. (1988). A Monte Carlo study of three odds ratio estimators and four tests of association in several 2×2 tables when the data are sparse, Communications in Statistics, *Simulation and*

Computing, Vol. 17, 813-835.

[12] Rosner, B.A. (1986). Fundamentals of Biostatistics, 2nd Ed. Boston: Duxbury Press.

[13] Upton (1982). A comparison of alternative tests for the 2×2 comparative trial, *Journal of Royal Statistical Society A,* Vol. 145, 86-105.

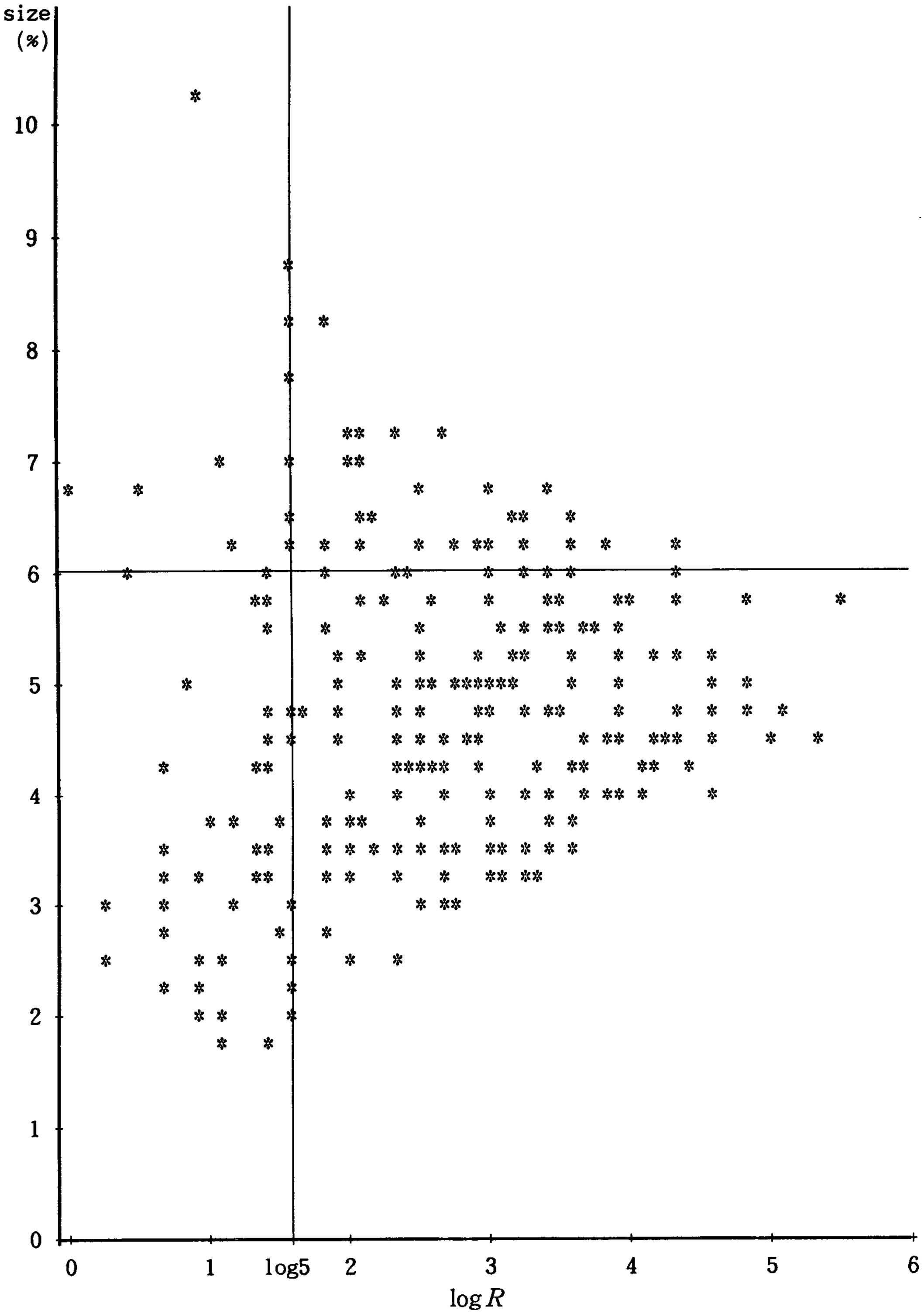

Figure 1. Size of the statistic (uncorrected) by log$R$. $R = \min(\mathrm{E}(X) - L,\ U - \mathrm{E}(X))$

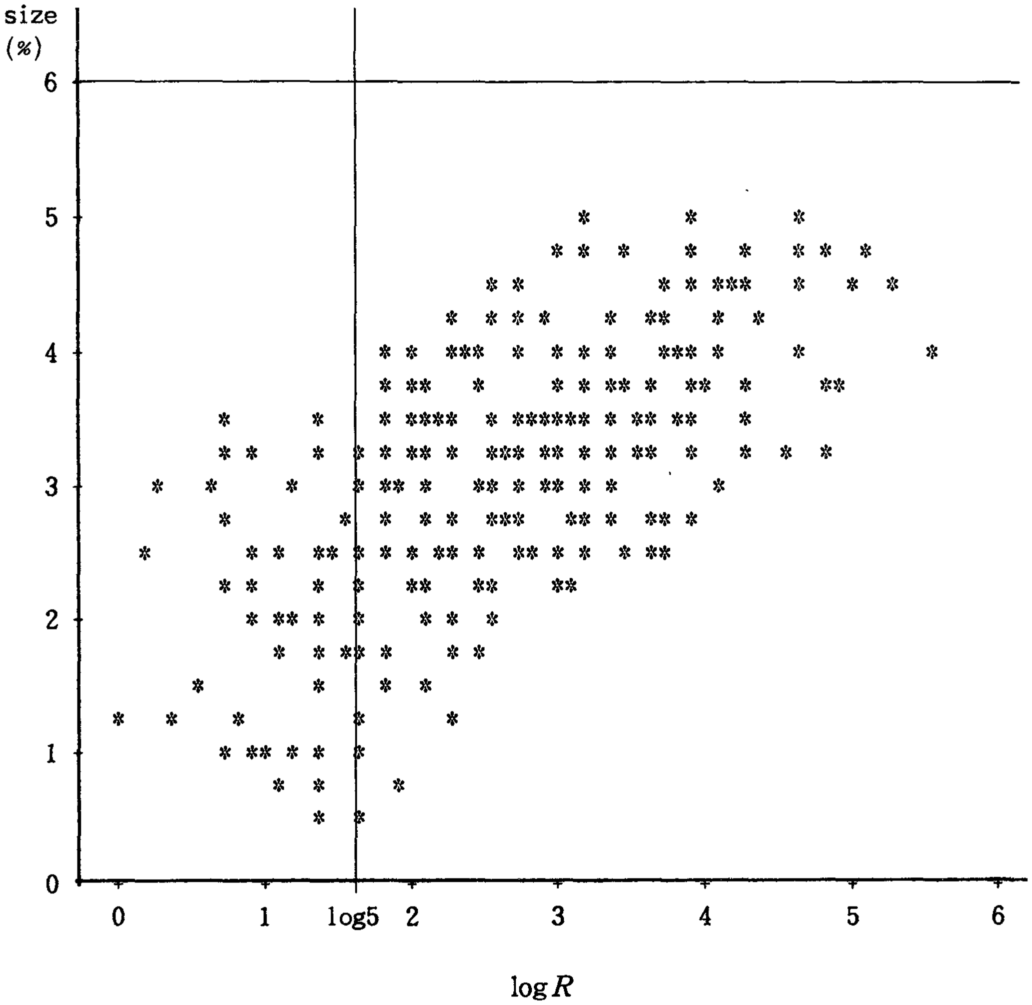

Figure 2. Size of the statistic (corrected) by log$R$. R=min(E(X)−L, U− E(X))