DOCUMENT RESUME

ED 064 318 TM 001 483

AUTHOR Kittleson, Howard M.; Roscoe, John T.
TITLE An Empirical Comparison of Four Chi-Square and Kolmogorov Models for Testing Goodness of Fit to Normal.

PUB DATE Apr 72
NOTE 8p.; Paper presented at the annual convention of the American Educational Research Association, Chicago, April, 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Comparative Testing; *Goodness of Fit; *Hypothesis Testing; Measurement Techniques; *Predictor Variables; Research Methodology; Testing; Test Validity; Theories
IDENTIFIERS Chi Square Test; Kolmogorov Test

ABSTRACT

This study compares the relative power and robustness of the chi-square and Kolmogorov statistics with both the linear score scale and equal areas models. It is limited to the situation in which the mean and standard deviation are fixed by the hypothesis (a necessary constraint with the Kolmogorov tests). Two tables are presented which report the findings for the null hypothesis and the findings for the false hypothesis (sampling from a uniform distribution and testing for normality). In each case the table entry is the number of rejections in 10,000 samples. Conclusions of the study proved the chi-square equal areas model to be superior to the chi-square linear score scale model and to both the Kolmogorov tests. (Author/LS)

AN EMPIRICAL COMPARISON OF FOUR CHI-SQUARE AND KOLMOGOROV MODELS FOR TESTING GOODNESS OF FIT TO NORMAL[1]

Howard M. Kittleson and John T. Roscoe, Kansas State University

1. BACKGROUND

The traditional statistical procedure for testing goodness of fit to normal has used the chi-square approximation of the multinomial and a model in which cell limits are defined by dividing a standard score scale into equal parts (a linear socre scale model). This model has been criticized because the expected frequencies in the tails of the distribution tend to be very small with samples of reasonable size (say n = 100 or less). However, recent research by a number of investigators indicates that small expected frequencies are not the handicap they have long been believed to be.

Several textbook authors (Hays (1963) and Roscoe (1969), for example) have suggested an alternative chi-square model in which cell limits are defined by dividing the area under the curve into equal parts (an equal areas model). In addition to overcoming the problem of small expected frequencies in the tails, this procedure focuses on the added power characteristic of the chi-square approximation with uniform expected frequencies. This model, however, has been criticized for lack of discrimination in the tails of the distribution.

A number of authors (Massey (1942) and Goodman(1954), for example) have suggested the Kolmogorov statistic as an alternative to chi-square in situations of the sort described. Some researchers have raised serious doubts about the utility of the Kolmogorov tests with samples of reasonable size (again, n = 100 or less).

[1]Paper presented at the annual convention of the American Educational Research Association, Chicago, April, 1972.

The problem arises because the chi-square test is an exact test as the sample size (n) approaches infinity, and the Kolmogorov test is an exact test as the number of cells (k) approaches infinity. Under all other circumstances, the two tests are approximations.

This study was undertaken to compare the relative power and robustness of the chi-square and Kolmogorov statistics with both the linear score scale and equal areas models. It is limited to the situation in which the mean and standard deviation are fixed by the hypothesis (a necessary constraint with the Kolmogorov tests). The authors were primarily concerned with applications of the sort encountered by behavioral scientists, but they believe their research will be of interest to scientists of other disciplines.

Cochran (1952) reviewed the historical development of chi-square tests of goodness of fit and related research dealing with such issues as minimum expected frequencies. This article, plus a later one by the same author (1954), has been most often cited by textbook authors with respect to these topics. Cochran indicates that one or two expectations may fall as low as one-half providing the remainder are above the conventional limits of five or ten, and he drew a tentative conclusion that the approximation might be acceptable if all expected frequencies were small but at least equal to two. For Cochran, the approximation was acceptable if the true probability fell within the range 0.04 to 0.06 for the 0.05 tabular value and within 0.007 to 0.015 for the 0.01 tabular value. He also suggests some investigators would be content with less restrictive limits.

Mann and Wald (1942) demonstrated that optimum power for the chi-square test of goodness of fit to some continuous distribution is achieved when the expectancies are equal. They also derived a mathematical strategy for selecting the optimum number of cells with very large samples (n = 200 or more). Watson (1957) appears to be the first to have suggested the equal areas model for the chi-square test of goodness of fit to normal. He also suggested that the number of cells should be at least ten. Kempthorne (1967) also recommended the use of the equal areas model, especially for goodness of fit to normal. His findings (based in part upon mathematical considerations and in part upon a small Monte Carlo study) suggest that a good approximation is achieved when the number of cells (k) is set equal to the sample size (n). In another Monte Carlo study of limited scope, Dahiya (1971), found that the approximation tends to be liberal if the value of k is set too high, specifically if k is larger than n.

The most extensive empirical study of the questions of sample size, minimum expected frequencies and number of cells for use with chi-square tests of goodness of fit appears to be that of Roscoe and Byars (1971). They demonstrated that an acceptable approximation (using Cochran's standards) is achieved with expectancies as small as one when testing goodness of fit to uniform. The approximation is not quite so good with non-uniform hypotheses. With moderate departures from the uniform case, they found an acceptable approximation is achieved at the 0.05 level with average expected frequencies as small as one, but with extreme departures from uniform they recommend expectancies be held to two or more. In either case, the average expected frequencies must be doubled to insure a good approximation at the 0.01 level. The approximation tends to be liberal if the expectancies are permitted to fall below these recommendations. They did not examine the hypothesis of goodness of fit to normal.

Massey (1951) established the superiority of the Kolmogorov test to the chi-square approximation for very large samples (n = 200 to 2000). However, it is the common experience of other investigators that his findings do not generalize to smaller samples. For example, Slakter (1965) compared chi-square to Kolmogorov tests of goodness of fit in a Monte Carlo study. In sampling from a uniform distribution, he found the Kolmogorov test to be markedly and consistently conservative under conditions most favorable to the test, and chi-square proved to be quite robust even with very small samples.

2. PROCEDURE AND FINDINGS

Uniformly distributed pseudo-random numbers were generated using the power residue method. An algebraic transformation was used to derive normally distributed random numbers for testing under the null hypothesis; the uniformly distributed numbers were retained for the test of the false hypothesis. Ten thousand sets of samples were drawn for each combination of sample size and number of cells used in the research. For samples of size 10, 20, 30, and 50, the number of cells was set equal to 6, 10, and 20. For samples of size 50, the number of cells was also set equal to 50. Both true and false hypotheses were tested for all four models (linear score scale and equal areas for both the chi-square and Kolmogorov tests) and for each combination of n and k.

Table 1 reports the findings under the null hypothesis. Table 2 reports the findings for the false hypothesis (sampling from a uniform distribution and testing for normality) In each case, the table entry is the number of rejections in 10,000 samples. The expected table values are, of course, 500 at the 0.05 level and 100 at the 0.01 level under the null hypothesis.

TABLE 1

Goodness-of-fit to Normal: Number of Rejections in 10,000 Samples under the Null Hypothesis

n, k	Chi-square linear scale		Chi-square equal areas		Kolmogorov linear scale		Kolmogorov equal areas	
	.05	.01	.05	.01	.05	.01	.05	.01
10, 6	646	384	430	175	70	28	133	36
10, 10	888	524	391	116	78	38	74	74
10, 20	777	374	312	154	115	68	219	69
20, 6	546	245	504	96	120	3	177	25
20, 10	977	416	510	105	177	29	345	16
20, 20	680	246	409	91	271	15	413	17
30, 6	575	186	503	107	56	14	111	31
30, 10	1022	386	442	105	132	26	145	43
30, 20	667	256	443	122	154	52	275	54
50, 6	543	161	485	97	70	10	133	20
50, 10	1063	388	489	98	138	22	183	136
50, 20	610	180	496	119	169	28	199	25
50, 50	1358	413	874	131	297	48	345	59

TABLE 2

Goodness-of-fit to Normal: Number of Rejections in 10,000 Samples when Sampling from Uniform Distribution

n, k	Chi-square linear scale .05	.01	Chi-square equal areas .05	.01	Kolmogorov linear scale .05	.01	Kolmogorov equal areas .05	.01
10, 6	886	263	635	278	196	50	335	69
10, 10	1008	422	511	179	252	87	167	167
10, 20	1165	676	672	351	293	145	441	162
20, 6	1483	526	1036	269	188	26	348	93
20, 10	1643	688	944	264	446	73	743	69
20, 20	1965	942	1277	479	519	88	906	71
30, 6	2326	966	1355	425	163	42	337	115
30, 10	2288	989	1100	337	442	107	437	165
30, 20	2877	1550	2030	955	471	172	751	177
50, 6	3995	2045	2076	757	277	47	567	103
50, 10	3902	1896	1577	501	549	130	684	170
50, 20	4987	3081	3579	1937	798	219	780	175
50, 50	3406	1573	4846	1895	1000	246	1099	275

3. CONCLUSIONS

The interpretation of the findings requires some convention with respect to what constitutes an acceptable approximation. The authors have elected to follow Cochran's recommendations cited earlier (0.04 to 0.06 for the 0.05 level and 0.007 to 0.015 for the 0.01 level) though they suspect some investigators will be content with less restrictive limits.

The chi-square equal areas model proved to be superior to the chi-square linear score scale model and to both of the Kolmogorov tests. In every case studied, the chi-square test utilizing the traditional linear score scale model was liberal with respect to Type I errors. The Kolmogorov test was clearly inferior in every respect, being so conservative as to invalidate its use. This is consistent with the findings of Slakter and others.

The chi-square equal areas model was erratic with samples of size 10; however, an acceptable approximation was achieved with all other sample sizes (n = 20, 30, and 50). The test was also liberal with n = 50 and k = 50; this is consistent with the findings of Dahiya and contrary to those of Kempthorne. The power appears to optimize with k set equal to 20. The authors are tempted to suggest that chi-square tests of goodness of fit to normal be standardized to use the equal areas model with 20 cells. In addition to the robustness and power evidenced by this strategy, it has the added advantage of removing the arbitrary element so characteristic of current practice.

While the research was limited to the situation in which the mean and standard deviation are fixed by the hypothesis, Watson's manuscript suggests the findings with respect to the chi-square equal areas model should generalize to the situation in which the mean and standard deviation are estimated from sample data.

REFERENCES

COCHRAN, W. G. The chi-square test of goodness of fit. Annals of Mathematical Statistics, 1952, 23, 315-45.

COCHRAN, W. G. Some methods for strengthening the common chi-square tests. Biometrics, 1954, 10, 415-51.

DAHIYA, R. C. On the Pearson chi-squared goodness of fit test statistic. Biometrika, 1971, 58, 685-86.

GOODMAN, L. A. Kolmogorov-Smirnov tests for psychological research. Psychological Bulletin, 51, 160-68.

HAYS, W. L. Statistics for Psychologists. New York: Holt, Rinehart and Winston, 1963.

KEMPTHORNE, O. The classical problem of inference--goodness of fit. Fifth Berkely Simposium on Mathematical Statistics and Probability, 1967, 1, 235-49.

MASSEY, F. J. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 1951, 46, 68-78.

ROSCOE, J. T. Fundamental Research Statistics for the Behavioral Sciences. New York: Holt, Rinehart and Winston, 1969.

ROSCOE, J. T. and BYARS, J. A. 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, 1971, 66, 755-59.

SLAKTER, M. J. A comparison of the Pearson chi-square and Kolmogorov goodness of fit tests with respect to validity. Journal of the American Statistical Association, 1965, 60, 854-58.

WATSON, G. S. The chi-square goodness of fit test for normal distributions. Biometrika, 1957, 44, 336-48.