ON THE CHOICE OF THE NUMBER AND WIDTH OF CLASSES FOR THE CHI-SQUARE TEST OF GOODNESS OF FIT*

C. Arthur Williams, Jr.
Columbia University

This article describes in non-mathematical fashion the technique suggested by H. B. Mann and A. Wald for selecting the number and width of class intervals for the chi-square test of goodness of fit when the null hypothesis distribution is continuous and completely specified. The number of classes is selected by means of a formula depending upon the sample size and the level of significance and the class limits are chosen such that each class contains the same number of items under the null hypothesis. Finally it is suggested that the number of classes as given by the formula may be halved for practical purposes.

In most statistical problems the distribution of the universe from which a sample has been drawn is unknown. To test whether or not this sample was drawn from a population having a specified distribution, statisticians commonly employ the chi-square test of goodness of fit.

In order to carry out this test, one first sets up a null hypothesis stating that the sample was drawn from a universe with a known distribution. If the parameters are based on standards, theory, or past experience, the distribution is completely specified. If the parameters are estimated from the sample, only the type of the distribution is specified. Next one computes

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - Np_i)^2}{Np_i} \tag{1}$$

where f_i is the actual or observed number of frequencies in the ith class, p_i the probability under the null hypothesis that an observation will fall into the ith class, N the number of observations in the sample, and k the number of classes. It can be shown that as the size of the sample approaches infinity the distribution of this statistic approaches the chi-square distribution with $k-1-s$ degrees of freedom where s is the number of parameters estimated from the sample. In practice the chi-square distribution is assumed to hold for finite values of N and one ascertains the value of $\chi^2_{k-1-s}(\alpha)$ such that the probability of χ^2 being

* This article is based on a Master's Essay written at Columbia University under Professor T. W. Anderson, Jr.

greater than or equal to $\chi^2_{k-1-s}(\alpha)$ is equal to α, the level of significance or probability of rejecting the null hypothesis when it is true. If the computed value of χ^2 is equal to or exceeds $\chi^2_{k-1-s}(\alpha)$, the null hypothesis is rejected. If the computed value is less than $\chi^2_{k-1-s}(\alpha)$, one accepts the null hypothesis.

Despite its wide use this test has some serious limitations. Firstly there are many distributions which will give the same theoretical class frequencies as the null hypothesis distribution. It is quite possible that we may accept a hypothesis stating that a particular sample is drawn from a normal population say, when in fact it belongs to an alternative population which gives the same theoretical class frequencies. If the roles of the null hypothesis distribution and an alternative distribution of this type were reversed, the new null hypothesis would also be accepted, the computed value of χ^2 being the same as in the first case. Such a null hypothesis distribution and an alternative distribution are shown in Diagram 1. Secondly, it is possible to choose the number and width of class intervals for the test in many different ways, some of which may change the result of the test. This is demonstrated in Diagram 2. Let the dashed line represent the null hypothesis distribution and the blocks the observed frequencies. If in conducting the test we used classes I, II, III, and IV, we would reject the null hypothesis. If we used classes I′ and II′, we would accept the null hypothesis. Thus it is clear that as long as this subjective element remains we always run the risk of influencing our results by the choice of the class interval. Thirdly, we know nothing about the power of the test, the probability of rejecting the null hypothesis when it is false.

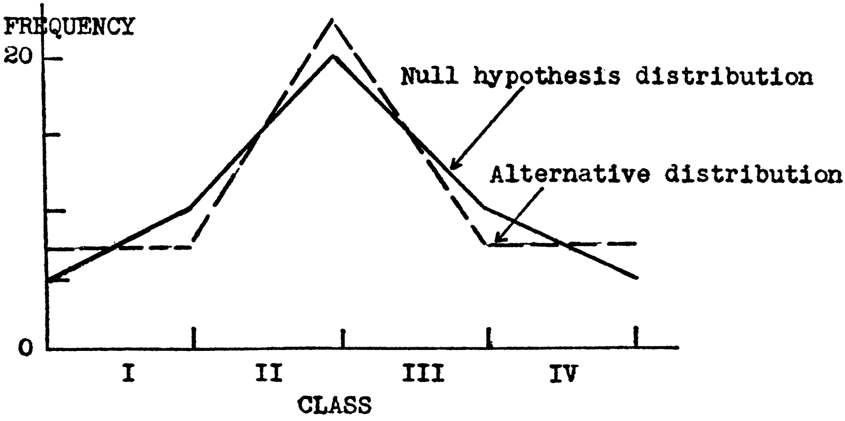

DIAGRAM 1

In the September 1942 issue of the *Annals of Mathematical Statistics* there appeared an article by H. B. Mann and A. Wald of Columbia University entitled "On the Choice of the Number of Intervals in the Application of the Chi-Square Test." In their article Mann and Wald proposed a solution to this problem in the case where the population parameters are not estimated from the sample but are based on standards, theory, or past experience. It will be the purpose of this article to point out the implications of the aforementioned article for the practical statistician and to suggest a reduction in the number of classes.

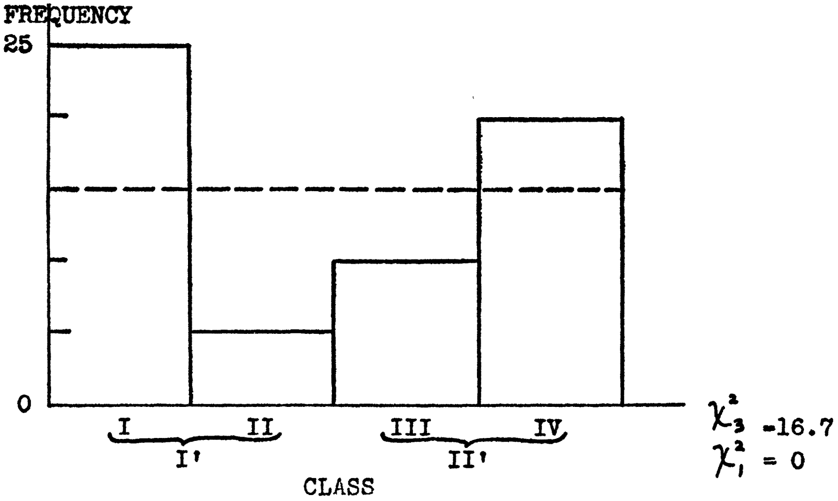

DIAGRAM 2

I. HOW TO USE THE MANN-WALD TECHNIQUE

The following mechanical procedure is necessary to carry out the suggestions of Mann and Wald:

1) Set up a null hypothesis stating that the sample was drawn from a universe with a completely specified probability distribution.
2) Compute the number of classes to be used by means of the following formula:

$$k = \left[4 \sqrt[5]{\frac{2(N-1)^2}{c^2}} \right]$$

where k is the number of classes, N is the number of items in the sample, and c is obtained from a table of areas under the normal curve such that

$$\int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \alpha,$$

the level of significance. A sample calculation follows:

Let us suppose that a sample of 1000 items is drawn and the test is to be conducted using the 5 per cent level of significance. The formula for k may be rewritten as follows:

$$k = \left[4 \text{ antilog} \left(\frac{1}{5} \log \frac{2(N-1)^2}{c^2} \right) \right].$$

From a table of areas under a normal curve we find that for the

TABLE 1

NUMBER OF CLASSES (k) AND DISTANCE (Δ) FOR THE 5% LEVEL OF SIGNIFICANCE AND FOR SELECTED VALUES OF N BETWEEN 200 AND 2000

N	k*	Δ†
200	31	.1605
250	34	.1469
300	36	.1343
350	39	.1284
400	41	.1213
450	43	.1157
500	45	.1112
550	46	.1052
600	48	.1024
650	50	.1000
700	51	.0961
750	53	.0945
800	54	.0914
850	55	.0887
900	57	.0877
950	58	.0855
1000	59	.0834
1100	62	.0812
1200	64	.0782
1300	66	.0757
1400	68	.0734
1500	70	.0715
2000	78	.0629

* Values of k were obtained by taking the greatest integer contained in the value given by the formula.

† Values of Δ are those for which the corresponding integer k gives the maximum power for the "worst distribution."

TABLE II

NUMBER OF CLASSES (k) AND DISTANCE (Δ) FOR THE 1% LEVEL OF SIGNIFICANCE AND FOR SELECTED VALUES OF N BETWEEN 200 AND 2000

N	k*	Δ*
200	27	.1847
250	29	.1657
300	32	.1577
350	34	.1479
400	35	.1369
450	37	.1315
500	39	.1273
550	40	.1209
600	42	.1184
650	43	.1137
700	45	.1120
750	46	.1083
800	47	.1051
850	48	.1022
900	49	.0997
950	50	.0974
1000	51	.0953
1100	53	.0918
1200	55	.0888
1300	57	.0862
1400	59	.0841
1500	61	.0823
2000	68	.0728

* See notes to Table I.

5 per cent level of significance $c = 1.64$. Substituting the values for N and c in the above formula, we find that

$$k = \left[4 \text{ antilog}\left(\frac{1}{5} \log \frac{2(1000-1)^2}{(1.64)^2}\right)\right]$$

$$k = [4 \text{ antilog } (\tfrac{1}{5} \log 742{,}008.08)]$$

$$k = [4 \text{ antilog } (1.174104)]$$

$$k = [4(14.9)] = [59.6] = 59.$$

Since this procedure is somewhat laborious, tables of N and k (Tables I and II) are provided for the 5 and 1 per cent levels of significance in this article. These tables are not complete since they deal only with selected values of N between 200 and 2000 but for practical purposes simple interpolation will give the desired number of classes for any sample size within this range. Also included in the tables are values of Δ or "distance," a term which will be introduced later.

3) Choose the class limits such that the number of theoretical frequencies in each class is equal to N/k. Notice that in this procedure the class limits and not the class frequencies vary from one class to the next.
4) From the data determine the actual number of observations falling in each of the classes.
5) Compute

$$\chi^2 = \frac{k}{N} \sum_{i=1}^{k} f_i^2 - N. \tag{2}$$

This formula is obtained by substituting $p_i = 1/k$ in formula (1) and simplifying it.

6) Determine $\chi^2_{k-1}(\alpha)$ using a chi-square table. There are $k-1$ degrees of freedom and the level of significance is α. Use the critical region $\chi^2 \geqq \chi^2_{k-1}(\alpha)$ to test the null hypothesis.

II. ADVANTAGES AND LIMITATIONS OF THE MANN-WALD TECHNIQUE

At first glance this procedure seems much more complicated han the ordinary one of selecting equal class intervals along the horizontal axis and it is reasonable to ask what advantages are to be obtained by using this suggested technique. Before discussing these advantages, however, it is necessary to define two terms which will be used throughout the following discussion.

1) The cumulative distribution function (*cdf*) is defined as the probability that a random variable X be less than or equal to a given value, say "a". For example, consider the familiar bell-shaped normal distribution. If we were to graph this as an ogive letting the vertical axis denote the percentage of total items having a value equal to or less than the value stated on the horizontal axis, we would have a *cdf* for the normal distribution. We could do the same for each probability distribution—binomial, Poisson, etc.—and in this discussion we will be referring to the *cdf* when we speak of a probability distribution. A typical *cdf* is shown in Diagram 3.
2) Consider two such *cdf*'s. For each value of "a" along the horizontal scale there is a numerical difference between the two *cdf*'s along the vertical scale. The absolute value of the greatest such

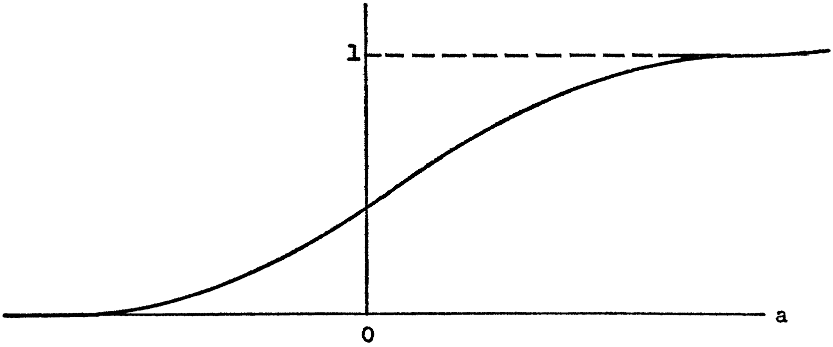

DIAGRAM 3

numerical distance will be defined as the distance between the two *cdf*'s. It is this distance that we will refer to when we speak of the power of the test. To clarify this point see Diagram 4.

It should be noted at this time that the Mann-Wald theory is an asymptotic theory and that it has been rigorously proven only for sample sizes greater than or equal to 450 and the 5 per cent level of significance, and for sample sizes greater than or equal to 300 and the 1 per cent level of significance. However the authors state their belief that the results hold approximately for sample sizes as low as 200 and may be true for considerably smaller samples. With this qualification and the aid of the definitions given, the advantages of the test may now be considered.

1) By obtaining k from a formula or table and by choosing the class limits such that there are N/k theoretical frequencies in each class, the subjective element is removed from the choice of the number and width of the classes.

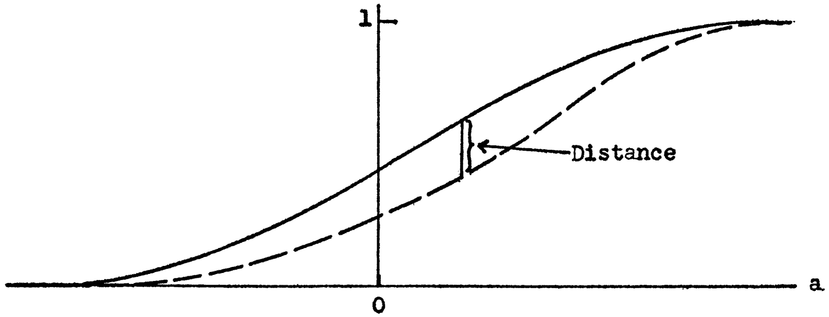

DIAGRAM 4

2) The maximum distance of those *cdf*'s which have the same class frequencies as the null hypothesis *cdf* is minimized. In other words, the Mann-Wald technique does not eliminate the first limitation to the chi-square test but it does minimize the maximum distance of such alternative distributions.

   This follows from the fact that the maximum distance between such *cdf*'s and the null hypothesis *cdf* is equal to the maximum class probability since by definition the ogive or *cdf* must rise or remain at the same level as we move from left to right along the horizontal axis and two *cdf*'s having the same class frequencies must intersect above the upper limit of each class. We minimize this maximum class probability by setting all the class probabilities equal for if we choose the class probabilities in any other way there will be at least one class probability greater than $1/k$.

3) The power of the test for those *cdf*'s whose distance from the null hypothesis distribution is greater than or equal to Δ as given in Tables I and II is for practical purposes greater than or equal to one-half. In other words, the probability of rejecting the null hypothesis when the universe actually has a *cdf* a distance Δ or greater away from the null hypothesis *cdf* is equal to one-half or more. We can make no such statement for the ordinary test.

   This statement and the two which follow require a mathematical proof which can be found in the article by Mann and Wald.

4) If a number of classes different from the one given by the formula is used to conduct the test, there will be at least one *cdf* whose distance from the null hypothesis *cdf* is greater than or equal to Δ such that the power of the test for that *cdf* is less than one-half.

5) The choice of equal class probabilities gives us an unbiased test in the sense that when the class frequencies in the universe are not equal, the probability of accepting the null hypothesis is less than when they are equal. In other words, when the null hypothesis is true, we have a larger chance of accepting it than when it is not true.

6) There is no need to worry about having more than five theoretical frequencies in each class for this condition is automatically fulfilled when the test is applicable.

Like all other tests, this one also has its limitations and these must now be considered.

1) As has already been explained, the theory is an asymptotic one. It has been proven rigorously only for large samples.

2) The procedure is more complicated than the ordinary one and the

choice of class limits which make the class probabilities equal is time consuming, especially since the number of classes given by the formula is quite high. However some time is saved since it is not necessary to compute theoretical frequencies as can be seen by noting formula (2).

The class intervals used for the test are not suitable for visual presentation and another grouping must be made for that purpose.

3) In order to conduct the test, ungrouped data are required since it is necessary to compute the actual class frequencies using class limits other than those given by an ordinary frequency distribution. Furthermore, in most cases the unclassified data must have as many or more significant figures than the class limits in order to decide into which class an observation falls. For a given sample size, as the range of the data increases, this becomes less important.

4) The power of the test for those distributions whose distance from the null hypothesis is less than Δ is not known.

Even more serious in this regard is the question as to whether this "distance" is a useful criterion. It may be more valuable to talk about the power of the test for those *cdf*'s which are similar in some other respect. For example, area between the alternative *cdf*'s and the null hypothesis *cdf* may be more important than distance.

5) The most serious limitation is the fact that the parameters are assumed to be known and the distribution under the null hypothesis must be continuous.

III. EFFECT OF A REDUCTION IN THE NUMBER OF CLASSES

All that has been said heretofore applies to the case where we use the number of classes specified by the formula. Two interesting problems to consider are 1) the effect on the distance when the power is required to be one-half or greater but a smaller number of classes is desired and 2) the effect on the power when the distance is held constant and a smaller number of classes is used.

Such a study has been made with the results listed below. These results apply to the "worst alternative distribution"—the alternative distribution with respect to which the power of the test is a minimum.

1) For a given distance, the power of the test is reduced a relatively small amount by cutting k in half, the reduction of the power becoming smaller as N increases. For example, when $N=1000$ and

$\Delta = .083$, the power drops from .50 to .40 when the number of classes is reduced from 59 to 30.

2) When the power of the worst alternative distribution is equal to one-half, the distance increases slightly when k is cut in half, the increase becoming smaller as N increases. For example, when $N = 1000$, the distance increases from .083 to .089 when the number of classes is cut from 59 to 30.
3) The power function of the test with respect to this worst alternative distribution is very steep, the power decreasing quite rapidly as the distance decreases.

These results indicate that for practical purposes, the number of classes suggested by the formula may be cut in half with a relatively small effect on the power or the distance.

IV. SUMMARY

If the chi-square test of goodness of fit is to be used with ungrouped data, a large number of observations, a continuous distribution and known parameters, the Mann-Wald method may be used. This procedure is the usual one with the exception that the number of classes is given by the formula

$$k = \left[4\sqrt[5]{\frac{2(N-1)^2}{c^2}}\right],$$

and the class limits are chosen such that the theoretical class probabilities equal N/k. This technique enables one to state that the power of the test for a family of cumulative distribution functions a distance Δ or greater away from the null hypothesis *cdf* is for practical purposes greater than or equal to one-half. If it is not required that the power of the test be one-half or more, the statistician may cut the number of classes in half for large sample sizes without greatly affecting the power. If the power of the test is required to be one-half but for a class with a distance greater than Δ away from the null hypothesis distribution, the number of classes may be halved without greatly increasing the distance.