# ON THE RELIABILITY OF THE CLASSICAL CHI-SQUARE TEST

By E. J. Gumbel

*New School for Social Research*

For a given set of observations and for a continuous variate, different classifications lead to different observed distributions and to different values of $\chi^2$. This shortcoming has been vaguely felt by statisticians. We shall explain how these differences arise and show that they are important enough to cast a great deal of doubt on the validity of the application of the usual $\chi^2$ method to a continuous variate. Finally, we propose a procedure which is free from these difficulties.

**1. The observed distributions.** The $\chi^2$ method gives a numerical measure of the differences between the observed and the theoretical distribution. A theoretical distribution is completely determined once the constants are known. For a discontinuous variate the observed distribution is also well defined; but for a continuous variate the concept "observed distribution" is vague. To classify $N$ observations, $x_1, x_2, \cdots x_m, \cdots x_N$ arranged in increasing order, we introduce two arbitrary actions: the choice of the intervals and the beginning of the first cell. As a rule, all cells have the same length, and they are bounded by integral numbers, or even numbers, or round numbers, 0, 5, 10, of the variate. But these classifications and the preference given to round numbers for the starting point have no theoretical foundation.

A certain guide for the systematic choice of the class length and the beginning of the first cell may be found by turning to the theory. Many theoretical distributions of a continuous variate $x$ have only two constants, and permit the introduction of a reduced variate $y$ with the dimension zero, where

$$y = \frac{x - a}{b}. \tag{1}$$

The constant $a$ is a mean, and $b$ is a measure of dispersion. The probabilities $W(x)$ (or $F(y)$) for values equal to or less than $x$ (or $y$) are

$$W(x) = F(y). \tag{2}$$

For most distributions, for which the above transformation is possible, tables for $F(y)$ exist, in which the argument progresses by a fixed interval $\Delta y$. By taking an initial value $y_0$ and a fixed interval $\Delta y$, the differences

$$NF(y_0 + i\Delta y) - NF(y_0 + (i - 1)\Delta y) = Np_i \qquad (i = 1, 2, \cdots k) \tag{3}$$

may be interpreted as being the theoretical distribution. The corresponding values of the variate, by (1), are

$$x(i) = a + b(y_0 + i\Delta y); \qquad x(i - 1) = a + b(y_0 + (i - 1)\Delta y) \tag{4}$$

and the cell length is

$$\Delta(x) = b\Delta y. \tag{5}$$

In (3) $k$ is the number of cells. In general, $x(i)$ and $x(i - 1)$ will not exist among the observed values $x_m$. By arranging the observations in the cells given by the theoretical values (4), we obtain an observed distribution consisting of the contents $a_i$ of the cell $i$. This procedure prescribes a classification of the observations according to the theory. The intervals selected are multiples of some measure of dispersion. In principle, the choice of $\Delta y$ and of the starting point $y_0$ remain arbitrary; in practice, the selection of $\Delta y$ is limited by the intervals given in the probability tables.

This natural classification may be used for constructing *different* observed distributions from the same set of observations. We determine the constants, then choose a small interval and a starting point which is below the smallest observation $x_1$. The last cell is such that it contains the largest observation $x_N$. In this way, we obtain the initial observed distribution, consisting of $k$ cells.

If we combine $h$ cells ($h = 2, 3, \cdots \frac{1}{2}k$), we obtain $h$ different observed distributions: We combine $h - 1$ void cells with the first cell of the initial distribution, we combine the second cell and the following $h - 1$ cells of the initial distribution, and so on. Generally, we combine $q$ void cells ($q = h - 1, h - 2, \cdots 1, 0$) with the first $h - q$ cells of the initial distribution, then the next $h$ cells of the initial distribution, and so on. The last of these $h$ distributions starts with the first $h$ cells of the initial distribution.

If we combine more and more cells, the number of observed distributions, having the same intervals, increases. The larger the intervals the larger is the influence of the starting point, and the more the observed distributions become dissimilar. To see this influence of classification on the shape of the observed distributions, consider the extreme case for a symmetrical theoretical distribution of an unlimited variate. Let the observed distribution consist of two cells. Assume besides that the observed median is close to the theoretical one. If the cut between the cells is identical with the theoretical median, the two cells have the contents $\frac{1}{2}N + \epsilon$ and $\frac{1}{2}N - \epsilon$, where $\epsilon$ is small. If the cut is shifted sufficiently far to the left or right of the median, the cell contents will be 0, $N$ and $N$, 0. These two distributions are completely different.

To each observed distribution corresponds a theoretical one obtained from (3) by the same combination of cells as the observed distribution. In the graphical representation, the same continuous theoretical distribution may be used for all observed distributions by choosing the scale of the ordinate properly. The length chosen for representing one observation in the initial distribution will represent $h$ observations for the $h$ distributions obtained by the combination of $h$ cells.

The different observed distributions corresponding to the same observations and to the same theory will give different values of

$$\chi^2 = \sum_{i=1}^{k} \frac{(a_i - Np_i)^2}{Np_i}. \tag{6}$$

The expected contents of the first and last cell are

$$Np_1 = NF(y_0 + \Delta y), \tag{7}$$

$$Np_k = N(1 - F(y_0 + .(k - 1)\Delta y)). \tag{8}$$

Since the total expected frequency must be equal to the number of observations

$$\sum_{i=1}^{k} Np_i = \sum_{i=1}^{k} a_i\,, \tag{9}$$

formula (6) may be written

$$\chi^2 = \sum_{i=1}^{k} \frac{a_i^2}{Np_i} - N. \tag{10}$$

This formula, being simpler than (6), will be used in the numerical example.

An upper limit for $\chi^2$ is furnished by the case that one cell $j$ contains all observations. Then

$$a_j = N; \qquad a_i = 0 \quad \text{for} \quad i \neq j,$$

whence from (10)

$$0 \leqq \chi^2 \leqq \frac{N}{p_j} - N. \tag{11}$$

The upper limit depends again upon the intervals and the starting point of the classification. If the probability for an observation to be contained in the cell $j$ is small, the upper limit is large.

The exact distribution of $\chi^2$ has not yet been established. To obtain an approximation, it is assumed that a binominal distribution may be replaced by a normal distribution. As this does not hold for cells with a small expected frequency, the contents of such cells must be combined. This prescription, which is also valid for a discontinuous variate, constitutes a third arbitrary action in the calculation of $\chi^2$. It invalidates the prior postulate that all cells ought to have the same length.

The approximation used for the probability $P$ of obtaining a value of $\chi^2$, equal to or larger than the observed one, is

$$P(\chi^2, \nu) = K \int_{\chi^2}^{\infty} z^{2\,\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}z^2}\, dz^2 \tag{12}$$

where $\nu$ is the number of degrees of freedom. Since

$$\frac{\partial P}{\partial \chi^2} < 0; \qquad \frac{\partial P}{\partial \nu} > 0, \tag{13}$$

$P$ diminishes as $\chi^2$ increases, $\nu$ being given, but $P$ increases as $\nu$ increases, $\chi^2$ being given. By choosing larger cells, the number $\nu$ diminishes, and $P$ may remain the same if $\chi^2$ diminishes adequately.

It is easy to see that $\chi^2$ cannot increase as a result of the combination of cells and will, in general, decrease. Let $a_1$ and $a_2$ represent the actual number of observations in two cells that are to be combined. Let $Np_1$ and $Np_2$ be the expected numbers. Then, the contribution of the two separate cells to $\chi^2$ minus the contribution of the two combined cells is, by (10)

$$\frac{a_1^2}{Np_1} + \frac{a_2^2}{Np_2} - \frac{a_1^2 + 2a_1a_2 + a_2^2}{N(p_1 + p_2)}.$$

As $a_1$ and $a_2$ are positive or zero, the difference is proportional to

$$a_1^2p_2^2 + a_2^2p_1^2 - 2a_1a_2p_1p_2 = (a_1p_2 - a_2p_1)^2 \geqq 0.$$

The equality holds only when $a_1:a_2 = p_1:p_2$. Then, the combination of cells has no influence on $\chi^2$, but it reduces the number of degrees of freedom by one, and diminishes the probability $P$. In the general case, the combination of cells diminishes $\chi^2$ and diminishes $\nu$ at the same time. According to (13), the first influence tends to increase the probability $P$, the second to diminish it. It cannot be stated a priori which influence is stronger.

For a given set of observations, a continuous variate and a given theory, which includes given estimates of the constants, the probability $P$ depends upon three arbitrary actions. If a certain choice of the intervals gives a good fit, it cannot be concluded that a broader classification gives the same or a better fit [4]. For a given interval, $P$ may vary considerably with the starting point. This influence cannot be allowed for by any formula as the number of degrees of freedom does not depend upon the starting point. Finally, the term "small expected numbers" is vague. Different combinations of cells lead to different probabilities. It is generally assumed that these influences remain within reasonable limits and that $P$ does not vary considerably if we change the class length or the starting point. In the following example, we shall show that this opinion is erroneous.

**2. Numerical example.** The flood discharge of the Mississippi River at Vicksburg for each of the fifty years 1890–1939 will be used to illustrate the extent to which the observed distributions and $P$ vary with the choice of cell length and the starting point. The observed flood discharges $x_m$ measured in 1,000 cubic feet per second are given in Table VI of a previous article [2], and are not repeated here. The expected distribution is given by the theory of largest values which states that the probability $\mathfrak{W}(x)$ of a flood discharge equal to or less than $x$ is

$$(14) \qquad \mathfrak{W}(x) = e^{-e^{-\alpha(x-u)}}$$

Values of $\mathfrak{W}(x)$ as a function of the reduced variate

$$(15) \qquad y = \alpha(x - u),$$

are given in Table II of the reference first cited.

Calculation of the constants $\alpha$ and $u$ leads to the theoretical value of the flood discharge

$$x = 1201.9 + 266.1y \tag{16}$$

associated with a given probability $F(y) = \mathfrak{W}(x)$.

TABLE I

*Observed and theoretical distribution (1) for the interval $\Delta y$ = .25; $\Delta x$ = 66.525*

| Variates | | Distributions | |
|---|---|---|---|
| Reduced<br>$y$ | Absolute<br>$x$ | Observed<br>$a_i$ | Theoretical<br>$Np_i$ |
| 1 | 2 | 3 | 4 |
| | 736.2 | 1 | .5655 |
| ≦ −1.50 | 802.8 | 1 | .959 |
| −1.25 | 869.3 | 3 | 1.775 |
| −1.00 | 935.8 | 3 | 2.720 |
| −.75 | 1002.3 | 5 | 3.5955 |
| −.50 | 1068.9 | 1 | 4.2315 |
| −.25 | 1135.4 | 3 | 4.5475 |
| .00 | 1201.9 | 3 | 4.554 |
| .25 | 1268.4 | 3 | 4.314 |
| .50 | 1334.9 | 6 | 3.914 |
| .75 | 1401.5 | 6 | 3.434 |
| 1.00 | 1468.0 | 4 | 2.934 |
| 1.25 | 1534.6 | 2 | 2.4565 |
| 1.50 | 1601.1 | 0 | 2.0235 |
| 1.75 | 1667.6 | 2 | 1.647 |
| 2.00 | 1734.1 | 0 | 1.3270 |
| 2.25 | 1800.6 | 2 | 1.0615 |
| 2.50 | 1867.2 | 2 | .844 |
| 2.75 | 1933.7 | 0 | .668 |
| 3.00 | 2000.2 | 2 | .527 |
| 3.25 | 2066.7 | 0 | .414 |
| 3.50 | 2133.3 | 0 | .325 |
| 3.75 | 2199.8 | 0 | .255 |
| 4.00 | 2266.3 | 0 | .1995 |
| ≧4.25 | 2332.8 | 1 | .708 |
| | | 50 | 50.000 |

The first observed distribution presented in Table I is obtained by letting $\Delta y = .25$; $\Delta x = 66.525$ and $y_0 = -1.75$. The expected number of observations for the first and last cell are $50F(-1.5)$ and $50\ (1 - F\ (4.25))$ respectively.

The expected frequencies (formula 4) for the other cells

$$np_i = 50\,[F(y + .25) - F(y)],$$

were obtained by successive substraction of two consecutive figures given in column 2, Table II [2]. The theoretical and the observed distribution are plotted in figure 1. The observed distribution given in Table I is very irregular. Evidently, the intervals are too small. Therefore, we construct the observed and theoretical distributions (2) and (3) for cells which are two times larger.

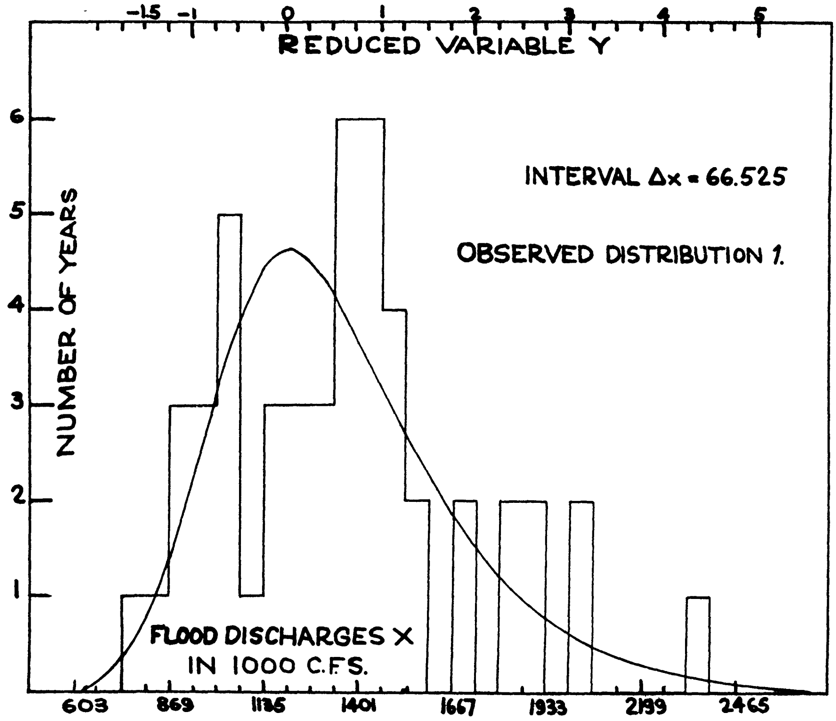

Fig. 1

The first cell in distribution (2) is obtained from distribution (1) by combining the first cell of (1) with the empty one before it; the second cell is obtained by combining the second and third cells of (1); and so on.

Distribution No. 3 is obtained by combining the first two cells of distribution No. 1, then the third and fourth, and so on. The observed distributions 2 and 3 and the theoretical distribution are plotted in figure 2. The scale of the ordinate is $\frac{1}{2}$ of the scale in figure 1. In the same way, the three observed distributions (4), (5), (6) for the interval $\Delta y = \frac{3}{4}$, $\Delta x = 199.57$ are obtained by combining either two void cells with the first cell of Table I, or one void cell with the first and second cell of Table III, or the first three cells of Table I (see fig. 3).

Finally, the four observed distributions (7), (8), (9), (10) for the interval

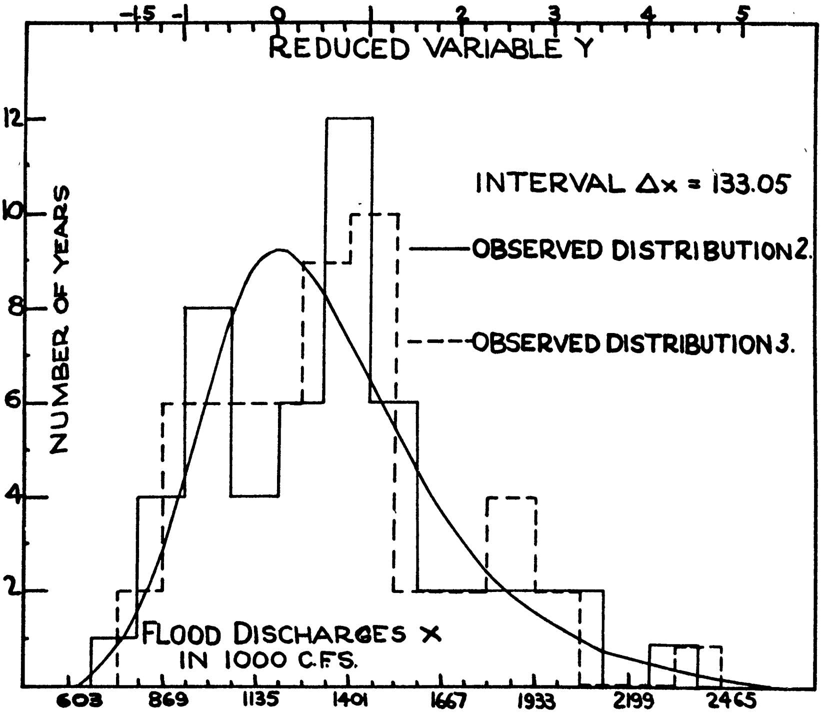
-1.5
-1
0
1
2
3
4
5
REDUCED VARIABLE Y
12
10
8
6
4
2
NUMBER OF YEARS
INTERVAL Δx = 133.05
OBSERVED DISTRIBUTION 2.
OBSERVED DISTRIBUTION 3.
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

Fig. 2

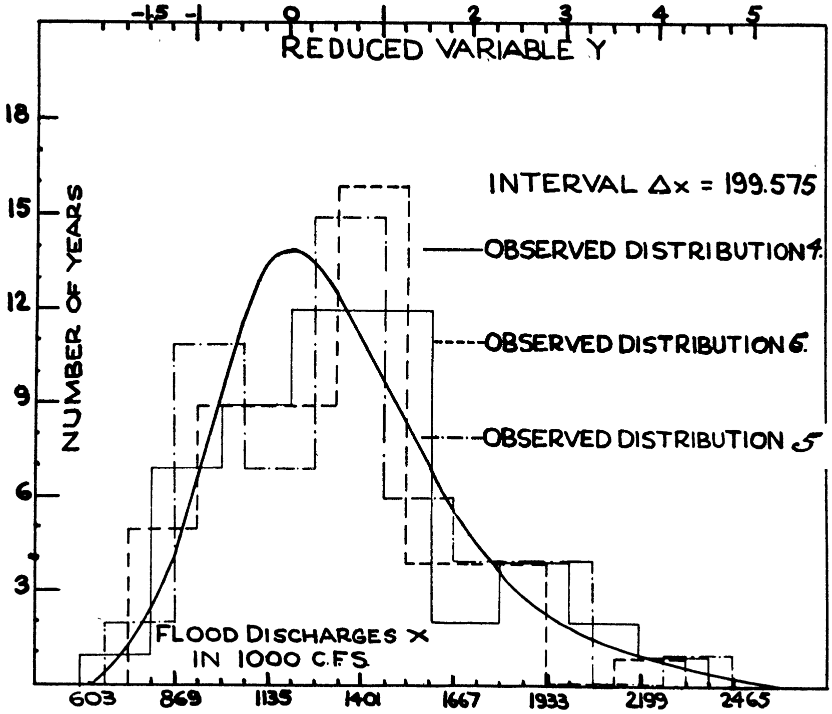
-1.5
-1
0
1
2
3
4
5
REDUCED VARIABLE Y
18
15
12
9
6
3
NUMBER OF YEARS
INTERVAL Δx = 199.575
OBSERVED DISTRIBUTION 4.
OBSERVED DISTRIBUTION 6.
OBSERVED DISTRIBUTION 5
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

Fig. 3

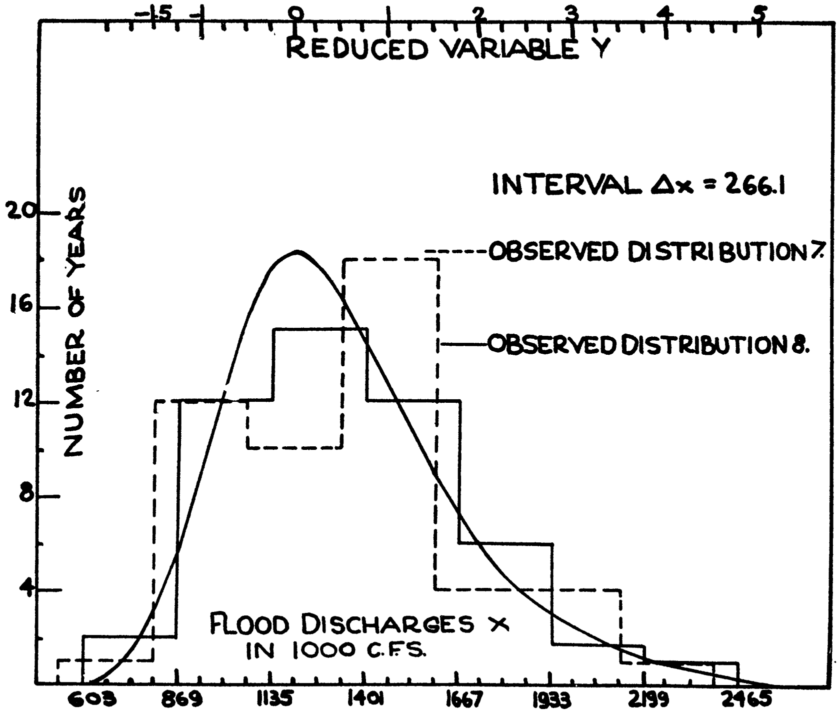
-1.5
-1
0
1
2
3
4
5
REDUCED VARIABLE Y
INTERVAL Δx = 266.1
OBSERVED DISTRIBUTION 7.
OBSERVED DISTRIBUTION 8.
NUMBER OF YEARS
20
16
12
8
4
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

Fig. 4

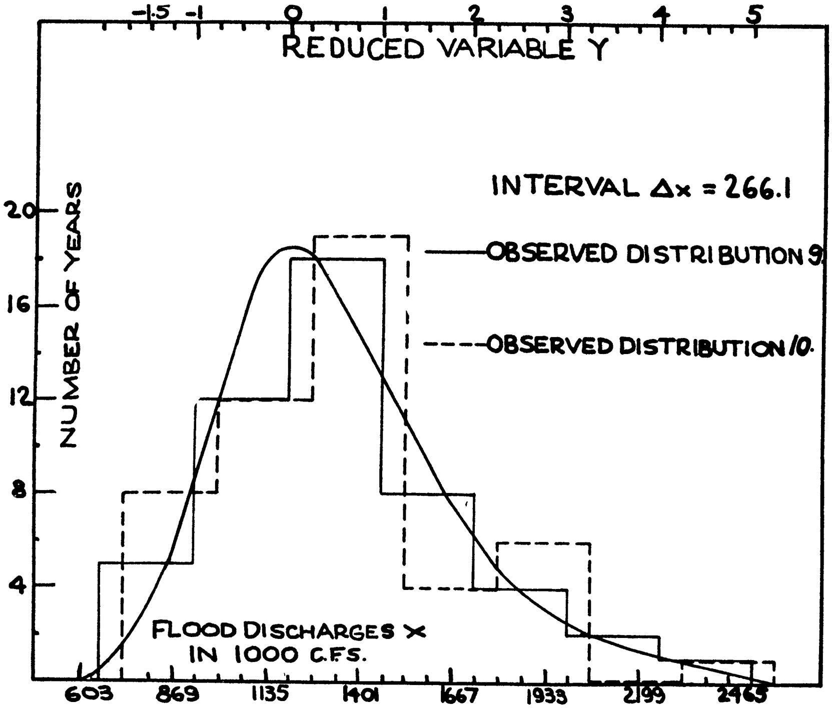
-1.5
-1
0
1
2
3
4
5
REDUCED VARIABLE Y
INTERVAL Δx = 266.1
OBSERVED DISTRIBUTION 9.
OBSERVED DISTRIBUTION 10.
NUMBER OF YEARS
20
16
12
8
4
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

Fig. 5

$\Delta y = 1$; $\Delta x = 266.1$ are compared with the theoretical distribution in figures 4 and 5. The four distributions 7–10 differ considerably. Distributions 8 and 9 indicate that the agreement between theory and observations is good, distribution 7 and 10 indicate that the fit is bad. The $\chi^2$ method must give the same contradictory results.

TABLE II

*Four values of $P(\chi^2)$ for the same observations and the same theory*

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| Mid-points | Observed Distributions, $a_i$ | | | | Theoretical Distributions, $Np_i$ | Components of $\chi^2 + N$ | | | |
| | (7) | (10) | (9) | (8) | | | | | |
| 803 | | | 5 | | 3.2995 | | | 7.577 | |
| 869 | | 8 | | | 6.0195 | | 10.632 | | |
| 936 | 13 | | | | 9.6150 | 17.577 | | | |
| 1002 | | | | 14 | 13.8465 | | | | 14.155 |
| 1069 | | | 12 | | 15.0945 | | | 9.540 | |
| 1135 | | 12 | | | 16.9285 | | 8.506 | | |
| 1202 | 10 | | | | 17.6470 | 5.667 | | | |
| 1268 | | | | 15 | 17.3295 | | | | 12.984 |
| 1335 | | | 18 | | 16.2160 | | | 19.980 | |
| 1401 | | 19 | | | 14.5960 | | 24.733 | | |
| 1468 | 18 | | | | 12.7385 | 25.435 | | | |
| 1534 | | | | 12 | 10.8480 | | | | 13.274 |
| 1601 | | | 8 | | 9.0610 | | | 7.063 | |
| 1667 | | 4 | | | 7.4540 | | 2.146 | | |
| 1734 | 4 | | | | 6.0590 | 2.641 | | | |
| 1800 | | | | 6 | 4.8795 | | | | 7.378 |
| 1867 | | | 7 | | 6.3290 | | | 7.742 | |
| 1933 | | 7 | | | 5.0020 | | 9.796 | | |
| 2000 | 5 | | | | 2.9405 | 6.344 | | | |
| 2066 | | | | 3 | 3.0965 | | | | 2.907 |
| $N$..... | 50 | 50 | 50 | 50 | 200.0000 | $\chi^2 + N = 57.664$ | 55.813 | 51.902 | 50.698 |
| $\nu$...... | 2 | 2 | 2 | 2 | $P$....... | .023 | .057 | .399 | .705 |

The details for the calculations of $\chi^2$ are given in Table II. The numbers of column 1 are the midpoints of the cells. To save space, the four theoretical distributions obtained from Table I, col. 4 are written in the same column (6) directly opposite the corresponding observed distributions given in columns 2 to 5. Through formula (10) we calculate the components of $\chi^2 + N$ (cols. 7 to 10). Although the four distributions differ only with respect to the beginning of the first cell, the value of $P$ for the observed distribution number (8) is more than thirty times the value of $P$ for the observed distribution number (7). In view of the fact that these values of $P$ are calculated for a fixed set of observations, for the same theory, the same constants, and the same number of degrees of freedom, the differences found are surprising.

**3. The probability integral transformation.** This example shows that the probability $P$ may vary with the starting point in such a way that no conclusion about the acceptance or rejection of a hypothesis can be obtained from the usual $\chi^2$ method. The three arbitrary steps described above may be avoided if we choose cells of equal probability instead of cells of equal length. The required intervals are obtained from the probability integral transformation, due to Karl Pearson [6]. Let $w(x)$ be a distribution of a continuous variate $x$, let $y = W(x)$ be the transformed variate, then the distribution $p(y)$ of the variate $y$ is

$$p(y) = 1. \tag{17}$$

In other words: The probabilities $W(x)$ are uniformly distributed. If a distribution $w(x)$ has been chosen for a given set of observations $x_m$, we can control this theory by investigating whether the "observations" $W(x_m)$, i.e., the theoretical cumulative frequencies of the observed values are uniformly distributed. Thus, the comparison of the observed distributions with any continuous theoretical distribution is reduced to the comparison of an "observed" with a theoretical uniform distribution. To a given set of observations and a given theory there is one, and only one, "observed" distribution. If we introduce within $w(x)$ another set of constants, or choose instead of $w(x)$ another theory $\varphi(x)$, we obtain, of course, other "observed" values [1].

The goodness of fit between this theory and these "observations" may be measured by the $\chi^2$ method. We divide the interval zero to $N$, which contains the $N$ "observed" numbers $NW(x_m)$ into $k$ cells of equal length, and enumerate the "observed" points $NW(x_m)$ contained in each cell. The starting point of the classification is always zero. The expected number of observations for each cell is always $N/k$. If we choose $k$ sufficiently small, the necessity for combining cells is eliminated. We have to choose $k$ in such a way that the conditions, under which formula (12) holds, are fulfilled. The question of the best choice for the number of cells has been studied by Wald and Mann [3]. Their solution is valid for small levels of significance and for large numbers of observations.

**4. Conclusion.** The usual $\chi^2$ test is unreliable for a continuous variate as it involves three arbitrary decisions. From the same observations, the same theory, and the same constants different statisticians, equally well trained and equally careful, may obtain different probabilities $P$, and may proclaim any one of these results as final. Therefore, the usual $\chi^2$ method does not lead to a decision whether a hypothesis has to be rejected or not. Such a decision is possible if we use the probability integral transformation. Unfortunately, the question of the best choice of the cells for small numbers of observations and large levels of significance is not yet solved.

## REFERENCES

[1] E. J. Gumbel, "Simple tests for given hypotheses," *Biometrika*, Vol. 32, Parts 3 and 4 (1942), pp. 317–333.

[2] E. J. Gumbel, "The return period of flood flows," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 163–190.

[3] H. B. Mann and A. Wald, "On the choice of the number of class intervals in the application of the chi square test," *Annals of Math. Stat.*, Vol. 13 (1942), p. 306–317.

[4] J. Neyman and E. S. Pearson, "Further notes on the $\chi^2$ distribution," *Biometrika*, Vol. 22, Parts 3 and 4 (1931), pp. 298–305.

[5] Karl Pearson, "On a method of determining whether a sample of size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random," *Biometrika*, Vol. 25, Parts III and IV (1933), pp. 379–410.