JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 294 JUNE, 1961 *Volume 56*

SIGNIFICANCE TESTS IN DISCRETE DISTRIBUTIONS

H. O. LANCASTER*
University of Sydney

In discrete distributions, it has often been recommended that an auxiliary random experiment should be carried out so as to make the size of the test equal to the significance level. In certain experimental situations, this procedure would be time-consuming and even embarrassing to the statistician. An alternative criterion of significance is suggested, the mid or median probability. To avoid computations $P(\chi)$, the probability of the square root of a chi-square variable, can be used in the binomial, Poisson and hypergeometric distributions as an approximation to the mid probability. In statistical control of counting experiments or where experiments are being repeated $P(\chi)$ rather than $P(\chi_c)$, the probability of χ corrected for continuity, gives acceptable approximations to size of significance levels. Some computations are given for the multinomial distribution which show that here $P(\chi^2)$ gives acceptable approximations.

1. INTRODUCTION

IN DISCRETE distributions, the cumulative sum of the probability of the observation itself and all more extreme has been used as the test function in the "exact test." If this "exact probability" is less than the level of significance, the observation is said to be significant or lie in the critical region. As a rule, there is a difference between the size of such a test and the significance level. Fisher [6] and elsewhere has maintained that the size must not exceed the significance level, whereas the Neyman-Pearson procedure is to equalize size and significance level by an auxiliary random experiment, for a description of which we may refer the reader to the review of Pearson [10]. In Section 4 we show that there are serious difficulties to be faced if this test is applied in practice. As an alternative procedure we introduce the mid- or median probability as a test function and show that it is plausible to consider it as the result of a randomization procedure carried out before the experiment. Its use can thus be reconciled with the Neyman-Pearson theory. In any of these three methods, there is a good deal of computation. In the commonly met discrete distributions, this can be avoided by the use of χ or χ^2 approximation. The exact probability corresponds closely with $P(\chi_c)$, the probability of χ, corrected for continuity. The mid-probability is closely approximated by $P(\chi)$ as a rule. We consider also the multinomials of the form, $(0.2+0.3+0.5)^N$, and show that $P(\chi^2)$ gives a reasonable approximation of size to the significance levels.

* Formerly Associate Professor in Medical Statistics at the School of Public Health and Tropical Medicine, University of Sydney, Australia.

Finally we lay down rules for criteria of significance and conclude that in many common situations, $P(\chi)$ or $P(\chi^2)$ have desirable features as a test function.

2. DISCRETE DISTRIBUTIONS

In discrete distributions, a random variable takes on finitely or denumerably many values $\{a_i\}$ with probabilities $\{p_i\}$, $i=0, 1, 2 \cdots ; \sum p_i=1$. There is no loss of generality in considering some transformation so that the random variable takes only non-negative integral values and can be written as i. i takes the values $0, 1, 2 \cdots n$ in the binomial and hypergeometric distributions but denumerably many integral values in the Poisson. In such cases the distribution is usually modified by taking the union of all except a finite number of events as a single event. We suppose that this has been done. Formally, we are dealing with one-sided tests. We define

$$P(i) = \sum_{j \geq i} p_j, \qquad P'(i) = P(i+1) = \sum_{j>i} p_j. \tag{2.1}$$

So defined $P(i)$ is the probability of the exact test and may be referred to as the exact probability. If there are $(n+1)$ values of i, then $P(0)=1$, $P_n=p_n$ and $P'(n)=0$. Let α be an arbitrary number, $0<\alpha<1$, chosen as the level of significance. Then the critical region for the exact test consists of all those values of i for which $P(i) \leq \alpha$. There is always a special or marginal value of i, I say, for which the inequality holds, $P(I) \geqq \alpha > P(I+1)$. For this marginal event, $i=I$, we define

$$\begin{aligned} \theta &= \{\alpha - P(I+1)\}/\{P(I) - P(I+1)\} \\ &= \{\alpha - P(I+1)\}/p_I. \end{aligned} \tag{2.2}$$

For a fixed distribution, for example, with $\{p_i\}$ given by the terms of the binomial

$$p_i = b(i \mid n, q) = \binom{n}{i} q^i (1-q)^{n-i}, \tag{2.3}$$

θ is not a random variable but with a mixture of binomials where n or q or both are random variables, θ will be a random variable. Now no *à priori* distributions can be postulated for n and q but a number of cases may be examined to see whether there is any tendency for θ to have special values. To this end, a fixed value of α, namely, 0.05 has been chosen; and we have given equal weight to the ten values of $n=40(1)49$ and to the forty values of $q=0.31(0.01).069$ with 0.50 counted twice. The results are detailed in Table I. The χ^2 of goodness of fit is 26.7 with 19 degrees of freedom and so it is a plausible guess that in discrete distributions treated by the statisticians the set of θ's will behave as though θ were rectangular in the interval (0, 1). In some physical contexts, there will be a randomising process; in taking parallel counts of blood cells or bacterial colonies, the set total

$$N = x_1 + x_2 + \cdots x_k, \tag{2.4}$$

will be a random variable, under the null hypothesis a Poisson variable with parameter, $k\lambda$, where λ is the parameter of the Poisson distribution for an individual count. Similar considerations hold for the fourfold tables.

TABLE I. THE VALUES OF θ IN 400 BINOMIAL DISTRIBUTIONS WITH THE SIGNIFICANCE LEVEL, $\alpha = 0.05$

Value of θ	Frequency	Value of θ	Frequency
0.00—	16	0.50—	19
0.05—	18	0.55—	22
0.10—	12	0.60—	15
0.15—	16	0.65—	24
0.20—	15	0.70—	20
0.25—	22	0.75—	24
0.30—	28	0.80—	34
0.35—	13	0.85—	21
0.40—	21	0.90—	24
0.45—	16	0.95—	20

"0.05 —" means $0.05 \leq \theta < 0.10$.
$p = 0.31(0.01)69$, 0.50 counted twice
$N = 40(1)49$.

3. THE SIZE OF THE TESTS IN DISCRETE DISTRIBUTION

The size of the exact probability test is never greater than the significance level, by definition. If a table of the binomial distributions is examined, it is easily verified that this results in great loss of power. In fact, if the null hypothesis specifies the value of the binomial index and the value of q, the power of the test may be less than α for moderate values of the difference between q of the null distribution and q' of the alternative.

Example

(i) $\alpha = 0.05, \quad n = 20, \quad q = 0.50; \quad P(15) = 0.0207, \quad P(14) = 0.0577$

Power of the test for $q' = 0.54$ is 0.0461.

(ii) $\alpha = 0.05, \quad n = 22, \quad q = 0.50; \quad P(16) = 0.0262, \quad P(15) = 0.0669$

Power of the test for $q' = 0.53$ is 0.0486.

Most authors are agreed that in the null case the (effective) rejection rate or size of the test should be equal to the (nominal) significance level, when the distribution is continuous. For discontinuous distributions, there is not the same degree of agreement. Fisher [6], indeed, criticizes authors for "laying down axiomatically, what is not agreed or generally true, that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. This intrusive axiom . . . seems to be a real bar to progress." On the other hand, even if this axiom is not admitted, it is essential to have a test which will assign the correct proportions of sets to the various probability classes in the null case, for example, when one is using the technique of statistical control of counting experiments, (introduced by Fisher, Thornton and Mackenzie [7]). The most convenient form for routine use is given in their Table 19. They are considering three parallel counts of bacterial colonies, x_1, x_2 and x_3. Under ideal conditions, x_1, x_2 and x_3 can be considered to be a sample of three independent drawings from a Poisson distribution with unknown parameter, λ. If only

variation between x_1, x_2 and x_3 for a fixed sample total, or the joint distribution of x_1, x_2 and x_3 conditional on $x_1+x_2+x_3=N$, is considered, then Fisher, Thornton and Mackenzie [7] show that the test criterion,

$$\chi^2 = \sum_j (x_j - \bar{x})^2/\bar{x}, \qquad n\bar{x} = \sum_1^n x_j, \tag{3.1}$$

is distributed approximately as χ^2 with 2 d.f. $P(\chi^2)$ is then used to classify the counts into probability classes, with points of division corresponding to convenient significance levels of χ^2, say 0.1(0.2)0.9. Their test of consistency between counts finally resolves itself into a test of goodness of fit of the observed assignments in the probability classes to the theoretical. The justification of this procedure can only be that the discontinuous χ^2 approximates in distribution to that of the continuous χ^2 and that a $100\alpha\%$ significance test will reject $100\alpha\%$ of sets and this argument is used by Fisher, Thornton and Mackenzie [7] on their pages 335 and succeeding.

4. THE METHOD OF THE AUXILIARY RANDOM EXPERIMENT

The possibility of having an experimental outcome, $i=I$, where $P(I)>\alpha>P'(I)=P(I+1)$ is considered in the Neyman-Pearson theory and an auxiliary randomisation procedure is recommended to make the size of the test equal to α. An auxiliary sampling experiment is carried out after the experiment has been performed using random sampling numbers and a proportion,

$$\{\alpha - P'(I)\}/\{P(I) - P'(I)\}$$

of the event, $i=I$, rejected. This method gives a size equal to the significance level, but at a price. Given the same experimental data, the statistician will not always give the same result even using a test of his own choice. Another form of inconsistency may be considered. Let the null hypothesis specify that $q=0.8$ and $N=20$ in the binomial, $\{q+(1-q)\}^N$. $P(20)=0.01153$ and P(19) $=0.06918$. An auxiliary random experiment is carried out and the event,$i=19$, is found to be significant at the significance level, $\alpha=0.05$. In the same report, there might be an experiment with $N=21$, $q=0.8$ and $i=20$. $P(21)=0.00922$ and $P(20)=0.05765$ and the auxiliary random experiment decides that $i=20$ is not significantly different from the expected. Together these two results mean that the statistician has decided that $19/20=0.950$ is significantly greater than 0.8 but that $20/21=0.953$ is not. Comparing the experiments, we find that the difference between them was a single additional observation which was in favour of the hypothesis being rejected but which had the opposite effect.

The computations for the auxiliary randomization would be found to be rather irksome. For example, a bacteriologist may be carrying out two parallel plate counts at a technically convenient density, say a mean number of 25 colonies per plate. He is using the boundaries 0.1(0.2)0.9 for his probability classes in the manner explained in Section 3. With an observed combined count of 35, only a difference of 5 or of more than 11 will not need to have an auxiliary random experiment so that the auxiliary experiment will be carried out in 79.4% of counts under a true null hypothesis. With a difference $|x_1-x_2|=3$, the auxiliary random experiment may refer the observation to any one of the

classes, 0.9 to 1.0, 0.7 to 0.9 or 0.5 to 0.7. With a combined count of 36, only differences $|x_1-x_2|$ of 6 or of more than 12 will escape the need for a random experiment, which will be needed in 80.9% of counts. With a combined count of 37, 90.1% of counts will require a random experiment. Suppose that a bacteriologist carrying out statistical control were exhibiting such results. An intelligent biologist present might ask, in what proportion of cases is the assignment to a probability class due to the auxiliary random experiment. It will surely weaken the statistician's position if he is compelled to admit that the assignment was due to the auxiliary random experiment in over 60% of all sets.

5. THE MEDIAN PROBABILITY AS A TEST FUNCTION

The median probability, or perhaps more appropriately the mid-probability, defined in Lancaster [8], by

$$P_m(i) = \tfrac{1}{2}\{P(i) + P(i+1)\} = \tfrac{1}{2}\{P(i) + P'(i)\} \tag{5.1}$$

may be considered as a test function for single experiments, with a rule of rejection.

$$P_m(i) \leq \alpha, \tag{5.2}$$

whereas the rule of rejection with the exact probability is to reject when $P(i) \leq \alpha$. For a given hypothesis and experimental result it will always give the same answer. A comparison of (5.1) and (2.2) shows that the rule of rejection (5.2) applied to the marginal event, I, is equivalent to a rule of rejection, when $\theta \geq \frac{1}{2}$.

Let a test be defined for the marginal event so that the null hypothesis is rejected when $\theta > \theta_0$ and let us determine how often it would agree with the auxiliary random sampling method in a mixture of populations such that θ is a rectangularly distributed variable in the range, zero to unity. The agreement will occur in a proportion,

$$\int_0^{\theta_0} (1-\theta)d\theta + \int_{\theta_0}^{1} \theta d\theta = 0.75 - (0.5-\theta_0)^2, \tag{5.3}$$

under these idealised conditions. For with $\theta \leq \theta_0$, the test will accept the null hypothesis and the auxiliary sampling method will accept in a proportion $(1-\theta)$. Similarly for $\theta > \theta_0$, the auxiliary sampling method test will reject the null hypothesis in a proportion, θ. So that, integrating over the postulated rectangular distribution of θ, the two tests would agree in about 75% of cases. Of all such tests with choice of θ_0, the median probability obtained by taking θ_0 to be 0.5 will agree most often with the auxiliary random sampling method.

Wallis [11], on his page 245, uses, in effect, the geometric mean of $P(I)$ and $P(I+1)$ as his criterion, which procedure also allows the size of the test to exceed the significance level and gives a rejection rate higher than that given by the median probability, since the geometric mean is less than the arithmetic, that is,

$$\sqrt{\{P(I)P(I+1)\}} < \tfrac{1}{2}\{P(I) + P(I+1)\}. \tag{5.4}$$

The results of the two methods will usually agree in marginal case.

Both the random auxiliary experiment and the median probability tests may require rather lengthy computations; however, $P(\chi^2)$ or $P(\chi)$ is shown in the next section to be a good approximation to the median probability.

The conclusion of this section is that in cases likely to be met by the statistician, the mid-probability test will agree with the Neyman-Pearson procedure in about 75% of the marginal cases and in all other cases.

6. NORMAL APPROXIMATIONS TO THE PROBABILITIES

Of the discrete distributions arising in practice, the binomial, the Poisson and the hypergeometric are the most commonly met. The approximations are now considered in detail for the binomial but the treatment is applicable to the other two. The possible events are $i=0, 1 \cdots n$. By Stirling's approximation to the factorial or by other means, there is obtained

$$p_i \simeq \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left\{-\tfrac{1}{2}(i - nq)^2/\sigma^2\right\}, \qquad i = 0, 1, 2, \cdots n, \tag{6.1}$$

where $\sigma^2=nq(1-q)$. It seems natural to equate these values to the areas in a histogram, where rectangles of the heights, p_i, given in (6.1) are erected on the base $(i-\frac{1}{2}, i+\frac{1}{2})$. The rectangular areas are then approximately equal to the area under the normal curve with mean, nq, and variance, σ^2. Yates [12] noted that the correct area under the normal curve to be equated to the exact probability, $P(i)$, would correspond to the area from $+\infty$ to $i-\frac{1}{2}$ and not to i. Yates [12] wrote

$$\chi_c = (i - \tfrac{1}{2} - nq)/\sigma. \tag{6.2}$$

The corresponding probability, $P(\chi_c)$, corresponds closely to the exact probability in any reasonable cases examined. We may write $\chi_c(i)$ for the value of χ_c at the point i and note that

$$\begin{aligned} P(\chi_c(i)) &\simeq P(i) \\ P(\chi_c(i+1)) &\simeq P(i+1). \end{aligned} \tag{6.3}$$

The crude or uncorrected χ may be defined by

$$\chi(i) = (i - nq)/\sqrt{\{nq(1-q)\}}. \tag{6.4}$$

It is evident that

$$\chi(i) = \tfrac{1}{2}\{\chi_c(i) + \chi_c(i+1)\}. \tag{6.5}$$

To what probability does $P(\chi(i))$ approximate? It seems reasonable to take a simple linear interpolation and write

$$\begin{aligned} P(\chi(i)) &\simeq \tfrac{1}{2}\{P(\chi_c(i)) + P(\chi_c(i+1))\} \\ &= \tfrac{1}{2}\{P(i) + P(i+1)\} = P_m(i). \end{aligned} \tag{6.6}$$

The $P(\chi(i))$ can thus be expected to be not greatly different from the median probability. To see this verified empirically let us find the size of the critical regions defined by these two tests for the two-sided test on the binomials, $(\frac{1}{2}+\frac{1}{2})^N$, $N=26(1)120$. At the 1% level, they agree in every case and the aver-

age size is 0.96%. At the 5% level they disagree for $N=26$, 31, 74, 84 and 95. The average sizes are 5.02% and 4.84% for the uncorrected $P(\chi)$ test and the median probability test, respectively. At the 10% level, they disagree for $N=72$, 83 and 94. The average sizes are 9.99% and 9.85%. At the 30% level, there is disagreement for $N=44$ and 74. The average sizes are 30.08% and 29.82%. At 50% there is one disagreement for $N=106$ and the average sizes are 50.10% and 49.97%. At the 70% level there is a disagreement for $N=26$ and the average sizes are 69.77% and 69.47%. At the 90% level there is one disagreement for $N=63$, and the average sizes are 90.06% and 89.85%. In all we have 570 tests with 13 disagreements. This is only a probable result if the $P(\chi)$ is usually close to the median probability.

Another test of the use of $P(\chi)$ to approximate to the median probability is to see how often these tests disagree for the marginal event at the 5% level in the binomials with $q=0.20(0.01)0.50$ and $n=40(1)49$ a total of 620 significance tests. In 40 cases out of the 620, the $P(\chi)$ test rejected the hypothesis and the median probability did not—a disagreement in 6.45% of the marginal cases.

It can be seen that there will be such disagreement because the interpolation between $\chi_c(i)$ and $\chi_c(i+1)$ will not be strictly linear. Average values for the doubtful event, I, in this series might be taken as an n of 45 and a q of 0.36 and so an average value of $\{P(I)-P(I+1)\}=p_I$ would be

$$\{2\pi nq(1-q)\}^{-1/2}\exp -\tfrac{1}{2}(1.645)^2 = 0.10311/3.200 = 0.0322.$$

In other words an average sort of finding is $P(I)=0.0661$, $P(I+1)=0.0339$. The corresponding normal deviates are 1.506 and 1.825, the mid-point of this range, corresponding to a mean-value of χ is 1.666 and $P(1.666)$ is 0.0486. $P(\chi)$ will be lower than the median probability. The $P(\chi)$ test will correspond to a value θ_0 of $(0.0486-0.0339)/0.0322=0.457$. The median probability will reject the event, I, in about 50% of cases, the $P(\chi)$ test will do so in about 54.3% of cases. There will be disagreement in about 4.3% of cases. On the other hand, the $P(\chi)$ test will agree in over 74% of cases with the auxiliary random sampling experiment by formula (5.3) and so would be justified alternatively as giving a greater agreement with the Neyman-Pearson procedure than any other fixed rule of the form, reject if $\theta>\theta_0$. It should be noted that at the 5% level there is only disagreement in say 4% of the marginal cases with I and with all other values of i, there will be agreement.

The good approximation of $P(\chi)$ to the median probability is in accord with intutition. The approximation (6.6) does not lead to a definite statement on direction. The observations in the binomial may be ordered by numbers of successes or by number of failures. It is quite unreasonable to suppose that $P(\chi)$ would approximate to the exact probability because if it did so for successes it certainly would not give a good approximation for failures. The notion that it should approximate to the median probability introduces symmetry into the considerations.

If it be accepted that the best test criterion is the median probability, the conclusions of this section lead us to assert that the most satisfactory practical test of significance in the binomial, Poisson and hypergeometric distributions is the crude χ. The crude χ always attaches the same probability to the same

experimental data, it is easy to compute and gives approximately the correct size for the critical region. It should be stressed, however, that this is only so if the particular distribution given by the $\{p_i\}$ can be regarded as chosen randomly from an appropriate class. These conditions are often fulfilled in a natural sort of way in experimental work; in control of counting, N, the total of counts within the sample, is a random variable from a Poisson, whose parameter is again a random variable. For a fourfold table, attention may be restricted to the case where the choice of rows has been made arbitrarily. The first column total is then a random variable and so a number of hypergeometric distributions are generated. $a_{.1}$ is a binomial variable in fact with unknown parameter, q, and index, $a = a_{1.} + a_{2.}$. The distribution of χ is then a mixture of the distributions,

$$F_z(\chi) = F(\chi \mid a_{i.}, a_{.1} = z), \text{ namely} \tag{6.7}$$

$$F(\chi) = \sum_z P(z \mid a, q) F_z(\chi). \tag{6.8}$$

At any given level, α, for a fixed z, the size of the critical region may be too large or too small but the averaging out effect of (6.8) will make the size acceptably close to the theoretical. This will be even more the case if the model used does not fix $a_{1.}$ and $a_{2.}$ before the experiment; with this model, a second summation over possible row totals would be introduced into (6.8). This natural form of randomisation may not be available. For example, the q of the binomial may be given by hypothesis; a convenient range for N may then be chosen and a random sampling experiment carried out to determine which N to use. Although this randomization appears artificial, it may be regarded as not inconsistent with the argument of pages 52 and 53 of Fisher [5].

7. χ^2 IN THE MULTINOMIALS, $(0.5+0.3+0.2)^N$

These multinomials have been used by previous investigators because of the relative ease of computation, for the three numbers provide the only solution of the partition of unity into three pairwise different one-place decimals, no one of which is too small. The outcome of a multinomial experiment is an ordered set of numbers, or configuration. The relative frequency of each configuration is a term of the multinomial. The terms are most conveniently calculated as a product of terms from the binomials $(0.5+0.5)^N$ and $(0.6+0.4)^{x_2+x_3}$, the values of which are tabulated in Eisenhart [3]. The x_i can be plotted on a triangular grid as homogeneous co-ordinates and the probabilities written alongside the points representing the configurations, (x_1, x_2, x_3). Noting also that

$$3N\chi^2 = 6x_1^2 + 10x_2^2 + 15x_3^2 - 3N^2, \tag{7.1}$$

the contours for the different percentage points of χ^2 can be marked out and the percentage assignment to each probability class obtained by summation in the different regions. This percentage is the calculated or effective assignment. Previous authors, El Shanawany [4] and Neyman and Pearson [9] have compared the assignments using χ^2 with the results of the exact test the events being ordered by the probability and not by their value of χ^2. The procedure

TABLE II. SIZES OF THE χ^2 TEST IN THE MULTINOMIAL DISTRIBUTIONS, $(0.2+0.3+0.5)^N$

N	The Percentage Rejections of the True Null Hypothesis, with a Nominal Significance Level of—						
	1%	5%	10%	30%	50%	70%	90%
9	0.82	4.22	8.02	33.66	45.71	74.49	91.50
10	0.96	5.02	9.03	31.34	53.22	65.27	91.50
11	0.86	3.68	9.67	29.23	47.88	78.95	92.20
12	0.69	4.47	10.20	31.81	48.61	74.45	92.98
13	1.38	5.28	8.37	27.70	47.03	70.71	87.40
14	0.65	4.51	10.54	31.09	50.07	68.31	93.92
15	0.94	4.74	8.67	25.85	54.59	70.55	88.83
16	1.00	4.42	8.79	30.48	50.93	71.58	94.53
17	0.90	4.30	11.33	36.21	50.31	73.19	90.18
18	0.83	4.36	8.87	28.12	49.21	71.04	90.70
19	0.88	5.06	10.47	33.02	49.13	68.81	86.74
20	0.78	4.75	10.04	29.13	51.02	72.90	95.58
21	0.94	4.83	10.03	30.98	52.78	67.90	88.09
22	0.87	5.33	10.62	29.87	53.98	75.25	84.94
23	0.83	4.04	9.84	29.00	48.96	73.57	88.97
24	0.96	5.29	11.31	30.74	50.77	71.36	93.05
25	0.92	4.67	10.05	29.07	44.81	69.55	93.05
26	0.96	4.53	9.87	32.45	51.42	70.74	87.28
27	0.79	5.11	9.54	30.35	52.79	69.39	90.52
28	1.10	4.62	9.98	30.86	49.31	68.16	87.85
29	0.85	4.68	10.05	30.09	48.59	73.35	91.05
30	0.89	4.64	9.20	30.45	49.87	67.74	91.53
31	1.14	4.52	9.83	30.71	50.77	72.60	91.69
91.89P	0.92	4.59	9.19	27.57	45.95	64.32	82.70

The percentage of rejections, due to the set with the χ^2 closest to the $100P\%$ point, is equal approximately to 91.89 P/N.

here is to order the various configurations (x_1, x_2, x_3) according to the χ^2 value and to compare the cumulative functions of the discrete distributions with the theoretical at arbitrary points, 0.01, 0.05, 0.1(0.2)0.9, 1.0. The results are given in Tables II and III for $N=9(1)31$. In Table II, we find that for normal significance levels of 0.9, 0.7, 0.5, 0.3 and 0.1, the calculated frequencies of assignment or sizes at each level are quite close to the theoretical. The calculated and theoretical values are still reasonably close even at the 5% level and the calculated proportion is usually below the theoretical. At the 1% level there is greater relative differences between the calculated and theoretical but since the values for the calculated vary only from 0.65 to 1.38, they too may be regarded as not unreasonable approximations. And yet, these are much lower values of N than will usually be met in practice.

Table III shows that although the total number of configurations is equal to $\frac{1}{2}(N+1)(N+2)$, the numbers of configurations in the range, 0.1 to 1.0, increase much more slowly, in fact linearly as N. A neighbourhood of a point, 0.8 for example, may be considered. Then a single configuration will be repre-

TABLE III. THE NUMBER OF CONFIGURATIONS ASSIGNABLE TO THE PROBABILITY CLASSES BY THE χ^2 TEST IN THE MULTINOMIALS, $(0.5+0.3+0.2)^N$.

N	The Probability Class														
	0 to 0.01	0.01 to 0.05		0.05 to 0.10		0.10 to 0.30		0.30 to 0.50		0.50 to 0.70		0.70 to 0.90		0.90 to 1.00		Total
	O[1]	O	C[2]	O	C	O	C	O	C	O	C	O	C	O	C	O
9	18	9	16	5	7	12	11	3	5	5	3	2	2	2	1	55
10	25	11	18	5	8	12	12	6	6	2	3	4	3	4	1	66
11	31	11	19	9	8	12	13	6	6	6	4	2	3	2	1	78
12	37	16	21	9	9	14	14	5	7	6	4	3	3	3	1	91
13	51	14	23	6	10	16	16	7	7	6	4	3	4	2	1	105
14	55	21	25	10	11	16	17	7	8	5	5	5	4	1	2	120
15	67	21	26	8	11	18	18	12	8	4	5	4	4	2	2	136
16	79	22	28	10	12	21	19	9	9	6	5	5	4	1	2	153
17	91	24	30	16	13	21	20	6	9	7	6	4	5	2	2	171
18	103	27	32	14	14	21	22	11	10	7	6	5	5	2	2	190
19	117	32	33	15	14	23	23	8	11	7	7	5	5	3	2	210
20	133	33	35	16	15	22	24	12	11	8	7	6	5	1	2	231
21	152	33	37	16	16	25	25	12	12	6	7	6	6	3	2	253
22	169	37	39	17	17	24	26	14	12	8	8	3	6	4	3	276
23	187	35	40	21	17	26	27	13	13	10	8	5	6	3	3	300
24	208	42	42	18	18	26	29	13	13	9	8	7	7	2	3	325
25	227	42	44	20	19	28	30	12	14	12	9	8	7	2	3	351
26	250	43	46	20	20	33	31	13	14	9	10	6	7	4	3	378
27	268	52	47	17	20	34	32	16	15	8	10	8	7	3	3	406
28	300	43	49	23	21	33	33	14	16	10	10	8	8	4	3	435
29	318	51	51	24	22	34	35	15	16	13	11	7	8	3	3	465
30	345	52	53	22	23	38	36	16	17	10	11	10	8	3	3	496
31	377	48	54	25	23	38	37	17	17	12	11	8	8	3	3	528
$c \log_e(\pi_2/\pi_1)$[3]	1.752			0.754		1.196		0.556		0.366		0.274		0.115		—

[1] O = observed [2] C = calculated.

[3] Values calculated from the formula, $2\pi N(0.03)^{1/2} \log_e(\pi_2/\pi_1)$, are given in the columns under C; π_2 and π_1 are the upper and lower boundaries of the probability classes; numerical values of the multiplier of N are given in the last line of the table.

sented by a frequency proportional approximately to $(0.2\times0.3\times0.5)^{-1/2}(2\pi N)^{-1} \exp -\frac{1}{2}\chi^2_{0.8}$, where $\chi^2_{0.8}$ is the χ^2 corresponding to the 0.8 probability level. The density of configurations in the neighbourhood of the 0.8 point will be the inverse of this and so the number of configuration assigned to any class in this neighbourhood will increase linearly with N.

Owing to the special relation between χ^2 for two degrees of freedom and the corresponding $P(\chi^2)$, namely

$$P(\chi^2) = 2\int_{\chi^2}^{\infty} \exp -\tfrac{1}{2}\chi^2 d\chi^2 = \exp -\tfrac{1}{2}\chi^2 \tag{7.2}$$

a simple analytic form can be derived for, ν, the number of configurations in a given probability class, (π_1, π_2).

$$\begin{aligned} \nu(\pi_1, \pi_2, N) &= (0.03)^{1/2} 2\pi N \int_{\pi_1}^{\pi_2} \frac{dP(\chi^2)}{\exp -\frac{1}{2}\chi^2} \\ &= (0.03)^{1/2} 2\pi N \int_{\pi_1}^{\pi_2} P^{-1} dP \\ &= cN \log_e(\pi_2/\pi_1), \end{aligned} \tag{7.3}$$

where c is a constant, $2\pi(0.03)^{1/2} = 1.08828$.

This rule is quite effective as can be seen by comparing observed and calculated numbers of the configuration assigned to the probability classes, the numerical values of $c \log_e \pi_2/\pi_1$ are shown at the foot of the columns of Table III and the calculated numbers, $cN \log_e \pi_2/\pi_1$, are given in the main body of the tables.

What is reasonable agreement between the calculated and theoretical values in Table II? Differences up to at least the value of the frequency of a single configuration (x_1, x_2, x_3) in the neighbourhood of the probability level being considered must be expected, that is, differences of the order of

$$(0.03)^{-1/2}(2\pi N)^{-1} \exp -\tfrac{1}{2}\chi_P^2 = 0.9189P/N. \qquad (7.4)$$

$0.9189\,P$ is tabulated at the foot of Table II. It will be found that comparatively few of the differences of the calculated from the theoretical in Table II are greater than $0.9189\ P/N$ and very few greater than twice this number. This section illustrates the unexpectedly accurate assignment by $P(\chi^2)$ of the different configurations (x_1, x_2, x_3) to the probability classes. This conclusion is consistent with those of Cochran [1] and [2].

Results, rather less favourable than these, would be obtained in the symmetrical multinomials since the number of different partitions of a number N will be smaller than the number of configurations, so that there will be fewer points of increase for the distribution of the discrete χ^2. But even in these multinomials, the assignments to the probability class were found to give reasonable agreement by Lancaster [8].

8. THE CONDITIONS FOR A CRITERION OF SIGNIFICANCE

It is possible to lay down some conditions which the probability, ω, assignable to any set of experimental results should possess.

(i) $\omega = 0$ should be impossible. This is necessary if $-2 \log_e \omega$ is to have finite expectation and variance. It is desirable also that $\omega = 1$ should be impossible in view of the possibility that $-2 \log_e (1-\omega)$ may be required.

(ii) In the null case, ω should be distributed rectangularly in the range, 0 to 1. In a single discrete population, this will not be possible but for a class of distributions it may be true for all a and b such that $0 \leq a < b \leq 1$, that the expectation of ω falling in the interval, (a, b), is proportional to $(b-a)$ approximately.

(iii) ω should be easy to compute.

(iv) The same judgment must always be made on the same data. ω should be uniquely determined by the data.

(v) If an hypothesis is rejected after an experiment at a certain level, further results unfavourable to the hypothesis should not have been able to reverse this judgment.

(vi) The statistical judgment should not be dominated by the auxiliary random experiment.

(vii) There should not be a great discrepancy between the size and (nominal) significance levels.

Now it is easy to see that the exact probability violates (i), (ii), (vii) and sometimes (iii). The χ corrected for continuity violates (i), (ii) and (vii). The auxiliary random experiment violates (iv), (v) and (vi) and sometimes (iii), (vi) being violated especially in statistical control of counting with small numbers because the random sampling will have to be used frequently as in the examples above.

In conclusion, the $P(\chi)$ test seems to have many desirable properties as a test criterion and might well be used in the commonly occurring situations.

REFERENCES

[1] Cochran, W. G., "The χ^2 distribution for the binomial and Poisson series, with small expectations," *Annals of Eugenics (Lond.)*, 7 (1936), 207–17.

[2] Cochran, W. G., "The χ^2 correction for continuity," *Iowa State College Journal of Science*, 16 (1942), 421–36.

[3] Eisenhart, C. (Editor), *Tables of the Binomial Probablity Distribution*, National Bureau Standards, Applied Mathematics Series 6, U. S. Dept. Commerce, Washington, D. C., 1949.

[4] el Shanawany, M. R., "An illustration of the accuracy of the χ^2 approximation," *Biometrika*, 28 (1936), 179–87.

[5] Fisher, R. A., "The logic of inductive inference," *Journal of the Royal Statistical Society*, 98 (1935), 39–82.

[6] Fisher, R. A., "The logical inversion of the notion of the random variable," *Sankhyā*, 7 (1945), 129–32.

[7] Fisher, R. A., Thornton, H. G., and Mackenzie, W. A., "The accuracy of the plating methods of estimating the density of bacterial populations," *Annals of Applied Biology*, 9 (1922), 325–59.

[8] Lancaster, H. O. "Statistical control of counting experiments," *Biometrika*, 39 (1952), 419–22.

[9] Neyman, J. and Pearson, E. S. "Further notes on the χ^2 distribution," *Biometrika*, 22 (1931), 298–305.

[10] Pearson, E. S. "On questions raised by the combinations of tests based on discontinuous distributions," *Biometrika*, 37 (1950), 383–98.

[11] Wallis, W. Allen, "Compounding probabilities from independent significance tests," *Econometrica*, 10 (1942), 229–48.

[12] Yates, F. "Contingency tables involving small numbers and the χ^2 test," *Journal of the Royal Statistical Society. Suppl.* 1 (1934), 217–35.