DISCRIMINATING BETWEEN BINOMIAL DISTRIBUTIONS

By Paul G. Hoel

University of California at Los Angeles

1. Summary. Given a set of k random samples, $x_1, x_2, \cdots, x_k$, from a binomial distribution with parameters p and n, it is shown that the familiar binomial index of dispersion

$$z = \frac{\sum_1^k (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n_0}\right)}$$

yields an approximate best critical region independent of p for testing the hypothesis $n = n_0$ against the alternative hypothesis $n > n_0$, provided $\bar{x}$ and $n_0 - \bar{x}$ are not small. Because of the nature of the test, its optimum properties also apply to testing whether the data came from a binomial population with $n = n_0$ or from a Poisson population.

2. Introduction. A problem of considerable interest in certain fields is that of deciding whether a set of observations should be treated as having come from either a binomial population or from a Poisson population. Although there was much discussion a few years ago concerning the best method for making such a decision [1], [2], [3], no solution of the problem was presented. In this paper a test that possesses certain optimum properties is derived for discriminating between two binomial populations. This test, however, is also capable of solving the problem of how to discriminate between a binomial and a Poisson population. The methods that are employed in the derivation of this test are similar to those of an earlier paper [4] in which the problem of discriminating between two Poisson populations was studied.

3. Similar regions. Let n denote the number of trials and p the probability of success in a single trial for a binomial distribution. Let $x_1, x_2, \cdots, x_k$ represent the observed frequencies in k random samples from this binomial population. Now consider the two alternative hypotheses

$$H_0 : n = n_0, p = p_0$$

and

$$H_1 : n = n_1 > n_0, p = p_1.$$

The purpose of this paper is to construct a test for discriminating between the two values of n regardless of the values of p; however it is convenient to begin with these more restrictive hypotheses

For the purpose of finding a critical region for testing H_0 against H_1, the x_i will be treated as the coordinates of a point in k dimensions. The probability of obtaining the particular point $x_1, \cdots, x_k$ when H_0 is true will be denoted by $P_0[x_i]$. Since the probability of obtaining x successes in n trials is given by

$$\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

it follows that

$$P_0[x_i] = \frac{(n_0!)^k}{\prod_1^k x_i!(n_0 - x_i)!} p_0^{\sum_1^k x_i} q_0^{\sum_1^k (n_0 - x_i)}. \tag{1}$$

In searching for a critical region that will be independent of p_0, it is illuminating to study the methods that were designed by Neyman and Pearson [5] for continuous distributions. These methods suggest that one should look for critical regions on the surfaces $\sum_1^k x_i = \text{constant}$. For this reason, instead of using (1) for constructing critical regions, it is desirable to study the conditional probability distribution of the points lying in the plane $\sum_1^k x_i = N$, where N is a positive integer not exceeding kn_0. The conditional probability of obtaining the point $x_1, \cdots, x_k$, when the point is restricted to lie in the plane $\sum_1^k x_i = N$, will be denoted by $P_0[x_i \mid N]$. Its value may be obtained by dividing the probability (1) by the probability that the point will lie in the plane $\sum_1^k x_i = N$. If this latter probability is denoted by $P_0[N]$, then

$$P_0[x_i \mid N] = \frac{P_0[x_i]}{P_0[N]}. \tag{2}$$

Since the sum of k independent variables each possessing the same binomial distribution has a binomial distribution with n replaced by kn, it follows that N possesses a binomial distribution and that

$$P_0[N] = \frac{(kn_0)!}{N!(kn_0 - N)!} p_0^N q_0^{kn_0 - N}. \tag{3}$$

If (1) and (3) are substituted in (2), it will reduce to

$$P_0[x_i \mid N] = \frac{(n_0!)^k N!(kn_0 - N)!}{(kn_0)! \prod_1^k x_i!(n_0 - x_i)!}. \tag{4}$$

This conditional probability distribution in the plane $\sum_1^k x_i = N$ is independent of p_0 and therefore may serve as the basis for constructing a critical region that

is independent of p_0 for testing H_0 against H_1. It will therefore be possible to test the less restrictive hypothesis

$$H_0' : n = n_0$$

against

$$H_1' : n = n_1 > n_0 .$$

4. Best critical region. Although a best critical region does not exist for testing H_0' against H_1', it is helpful to proceed as though one did.

If a critical region of size α could be selected in each plane $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$, then the totality of such critical regions would constitute a critical region of size α that is independent of p_0 and which therefore could be used to test H_0' against H_1'. For, if P_0 [$X \in$ C.R.] denotes the probability that the sample point, which will be denoted by X, will lie in the critical region, it follows that

$$\begin{aligned} P_0[X \in \text{C.R.}] &= \sum_{N=0}^{kn_0} P_0[N] P_0[X \in \text{C.R.} \mid N] \\ &= \sum_{N=0}^{kn_0} P_0[N]\alpha \\ &= \alpha. \end{aligned} \tag{5}$$

This last equality follows from the fact that the sample point must lie in one of the planes $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$.

Furthermore, this would be the only critical region of size α independent of p_0, because if a critical region of size α_N, $(N = 0, 1, \cdots, kn_0)$, were selected in the plane $\sum_1^k x_i = N (N = 0, 1, \cdots, kn_0)$, it would be necessary that

$$\sum_{N=0}^{kn_0} P_0[N]\alpha_N = \alpha,$$

independent of the value of p_0. From (3) this is equivalent to requiring that

$$\sum_{N=0}^{kn_0} \frac{(kn_0)!}{N!(kn_0 - N)!} p_0^N (1 - p_0)^{kn_0 - N} \alpha_N = \alpha, \tag{6}$$

independent of the value of p_0. Since the left side of (6) is a polynomial in p_0, its constant term must equal α and all other coefficients must vanish. It will be observed that no terms of the sum in (6) that arise from $N > r$ will contribute to the coefficient of p_0^r; consequently this coefficient will not contain the unknowns $\alpha_{r+1}, \cdots, \alpha_{kn_0}$. These considerations show that the α_N must satisfy equations of the form

$$\begin{aligned}
\alpha &= c_{00}\alpha_0 \\
0 &= c_{10}\alpha_0 + c_{11}\alpha_1 \\
&\cdots\cdots \\
0 &= c_{kn_0 0}\alpha_0 + c_{kn_0 1}\alpha_1 + \cdots + c_{kn_0 kn_0}\alpha_{kn_0}.
\end{aligned}$$

It will also be observed that $c_{rr} = (kn_0)!/r!(kn_0 - r)!$; consequently the triangular matrix of the coefficients in these $kn_0 + 1$ non-homogeneous equations is non-singular. The equations therefore possess a unique solution, namely the known solution of $\alpha_N = \alpha$.

The preceding discussion shows that it is necessary to find critical regions of size α in each plane $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$, if a critical region independent of p_0 is desired. If each such planar critical region were a best critical region for that plane, then the totality of such regions would constitute a best critical region independent of p_0 for testing H_0' against H_1'.

It follows from the theory of best critical regions [5] that if a best critical region in the plane $\sum_1^k x_i = N$ did exist, it would be determined by the inequality

$$\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} < K, \tag{7}$$

where P_1 corresponds to P_0 when H_1 is true and where K is a constant whose value is chosen to make the critical region one of size α. Now from (4),

$$\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} = \frac{(n_0!)^k(kn_0 - N)!\,(kn_1)!\,\Pi(n_1 - x_i)!}{(n_1!)^k(kn_1 - N)!\,(kn_0)!\,\Pi(n_0 - x_i)!}. \tag{8}$$

In order to study the possibility of a best critical region, it is therefore necessary to study the possibility of (8) satisfying inequality (7).

5. Approximate best critical region. Unfortunately, because the variables x_i are discrete, it is not possible to find critical regions of exactly size α for arbitrary α as required in (5). Consequently it is necessary to introduce continuous approximating functions for discrete probability functions or to resort to other devices if critical regions of the type discussed in the preceding section are to be obtained.

For the purpose of introducing such approximations, (8) will be written in the following form:

$$\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} = c_1 \frac{(kn_0 - N)!}{\Pi(n_0 - x_i)!}\left(\frac{1}{k}\right)^{kn_0 - N} \div \frac{(kn_1 - N)!}{\Pi(n_1 - x_i)!}\left(\frac{1}{k}\right)^{kn_1 - N}, \tag{9}$$

where c_1 is independent of the variables x_i. It will be observed that the ratio on the right is a ratio of two multinomial functions. Now the multinomial function

$$\frac{N!}{x_1!x_2!\cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

where $\sum_1^k x_i = N$, can be approximated by the multivariate normal function

$$\frac{e \exp\left[-\frac{1}{2}\sum_1^k \left(\frac{x_i - Np_i}{\sqrt{Np_i}}\right)^2\right]}{(2\pi N)^{\frac{1}{2}(k-1)}\sqrt{p_1 p_2 \cdots p_k}}. \tag{10}$$

The approximation is good provided the Np_i are large and the x_i remain away from their extreme values. If this approximation is applied to both numerator and denominator of (9), to this order of approximation,

$$\begin{aligned}\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} &= c_1 \frac{k^{k/2} e \exp\left[-\frac{1}{2}\sum_1^k \left(\frac{x_i - N/k}{\sqrt{n_0 - N/k}}\right)^2\right]}{[2\pi(kn_0 - N)]^{\frac{1}{2}(k-1)}} \\ &\quad \div \frac{k^{k/2} e \exp\left[-\frac{1}{2}\sum_1^k \left(\frac{x_i - N/k}{\sqrt{n_1 - N/k}}\right)^2\right]}{[2\pi(kn_1 - N)]^{\frac{1}{2}(k-1)}} \\ &= c_1 \left[\frac{kn_1 - N}{kn_0 - N}\right]^{\frac{1}{2}(k-1)} e \exp\left[-\frac{1}{2}\frac{n_1 - n_0}{(n_1 - N/k)(n_0 - N/k)} \cdot \sum_1^k (x_i - N/k)^2\right].\end{aligned} \tag{11}$$

Since, by hypothesis, $n_1 > n_0$ and $n_0 > N/k$, except for the case of $n_0 = N/k$, which will be considered later, it follows that

$$\frac{n_1 - n_0}{(n_1 - N/k)(n_0 - N/k)} > 0.$$

As a consequence, the right side of (11) will decrease in value as $\sum_1^k (x_i - N/k)^2$ increases in value. If $(x_1, \cdots, x_k)$ is a point lying on the sphere

$$\sum_1^k (x_i - N/k)^2 = R \tag{12}$$

and if the coordinates of this point satisfy inequality (7) when approximation (11) is used, then all points outside this sphere will also satisfy (7) to this same order of approximation. A best critical planar region of size α in this approximate sense can therefore be obtained in the plane $\sum_1^k x_i = N$ by determining a

sphere with center at $(N/k, \cdots, N/k)$ such that when H_0' is true the probability is α that a point lying in the plane will lie outside this sphere. Furthermore, such a region will be a common best critical region for all values of $n_1 > n_0$ because the preceding arguments do not require the value of n_1 but merely the knowledge that $n_1 > n_0$.

For the purpose of determining the radius of the sphere that will yield the desired critical region, (4) will be expressed as follows:

$$P_0[x_i \mid N] = c_2 \frac{N!}{\Pi x_i!}\left(\frac{1}{k}\right)^N \frac{(kn_0 - N)!}{\Pi(n_0 - x_i)!}\left(\frac{1}{k}\right)^{kn_0 - N}, \tag{13}$$

where c_2 is independent of the x_i. If these multinomials are replaced by their multivariate normal approximations as given by (10), to this approximation (13) will reduce to

$$\begin{aligned} P_0[x_i \mid N] &= c_3\, e \exp\left[-\tfrac{1}{2}\sum_1^k \left(\frac{x_i - N/k}{\sqrt{N/k}}\right)^2\right] e \exp\left[-\tfrac{1}{2}\sum_1^k \left(\frac{x_i - N/k}{\sqrt{n_0 - N/k}}\right)^2\right] \\ &= c_3\, e \exp\left[-\tfrac{1}{2}\frac{\sum_1^k (x_i - N/k)^2}{\frac{N}{k}\left(1 - \frac{N}{kn_0}\right)}\right] \end{aligned} \tag{14}$$

where c_3 is independent of the x_i. Since $\sum_1^k x_i = N$ here, x_k may be expressed in terms of the remaining variables; consequently (14), except for a constant factor, may be treated as a normal distribution in the variables $x_1, \cdots, x_{k-1}$. If the factorials in c_3 are replaced by their Stirling approximations, it will be found that c_3 is the correct constant for the normal distribution.

Since it is known [6] that -2 times the exponent in a normal distribution function possesses a chi-square distribution, it follows that to this order of approximation

$$\frac{\sum_1^k (x_i - N/k)^2}{\frac{N}{k}\left(1 - \frac{N}{kn_0}\right)} \tag{15}$$

possesses a chi-square distribution with $k - 1$ degrees of freedom. If χ_α^2 is a value such that $P[\chi^2 > \chi_\alpha^2] = \alpha$, then

$$\frac{\sum_1^k (x_i - N/k)^2}{\frac{N}{k}\left(1 - \frac{N}{kn_0}\right)} = \chi_\alpha^2 \tag{16}$$

determines a sphere such that to this order of approximation the probability is α that a point lying in the plane $\sum_1^k x_i = N$ will lie outside the sphere. From

the arguments following (12), it therefore follows that a common best critical region in this approximate sense for testing H_0' against H_1' will consist of that part of each plane $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$, which lies outside the corresponding sphere given by (16). Since the x_i are non-negative and do not exceed n_0, the planes corresponding to $N = 0$ and $N = kn_0$ contain a single point; therefore it is necessary to adopt some convention that assigns 100α percent of the samples with $N = 0$ and $N = kn_0$ to a critical region in order to obtain critical regions of size α in these two cases.

For a given set of data, the procedure to be followed then consists in calculating the statistic

$$z = \frac{\sum_1^k (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n_0}\right)},$$

where $\bar{x} = \sum_1^k x_i/k$, and agreeing to reject the hypothesis that $n = n_0$ in favor of the alternative hypothesis that $n > n_0$ if and only if $z > \chi_\alpha^2$, where $P[\chi^2 > \chi_\alpha^2] = \alpha$ for $k - 1$ degrees of freedom. Because of the nature of the approximations used in (10) and (14), this result may be expected to be accurate only if $\bar{x}$ and $n_0 - \bar{x}$ are large.

The interesting feature of this result is that the familiar binomial index of dispersion, z, possesses optimum properties in this approximate sense for testing $n = n_0$ against $n > n_0$.

6. Poisson application. Since the preceding test will possess approximate optimum properties for n as large as desired, independent of the value of p, and since a Poisson distribution with parameter m can be approximated as closely as desired by means of a binomial distribution with $np = m$ by allowing n to increase sufficiently, it follows that the test will also possess approximate optimum properties for deciding between a binomial distribution with $n = n_0$ and a Poisson distribution.

7. Estimation of n. Although the purpose of this paper has been accomplished in the preceding sections, it is interesting to observe the role played by the closely related Poisson index of dispersion in the extimation of n.

Approximate confidence limits for n may be obtained by means of (16). If $\chi_{1-\alpha}^2$ is a value of χ^2 such that $P[\chi^2 > \chi_{1-\alpha}^2] = 1 - \alpha$, then, to this same order of approximation, the probability is $1 - 2\alpha$ that

$$\chi_{1-\alpha}^2 < \frac{\sum_1^k (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n}\right)} < \chi_\alpha^2.$$

If these inequalities are solved for n, the following $100(1 - 2\alpha)$ percent approximate confidence limits for n will be obtained:

$$\frac{\bar{x}\chi_\alpha^2}{\chi_\alpha^2 - \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}} < n < \frac{\bar{x}\chi_{1-\alpha}^2}{\chi_{1-\alpha}^2 - \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}}. \tag{17}$$

Only the lower limit here will possess optimum properties. Now it will be observed that only positive values of n will be admissible if

$$\frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} \leq \chi_{1-\alpha}^2,$$

whereas only negative values will be admissible if

$$\frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} \geq \chi_\alpha^2.$$

The range of values will be infinite in each case if there is equality rather than inequality. If, however,

$$\chi_{1-\alpha}^2 < \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} < \chi_\alpha^2,$$

then both positive and negative values of n over infinite ranges will be admissible. Since n increases as the Poisson index $\Sigma(x_i - \bar{x})^2/\bar{x}$ increases until it becomes infinite and then increases from minus infinity through negative values, (17) may still be thought of as giving an interval (infinite) of values with a positive "lower" limit and a negative "upper" limit. Thus, the familiar Poisson index of dispersion plays an interesting role in determining whether a Poisson assumption is reasonable as far as admissible values of n are concerned.

If the population is truly binomial, negative values of n must be ruled out; consequently a Poisson assumption becomes increasingly tenable as the Poisson index increases. However, experience has shown [7] that a negative binomial distribution is often more realistic in describing data supposedly drawn from a binomial or Poisson population than is the assumed distribution; consequently a negative binomial should be given consideration if (17) yields only negative values or if it yields a negative "upper" limit that is numerically small relative to a positive "lower" limit.

It is also interesting to consider the point estimation of n. Here, it is customary [7] to estimate n by means of

$$\frac{k\bar{x}}{k - \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}}.$$

Thus, a positive, infinite, or negative estimate for n will be obtained according as the Poisson index is less than, equal to, or greater than k.

REFERENCES

[1] J. BERKSON, "Some difficulties of interpretation encountered in the application of the chi-square test," *Jour. Amer. Stat. Assoc.*, Vol. 33 (1938), pp. 526–536.

[2] B. H. CAMP, "Further interpretations of the chi-square test," *Jour. Amer. Stat. Assoc.*, Vol. 33 (1938), pp. 537–542.

[3] J. BERKSON, "A note on the chi-square test, the Poisson, and the binomial," *Jour. Amer. Stat. Assoc.*, Vol. 35 (1940), pp. 362–367.

[4] P. G. HOEL, "Testing the homogeneity of Poisson frequencies," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 362–368.

[5] J. NEYMAN AND E. S. PEARSON, "On the problem of the most efficient tests of statistical hypotheses," *Roy. Soc. Phil. Trans.*, Vol. 231 (1933), pp. 289–337.

[6] S. S. WILKS, *Mathematical Statistics*, Princeton Univ. Press, 1943, p. 104.

[7] "STUDENT," "An explanation of deviations from Poisson's law in practice," *Biometrika*, Vol. 12 (1919), pp. 211–215.