DETERMINING SAMPLE SIZE TO BOUND THE PROBABILITY OF CLASSIFYING A SAMPLE INTO THE WRONG ONE OF TWO MULTINOMIALLY DISTRIBUTED POPULATIONS

C. Philip Cox
DETERMINING SAMPLE SIZE TO BOUND THE PROBABILITY OF CLASSIFYING A SAMPLE INTO THE WRONG ONE OF TWO MULTINOMIALLY DISTRIBUTED POPULATIONS

C. Philip Cox
Iowa State University
Department of Statistics, Ames, IA 50011-1210

ABSTRACT

The problem considered is that of choosing between the two
\[ \pi_{ij}, \Sigma \pi_{ij} = 1, i = A, B, \] of known multinomial
probabilities on the basis of sample values \( x_j \), the observed counts in the
\[ j = 1, \ldots, k, \] classes, with \( \Sigma x_j = N \). The particular question examined is
'how large should \( N \) be to achieve reliable differentiation?'. It is shown
how to find \( N \) such that the probability of misclassification does not
exceed a prescribable value. The method is exemplified in a genetic
context.

KEY WORDS: categorized data, \( \chi^2 \), cytogenetics, goodness-of-fit,
misclassification probabilities, multinomial distributions,
sample size, soybean breeding.

1. Introduction

The problem to be considered arose in an agricultural context,
specifically in a genetic study (Hedges, 1989) of the occurrence of those
soybean mutants (trisomics) which contain an extra chromosome and are 'not
inherited in a normal Mendelian manner'. Hedges noted that disomic and
trisomic soybean individuals can be identified only by chromosome counts
and for both types he calculated the expected segregation ratios, that is,
the proportions of \( F_2 \) progeny, to be expected in three classes. The
question which then naturally arises is - how many individuals should be
examined to obtain a reliable choice between the two types of segregation?
It is widely appreciated - if less widely implemented - that sample size
determinations are essential to the planning of efficient experimentation
and their importance is now increasing with sensitivity to ethical
considerations in, for example, clinical and other trials using animal
subjects.

Instead of the commonly exposited statistical context wherein the test and
alternative hypotheses are simple and composite respectively, both are now
simple and such situations are usually treated as classification problems
in the multi-continuous-variate literature. The hypothesis testing
approach can, however, be retained by symmetrically regarding the
erroneous classification of a sample from either one into the other
population as analogous to the usual Type I error. Accordingly the
multi-discrete-variate case considered here entails the choice between
the two specifications \( \pi_{ij}, \sum_{j=1}^{k} \pi_{ij} = 1, i = A, B, \) of known multinomial
probabilities on the basis of a data set of values \( x_j, \) the observed counts
in the \( j = 1, \ldots, k \) classes. In the genetic context Mather (1938, 1951)
has given a solution for the \( k = 2 \) case; Hanson (1959) has summarized some
related studies; solutions for \( k \geq 2 \) classes are presented here.

2. Theoretical aspects

Suppose that a total of \( N \) values are distributed into \( k \) classes, that \( x_j \)
is the number in the \( j \)th class and that \( p_j = x_j/N, j = 1, \ldots, k. \) Because
\( \sum x_j = N \) and equivalently \( \sum p_j = 1, \) it is sufficient to consider only the
first \( k - 1 \) classes and, on the assumption that \( N \) is large enough, the
mean of the multivariate normal distribution of the vector \( \pi =
[p_1, p_2, \ldots, p_{k-1}]' \) is \( \pi = [\pi_1, \pi_2, \ldots, \pi_{k-1}]' \) where \( \pi_j \) is the population
probability for the occurrence of a value in the \( j \)th class. The
covariance matrix \( \Sigma, \) of the vector has diagonal elements \( \pi_j(1-\pi_j)/N \) and
off-diagonal elements \( -\pi_i \pi_j/N, i \neq j \) and it is easily shown that \( |\Sigma| =
\pi_1, \pi_2, \ldots, \pi_{k-1}/N \) and that \( \Sigma = \sum^{-1}[D - \pi\pi'] \) where the \( j,j \)th element of the
diagonal matrix \( D \) is \( \pi_j. \) It then follows, e.g., from Theorem 3.3.3 in
Anderson (1984) that
\[
[p - \pi] \Sigma^{-1} [p - \pi] \sim \chi^2_{k-1}
\]
Hence it seems intuitively reasonable that a \( p \)-vector can be classified as
a member of population \( i, \) with probability of misclassification \( \alpha, \) if 'the
test statistic'
\[
[p - \pi_i] \Sigma^{-1}_i [p - \pi_i] < \chi^2_{(k-1; \alpha)}.
\]  
(1)

When as here \( \Sigma_A \neq \Sigma_B, \) however, difficulties arise because it is
conceivable that (1) may be either true or false for both of \( i = A \) and \( i = B. \) To examine this we first note that the inverse of \( \Sigma \) is \( \Sigma^{-1} = \sum^{-1}
+ (1/\pi_k)J \) where \( J \) is the uniform matrix. Hence or otherwise, the test
statistic in (1) can be expressed in the standard symmetrical form as
\[
\chi^2_{IT} = \sum_{j=1}^{k} \left( \frac{p_j - \pi_{ij}}{\pi_{ij}} \right)^2 / \pi_{ij}
\]
which is easily reduced to the equivalent (and computationally more
convenient) form:
The surfaces $\chi^2_{i1T} = \text{constant}$ are hyper-ellipsoids in $R^k$ and these intersect in ellipsoids of $k - 1$ dimensions with the hyperplane

$$\Sigma \pi_{ij} = 1$$

which contains the points $(p_1, p_2, \ldots, p_k)$ and $(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})$, $i = A, B$. It is then easily shown that the locus of points in this plane for which

$$\chi^2_{AT} = \chi^2_{BT}$$

is

$$\sum_{j=1}^{k} p^2_j \left( \frac{1}{\pi_{Aj}} - \frac{1}{\pi_{Bj}} \right) = 0$$

which, necessarily, passes through the origin $(0,0,\ldots,0)$ and does not depend on $N$. To this stage therefore, with $K$ equal to the left hand side of (3), the decision rule:

Take A as the parent population if $K \leq 0$, if not take B and if $K > 0$ take B as the parent, if not take A,

has the attribute that, if the two parent populations are 'equally likely', the probabilities of misclassification are equal. Practical implementation of this apparently commonsensical procedure has the drawback that, except for the, could-be-inefficient, professional axiom - the larger $N$ is, the better - there is no control over the actual size of the probability of misclassification. A resolution applicable for $k = 3$, is next considered.

3. The $k = 3$ case - a geometrical approach

When $k = 3$ at least one of the coefficients of $p^2_j$ in (3) must be negative so that, multiplying through by $-1$ and relabelling if necessary, the surface (3) can be written as

$$a^2_1 p^2_1 - a^2_2 p^2_2 - a^2_3 p^2_3 = 0,$$

where

$$a^2_j = \left| \frac{1}{\pi_{Aj}} - \frac{1}{\pi_{Bj}} \right|,$$

which defines a degenerate surface in $R^3$, specifically that generated by the line of intersection of two planes. With (4) as

$$(a_1 p_1 - a_2 p_2)(a_1 p_1 + a_2 p_2) = (a_3 p_3)^2$$
the two planes are
\[
\begin{align*}
  a_1 p_1 - a_2 p_2 - \gamma a_3 p_3 &= 0 \\
  a_1 p_1 + a_2 p_2 - \frac{1}{\gamma} a_3 p_3 &= 0
\end{align*}
\]
(5)
where, in general, \( \gamma \) is an arbitrary constant.

The line through the origin defined by (5) will intersect the plane
\[ p_1 + p_2 + p_3 = 1 \]
(6)
in a single point \( P_\gamma \) say and the locus of \( P_\gamma \) as \( \gamma \) changes will be the intersection of the surface (3) with the plane (6). The coordinates \( (P_{\gamma 1}, P_{\gamma 2}, P_{\gamma 3}) \), abbreviated as \( (P_1, P_2, P_3) \), are
\[
\begin{bmatrix}
  P_1 \\
  P_2 \\
  P_3
\end{bmatrix} = \begin{bmatrix}
  1 & 1 & 1 \\
  a_1 & -a_2 & -\gamma a_3 \\
  a_1 & a_2 & -a_3/\gamma
\end{bmatrix}^{-1} \begin{bmatrix}
  1 \\
  0 \\
  0
\end{bmatrix}
\]
so that
\[
\begin{align*}
P_1 &= a_2 a_3 (1+\gamma^2)/\gamma \Delta, \\
P_2 &= a_1 a_3 (1-\gamma^2)/\gamma \Delta, \\
P_3 &= 2 a_1 a_2/\Delta
\end{align*}
\]
(7)
where, directly or because \( \Sigma P_i = 1 \),
\[
\Delta = a_2 a_3 (1+\gamma^2)/\gamma + a_1 a_3 (1-\gamma^2)/\gamma + 2 a_1 a_2
\]
\[
= a_1 a_2 a_3 \left\{ \left( \frac{1}{\gamma} + \gamma \right)/a_1 + \left( \frac{1}{\gamma} - \gamma \right)/a_2 + 2/a_3 \right\}
\]
and, \( 0 \leq \gamma \leq 1 \) because the coordinates must here be positive.

At each point \( P_\gamma \) defined by (7) the values of the \( \chi^2_T \) 'test statistics' - for departure from populations A and B - will be equal and, for some \( P_\gamma = P_{\min} \) 'between' the points \( \pi_{A1}, \pi_{A2}, \pi_{A3} \) and \( \pi_{B1}, \pi_{B2}, \pi_{B3} \), the value of the test statistic will achieve its minimum value. The probabilities of misclassification may then be controlled by designating an \( N \) so large that, evaluated at \( P_{\min} \), the probabilities do not exceed a prescribed value.

Finding \( P_{\min} \)

Noting that on the locus of equal \( \chi^2_T \)-values,
\[
1 + N^{-1} \chi^2_T = \Sigma p_j^2/\pi_{Aj} = \Sigma p_j^2/\pi_{Bj}
\]
(8)
it suffices to minimize, with respect to \( \gamma \),

\[
H = \sum b_j P_j^2
\]

(9)

wherein \( b_j = 1/\pi_{A_j} \) and the \( P_j \) are obtained from (7).

Accordingly the equation \( \frac{dH}{d\gamma} = 0 \) gives, after algebraic reductions, the stationary points of \( H \) as the solutions of the quartic equation:

\[
a_1a_2(c_1+c_2)\gamma^4 + 2a_3(a_1c_1+a_2c_2)\gamma^3 + 2a_3(a_1c_1-a_2c_2)\gamma^2 - a_1a_2(c_1+c_2) = 0
\]

(10)

wherein

\[
c_1 = \frac{b_1}{a_1} + \frac{b_2}{a_2} \quad \text{and} \quad c_2 = \frac{b_2}{a_2} - \frac{b_3}{a_3}
\]

(11)

Since the expression on the left of (10) is negative at \( \gamma = 0 \) and positive at \( \gamma = 1 \) it does have a root giving positive values for the \( P_{\min} \) coordinates in (7). With these and the specifiable value of \( \chi^2_{T} \), (8) can then be solved for the required value of \( N \). The development to this stage is next exemplified.

**Example 1**

'The expected genotypic frequencies in the \( F_2 \) progeny of an \( A_1A_1A_2 \) individual assuming maximal equational reduction' were given in Hedges (1989), Table 2, as

<table>
<thead>
<tr>
<th></th>
<th>( A_1^- )</th>
<th>( A_1A_2^- )</th>
<th>( A_2^- )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trisomics</td>
<td>10</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Disomics</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

so that the population probabilities for the three classes are (10/36, 25/36, 1/36) and (4/9, 4/9, 1/9) for the trisomics and disomics respectively. Since 25/36 > 4/9 and the other two such differences are negative, the first two classes are first interchanged to give the specification:

<table>
<thead>
<tr>
<th></th>
<th>( \pi_1 )</th>
<th>( \pi_2 )</th>
<th>( \pi_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>trisomics (A)</td>
<td>25/36</td>
<td>10/36</td>
<td>1/36</td>
</tr>
<tr>
<td>disomics (B)</td>
<td>16/36</td>
<td>16/36</td>
<td>4/36</td>
</tr>
</tbody>
</table>

so that (4) becomes
0.81p_1^2 - 1.35p_2^2 - 27p_3^2 = 0

with

\[ a_1^2 = (36/16 - 36/25) = 0.81, \quad a_2^2 = 1.35, \quad a_3^2 = 27 \]

and

\[ b_1 = 1.44, \quad b_2 = 3.6, \quad b_3 = 36 \]

and, from (10),

\[ c_1 = 28/9 \quad \text{and} \quad c_2 = 12/9. \]

Substitutions in (10) then give the following quartic equation for \( \gamma \):

\[ \gamma^4 + 9.72509\gamma^3 + 2.79689\gamma - 1 = 0 \]

of which the root \( 0 \leq \gamma = 0.2795 \leq 1 \) is the one required. The corresponding coordinates of \( P_{\min} \) from (7) are

\[ (P_1, P_2, P_3) = (0.5707, 0.3780, 0.0513). \]

Finally, using (9) and (8), the minimum value of \( N^{-1}X_T^2 \) is calculated as 0.0781 which exceeds \( \chi^2(2, 0.05) \) if \( N > 76.7 \).

Example 2 – a degenerate trinomial case

If one of \( a_2^2 \) and \( a_3^2 \) in (4) is zero the quartic equation (10) does not properly reduce to give the required solution. In this case, however, the proper solution can be obtained as follows.

Suppose that \( \pi_{A3} = \pi_{B3} \) so that, because \( a_3^2 = 0 \), the two \( \chi_T^2 \)'s are equal if

\[ a_1^2p_1 - a_2^2p_2 = (a_1p_1 - a_2p_2)(a_1p_1 + a_2p_2) = 0 \]

Because neither of \( p_1 \) and \( p_2 \) can be negative the locus of points giving equal \( \chi_T^2 \)'s is therefore the line of intersection of the two planes,

\[ a_1p_1 - a_2p_2 = 0 \quad \text{and} \quad p_1 + p_2 + p_3 = 1 \]

The coordinates of a point on this line are then

\[ P_1 = \gamma a_2/(a_1 + a_2), \quad P_2 = \gamma a_1/(a_1 + a_2), \quad P_3 = 1 - \gamma \]

(12)

and, with \( H \) from (9), the \( \gamma \)-value which minimizes \( \chi_T^2 \) is easily obtained from \( \frac{dH}{d\gamma} = 0 \) or directly because \( H \) is quadratic in \( \gamma \). The results are
that:

\[
\gamma = b_3(a_1 + a_2)^2 / \left( a_1^2 b_2 + a_2^2 b_1 + b_3(a_1 + a_2)^2 \right)
\]

\[
H_{\min} = b_3(1-\gamma) = \frac{b_3(a_1^2 b_2 + a_2^2 b_1)}{a_1^2 b_2 + a_2^2 b_1 + (a_1 + a_2)^2 b_3}
\]

(13)

The determination of the value of N required then proceeds, via (8), as before.

Example 2 (Hedges 1992)

The specifications for the populations A and B were

<table>
<thead>
<tr>
<th></th>
<th>(\pi_1)</th>
<th>(\pi_2)</th>
<th>(\pi_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1/2</td>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>B</td>
<td>13/18</td>
<td>1/36</td>
<td>1/4</td>
</tr>
</tbody>
</table>

from which are calculated:

\[a_1^2 = 2 - (18/13) = 8/13, a_2^2 = |4-36| = 32, a_3^2 = 0\]

\[b_1 = 2, b_2 = 4, b_3 = 4\]

\(H_{\min}\) is then found directly from (13) to be 1.1438 whence (8) gives \(N > 41.7\) for \(\chi^2_T = \chi^2(2;0.05)\). Calculation from (12) incidentally shows that the minimum \(\chi^2_T\) - value occurs at the point (0.627, 0.087, 0.286).

4. A general method for any number of classes

With the slightly revised notation

\[d_j = \frac{1}{\pi_{Aj}} - \frac{1}{\pi_{Bj}}, b_j = \frac{1}{\pi_{Aj}}, j = 1,\ldots,k\]

(14)

so that \(d_j\) is no longer necessarily positive, the general problem is the minimization of

\[H = \sum_{j=1}^{k} b_j p_j^2\]

subject to the constraints that \(p_j \geq 0\) and,
Using the Lagrangian procedure we accordingly seek to minimize

\[ \phi = H - \lambda_1 \sum_j p_j^2 - 2\lambda_2 (\sum_j p_j - 1) \]

which from \( \frac{\partial \phi}{\partial p_j} = 0 \) gives

\[ (b_j - \lambda_1 d_j)p_j = \lambda_2 \]  

(17)

and hence, from (16), the appropriate solution \( \lambda_1 \) of

\[ f(\lambda_1) = \sum_j (b_j - \lambda_1 d_j)^2 = 0 \]

(18)

is required.

At \( \lambda_1 = 0 \)

\[ f(\lambda_1) = \sum_j b_j^2 = \sum_j \frac{1}{\pi_Aj} - \frac{1}{\pi_Bj} \]

\[ = 1 - \sum_j \frac{\lambda_j}{\pi_Bj} \]

\[ = -N^{-1} \chi_{BT}^2 \]

using (2), where \( \chi_{BT}^2 \) is the necessarily positive test statistic for examining the significance of the deviation of the point \( (\pi_{A1}, \ldots, \pi_{Ak}) \) from the point \( (\pi_{B1}, \ldots, \pi_{Bk}) \). A similar argument shows that \( f(\lambda_1) \) is positive at \( \lambda_1 = 1 \). There is therefore at least one real root in \( 0 \leq \lambda_1 \leq 1 \). Further, in

\[ f'(\lambda_1) = 2\sum_j (b_j - \lambda_1 d_j)^3, \]

\[ b_j - \lambda_1 d_j = \frac{(1-\lambda_1)}{\pi_Aj} + \frac{\lambda_1}{\pi_Bj} \]

is positive so that \( f(\lambda_1) \) is monotonic and the root in the interval is unique. Equation (18) is of degree \( 2(k-1) \) in \( \lambda_1 \) and numerical solution is indicated for \( k > 2 \); the iterations using Newton's method are very simple. When applied to the data in Example 1, the following results were obtained

\[
\begin{array}{cccc}
\lambda_1 & 0.5 & 0.45 & 0.55 & 0.55 \\
f(\lambda_1) & -0.03 & -0.5 & -0.0026 & +0.0030
\end{array}
\]
Hence, taking $A = 0.555$ gave the coordinates of $P_{\text{min}}$ as $(0.5705, 0.3782, 0.0513)$, values which are agreeably close to those obtained by the geometric method (Example 1), as also is the minimum sample size here determined as $N > 76.5$.

Although, (18) may be used for $k = 2$ it is simpler to note that, taking $d_2$ to be negative, (16) gives

$$p_1 + p_2 = 1, \quad p_1\sqrt{d_1} = p_2\sqrt{-d_2}$$

so that

$$p_1 = (1 + \sqrt{d_1/-d_2})^{-1}, \quad p_2 = (1 + \sqrt{-d_2/d_1})^{-1}$$

from which $N$ can be calculated via (15) and (8) as before. In essence, although slightly simpler computationally, this is equivalent to the methods given in Mather (1951).

5. Interpretation

With $N$ chosen so that the equal test statistics $\chi^2_{AT}$ and $\chi^2_{BT}$ defined in (2) and evaluated at $P_{\text{min}}$ exceed the 'critical value' $\chi^2_\alpha = \chi^2(k-1;\alpha)$ the procedure is to classify a sample point $P$ with coordinates $(p_1, \ldots, p_k)$ as belonging to population $A$ if

$$\chi^2_{AP} = N\left(\sum_{j=1}^{k} \frac{p_{j}^2}{\pi_j} - 1\right) < \chi^2_{BP} = N\left(\sum_{j=1}^{k} \frac{p_{j}^2}{\pi_{Bj}} - 1\right)$$

and as belonging to population $B$ if $\chi^2_{AP} > \chi^2_{BP}$.

Then, provided:

i) $N$ not only satisfies the foregoing requirement but is also large enough to support the normality approximation and,

ii) it is certain that a sample point must belong to one of the two populations $A$ and $B$,

the probability of misclassification is $\alpha/2$. This follows because, as Mather (1951) noted for the $k = 2$ case, '... deviations in but one of the two possible directions are misleading'; Figure 1 illustrates this case. For $k = 2$ classes, the points $A$, $(\pi_{A1}, \pi_{A2})$ and $B$, $(\pi_{B1}, \pi_{B2})$ lie on the line $\pi_{11} + \pi_{12} = 1$, illustrated in Figure 1, as does the sample point $P$, $(p_1, p_2)$ to be classified. Then, if $P$ belongs to population $A$, for example, and $N$ is large enough, the distance $AP$ will have the Gauss
distribution with mean zero and variance $\pi A_1 \pi A_2 / 2N$. The points $C_1$ and $C_2$ such that $P[A^2 > A C_1^2 = A C_2^2] = P[x^2 > (x^2_1; \alpha)] - \alpha$ can then be located and it is seen that although this probability statement holds for points $P$ which are either to the left of $C_2$ or to the right of $C_1$, the former do not lead to misclassification because $x^2_{TA} < x^2_{TB}$ for such points.

In the general case, the points $A$ and $P$ lie in the hyper-plane $\Sigma$

\[ p_j = 1, \text{ distances } AP \text{ have Gaussian distributions and } x^2_{TA} - \text{ values are equal on hyper-ellipses in the plane. Hence again there are two regions for which } x^2_{TA} > x^2(k-1;\alpha) \text{ but } x^2_{TA} \text{ will exceed } x^2_{TB}, \text{ thus leading to misclassification, in only one of the regions.} \]

Finally it is to be noted that it is the total probability of misclassification which is at most $\alpha/2$ because this probability is

\[ f_1 P[B|A] + f_2 P[A|B] \]

where $P[B|A]$ is the probability of misclassifying a sample from population $A$ into population $B$ and $f_1$ and $f_2$ are the relative frequencies - or probabilities - with which the two, and only two, populations $A$ and $B$ occur so that $f_1 + f_2 = 1$.

6. Conclusions

Although the preceding development importantly depends on the multinomial approximation to Gaussian distribution, it is suggested that the sample sizes needed to control the probability of misclassification will be large enough to sustain the validity of the approximation in many practical cases. On this, one specific criterion, Yarnold (1970), is that the minimum value of $N \pi_j$, $j = 1, \ldots, k$, can be as small as

\[ (5/k)(\text{The number of classes for which } N \pi_j < 5) \]

without vitiating the assumption. Thus, for the situation in Example 1, only one class, that for which $\pi A_3 = 1/36$ would appear to be 'at risk'; here Yarnold’s criterion requires $N$ to exceed $(5/3)(36) = 60$ which, at $N = 77$, it safely does. The value of $N$ does, however, also depend on the prescribed probability of misclassification so that ad hoc examinations can be recommended in some cases and, more generally, to investigate the dependence of $N$ on the positions of, and the divergence between, the vectors $(\pi A_1, \ldots, \pi A_k)$ and $(\pi B_1, \ldots, \pi B_k)$.

Further useful investigation could examine the 'mechanics' of the general solution which involves optimization subject to explicit linear
and non-linear constraints and, less tractably, to the inequalities $p_j \geq 0$, $j = 1, \ldots, k$.

Lastly here it may be noted that by minimizing subject to the more general constraint \( \chi^2_{TA} = C \chi^2_{TB} \), where \( C \) is a selectable constant, the method may be at least approximately extensible to cases for which the misclassification probabilities \( P[B|A] \) and \( P[A|B] \) are unequal. Also feasible, mutatis mutandis is extension to the continuous multivariate cases when all the parameters of the two putative parent populations are known.

Summary

A method – for determining the minimum sample size required to control the probability of misclassifying a sample from one into the other of two multi-discrete-variate populations – is given.

Acknowledgement

It is a pleasure to acknowledge critical encouragement throughout the preparation of this paper from colleague Dr. Edward Pollak who also suggested the genetic application.

References


Figure 1

The probability of misclassification