Group Testing, the Pooled Hypergeometric Distribution, and Estimating the Number of Defectives in Small Populations

C. M. Theobald & A. M. Davie

a School of Mathematics, University of Edinburgh, Edinburgh, Scotland
b Biomathematics and Statistics Scotland, Edinburgh, Scotland

Published online: 13 Jun 2014.

To cite this article: C. M. Theobald & A. M. Davie (2014) Group Testing, the Pooled Hypergeometric Distribution, and Estimating the Number of Defectives in Small Populations, Communications in Statistics - Theory and Methods, 43:14, 3019-3026

To link to this article: http://dx.doi.org/10.1080/03610926.2012.687066
Group Testing, the Pooled Hypergeometric Distribution, and Estimating the Number of Defectives in Small Populations

C. M. THEOBALD¹,² AND A. M. DAVIE¹

¹School of Mathematics, University of Edinburgh, Edinburgh, Scotland
²Biomathematics and Statistics Scotland, Edinburgh, Scotland

The testing of combined bacteriological samples – or “group testing” – was introduced to reduce the cost of identifying defective individuals in populations containing small proportions of defectives. It may also be applied to plants, animals, or food samples to estimate proportions infected, or to accept or reject populations. Given the proportion defective in the population, the number of positive combined samples is approximately binomial when the population is large: we find the exact distribution when groups include the same number of samples. We derive some properties of this distribution, and consider maximum-likelihood and Bayesian estimation of the number defective.

Keywords Bayesian estimation; Combined bacteriological samples; Group testing; Maximum-likelihood estimation; Pooled hypergeometric distribution.

Mathematics Subject Classification 62E15; 62P10.

1. Introduction

Grouping (or pooling) of bacteriological samples was introduced by Dorfman (1943) in order to reduce the cost of testing large numbers of individuals for a disease with low prevalence. Samples taken from several individuals are combined and tested together. Assuming that the diagnostic test is sufficiently sensitive and specific, evidence of a disease or some other defect is provided when and only when at least one of the grouped samples is positive. If the purpose of testing is to identify all the individuals in a population that show signs of the disease, those contributing to grouped samples found to be positive may be retested individually. Dorfman (1943) derived optimum group sizes assuming that the population is large enough for a binomial model to apply, and identified savings of up to 80% in the numbers of tests required for detecting syphilis among US conscripts. Johnson et al. (1991) describe numerous generalizations of Dorfman’s method.
Reduction in cost or effort provides the obvious motive for grouping samples, but other reasons may apply: for example, if bats in a colony are tested for rabies, the amount of blood taken per individual may be too little for a test to be performed on each of them.

Whereas Dorfman (1943) sought to identify all the diseased individuals in a population, grouped tests are also carried out with no intention of identifying individuals or items. The purpose is instead efficient estimation of the prevalence of infection (or more generally the proportion of defective items in a population), or deciding whether to remove a product from the market. For example, individuals need not be identified when farm animals are tested for Salmonella infection. Also when donated human blood is tested for HIV antibodies, there may be no intention to identify the affected donors: indeed, it may be desirable to preserve their anonymity (Gastwirth and Hammick, 1989). Since no retesting is required for estimating the proportion of defectives, there is no need for this proportion to be low.

Chen and Swallow (1990) considered estimation of the proportion of individual samples that are defective when they have been taken at random from a large population and grouped. If \( mk \) individual samples are combined \( k \) at a time to give \( m \) grouped samples, then the number of positive grouped samples has approximately the binomial distribution \( \text{Bi}(m, \theta_k) \), where the probability \( \theta_k \) that a grouped sample is positive is given by

\[
\theta_k = 1 - (1 - \theta)^k,
\]

and \( \theta \) is the probability that an individual sample chosen at random from the population is defective.

Possibly because of the emphasis on large populations and on identifying defective individuals, there appears to have been no consideration in the literature of the distribution of the number of positive grouped samples when the proportion of individuals tested is not small. Large proportions may sometimes be required for testing, though: for example, the regulations in England for taking feces samples from poultry breeding flocks demand large proportions when the number of birds kept in a building is small (Anonymous, 2007).

In the next section, we derive the distribution of the number of positive grouped samples for groups of equal size. We then examine some properties of this distribution, including maximum-likelihood and Bayesian estimation of the number of defective individuals in the population.

We suggest the name “pooled hypergeometric” for this distribution: another possible name is “grouped hypergeometric”, but the various meanings of “grouped” in statistics are more likely to lead to ambiguity.

2. Derivation of the Distribution

Suppose that a population of \( N \) individuals contains a number \( \delta \) that are defective or infected, so that \( \theta \) in (1) equals \( \delta/N \); \( mk \) (\( \leq N \)) individuals are selected at random from the population, and samples from them are combined \( k \) at a time to form \( m \) grouped samples. A grouped sample is assumed to test positive if and only if it includes at least one defective individual sample.

**Theorem 1.** The probability that exactly \( r \) of the \( m \) grouped samples tests positive is given by

\[
f(r \mid \delta) = \binom{m}{r} \sum_{i=0}^{r} (-1)^i \binom{r}{i} \frac{\left( N - \delta \right)}{(m-r+i)k} \bigg/ \binom{N}{(m-r+i)k},
\]

where \( \max[0, m - (N - \delta)/k] \leq r \leq \min(m, \delta) \).
Proof. Suppose that the $m$ grouped samples are labeled $G_1, \ldots, G_m$. The probability $q_t$, say, that a given set of $t$ grouped samples each contain no defectives is the hypergeometric probability that none of its $tk$ members is defective, so that

$$q_t = \binom{N - \delta}{tk} / \binom{N}{tk}.$$ 

Let $A$ denote the event that the $m - r$ grouped samples $G_{r+1}, \ldots, G_m$ contain no defectives and $B_i$ the event that these grouped samples and $G_i$ contain no defectives ($i = 1, \ldots, r$). Then the probability that $G_1, \ldots, G_r$ test positive and $G_{r+1}, \ldots, G_m$ test negative is (using the Inclusion–Exclusion Principle)

$$P(A) - P(B_1 \cup \ldots \cup B_r) = P(A) - \sum_{i=1}^r P(B_i) + \sum_{i<j} P(B_i \cap B_j) - \cdots + (-1)^r P(B_1 \cap \ldots \cap B_r)$$

$$= q_{m-r} - \binom{r}{1} q_{m-r+1} + \binom{r}{2} q_{m-r+2} - \cdots + (-1)^r q_m$$

$$= \sum_{i=0}^r (-1)^i \binom{r}{i} \binom{N - \delta}{i} \binom{N}{m-r+i} / \binom{N}{m-r+i}.$$

The above labeling of the grouped samples is only one of the $\binom{m}{r}$ possible labelings with exactly $r$ positives, so the probability that exactly $r$ of the grouped samples tests positive is given by (2), as required.

3. Some Properties of the Distribution

First note that when $r$ is zero, (2) reduces to the hypergeometric probability $\binom{N - \delta}{mk}/\binom{N}{mk}$, since the absence of defectives among the $mk$ individual samples is equivalent to the absence of positive grouped samples. Also setting $k$ to 1 provides a combinatorial identity for the hypergeometric probability $\binom{m}{r} (\binom{N - \delta}{r}) / \binom{N}{r}$.

**Theorem 2.** If $p_u$ denotes $\binom{N - \delta}{u}/\binom{N}{u}$ ($u = 0, 1, \ldots, N - \delta$), then the first and second moments of the distribution defined by Equation (2) are

$$E(X) = mp_k = m \binom{N - \delta}{k} / \binom{N}{k},$$

$$E(X^2) = mp_k + m(m-1) p_{2k},$$

$$Var(X) = mp_k + m(m-1) p_{2k} - m^2 p_k^2$$

$$= m \binom{N - \delta}{k} / \binom{N}{k} + m(m-1) \binom{N - \delta}{2k} / \binom{N}{2k} - m^2 \left[ \binom{N - \delta}{k} / \binom{N}{k} \right]^2.$$

Proof. Define random variables $X_i$ equal to 1 if the $i$th grouped sample contains no defective and 0 otherwise ($i = 1, \ldots, m$), and let $X$ denote their sum (so that $f(r \mid \delta) =$
Theorem 3. If $N$ and $\delta$ tend to $\infty$ in such a way that $\delta/N$ tends to $\theta$, then the pooled hypergeometric distribution defined by Equation (2) tends to the binomial distribution $Bi(m, \theta_k)$. Also the moments given in Theorem 2 tend to those of $Bi(m, \theta_k)$.

Proof. Since $p_u$ in Theorem 2 is the product of terms $\frac{N-\delta-v}{N-v}$, $(v = 0, 1, \ldots, u - 1)$, it tends to $(1 - \theta)^u$. Also $f(r \mid \delta)$ equals $(\binom{m}{r}) \sum_i (-1)^i \binom{r}{i} p_{(m-r+i)k}$, whose limit as $N \to \infty$ is

\[
\binom{m}{r} \sum_{i=0}^{r} (-1)^i \binom{r}{i} (1 - \theta_k)^{m-r+i} = \binom{m}{r} (1 - \theta_k)^{m-r} \sum_{i=0}^{r} \binom{r}{i} (-1)^i (1 - \theta_k)^i
\]

\[
= \binom{m}{r} (1 - \theta_k)^{m-r} \theta_k^r.
\]

Since the limits of $p_k$ and $p_{2k}$ are $1 - \theta_k$ and $(1 - \theta_k)^2$, respectively, those of $E(X)$ and $\text{Var}(X)$ are $m (1 - \theta_k)$ and $m \theta_k (1 - \theta_k)$, respectively, in agreement with the moments of the binomial. □

Figure 1 compares the pooled hypergeometric probability function (2) with the binomial approximation based on Eq. (1) for grouped samples of size 5 and population sizes $N$ equal to 40, 80, 160, and 320; within each panel, the proportions $\theta$ of defective individual samples are 0.05, 0.1, 0.2, and 0.4. The numbers of grouped samples illustrated for these four population sizes are, respectively, 6, 8, 10, and 11; these numbers and the group size are those specified for bacteriological testing of feces samples from poultry breeding flocks for Salmonella in England (Anonymous, 2007). The method for calculating the numbers of grouped samples to be taken is not stated, but can be interpreted as ensuring that the probability of finding no positive grouped sample is about 0.05 when the proportion of samples that are positive in the population is also 0.05.

The binomial approximation is seen to be poor for population size 40, and better for defective proportion 0.4 than for the lower proportions.

4. Inference About the Number of Defectives in the Population

We consider inferring the number $\delta$ and proportion $\theta$ of defective individuals in the population from the number of grouped samples testing positive.

The method of maximum likelihood may be applied under the pooled hypergeometric distribution, but the standard asymptotic properties of the method do not apply because the parameter space is discrete. If no grouped sample tests positive, then the likelihood has a single maximum at zero since $f(0 \mid 0) = 1$, but $f(0 \mid \delta) < 1$ for positive $\delta$. Also $f(m \mid \delta) = 1$ for $\delta > N - k$, but $f(m \mid \delta) < 1$ for $\delta \leq N - k$, so that if all $m$ grouped samples are positive then the likelihood is maximized at the largest $k$ values of $\delta$.

Figure 2 shows the possible exact and approximate likelihood functions when $N$ is equal to 40 and 6 grouped samples of size 5 are formed. It suggests that – unless all the grouped samples are positive – any likelihood-based confidence regions for $\delta$ are smaller under the exact distribution than under the binomial approximation. The latter regions can even include values of $\delta$ that are less than the number of positive grouped samples.
Figure 1. Pooled hypergeometric probability functions (thick lines) and their binomial approximations (thin, dashed lines) for population sizes 40, 80, 160, and 320 and grouped samples of size 5. Within each panel, the proportions $\theta$ of defective individual samples are 0.05, 0.1, 0.2, and 0.4.

Under the binomial approximation to the distribution of the number of positive grouped samples, the maximum-likelihood estimate $\hat{\theta}$ of the proportion of defective individuals is found by equating $\theta_k$ in (1) to the proportion $r/m$ of grouped samples that are positive: thus the number defective in the population is estimated by

$$\hat{\delta} = N\hat{\theta} = N[1 - (1 - r/m)^{1/k}].$$  

(3)

When all $m$ grouped samples are positive, the estimate in (3) equals $N$, although as few as $m$ of the $mk$ individual samples may be defective. Estimates of $\delta$ maximizing the exact and approximate likelihoods may both be regarded as unsatisfactory: the exact distribution fails to provide a unique estimate when all grouped samples are positive, and both estimated proportions may greatly exceed the true proportion of defectives in the population. Richards (1991) suggests an ad hoc adjustment to the binomial estimate in this case.

The specification of a sampling scheme for detecting defective samples requires some knowledge of the proportion of defectives to be expected in the relevant populations. Given that such knowledge is available, a Bayesian approach might be taken to estimating the number of defectives in the population. We therefore consider using the posterior expected
values $E(\delta | r)$ to estimate $\delta$ under various prior distributions for the number of defectives. The prior distribution for $\delta$ in a population of size $N$ should reflect knowledge of the numbers typically occurring in such populations and their relation to $N$. A very simple model for the prior probability function $p_N(\delta)$ supposes that the $N$ individuals in a population form a random selection from the individuals of that type in some geographical region, leading to a binomial prior distribution $\text{Bi}(N, p)$, where $p$ denotes the overall proportion of defectives in the region. To account for additional variation between populations in the proportions defective, a more realistic model allows the probability that an individual is defective to vary between populations. Assuming a beta distribution for $p$ results in the beta-binomial distribution $\text{BeBi}(N, a, b)$ with expectation $Na/(a + b)$ for the number of defectives in a population of size $N$. If $a$ and $b$ are less than 1, the distribution has modes at 0 and $N$: such highly dispersed prior distributions may be appropriate for animal studies in which the population comprises individuals in close contact.

Table 1 shows the maximum-likelihood estimates of the numbers of defective individuals given the numbers of positive grouped samples, based on the pooled hypergeometric distribution of (2) and the binomial approximation (3) for a population of size 40 when 30 samples are combined into 6 groups of 5. Note that if all the grouped samples are positive, then the 5 largest values of $\delta$ all maximize the likelihood.

In Table 1, the maximum-likelihood estimates are compared with the posterior expected values of the number of defectives assuming a pooled hypergeometric distribution and beta-binomial prior distributions. The prior distributions in rows 3 to 5 have common expectation 0.05, standard deviations 8.0, 5.2, and 2.3 and probabilities 0.91, 0.72, and 0.32 that no individual in the population is defective; those in rows 6 to 8 have common expectation 0.01, standard deviations 2.8, 1.3, and 0.7 and probabilities 0.96, 0.85, and 0.71 of no defective in the population.
Table 1

Maximum-likelihood estimates (MLE) of the number defective $\delta$ in a population of size 40 based on 6 grouped samples of 5 each, assuming a pooled hypergeometric distribution and a binomial approximation, along with posterior expectations (PE) of $\delta$ assuming a pooled hypergeometric distribution and beta-binomial prior distributions, three with expected proportion 0.05 and three with expected proportion 0.01

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>Prior distribution</th>
<th>Number of positive grouped samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>MLE</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Binomial MLE</td>
<td>—</td>
<td>0.00</td>
</tr>
<tr>
<td>PE</td>
<td>BeBi(40, 0.01, 0.19)</td>
<td>0.00</td>
</tr>
<tr>
<td>PE</td>
<td>BeBi(40, 0.1, 1.9)</td>
<td>0.03</td>
</tr>
<tr>
<td>PE</td>
<td>BeBi(40, 0.1, 1.9)</td>
<td>0.20</td>
</tr>
<tr>
<td>PE</td>
<td>BeBi(40, 0.01, 0.99)</td>
<td>0.00</td>
</tr>
<tr>
<td>PE</td>
<td>BeBi(40, 0.1, 0.9)</td>
<td>0.02</td>
</tr>
<tr>
<td>PE</td>
<td>BeBi(40, 0.1, 1.9)</td>
<td>0.08</td>
</tr>
</tbody>
</table>

The maximum-likelihood estimates for the pooled hypergeometric distribution are necessarily integer-valued: here most of them equal the rounded values of the binomial estimates.

The posterior expectations based on the beta-binomial prior distributions in Table 1 differ substantially from each other and from the maximum-likelihood estimates when the number of positive grouped samples is large, reflecting the flatness of the likelihood function for large proportions of defectives in these cases. Thus, it appears worthwhile to incorporate a prior distribution into the analysis, provided that this distribution is carefully chosen.

More complex inferences about numbers of defectives based on grouped sampling, such as comparisons between populations, are complicated by the discreteness of the parameter space, since this prevents generalized linear models being applied directly to these numbers. Instead the suggestion made above that $\delta$ be treated as binomial for given $N$ might be extended to a hierarchical model that assumes a joint normal distribution for the logits of the probabilities $p_i$ that individuals in population $i$ are defective.

Funding

The work of CMT was supported by the Scottish Government.

References


