Wiley StatsRef:
Statistics Reference Online

Group Testing for Estimation

By *Christopher R. Bilder*

Keywords: *binary response, infectious disease, pooled testing, R, regression, screening, sensitivity, specificity, testing*

Abstract: Group (pooled) testing involves combining a set number of items together and performing one test for a binary trait (e.g., positive or negative for a disease) on that combination. While testing is performed on the group, the purpose is still to understand this trait relative to each individual item. When compared to testing a large number of items separately, testing instead via groups can result in a significant reduction in the total number of tests, provided that the prevalence of a particular level of the trait (e.g., positive for a disease) is small. Furthermore, group testing can lead to more efficient estimators than individual testing if testing error is possible. The purpose of this article is to introduce how estimation can be performed with data arising through group testing. Focus is on estimating the probability an item has a particular level of the trait rather than determining the trait status of individual items.

1 Introduction

Estimating a probability associated with a binary trait, such as the probability an individual is positive or negative for a disease, is of interest in a wide variety of applications. Unfortunately, there are many situations when estimation is difficult, and perhaps even impossible, due to the cost and time associated with testing for the trait. When the overall prevalence for the binary level of interest is small, a process known as *group testing* (also known as *pooled testing* or simply as *pooling*) can make the impossible possible. One of the largest applications of group testing in the world is the screening of blood donations by the American Red Cross (ARC). The ARC receives millions of donations per year and each donation needs to be declared disease free to prevent the spread of disease from donor to recipient. To handle its high volume of clinical specimens, the ARC pools portions of specimens from 16 donors and performs one test upon this pooled specimen[1, 2].[1] If the test is negative for a disease, the corresponding donations represented within the group can be considered free of that particular disease. If the test is positive for a disease, the remaining portions of each specimen are retested separately to determine who is positive and who is negative for that disease. By applying this process across all blood donations, the ARC significantly reduces its testing load because most groups will test negatively due to the low disease prevalence. Testing items in groups rather than separately is used in a diverse set of other applications, including infectious disease detection

University of Nebraska-Lincoln, Lincoln, NE, USA

DOI: 10.1002/9781118445112.stat08231

in animals[3], determining virus transmission from an insect to a plant[4], bacteria screening for food[5], discovery of new pharmaceuticals[6], and verification of computer network security[7]. Due to this diversity, we will use "positive" or "negative" in this article as the outcomes associated with testing for a binary trait, where the positive outcome has a small overall prevalence. For example, this would coincide with HIV testing to determine which individuals are positive or negative for that infection.

The group testing research problem is divided into two separate areas: identification and estimation. The identification area involves determining what items are positive or negative. Items are first tested within groups and retests are performed in an algorithmic manner as needed to decode the positives from the negatives. The algorithm described for the ARC example is often referred to as Dorfman's algorithm in honor of Robert Dorfman who was one of the originators of the group testing idea[8–10]. Other algorithms exist and can result in a smaller number of tests. Please see our companion article of Bilder[11] for more information.

We focus in this article on the estimation area. The goal for estimation is to estimate an overall prevalence of positivity or, equivalently, the probability a randomly selected item from a homogeneous population is positive. More generally, regression models can be used to estimate this probability when information on factors that may affect this probability are available. By using the group outcomes alone, estimation and associated inference procedures can be performed. Frequently, retests involving members of positive groups are available as well because identification is a simultaneous interest. When testing error (false positives or false negatives) is possible and the identification of positive/negative individuals is performed, group testing can result in estimators with similar or even smaller variances than if every item was tested separately at the start. In other words, a smaller sample size (fewer tests) can lead to more efficient parameter estimators than would be obtained from individual testing.

The order of our article is as follows. Section 2 introduces the estimation area for when only the overall prevalence is of interest. This involves observing group outcomes alone without any retesting. Section 3 presents estimation in the context of regression models. For this situation, a much more general setting is described with regard to retesting items within positive groups and the possibility of testing error. Section 4 provides examples of how to implement the methods described in Sections 2 and 3. Finally, Section 5 summarizes and introduces readers to additional topics involving group testing.

2 Prevalence Estimation

The simplest application of group testing involves testing items in nonoverlapping groups without any subsequent retesting. We will focus on how plant pathologists use this type of application to estimate the probability that a randomly selected insect vector infects a plant with a disease (or the pathogen that leads to a disease). Because the disease transmission rate can be small, a one insect per plant experimental design is most often not a feasible way to estimate this probability. Instead, multiple vectors can be transferred to each plant to increase the probability of disease acquisition per plant. Figure 1 illustrates how this type of design can be implemented inside a greenhouse. In the context of group testing, the plant is the group and the individual items are the insects. After observing the plants for a period of time, the positive or negative disease status can be determined for each plant, often without the possibility of testing error.[2] Using the probability of infection relationship between plants and individual vectors (to be discussed shortly), one can estimate the probability of interest.

Define $\tilde{Z}_k$ as a binary random variable representing the true disease status for the kth plant, where $k = 1, \ldots K$. We will always denote binary response values as a 1 to represent positive and a 0 to represent negative. The probability that the kth plant is positive is $P(\tilde{Z}_k = 1) = \tilde{\theta}_k$. Each plant status is a function of whether or not an individual insect vector infects the plant. Therefore, define $\tilde{Y}_{i(k)}$ as a binary random

Wiley StatsRef: Statistics Reference Online, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08231

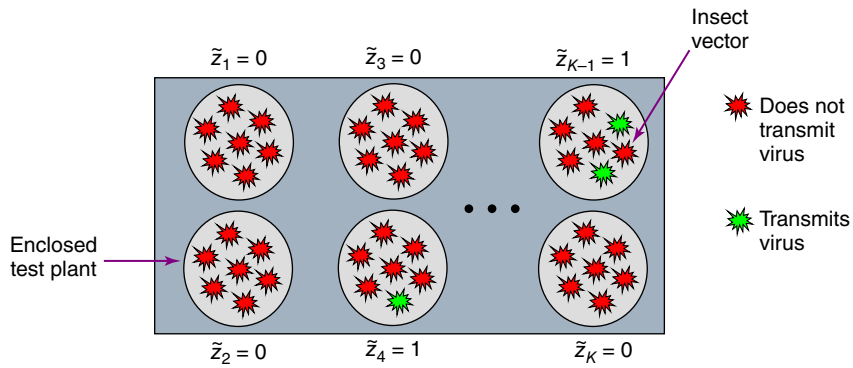

Figure 1. A greenhouse configuration for a multiple vector transfer design experiment. *Source:* Reproduced by permission of Christopher R. Bilder.

variable for the true positive (disease transmitted) and negative (disease not transmitted) status for insect i on the kth plant, where $i = 1, \ldots, I_k$. The probability that insect i on the kth plant infects a plant is $P(\tilde{Y}_{i(k)} = 1) = \tilde{\pi}$. We assume that the insect vector statuses are independent and identically distributed.

The relationship between the group and individual statuses is given by $\tilde{Z}_k = I\left(\sum_{i=1}^{I_k} \tilde{Y}_{i(k)} \geq 1\right)$, where $I(\cdot)$ is the indicator function. Similarly, the relationship between the corresponding probabilities is

$$\begin{aligned}\tilde{\theta}_k &= 1 - P(\tilde{Z}_k = 0) \\ &= 1 - P(Y_{1(k)} = 0, \ldots, Y_{I_k(k)} = 0) \\ &= 1 - (1 - \tilde{\pi})^{I_k} \end{aligned} \tag{1}$$

The likelihood function for $\tilde{\pi}$ is the product of Bernoulli distributions

$$L(\tilde{\pi}|\tilde{z}_1, \ldots, \tilde{z}_K) = \prod_{k=1}^{K} [1 - (1 - \tilde{\pi})^{I_k}]^{\tilde{z}_k} (1 - \tilde{\pi})^{I_k(1-\tilde{z}_k)}$$

where $\tilde{z}_1, \ldots, \tilde{z}_K$ are the observed plant responses. The maximum likelihood estimator (MLE) for $\tilde{\pi}$, $\hat{\tilde{\pi}}$, is obtained through using numerical iterative methods. When each plant has the same number of insect vectors (equal group sizes), say I, so that $\tilde{\theta} = 1 - (1 - \tilde{\pi})^I$, the closed-form expression for the MLE is

$$\hat{\tilde{\pi}} = 1 - (1 - \hat{\tilde{\theta}})^{1/I} \tag{2}$$

where $\hat{\tilde{\theta}} = \sum_{k=1}^{K} Z_k / K$ is the observed proportion of positive plants.

Unfortunately, the MLE for $\tilde{\pi}$ is biased for a fixed sample size. There have been a number of other estimators proposed to reduce this bias. For the equal group size case, a frequently used alternative was given by Burrows[12] as

$$\hat{\tilde{\pi}}_B = 1 - \left[1 - \frac{\sum_{k=1}^{K} Z_k}{K + b}\right]^{1/I}$$

where $b = (I - 1)/(2I)$. Others have approached estimation through a Bayesian formulation. For example, Bilder and Tebbs[13] proposed an empirical Bayesian motivated estimator of

$$\hat{\tilde{\pi}}_{\text{EB}} = 1 - \left[1 - \frac{1 + \sum_{k=1}^{K} Z_k}{K + 1 + \hat{\delta}/I}\right]^{1/I}$$

DOI: 10.1002/9781118445112.stat08231

where $\hat{\delta}$ is the marginal maximum likelihood estimator for the distribution of $\sum_{k=1}^{K} Z_k$ when using a beta$(1, \delta)$ prior for $\tilde{\pi}$. Both $\hat{\tilde{\pi}}_B$ and $\hat{\tilde{\pi}}_{\text{EB}}$ can have substantially less bias than $\hat{\tilde{\pi}}$. For the unequal group size case, Hepworth and Biggerstaff[14] proposed a numerical iterative method based on the work of Firth[15] for general maximum likelihood estimation problems. In summary, the expression of

$$S(\tilde{\pi}) - F(\tilde{\pi})b(\tilde{\pi}) = 0$$

is solved for $\tilde{\pi}$, where $S(\tilde{\pi})$ is the score function, $F(\tilde{\pi})$ is the Fisher information matrix, and $b(\tilde{\pi})$ is an approximate bias of the MLE.

A Wald confidence interval for $\tilde{\pi}$ is

$$\hat{\tilde{\pi}} - Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\tilde{\pi}})} < \tilde{\pi} < \hat{\tilde{\pi}} + Z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\tilde{\pi}})}$$

where $\widehat{\text{Var}}(\hat{\tilde{\pi}}) = \left\{ \sum_{k=1}^{K} I_k^2(1-\hat{\tilde{\pi}})^{I_k-2}/[1-(1-\hat{\tilde{\pi}})^{I_k}] \right\}^{-1}$ and $Z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from a standard normal distribution. Unfortunately, this interval can have rather poor coverage properties[16, 17]. For equal group sizes, Tebbs and Bilder[16] show that intervals formed for $\tilde{\theta}$ and then transformed to the $\tilde{\pi}$-scale using the inverse of Equation (1) can greatly improve coverage. For example, a score-based interval would be

$$1-(1-\tilde{\theta}_L)^{1/I} < \tilde{\pi} < 1-(1-\tilde{\theta}_U)^{1/I}$$

where $\tilde{\theta}_L$ and $\tilde{\theta}_U$ are the lower and upper bounds, respectively, for a score interval (also known as a *Wilson interval*) for $\tilde{\theta}$ (see Section 1.1 of Bilder and Loughin[18]). The unequal group size situation is a little more complex because the relationship between $\tilde{\theta}_k$ and $\tilde{\pi}$ is not only one monotone transformation for all k. Hepworth[19] proposed a score interval that also included a correction for the skewness $\gamma(\tilde{\pi})$. Thus, solve for $\tilde{\pi}$ in

$$S(\tilde{\pi}) - \frac{1}{6}\gamma(\tilde{\pi})(Z_{1-\alpha/2}^2 - 1) = \pm Z_{1-\alpha/2}$$

once using $+Z_{1-\alpha/2}$ and once using $-Z_{1-\alpha/2}$ on the right side of the expression. Other intervals were also examined by Hepworth[19] and Biggerstaff[17], but generally this interval performed at least similar to or better than others.

3 Regression Models

The goal for this section is to estimate the probability an individual item is truly positive given a set of covariates. For example, we may want to determine what factors (e.g., risky behavior, clinical observations) are related to the probability an individual has a sexually transmitted disease. In comparison to the previous section, we generalize our discussion to allow for retesting items in positive-testing groups. Because there are many retesting algorithms, we focus on the algorithm most often used – retest separately those items within these groups (i.e., Dorfman's algorithm). Also, we generalize to allow for testing error. Thus, false-positive or false-negative test outcomes may occur.

Define $\tilde{Y}_{i(k)}$ again as a binary random variable representing the true positive/negative status for item i in the kth group with $i = 1, \ldots, I_k$ and $k = 1, \ldots, K$. The associated probability of being truly positive is now defined as $P(\tilde{Y}_{i(k)} = 1) = \tilde{\pi}_{i(k)}$. We relate these probabilities to a function of covariates $\mathbf{x}_{i(k)} = (1, x_{i(k)1}, \ldots, x_{i(k)p})$. The function class that we focus on here involves a monotone, differentiable link function $g(\cdot)$, such as the logit link function, and set $\tilde{\pi}_{i(k)} = g^{-1}(\mathbf{x}_{i(k)}\boldsymbol{\beta})$ for a vector of parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$. The likelihood function for $\boldsymbol{\beta}$ is the product of Bernoulli probability mass functions:

Wiley StatsRef: Statistics Reference Online, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08231

$$L(\boldsymbol{\beta}|\tilde{\mathbf{y}}) = \prod_{k=1}^{K}\prod_{i=1}^{I_k} \tilde{\pi}_{i(k)}^{\tilde{y}_{i(k)}}(1-\tilde{\pi}_{i(k)})^{1-\tilde{y}_{i(k)}} \tag{3}$$

where $\tilde{\mathbf{y}} = (\tilde{y}_{1(1)}, \ldots, \tilde{y}_{I_K(K)})$. Because of potential testing error, we cannot observe $\tilde{\mathbf{y}}$, which makes this a nonstandard estimation and inference problem.

What we observe are the group test and individual retest outcomes. Define Z_k as the binary outcome for group k. Similarly, define $Y_{i(k)}$ as the binary retest outcome for item i within group k. Accuracy for true positives is measured by the sensitivity $S_{e,z} = P(Z_k = 1|\tilde{Z}_k = 1)$ for a group test, where $\tilde{Z}_k$ denotes the true binary status again for group k, and the sensitivity $S_{e,y} = P(Y_{i(k)} = 1|\tilde{Y}_{i(k)} = 1)$ for an individual retest. Accuracy for true negatives is defined analogously as $S_{p,z} = P(Z_k = 0|\tilde{Z}_k = 0)$ and $S_{p,y} = P(Y_{i(k)} = 0|\tilde{Y}_{i(k)} = 0)$ to represent the specificity for the group tests and individual retests, respectively. For now, we treat these accuracy measures as equal across all individual retests and across all group tests, despite potentially different group sizes. Also, we assume values for these measures as being known. This latter assumption can be reasonable for disease-testing applications by using the observed accuracy rates from large clinical trials as the actual conditional probabilities. Alternatives to these assumptions are discussed later in this section.

Xie[20] proposed to maximize Equation (3) by using the expectation-maximization (EM) algorithm. The function that is maximized becomes

$$E[\log(L(\boldsymbol{\beta}|\tilde{\mathbf{Y}}))|\mathcal{I}] = \sum_{k=1}^{K}\sum_{i=1}^{I_k} \omega_{i(k)}\log(\tilde{\pi}_{i(k)}) + (1-\omega_{i(k)})\log(1-\tilde{\pi}_{i(k)})$$

where $\mathcal{I}$ represents the information observed from the group tests and individual retests and $\omega_{i(k)} = E(\tilde{Y}_{i(k)}|\mathcal{I})$. Closed-form expressions for $\omega_{i(k)}$ can be derived. For example, when group k tests negatively, we have $\omega_{i(k)} = P(\tilde{Y}_{i(k)} = 1|Z_k = 0) = (1 - S_{e,z})\tilde{\pi}_{i(k)}/(1-\theta_k)$, where

$$\theta_k = S_{e,z} + (1 - S_{p,z} - S_{e,z})\prod_{i=1}^{I_k}(1-\tilde{\pi}_{i(k)})$$

Maximum likelihood estimates for $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, are achieved at convergence of the EM algorithm, and their standard errors follow from the methods of Louis[21].

There has been a considerable amount of research in this area since Xie[20]. In particular, Zhang *et al.*[22] derived $\omega_{i(k)}$ when retesting is performed by halving positive-testing groups over multiple retesting stages. For array testing, closed-form expressions for $\omega_{i(k)}$ are not possible, so Zhang *et al.*[22] developed a Gibbs sampling algorithm to estimate $\omega_{i(k)}$. McMahan *et al.*[23] generalized previously proposed Bayesian approaches to allow for retests. One advantage of a Bayesian approach is that priors can be placed on the sensitivity and specificity parameters so that specific values are not assumed. In particular, these priors can be made informative by using data obtained during test calibration, such as when clinical trials are performed by diagnostic testing companies.

Others have approached challenges caused by testing error through allowing for a potential dilution effect (i.e., sensitivity may decrease as the number of group members increase). For example, Delaigle and Hall[24] assumed that an underlying continuous response is observed for a test outcome, such as an optical density reading for an enzyme-linked immunosorbent assay, and subsequently turned this measurement into a positive/negative result by using a threshold. With this information, they estimate densities for the continuous response given the true binary status of an individual/group and use these densities within a nonparametric regression framework to estimate $\tilde{\pi}_{i(k)}$. This paper was awarded the George W. Snedecor Award in 2017 for its important contribution to biometric research. In other applications, the underlying continuous response may not be available. For this reason, Warasi *et al.*[25]

DOI: 10.1002/9781118445112.stat08231

proposed a submodel for the sensitivity that is included within the regression model for $\tilde{\pi}_{i(k)}$. This submodel allows the sensitivity to change as a function of the number of items being pooled together. The innovative aspect of this work was equating the sensitivity of a group with all positive items to having essentially the same sensitivity as when testing an individual item.

Group composition and whether retests are performed play an important role in estimator precision and agreement when compared to individual testing. When only group responses are observed, Vansteelandt *et al.*[26] showed that creating groups consisting of as alike covariate values as possible will lead to the smallest variances for estimators. Bilder and Tebbs[27] showed the agreement of estimators is best in the alike covariate situation as well, but also showed there are benefits from other group compositions. This is important because groups most often cannot be constructed in this alike manner. When retests for identification purposes are included with the data, Zhang *et al.*[22], McMahan *et al.*[23], and others have shown that the variance of estimators can be approximately the same or even smaller than those from individual testing. This result is quite remarkable because group testing will have fewer tests (i.e., smaller sample size) than individual testing in appropriately applied settings. Thus, a smaller sample size can lead to more efficient estimators.

4 Examples

Estimation is performed by a number of functions in the `binGroup` package of R[28]. In particular, the `pooledBin()` function finds estimates of $\tilde{\pi}$ along with corresponding confidence intervals. To illustrate this function, consider the multiple vector transfer design experiment performed by Gildow *et al.*[29]. The purpose of this experiment was to estimate the probability that particular species of aphids would transfer the Cucumber Mosaic virus to snap bean plants. For the *Aphis glycines* species, there were 50 plants with 10 aphid vectors placed upon each plant for a 16-hour time period. The plants were subsequently monitored over a four-week time period, and 30 plants were found to be infected.

```
> # MLE
> 1 - (1 - 30/50)^(1/10)
[1] 0.08755646

> library(package = binGroup)

> # Estimates and confidence intervals
> pooledBin(x = 30, m = 10, n = 50, alpha = 0.05, pt.method = "mle",
    ci.method = "wald")
PointEst    Lower    Upper
  0.0876   0.0566   0.1185
> pooledBin(x = 30, m = 10, n = 50, alpha = 0.05, pt.method = "firth",
    ci.method = "skew-score")
PointEst    Lower    Upper
  0.0863   0.0599   0.1220
```

The maximum likelihood estimate for the probability of infection is $\hat{\tilde{\pi}} = 0.0876$. The `pooledBin()` function allows for other estimation approaches, including the bias reduction method of Firth[15] that results in an estimate of 0.0863. The 95% score interval with a skewness correction is $0.0599 < \tilde{\pi} < 0.1220$. When unequal group sizes are used, vectors containing the number of positive groups, the group

Wiley StatsRef: Statistics Reference Online, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08231

sizes, and the number of groups for each unique group size can be included in the `x`, `m`, and `n` arguments of `pooledBin()`, respectively. For example, `pooledBin(x = c(4,7,4,8,7), m = rep(10, times = 5), n = rep(10, times = 5), alpha = 0.05, pt.method = "firth", ci.method = "skew-score")` will result in the same estimate and interval as provided in the last example within the output.

The `gtreg()` function in `binGroup` is used to estimate a regression model with the group tests only or with the group tests and subsequent retests. We demonstrate its use by fitting a model to the `hivsurv` data in `binGroup` to estimate the probability of HIV infection among pregnant women from a region of Kenya[26, 30]. The purpose of the associated study was to compare estimates obtained by both group and individual testing. The `HIV` variable in `hivsurv` gives the individual test outcomes. Unfortunately, group outcomes were not available, so the data set provides outcomes from artificially created groups. The variable `groupres` contains these outcomes with the values repeated over individuals within the same group. Also, retests were not performed in the study, so we also artificially create them here for demonstration purposes. We estimate a model of the form $\text{logit}(\tilde{\pi}_{i(k)}) = \beta_0 + \beta_1 x_{i(k)}$, where $x_{i(k)}$ represents the education level of an individual (`EDUC.`; coded as 1, 2, 3, and 4 to represent more education as having higher values).

```
> # Artificially create retests
> sens.ind <- 0.99  # Sensitivity of retest
> spec.ind <- 0.99  # Specificity of retest
> set.seed(7342)
> hivsurv$retest <- rep(x = NA, times = nrow(hivsurv))
> for(a in 1:nrow(hivsurv)) {
   # Simulate retest by conditioning on observed individual outcome
   if(hivsurv$groupres[a] == 1 & hivsurv$HIV[a] == 1)
     hivsurv$retest[a] <- rbinom(n = 1, size = 1, prob = sens.ind)
   if(hivsurv$groupres[a] == 1 & hivsurv$HIV[a] == 0)
     hivsurv$retest[a] <- rbinom(n = 1, size = 1, prob = 1 - spec.ind)
  }

> # A negative group (group #1)
> hivsurv[hivsurv$gnum == 1, c("HIV", "EDUC.", "gnum", "groupres",
   "retest")]
  HIV EDUC. gnum groupres retest
1   0     4    1        0     NA
2   0     2    1        0     NA
3   0     1    1        0     NA
4   0     2    1        0     NA
5   0     1    1        0     NA

> # A positive group (group #3)
> hivsurv[hivsurv$gnum == 3, c("HIV", "EDUC.", "gnum", "groupres",
    "retest")]
   HIV EDUC. gnum groupres retest
11   0     3    3        1      0
12   1     3    3        1      1
```


DOI: 10.1002/9781118445112.stat08231

```
13   0     3    3        1      0
14   1     2    3        1      1
15   0     3    3        1      0

> # Estimate model with group tests
> fit.group <- gtreg(formula = groupres ~ EDUC., data = hivsurv, groupn =
    gnum, sens = 0.99, spec = 0.99, linkf = "logit", method = "Xie")
> round(summary(fit.group)$coefficients, 4)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7368     0.9812 -3.8085   0.0001
EDUC.         0.5857     0.3838  1.5259   0.1270

> # Estimate model with group tests and individual retests
> fit.group.retest <- gtreg(formula = groupres ~ EDUC., data = hivsurv,
    groupn = gnum, sens = 0.99, spec = 0.99, linkf = "logit",
    method = "Xie", retest = retest, sens.ind = sens.ind,
    spec.ind = spec.ind)
> round(summary(fit.group.retest)$coefficients, 4)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7493     0.5731 -6.5420   0.0000
EDUC.         0.5556     0.2178  2.5508   0.0107

> # Estimate model with individual tests
> fit.ind <- glm(formula = HIV ~ EDUC., data = hivsurv,
    family = binomial(link = logit))
> round(summary(fit.ind)$coefficients, 4)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8853      0.570 -6.8164   0.0000
EDUC.         0.6239      0.214  2.9149   0.0036

> nrow(hivsurv)  # Number of individual tests
[1] 428
> max(hivsurv$gnum)  # Number of group tests
[1] 86
> max(hivsurv$gnum) + sum(!is.na(hivsurv$retest))
           # Number of group tests and retests
[1] 241
```

All three estimation approaches result in similar estimates of the regression parameters. The standard errors increase by about 75% when using the group tests only as compared to when using the individual tests. However, the estimation from the group tests used only 86/428 = 20% of the tests that individual testing required. When retests are included with group tests, the standard errors become quite close to those from the individual tests despite using approximately half as many tests. Thus, essentially the same amount of information for the study would be obtained by using only the group tests and retests but likely at a significantly lower cost and with a reduction in testing time.

The `binGroup` package also provides functions to estimate regression models using the halving (`gtreg.halving`) and the array testing (`gtreg.mp`) algorithms for group testing. Bayesian methods

Wiley StatsRef: Statistics Reference Online, © 2014–2019 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08231

for estimation and methods to take into account a dilution effect are not available in `binGroup`, but R code often can be found in the supplementary materials of their corresponding research articles.

5 Conclusion

In comparison to testing items individually, group testing produces similar estimates for probabilities associated with a binary trait, while also significantly reducing the number of tests. Group testing can also lead to a smaller variance for estimators when both identifying the positive/negative status for each item is of interest and testing error is possible.

Because groups are chosen prior to implementation, group members and overall group sizes can be chosen to be optimal in some manner. Section 3 discusses the work of Vansteelandt *et al.*[26] regarding how to make estimator variances for regression as small as possible. With respect to estimating an overall prevalence, Swallow[31] examined the optimal number of insect vectors per plant relative to minimizing the mean square error of $\hat{\hat{\pi}}$ (method available in the `estDesign()` function of the `binGroup` package). When identification of positive/negative items is of interest as well, how to choose group sizes is not necessarily as clear. Bilder[11] discusses how to choose groups relative to minimizing the expected number of tests per individual for the identification problem alone. Whether these optimal groups are similar for the estimation problem is unknown. And, if they are not similar, determining how to balance optimality for estimation and identification could be an area for future research.

The outcome for group testing applications does not need to be of a Bernoulli response form. Xie *et al.*[32] examine how to estimate probabilities associated with a three-category response. This type of situation occurs in pharmaceutical drug discovery when finding a potent (positive) chemical compound is of interest, but there are also negative compounds and blocker compounds (i.e., those compounds that block a positive group outcome from being observed). Other researchers have examined estimation in the context of testing for multiple diseases simultaneously, such as with a multiplex assay for infectious disease testing. Hughes-Oliver and Rosenberger[33] found optimal group sizes based on the D-optimality criteria when there are no retests and no testing error. Zhang *et al.*[34] developed an expectation-solution algorithm to estimate regression models for the multiple disease situation when testing error was possible. Finally, a number of researchers have used a negative binomial approach when groups are tested until a fixed number of groups are observed to be positive. Similar to the Firth-based method given in Section 2, Hepworth[35] proposed an estimator that greatly reduces the bias associated with estimating the prevalence.

Acknowledgment

This research was supported in part by Grant R01 AI121351 from the National Institutes of Health.

End Notes

1. The ARC tests for HIV, hepatitis B, hepatitis C, and West Nile virus in groups. Testing for West Nile virus is performed individually in areas when an outbreak occurs. Testing for other diseases, such as syphilis, is performed on individual specimens.
2. Diagnostic tests, such as a nucleic acid amplification test, may not be necessary. After removal of insect vectors, plants can be observed for a sufficient period of time to make an accurate determination of disease status.

DOI: 10.1002/9781118445112.stat08231

Related Articles

Dorfman-Type Screening Procedures; **EM Algorithm**; **Gibbs Sampling**; **Group Testing for Identification**; **Proportions, Inferences and Comparisons**; **Statistics in Drug Discovery**.

References

[1] Saá, P., Proctor, M., Foster, G., *et al.* (2018) Investigational testing for Zika virus among US blood donors. *N. Engl. J. Med.*, **378**, 1778–1788.
[2] American Red Cross (2019) *Infectious Disease Testing*, https://www.redcrossblood.org/biomedical-services/blood-diagnostic-testing/blood-testing.html (accessed 8 April 2019).
[3] Colorado State University (2019) *Pooled Testing on Three Fronts to Improve Test Affordability*, http://csu-cvmbs.colostate.edu/vdl/Pages/pooled-testing-improves-test-affordability.aspx (accessed 8 April 2019).
[4] Shah, D., Dillard, H., and Nault, B. (2005) Sampling for the incidence of aphid-transmitted viruses in snap bean. *Phytopathology*, **95**, 1405–1411.
[5] Mester, P., Witte, A., Robben, C., *et al.* (2017) Optimization and evaluation of the qPCR-based pooling strategy DEP-pooling in dairy production for the detection of Listeria monocytogenes. *Food Control*, **82**, 298–304.
[6] Salzer, E., Nixon, E., Drewes, G., *et al.* (2016) Screening pools of compounds against multiple endogenously expressed targets in a chemoproteomics binding assay. *J. Lab. Autom.*, **21**, 133–142.
[7] Thai, M. (2011) *Group Testing Theory in Network Security: An Advanced Solution*, Springer, New York.
[8] Dorfman, R. (1943) The detection of defective members of large populations. *Ann. Math. Stat.*, **14**, 436–440.
[9] Johnson, N., Kotz, S., and Wu, X. (1991) *Inspection Errors for Attributes in Quality Control*, CRC Press, London.
[10] Ding-Zhu, D. and Hwang, F. (2000) *Combinatorial Group Testing and its Applications*, World Scientific, Singapore.
[11] Bilder, C. (2019) *Group Testing for Identification*, Wiley StatsRef: Statistics Reference Online.
[12] Burrows, P. (1987) Improved estimation of pathogen transmission rates by group testing. *Phytopathology*, **77**, 363–365.
[13] Bilder, C. and Tebbs, J. (2005) Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biom. J.*, **47**, 502–516.
[14] Hepworth, G. and Biggerstaff, B. (2017) Bias correction in estimating proportions by pooled testing. *J. Agric. Biol. Environ. Stat.*, **22**, 602–614.
[15] Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.
[16] Tebbs, J. and Bilder, C. (2004) Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *J. Agri. Biol. Environ. Stat.*, **9**, 75–90.
[17] Biggerstaff, B. (2008) Confidence intervals for the difference of two proportions estimated from pooled samples. *J. Agri. Biol. Environ. Stat.*, **13**, 478–496.
[18] Bilder, C. and Loughin, T. (2014) *Analysis of Categorical Data with R*, CRC Press, Boca Raton, FL.
[19] Hepworth, G. (2005) Confidence intervals for proportions estimated by group testing with groups of unequal size. *J. Agri. Biol. Environ. Stat.*, **10**, 478–497.
[20] Xie, M. (2001) Regression analysis of group testing samples. *Stat. Med.*, **20**, 1957–1969.
[21] Louis, T. (1982) Finding the observed information matrix when using the EM algorithm. *J. Roy. Stat. Soc. B*, **44**, 226–233.
[22] Zhang, B., Bilder, C., and Tebbs, J. (2013) Group testing regression model estimation when case identification is a goal. *Biom. J.*, **55**, 173–189.
[23] McMahan, C., Tebbs, J., Hanson, T., and Bilder, C. (2017) Bayesian regression for group testing data. *Biometrics*, **73**, 1443–1452.
[24] Delaigle, A. and Hall, P. (2015) Nonparametric methods for group testing data, taking dilution into account. *Biometrika*, **102**, 871–887.
[25] Warasi, M., McMahan, C., Tebbs, J., and Bilder, C. (2017) Group testing regression models with dilution submodels. *Stat. Med.*, **36**, 4860–4872.
[26] Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000) Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*, **56**, 1126–1133.
[27] Bilder, C. and Tebbs, J. (2009) Bias, efficiency, and agreement for group-testing regression models. *J. Stat. Comput. Simul.*, **79**, 67–80.
[28] Bilder, C., Zhang, B., Schaarschmidt, F., and Tebbs, J. (2010) binGroup: A package for group testing. *R J.*, **2**, 56–60.
[29] Gildow, F., Shah, D., Sackett, W., *et al.* (2008) Transmission efficiency of Cucumber mosaic virus by aphids associated with virus epidemics in snap bean. *Phytopathology*, **98**, 1233–1241.

DOI: 10.1002/9781118445112.stat08231

[30] Verstraeten, T., Farah, B., Duchateau, L., and Matu, R. (1998) Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district. *Trop. Med. Int. Health*, **3**, 747–750.
[31] Swallow, W. (1985) Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology*, **75**, 882–889.
[32] Xie, M., Tatsuoka, K., Sacks, J. and Young, S. (2001) Group testing with blockers and synergism. *J. Am. Stat. Assoc.*, **96**, 92–102.
[33] Hughes-Oliver, J. and Rosenberger, W. (2000) Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, **87**, 315–327.
[34] Zhang, B., Bilder, C.R., and Tebbs, J.M. (2013) Regression analysis for multiple-disease group testing data. *Stat. Med.*, **32** (28), 4954–4966.
[35] Hepworth, G. (2019) Bias correction of estimated proportions using inverse binomial group testing. *Aust. NZ J. Stat.*, **1** (1), 51–60.

DOI: 10.1002/9781118445112.stat08231