Published in final edited form as:

Am Stat. 2015 ; 69(1): 45–52. doi:10.1080/00031305.2014.983545.

A Note on the Minimax Solution for the Two-Stage Group Testing Problem

Yaakov Malinovsky and

Assistant Professor, Department of Mathematics and Statistics, University of Maryland, Baltimore, MD 21250

Paul S. Albert

Chief and Senior Investigator, Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892

Abstract

Group testing is an active area of current research and has important applications in medicine, biotechnology, genetics, and product testing. There have been recent advances in design and estimation, but the simple Dorfman procedure introduced by R. Dorfman in 1943 is widely used in practice. In many practical situations, the exact value of the probability p of being affected is unknown. We present both minimax and Bayesian solutions for the group size problem when p is unknown. For unbounded p, we show that the minimax solution for group size is 8, while using a Bayesian strategy with Jeffreys' prior results in a group size of 13. We also present solutions when p is bounded from above. For the practitioner, we propose strong justification for using a group size of between 8 and 13 when a constraint on p is not incorporated and provide useable code for computing the minimax group size under a constrained p.

Keywords

Loss function; Optimal design; Optimization problem

1. INTRODUCTION

The purpose of this article is to propose a practical and simple group testing procedure that performs well in a wide range of situations. Group testing procedures save cost and time and have wide spread applications, including blood screening (Dorfman 1943; Finucan 1964; Gastwirth and Johnson 1994; Litvak, Tu, and Pagano 1994; Delaigle and Hall 2012; McMahan, Tebbs, and Bilder 2012; Tebbs, McMahan, and Bilder 2013), quality control in product testing (Sobel and Groll 1959, 1966), computation biology (De Bonis, Gasieniec, and Vaccaro 2005), DNA screening (Du and Hwang 2006; Golan, Erlich, and Rosset 2012), and photon detection (van den Berg et al. 2013). According to Hughes-Oliver (2006), group

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/tas.

testing began as early as 1915, when it was used in dilution studies for estimating the density of organisms in a biological medium.

In his 1950 book, William Feller nicely described the group testing problem as (Feller 1950):

> A large number, N, of people are subject to a blood test. This can be administered in two ways. (i) Each person tested separately. In this case N tests are required. (ii) The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for the k people. If the test is positive, each of the k persons must be tested separately, and in all $k + 1$ tests are required for the k people. Assume the probability p that the test is positive is the same for all and that people are stochastically independent.

Procedure (ii) is commonly referred to as the Dorfman two-stage group testing procedure (DTSP; Dorfman 1943; Samuels 1978). Interesting historical comments related to this problem can be found in the introduction of the book by Du and Hwang (1999).

Let $E(k, p)$ be the expected number of tests per person using DTSP with a group size k and probability of infection p. Then $E(1, p) = 1$ and $E(k, p) = 1 - (1 - p)^k + k^{-1}$ for $k \geq 2$. An important issue for the DTSP is to find an optimal value of k, $k^* = k^*(p)$, that minimizes the expected number of tests for a given p.

Samuels (1978) solved this optimization problem as follows. Let $[x]$ and $\{x\} = x - [x]$ denote the integer and fractional parts of x, respectively. k^* is a nonincreasing function of p, which is 1 for $p > 1 - 1/3^{1/3} \approx 0.31$, and otherwise is either $1 + [p^{-1/2}]$ or $2 + [p^{-1/2}]$. If $\{p^{-1/2}\} < [p^{-1/2}]/(2[p^{-1/2}] + \{p^{-1/2}\})$, then $k^* = 1 + [p^{-1/2}]$. If $\{p^{-1/2}\} > [p^{-1/2}]/(2[p^{-1/2}] + \{p^{-1/2}\})$, then to find out which of the values $k' = 1 + [p^{-1/2}]$ or $k'' = 2 + [p^{-1/2}]$ is optimal, one plugs them into $E(k, p)$.

The DTSP is not an optimal procedure and can be improved by introducing more than two stages (e.g., Sobel and Groll 1959 for known p). However, the optimal testing algorithm (unknown p) is unknown, and it is a difficult optimization problem (Du and Hwang 1999). Ungar (1960) proved that if $p > (3 - 5^{1/2})/2 \approx 0.38$, then there does not exist an algorithm that is better than individual one-by-one testing. Sobel and Groll (1966) presented a Bayesian model for the multistage group testing problem based on a prior distribution of p. Schneider and Tang (1990) derived adaptive procedures for the two-stage group testing problem based on a beta prior distribution. Although the DTSP is not optimal, it is often used in practice due to its simplicity (Tamashiro et al. 1993; Moore et al. 2000; Westreich et al. 2008).

In many practical situations, the exact value p of the probability of being affected (e.g., disease prevalence) is unknown, and therefore the optimal group size cannot be calculated. In this note, we derive both the minimax group size as well as the Bayesian solution under reasonable prior distributions for the DTSP. A comparison of the solutions under these different alternatives will aid the practitioner in the design of future applications of group testing. We first present the loss function needed for the minimax solution.

2. LOSS FUNCTION

If we know the value of *p*, then by using the result from Samuels (1978) we achieve the minimum value for *E*(*k, p*) in DTSP by

$$E(k^*(p),p)=\begin{cases}1-(1-p)^{k^*(p)}+\frac{1}{k^*(p)} & \text{for } 0<p\leq 1-(1/3)^{1/3}\\ 1 & \text{otherwise.}\end{cases} \quad (1)$$

where we can write the expected number of tests per person in a group of size *k* as

$$E(k,p)=\begin{cases}1-(1-p)^{k}+\frac{1}{k} & \text{for} \quad k>1\\ 1 & \text{for} \quad k=1.\end{cases} \quad (2)$$

It is important to note that according to Samuels (1978), $k^*(p)$ is a nonincreasing function of *p*, and $k^*(p) \to \infty$ as $p \downarrow 0$. Combining this fact with (1) and using the particular form of $k^*(p)$, it is possible to show (see Appendix B) that $E(k^*(p), p) \to 0$, as $p \downarrow 0$. Also, from Comment A.2 (Appendix A) it follows that $E(k^*(p), p)$ is a nondecreasing function of *p*.

We define the loss in DTSP as a difference between the expected number of tests and the expected number of tests under an optimal DTSP. Specifically,

$$L(k,p)=E(k,p)-E(k^*(p),p). \quad (3)$$

It is important to note that this type of loss function was considered by Robbins (1952) for the two-armed bandit problem. Although Robbins (1952) discussed a different problem from ours, there is a similarity between the two problems in that both try to quantify the loss “due to ignorance of the true state of affairs” and to seek a minimax solution.

Another loss function is

$$L_2(k,p)=\frac{E(k,p)}{E(k^*(p),p)}-1, \quad (4)$$

which reflects a relative rather than absolute change. However, it cannot be used since $\lim_{p\downarrow 0} E(k^*(p), p) = 0$ and the measure becomes undefined when *p* is near 0. Also, it is important to note that the expected number of tests per person, *E*(*k, p*), itself is not an appropriate loss function for obtaining the minimax solution since *E*(*k, p*) is a nondecreasing function of *p* for any *k* (i.e., *E*(*k, p*) is maximized at the upper bound of *p* for all *k*).

3. MINIMAX SOLUTION FOR UNBOUNDED *P*

We define the minimax group size as a group size that minimizes the largest loss $L(k, p)$,

$$k^{**}=\arg\min_{k\in\mathbb{N}^+}\sup_{p\in(0,1-(1/3)^{1/3}]}L(k,p). \quad (5)$$

From Equation (1) it follows that the DTSP has utility only when p is in the range $(0, 1 - (1/3)^{1/3}]$, while individual testing is optimal outside this range. Also it is important to note that

$$\sup_{p\in(0,1-(1/3)^{1/3}]}L(k,p)=\sup_{p\in(0,1)}L(k,p) \quad (6)$$

(see Appendix B, Result B.1). In Appendix C, Lemma C.1, we are able to obtain a closed-form expression for k^{**} in (5). The proof uses the following two steps to show that $k^{**} = 8$.

Step 1: Fix k, $k \in \{1, 2, 3,...,\}$. Find $p^*(k) = \arg\sup_{p\in(0,1-(1/3)^{1/3}]} L(k, p)$.

Step 2: Find k^{**}, where

$$k^{**}=\arg\min_{k\in\{1,2,3,...\}}L(k,p^*(k)). \quad (7)$$

In addition to the proof (see Appendix) we present a heuristic argument for finding the minimax solution, which we present as follows (we also use the same argument when additional information on an upper bound of p is incorporated: see Section 5).

In Step 1, we performed a grid search with incremental steps of 10^{-6}. We present the graphs of $L(k, p)$ as a function of $p \in (0, 1 - (1/3)^{1/3}]$ for $k = 1,..., 9$ (Figure 1). We noticed that $p^*(1) = p^*(2) = \cdots = p^*(7) = 0$ and therefore $L(k, p^*(k)) = 1/k$ for $k = 1, 2,..., 7$. It is clear that there is a jump at $k = 8$ for $p^*(k)$, reflecting the worst case for p (largest loss) in Step 1.

Note that the loss function (3) is not a smooth function of p. In Result C.1 (Appendix C), we prove that $L(k, p^*(k))$ is a unimodal function of k for $k \geq 1$. We also obtain a close-form expression for $p^*(k)$ (see Remark C.1, Appendix C).

Table 1 presents the results of the maximization of the loss function (Step 1) as a function of p for a given k. The minimax solution is then obtained by minimizing the loss function as a function of k (Step 2) (we present Matlab code for the two-step procedure in Appendix D).

As shown in Table 1, the minimax solution is $k^{**} = 8$. To evaluate the performance of the minimax solution, we compare it with the optimal solution assuming p is known. Specifically, we compare the optimal expected number of tests per individual $E(k^*(p), p)$

with the minimax expected number of tests $E(k^{**}, p)$. Table 2 presents the ratio of these two quantities defined as RE(p) = $E(k^{**}, p)/E(k^{*}(p), p)$ for different values of p.

Table 2 suggests that as long as p is not very small ($p < 0.005$), the minimax solution is close to optimal. In the situation where we believe that p is very small and we know the upper bound for p, we can obtain the minimax solution in the restricted parameter space. This is discussed in Section 5. In Table 2, we also present the relative performance of the Bayesian solution under a Jeffreys' prior to the optimal solution as the ratio $\mathrm{RE}_{J_1}(p)=E(k^{*}_{J_1},p)/E(k^{*}(p),p)$, where $k^{*}_{J_1}$ is the optimum group size under a Jeffreys' prior. This will be discussed in Section 4.

In practice, it is useful to identify the range of the values of p for which minimax design k^{**} = 8 is optimal. More specifically, for any group size l we can compute the range [p^{*}_{l}, p^{**}_{l}] of the values of p for which the group size l is optimal. The derivation of p^{*}_{l} and p^{**}_{l} is presented in Appendix A as Proposition A.1. As a consequence of Proposition 1, we find that the minimax design is optimal over a range of p of $p^{*}_{8}=0.0157$ to $p^{**}_{8}=0.0206$.

In the next section, we propose a Bayesian alternative to the minimax solution.

4. BAYESIAN SOLUTION

For the Bayesian design, we are required to specify a prior distribution $\pi(p)$ for the probability p of being affected. The basic assumption of our model is that the affected status X_i, $i = 1, 2, \dots$ are iid Bernoulli random variables with parameter p (i.e., $X_i \mid p \sim \mathrm{Ber}(p)$). Under the Bayesian consideration with the specified prior distribution p, the analog to the loss function (3) is

$$L_{\pi}(k)=\int L(k,p)\pi(p)dp=E_{\pi}(k)-E^{*}_{\pi}, \quad (8)$$

where $E_{\pi}(k) = \int E(k, p)\, \pi(p)\, dp$, $E^{*}_{\pi}=\int E(k^{*}(p),p)\pi(p)dp$. After choosing a prior, the loss function $L_{\pi}(k)=E_{\pi}(k)-E^{*}_{\pi}$ is a function of k only, and the optimal value of k is simply,

$$k^{*}_{\pi}=\min_{k} L(k,\pi)=\min_{k} E_{\pi}(k). \quad (9)$$

It is clear that the optimal group size k^{*}_{π} in (7) is a function of the prior distribution π.

4.1 Uniform Prior

If we do not have any prior information about the disease prevalence p, that is, all values of p are equally likely, then the uniform prior is reasonable. Denote the uniform prior by $\pi = I_U$, where U is an upper bound of the distribution support, and $E_{\pi}(k)$ can be written as

$$E_{I_U}(k)=1+\left\{\frac{1}{k}+\frac{1}{k+1}\left((1-U)^{k+1}-1\right)\right\}1_{\{k>1\}}. \quad (10)$$

For unbounded p, $U = 1$, $E_{I_1}(k)=1+\frac{1}{k}1_{\{k>1\}}$, and therefore $k^*_{I_1}=1$. As we will discuss in Section 5 (see Tables 3, 4, and E.1), the Bayesian solution is not 1 for bounded p (i.e., $U < 1$).

4.2 Jeffreys' Prior

Another possibility is to find a prior distribution $\pi(p)$ that has *a small effect* on the posterior $\pi(p|x)$ distribution. Thus, we want to find a prior distribution $\pi(p)$ that produces the maximum value of the Kullback–Leibler information, $K(\pi(p \mid x), \pi(p))$, for discrimination between two densities, $\pi(p)$ and $\pi(p \mid x)$, where the latter is the posterior that reflects the sampling density. The Kullback–Leibler information for discrimination between two densities, f and g is defined as

$$K(f,g)=E_f\left(\log\frac{f}{g}\right). \quad (11)$$

A prior that maximizes the Kullback–Leibler information is by default a *reference* prior (for the general discussion of choosing a reference prior, we refer to Jeffreys (1946), Berger, Bernardo, and Sun (2009), and Shemyakin (2014)). The above approach to the construction of the reference prior is due to Bernardo (1979) (an excellent summary is given in Lehmann and Casella 1998 and the recent developments in Berger, Bernardo, and Sun 2009, 2012). We cannot directly use $K(\pi(p \mid x), \pi(p))$ because it is a function of x. Therefore, we consider the expected value of $K(\pi(p \mid x), \pi(p))$ with respect to the marginal distribution of X, which is the Shannon information

$$S(\pi)=\int K(\pi(p|x),\pi(p))m_\pi(x)dx, \quad (12)$$

where $m_\pi(x) = \int f(x|p)\,\pi(p)dp$ is the marginal distribution of X.

The following lemma is due to Clarke and Barron (1990) and is taken from Lehmann and Casella (1998):

Lemma 1—Let $X_1,\ldots, X_n$ be an iid sample from $f(x|p)$, and let $S_n(\pi)$ denote the Shannon information of the sample. Then, as $n \to \infty$,

$$S_n(\pi)=\frac{1}{2}\log\frac{n}{2\pi e}+\int\pi(p)\log\frac{|I_n(p)|^{1/2}}{\pi(p)}dp+o(1), \quad (13)$$

where $I_n(p)$ is the Fisher information contained in the sample $X_1,\ldots, X_n$, and $o(1)$ is a notation of the function $h(n) = o(1)$ such that $h(n) \to 0$ as $n \to \infty$.

The Fisher information contained in the sample of size n is defined as

$$I_n(p)=-E\left\{\frac{\partial^2}{\partial p^2}\log f(X_1,\ldots,X_n|p)\right\}=\frac{n}{p(1-p)}. \tag{14}$$

The last equality in (14) is because X_i given p follows a Bernoulli distribution.

From Jensen's inequality, it follows that the right-hand side of Equation (13) is maximized when the prior distribution is proportional to the square root of the Fisher information, that is,

$$\pi_J(p)=\arg\sup_{\pi} S_n(\pi)\propto |I_n(p)|^{1/2}\propto\frac{1}{(p(1-p))^{1/2}}. \tag{15}$$

The last expression $\frac{1}{(p(1-p))^{1/2}}$ in (15) is known as Jeffreys' prior (see, e.g., Lehmann and Casella 1998), where the normalizing constant for Jeffreys' prior is

$$c_U=\int_0^U\frac{1}{(p(1-p))^{1/2}}dp=2\arcsin\sqrt{U}. \tag{16}$$

Under this approach if $0 < p \le U \le 1$, then

$$\pi_J(p)=\frac{1}{c_U}\frac{1}{(p(1-p))^{1/2}}.$$

Under this prior, the expected number of tests per person with a group size k is

$$E_{J_U}(k)=\frac{1}{c_U}\int_0^U E(k,p)\frac{1}{(p(1-p))^{1/2}}dp. \tag{17}$$

Numerically evaluating (17) (Appendix D), the optimum group size under a Jeffreys' prior is

$$k_{J_1}^*=\arg\min_k E_{J_1}(k)=13.$$

First, from a Bayesian perspective, the minimax design performs well relative to the optimal Bayesian design under a Jeffreys' prior (i.e., $E_{J_1}(13)/E_{J_1}(8) = 0.9219/0.9286 = 0.993$). From a frequentist point of view, the Bayesian design performs well for small p but not as well as the minimax design for $p > 0.01$ (see the third row of Table 2, where $\mathrm{RE}_{J_1}(p)=E(k^*_{J_1}, p)/E(k^*(p), p))$.

It is important to note that theoretically we can define the min-imax group size under the Bayesian setup (see, e.g., Ferguson 1967, p. 57) in the following way:

$$k^*_B=\min_k \sup_\pi L_\pi(k).$$

Unfortunately, maximization with respect to all possible prior distributions $\pi(p)$ is intractable. Even if we consider a beta prior, the problem remains intractable unless we limit the range of the parameters.

5. MINIMAX AND BAYESIAN SOLUTIONS FOR BOUNDED *p*

Define the minimax solution in the restricted parameter space (when we know an upper bound U of p) as

$$k^{**}_U=\arg \min_{k\in\mathbb{N}^+} \sup_{p\in(0,U]} L(k,p). \qquad (18)$$

The minimax solution subject to an upper bound on p denoted as k^{**}_U can be evaluated with a two-step procedure similar as to that presented in Section 3. The difference is that the support of p is changed to $(0, U]$ and the grid step is changed accordingly. Table 4 demonstrates both minimax and Bayesian solutions in the restricted parameter space of p (only the upper bound is specified).

Define $\mathrm{RE}_U(p)=E(k^{**}_U, p)/E(k^*(p), p)$ as an index of the efficiency of the minimax test (in the restricted parameter space) relative to that of the optimal DTSP test. In Table 4, we present $\mathrm{RE}_U(p)$ for different values of U and p.

A comparison between Tables 2 and 4 demonstrates the clear advantage of a minimax estimator in the restricted parameter space in comparison to the minimax estimator in the unrestricted parameter space. For example, when $p = 0.001$, bounding the parameter space by $U = 0.005$ increases the relative efficiency by 2.11(= 2.118/1.0028). Table 4 shows that the minimax solution in the restricted parameter space performs the worst relative to an optimal design when $p = U$.

In Table 4, we also present the performance of the optimal Bayesian (under Uniform and Jeffreys' priors) design in the restricted parameter space. In the restricted parameter space, the optimal group size under Jeffreys' prior is $k^*_{J_U}=\arg\min_k E_{J_U}(k)$, where $E_{J_U}(k)$ can be evaluated using (16) and (17) for a particular upper bound U. The ninth row in Table 4 presents $k^*_{J_U}$, and the fifth row presents the relative efficiency $\mathrm{RE}_{J_U}(p)=E(k^*_{J_U},p)/E(k^*(p),p)$. The optimal Bayesian group size for bounded *p* is similar to the minimax group size, with the relative efficiencies of both designs being near optimum.

6. SUMMARY

This note presents both unconstrained and constrained minimax group size solutions for group testing within the DTSP framework. We found that a group size of eight is the unconstrained minimax solution. We also present novel methodology to evaluate the range of the values of *p* for which the minimax design is optimal. When we have prior information that establishes an upper bound on *p*, we show that the constrained performs substantially better than the unconstrained minimax. An advantage of the minimax solution is their simplicity within the two-stage group testing framework. In addition, we developed a Bayesian design under both constrained and unconstrained settings, which in most cases performed similarly to the minimax design in terms of the relative efficiency. This article has important design implications in practice. For example, Pilcher et al. (2004) used a DTSP with a group size of 10 to detect acute HIV infection. This is consistent with our design result that suggested a group size of between 8 and 13 given that no information on the primary infection rate (*p*) was known a priori. Further, with known constraint on support of *p*, we provide computer code (Appendix D) that easily can be applied by the practitioner. Research in more general algorithms (e.g., more than two-stage) is needed, but any such algorithm will be complex and difficult for the practitioner to implement.

Acknowledgments

The work was partially supported by a 2013 UMBC Summer Faculty Fellowship grant and the Eunice Kennedy Shriver National Institute of Child Health and Human Development intramural program. The authors thank an editor, associate editor, and two referees for their thoughtful and constructive comments and suggestions that resulted in very significant improvements in the article. The authors also thank Sara Joslyn for editing the article and Abram Kagan and Yosef Rinott for discussions on the topic and comments on the article.

References

Berger JO, Bernardo JM, Sun D. The Formal Definition of Reference Priors. The Annals of Statistics. 2009; 37:905–938.

Berger JO, Bernardo JM, Sun D. Objective Priors for Discrete Parameter Spaces. Journal of the American Statistical Association. 2012; 107:636–648.

Bernardo JM. Reference Posterior Distributions for Bayesian Inference" (with discussion). Journal of the Royal Statistical Society, Series B. 1979; 41:113–147.

Clarke BS, Barron AR. Information-Theoretic Asymptotics of Bayes Methods. IEEE Transactions on Information Theory. 1990; 36:453–471.

De Bonis A, Gasieniec L, Vaccaro U. Optimal Two-Stage Algorithms for Group Testing Problems. SIAM Journal on Computing. 2005; 34:1253–1270.

Delaigle A, Hall P. Nonparametric Regression With Homogeneous Group Testing Data. The Annals of Statistics. 2012; 40:131–158.

Dorfman R. The Detection of Defective Members of Large Populations. The Annals of Mathematical Statistics. 1943; 14:436–440.

Du, D.; Hwang, FK. Combinatorial Group Testing and its Applications. Singapore: World Scientific; 1999.

Du, D.; Hwang, FK. Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing. Singapore: World Scientific; 2006.

Feller, W. An Introduction to Probability Theory and its Application. New York: Wiley; 1950.

Ferguson, TS. Mathematical Statistics: A Decision Theoretic Approach. New York and London: Academic Press; 1967.

Finucan HM. The Blood Testing Problem. Applied Statistics. 1964; 13:43–50.

Gastwirth J, Johnson W. Screening With Cost Effective Quality Control: Potential Applications to HIV and Drug Testing. Journal of the American Statistical Association. 1994; 89:972–981.

Golan D, Erlich Y, Rosset S. Weighted Pooling—Practical and Cost Effective Techniques for Pooled High Throughput Sequencing. Bioinformatics. 2012; 28:i197–i206. [PubMed: 22689761]

Hughes-Oliver, JM. Pooling Experiments for Blood Screening and Drug Discovery. In: Dean, AM.; Lewis, SM., editors. Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics. New York: Springer-Verlag, Inc; 2006. p. 48-68.

Jeffreys H. An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London, Series A. 1946; 186:453–461.

Lehmann, EL.; Casella, G. Theory of Point Estimation. 2. New York: Springer; 1998.

Litvak E, Tu X, Pagano M. Screening for the Presence of a Disease by Pooling Sera Samples. Journal of the American Statistical Association. 1994; 89:424–434.

McMahan C, Tebbs J, Bilder C. Informative Dorfman screening. Biometrics. 2012; 68:287–296. [PubMed: 21762119]

Morre SA, Meijer CJLM, Munk C, Krüger-Kjaer S, Winther JF, Jorgensens HO, Van den Brule AJC. Pooling of Urine Specimens for Detection of Asymptomatic Chlamydia Trachomatis Infections by PCR in a Low-Prevalence Population: Cost-Saving Strategy for Epidemiological Studies and Screening Programs. Journal of Clinical Microbiology. 2000; 38:1679–1680. [PubMed: 10747169]

Pilcher CD, Price MA, Hoffman IF, Galvin S, Martinson FE, Kazembe PN, Eron JJ, Miller WC, Fiscus SA, Cohen MS. Frequent Detection of Acute Primary HIV Infection in Men in Malawi. AIDS. 2004; 18:517–524. [PubMed: 15090805]

Robbins H. Some Aspects of the Sequential Design of Experiments. Bulletin of the American Mathematics Society. 1952; 58:527–535.

Samuels SM. The Exact Solution to the Two-Stage Group-Testing Problem. Technometrics. 1978; 20:497–500.

Schneider H, Tang K. Adaptive Procedures for the Two-Stage Group-Testing Problem Based on Prior Distributions and Costs. Techno-metrics. 1990; 32:397–405.

Shemyakin A. Hellinger Distance and Non-informative Priors. Bayesian Analysis. 2014; 9:923–938.

Sobel M, Groll PA. Group Testing to Eliminate Efficiently All Defectives in a Binomial Sample. Bell System Technical Journal. 1959; 38:1179–1252.

Sobel M, Groll PA. Binomial Group-Testing With an Unknown Proportion of Defectives. Technometrics. 1966; 8:631–656.

Tamashiro H, Maskill W, Emmanuel J, Fauquex A, Sato P, Heymann D. Reducing the Cost of HIV Antibody Testing. Lancet. 1993; 342:87–90. [PubMed: 8100916]

Tebbs J, McMahan C, Bilder C. Two-Stage Hierarchical Group Testing for Multiple Infections With Application to the Infertility Prevention Project. Biometrics. 2013; 69:1064–1073. [PubMed: 24117173]

Ungar P. Cutoff Points in Group Testing. Communications on Pure and Applied Mathematics. 1960; 13:49–54.

van den Berg E, Candàe E, Chinn G, Levin C, Olcott PD, Sing-Long C. Single-Photon Sampling Architecture for Solid-State Imaging Sensors. Proceedings of the National Academy of Sciences of the United States. 2013; 110:E2752–E2761.

Westreich DJ, Hudgens MG, Fiscus SA, Pilcher CD. Optimizing Screening for Acute Human Immunodeficiency Virus Infection With Pooled Nucleic Acid Amplification Tests. Journal of Clinical Microbiology. 2008; 46:1785–1792. [PubMed: 18353930]

APPENDIX A. INVERSE OF THE SAMUELS RESULT

Proposition A.1

Let $q = 1 - p$. Define $q_2^* = 1/3^{1/3}$ and q_l^*, $l \geq 3$ be the larger real root of equation $q^l(1-q) = \frac{1}{l(l+1)}$. If $q \leq 1/3^{1/3}$, then $k^*(q) = 1$. If $q > 1/3^{1/3}$ and $q_{l-1}^* < q < q_l^*$, $l \geq 3$, then $k^*(q) = l$.

For example, for the minimax group size $k^{**} = 8$, $q_8^* \approx 1 - 0.0157$ is the larger real root of equation $q^8(1-q) - \frac{1}{8(8+1)} = 0$, and $q_7^* \approx 1 - 0.0206$ is the larger real root of equation $q^7(1-q) - \frac{1}{7(7+1)} = 0$. Therefore, the minimax group size $k^{**} = 8$ is optimal for any q in the range $(1 - 0.0206, 1 - 0.0157)$ or, alternatively, for any p in the range $(0.0157, 0.0206)$.

Comment A.1

The Corollary in Samuels (1978) (Equation (2)) provides the lower and upper bounds for $E(k^*(p), p)$. The upper bound is not sharp, and in fact is invalid since it is greater than one when p is larger than 0.1485 (since 0.1485 is the real zero of $4x^3 + 2x - 1$, $x = p^{1/2}$). A simple upper bound for $E(k^*(p), p)$ follows directly from Proposition A.1: if $p \leq 1 - 1/3^{1/3}$, then $E(k^*(p), p) < E(3, p) = 1 - q^3 + 1/3$.

Proof of Proposition A.1

We obtain the proof from the analyzing Samuels (1978) method. Assume that $p \leq 1 - 1/3^{1/3}$. First, from the Samuels (1978) theorem (Section 1), it follows that $k^*(p) \geq 2$. Second, Proposition 1 in Samuels (1978) says that $k^*(p)$ is never 2. Therefore, $k^*(p)$ should be at least 3. Third, following Samuels (1978) notation, define $r_1(k)$ and $r_2(k)$ as the smaller and larger roots of the equation

$$\begin{aligned}\Delta_k(q) &= E_{k+1}(q) - E_k(q) \\ &= q^k(1-q) - \frac{1}{k(k+1)} = 0, \; k \geq 3.\end{aligned} \qquad (A.1)$$

Samuels (1978) showed that $\frac{k}{k+1} = \max_q \Delta_k(q)$, $\Delta_k(\frac{k}{k+1}) > 0$, $\Delta_k(0) = \Delta_k(1) < 0$ and that both $r_1(k)$ and $r_2(k)$ are increasing functions of k. Combining points one to three, we can conclude that if $q < r_2(3)$ then $E_3 < E_4$, if $q < r_2(4)$ then $E_4 < E_5$, and so on. The inequality $r_2(3) < r_2(4) < \cdots$ completes the proof.

Comment A.2

It follows from the proof of Proposition A.1 that $E(k^*(p), p)$ is a nondecreasing function of p, $p \in (0, 1)$.

APPENDIX B. SUPPORT OF THE LOSS FUNCTION WITH RESPECT TO p

Result B.1

$\sup_{p \in (0, 1-(1/3)^{1/3}]} L(k, p) = \sup_{p \in (0,1)} L(k, p)$.

Proof

First we show that $\lim_{p \downarrow 0} E(k^*(p), p) = 0$. Recall, that according to Samuels (1978), if $0 < p \leq 1 - (1/3)^{1/3}$ then k^* is $1 + [p^{-1/2}]$ or $2 + [p^{-1/2}]$ and in this case from (1) we have $E(k^*(p), p) = 1 - (1-p)^{k^*(p)} + \frac{1}{k^*(p)}$. Specifically, for any p satisfies $0 < p \leq 1 - (1/3)^{1/3}$, we have $1 > (1-p)^{k^*(p)} \geq (1-p)^{2+p^{-1/2}}$. From the facts that $\lim_{p \downarrow 0} (1-p)^2 = 1$ and $\lim_{p \downarrow 0} \frac{1}{p^{1/2}} \log(1-p) = 0$, it follow that $\lim_{p \downarrow 0} (1-p)^{k^*(p)} = 1$. Combining this with Samuels (1978) result that $k^*(p) \to \infty$ as $p \downarrow 0$, we prove that $\lim_{p \downarrow 0} E(k^*(p), p) = 0$.

For $k = 1$. We have $L(1, p) = 1 - E(k^*(p), p)$. Hence, $\sup_{p \in (0,1)} L(1, p) = 1 - \inf_{p \in (0,1)} E(k^*(p), p) = 1 - \lim_{p \downarrow 0} E(k^*(p), p)$, where the last equation follows from Comment A.2. Therefore,

$$\sup_{p \in (0, 1-(1/3)^{1/3}]} L(1, p) = \sup_{p \in (0,1)} L(1, p). \qquad \text{(B.1)}$$

For $k \geq 2$. Define $p_0 = 1 - (1/3)^{1/3}$. For any $p \in (p_0, 1)$, from (1) and (3) it follows that $L(k, p) = E(k, p) - 1 = 1/k - (1-p)^k$ is an increasing function of p, and therefore from (1) and (2) it follows that

$$\sup_{p \in (p_0, 1)} L(k, p) = E(k, 1) - 1 = 1/k. \qquad \text{(B.2)}$$

Again, from Comment A.2 and (2), it follows that $\lim_{p \downarrow 0} L(k, p) = 1/k$. Therefore,

$$\sup_{p \in (0, p_0]} L(k, p) \geq 1/k. \qquad \text{(B.3)}$$

Combining, (B.2) and (B.3), we get

$$\sup_{p \in (0, p_0]} L(k, p) \geq \sup_{p \in (p_0, 1)} L(k, p). \qquad \text{(B.4)}$$

Equations (B.1) and (B.4) complete the proof.

APPENDIX C. BEHAVIOR OF p*(k) AND UNIMODALITY OF L(k, p*(k))

Result C.1

$p^*(k) = 0$, for $k = 1, 2, \ldots, 7$, $p^*(8)=1-(\frac{3}{8})^{1/(8-3)} \approx 0.178$. Moreover, $L(k, p^*(k))$ is a unimodal function of k for $k \geq 1$.

Proof

Define $q_0 = 1 - p_0$. Recall that it was shown in (B.3) that $\sup_{p\in(0,p_0]} L(k, p) \geq 1/k$. Also we have $\lim_{p\downarrow 0} L(k, p) = 1/k$. Therefore, if $\sup_{p\in(0,p_0]} L(k, p) = 1/k$, then $p^*(k) = 0$, and if $\sup_{p\in(0,p_0]} L(k, p) > 1/k$ then $p^*(k) > 0$. Now $L(k, p) > 1/k$ if and only if

$$q^{k^*}-q^{k}>\frac{1}{k^*}. \quad \text{(C.1)}$$

It is clear, that the necessary condition for the inequality (C.1) is

$$k^*<k. \quad \text{(C.2)}$$

We know (Samuels 1978 and Proposition A.1 above) that for $q > q_0$, k^* is a piecewise nonincreasing (nondecreasing) function of $p(q)$ and k^* is at least 3. From this and (C.2), it follows that $p^*(k) = 0$, for $k = 1, 2, 3$. Using Proposition A.1, we can verify that for $k = 7$, there does not exist a q ($q \in [q_0, 1)$) such that (C.1) holds. From direct logic it follows that if there exists q such that Equation (C.1) holds for k', then (C.1) holds for any $k > k'$. It is the same to say that if for any q (in the appropriate range) the Equation (C.1) does not hold for k'', then it does not hold for any q and for any $k < k''$. Therefore, $p^*(k) = 0$ and $L(k, p^*(k)) = 1/k$, for $k = 1, 2, \ldots, 7$.

Again using Proposition A.1, we can verify that for $k = 8$, Equation (C.1) holds only for some values of q that corresponds to $k^* = 3$. Therefore, finding $\sup_{p\in(0,p_0]} L(8, p)$ is equivalent to finding $\sup_{q_2^*<q<q_3^*}\{q^3-q^8\}$, where q_2^* and $q^*(8)=q_3^*$ are defined in Proposition A.1. A simple calculation then shows that $q^*(8) = (3/8)^{1/5}$ (correspondingly $p^*(8) = 1- (3/8)^{1/5} \approx 0.1781$) and $L(8, p^*(8)) = 437/3152 \approx 0.1386$.

From above, it follows that for $k \geq 8$, there exists a q such that $q_2^*<q<q_3^*$ and Equation (C.1) holds. Now, for such a q, it follows from the last inequality from the last paragraph in the proof of Proposition A.1 that $L(9, p) - L(8, p) = E(9, p) - E(8, p) > 0$, $p = 1 - q$. Therefore,

$$\begin{aligned} L(8,\,p^*(8))=&\sup_{p\in(0,\,p_0]} L(8,\,p)<\sup_{p\in(0,\,p_0]} L(9,\,p) \\ =&L(9,\,p^*(9)). \end{aligned} \quad \text{(C.3)}$$

Proceeding by induction on $k = 10, 11, \ldots$ we complete the proof of unimodality of $L(k, p^*(k))$ for $k \geq 1$.

Remark C.1

The function $p^*(k)$ is a decreasing function of k for $k \geq 8$ and $p^*(k)$ has a form $1 - (k^*/k)^{1/(k-k^*)}$ for $k \geq 8$ and for k^* that satisfied Equation (C.1) and the condition $q^*_{k^*-1} < q < q^*_{k^*}$ of Proposition A.1. But we have not proved it rigorously.

Lemma C.1

$k^{**} = 8$.

Proof

Follows immediately from Result C.1 and the proof that $L(k, p^*(k))$ is the unimodal function of k with the minimum at $k = 8$.

APPENDIX D: MATLAB CODE

i. **Matlab function "MinMaxValue,"** which has input K and uses all three functions (ii), (iii), (iv) given below.

```
function minmax=MinMaxValue(K,U)
R=[];
step=1/1000000;
for k=1:1:K
output=zeros(1000001,3);
counter=0;
for p=0:step:U
counter=counter+1;
prophet=ProphetNumberTestPerPerson(p);
player=PlayerNumberTestPerPerson(p,k);
l=player-prophet; output(counter,:)=[k p l];
end
m=max(output(:,3)); pl=find(output(:,3)==m); R=[R;output (pl,:)];
end
mm=min(R(:,3)); mmm=find(R(:,3)==mm); minmax=R(mmm, 1);
```

ii. **Matlab function "OptimalGroupSize,"** which has input $p \in (0, 1)$ and the output is the optimum group size $k^*(p)$ (based on Samuels 1978):

function kOpt=OptimalGroupSize(p)

if $p <= 1 - (1/3)^{1/3}$;

$q = 1 - p; w = p^{(-1/2)}$;

int=floor(w); frl=w-int;

$I nd = (f rl < int/(2 * int + f rl))$;

$k1 = int + 1$; k2=int+2;

$f1 = 1/k1 + 1 - q^{k1}$; $f2 = 1/k2 + 1 - q^{k2}$;

$f = min(f1, f2)$; $I nd1 = (f == f1)$;

$kOpt = ((1 + int)^{I nd}) * (((k1)^{I nd1}) * ((k2)^{1-I nd1}))^{1-I nd}$;

else

$kOpt = 1$;

end

iii. **Matlab function “PlayerNumberTestPerPerson”** calculates the expected number of tests per person in a group of size k with probability of infection p.

function player=PlayerNumberTestPerPerson(p,k)

q=1−p;

player= $1 - q^k + 1/k$;

iv. **Matlab function “ProphetNumberTestPerPerson”** calculates the expected number of tests per person in a group of the optimal size $k^*(p)$ with probability of infection p.

function prophet=ProphetNumberTestPerPerson(p)

kOpt=OptimalGroupSize(p); q=1−p; $prophet = 1 - q^{k\,Opt} + 1/kOpt$;

v. **Matlab function “JeffreyOptGroupsize”** calculates the optimal group size under Jeffreys’ prior with upper bound for support of p equal to U.

function kj=JeffreyOptGroupsize(U)

S=[];

for k=1:1:400

s=PlayerNumTestPerBetta1(k,U,1/2,1/2); S=[S;k s];

end

m=min(S(:,2)); l=find(S(:,2)==m);kj=S(l,1);

vi. **Matlab function “PlayerNumTestPerBetta1”** calculates the expected number of tests under Beta prior with upper bound for support of p equal to U.

function player=PlayerNumTestPerBetta1(k,U,a,b)

$c = @(p)p^{(a-1)} * (1-p)^{(b-1)};$

$f = @(p)(1-((1-p)^{k}) + 1./k) * p^{(a-1)} * (1-p)^{(b-1)};$

$player = (quad(c, 0, U))^{(-1)} * quad(f, 0, U);$

APPENDIX E. RELATIVE EFFICIENCY

Table E.1

Relative efficiencies of restricted minimax and Bayesian designs

	***U* = 0.10**			***U* = 0.20**			***U* = 0.30**		
p	0.01	0.05	0.10	0.10	0.15	0.20	0.20	0.25	0.30
$\mathrm{RE}_{U}(p)$	1.0342	1.0830	1.1694	1.1694	1.1853	1.1655	1.1655	1.1244	1.0778
$\mathrm{RE}_{I_U}(p)$	1.2732	1	1.0263	1	1.0122	1.0232	1.0232	1.0243	1.0198
$\mathrm{RE}_{J_U}(p)$	1.0778	1.0429	1.1190	1.0263	1.0516	1.0621	1.0621	1.0562	1.0420
$k^*(p)$	11	5	4	4	3	3	3	3	3
k_U^{**}	8	8	8	8	8	8	8	8	8
$k_{I_U}^{*}$	5	5	5	4	4	4	4	4	4
$k_{J_U}^{*}$	7	7	7	5	5	5	5	5	5

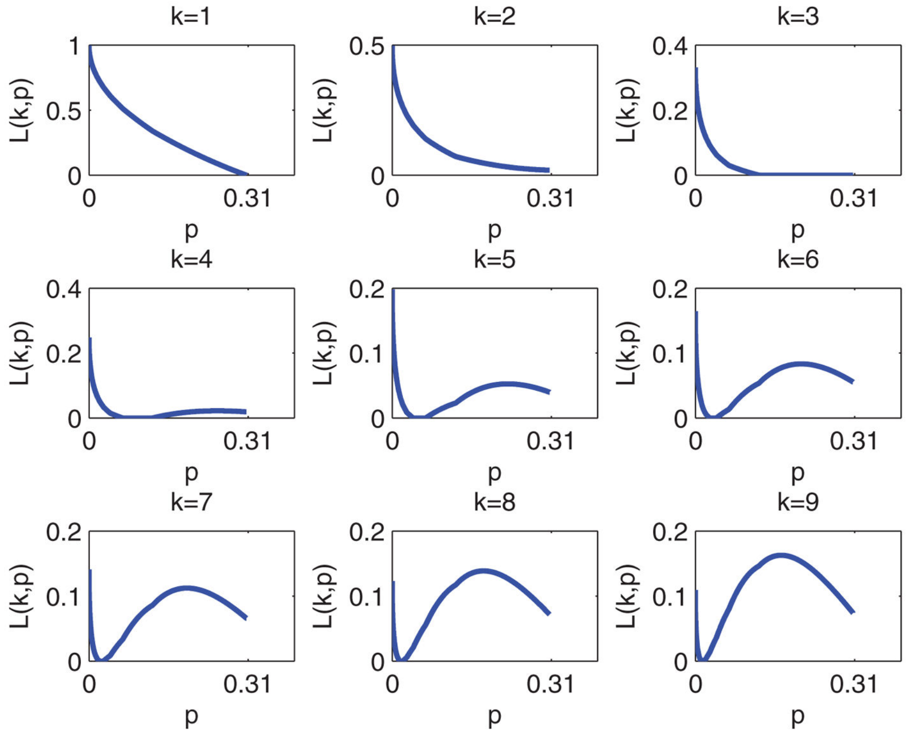

Figure 1.

$L(k, p)$ as a function of $p \in (0, 1 - (1/3)^{1/3}]$.

Table 1

Two-step (heuristic) solution

k	**1**	**2**	**3**	**···**	**7**	**8**	**9**	**10**
$p^*(k)$	0	0	0	···	0	0.178	0.167	0.158
$L(k, p^*(k))$	1	1/2	1/3	···	1/7	0.138	0.162	0.184
k	25	50	100	1000		10,000		
$p^*(k)$	0.083	0.049	0.029	0.004		0.0005		
$L(k, p^*(k))$	0.382	0.516	0.628	0.858		0.949		

Table 2

Relative efficiency of minimax design

p	0.0001	0.0005	0.001	0.005	0.01	0.05	0.10	0.25	0.30
RE(p)	6.305	2.900	2.118	1.181	1.034	1.082	1.169	1.124	1.078
RE_{J_1} (p)	3.921	1.875	1.432	1.007	1.020	1.322	1.385	1.156	1.078

Table 3

Minimax and Bayesian solution (group size) when upper bound U of p is specified

U	**0.0001**	**0.0005**	**0.001**	**0.005**	**0.01**	**0.05**	**0.10**	**0.15**	**0.30**
k_U^{**}	201	91	64	30	21	11	8	8	8
$k_{I_U}^{*}$	142	64	45	21	15	7	5	5	4
$k_{J_U}^{*}$	181	79	56	25	18	9	7	6	5

Table 4

Relative efficiencies of restricted minimax and Bayesian designs

	***U* = 0.0005**			***U* = 0.005**			***U* = 0.05**		
p	0.0001	0.0003	0.0005	0.001	0.003	0.005	0.005	0.01	0.05
$\mathrm{RE}_U(p)$	1.0048	1.0994	1.2474	1.0028	1.1055	1.2433	1.0392	1	1.2249
$\mathrm{RE}_{I_U}(p)$	1.1030	1.0044	1.0596	1.0901	1.0060	1.0606	1.2749	1.0778	1.0429
$\mathrm{RE}_{J_U}(p)$	1.0289	1.0461	1.1556	1.0310	1.0392	1.1343	1.1159	1.0103	1.1282
$k^*(p)$	101	58	45	32	19	15	15	11	5
k_U^{**}	91	91	91	30	30	30	11	11	11
$k_{I_U}^*$	64	64	64	21	21	21	7	7	7
$k_{J_U}^*$	79	79	79	25	25	25	9	9	9