CBCS
http://agrobiol.sggw.waw.pl/cbcs

Communications in Biometry and Crop Science
Vol. 2, No. 1, 2007, pp. 32–40

International Journal of the Faculty of Agriculture and Biology,
Warsaw University of Life Sciences, Poland

REGULAR ARTICLE

Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing

Frank Schaarschmidt

Institute of Biostatistics, Leibniz University Hannover, 30419 Hannover, Germany.
E-mail: schaarschmidt@biostat.uni-hannover.de

CITATION: Schaarschmidt, F. (2007). Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. *Comm. Biometry Crop Sci.* 2 (1), 32–40.

Received: 16 February 2007, Accepted: 19 April 2007, Published online: 16 May 2007

ABSTRACT

A common statistical issue in seed-quality control is to prove that the proportion of individuals showing an unwanted trait is less than a small threshold. Group testing can be used to reduce costs of assay and upper confidence limits for the proportion of detrimental individuals can be used for either estimation or hypothesis testing. A crucial problem of group testing is the appropriate choice of group size in dependence of the number of groups, an assumed true proportion, and the threshold. This paper reports on experimental design to achieve high power for tests or low confidence interval width. Two agricultural applications are presented for which experimental design is discussed.

Key Words: *group testing; confidence limit; experimental design; power; seed testing*.

INTRODUCTION

Group testing is used for inference on small binomial proportions if the assay method is sensitive but expensive (Thompson, 1962; Swallow, 1985; Tebbs and Bilder, 2004). Groups of individuals are characterized instead of single individuals, so that the possible outcomes are: a negative group if all individuals are negative or a positive group if at least one individual is positive. Interest might still be in hypothesis testing or estimation of the proportion of positive individuals. For example, a common problem in seed-quality control is to prove that a certain detrimental trait is rare in a population of individuals. Regulation No. 1829 of the European Union (Anonymous, 2003) requires that the proportion of genetically modified (GM) seed impurities be lower than 0.005 in a seed lot. In this case, a threshold is defined a priori, and interest is in rejecting the H_0: $\pi \geq 0.005$ in favor of H_1: $\pi < 0.005$. This can be performed by a one-sided test or, alternatively, an upper confidence limit for π can be estimated. The null hypothesis is rejected if the confidence limit is lower than 0.005. In other situations, no *a priori*-defined threshold is available. Similar problems arise in plant breeding,

and surveillance programs for plant, animal or human diseases (Gu et al., 2004), where a population is regarded as unsafe or not marketable if the proportion of an unwanted trait exceeds a small threshold level.

In group testing, experimental design is a crucial issue. The inappropriate choice of group size for a given number of groups and given proportion can result in bias of the point estimator (Swallow, 1985), decrease of power and increased confidence width. Therefore, this paper focuses on experimental design under different restrictions and considers different methods for estimation of upper confidence limits. Power is of main interest for hypothesis testing against a priori known thresholds. If the only objective is to estimate confidence limits, experimental design can be used to achieve a minimal expected confidence interval width.

ASSUMPTIONS AND NOTATION

Interest is in the estimation of the proportion π of a rare trait in a population. The event that an individual shows this trait is assumed to be independent and identically distributed (i.i.d.) Bernoulli for each individual in a population. In group testing, with equal group sizes, ns individuals are assigned randomly to n groups or bulks, each containing s individuals. Biological or biochemical assays are performed on each group, and the outcome is denoted 'positive' if at least one positive individual is in the group and is denoted 'negative' only if all individuals in the group are negative. It is assumed that the assay method has sensitivity and specificity 1, i.e., it is able to detect a group as positive if at least one positive individual is present in the group of size s. The number of positive groups is denoted y, and $t = y/n$ is the estimated fraction of positive groups. Here y is the realization of a binomial random variable $Y \sim Bin(n, \theta)$, where θ is the unknown probability of finding a positive group. In group testing, information on the proportion π is gained from negative groups. The probability of finding a negative group is $1-\theta = (1-\pi)^s$. An estimator for π can simply be derived by replacing θ by its estimator t: $p = 1-(1-t)^{1/s}$. This estimator is positively biased (Swallow, 1985); the bias becomes large as the probability of observing only positive groups becomes large.

Note that the statistical methods described below are valid only if assumptions of perfect sensitivity and specificity are met. Assays with sensitivity < 1 might lead to point estimates with less positive and even negative bias; the corresponding confidence intervals can have coverage probability < $(1-\alpha)$, and tests can have size > α. Assays with specificity < 1 lead to overestimation of π, corresponding confidence intervals and tests are too conservative. Especially, too large a group size might decrease assay sensitivity. Therefore, in addition to the statistical aspects discussed below, assay sensitivity should be considered when planning experiments.

ONE-SIDED CONFIDENCE LIMITS FOR A PROPORTION ESTIMATED FROM GROUP TESTING

Tebbs and Bilder (2004) investigated seven methods for construction of confidence intervals for π, which can be constructed directly on the scale of individuals or on the scale of groups. In the second approach, a confidence limit t_U is constructed for the proportion of 'positive groups' θ, and an upper confidence limit p_U for the probability π then is: $p_u = 1-(1-t_u)^{1/s}$. This approach is attractive because of computational simplicity and the possibility of making use of well-described interval methods for single binomial proportions, and, thus, it will be used below.

A large variety of methods is available for estimation of confidence intervals for a binomial proportion. The upper exact $(1-\alpha)$ confidence limit (Clopper and Pearson, 1934) can be calculated using the quantiles of the F-distribution:

$$\left[0;\left(1+\frac{n-y}{(y+1)F_{2(y+1),2(n-y),\alpha}}\right)^{-1}\right],$$

where n denotes the number of groups, y is the number of observed positive groups, and F is the quantile of the F-distribution.

The mean coverage probability of several recently recommended asymptotic confidence limits is closer to the nominal level than that of the exact Clopper-Pearson limit, but they allow violation of the nominal level. The widely known Wilson-Score limit performs well for even moderate to small sample sizes if two-sided estimation is considered, but the upper limit is conservative for proportions close to 0 and liberal for proportions close to 1 (Cai, 2005). The upper (1–α)-Wilson-Score confidence limit can be calculated using the quantiles z of the standard normal distribution:

$$\left[0;\left(t+\frac{z_{1-\alpha}^2}{2n}+z_{1-\alpha}\sqrt{\left(t(1-t)+\frac{z_{1-\alpha}^2}{4n}\right)/n}\right)\Bigg/\left(1+\frac{z_{1-\alpha}^2}{n}\right)\right],$$

with $t=y/n$.

Because asymmetry of coverage is undesired particularly for one-sided estimation, Cai (2005) proposed the Second-order-corrected limit. Its upper (1–α) - limit is given as:

$$\left[0;\frac{(y-c)}{(n+2c)}+z_{1-\alpha}\sqrt{t(1-t)+\frac{k_1t(1-t)+k_2}{n}}\Bigg/\sqrt{n}\right], \text{ where } c=\frac{1}{3}z_{1-\alpha}^2+\frac{1}{6},\ k_1=-\frac{13}{18}z_{1-\alpha}^2-\frac{17}{18}$$

and $k_2=\frac{1}{18}z_{1-\alpha}^2+\frac{7}{36}$, and z is the quantile of the standard normal distribution. The Second-order-corrected interval may give upper limits larger than 1 for small sample sizes and may exclude the estimate p=1 in some cases of $y=n$ for large sample sizes. Therefore, its upper limit should be restricted to [0, 1] and additionally is defined as 1 for the case $y=n$ in the following applications.

However, if in practice the case $y=n$ is observed for a certain group size s, one can perform an additional number of assays with a reduced group size. Subsequent point and interval estimation for π, based on the combined results of the two experimental steps, can be performed via the methods described in Hepworth (1996, 2005).

EXPERIMENTAL DESIGN

In the following section, experimental design for group testing will be considered from different viewpoints. Both number of groups n and group size s influence the performance of the confidence interval methods. Three different practical situations are considered: 1) Group size s might be restricted due to assay sensitivity, and the number of groups n has to be chosen. 2) The number of assays n is limited by cost, but the total number of units ns and the group size s can be chosen without serious limitation. 3) The total number of units ns might be limited and at the same time the number of assays n might be limited.

In group testing, the probability $\Pr(Y=y)$ depends on the binomial parameter θ, where $\theta = 1-(1-\pi)^s$:

$$\Pr(Y=y\mid n,s,\pi)=\binom{n}{y}\left(1-(1-\pi)^s\right)^y(1-\pi)^{s(n-y)} \quad (1)$$

By multiplying with an indicator function $I()$ for a certain event and summation across all possible realizations of y=0, …, n, the expectation of this event can be calculated for given parameters. Here, power will be defined as the probability $\Pr(p_U<\pi_0\mid\pi)$ that the upper limit p_U excludes a threshold proportion π_0. To calculate power for a given nominal α, the indicator function $I(y, n, s, \pi_0) = 0$ if the limit does not exclude the threshold π_0, and $I(y, n, s, \pi_0) = 1$ if the limit excludes the threshold and consequently H_0 is rejected.

$$power(n,s,\pi,\pi_0)=\sum_{y=0}^{n} I(y,n,s,\pi_0)\binom{n}{y}\left(1-(1-\pi)^s\right)^y(1-\pi)^{s(n-y)} \quad (2)$$

Expected interval width, bias of the point estimator (Swallow, 1985) and coverage probability of confidence intervals can be calculated by equation (2) together with an appropriate definition of indicator functions. Defining $I(y, n, s)=p_v-p_L$ is straightforward in the case of two-sided intervals. For upper confidence limits, the expected difference between the assumed true parameter π and the upper limit $I(n, s, y, \pi)= p_v-\pi$ can be considered a criterion for the precision of the confidence limit. However, in practice π is unknown, and experiments have to be planned based on an assumed value of π. Here, an upper bound of interest for π might be defined using data from previous trials, or in quality control, the largest expected impurity might be defined by convention. It is assumed that there is no interest in precise estimation if the true proportion is larger than that value. If a design {*n*, *s*} has sufficient power or acceptable interval width for this assumed π, it will have higher power and lower interval width for all smaller proportions. When interest is in a precise estimation across a broad range of π, experimental designs using different group sizes following Hepworth (1996) are recommended.

CHOICE OF N WHEN S IS FIXED

In the first situation of a certain fixed group size *s*, the interest lies in finding an appropriate number of groups *n*. This problem is similar to that of sample size calculation in simple binomial testing (*s*=1) (Chernick and Liu, 2002). The same transformation $p=1-(1-t)^{1/s}$ is always applied if the group size *s* is fixed; therefore, the power properties known from *s*=1 are truncated towards smaller values of π. Power increases for increasing *n* in a non-monotone manner until power close to 1 is achieved for *n* smaller than in the situation *s*=1. Expected width of confidence limits shows a monotone decrease for increasing *n*.

CHOICE OF S WHEN N IS FIXED

The second situation, i.e., finding an appropriate group size *s* while other parameters are fixed, is more interesting than the previous situation. This problem arises if the assay method is sufficiently sensitive and specific across a wide range of group sizes and single individuals can be supplied without serious limitations. Figure 1 illustrates how increasing the group size *s* influences the coverage probability, the power to reject H_0: π ≥ 0.005, the expected distance between true proportion π and the upper limit for the three methods and the bias of point estimator. The confidence level of upper limits is 95%, and the assumed true proportion is π=0.003. Here, the number of groups is kept constant at *n*=30, but the total number of individuals *ns*, which contribute to the observations, is increased from 30 to 60000 by increasing the group size *s* from 1 to 2000. This leads to an increasing power for small proportions in group testing. If the group size is too large for a given π and *n*, the risk to observe *y*=*n* is increased, resulting in a greatly biased estimator. Therefore, power decreases for too large group sizes; simultaneously bias increases. Which group sizes *s* are actually too large depends not only on π but also on *n*.

The power to reject H_0: π ≥ 0.005 differs between methods because of different coverage probabilities. For a given π and method, a local maximum of power is found for those combinations of *n*, *s*, α for which the coverage probability has a local minimum for the case that H_0: $\pi = \pi_0$ is true (see Schaarschmidt, 2005 for details). The interval first decreases in comparison to the situation *s*=1, stays on a low level across a large range of *s* and increases largely for those group sizes that lead to a large bias of the estimator. The methods differ only slightly with respect to their minimal expected width and the corresponding group size.

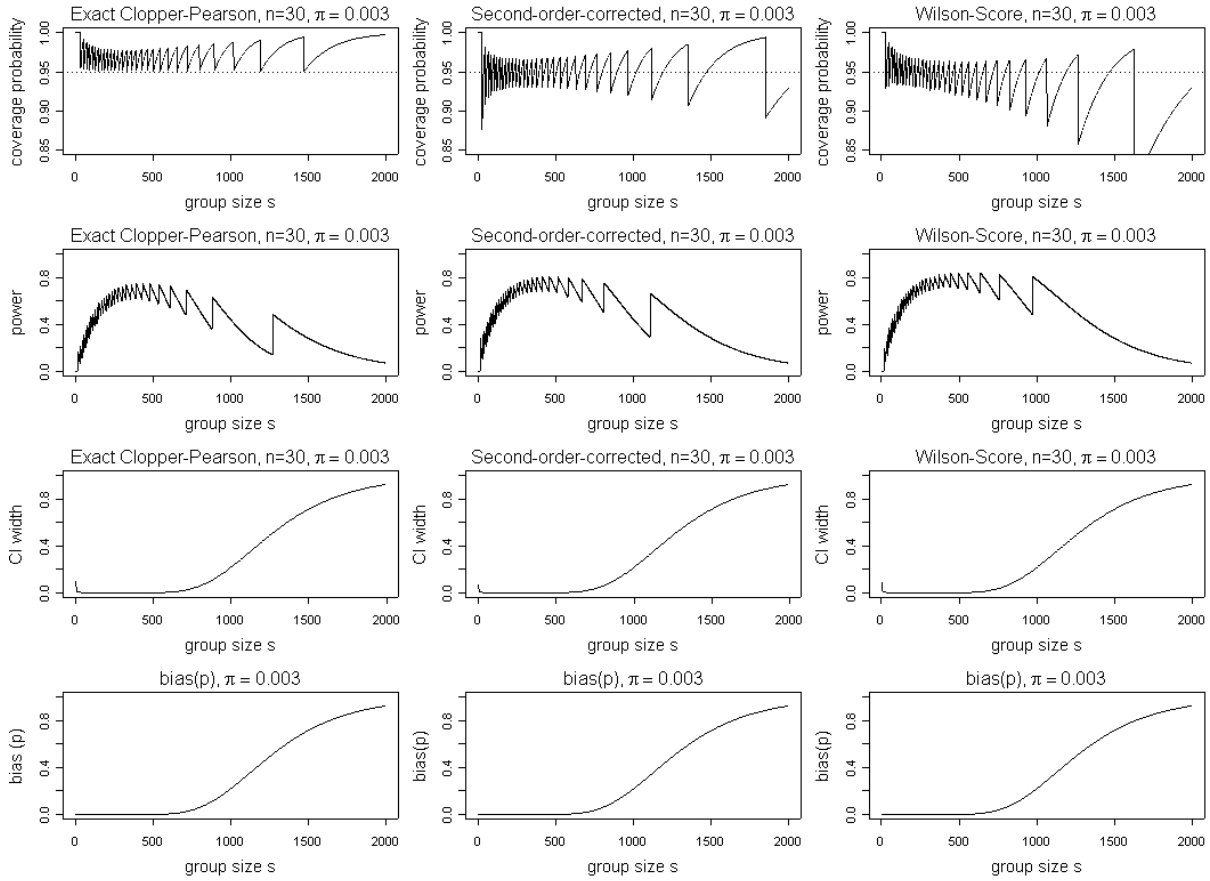

Figure 1. Coverage probability, power to reject H_0: $\pi \geq 0.005$, expected interval width and bias (p) for increasing group size s=1,…,2000, for the upper 95% exact Clopper-Pearson, Second-order-corrected and Wilson-Score limit in the situation n=30, π=0.003.

CHOICE OF S AND N WHEN TOTAL NUMBER OF INDIVIDUALS IS FIXED

In the third situation, supply of individuals for testing purposes is limited and also the assays are expensive and restrict the number of observations n. For example, in seed testing, only a limited number of seeds is available for destructive assay methods; or in surveillance of population of virus vectors (Gu et al., 2004), the sampling of large numbers of individuals might be very time consuming. Then, a given number of individual units might be divided into either many small groups or a few large groups. Here interest is in the question, which setting can be chosen to decrease assay costs without too large a loss of power or unacceptable increase of confidence interval width. Starting from the binomial case of evaluating all individuals separately, power does not decrease substantially if units are assigned to groups of increasing size, as long as the number of groups does not become too small and bias of estimator stays negligible. This is shown for one particular situation in Table 1: 2400 units are assigned to groups of different size; the probability of an upper 95% exact confidence limit to exclude 0.005 is calculated for the case that the true proportion is π = 0.002 for each resulting group testing design. For this small proportion, the assay affords (the number of assays n) can be reduced from n=2400 to n=200 by increasing group size from s=1 to s=12 without reducing power. Designs with lower number of groups and further increased group size lead to substantial loss of power compared to the binomial case n=2400, s=1.

Table 1. Upper 95% exact Clopper-Pearson limit: Power to reject H_0: $\pi \geq 0.005$ and expected interval width for the situation $\pi = 0.002$ and different number of assays n and group sizes s resulting in constant total number of units ns=2400.

Number of assays n	Group size s	Power	Expected limit width (p_U-π)	Bias (p)
2400	1	0.791	0.00223	0
1200	2	0.792	0.00223	0.0000007
800	3	0.793	0.00223	0.0000008
600	4	0.793	0.00224	0.0000013
400	6	0.795	0.00224	0.0000021
300	8	0.797	0.00225	0.0000029
200	12	0.800	0.00225	0.0000046
100	24	0.671	0.00228	0.0000099
50	48	0.693	0.00233	0.0000208
25	96	0.585	0.00246	0.0000449
10	240	0.679	0.00309	0.0002114
8	300	0.473	0.00509	0.0019196
6	400	0.556	0.03170	0.0281089

Unfortunately, the pattern of power development depends on the particular values of ns, α, π, π_0, and the method used. The positions $\{n, s\}$ of local power maxima depend on α, π_0, and the interval method only. They correspond to the local minima of coverage probabilities for these n, s, α for the case of $\pi = \pi_0$.

Software is needed to find appropriate designs under different restrictions. Functions for evaluation of group testing experiments, power calculation and sample size iteration are available for R (R Development Core Team, 2005) in the package binGroup (http://www.biostat.uni-hannover.de/research/thesis).

EXAMPLES

The first example is concerned with testing conventional seeds for presence of genetically modified (GM) organisms. Interest is in rejecting H_0: $\pi_{GM} \geq 0.005$ in favor of H_1: $\pi_{GM} < 0.005$. In an experiment (K. Weissleder and KWS Saat AG, 2005, personal communication), 63000 seeds were assigned to 21 groups, each containing s=3000 seeds. Twenty groups were found GM-negative, Y=1 group was GM-positive, resulting in the estimator $p_{GM} = 1-(1-1/21)^{1/3000} = 0.000017$. An exact upper 95% confidence limit for the proportion of GM-seeds in the population is $p_{U(GM)}$=0.000077. Based on the exact method, the null hypotheses can be rejected, because with 95% probability the proportion of GM-seed in the seed lot is not greater than about 0.00008, i.e., the threshold proportion 0.005 is not included in the upper confidence limit. The upper 95% confidence limits according to the Second-order-corrected and the Wilson-Score method are $p_{U(GM)}$=0.000064 and $p_{U(GM)}$=0.000069, respectively.

The above test controls the consumers risk at level α=5%, while the producers risk is the type-II error of the above procedure, i.e., the risk that a seed lot is considered to contain $\pi_{GM} \geq 0.005$ although this is not the case. Therefore, a sufficient power of tests in quality control is in the interest of companies concerned with marketing of seeds. Assume that not more than 21 assays per seed lot can be afforded, and interest is in a sufficiently low producer risk for true GM proportions up to π=0.003. Is the chosen group size s=3000 optimal or can a more appropriate group size be chosen?

In the left side of Figure 2, power for the given situation is calculated for increasing group size s=1,…,3000. The total number of seeds is increased from ns=21 to ns =63000. Power increases in a non-monotone manner by increasing group size and reaches its highest value for s=462. Power is decreasing for larger s because the probability, that all 21 groups are positive, increases. The right side of Figure 2 shows power and coverage probability to be dependent on π for group sizes s=462 (solid line) and s=461 (dashed line). The design with optimal group size s=462 has higher power than the design with group size s=461 for the whole range of π and the given alternative hypothesis. Note that coverage probabilities for both group sizes differ under the margin of the alternative hypothesis $\pi=\pi_0=0.005$. For s=462, coverage probability is 0.9503, while for s=461, coverage = 0.9861.

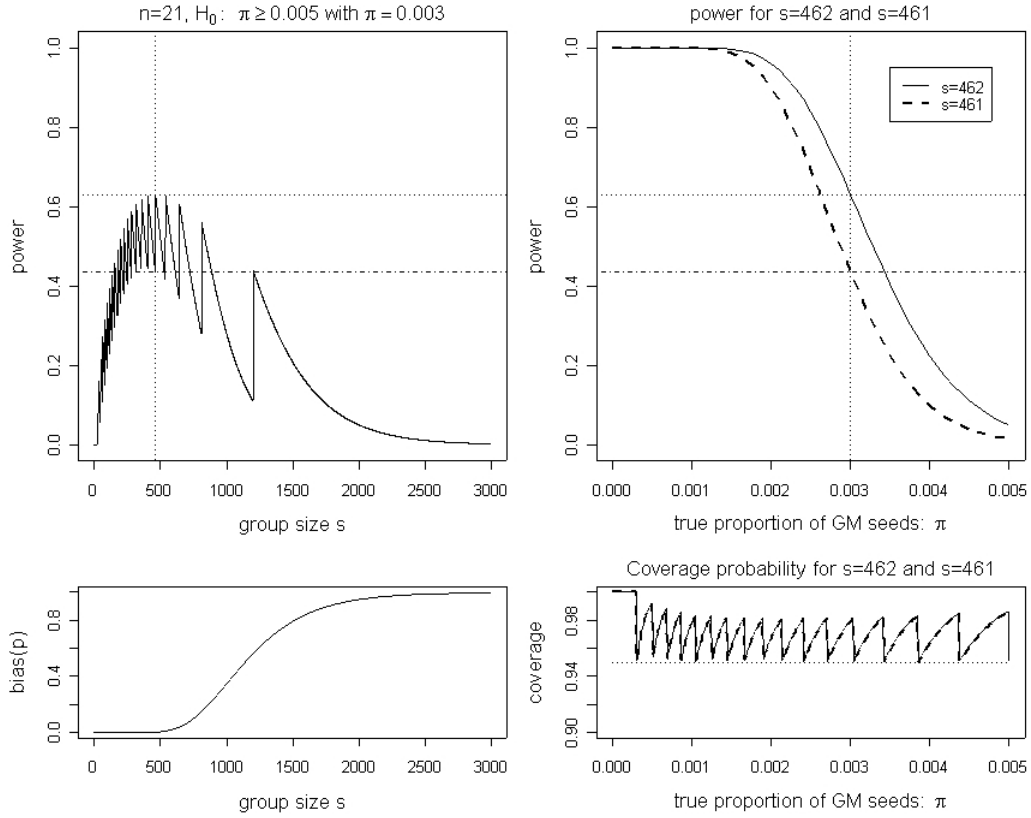

Figure 2. Dependency between power and choice of group size for the upper 95% exact Clopper-Pearson confidence limit for rejection of H_0: $\pi \geq 0.005$ if the number of groups is fixed n=21.

The second example relates to resistance breeding. A dominant gene-controlled resistance can be detected using a co-dominant molecular marker, i.e., single plants can be classified as resistant (RR, Rr) and susceptible (rr). The molecular marker is sensitive to detect a single R or r allele in bulk samples of up to 5 plants. Objective is to select lines with a low proportion of individuals carrying allele r, i.e., lines with a high proportion of true breeding RR individuals. In an experiment, 290 individual plants of a single inbred line were assigned to n=58 bulk samples, each consisting of s=5 plants (K. Weissleder and KWS Saat AG, 2005, personal communication). Fifty-seven groups showed response of R alleles only. The Y=1 group showed response of R and r allele, thus contained at least one r allele. The estimated proportion of non-RR individuals in the population is $p_{no\text{-}RR}=1-(1-1/58)^{1/5}= 0.003472$; an upper 95% Second-order-corrected confidence limit for the

proportion of non-RR groups is [0; 0.0665] and the corresponding 95% interval for the proportion of non-RR individuals is [0; 0.0137].

Here, experimental design might be based on expected interval width. If 290 individuals of an inbred line are used for testing, which combination {*n*, *s*} will result in limits that still show acceptable interval width? Below, it is assumed that the breeder is interested in a precise estimation of lines that contain $\pi \leq 0.01$, i.e., 1% of individuals with r-alleles.

Table 2. Expected interval width of upper 95% Second-order-corrected limits for π=0.01 and different combinations {*n, s*} resulting in a total number of units *ns* = 290.

Number of groups *n*	Group size *s*	Expected interval width for π=*0.01*	Interval width relative to *s*=1
290	1	0.01331	1
145	2	0.01336	1.004
58	5	0.01351	1.015
29	10	0.01377	1.035
10	29	0.01504	1.130
5	58	0.03436	2.581

If 290 individuals are assigned to groups of size *s*=1, 2, 5, 10, 29 or 58 individuals, upper 95% second-order-corrected limits will have the expected widths shown in Table 2, if the true proportion of individuals with susceptibility allele is π=0.01. Using a more sensitive assay method allowing larger group sizes, a design with *n*=29 assays and bulk sample size *s*=10 would result in only about 3.5% longer confidence limits than applying *n*=290 assays on each individual plant (*s*=1).

DISCUSSION

From a practitioner's point of view, the statistical assessment of small proportions of unwanted traits can be approached by estimation of upper confidence limits, which can be used for estimation and decision making concerning whether the proportion of detrimental individuals is lower than a certain threshold.

For estimation and hypothesis testing of small proportions using group testing, the choice of experimental design is crucial to achieve negligible bias (Swallow, 1985; Remund et al., 2001). We examined experimental design to achieve a sufficient power of the test procedure or a sufficient width of confidence intervals. Particularly, the group size is of importance: while too small groups can lead to the situation that the null hypothesis cannot be rejected at all, too large groups can result in a severe downturn of power and increase the interval width. Power functions in dependence of the number of groups or the group size are non-monotone because of the discrete nature of the binomial distribution. If a practitioner has at least some freedom to choose the number of groups *n* or the group size *s*, designs should be chosen that result in much higher power than others. For a given method and nominal level, which combination {*n,s*} results in a design with locally maximal power depends only on the *a priori*-defined threshold proportion.

REFERENCES

Anonymous (2003). Regulation (EC) 1829 of the European Parliament and the European Council of 22 September 2003 on genetically modified food and feed. *Official Journal of the European Union* L 268.

Cai, T.T. (2005). One-sided confidence limits in discrete distributions. *J. Stat. Plan. Infer.* 131, 63–88.

Chernick, M.R., Liu, C.Y. (2002). The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods. *Am. Stat.* 56, 149–155.

Clopper, C.J., Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404–413.

Gu, W., Lampman, R., Novak, R.J. (2004). Assessment of arbovirus vector infection rates using variable size pooling. *Med. Vet. Entomol.* 18, 200–204.

Hepworth, G. (1996). Exact confidence limits for proportions estimated by group testing. *Biometrics* 52, 1134–1146.

Hepworth, G. (2005). Confidence intervals for proportions estimated by group testing with groups of unequal size. *J.Agric. Biol. Envir. St.* 10, 478–497.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Remund, K.M., Dixon, D.A., Wright, D.L., Holden, L.R. (2001). Statistical considerations in seed purity testing for transgenic traits. *Seed Sci. Res.* 11, 101–119.

Schaarschmidt, F. (2005). *Group testing – design and analysis*. Thesis Fachbereich Gartenbau, Universität Hannover. http://www.biostat.uni-hannover.de/research/thesis

Swallow, W.H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology* 75, 882–889.

Tebbs, J.M., Bilder, C.R. (2004). Confidence limit procedures for the probability of disease transmission in multiple-vector-transfer designs. *J.Agric. Biol. Envir. St.* 9, 75–90.

Thompson, K.H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics* 18, 568–578.