Diplomarbeit im Lehrgebiet Bioinformatik

Betreuer und Erstprüfer: Prof. Dr. L. Hothorn

Zweitprüfer: Prof. Dr. B. Hau

Fachbereich Gartenbau

Universität Hannover

Binomial group testing – Design and Analysis

Frank Schaarschmidt

Dorotheenstr. 5A

30419 Hannover

Matrikelnummer: 1987106

9. Fachsemester

Abgabetermin: 14.03.2005

Index

Summary

The problem of estimating small binomial proportions and performing a proof of safety against very small threshold proportions is a methodology widely needed in seed production, breeding and epidemiology. For both estimation and performing a hypothesis test, confidence intervals can be used. Different methods for interval estimation for a binomial proportion are reviewed, discussed and compared with respect to coverage probability. All commonly used methods perform poor for small sample sizes and small proportions, but because of expensive assay methods a methodology with sufficient coverage probability and power is needed for these cases. The binomial group testing approach was shown to perform well in estimation of small proportions using a comparatively low number of observations. The main part of the thesis concerns about the use of binomial group testing for performing a proof of safety if contaminations exceeding a small threshold are regarded as unsafe. Different confidence interval methods are compared in binomial group testing with respect to coverage probability, power and sample size requirement. The effect of the different parameters of experimental design on estimation, coverage probability and power is discussed. Especially the possibility to greatly improve power for a limited number of assays and the need for controlling bias in experimental design is shown. Finally, real data examples from different areas of application are evaluated, including discussions of their experimental design.

1 Introduction

1.1 The problem: Testing for small proportions using expensive assays

This diploma thesis concerns about estimation and hypothesis testing of small proportions. The proportion here describes the probability of a single individual to exhibit one of two distinct stages: '0' or '1', 'infected' or 'not infected', 'positive' or 'negative' result of an assay. These so called dichotomous data can be natural or generated from continuous data by a fixed or optimal selected cut point. A main problem in analysis of dichotomous data is discreteness of observation: If a sample of n individuals is randomly taken from a population, one will observe a discrete number of positive individuals: 0,1,2,…,*n*.

If f.e. positive individuals are very rare, e.g. less than 1% in a population, one usually needs to evaluate many single individuals to estimate the true proportion. If only a small sample is taken from a population, the probability to observe positive individuals at all is low.

If the trait of interest of an individual can be observed very easily, evaluation of single individuals provides most information. This might be the case in assessing f.e. viability of seeds or visible disease symptoms on plants. If the character of interest can only be observed by application of laborious, expensive assays, it might be desired to perform only few assays, resulting in only few observations for estimation. Binomial group testing compromises the economical limitation of number of assays with the need to include many individuals into the experiment.

In a binomial group testing experiment, groups of individuals are pooled, and each pool is evaluated whether no or an unknown number of positive individuals are member of the pool. The group is classified positive, if one or more members are positive and is classified negative, if no member is positive. The few resulting observations are based on a larger number of individuals.

For compromising few observations and a high number of contributing individuals, a price has to be paid: information is lost, additional assumptions need to be fulfilled and experimental design has to be chosen carefully.

1.2 Areas of application

Binomial group testing can be useful, if low proportions shall be estimated or hypothesis tests against very small thresholds shall be performed and the character of interest can be observed only using expensive assays, f.e., PCR methods, other molecular markers or ELISA.

1.2.1 Contamination by GMO

The Regulation 1829 of the European Parliament and Council (Anonymous, 2003) requires labelling of food and feed containing GMO unless it contains less than 0.9 % GMO, if these GMO are approved in the European Union. Contaminations with GMO not approved in the European Union, which are assessed as favourable by the European scientific committee are allowed to be present until a threshold of 0.5%.

According to the Deutsche Gentechnikgesetz (GenTG) § 14 (2a), the regulations for releasing and marketing of a GMO containing product (§ 14, 15, 16) are not applied, if it contains less than 0.5% of the GMO and several additional conditions met.

Because of this, seed companies might want to ensure, that their marketed seed lots have contents of GMO below these thresholds. This is of special importance in species where outcrossing during mass production of seed can hardly be totally controlled, because pollen is disseminated by wind, as are maize (*Zea mays*), rhy *(Secale cereale)* or sugar beet *(Beta vulgaris*).

For any decision based on sampling from a population, false positive and false negative decisions may occur, the risks of erroneous decisions can not be avoided, unless $n \to \infty$.

In this context the terms ‘consumers risk’ and ‘producers risk’ are frequently used (Remund et al. 2001): The consumers risk is the probability that a seed lot is allowed to be marketed although its content of GMO is higher than the threshold, whereas the producers risk is the probability that a seed lot is not allowed to be marketed, although its GMO content is below the threshold. Depending on the formulation of hypotheses, these terms correspond to the false positive and false negative error rates in hypothesis testing, the focus of section 4.

Assays for GMO detection

Depending on the assay method, further sources of erroneous decisions can occur: The probability of an assays, to detect a positive individual as positive is called sensitivity, while the probability that the assay detects a negative individual as negative is called specificity (Sachs,1991).

For GMO detection, various methods are reviewed in the references, which differ in their sensitivity and specificity. GMO can be detected either by labelling special gene products with antibodies or by detection of special DNA sequences in a sample using PCR or related methods. DNA-based methods have the advantage to be very sensitive (Giovannini et al. 2002). Sensitivity is of main importance because a sufficient sensitivity of the assay method is crucial for the validity of group testing methods.

According to Holst-Jensen et al. (2003) the level of detection (LOD) of a PCR assay can be defined in different scales, either as absolute copy number of the target sequence or as percentage of the target molecule relative to the total DNA. The level of detection mainly depends on the quality of extracted DNA, on the presence of inhibitory substances, or can be reduced due to processing of the material in food production. Jankiewicz et al. (1999) thus distinguished between theoretical LOD in serial dilutions of DNA from seeds and practical LOD in processed material with known content of GMO. They found a theoretical LOD of 0.005 % or 30 copies for Round-up-Ready soybeans and 0.005 % or 9 copies for a Bt maize variety. The practical LODs were 0.1 % for both, which is close to the threshold.

Moreover, assay methods might differ in their specificity: Holst-Jensen et al. (2003) classify different approaches according to the target sequence and the resulting sensitivity. Primers for target sequences, which were introduced in most GMO (certain vector, promotor sequences or resistance genes) can be used for general screenings on GMO, but might have higher false positive rates. More specific primers for a certain gene, a certain construct of promotor, gene, terminator and vector sequences or for a special transformation event allow more specific screening, but require more prior knowledge of sequences. Although methods might differ in sensitivity and specificity, the statistical examinations in this thesis assume both =1. The user of the shown

methodology has to ensure that these assumptions come true for the chosen assay method, species, organ or tissue type and the size of bulk samples.

1.2.2 Plant Breeding

In plant breeding it can be of interest to show that a certain trait (f.e. susceptibility for a disease) reaches only a small incidence in breeding material. An objective might be selection of inbred lines or other breeding material with a low level of plants carrying a certain allele using molecular markers. Here a clear threshold is unlikely to exist, so estimation using confidence intervals is more important.

1.2.3 Vector transfer design

Many viral plant diseases are transmitted during the process of feeding of certain animal species on the host plant. The rate of disease transmission by these so-called vectors is of importance for the epidemiology of these diseases. Vector transfer designs have the aim to estimate the probability of a single vector to transmit the disease to a plant: a certain number of infected vectors are placed on each of n isolated healthy test plants. Then a known number of vectors have the opportunity to transmit the disease to each test plant by feeding on it. After sufficient periods of feeding and allowing appearance of disease symptoms, each single plant is evaluated for showing the disease or not. Each infected plant has received the disease from at least one of the vectors. The plants and their evaluation are the biological assay, where the number of plants might be limited due to capacity of greenhouses and isolation cages or by the costs of performing an ELISA or PCR on each single plant for disease detection (Tebbs and Bilder 2004, Swallow 1985).

1.2.4 Estimation of infection rates, incidence of viruses

In epidemiological studies it might be of interest to estimate the incidence of individuals carrying the pathogen in natural vector populations. If this incidence is expected to be small, group testing might be applied. The caught individual vectors can be randomly assigned to bulk samples and the assays for detecting the virus are performed on these single bulks. For example, Tedeschi et al. (2003) used group testing to estimate the rate of infection with the Apple

Proliferation phytoplasma in natural populations of the vector of this disease, a psyllid.

This application is also interesting for epidemiology of some human diseases: Gu et al. (2004) used group testing to assess the incidence of West Nile virus (WNV) in natural populations of mosquito.

1.2.5 Seed borne pathogens

For the control of seed borne diseases, incidence of the viable pathogen in a marketed seed lot has to be very low to avoid economical losses due to the disease. Seed companies might want to fulfil internal quality standards. Walcott (2003) reviewed methods for detection of seed borne pathogens. Combinations of PCR with pre-enrichment of viable cells using selective media (BIO-PCR), enrichment of cells using antibodies (IMS-PCR) or enrichment of DNA-fragments using single stranded probes (MCH-PCR) can achieve much higher sensitivity than conventional detection methods, but are expensive. Thus, group testing might be applied for screening seed lots on seed borne pathogens.

1.3 Assumptions

In the following it is assumed that:

1. The probability to show the trait of interest is independent and identically distributed Bernoulli (π) for each unit in the population. This is also assumed for simple binomial testing.

For group testing, additional assumptions are needed:

2. The units are randomly assigned to the groups.
3. All groups contain the same number of units.
4. The chosen group size does not influence the probability of a single unit to show the trait of interest. The sensitivity of the assay must be sufficient to detect a group as positive if only one single member is positive.
5. The assays do not misclassify groups, i.e. they have sensitivity=1 and specificity=1 and do not vary in specificity and sensitivity.

(see Tebbs and Bilder, 2004)

The experimenter has to ensure, that these assumptions are fulfilled. To fulfill assumption 1 the sample of individual units has to be taken representative for the population and has to account for possible clustering f.e. of GMO in big seed lots (Remund et al. 2001).

Assumption 3 might be problematic in vector transfer designs, where single vectors might escape or die and decrease the group size in this case. Hepworth (1996, 2004) reviews and proposes methods for evaluation of group testing experiments with variable group size (see section 5.1).

Assumption 4 greatly depends on the assay method: For vector transfer designs, the feeding behavior of insects might change with increasing number of individuals on one plant. Using ELISA, a decreasing sensitivity to detect single positive units in a group can be expected with increasing group size, due to dilution effects or an increasing proportion of inhibitory substances. Hung and Swallow (1999) examined the problem of dilution effects and testing errors for binomial group testing and give recommendations for experimental design (see section 5.2), if violation of assumptions 4 and 5 can not be excluded.

1.4 Notation and definitions

π denotes the binomial probability of a single unit to be 'positive', where π is assumed to be the same for all individuals in the population.

π_0 denotes the threshold proportion to test against in a hypothesis test. π_0 corresponds to the LQL (lowest quality level) in guidelines and papers on seed testing in common and GMO-testing in special (Remund et al. 2001).

n denotes the number of observations, i.e. for binomial group testing the number of groups tested or the number of assays performed.

s is the number of individuals per group i.e. the group size. It is assumed to be equal for all groups. *n*s* then is the total number of individuals in the experiment. The observations are performed on the group level: A group is counted positive if at least one individual in the group is positive. These observations are assumed to be free of misclassifications.

θ binomial probability of a group to be positive, depending on π and the group size *s*.

Y *Y* is the observed number of positive groups, i.e. a certain realization of the random variable *y* following a binomial distribution *Bin(n, θ)*.

t *t=Y/n* is the observed fraction of positive tested groups, the estimator of θ.

p *p* is the estimated probability π of a single individual to be positive.

α denotes the probability of rejecting a null hypothesis in case that it is true, commonly called error of first kind.

β denotes the probability not to reject the null hypothesis although the alternative hypothesis is true, commonly called error of second kind.

E denotes the expected value of the random variable, which is defined as $E(y)=\sum_i Y_i P(y=Y_i)$ for a discrete random variable *y* (Sachs, 1991).

Var denotes the variance of a random variable, i.e. the expected value of the squared difference between its single realizations and its expected value, which is $Var(y)=\sum_i (Y_i - E(y))^2 P(y=Y_i)$ for a discrete random variable *y* (Sachs, 1991).

Bias denotes the bias of an estimator, i.e. the difference between the true unknown parameter and the expected value of its estimator (Sachs, 1991).

MSE denotes the mean square error of an estimator, which is the expected value of the squared difference between the true unknown parameter

and its estimator and is used as a measure for the goodness of an estimator. In case of an unbiased estimator, *MSE*=*Var*, whereas for a biased estimator it includes the variance and the square of its bias (Kotz and Johnson 1985, Sachs 1991).

'(*1-α*)*100%-confidence interval' (CI) for a parameter π is defined as an estimated interval [p_L,p_U] fulfilling the condition that $P\left(\pi \in \left[p_L, p_U\right]\right) \geq 1-\alpha$

'Consumers risk' denotes the probability that a marketed product does contain contaminations higher than the threshold.

'Producers risk' denotes the probability that a product is not allowed for marketing although its contamination is below the threshold.

'Coverage probability' denotes the actual probability of a confidence interval to contain the true parameter $P\left(\pi \in \left[p_L, p_U\right]\right)$.

'Power' denotes the actual probability to reject the null hypothesis in a case where the alternative hypothesis is true, i.e. the probability 1-β, in case of a confidence interval it is $P\left(\pi_0 \notin \left[p_L, p_U\right]\right)$

2. Point estimation

2.1 Estimator for π

A single group is assumed to be positive in the assay if *1,…,s* individuals units of the group are positive. Thus, several events on the individual scale can lead to the appearance of positive groups. According to the assumptions, negative groups can only result from one event on the individual scale, which is: none of the s individual units in the group is positive.

Therefore the probability *1-θ* of a group to be negative equals the probability that s negative individuals were assigned independently to the group:

$$1-\theta=(1-\pi)^s$$

From this equation the estimator p is derived by solving for π and replacing θ by its estimator t:

$$p=1-(1-t)^{1/s}$$

In group testing, the number of negative groups provides the information, that no individual in this group was positive, while positive groups can result from 1 to s positive individuals in a group. The estimator p is biased due to the fact that the information about the events leading to a positive group is very limited. Bias increases as the probability of observing only positive groups increases. This occurs if the group size s is chosen too high in relation to the unknown probability π. Bias and variance of the estimator can be calculated from n, π and s using the formulas in Swallow (1985).

2.2 Probability to observe y=Y

The probability to observe a certain realization *Y* in a group testing experiment depends on n and the binomial probability θ to observe a positive group:

$$\Pr(y=Y\mid n,\theta)=\binom{n}{Y}\theta^Y(1-\theta)^{n-Y}$$

Because θ itself depends on π and the group size *s*, it can be calculated from *n, s,* π as

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

This can be used for closed calculation of the expected value of the estimator *p*, its bias, variance and mean square error and later on for the closed calculation of coverage probability, power and interval length of confidence intervals in group testing experiments.

2.3 Expected value of p

The expected value of the estimator p then is:

$$E(p) = \sum_{y=0}^{n} 1 - \left(1 - \frac{y}{n}\right)^{\frac{1}{s}} \binom{n}{y} \left(1 - (1-\pi)^s\right)^y (1-\pi)^{s(n-y)}$$

(see Swallow, 1985, with different notation)

2.4 Bias of estimator p

$$Bias(p) = E(p) - \pi$$

2.5 Variance of estimator p

$$Var(p) = E\left(p - E(p)\right)^2$$

$$Var(p) = \sum_{y=0}^{n} \left(1 - \frac{y}{n}\right)^{\frac{2}{s}} \binom{n}{y} \left((1-\pi)^s\right)^{n-y} \left(1 - (1-\pi)^s\right)^y - \left(1 - E(p)\right)^2$$

see Swallow (1985), with different notation.

2.6 MSE of the estimator p

$$MSE(p) = Var(p) + \left(Bias(p)\right)^2$$

3 Methods for interval estimation for the binomial parameter π

Confidence intervals take the uncertainty of estimation into account. Assuming a certain distribution of the estimator, the confidence interval promises to contain the true, unknown parameter within its limits with a high, pre-specified probability. Thus confidence intervals allow for both, estimation with a specified probabilistic precision and hypothesis testing with a controlled type-I-error rate α.

Hypotheses of interest in testing might be one-sided

H_0: $\pi \geq \pi_0$ vs. H_1: $\pi < \pi_0$

H_0: $\pi \leq \pi_0$ vs. H_1: $\pi > \pi_0$

or two-sided

H_0: $\pi = \pi_0$ vs. H_1: $\pi \neq \pi_0$

The null hypothesis is rejected with error probability α, if the (1-α)-confidence interval does not contain the threshold value π_0.

The proof of safety

In the case of testing a seed lot for a low content of contaminations, the decision on the hypotheses

H_0: $\pi \geq \pi_0$ vs. H_1: $\pi < \pi_0$,

requires a upper (1-α)-confidence limit [0; p_U]. The probability to erroneously decide against H_0 by excluding π_0 is then controlled only upwardly and the consumers risk =α. This is known as 'proof of safety': the seed lot is assumed to be hazardous under H_0 and is only allowed for marketing, if H_0 is rejected, i.e. the seed lot has been shown to be 'safe'. The producer of seed is interested in controlling the risk to classify a seed lot as contaminated although its proportion of GMO is in fact lower than the threshold, what corresponds to controlling β via choice of experimental design (section 4.5).

The proof of hazard

Oppositely, one might approach this problem testing the hypotheses

H_0: $\pi \leq \pi_0$ vs. H_1: $\pi > \pi_0$,

then H_0 will be rejected if the threshold π_0 is not included in the lower confidence limit of a confidence interval $[p_{L,1}]$. This is known as a 'proof of hazard': under H_0, the seed lot is assumed to be 'safe', i.e. equal or below the threshold. If H_0 is rejected, one can conclude that the proportion of GMO is over the threshold, i.e. the seed lot is 'hazardous'. Because of the falsification principle one can not conclude that the seed lot is 'safe' in case that H_0 is not rejected. Here the consumers risk equals β, while the producers risk is controlled via the pre-specified error rate α.

Also for testing against other contaminations than GMO or in situations where low contents of an undesired trait shall be estimated, it seems more reasonable to perform a proof of safety than to perform a proof of hazard. Because of this, main focus will be on the performance of upper confidence limits: with respect to the actual consumers risk (section 4.3, 4.4) and to experimental design with the objective of a low producers risk (section 4.5).

3.1 Confidence intervals for simple binomial testing

The following methods can be applied for binomial probabilities, thus either for CI estimation for a binomial probability π in a simple binomial experiment with evaluation of each unit (s=1), or for CI estimation for the probability θ in a group testing experiment. In the following, they are given in the notation to estimate a simple binomial probability π.

3.1.1 Confidence intervals based on normal approximation

The following asymptotic methods all assume gaussian distribution of their estimator.

Wald interval

The Wald CI is based on the large sample approximation that

$$\frac{(p-\pi)}{\sqrt{p(1-p)/n}} \sim N(0,1).$$

Then a nominal 95%-confidence interval for π is

$$\left[p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right]$$ (Santner and Duffy,1989)

This interval is known to be liberal (Santner and Duffy, 1989, Brown et al., 2001) and will therefore not be used for construction of CI on the group scale. But it will be shown for comparison of CI for simple binomial experiments because it was often used as a standard method in textbooks and software.

Wilson score interval

This interval introduced by Wilson (1927), is derived from the inversion of the Score test, which compares the observed proportion with a hypothetical proportion using the standard error under the null hypothesis. Blyth and Still (1983, normal approximation I) derive the Wilson Score Interval and the Wilson Score Interval with continuity correction from the following test:

H_0: $\pi = \pi_0$ is rejected if

$$\frac{|X - n\pi_0|}{\sqrt{n\pi_0(1-\pi_0)}} > c \text{, with } P\{|Z| \le c\} = 1 - \alpha \text{ for a Standard Normal Z.}$$

The standard error under a hypothetical binomial parameter $\sqrt{n\pi_0(1-\pi_0)}$ is used here instead of the standard error estimated from the sample $\sqrt{np(1-p)}$, which is used in the test underlying the Wald interval for a simple binomial parameter. This main difference to the Wald interval for a simple binomial proportion results in much better coverage probabilities of the two-sided Wilson Score Interval and the methods derived from it. The midpoint of this interval is shifted towards 0.5, depending on the chosen α. The shifting of the interval midpoint compared to the point estimate is stronger for small point estimates and small n. Additionally, the width of the interval becomes larger for small point estimates and smaller for point estimates near 0.5, compared to the Wald-Interval. For details see Agresti and Coull (1998), Brown et al. (2001).

The two-sided (*1-*α)-Wilson score Interval for π is

$$\left[\left(p+\frac{z_{\alpha/2}^2}{2n}\pm z_{\alpha/2}\sqrt{\left[p(1-p)+\frac{z_{\alpha/2}^2}{4n}\right]\Big/ n}\right)\Big/\left(1+z_{\alpha/2}^2/n\right)\right]$$

(Agresti and Coull, 1998; Piegorsch, 2004)

A modification exists, removing the most extreme violations of nominal confidence level for small values of π (see Brown et al. 2001).

Generalized Agresti-Coull-Interval

The generalized Agresti-Coull Interval (Piegorsch, 2004, Brown et al., 2001) uses a re-centered estimator $\tilde{p}=\left(y+\frac{1}{2}z_{\alpha/2}^2\right)\Big/\left(n+z_{\alpha/2}^2\right)$. Then the two-sided (*1-*α)-confidence interval for π is:

$$\left[\tilde{p}\pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}\right] \text{ where } \tilde{n}=n+z_{\alpha/2}^2$$

Since this interval is derived from the Wilson Score interval, it shows similarly good coverage probabilities, but is more conservative for binomial probabilities near 0 (Brown et al., 2001). Additionally it has the disadvantage to have lower limits p_L < 0 and upper limits p_U > 1 for *Y* close to 0 and *Y* close to *n*, respectively.

Add-4-Interval

The add-4 method is a simplification of the generalized Agresti-Coull method for the two-sided case and α = 0.05. The term $z_{\alpha/2}^2=1.96^2$ is replaced by 4, the re-centered estimator is simply $\tilde{p}=(y+2)/(n+4)$ and $\tilde{n}=n+4$.

Because this two-sided CI only is a special case for α = 0.05, it cannot be expected to maintain good coverage properties for other cases, f.e. if a one-sided 95%-CI shall be constructed or another confidence level is required. Therefore it will not be used for further examinations.

The add-4- and generalized Agresti-Coull-Interval have the same re-centered midpoint $\tilde{p}$ as the Wilson-interval (Agresti and Coull, 1998, Brown et al., 2001)

and are slightly broader in length. Thus they show the same shifting of the interval towards 0.5, depending on the estimator, *n* and α, but are slightly more conservative.

Wilson Score Interval with continuity correction

Blyth and Still (1983) recommend this interval for moderate sample sizes ($n \geq 30$), for which their improved exact interval is not tabulated anymore. It is derived from the same approximation as the Wilson Score Interval but *Y* is replaced by (*Y*+0.5) for calculating the upper bound and by (*Y*-0.5) for calculating the lower bound. Additionally, it is defined to have a lower bound = 0 for *Y* = 0 and an upper bound = 1 for *Y*=*n*. The two-sided (*1*-α)- Confidence Interval for π then is.

$$\left[\frac{(Y \pm 0.5) + z_{\alpha/2}^2 \Big/ 2 \pm z_{\alpha/2}\sqrt{(Y \pm 0.5) - (Y \pm 0.5)^2 \Big/ n + z_{\alpha/2}^2 \Big/ 4}}{n + z_{\alpha/2}^2}\right]$$

The continuity correction results in an actual confidence deviating only upwardly from the nominal level (for details see Blyth and Still, 1983), so this CI shows an actual coverage probability much higher than the nominal level (Brown et al. 2001b), thus cannot be recommended.

3.1.2 Further approaches for binomial confidence intervals

Brown et al. (2001) reviewed additional methods as the Jeffreys prior interval (an originally bayesian approach, but corresponding to a mid-p version of the Clopper-Pearson CI) and its modification, the Logit interval and the arcsine interval. In terms of coverage, they are not superior to the intervals recommended, because for some values of π they are more conservative for others they are more liberal than the Wilson or Agresti-Coull interval (Brown et al., 2001).

3.1.3 Exact confidence intervals

Exact confidence intervals guarantee the nominal confidence level. But because of discreteness of the binomial distribution they tend to be conservative. An exact confidence interval is derived from an exact test for π. The probability of a certain realization *Y* under the null hypothesis $\pi = \pi_0$ can be

calculated exactly. The null hypothesis is rejected if the probability of occurrence of the observed *Y* under the null hypothesis is lower than a pre-specified level α.

The idea of exact confidence intervals for π is then to perform an exact tests on each null hypothesis $\pi = \pi_0$ in $\pi_0 = 0,\ldots,1$. The confidence region for π consists of all values of π_0 for which the null hypothesis was not rejected for the observed *Y*. The resulting confidence interval has a confidence coefficient $\geq 1-\alpha$ if the family of tests performed has a size $< \alpha$ (Blyth and Still, 1983).

Exact confidence intervals can be calculated fulfilling different conditions (see Blyth and Still, 1983). If the interval requires a probability to exclude true π of less than $\alpha/2$ in each of both tails, this results in the longer, more conservative Clopper-Pearson Interval, whereas requiring a probability to exclude true π of less than α for both tails together results in shorter intervals, which are subsets of the Clopper Pearson Interval (Blyth and Still, 1983).

Because of this, the upper or lower bound of a two-sided $(1-2\alpha)$- Clopper-Pearson-confidence-interval can be used as a one-sided $(1-\alpha)$-Confidence interval, as recommended by Reiczigel (2003). This cannot be done using the bounds of the Blyth-Still Interval (see Agresti and Min, 2001). Agresti and Min (2001) recommend to define special one-sided CI for one-sided hypothesis.

Clopper-Pearson interval

The Clopper-Pearson Interval $[p_L, p_U]$ is constructed by inversion of an exact test to fulfill the conditions, that the probability of p_L to be greater than the true value is $\leq \alpha/2$ and the probability of p_U to be smaller than the true value is $\leq \alpha/2$, too. The upper and lower limit can also be calculated using the beta-distribution or the F-distribution (Santner and Duffy, 1989, Tebbs and Bilder, 2004).

Using the quantiles of the F-distribution, the Clopper-Pearson CI is denoted as:

$$\left[\frac{1}{1+\frac{n-Y+1}{Y}F_{\alpha/2,df1=2(n-Y+1),df2=2(Y)}};\frac{\frac{Y+1}{n-Y}F_{\alpha/2,df1=2(Y+1),df2=2(n-Y)}}{1+\frac{Y+1}{n-Y}F_{\alpha/2,df1=2(Y+1),df2=2(n-Y)}}\right]$$

(Santner and Duffy, 1989, Remund et al., 2004)

As can be seen from the formula, the lower bound is not defined for Y=0 and the upper bound is not defined for Y=n. Thus in case that Y=0, the lower bound is set 0 (as the estimator) and for Y=n, the upper bound is set 1.

Improved exact confidence intervals: Blyth-Still-Casella, Blaker

Blyth and Still (1983) and Casella (1986) describe computational methods to derive two-sided intervals which are exact but less conservative than the Clopper-Pearson Interval and are improvements of the method given by Sterne (1954). Tables of the Blyth-Still-Interval for $n \leq 30$ for confidence levels of 0.95 and 0.99 are given in Blyth and Still (1983) and Duffy and Santner (1989). Its modification by Casella is tabulated in Casella (1986). They have in common not to require equal probabilities $\alpha/2$ to exclude the true parameter for the upper and the lower bound. Thus, these methods cannot be applied as one-sided intervals in the usual way of replacing $\alpha/2$ by α (Reiczigel, 2003).
According to Reiczigel (2003), the exact confidence Interval proposed by Blaker (2000) is nearly equivalent (fulfilling slightly different conditions) to that of Sterne, but is easier to compute.

3.2 Confidence intervals for binomial group testing

In group testing experiments, confidence intervals for π can be constructed in two basic ways: on the individual scale or on the group scale.

1. On the individual scale, the estimator p and its variance are used for construction of the confidence limits. The structure of the variance of p (including Y,n,s) makes it complicated to implement adjustments for improved confidence intervals straightforward into these methods.

2. The construction of intervals on the group scale uses the fact that θ is a binomial proportion too. A confidence interval $[t_L,t_U]$ for θ is first constructed using the estimator t=Y/n. This interval is transformed in a second step to a confidence interval $[p_L,p_U]$ for the individual probability by applying $p = 1-(1-t)^{1/s}$ on the confidence limits of $[t_L,t_U]$. This is possible since the relation between p and t is monotone for all sets of p

and *s* (Tebbs & Bilder, 2004). Because *t* is the estimator of a simple binomial proportion with the simple variance term $\mathrm{var}(t) = t(1-t)/n$, the usual methods for construction of confidence limits for the binomial proportion can be applied this way.

3.2.1 Asymptotic confidence interval constructed on the individual scale

The individual scale Wald CI

A Wald-type confidence interval can be constructed using the estimator *p* and its variance multiplied with the appropriate quantile of the standard normal distribution.

The two-sided Wald-Interval is

$$\left[p \pm z_{\alpha/2} \sqrt{\hat{V}ar(p)}; \right]$$

where $\hat{V}ar(p) = \left[1-(1-p)^s\right] / \left[ns^2(1-p)^{s-2}\right]$ is the asymptotic variance of *p*, recommended by Thompson (1962) and Swallow (1987) for appropriate sets of *n, s* and *p*. This asymptotic variance estimator is problematic for small n and large values of *s* and π, as shown by Thompson (1962). Alternatively, the Wald-Interval can be calculated using the formula for closed calculation of the variance given in section 2.5.

Tebbs and Bilder (2004) examined further confidence intervals constructed on the individual scale, which did not show advantages compared to some of the intervals constructed on the group scale.

3.2.2 Confidence intervals constructed on the group scale

The confidence intervals recommended for simple binomial testing (section 3.1) can be applied here to estimate an interval [t_L, t_U] for the group scale parameter θ. The methodology of transferring confidence intervals from the group scale [t_L, t_U] to the individual scale [p_L, p_U] assumes that the relation between *t* and *p* is monotone for fixed values of *s* (Tebbs and Bilder 2004, Tebbs and Swallow 2003b). Then it is provided, that a positive difference $t_2 - t_1$ will result in a positive difference $p_2 - p_1$, and that a confidence interval for θ which excludes a

certain θ_0 will result in a corresponding interval for π which as well excludes the corresponding π_0.

Figure 1 illustrates the relationship $p = f(t,s) = 1-(1-t)^{1/s}$ for different group sizes *s*.

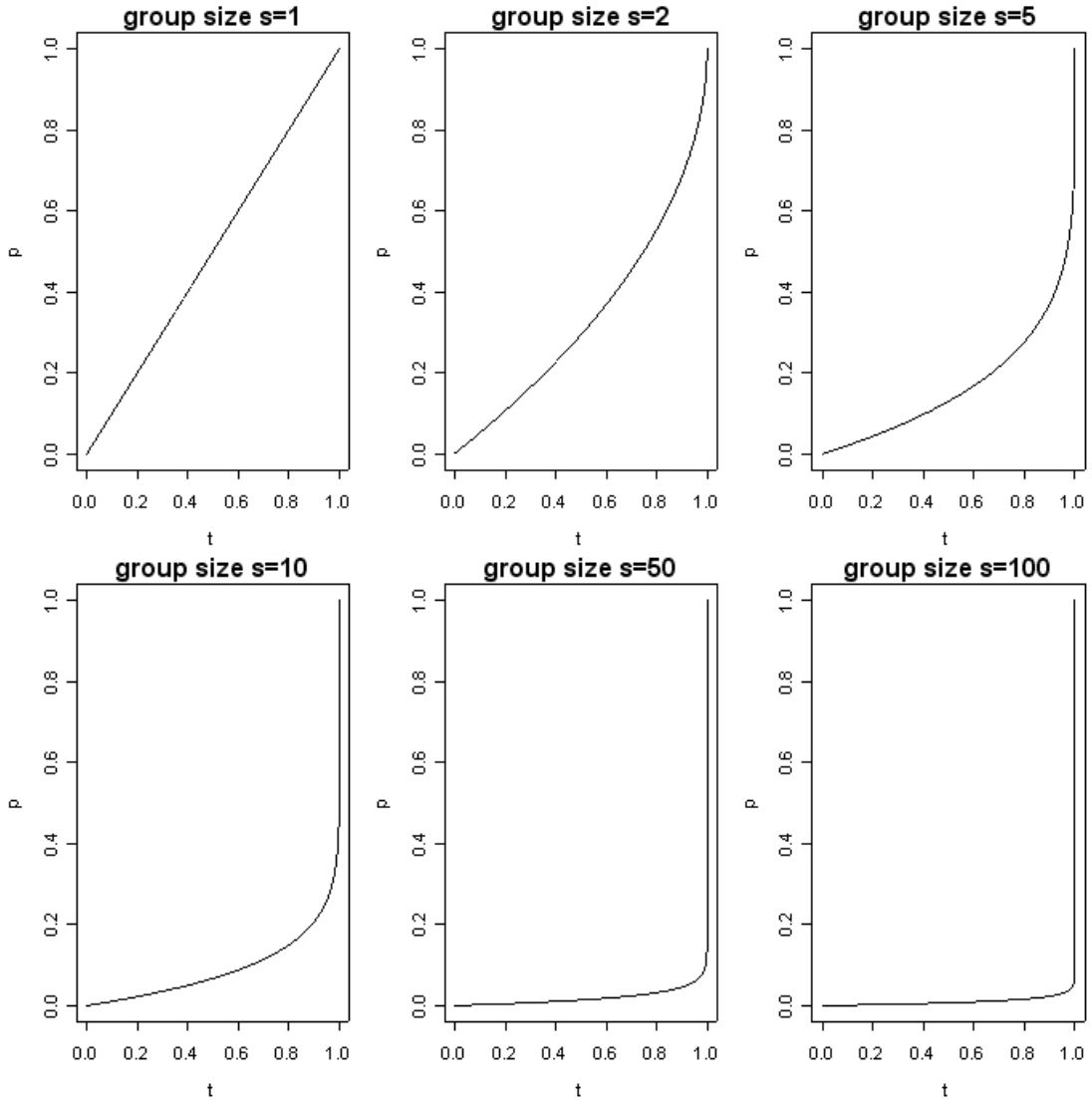

Figure 1: relation between p and t for different group size s= 1,2,5,10,50, 100

As expected there is a linear relation between *t* and *p* for group size *s*=1. For group sizes *s*>1 the relation is not linear but still monotone. Thus t ranges from 0 to 1 and can have only $i = n + 1$ possible values, where $(t_i - t_{i-1})$ always equals $1/n$, the possible estimators *p* of a group testing experiment also can have only $i=n+1$ values. For large group sizes, the differences $(p_i - p_{i-1})$ become much smaller than $(t_i - t_{i-1})$, so rounding of outcomes should be avoided for large *n* and *s*.

That the assumption of monotony is true for all sets of *p* and *s* can be shown by application of the resampling function given in section 6. Ordering and construction of a confidence interval on the group scale always results in the same order of outcomes and thus in the same confidence interval as if this is done on the group scale.
Because of this monotony, the performance of confidence interval procedures in simple binomial testing can be expected to be transferred from the interval on the group scale to the interval on the individual scale.

3.3 Binomial confidence intervals and tests used in statistical standard software

StatXact 6

For a single binomial proportion, StatXact 6 allows computation of the Clopper-Pearson confidence interval and the improved method of Blyth and Still (1983) and Casella (1986), as well as one- and two-sided p-values of the exact test.

R 2.0.1

In R the standard method in the package ‘stats’, implemented in the function binom.test, is the Clopper-Pearson CI and p-values of the exact test. The CRAN package ‘epitools’ additionally provides the functions binom.wilson and binom.approx for calculation of the Wilson and Wald CI, respectively.

SAS 8.2

In SAS, methods for a single binomial proportion are implemented in the procedure FREQ. Here, either the Clopper-Pearson or the Wald CI can be calculated while p-values are available for the Score test (corresponding to the Wilson CI) and the exact test.

At the moment, methods for group testing are not implemented in these programs, neither for analysis of experiments, nor for experimental design depending on bias, MSE or power.

4 Comparison of the methods

4.1 Criteria

Because main interest is in a proof of safety, focus will be on upper confidence limits, which will be compared with respect to their coverage probabilities and their power against a null hypothesis of interest. If the intended use of the interval is mainly estimation and not decision on a hypothesis, interval widths might be compared instead of power, as done in the references (Brown et al. 2001, Tebbs and Bilder, 2004).

Since objective is a proof of safety using upper confidence limits, the actual consumers risk corresponds to *α*, i.e. methods with an actual coverage probability close to (*1-α*) are required. Since the producers risk for the proof of safety H_0: $\pi \geq \pi_0$ vs. H_1: $\pi < \pi_0$ corresponds to *β*, a high power *1-β* is required what can be achieved by application of an appropriate experimental design. If the actual coverage probability is smaller than the nominal level (*1-α*), the method is called 'liberal', if the actual coverage of a method is greater than this pre-specified confidence level, it is called 'conservative'. In this second case, the actual power (*1-β*) will be lower than necessary, whereas liberal methods exhibit a higher tendency to reject the null hypothesis, because they allow a higher type-I-error *α*.

4.2 Methods for calculation of power and coverage probability

Coverage probability denotes the actual probability of a confidence interval to contain the true parameter which it is supposed to contain with probability *1-α*, i.e. in case of using an upper limit $P\left(\pi \in \left[0, p_U\right]\right)$. Power denotes the actual probability *1-β*, i.e. the probability to reject the null hypothesis in a case where the alternative hypothesis is true. If the hypotheses H_0: $\pi \geq \pi_0$ vs. H_1: $\pi < \pi_0$ are tested using the upper limit of a confidence interval, power is the probability that the confidence interval [*0, p_U*] does not contain the threshold π_0.

Coverage probability, power and interval length can be either simulated or calculated closed. The closed calculation is preferable in terms of calculation time at least for small *n*. Here closed calculation will be used, but the underlying R code was checked by comparing the results with simulations. Due to overflow

of the value of the binomial coefficient $\binom{n}{k}$ for *n*>1020, closed calculation can not be used for very large sample sizes. In this case simulation will be applied.

Simulation

For simulations a single binomial or binomial group testing experiment was created using the binomial pseudo random numbers of the function rbinom() in the 'stats' package of R 2.0.1. The event of interest ("CI contains the true parameter", "H_0 is rejected" or "actual interval length") is evaluated and saved. Experiments with a given, common set of parameters are repeated a sufficient number of times, and the probabilities of events or the expected lengths can be calculated from the saved values.

The R code (programmed under R.2.0.1) is given in the Annex, section 11.

Closed calculation

Closed calculation uses the fact that the observation *y* can only have the realizations *Y = 0,…,n*. The probability of a certain realization of *y* can be calculated for a given set of parameters from binomial distribution. Furthermore, for each realization of *y* it can be calculated which event ("CI contains the true parameter", "H_0 is rejected", "actual interval length") comes true for the given parameters. Expected value of an event for a given CI method applied for *n* observations can then be calculated by summation over all possible realizations of *y*:

for simple binomial testing (*s=1, $\pi=\theta$*):

$$E\left(I \mid \pi, n\right) = \sum_{y=0}^{n} I(y,n) \binom{n}{y} \pi^{y} \left(1-\pi\right)^{n-y}$$

(Brown et al. 2001)

for binomial group testing:

$$E\left(I \mid \pi, n, s\right) = \sum_{y=0}^{n} I\left(y,n,s\right) \binom{n}{y} \left(1-\left(1-\pi\right)^{s}\right)^{y} \left(1-\pi\right)^{s(n-y)}$$

(Tebbs and Bilder, 2004)

where *I(y,n,s)* is an indicator function which gives a 'value' of the confidence interval constructed for a certain value *y* and given *n, s*. This value might be the length of the interval for a certain *Y* or it might be 1 or 0 if this interval contains a hypothetical value π_0 or not.

The following formulas can be used for comparing coverage probabilities (see Tebbs and Bilder, 2004) and power of the different methods for construction of confidence intervals for group testing.

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

calculates the exact coverage probability, where *I(y,π,n,s)* is the indicator function, which has the value 1 if the confidence interval contains π and has the value 0 if not (Tebbs & Bilder, 2004).

Analogously, the exact power for a group testing confidence interval is given by

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

where *I(y,π_0,n,s)* denotes an indicator function with the value 1 if a confidence interval does not contain the hypothetical value π_0 and has the value 0 if it contains the value π_0.
The code for calculation is given in the Annex, section 11.

4.3 Comparison of intervals for simple binomial testing

The CI for simple binomial testing are the basis of the group testing CI which are constructed first on the group scale. Although interest is in small proportions π, the interval procedures have to be compared over the entire range (0,1) because these methods are later on used to construct intervals for θ and the range of θ depends on the chosen group size *s*.

4.3.1 Two-sided intervals

The coverage probabilities of two-sided intervals were examined by Agresti and Coull (1998) for the Wilson Score and Clopper-Pearson interval; Brown et al. (2001a, 2002) reviewing many methods including the Wilson, the Clopper-Pearson; Jeffreys prior and the generalized Agresti-Coull interval; Casella (2001) and Blyth and Still (1983) for the Blyth-Still-interval; Brown et al. (2001b), Blyth and Still (1983) for the continuity corrected Wilson Score interval, Reiczigel (2003) for the Sterne, Blaker and Wilson interval.

Although recommendations are discussed controversy, the following can be resumed:

A coverage probability $\geq$ (*1-*α) for all *n* and π is only guaranteed by the exact methods as Clopper-Pearson, Sterne, Blyth-Still, Casella and Blaker. Because of discreteness, all of them tend to be conservative, with an actual mean coverage higher than the nominal. The two-sided (*1-*α)-Clopper-Pearson confidence interval guarantees to exclude the true parameter with a probability $\leq \alpha/2$ for each bound of the interval (Santner and Duffy, 1989, Blyth and Still, 1983). The actual coverage of the two-sided Clopper-Pearson CI is higher than *1-*α. Especially for small *n*, it is between (*1-*$\alpha/2$) and *1* (see Blaker, 2000).

Several Comments on the paper of Brown et al. (2001a) recommend the confidence interval of Blyth and Still (1983) which was modified by Casella (1986), because it holds the nominal confidence level and is less conservative than the Clopper-Pearson Interval (Agresti and Min, 2001). The Blyth-Still-Casella interval is also used as less conservative option in StatXact 6.

For large sample sizes also authors favoring exact solutions recommend approximative methods. F.e. for π between 0.3 and 0.7, Blyth and Still (1983) recommend the Wilson CI for n > 90, 140, 550 and 2200, if exceeding of the nominal α=0.05 by 25, 20, 10 and 5 % is acceptable, respectively.

If a mean coverage close to and slightly higher than ($1-\alpha$) and a low probability of heavily violating ($1-\alpha$) is the criterion for a recommended interval, the Wilson Score and generalized Agresti-Coull are recommended (Brown et al. 2001, Agresti and Coull, 1998). Here, the Agresti-Coull CI has the advantage to avoid severe violation of the nominal confidence level for π close to 0 or 1. Reiczigel (2003) proposes a computational intensive method to reduce the nominal level of exact CI-methods until the actual mean coverage is close to but higher than the required ($1-\alpha$) and shows that then exact Sterne / Blaker intervals are better than Wilson Score and Agresti-Coull CI.

Since main interest is in one-sided CI, the improved exact confidence intervals of Blyth, Still, Casella and Blaker will not be shown because they are inherently two-sided procedures (Reiczigel, 2003). Their performance is discussed and illustrated in the references. The continuity corrected Wilson CI will be omitted because of its conservatism (Brown et al. 2001b). Thus, from the methods discussed in section 3, only the Wald-, Wilson Score, Agresti-Coull and Clopper-Pearson interval will be compared according to their one-sided performance.

For the asymptotic methods, the question remains, for which n and π the central limit theorem suffices. The calculations of Blyth and Still (1983) for the Wilson CI reveal that even for intermediate π sample sizes of more than 1000 will be needed for an actual coverage probability close to the nominal level. For small π, the required sample size might be much higher, because the binomial distribution becomes asymmetric for extreme π, and a normal approximation becomes more unlikely to be sufficient.

The following graphs show simulated (20000 times for each point) coverage probabilities of two-sided 95% Wald, Wilson, Agresti-Coull and Clopper-Pearson CI for n=10000 and n=5000 and values of π between 0 and 0.5. All

methods differ only slightly, the approximate methods and the exact Clopper-Pearson both approach the nominal confidence level.

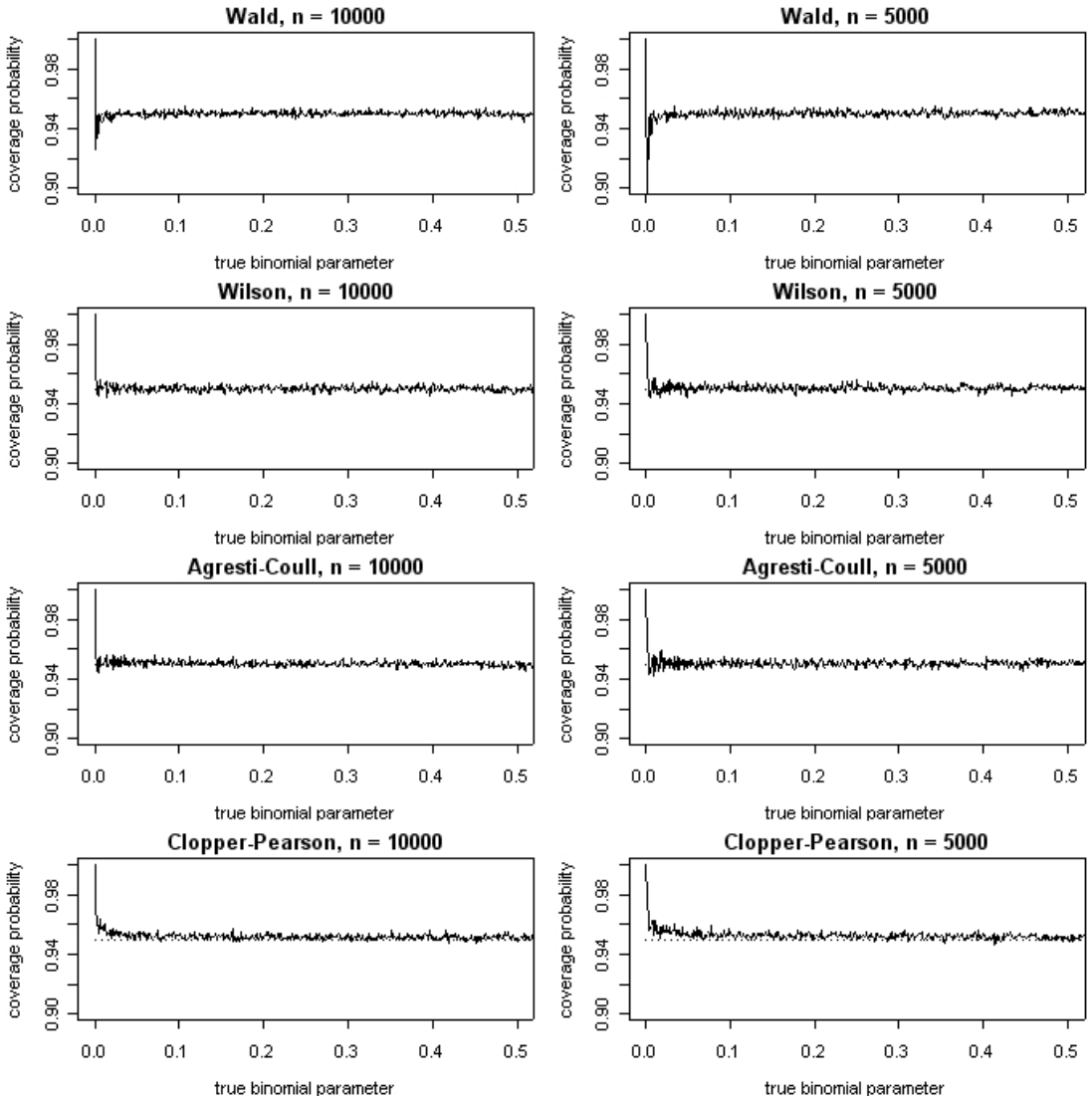

Figure 2: Coverage probabilities of two-sided 95% Wald, Wilson, Agresti-Coull and Clopper Pearson CI, n=10000 and 5000, π=0 ,…,0.5

Figure 3 shows calculated coverage probabilities for two-sided 95% Wald, Wilson, Agresti-Coull and Clopper-Pearson CI for *n*=1000 and n=500 and values of π between 0 and 0.5. Due to increasing effect of discreteness, the coverage probabilities increasingly oscillate about the nominal level, where Wald especially for small values of π has lower coverage than the nominal level, while Wilson is in average close to the nominal level and Agresti-Coull and Clopper-Pearson become conservative. The performance is symmetric around π=0.5, so the same happens for π close to 1. For intermediate values of

π, the three approximate methods are close to the nominal level, whereas the Clopper-Pearson CI becomes slightly conservative even for these large *n*.

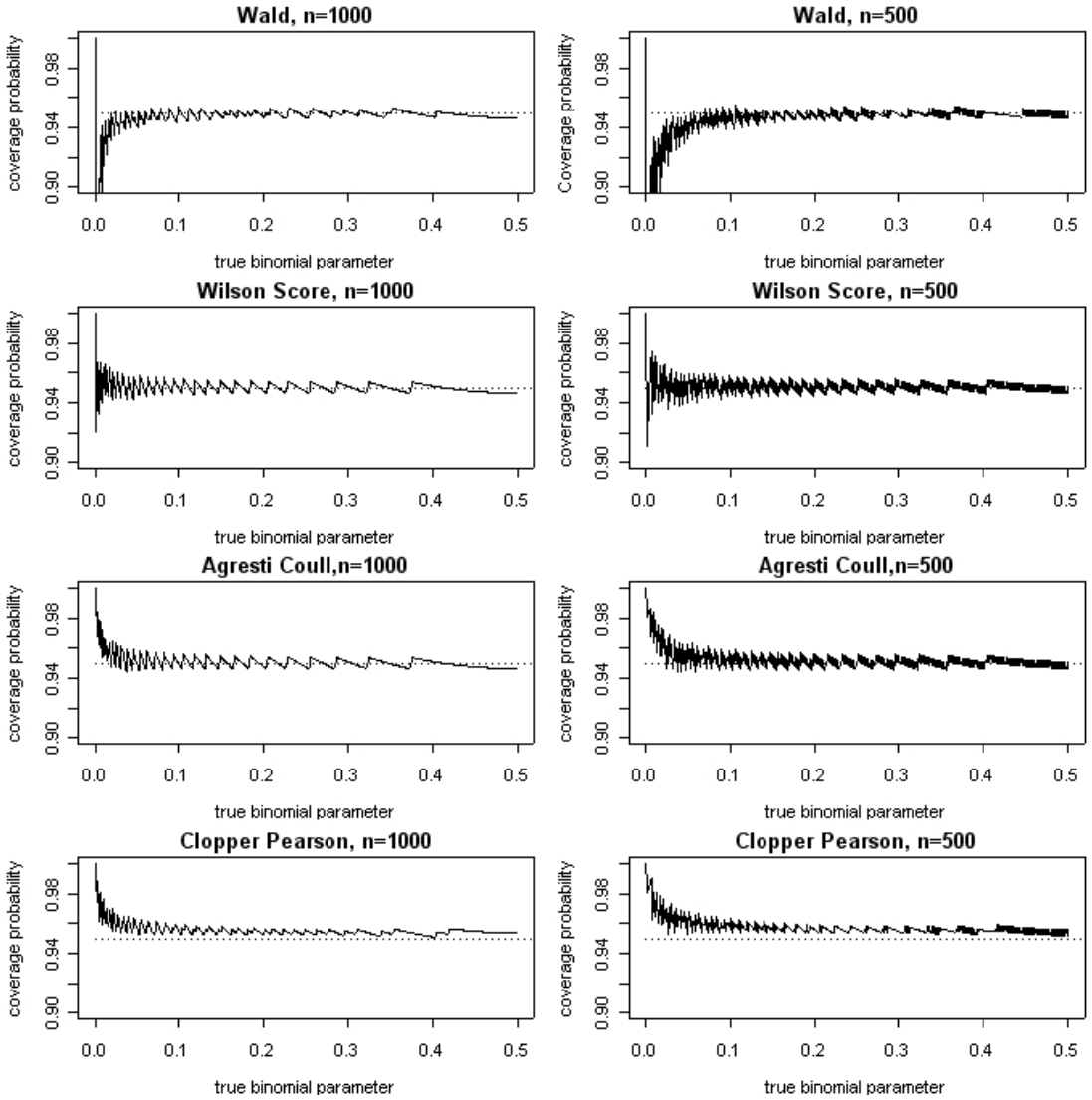

Figure 3: Coverage probabilities of two-sided 95% Wald, Wilson, Agresti-Coull and Clopper Pearson CI, n=1000 and 500, π=0 ,…,0.5

However, there is a need for smaller sample sizes than 500, if assay methods are expensive and laborious or sample size is simply limited by space capacities or the number of isolation cages in vector transfer designs. These are situations where group testing might be applied.

Figure 4 shows coverage probabilities of two-sided nominal 95%-CI for *n*=50 and 20 and π=0, …,0.5

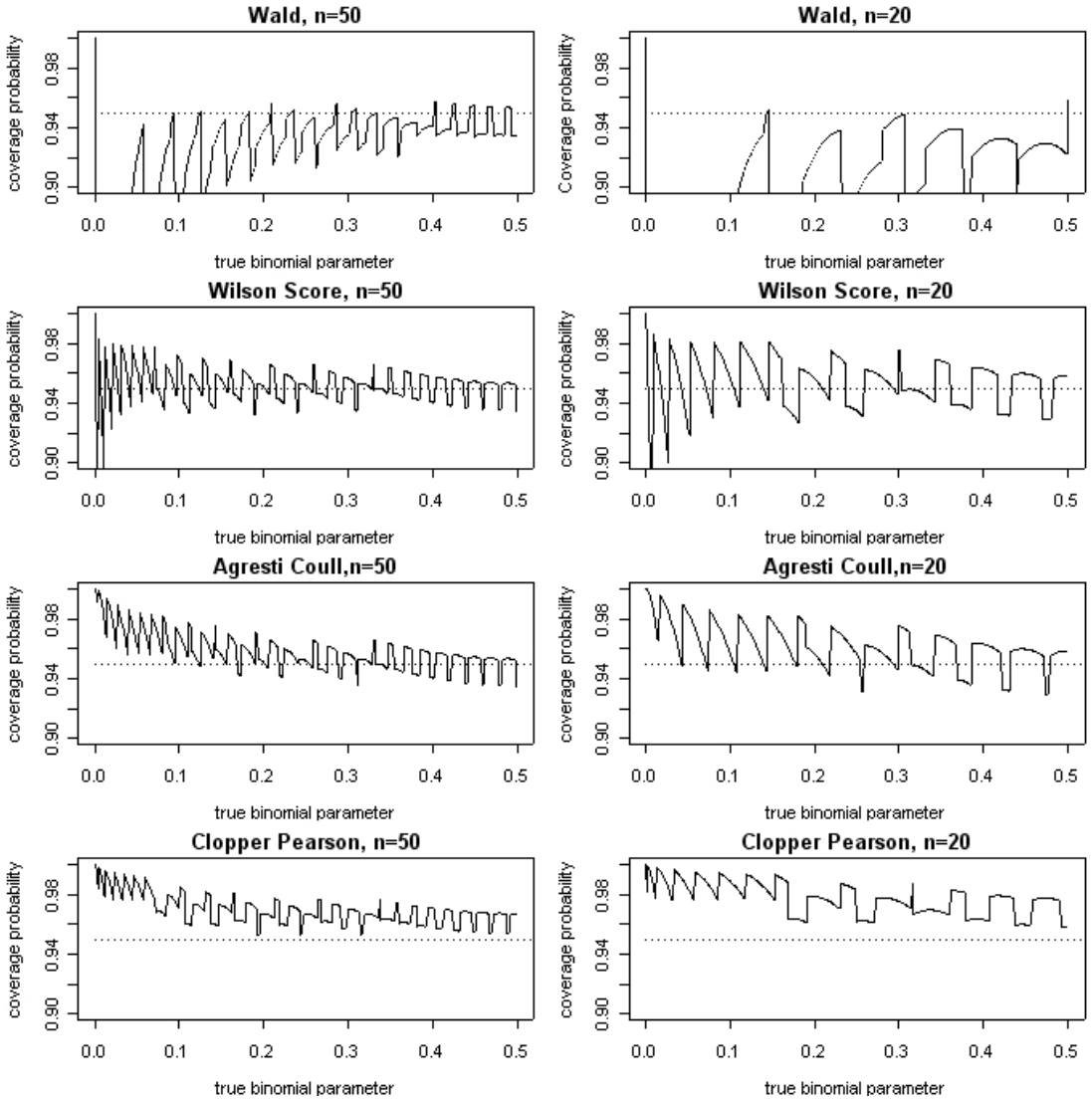

Figure 4: Coverage probabilities of two-sided 95% Wald, Wilson, Agresti-Coull and Clopper Pearson CI, n=50 and 20, π=0 ,…,0.5

The Wald interval is much too liberal, while the Wilson Score Interval uses a much more appropriate normal approximation but still shows downward spikes of coverage for small values of π. The Agresti-Coull interval avoids these downward spikes for extreme π but is still much closer to the nominal level than the conservative Clopper-Pearson CI. For extreme π, the Clopper-Pearson CI even shows actual coverage between (*1-α/2*) and *1*.

In practice, even smaller sample sizes might be required; then the shown properties become more extreme.

4.3.2 Upper confidence limits

None of the publications mentioned in sections 3 and 4.3.1 describes coverage probabilities of upper and lower bound separately. If group testing is applied in plant breeding and GMO-testing, objective is performing a proof of safety: to decide on H_0: $\pi \geq \pi_0$ vs. H_1: $\pi < \pi_0$. Then we are mainly interested in the properties of upper confidence limits. The coverage of lower limits will not be shown but their performance is symmetric on π=0.5 for Wald, Wilson, Agresti-Coull and Clopper-Pearson, i.e. they exhibit the same problems for π close to 1 as upper limits do for π close to 0 and the other way round.

Again, the methods all perform well for large sample sizes. In figure 5 their simulated coverage probabilities (20000 times for each point) will be compared only for small π=0,…0.05 and n =10000, 5000.

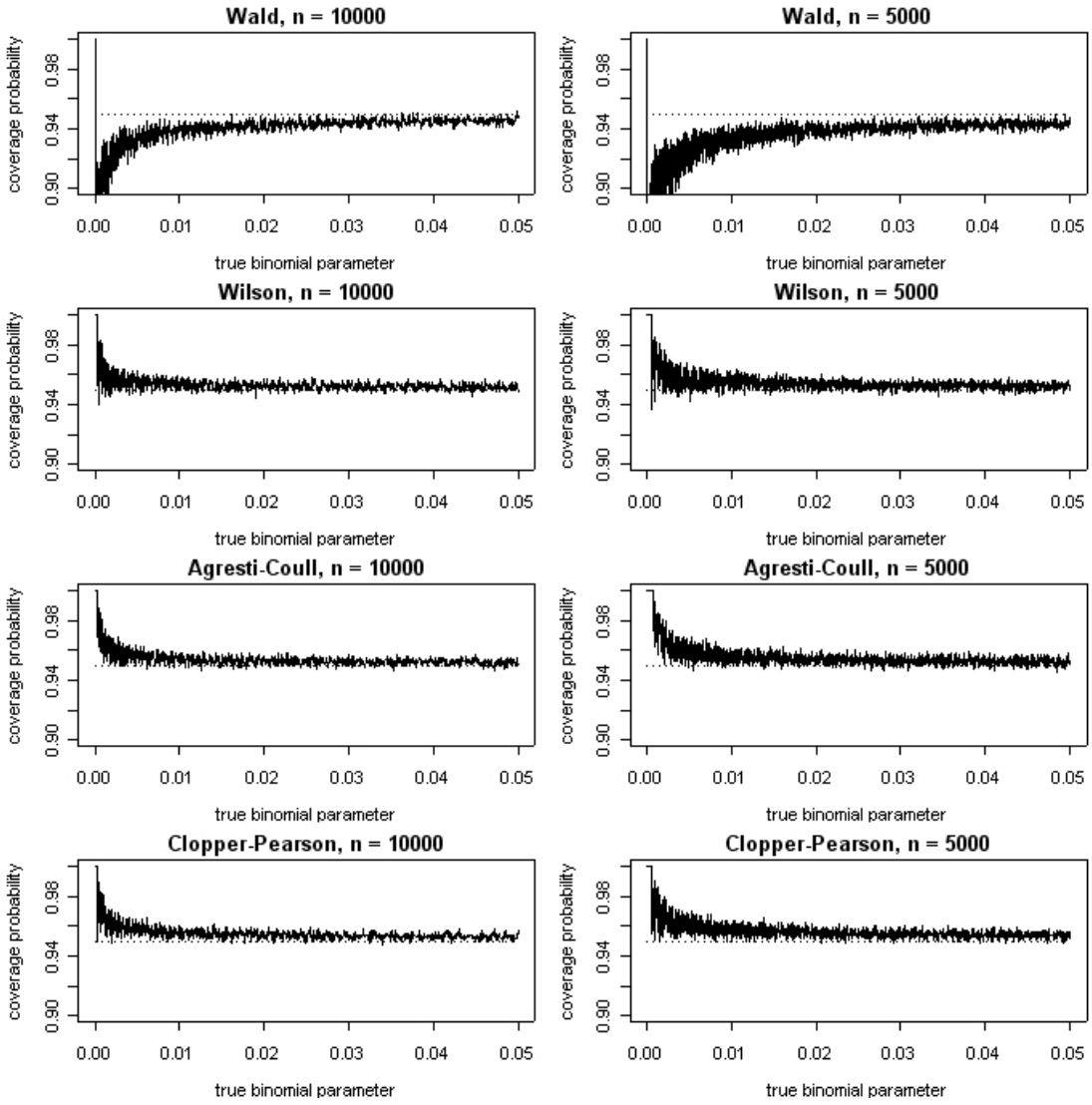

Figure 5: Coverage probabilities of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits, n=10000 and 5000, π=0 ,…,0.05

For *n*=10000 and 5000, the Score type methods have an actual coverage very close to the nominal level and only become slightly conservative for π <0.0025 and 0.005, respectively. The Wald interval already for this large sample size slightly violates the nominal level. This performance becomes more extreme as π decreases. Also Clopper-Pearson appears slightly conservative compared to the Wilson CI.

Figure 6 gives the calculated coverage probabilities of upper nominal 95% limits for *n*=1000 and 500 and π =0,…0.05.

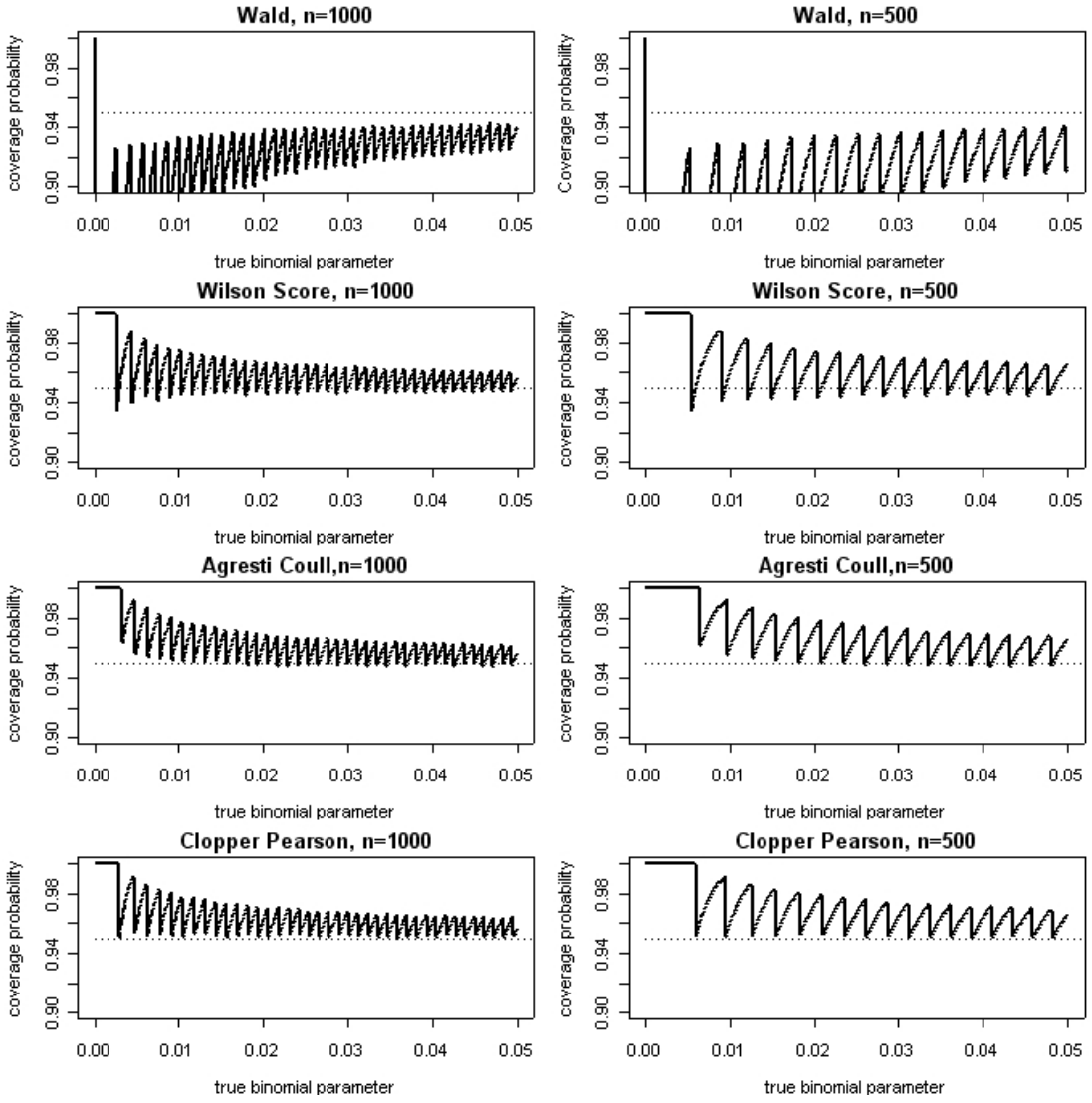

Figure 6: Coverage probabilities of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits, n=1000 and 500, π=0 ,…,0.05

Obviously, the shape of coverage probabilities is different from the two-sided CI: for small n and π close to 0, a very conservative region appears. In other words, some values of π are always included in the confidence limit, so that a null hypothesis $\pi \geq \pi_0$ with π_0 included in this region can never be rejected. This very conservative region increases with decreasing sample size. Using a 95% upper confidence limit, even for n=500 it is not possible anymore to reject H_0: $\pi \geq \pi_0$ with π_0 < 0.0054, 0.0065, 0.0060 using 95% upper limits of the Wilson, Agresti-Coull and Clopper-Pearson CI, respectively. Thus, using simple binomial testing for a proof of safety in GMO-testing, also using 500 observations will never result in a rejection of H_0: $\pi_{GMO} \geq 0.5\%$.

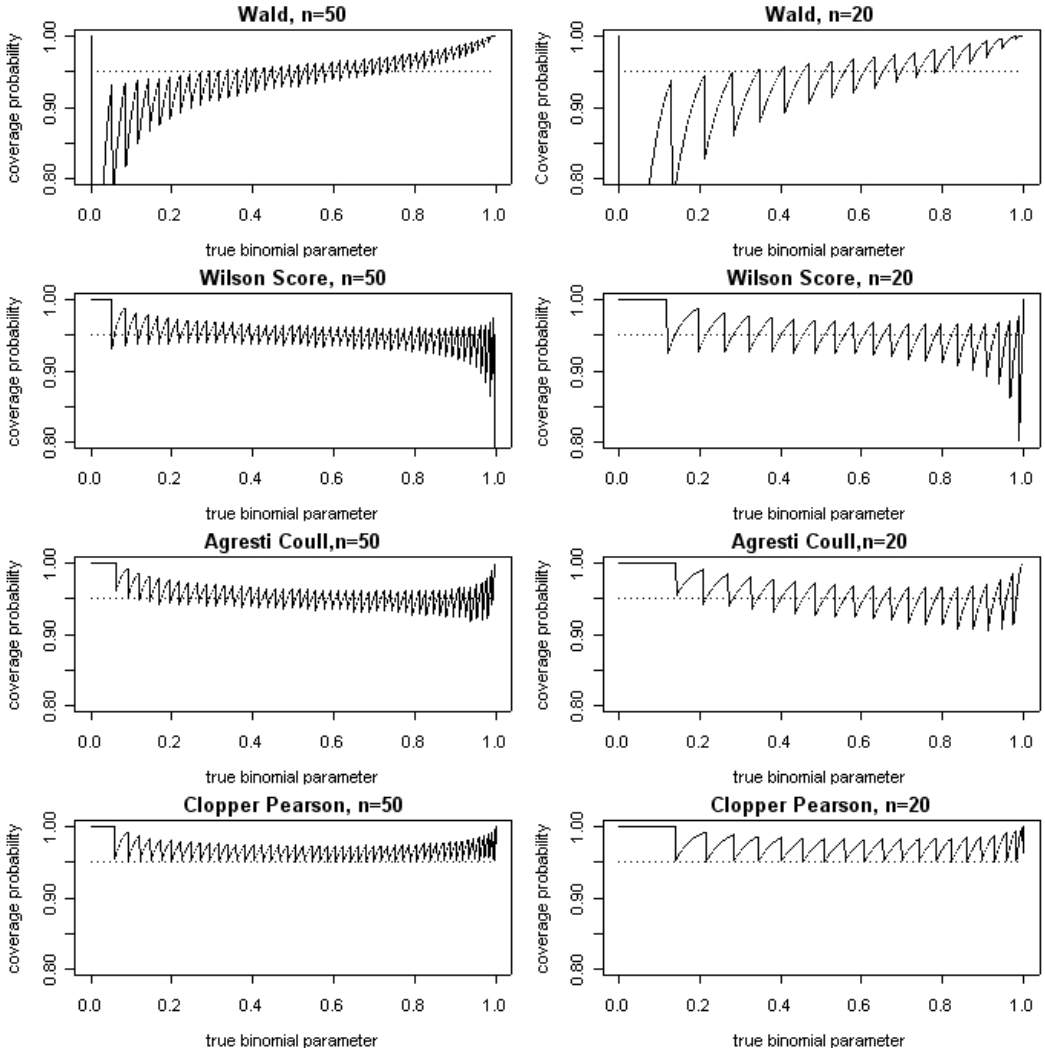

Figure 7: Coverage probabilities of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits, n=50 and 20, π=0 ,…,0.05

Again in practical application much smaller sample sizes might be required. For these small sample sizes as n = 50, 20, 10, 5, coverage probabilities will be calculated over the entire range of π from 0 to 1.

Figure 7 shows coverage probabilities of upper nominal 95% confidence limits for n = 50 and 20. The upper bound of the Wald CI again is much too liberal for small π, but becomes conservative for π near 1. The other asymptotic methods are very conservative for a range of π near 0, and become liberal for large π. Again, Wilson shows larger downward spikes than Agresti-Coull, but has a slightly shorter conservative area for π near 0. Clopper-Pearson also shows the very conservative range for small π. For other values of π, Clopper-Pearson is less conservative than in the two-sided case, because the one-sided Clopper-Pearson corresponds to the inversion of a one-sided niveau-α-test, while the two-sided Clopper-Pearson is derived from two one-sided tests, each with niveau $\alpha/2$ (see Agresti and Min, 2001, Blyth and Still, 1983).

If an experiment is limited to 50 assays, π < 0.051, 0.061, 0.058 will be always included in the upper 95% limit of Wilson, Agresti-Coull and Clopper-Pearson CI, respectively.

These properties become more extreme for small sample sizes, because of discreteness as shown in Figure 8 for n=10 and 5:

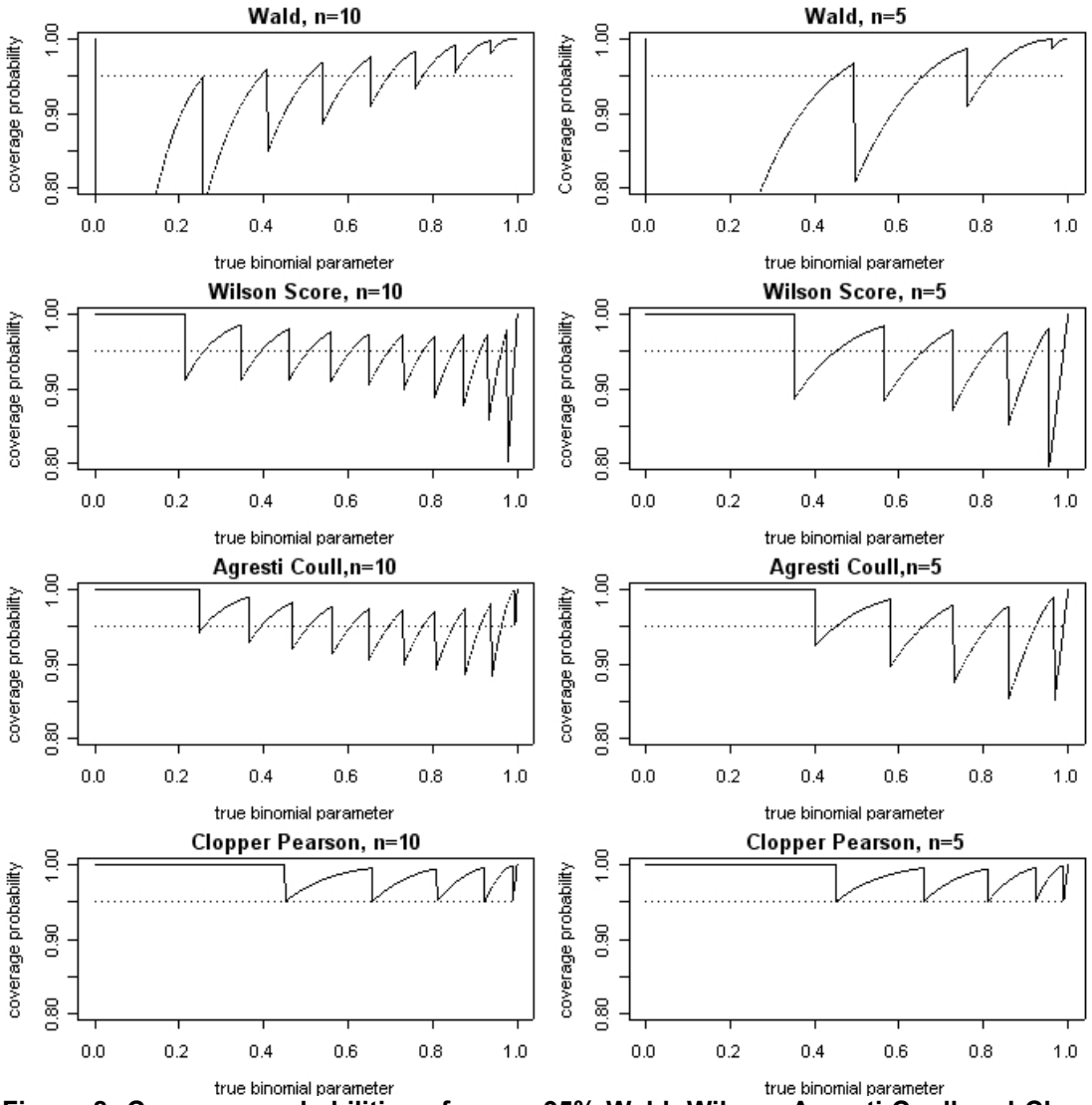

Figure 8: Coverage probabilities of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits, n=10 and 5, π=0 ,…,0.05

For this sample size *n*, the normal (large sample) approximations are clearly inappropriate. The Wald and to smaller extend the Wilson Score and Agresti-Coull method violate the nominal confidence level to a not acceptable extend. Thus, for these small sample sizes, only the Clopper-Pearson CI can be recommended. For *n*=5, this method shows a very conservative performance if $\pi < 0.45$.

Power to reject a null hypothesis

The upper bounds of Clopper-Pearson, Agresti-Coull and Wilson Score show a very conservative performance for small π, where the probability to exclude the true parameter is 0. These regions are necessary and natural, because

obviously a CI has to be broader than 0 even if no positive individual was observed (*Y=0*), to take uncertainty of sampling into account. Upper confidence limits increase for increasing *Y*, and all intervals for *Y>0* will also include those π included in the limit for *Y=0*. Thus, it is impossible that an upper confidence limit (*n, Y*) has any power to exclude $\pi_0 \in CI\ (n, Y=0)$.

The first plot in figure 9 shows the coverage probabilities of 95%-Wilson, Agresti-Coull (AC) and Clopper-Pearson (CP) CI for *n* =20 for values of π = 0,…,0.2. For this *n*=20, the upper 95% confidence limit (*Y=0*) is for Wilson [0, 0.119], for Clopper-Pearson [0, 0.139] and for Agresti-Coull [0, 0.141].

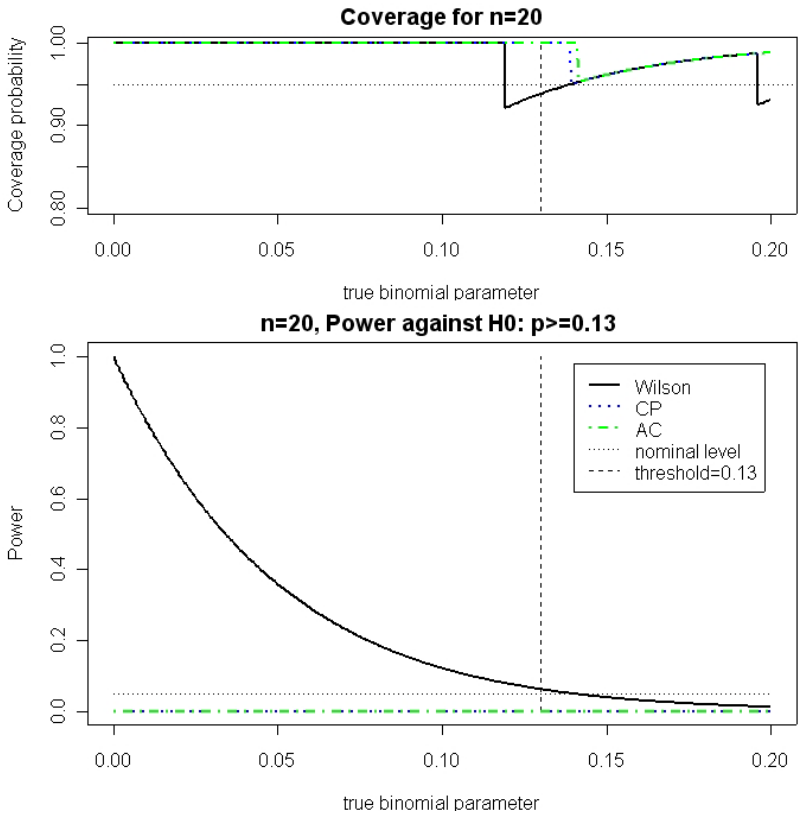

Figure 9: Coverage probability and power of upper 95% Wilson, Agresti-Coull and Clopper Pearson limits, n=20, π=0 ,…,0.2

If one wants to test H_0: $\pi \geq \pi_0$, where π_0 is included in the upper limit for *n*=20, *Y*=0 of a given method, the power to reject the null hypothesis is 0 for any π. The second plot in figure 9 shows that for π_0=0.13 this is the case for Clopper-Pearson and Agresti-Coull, but not for the Wilson-CI, which allows a proof of safety with a sufficient power at least for small π.

As shown in the figures above, the range of π always included in the upper confidence limit strongly decreases with increasing *n*. Figure 10 shows for

Wilson, Agresti-Coull and Clopper-Pearson the smallest thresholds π_0, for which H_0: $\pi \geq \pi_0$ can be rejected using an upper 95 % confidence limit. Generally, Wilson has a slightly shorter conservative region than Clopper-Pearson and Agresti-Coull is similar to Clopper-Pearson or more conservative. With a sample size n=50, null hypotheses H_0: $\pi \geq \pi_0$ with $\pi_0 < 0.06$ for Clopper-Pearson or $\pi_0 < 0.052$ for Wilson will never be rejected. Or, if a H_0: $\pi \geq 0.01$ shall be rejected, one will need at least about n=270 if Wilson is used, n=300 for Clopper-Pearson and even n=340 for Agresti-Coull.

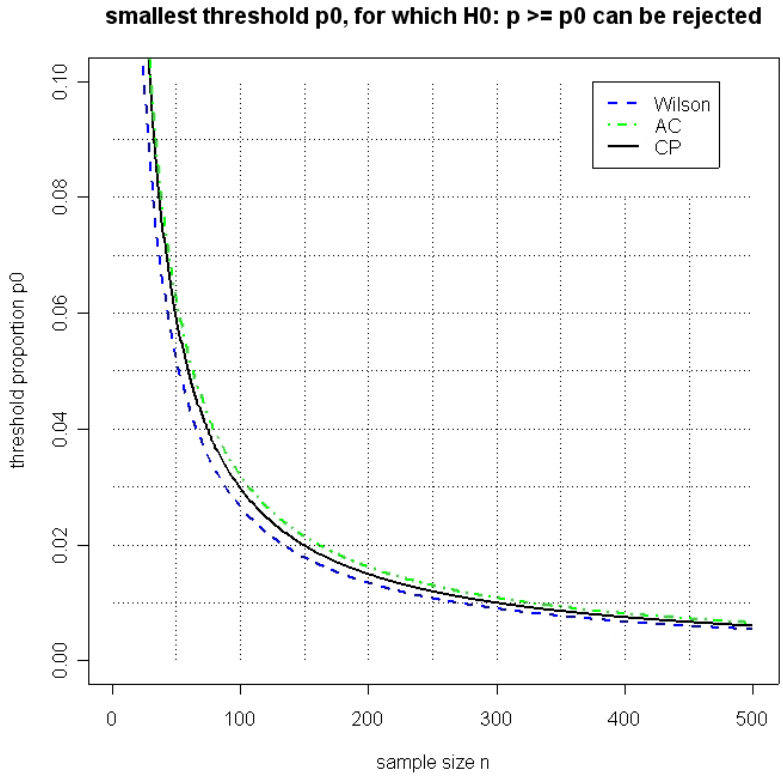

Figure 10: the smallest threshold π_0 for which H_0: $\pi \geq \pi_0$ can be rejected using upper 95% Wilson, Agresti-Coull and Clopper Pearson limits, n=2,…,500

In section 4.3 it was shown, that for performing a proof of safety if small threshold proportions regarded as unsafe, simple binomial procedures require a large number of observations. If expensive assays are necessary for detection of the unsafe trait, these large sample sizes can hardly be provided.

In the sections 4.4 and 4.5 it will be shown, that the performance of CI methods for small numbers of observations and small threshold proportions can be greatly improved, if they are applied in binomial group testing experiments.

4.4 Confidence intervals for binomial group testing: Coverage probability

The consideration of coverage probability and power of CI for simple binomial testing was important for explanation of the corresponding characteristics of group testing CI, which are first constructed for the group scale parameter θ and then transformed to the individual scale π. If the CI performs poor for a given parameter θ on the group scale, the transformed individual scale CI will perform similarly, because of monotony. The advantage of group testing now is that θ depends on π and the group size *s*. Thus, in case that it makes sense to assume a certain range of π, group size *s* can be chosen so that θ has values for which the known procedures perform well.

Another factor will influence coverage probability and power of group testing CI: bias of the estimator. A large positive bias means that the expected value of the estimator *p* is much higher than the true unknown parameter π. Thus, also the CI constructed according to the estimator will tend to have the wrong position. Because of this, one-sided CI [*0, p_U*] can be expected to become very large and conservative if the estimator is positively biased. For one-sided CI [*p_L,1*] the same happens because these are very conservative for values of π close to *1* (not shown).

The methods of individual scale Wald interval, and the group scale Wilson Score, Agresti-Coull and Clopper-Pearson CI will be compared depending on three different scenarios:

1) The number of assays is limited (by costs), but the total number of units can be chosen without serious limitation. Then the group size *s* might be chosen in a way that maximizes power. In the same time it has to be controlled, that bias does not exceed a certain level. In this case of fixed *n* and varied *s*, of course the total number of observed units increases linear with *s*.

2) The group size might be restricted to fulfill the assumptions 4 and 5 of assay sensitivity for even a single individual and to avoid misclassifications in the

assay. Then the number of assays *n* might be the only factor, which can be varied to improve the power of an experimental design. As increasing *n* decreases the bias of *p*, bias does not have to be considered in case that

3) In some practical applications, the total number of units might be limited. Then a large group size *s* corresponds to a low number of assays *n*, and the other way round. In this case, objective of experimental design is to allocate the available units to *n* and *s* in a way that achieves maximal power within a restricted bias. If the total number *n*s* is fixed at one certain value, the set of possible integers (*n, s*) is very restricted, so that additional discreteness is added to the problem.

4.4.1 Restriction 1: Limited number of assays n

If the number of assays is limited by costs, one should usually perform the maximal number of assays, which still might be very small. Then increasing group size is the only way to further improve performance of a given statistical method.

Increasing group size *s* results in higher expected values of θ, thus characteristics of the CI more correspond to the characteristics for higher values of binomial parameter. In other words, the group scale interval for [t_L, t_U] is constructed according to $t=Y/n$ and is transformed to individual scale π by application of $p = 1 - (1 - t)^{1/s}$. Because this function is non-linear, the properties of the group scale interval are distorted towards smaller values of π if group size is increased (compare figure 1 in section 3.2.2). The methods become appropriate for small π, but insensitive for large π. This occurs for all CI methods in the graphs shown later. The magnitude of up- and downward spikes depends on *n*, because *n* determines the number of possible observations and thus the discreteness of the problem.

4.4.1.1 Comparison of upper confidence limits

Figures 11 and 12 show the coverage probabilities of the upper limits of the four interval methods for n=10, a nominal coverage of 95%, a range of π = 0, …, 0.05 and group sizes of s=10, 20, 50 and 100. At the same time, the total

sample size *n*s* = 100, 200, 500 and 1000, respectively. Thus, the number of individual units contributing to the 10 observations is increased.

Although the range of π is the same, increasing group size results in coverage probabilities similar to those for higher values of π for simple binomial CI: The magnitude of the conservative region near π=0 decreases.

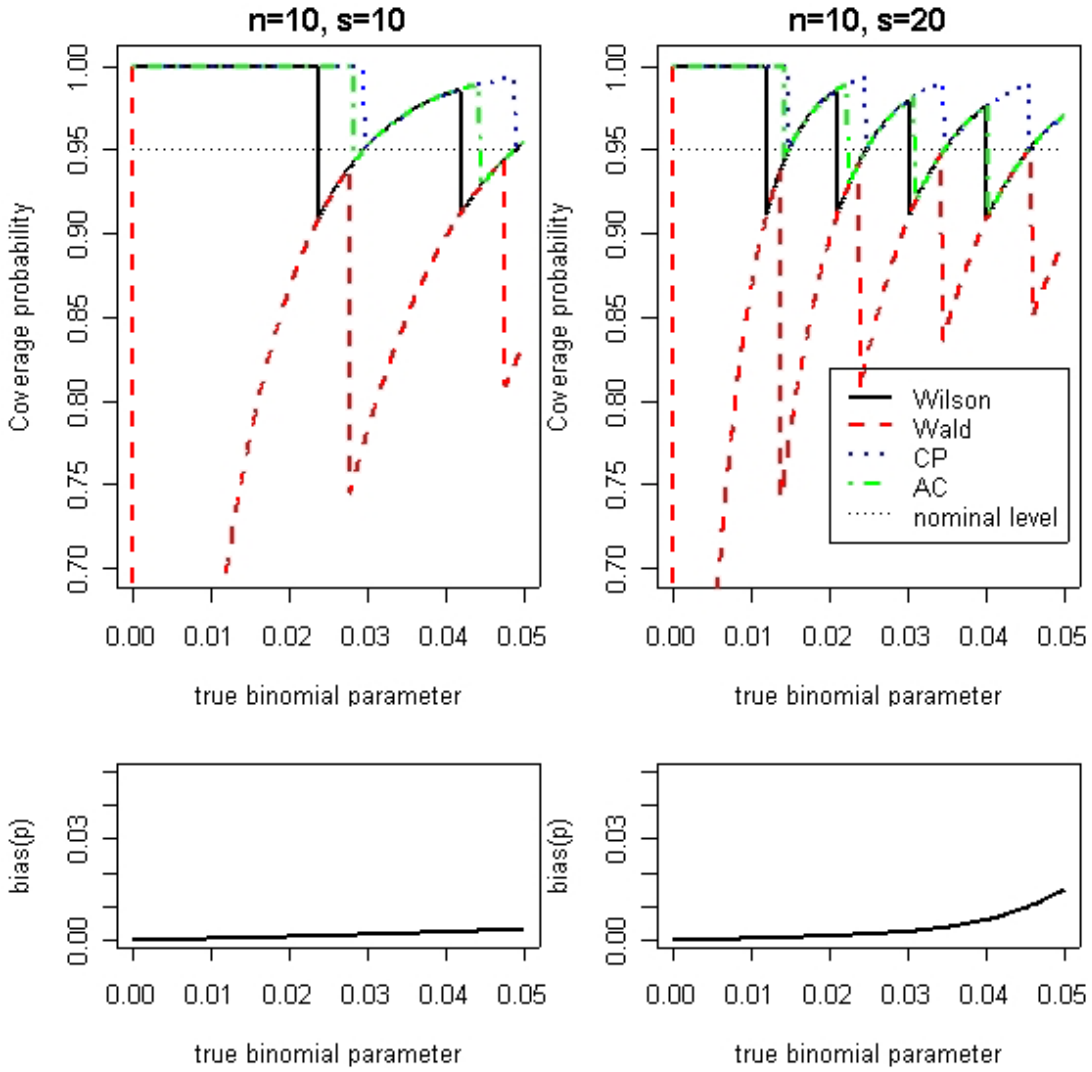

Figure 11: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits and bias of the estimator p for n=10, s=10,20, π=0,…,0.05

Comparing the four CI methods reveals that:

The upper bound of the individual scale Wald Interval is again liberal to inacceptable extend (see Tebbs and Bilder, 2004 for the two-sided case). The upper bound of Wilson Score interval has the shortest very conservative region near π=0, but also shows larger downward spikes than Agresti-Coull or Clopper-Pearson. Clopper-Pearson always holds the nominal confidence level,

also for clearly inappropriate combinations of π and *s* (π>0.015, s=100). The upper bound of the Agresti-Coull-CI has coverage probabilities close to Clopper-Pearson for small π but becomes more close to Wilson for larger π.

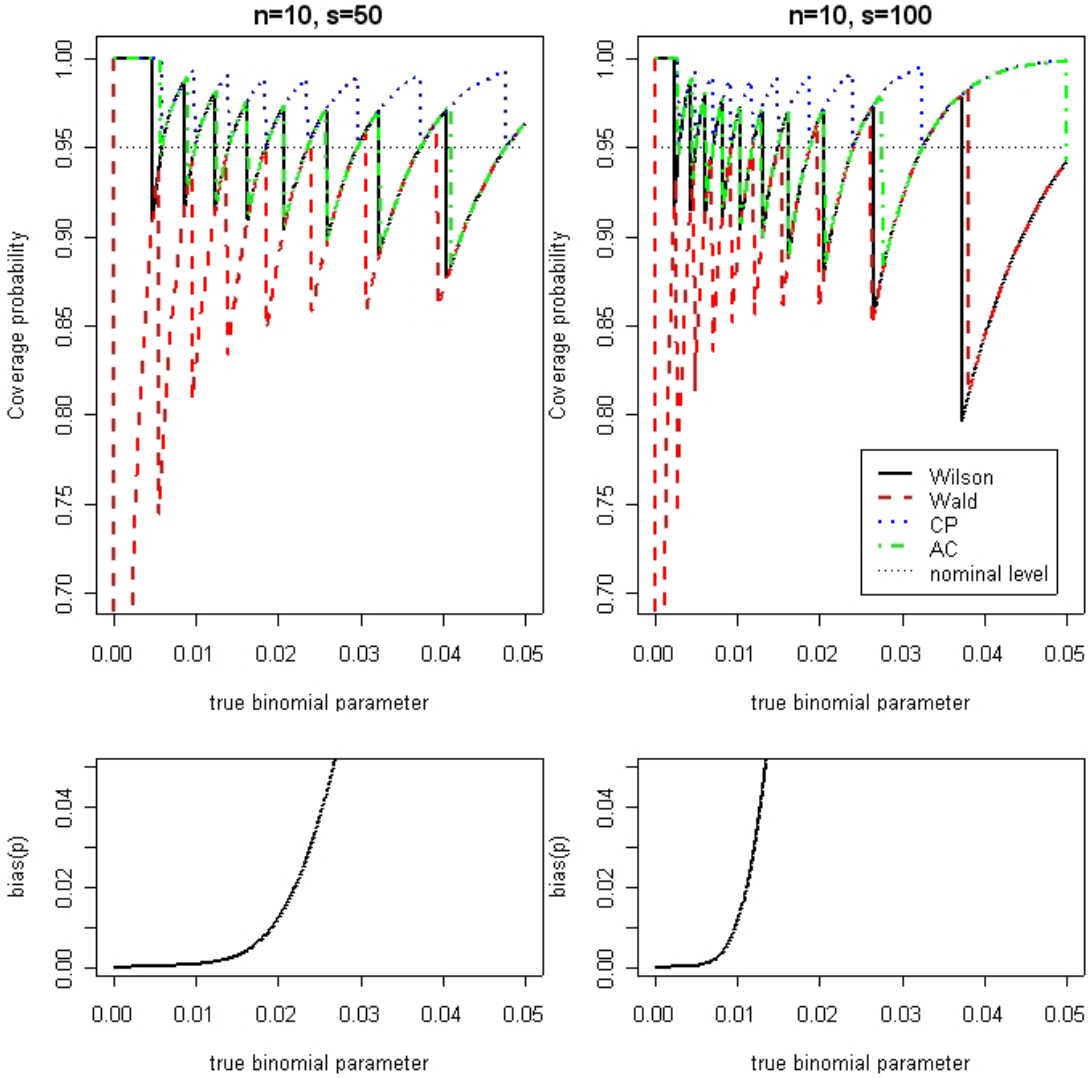

Figure 12: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits and bias of the estimator p for n=10, s=50,100, π=0,…,0.05

If the group size *s* is inappropriate for π, the intervals perform similar to the corresponding CI methods for simple binomial experiments do for binomial parameters close to 1. Thus, Wilson Score and Agresti-Coull become liberal (Compare figures 7 and 8).

If *s* becomes much too large for a given group size, the probability to yield any negative groups becomes very low. If all outcomes are positive, the estimator will always be *t=p=1*, resulting in a huge bias and very broad CI that always includes *1*. The same happens for two-sided and lower bounds coverage,

whereas for lower bounds this phenomenon is more extreme because the very conservative region for π close to *1* becomes enlarged.

Figure 13 shows coverage probabilities of upper confidence limits and bias for π = 0,…, 0.4 in a design *n*=10, *s*=50.

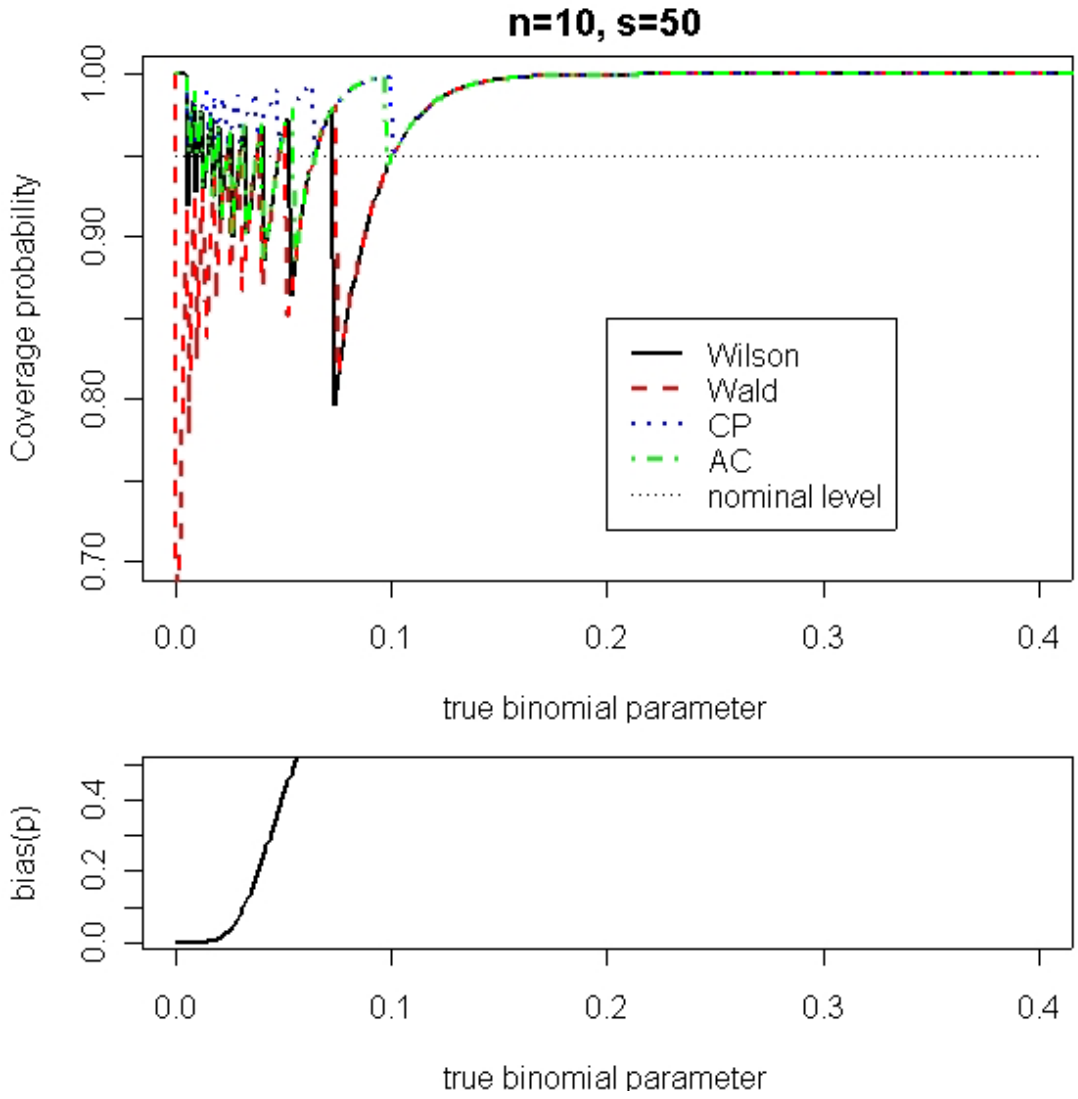

Figure 13: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits and bias of the estimator p for n=10, s=50, π=0,…,0.4: The effect of a group size inappropriate for π.

The amplitude of coverage probability does not depend on the group size *s* but on the chosen *n*. F.e. for the upper limit of the Wilson CI, the first downward spike shows minimal coverage of approximately 0.89 for *n*=5, 0.91 for *n*=10 and 0.93 for *n*=50 independent of the group size.

4.4.1.2 Two-sided confidence intervals

The performance of two-sided CI is to smaller extend shown in Tebbs and Bilder (2004). What happens to coverage probability is basically the same as for the upper limit: the 'shape' of coverage probabilities depending on θ is shifted to smaller π if group size s is increased.

Figures 14 and 15 show the coverage probabilities of the individual scale Wald interval, the Wilson, Agresti-Coull and Clopper-Pearson-method applied as two-sided 95%-CI for n=10, different group sizes and a range of π=0,…,0.05.

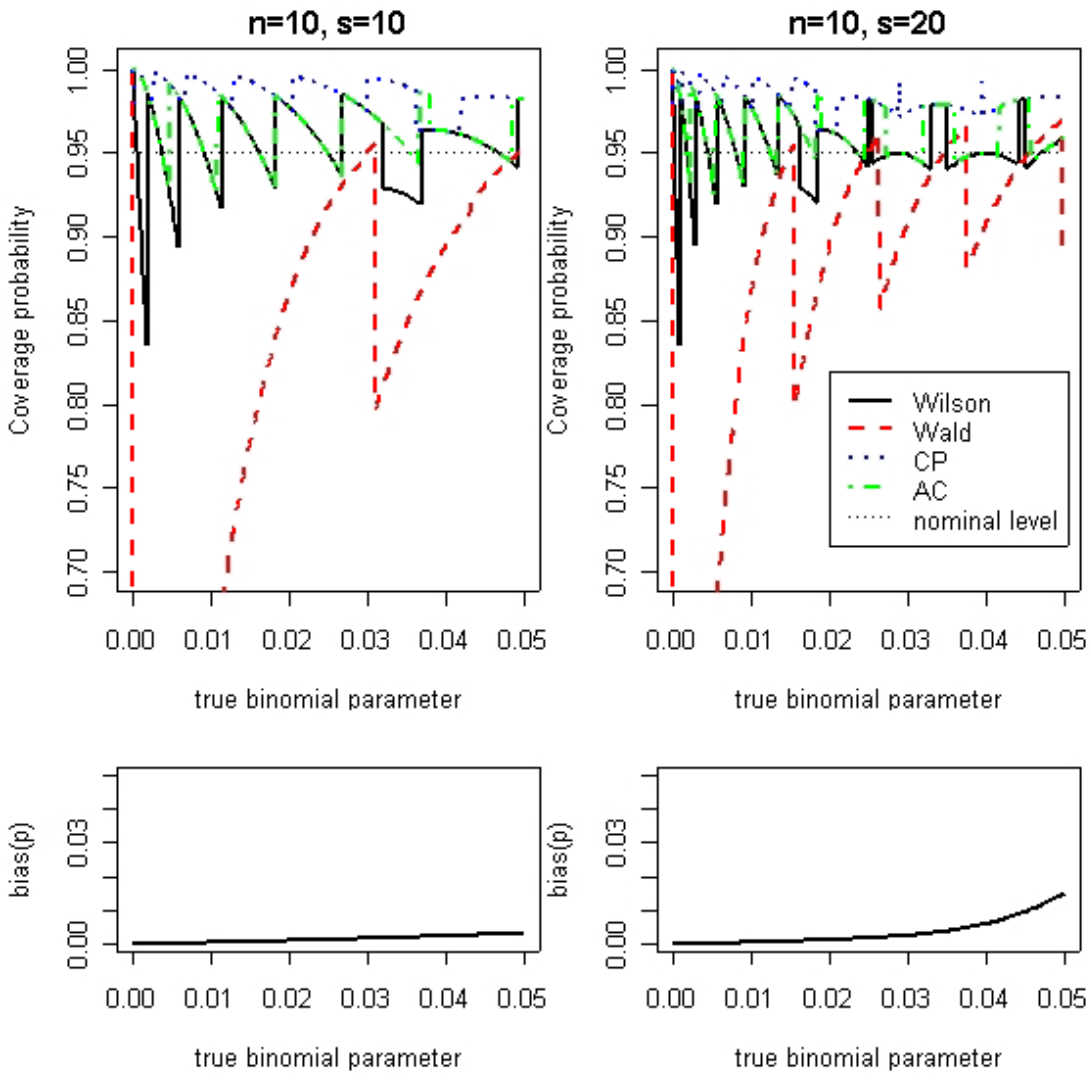

Figure 14: Coverage probability of two-sided 95% Wald, Wilson, Agresti-Coull and Clopper Pearson CI and bias of the estimator p for n=10, s=10,20, π=0,…,0.05

Comparing the different methods results in more messy graphs than for the upper limit, but still the shape of coverage probabilities corresponds to those of the simple binomial methods (compare Figure 4 in section 4.3.1).

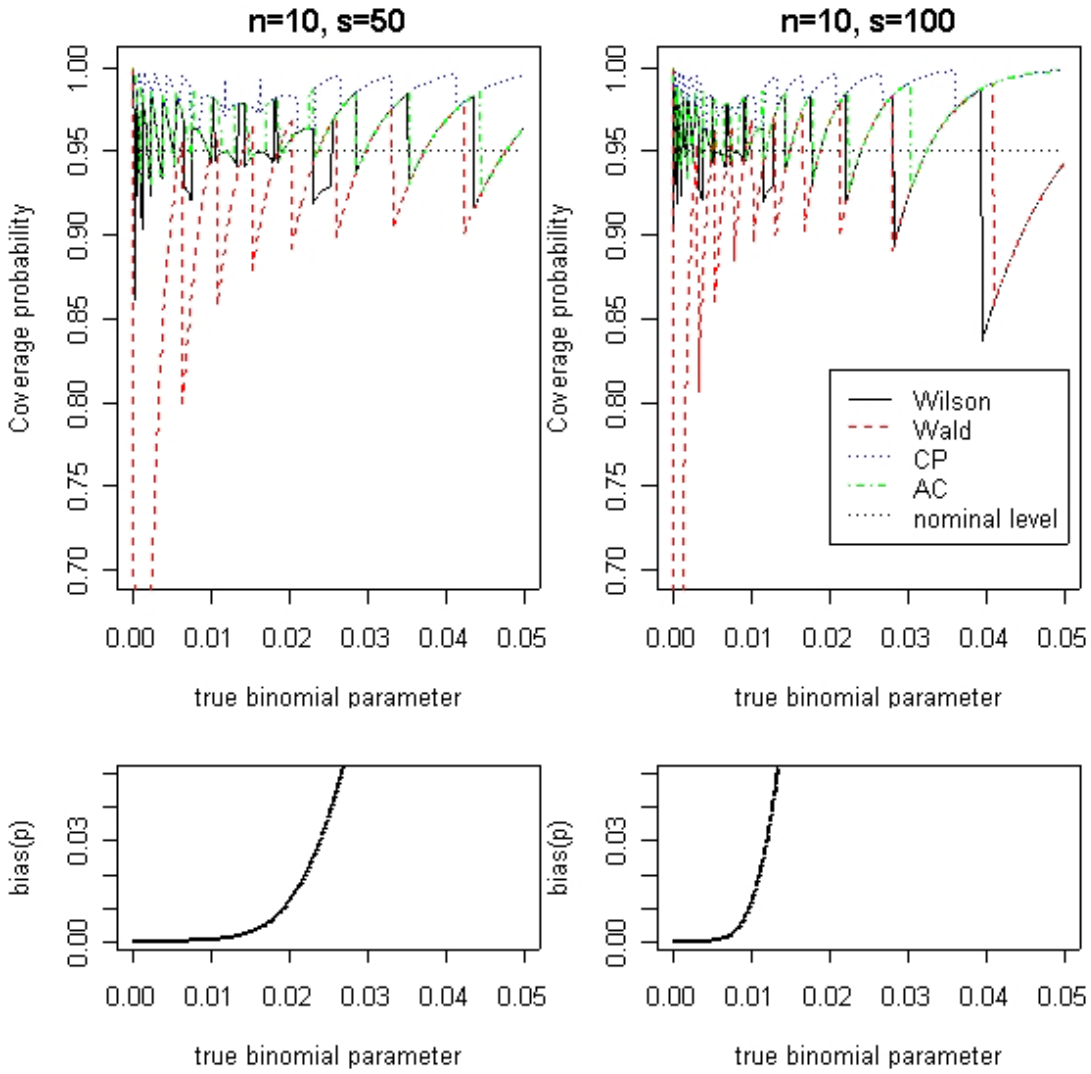

Figure 15: Coverage probability of two-sided 95% Wald, Wilson, Agresti-Coull and Clopper Pearson CI and bias of the estimator p for n=10, s=50, 100, π=0,…,0.05

As in application for simple binomial estimation, the two-sided Clopper-Pearson CI is much more conservative than in the one-sided case because of the stronger condition to invert two one-sided niveau-$\alpha/2$-tests instead of inverting on single niveau-α-test. Wilson shows the same downward spikes of coverage as in simple binomial testing if π is small, whereas Agresti-Coull tends to be conservative for small π. The Wald interval shows a not acceptable low coverage. For obviously inappropriate group sizes, indicated by a large bias, all methods fail, because they show the same performance as the corresponding

simple binomial methods do for π close to *1*: Clopper-Pearson and for larger π also Agresti-Coull again become conservative, while Wilson again exhibits the liberal spikes as it does in simple binomial testing for large π.

4.4.2 Restriction 2: Limited group size s

What happens if number of assays is increased is basically the same as increasing the sample size in simple binomial testing. For comparison with the situation of varying group size, *n* and *s* were chosen to result in the same total numbers *n*s*=100, 200, 500, 1000 as in figures 11 and 12 in section 4.4.1.1.

The following graphs show the coverage probabilities for π=0,…, 0.05 and a fixed group size *s*=10 for different sample sizes.

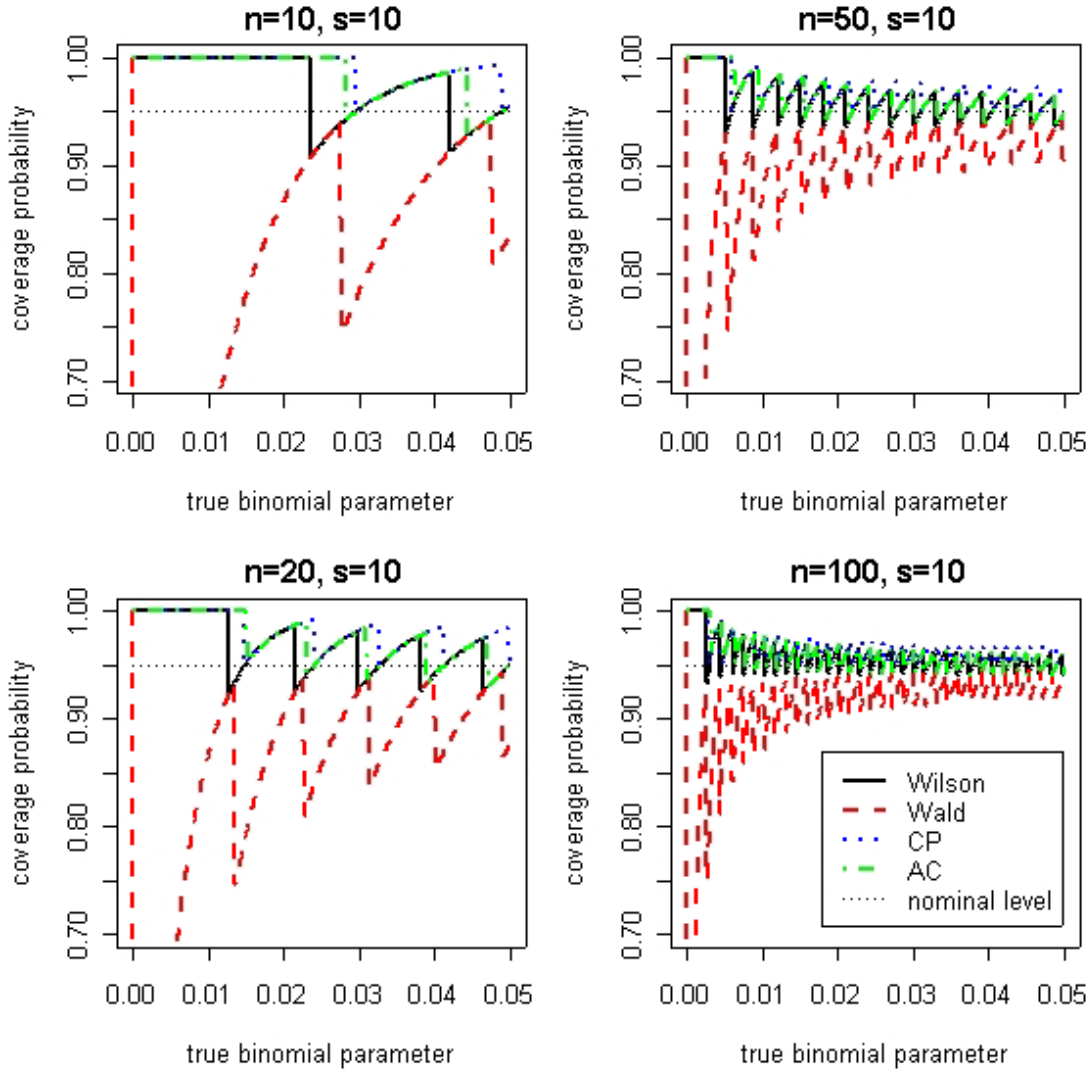

Figure 16: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for constant group size s=10, n=10,20,50,100 π=0,…,0.05

Increasing *n* while *s* is constant decreases the very conservative region near π*=0*, decreases the oscillation of coverage probability for different π and actual confidence levels become closer to the nominal level.

*4.4.3 Restriction 3: Limited total number of units n*s*

As shown in the previous sections, increasing *n* and *s* have similar effects on the performance of the methods, as far as the group size is appropriate for π. This leads to the question, whether it makes sense to look for an optimal allocation of a limited total number of units to *n* and *s*.

In figure 17 the first plot shows coverage probability for a fixed *n*=20, while increasing *s*= 5,…, 100, in the second plot group size is kept at *s*=20 and the number of *n* is increased from 5 to 100, resulting in equal total sample sizes *n***s*=100,…, 2000, for constant π =0.0025.

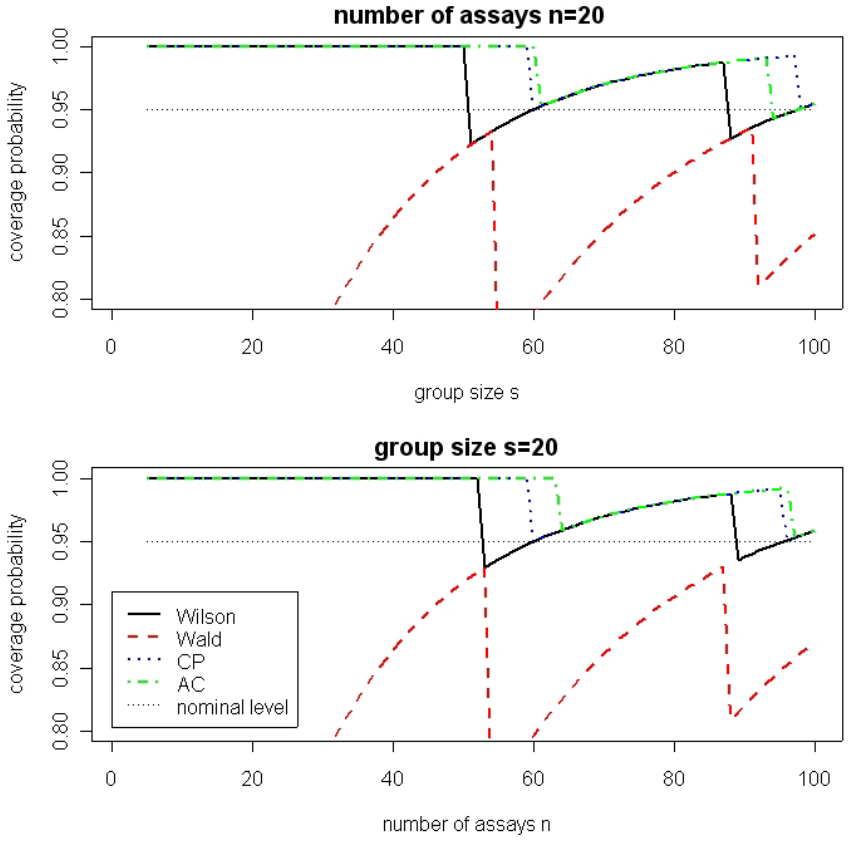

Figure 17: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits π=0,…,0.05 for 1) constant group size s=20, n=5,…100, 2) constant number of assays n=20, s=5,…100

The resulting coverage probabilities are very similar. From this, it seems reasonable that there is no clear optimal combination (*n,s*) if the total number of units *n*s* is fixed.

The first row of plots in figure 18 shows coverage probabilities of 95%-CI [*0, p_U*] for π=0.0025 and in dependence of sets of *n* and *s* resulting in a total sample size of 1000, 5000 and 10000. The number of assays *n* starts from 1000, 5000, 10000 in the beginning of the plots and ends up at 5 in all plots. At the same time *s* increases from 1 to 400, 1000 and 2000. Both results in an increasing bias, plotted in the second row of plots. Thus, the starting points of each plot show the performance of methods for simple binomial testing. The restriction of a fixed total sample size adds additional discreteness. The plots are based on 13, 17 and 22 (*n,s*)-combinations for *n*s*=1000, 5000 and 10000, respectively.

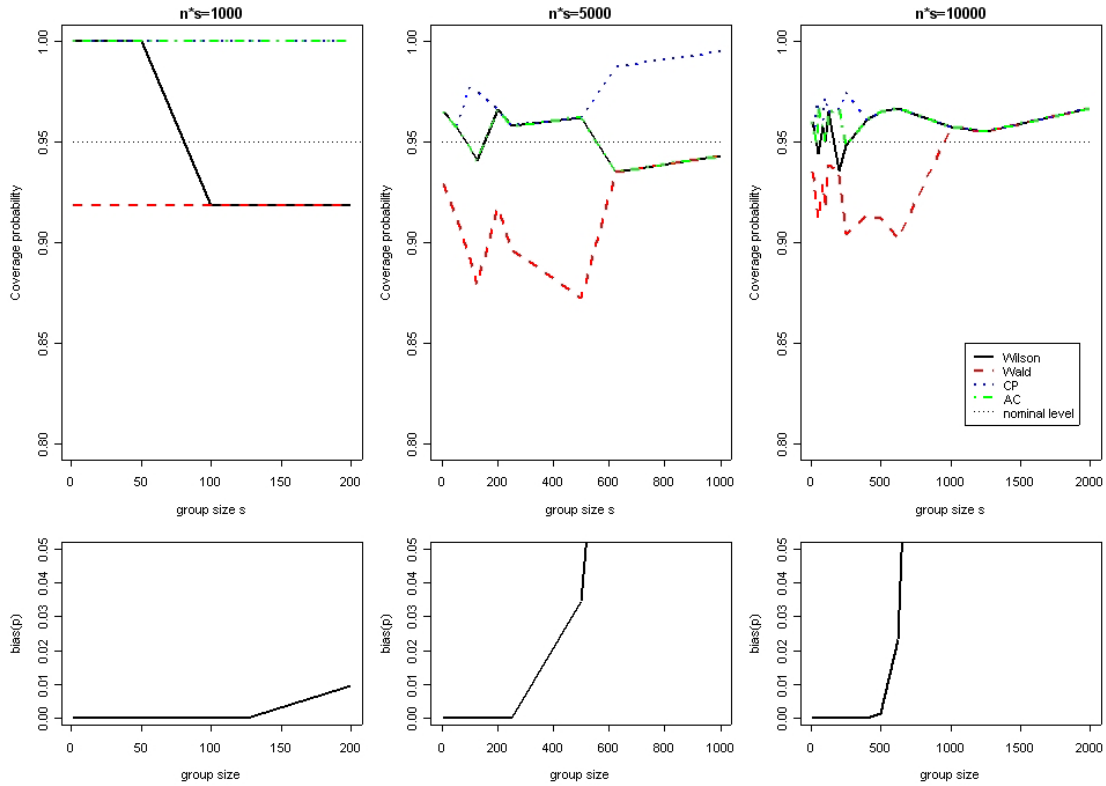

Figure 18: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits and bias (p) for combinations (n,s) resulting in constant total sample size n*s=1000, 5000, 10000, π=0.0025

Obviously, the coverage probability does not vary with a clear tendency between different sets of *n, s* as long as the group size s is appropriate for the given π. This is due to the fact that the effect of reducing *n* counteracts the effect of increasing *s*.

4.4.4 Conclusions

Which interval method to choose first of all depends on the level of security wanted for the decision or estimation. The Clopper-Pearson always guarantees the nominal level, whereas Wilson Score and Agresti-Coull do not. As will be shown in the section 4.5, the Wilson Score CI has in average a higher power than Clopper-Pearson, because of its lower mean coverage. Whether Wilson Score has lower or same actual coverage as Clopper-Pearson, depends on the particular combination of *n, s* and π. Correspondingly, whether Wilson Score has higher or same power as Clopper-Pearson depends on the combination of *n, s,* π_0 as will be shown in the next section.

Additionally, the decision depends on whether one- or two-sided hypothesis are tested. Unless Clopper-Pearson is much more conservative if applied two-sided, the Wilson Score and Agresti-Coull loose much of their superiority compared to Clopper-Pearson, if they are applied as upper limit: for small π, Agresti-Coull is as conservative as Clopper-Pearson, for some cases even slightly more conservative. In group testing, the differences between the methods become less important if design can be chosen appropriate: the very conservative region decreases in length with increasing *s*.

4.5 Confidence intervals for binomial group testing: Power and experimental design

4.5.1 Criteria for choice of n and s

Bias(p) and MSE(p)

One criterion for choice of optimal group size might be minimization of the point estimates bias, as discussed f.e. in Thompson (1962) and Swallow (1985). Bias is of main importance, if estimation is main objective of the study and decision on a hypothesis is less important.

Bias of p increases for increasing group sizes *s* or increasing probabilities π. If group size is chosen to large for a given π, the expected value of the estimator is much higher than the true value.

Oppositely, the variance of *p* decreases for large group sizes *s*. Swallow (1985) thus gives the mean squared error of the estimator *p*

$$MSE(p) = variance(p) + [bias(p)]^2$$

as main criterion for the goodness of the estimator *p*. Thompson (1962) and Swallow (1985) recommend to choose *s* for a given *n* and an expected π so that a minimal MSE(*p*) results. Thompson (1962) gives a simple approximation to solve this problem.

Increasing *n* decreases the MSE(*p*) and the advantage of group testing compared to conventional binomial testing (*s=1*) in terms of MSE(*p*). If the total number of individuals *n*s* is limited by costs and n is not limited, *s=1* is the optimal choice. But for small values of π there are $s > 1$ for which the MSE(*p*) is only slightly higher (Swallow,1985).

If the total number of units is fixed, then a group size of *s*=1 generally results in minimal MSE(*p*). Then cost relation of assays vs. units might be an additional criterion for optimization of the design (Thompson, 1962).

Power

A second criterion for choice of the experimental design might be the power to decide against the null hypothesis. The power again depends on *n, s,* π*,* but additionally on the required confidence level (*1-*α), the difference between the true proportion, the threshold π_0 and on the confidence interval or testing procedure used for evaluation.

Power will be examined for the proof of safety in testing against GMO contamination, where the hypotheses are

$H_0: \pi \geq \pi_0$ vs. $H_1: \pi < \pi_0$, with thresholds π_0 = 0.005 or 0.009.

Again, the design is examined for the 3 relevant situations of 1) limited number of assays and free group size, 2) limited group size and variable number of assays, and 3) limited total number of units.

4.5.2 Restriction 1: Limited number of assays, variable group size

Figure 19 shows power for *n*=10, *s*=10, 20, 50, 100 and relevant values of the binomial parameter π against the hypothesis $H_0: \pi \geq 0.005$ (compare the coverage probability and bias in figures 11 and 12 for the same *n* and *s* but π<0.05). For the shown values of π*, n, s*, bias is negligible. Obviously it is impossible to show significance using very small values of *s* for this hypothesis,

using the valid methods Wilson, Agresti-Coull or Clopper-Pearson because of the very conservative performance of these procedures for small *n, s* and π. The Wald interval has higher power to decide against H_0 due to its violation of the nominal coverage probability. Whether a method has equivalent or superior power compared to the others depends on *n, s,* π_0. Here, for *n*=10, *s*=50, Wald and Wilson Score CI show the same power whereas Clopper-Pearson and Agresti-Coull perform similarly poor. With *s*=100 Clopper-Pearson, Wilson and Agresti-Coull perform similar and Wald is different.

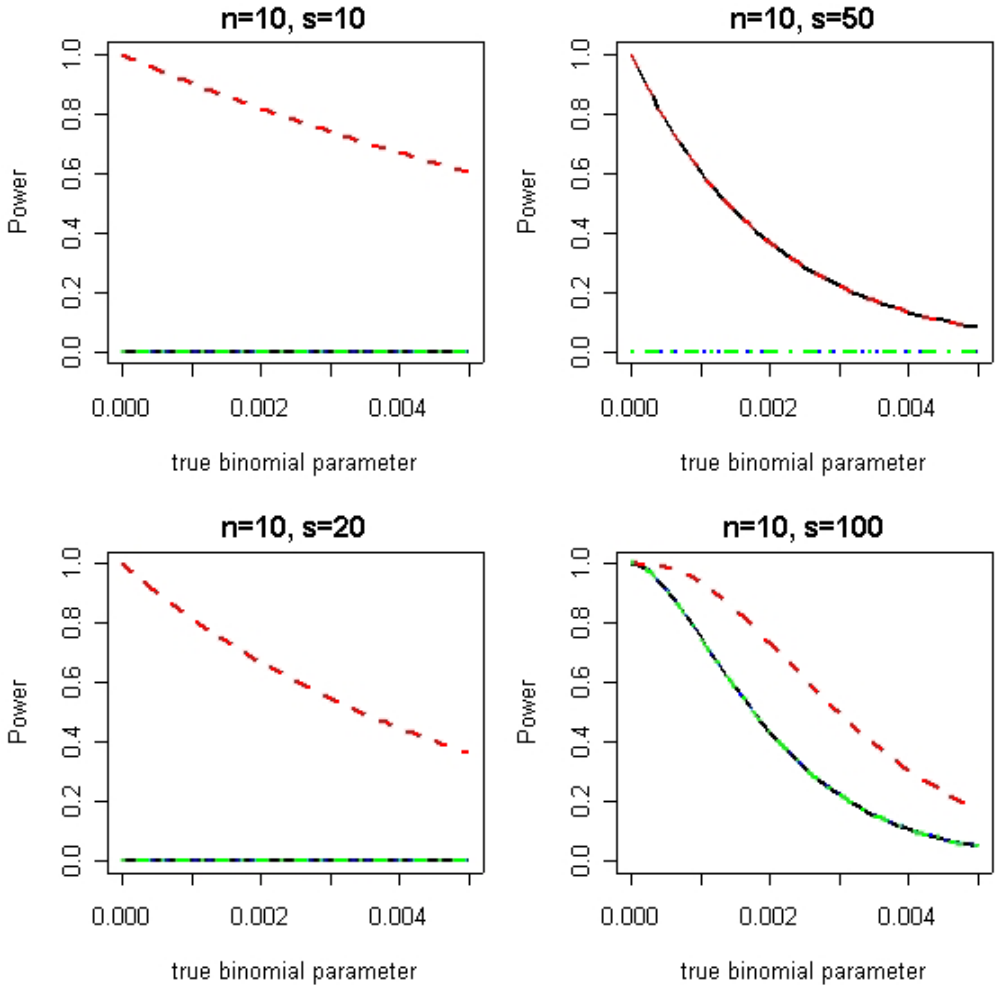

Figure 19: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for constant number of assays n=10, s=10, 20, 50, 100 π=0,…,0.05

Superiority in terms of power also depends on the hypothesis, what is shown in figure 20 for H_0: $\pi \geq 0.009$ and the same combinations of *n* and *s*.

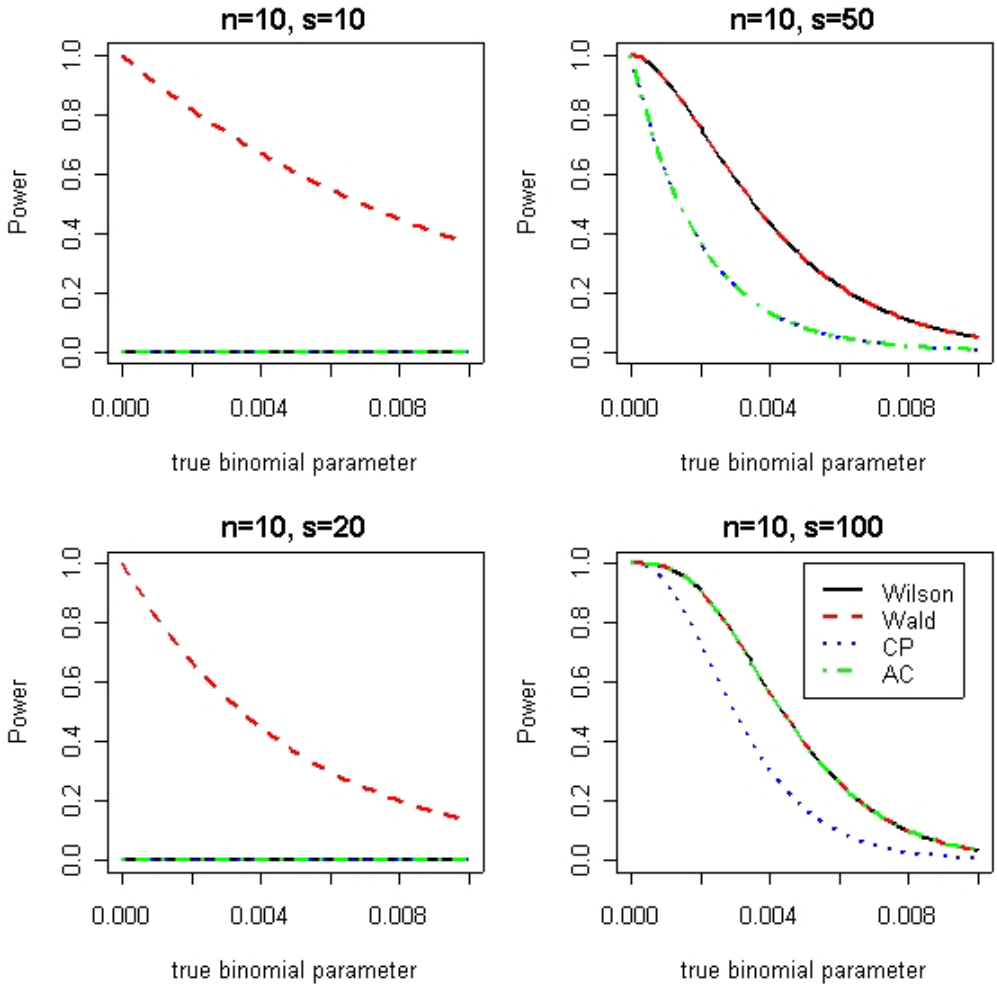

Figure 20: Power to reject H_0: $\pi \geq 0.009$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for constant number of assays n=10, s=10, 20, 50, 100 π=0,…,0.05

The objective in design of experiments with limited number of assays can be to find a group size promising sufficient power. In this situation the exact value of the true parameter is of course unknown, so the decision on *s* must be made for an assumed range of values of π. Figure 21 illustrate this problem for the hypothesis H_0: $\pi \geq 0.005$, with a fixed number of assays *n*=10 and increasing group size *s* from 1 to 200, resulting in increasing total number of units *n*s*=10,…, 2000 These combinations *n, s* are shown for four different values of π = 0.001, 0.002, 0.003, 0.004. For the shown combinations of *n, s* and π, bias is negligible.

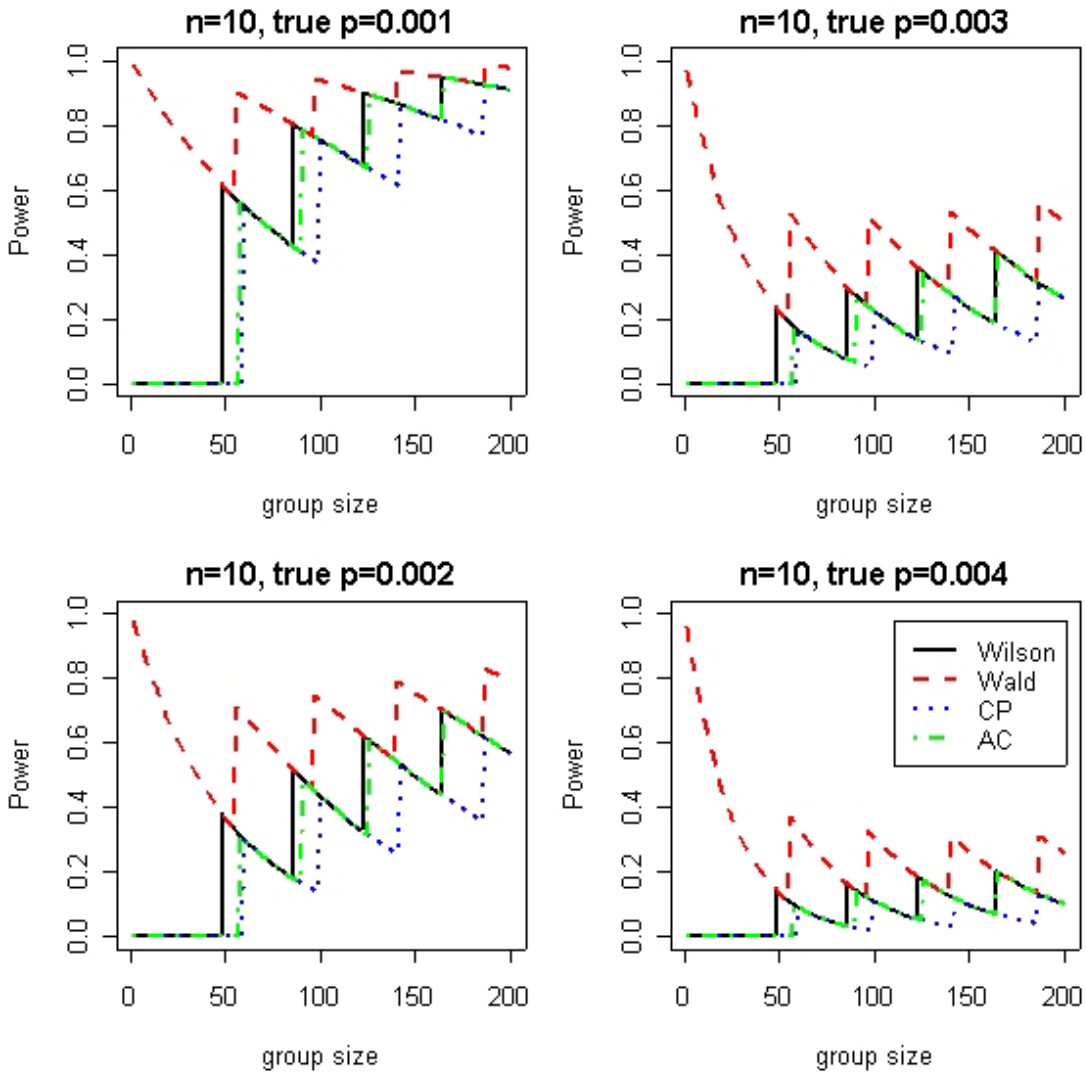

Figure 21: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for constant number of assays n=10, s=1,…,200, π=0.001, 0.002, 0.003, 0.004

It reveals that for these combinations power increases with increasing group size s and increasing difference π_0 - π, i.e. increasing non-centrality. Whereas Wald has highest power in nearly all cases due to its low coverage probability, Clopper-Pearson has either same or lower power than Wilson and Agresti-Coull. Important for experimental design is the issue that with increasing group size s, power does not increase monotonous, but with local maxima and minima. The position of maxima stays the same for different values of π, but changes for other hypotheses (f.e. π_0=0.009, π_0= 0.01). This is shown by the following graphs: In the first row, power against the hypotheses H_0: $\pi \geq 0.005$ and H_0: $\pi \geq 0.009$ is plot for n=10, s=1,…,200 and π=0.0025. In the second row,

the corresponding coverage probabilities for π = 0.005 and π = 0.009 and the same *n* and *s* are shown. Local maxima of power appear for those combinations of *n* and *s*, for which coverage at π_0 shows local minima.

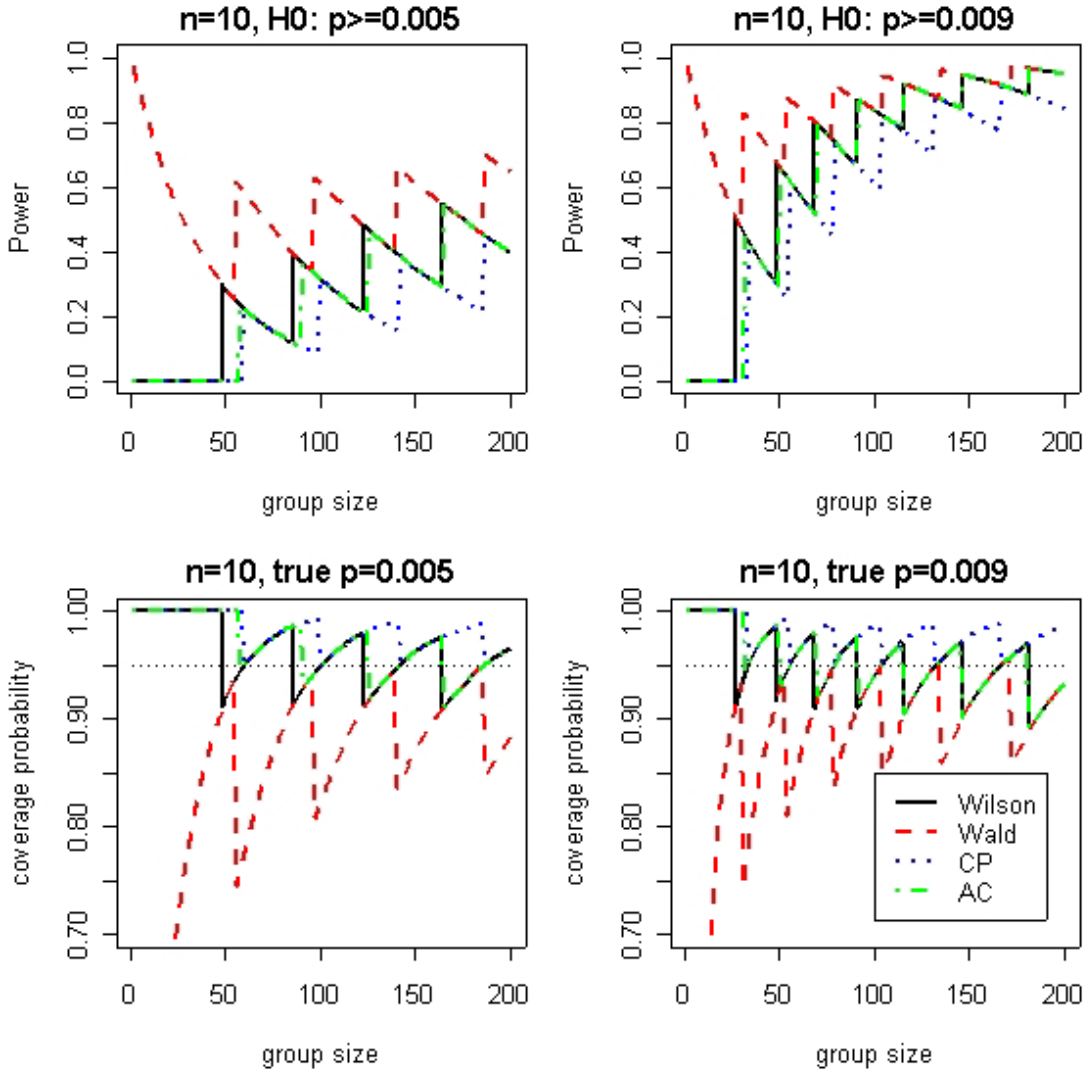

Figure 22: constant number of assays n=10, s=1,…,200, upper 95% Wald, Wilson, Agresti-Coull and Clopper-Pearson limits
First row: power to reject H_0: $\pi \geq 0.005$ and H_0: $\pi \geq 0.009$ at $\pi = 0.0025$
Second row: Coverage in case that H_0 is true: $\pi = 0.005$ and $\pi = 0.009$

The actual coverage probability of a CI method in a given group testing trial with a given design (*n,s*) depends on the true, unknown parameter π. This is illustrated by figure 23, showing the coverage probabilities of upper 95%-confidence limits for *n*=10, increasing group size *s*=1,…, 200 and values of π= 0.001, 0.002, 0.003 and 0.004.

Obviously, it depends on π, which actual coverage probability results for a given design (*n, s*).

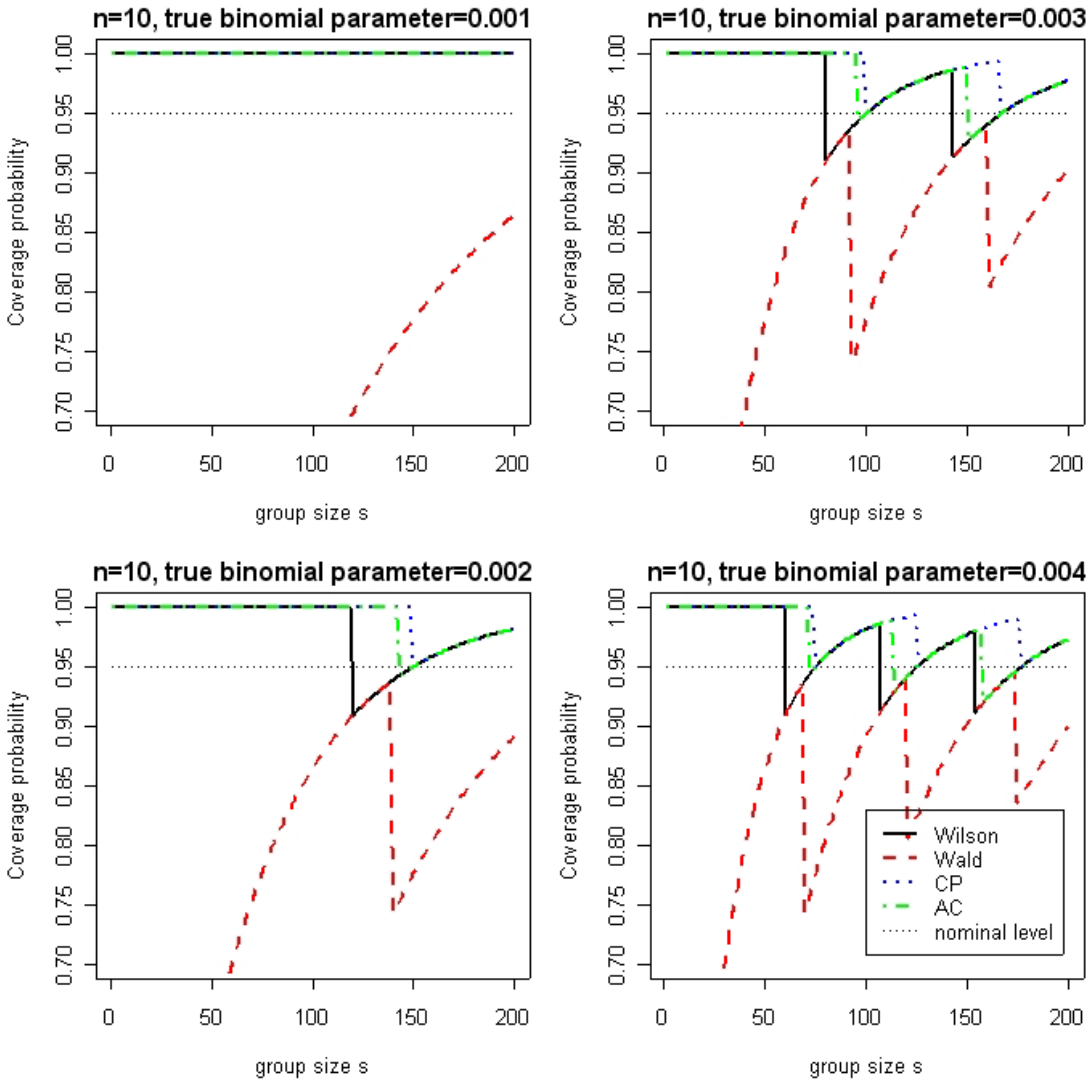

Figure 23: Coverage probability of upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for constant number of assays n=10, s=1,…,200, π=0.001, 0.002, 0.003, 0.004

Thus, increasing the group size *s* until a certain power is achieved will end up at local maxima, which depend on the given hypothesis and CI-method. This procedure does not select the group size for which a procedure is generally most liberal, but these values of *s* are those for which the coverage probability is minimal for $\pi=\pi_0$. In case of Clopper-Pearson, this means choosing the group size for which coverage is closest to the nominal level, whereas Wilson or Agresti-Coull will be most liberal for this *s* if $\pi=\pi_0$.

4.5.3 Restriction 2: Limited group size, variable number of assays

For *s*=10 and *n*=10, 20, 50, 100, power was calculated for the null hypothesis H_0: $\pi \geq 0.005$. Obviously, the group size *s* was chosen too small: even with 50 assays Wilson, Agresti-Coull and Clopper-Pearson failed to show any significance. Again, whether a single CI-method has the same of different power than another, depends on the chosen *n, s* and π_0.

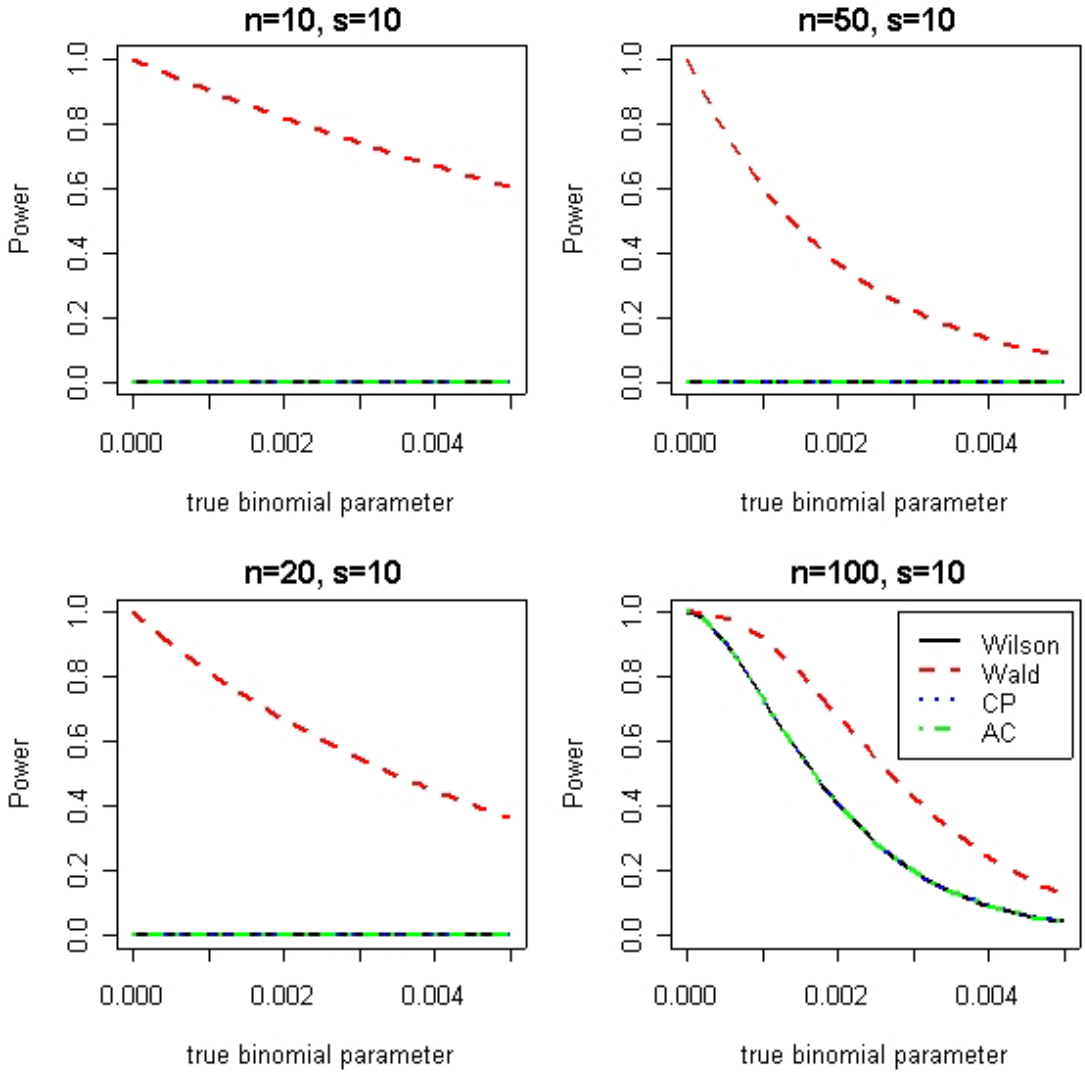

Figure 24: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for constant group size n=10, n=10,20,50,100, π=0,…,0.05

Objective of planning an experiment with a fixed group size *s* might be to increase the number of assays *n* until a sufficient power is achieved. If this is done, one will end up again at the local maxima of power. The values of *n* for

which power has a local maximum again depend on the hypothesis and are those *n*, for which the given method shows minimal coverage probability if $\pi=\pi_0$. The following graphs show that the position of local optima of power are not dependent on the true parameter π. Here always power against H_0: $\pi \geq 0.005$ is calculated for *s*=10, *n*=5,…,100 and π=0.001, 0.002, 0.003, 0.004.

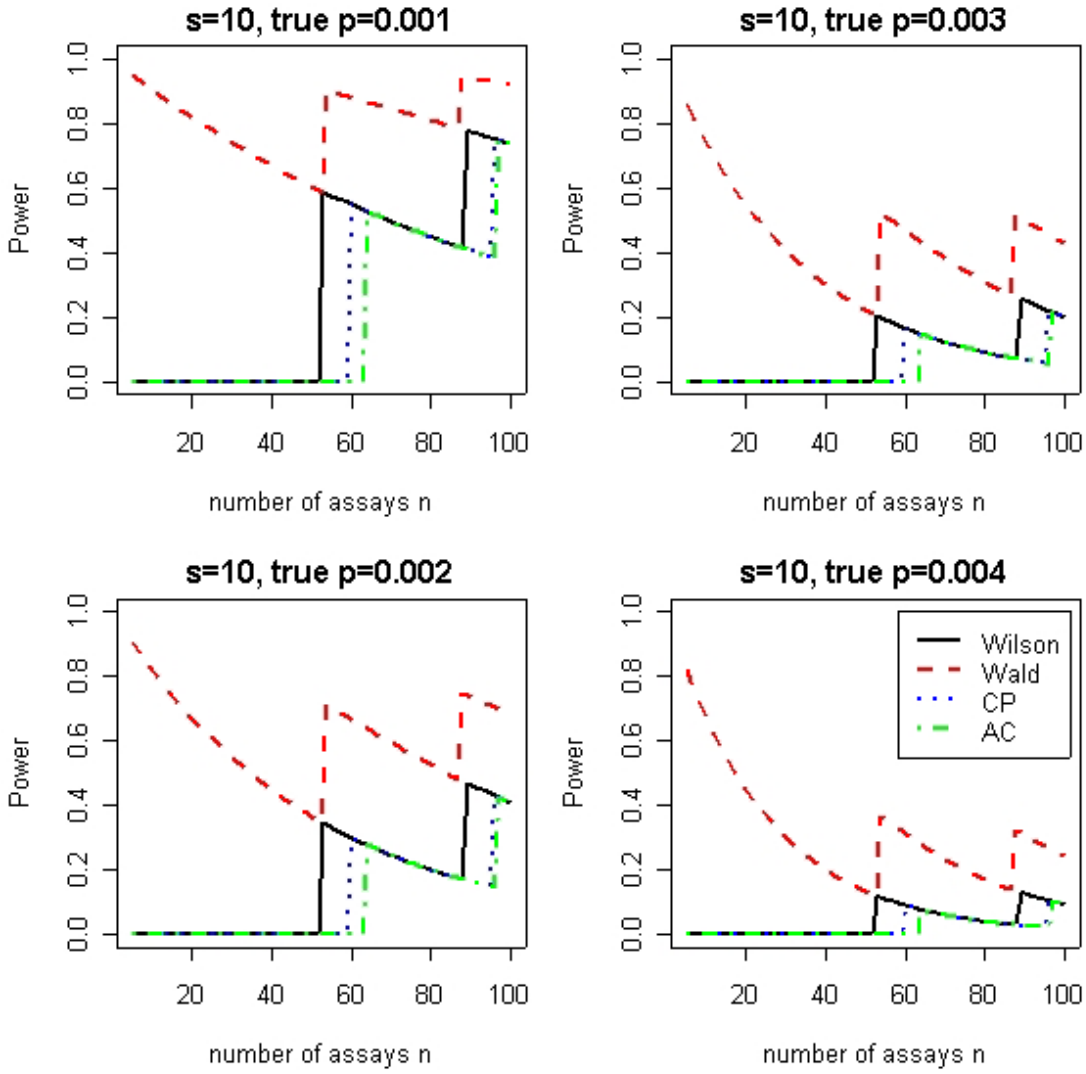

Figure 25: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper-Pearson limits for constant group size s=10, n=2,…,100, π=0.001, 0.002, 0.003, 0.004

For decreasing non-centrality, the general level of power decreases, again Wald has much higher power than exact and Score type CI because of its low coverage probability. Here also the conservative performance of Agresti-Coull CI becomes more obvious: Its very conservative region is longer than that of Clopper-Pearson.

The first row of figure 26 shows the power for a fixed *s*=10 and *n* increased from *n*=5 until *n*=100, where the true binomial parameter is always kept π=0.0025. Power is shown for two hypotheses: H_0: $\pi \geq 0.005$ and H_0: $\pi \geq 0.009$. The second row then shows the corresponding coverage probabilities of the methods for the same s and n but $\pi=\pi_0=0.005$ and $\pi=\pi_0=0.009$. Obviously, the position of local power optima corresponds to the local minima of coverage probability at the hypothetical parameter π_0.

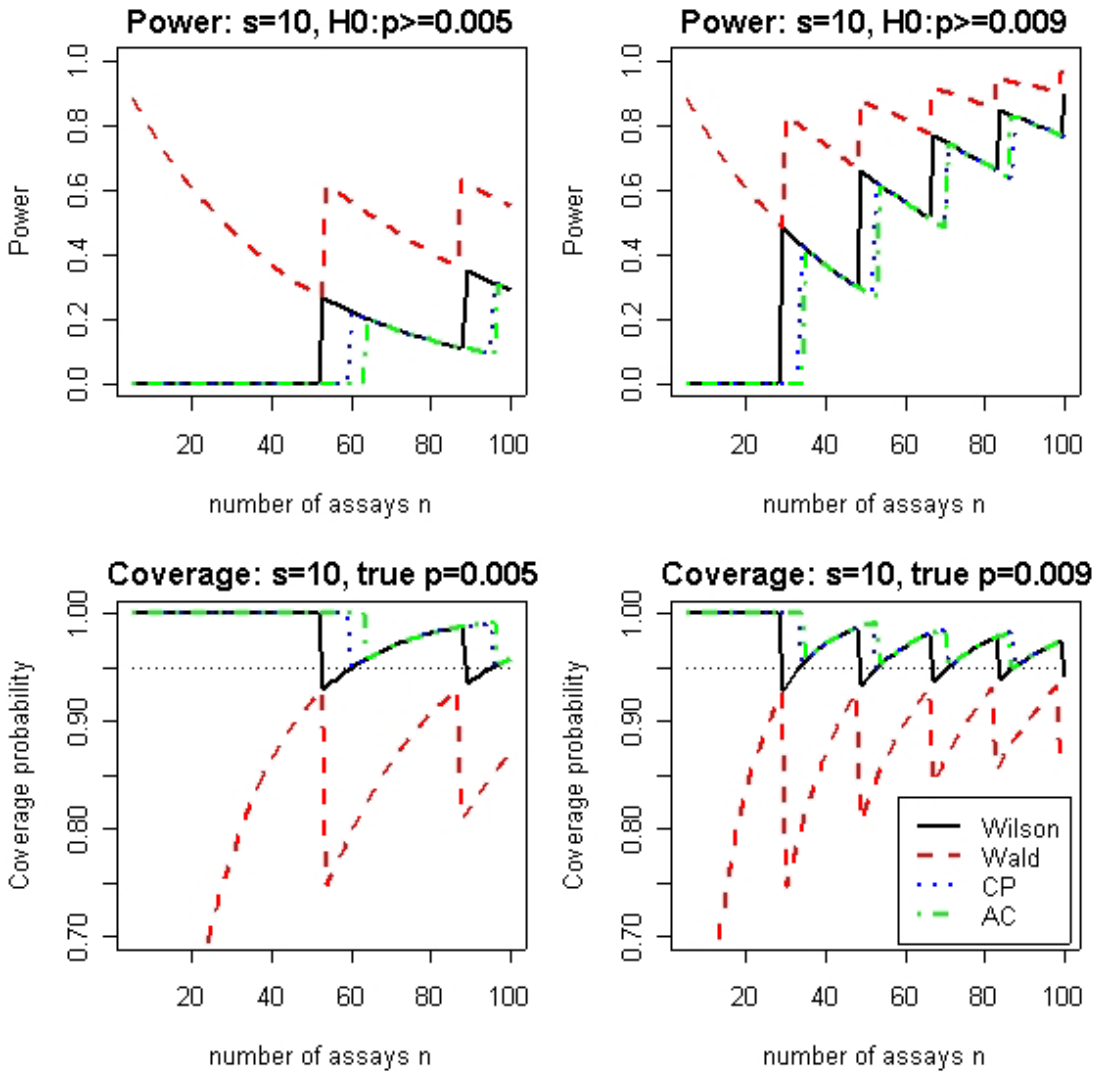

Figure 26: constant group size s=10, s=1,…,100, upper 95% Wald, Wilson, Agresti-Coull and Clopper-Pearson limits
First row: power to reject H_0: $\pi \geq 0.005$ and H_0: $\pi \geq 0.009$ at $\pi = 0.0025$
Second row: Coverage in case that H_0 is true: $\pi = 0.005$ and $\pi = 0.009$

Minimal *n* and *s* in the proof of safety

In section 4.3.2 it was shown that for small and intermediate numbers of observations *n*, a certain range of small π is always smaller than the upper bound of Wilson, Agresti-Coull and Clopper-Pearson CI. This very conservative performance of the valid methods results in the fact that the null hypothesis of a proof of safety H_0: $\pi \geq \pi_0$ can never be rejected for small π_0. As shown in the previous figures of section 4.4.1 and 4.4.2, increasing the group size *s* leads results in decreasing this conservative range of π as well as increasing the number of observations *n* if group size *s* is fixed. Thus, even for a very limited number of assays, H_0: $\pi \geq \pi_0$ can be rejected if the group size is chosen appropriate. Figure 27 summarizes the relation between *n, s*, and the smallest rejectable π_0 for a 95% upper Clopper-Pearson and Wilson confidence limit, respectively.

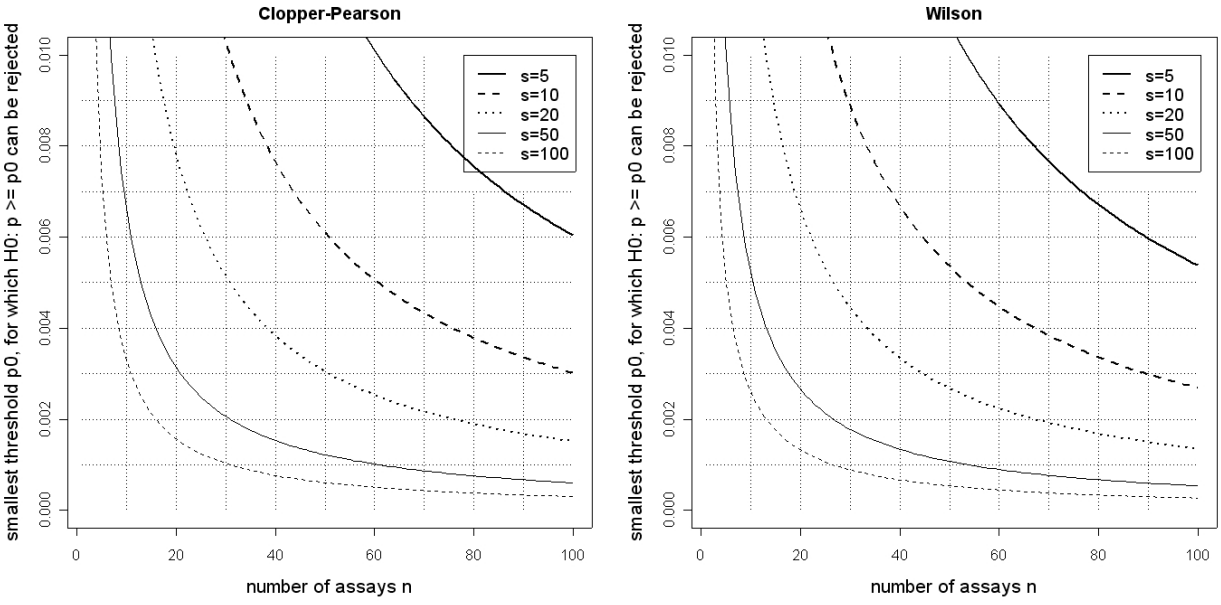

Figure 27: smallest threshold π_0 for which H_0: $\pi \geq \pi_0$ can be rejected for variable n and s= 5, 10, 20, 50, 100

To reject H_0: $\pi \geq 0.005$ in a group testing experiment, using a common group size *s*=10, at least *n* > 61, 55 groups will be required for Clopper-Pearson and Wilson respectively. If group size *s*=20 is used, only *n*>31 groups are needed for Clopper-Pearson and about *n*>27 if the upper Wilson limit is used. Clopper-Pearson in general will require higher minimal sample sizes because of its more conservative performance. Using larger groups enables to perform a

proof of safety for even smaller thresholds, as long as assumptions for group testing can be fulfilled. In simple binomial testing, much higher minimal numbers of observations *n* are needed for rejection of the same hypotheses (compare figure 10).

4.5.4 Restriction 3: limited total number of units: allocation to n and s

The two previous sections showed the effect of increasing total number of units either by increasing number of assays *n* while the group size is constant or by increasing the group size *s* while the number of assays is kept the same. Both has similar effects on power. In case of a fixed total number of units different allocations either to higher number of groups or a higher group size are possible. The question arises whether an optimal design can be found in this case. Figure 18 revealed, that coverage probability does not show a clear tendency for increasing group size for a constant total sample size *n*s*.

Figure 28 shows the power of the CI methods against H_0: $\pi \geq 0.005$ for different sets of *n, s* resulting in a total sample size of 1000, 5000 and 10000. The true probability in all plots is $\pi = 0.0025$. While increasing group size *s* from 1,2,4,8 to 200 in the first plot, the number of assays *n* is decreased from 1000, 500, 250, 125 to *n*=5. Each plot starts in the left side with simple binomial testing of *n*s* units, in the middle of each plot the power for *n*=10 is shown. If the number of assays *n* is decreased in a constant total number of units, power remains close to the power of simple binomial testing, as long as bias stays negligible and number of assays does not become too small (<10). Compared to simply increasing the group size *s* while keeping *n* constant (5.2.1), here the increase of bias is accelerated because both increasing *s* and decreasing n leads to higher bias. Figure 29 shows power for the same $\pi = 0.0025$ and the same sets of *n* and *s* but against H_0: $\pi \geq 0.009$.

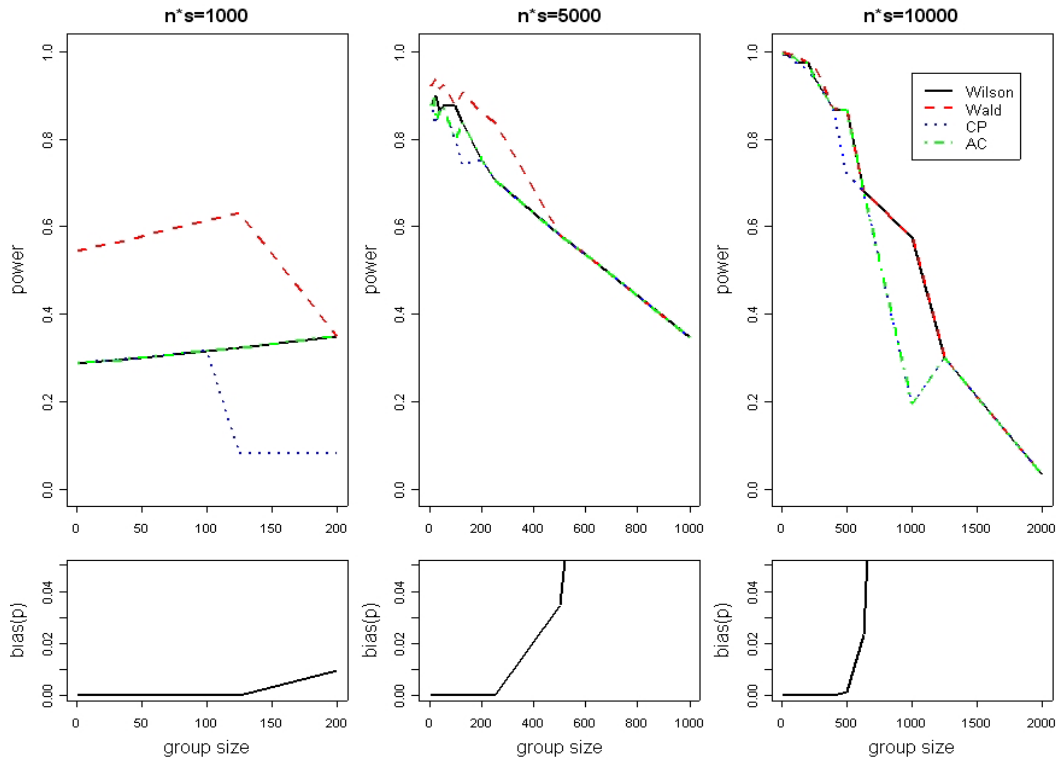

Figure 28: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits and bias (p) for combinations (n,s) resulting in constant total sample size n*s=1000, 5000, 10000, π=0.0025

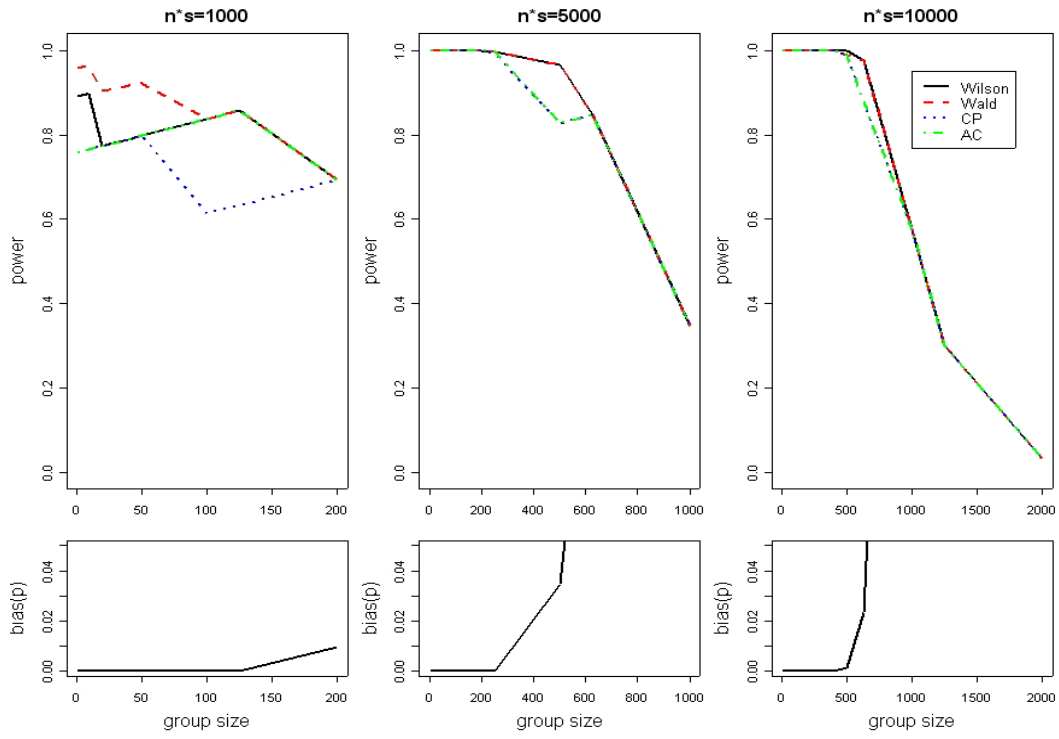

Figure 29: Power to reject H_0: $\pi \geq 0.009$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits and bias (p) for combinations (n,s) resulting in constant total sample size n*s=1000, 5000, 10000, π=0.0025

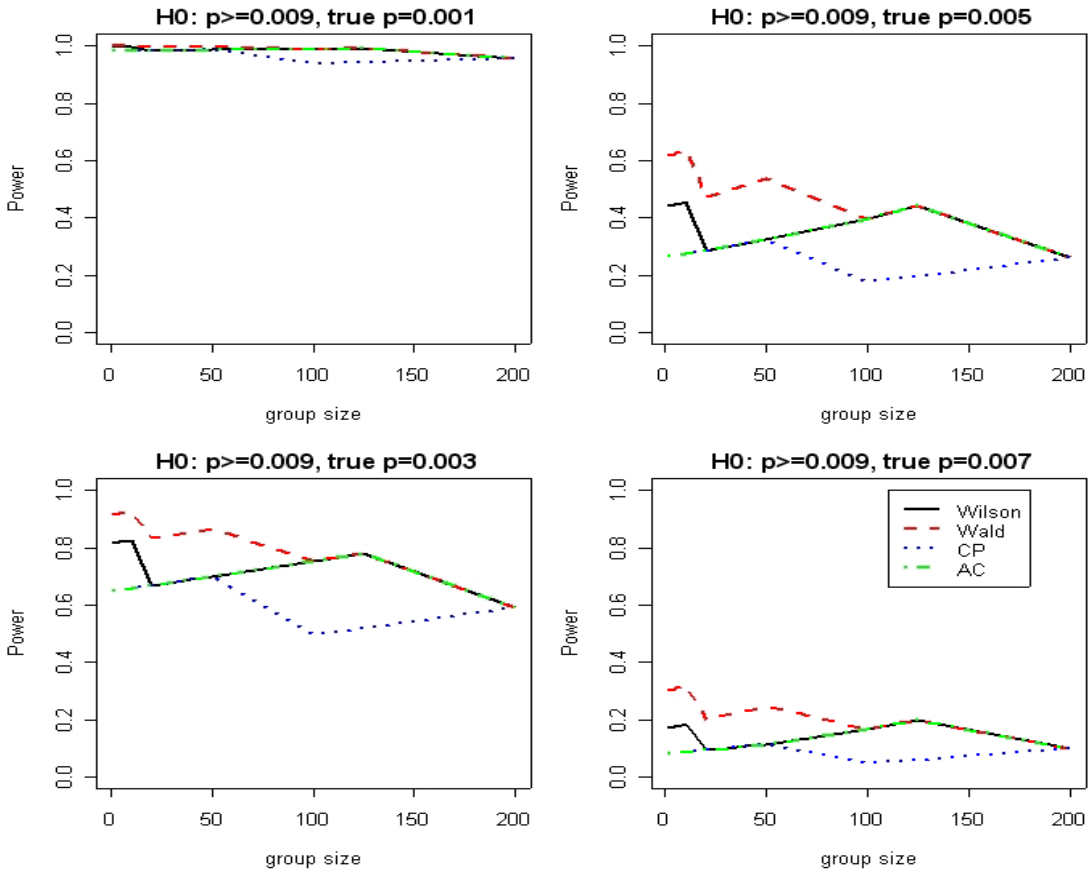

Figure 30: Power to reject H_0: $\pi \geq 0.009$ using upper 95% Wald, Wilson, Agresti-Coull and Clopper Pearson limits for combinations (n,s) resulting in constant total sample size n*s=1000, for π=0.001, 0.003, 0.005, 0.007

Figure 30 shows the power of the 95%-CI against a given H_0: $\pi \geq 0.009$ and a total number of units $n*s$ = 1000, but different values of the true parameter π. The graphs show power in dependence of s=1,…, 200, but in the same time, number of assays n reduces from 1000 for s=1 to 5 for s=200. The power in average only slightly decreases with decreasing n, but local optima may exist. Again, the local optima seem to depend not on the true parameter π, but are the same, as long as the hypothesis and total number of units are the same.

For experimental design, it can be concluded: Beside the required α and the difference between π and π_0, a sufficient total number of units in the experiment is most important for a high power. For a fixed total number of units, the mean power stays nearly constant, independent of the allocation to either group size or number of assays, as long as the group size is appropriate for the given π. If the number of assays is chosen to small, power decreases. In other words, starting from a simple binomial experiment (group size s=1), the number of

assays *n* can be reduced by increasing group size without greatly reducing power. The minimal n (or the maximal *s*) until which power does not decrease greatly, depends on the total number of units available, the true but unknown π and of course the difference between π and π_0.

The graphs illustrate again the dependency of the general level of power on the difference between π and π_0 and that Clopper-Pearson in average has a lower power than Wilson Score.

4.5.5 Approximate sample size calculation

As shown in sections 4.5.3 and 4.5.4, simply increasing *n* until the desired power is achieved for a fixed *s* or *n*s*, will result in the *n*, for which the CI method shows the lowest actual coverage in case that H_0 is true. This is appropriate for exact methods as Clopper-Pearson, but for the Wilson and Agresti-Coull CI sample sizes will be chosen for which these methods are most liberal under H_0. Alternatively, an approximate sample size calculation can be derived for the Wilson CI, using its correspondence to the Score test. Corresponding to the derivation of an approximate formula for sample size calculation for the one-sided Wald test (shown in Bock, 1998), a formula for the Score test will be derived.

The Score test for simple binomial testing is based on the assumption:

$$\frac{p-\pi}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0,1)$$

For performing a test with a type I error rate α in case that H_0 is true ($\pi=\pi_0$) and a type II error rate of β in case that π has a certain value π_1 under H_1, the following is required to hold:

$$\frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = z_{1-\alpha} \text{ and } \frac{p-\pi_1}{\sqrt{\pi_0(1-\pi_0)/n}} = z_{1-\beta}$$

resulting in the relation:

$$n = \frac{\pi_0(1-\pi_0)\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{\left(\pi_1-\pi_0\right)^2}$$

from which the required sample size *n* for the Score test and the corresponding Wilson CI can be approximated for a given (π_1-π_0) and a required α and β. The power (*1-*β) for given *n,* α*,* π_1*,* π_0 can be calculated from:

$$z_{1-\beta} = \sqrt{\frac{n(\pi_1 - \pi_0)}{\pi_0(1-\pi_0)}} - z_{1-\alpha}$$

The Score test corresponding to the Wilson CI for binomial group testing is a Score test for the group scale proportion θ:

$$\frac{t-\theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} \sim N(0,1)$$

Then, the sample size can be approximated using

$$n = \frac{\theta_0(1-\theta_0)\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(\theta_1 - \theta_0\right)^2} \text{ where } \theta_0 = 1-\left(1-\pi_0\right)^s \text{ and } \theta_1 = 1-\left(1-\pi_1\right)^s$$

Since this approximation does not take the binomial distribution of the observed random variable *Y* into account, the resulting relations are monotone, without local maxima and minima. As based on the standard normal distribution, this approximate sample size calculation does not reflect the asymmetry of the Wilson CI coverage for small π and thus results in a biased estimation of sample size, depending on π, and whether upper or lower confidence limit is used. In fact this equation calculates the required sample size, which would be needed for the Score test, if the assumption of standard normal distribution would be exact.

Figure 31 shows the relation between power and number of groups for one-sided hypotheses and a fixed group size s=20 as calculated from the approximate formula and from closed calculation. The first plot shows the power to reject H_0: $\pi \geq 0.005$ if π=0.003 (power of the upper confidence limit), the second plot shows power to reject H_0: $\pi \leq 0.005$ if π= 0.007 (power of the lower confidence limit). So, for both plots the absolute value of (π-π_0) is the same, resulting in the same approximate power. Since the upper limit of Wilson tends to be conservative for small π and *n,* while the lower limit is liberal for small π and *n*, the actual power from closed calculation is lower for the upper and higher for the lower bound in these cases. For larger *n*, the opposite occurs.

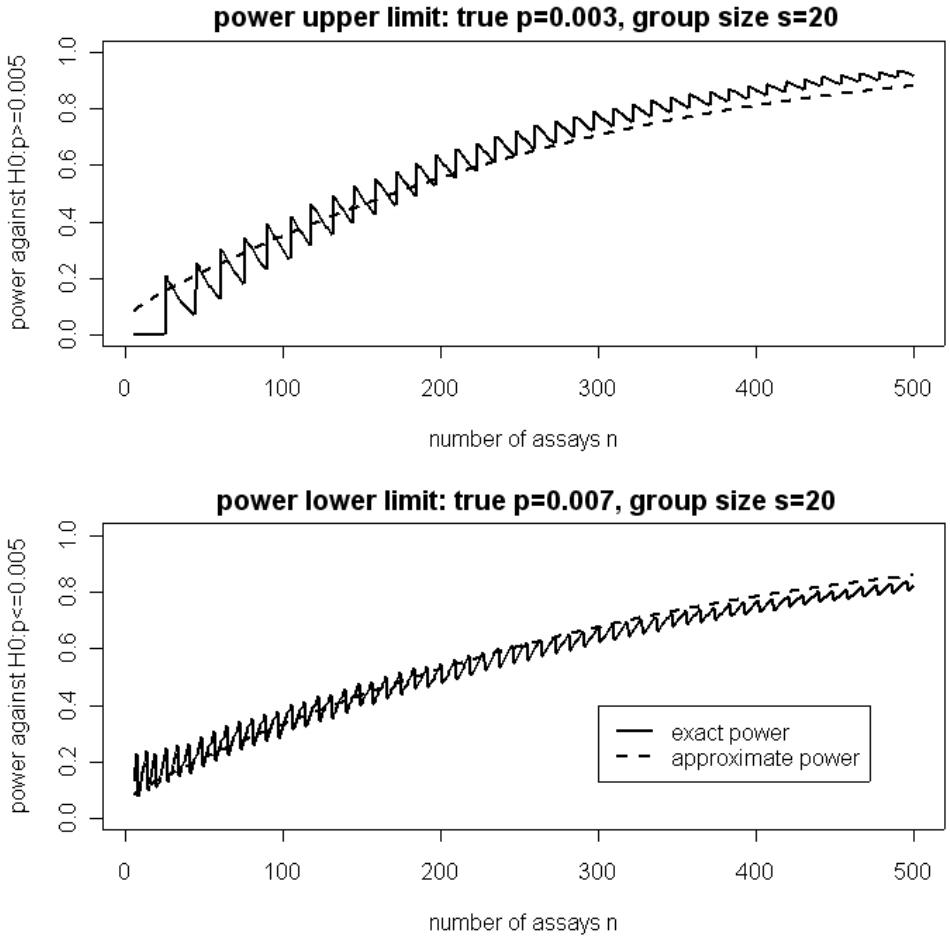

Figure 31: Power of Wilson CI to reject H_0: $\pi \geq 0.005$ for π=0.003 and H_0: $\pi \leq 0.005$ for π=0.007 for α=0.05, s=20, n=5,…,500

This leads to the conclusion, that approximate power or sample size calculation is not an improvement of closed calculation, because it might lead to truncated estimation of required sample size. It can only be recommended for use if the sample sizes n are to large for closed calculation.

4.5.6 Conclusions

Experimental design

The minimal sample size required for rejection of a null hypothesis in a proof of safety is much lower in binomial group testing than in simple binomial evaluation of each single unit. For a given number of individuals the number of observations *n* can be reduced greatly via assigning the individuals to groups of increasing size without greatly reducing power. To give a rule of thumb for π=0.001,…, 0.005, this might be done as long as the bias(*p*) should not exceed

100*π % and *n*>10. But for optimal decisions it is always recommended to directly calculate power and bias(*p*) of a design using the given functions.

Which combination *n* and *s* results in optimal power, depends on the threshold π_0 in the hypotheses, not on the actual π in the experiment, while the actual coverage of a design n, s of course depends on the actual π. The actual value of power of course depends on *n, s,* α and the difference π_0-π. If the design *n, s* is chosen iterative to achieve optimal power, one can expect minimal coverage in case that π equals the threshold π_0 but not for other values of π. This is acceptable for exact methods, while for asymptotic methods the choice of a design results which is liberal under H_0. The proposed approximate sample size calculation does not generally result in a more reasonable choice of sample size.

Generally, for the same underlying number of units, all methods perform better for high number of assays and low group size than vice versa. The highest number of assays payable is always recommended.

CI methods

In average Wilson Score has higher power than Clopper-Pearson, and Agresti-Coull is mainly between both, but close to Clopper-Pearson for small π and close to Wilson for medium values of π. Because of this, among the methods compared for the upper limits, Clopper-Pearson is recommended if the nominal level shall be guaranteed and if the number of groups is very small (*n<30*), whereas Wilson might be used to improve power if actual coverage shall not be strictly guaranteed and sample size *n* is not too small. Whether a special method is superior in a particular situation, depends on the particular choice of *n, s* and π_0.

A main problem in experimental design remains: π is unknown:

If a certain hypothesis H_0: $\pi \geq \pi_0$ shall be tested, it is secure but suboptimal to choose the design appropriate for the highest π which shall be shown to differ significantly from π_0, because for smaller π, bias will decrease. If estimation is main aim, it is a secure but conservative choice to optimize the design for the

upper bound of the range where π is expected in order to achieve a minimal MSE(*p*), as recommended by Swallow (1985).

5 Violation of assumptions

5.1 Unequal group size

In the consideration of coverage and power above, group size was assumed to be equal for all n observations. If group size is not equal, the methods for point estimation and even more interval estimation are not necessarily valid anymore.

There are two reasons for unequal group sizes in group testing experiments might be:

1) By accident, f.e. in vector transfer designs, single vectors might escape, die or only available with limitation resulting in slightly reduced group sizes for some observations.
2) If there is no reason to expect the true, unknown incidence within a certain range, choosing two or three different group sizes can be the most appropriate experimental design, as shown by Hepworth (1996). If true π is very small, only large group sizes will result in positive observations, but if true π is comparatively large, only small groups will provide some negative observations and thus an estimator different from 1.

Hepworth (1996, 2004) reviews methods for point estimation and proposes two methods for interval estimation.

5.2 Limited assay sensitivity and specificity

For biochemical assays the specificity and sensitivity might not 1. That is, groups might be misclassified erroneously as positive or negative, what is a violation of the assumptions 4 and 5 given in section 1.3. Rates of false positive or false negative assay decisions might be known from pilot studies (Xie et al. 2001). Xie et al. 2001, Remund et al. 2001 include this problem in their discussion of group testing analysis, Hung and Swallow (1999) discuss the special problem of dilution effects with regard to the optimal choice of group size.

6 A resampling confidence interval: an alternative?

The distribution of the estimator *p* is only approximately gaussian distributed for large n. For inappropriate sets of *s* and π the distribution of the group-scale-estimator *t* is even not symmetric anymore.

Therefore, an alternative confidence interval can be derived from the simulated distribution of the estimator *p* under the assumption that the observed *p* is true.

To do this, the observed value of *Y* is first used to calculate the estimator *p*. This estimator is used for random experiments, which simulate the group testing experiment under the assumption that the estimated *p* is the true individual probability π. The population of estimators resulting from numerous resamplings then is used to calculate confidence limits: the most extreme percentiles of this simulated distribution are 'cut off' so that the remaining part include at least (*1-α*)*100% of the simulated estimators. The R code for this procedure is given in the annex.

This procedure is used in the group testing context only to show that ordering of outcomes on group or individual scale is equivalent. It is inappropriate for interval estimation in group testing, because group testing is applied for small number of observations, for example *n*=5,…,50. In Efron and Tibshrani (1993) the use of resampling methods for binomial problems is shown for high numbers of observations. But for an experiment with only a small number of observations *n*, only *n+1* different outcomes can be present in the resampled populations: The resulting distribution is as discrete as the exact distribution, but the CI is constructed using the resampled populations and will also have only *n+1* possible interval bounds. Thus it will either be very conservative or very liberal, and can never be superior to exact calculations for small *n*.

7 A confidence interval explicitly constructed for one-sided hypotheses

7.1 A new confidence interval and a deviating recommendation

Recently, when this thesis was nearly finished, Cai (2005) published a paper concerning on “One-sided confidence intervals in discrete distributions”. In this he considers the upper coverage probabilities of known confidence interval methods for the binomial, poisson and negative binomial distribution. His findings agree with the findings of this thesis: The Wald intervals exhibits too low coverage for the upper bound for π close to 0 and the lower bound for π close to 1, whereas its lower bound is conservative for π close to 0 and its upper bound is conservative for π close to 1. Also the Wilson CI exhibits an asymmetric coverage probability, but here the estimated interval is biased to lower extend and in the opposite direction: the upper bound is conservative for small π and liberal for π close to 1, while the lower bound is liberal for small π and conservative for π close to 1. The corresponding Wald and Score intervals for the poisson and negative binomial distribution show an even worse asymmetric coverage for upper and lower confidence limit.

For comparison of the methods the natural oscillation of coverage probability due to discreteness of the underlying distributions makes it difficult to decide clearly for one method (compare Brown et al. 2001 for the two-sided case). Cai (2005) examined different CI methods by comparing the non-oscillating terms of the coverage probabilities using Edgeworth expansion. Among the known methods he recommends the Jeffreys prior interval because its coverage is the most symmetric. This is rather easy to compute for the simple binomial case using the quantiles B of the beta distribution:

$\left[0, B_{1-\alpha, Y+1/2, n-Y+1/2}\right]$ for the upper and $\left[B_{\alpha, Y+1/2, n-Y+1/2}, 1\right]$ for the lower bound.

Since the Jeffreys prior interval is not a satisfying solution because of slightly liberal performance, Cai used the Edgeworth expansion to derive the so called second-order corrected interval. This is based on the normal approximation, but uses a shifted midpoint and additional correction terms.

The upper bound of Cais second-order corrected interval for a binomial proportion π is:

$$\left[0, \tilde{p} + z_{1-\alpha}\sqrt{p(1-p) + \frac{\gamma_1 p(1-p) + \gamma_2}{n}} \Big/ \sqrt{n}\right],$$

and the lower bound is:

$$\left[\tilde{p} - z_{1-\alpha}\sqrt{p(1-p) + \frac{\gamma_1 p(1-p) + \gamma_2}{n}} \Big/ \sqrt{n}, 1\right], \text{where} \quad \tilde{p} = \frac{(Y-\eta)}{(n+2\eta)}, \quad \eta = \frac{1}{3}z_{1-\alpha}^2 + \frac{1}{6},$$

$\gamma_1 = -\frac{13}{18}z_{1-\alpha}^2 - \frac{17}{18}$ and $\gamma_2 = \frac{1}{18}z_{1-\alpha}^2 + \frac{7}{36}$, and z is the quantile of the standard normal distribution. Both can be easily applied as two-sided intervals by replacing *α* by *α/2*.

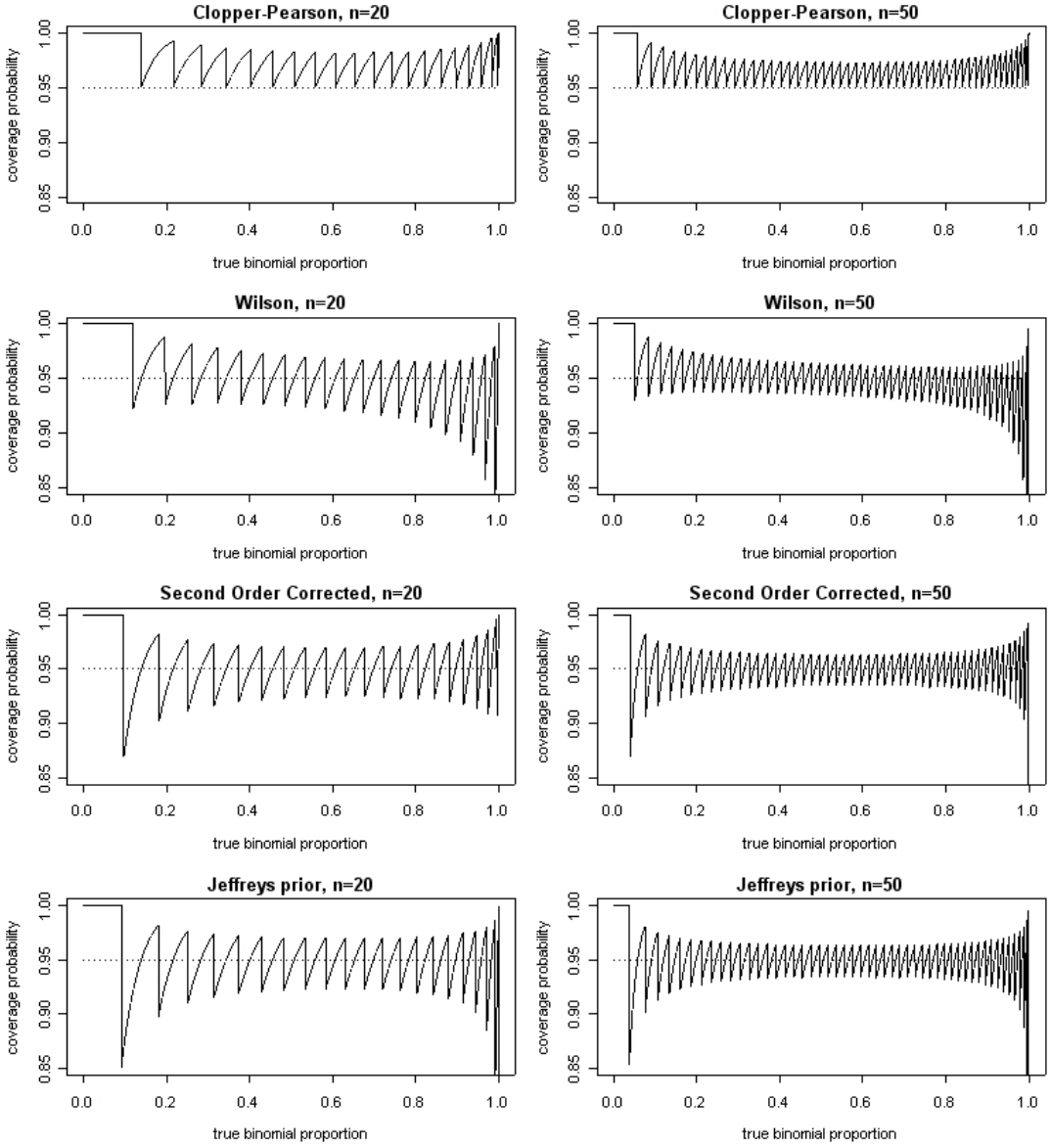

Figure 32: Coverage probability of upper 95%-confidence limits of Clopper-Pearson, Wilson, Second-order corrected and Jeffreys prior interval for π=0,…,1, n=20 and 50.

Figure 32 compares the coverage probabilities of upper 95% confidence limits of Clopper-Pearson, Wilson, Second-order Corrected and Jeffreys prior intervals.Obviously, the upper bounds of the two methods recommended by Cai (2005) perform more symmetric over the total range of π. Especially, the second-order corrected interval (SOC) performs much better for π close to 1. SOC and Jeffreys prior are more liberal than the Wilson CI for small π. Cai (2005) mainly compares the methods on the 99% level: then Wilson is more conservative for small π and Cais recommendation of SOC and Jeffreys prior is reasonable. The performance of the four methods thus will also be compared for the upper 99% limits and sample sizes of n=20 and 50 in figure 33.

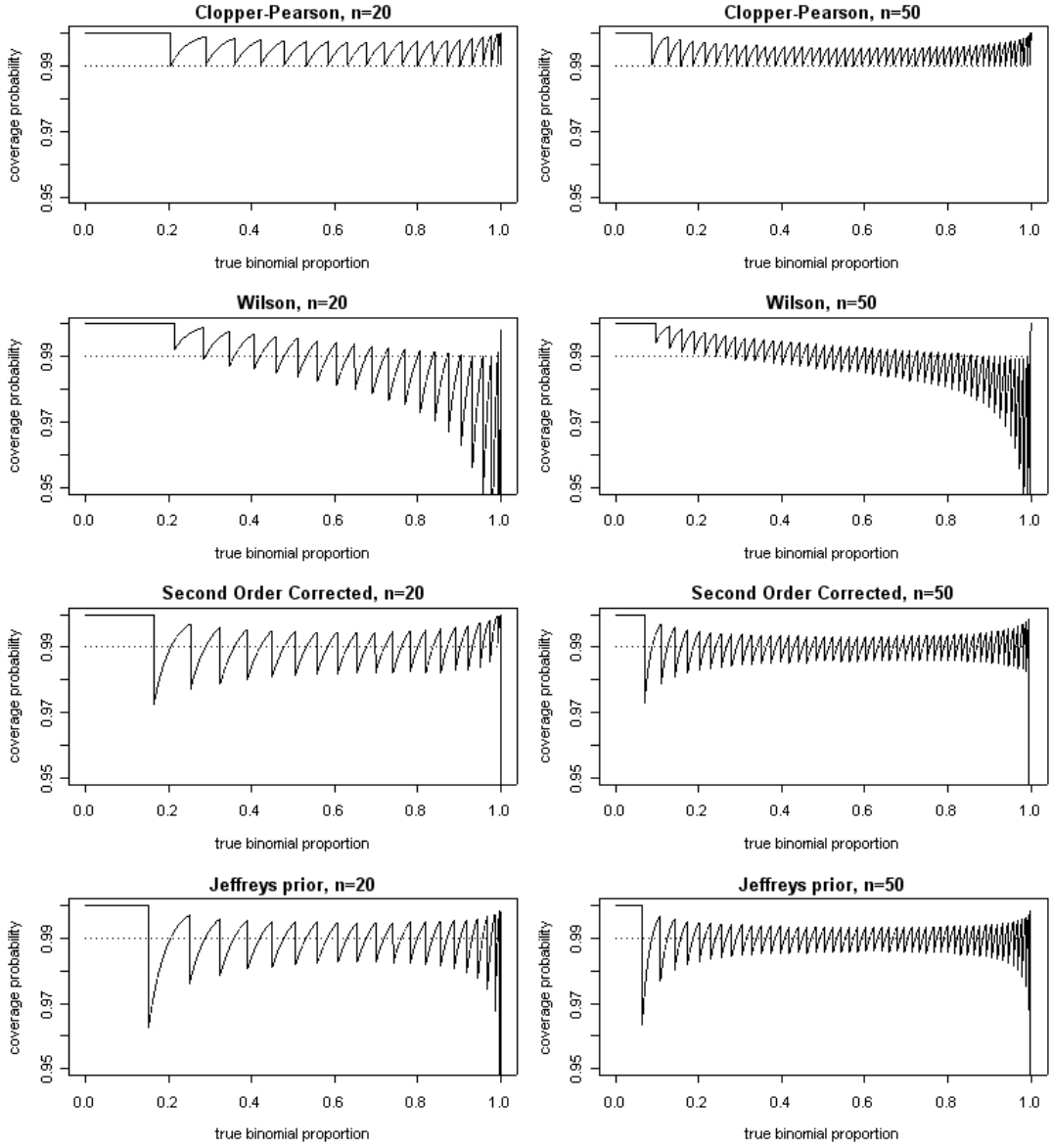

Figure 33: Coverage probability of upper 99%-confidence limits of Clopper-Pearson, Wilson, Second order corrected and Jeffreys prior interval for π=0,…,1, n=20 and 50.

Due to the shorter very conservative region (accompanied with a more liberal performance in close connection to it) of the SOC and Jeffreys prior interval, these can expected to require a lower minimal sample size for rejection of a certain null hypothesis H_0: $\pi \geq \pi_0$. Figure 34 compares SOC, Jeffreys prior and Wilson Score intervals with respect to the smallest threshold π_0 for which H_0: $\pi \geq \pi_0$ can be rejected for a given sample size n. Compare figure 10, for Wilson, Agresti-Coull and Clopper-Pearson.

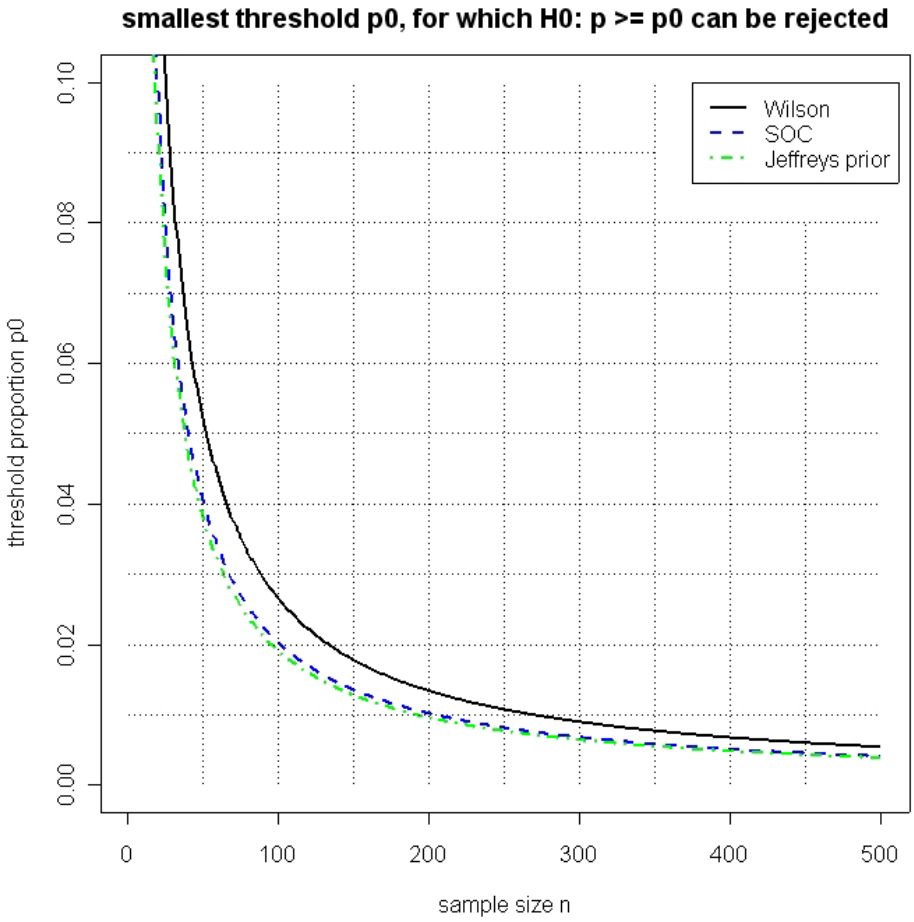

Figure 34: the smallest threshold π_0 for which H_0: $\pi \geq \pi_0$ can be rejected using upper 95% Wilson, Second-order corrected (SOC) and Jeffreys prior confidence limits, n=2,…,500

F.e. to reject H_0: $\pi \geq 0.01$ with 95% confidence level, SOC will at least need n≥203, Jeffreys prior will at least need n≥191, whereas usage of the Wilson CI requires n≥268.

A problem of the second-order corrected interval (SOC) is that the upper 95%-bound may not include the estimator p=1 for Y=n, especially if n becomes larger than shown here. This leads to the downward spikes of the coverage near π=1 for the upper bound. This might be easily avoided by setting p_U=1 for

$Y=n$ and correspondingly $p_L=0$ for $Y=0$, as is done for the Wilson interval with continuity correction. For a confidence level of 99 %, the upper bound can be slightly greater than 1 for at least $Y=n$, n-1, n-2, resulting in problems in calculation of SOC for group testing. To avoid leaving the range of definition for t in the transformation from group scale to individual scale, the SOC interval in the following will used with the correction that any upper limits $t_U>1$ will be set $t_L=1$.

7.2 Application for group testing

The two intervals recommended by Cai (2005) can straightforward be applied for group testing as described in general for CI construction on the group scale (section 3.2).

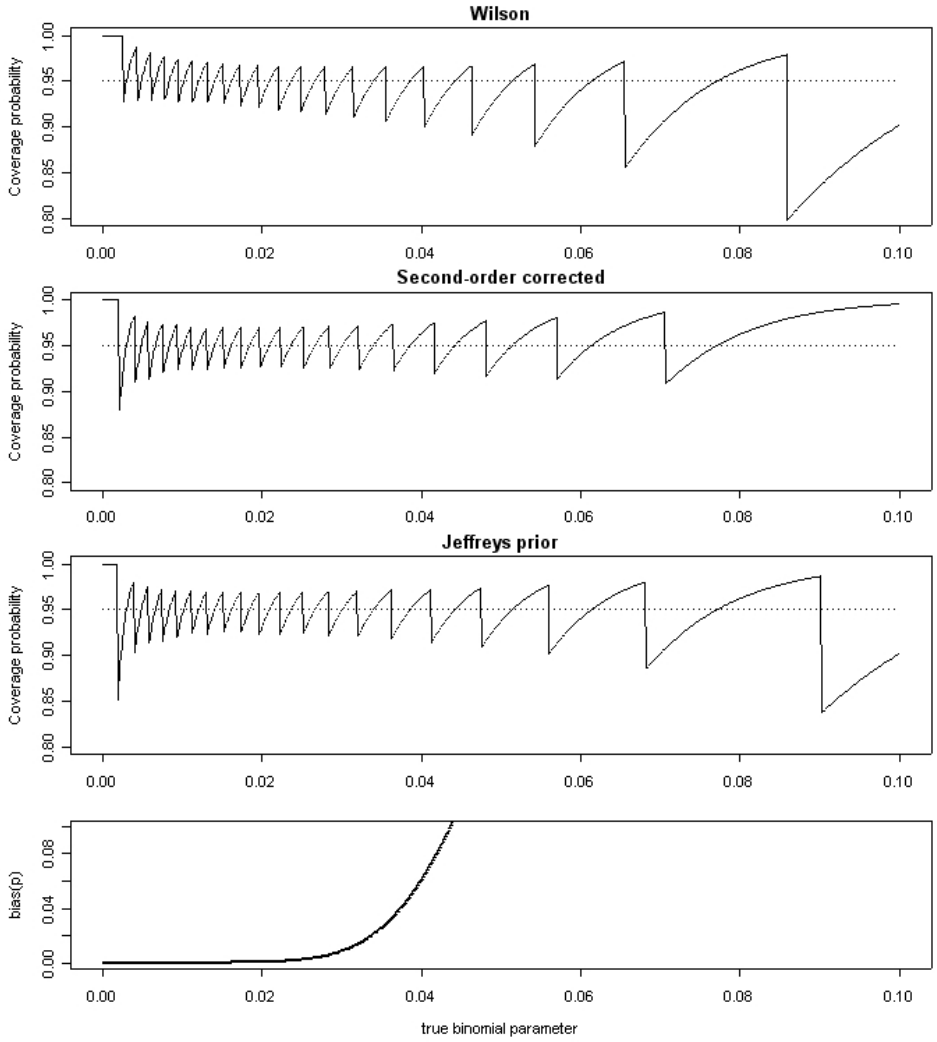

Figure 35: Bias(p) and Coverage probability of upper 95%-confidence limits of Wilson, Second order corrected and Jeffreys prior interval applied in binomial group testing for π=0,…,0.1, n=20, s= 50.

As shown in section 4, the performance of binomial intervals used for group testing is the same as for the simple binomial case, but truncated towards smaller values of π as group size *s* increases. Because of its symmetrical coverage, the SOC interval is superior over Wilson in case that group size was chosen too high for the actual proportion π: SOC will not become liberal as Wilson does in this case. Figure 35 shows this for n=20, a group size of s=50 and π between 0 and 0.1.

For values of π>0.04, for which group size s=50 becomes clearly inappropriate because E(*p*) is more than double of π, Wilson and to smaller extend Jeffreys prior CI become liberal, while the SOC interval in average is close to the nominal level. Therefore, the SOC interval is recommended for one-sided estimation of rare traits without testing of a clear threshold. Those problems may occur in plant breeding, if lines with low proportion of a special trait shall be selected. However, the situation of inappropriate group size stays undesirable and should be avoided by a conservative choice of group size if possible. In case that a clear threshold exists (as in GMO-testing) and design can be chosen appropriate for a certain range of $\pi<\pi_0$, and too large group sizes are avoided, also the Wilson CI can be used.

The results for the power and experimental design in section 4.5 as well are transferable to the SOC and Jeffreys interval. Because of the slightly more liberal performance for small π, they have a higher power to reject null hypothesis H_0: $\pi \geq \pi_0$ for small threshold proportions and small n and s. Correspondingly, for rejection of a certain nul hypothesis, SOC and Jeffreys prior CI need lower number of assays for a certain group size or lower group size in case that the number of assays is limited than the Wilson CI requires for rejection of the same null hypothesis. Figure 36 shows coverage probabilitys and power of 99% upper confidence limits of Wilson, Jeffreys prior and SOC interval and bias(*p*) for a fixed number of assays *n*=20 and increasing group size *s*=1,…,200. The power to reject H_0: $\pi \geq 0.005$ is shown for π=0.002, whereas coverage is shown for π=0.002. Bias(*p*) is negligible over the whole range of *s*. The differences between Wilson and SOC, Jeffreys are big for small

group sizes *s*, i.e. for small total number of units *n*s,* while they become less important for increasing group size and increasing total number of units.

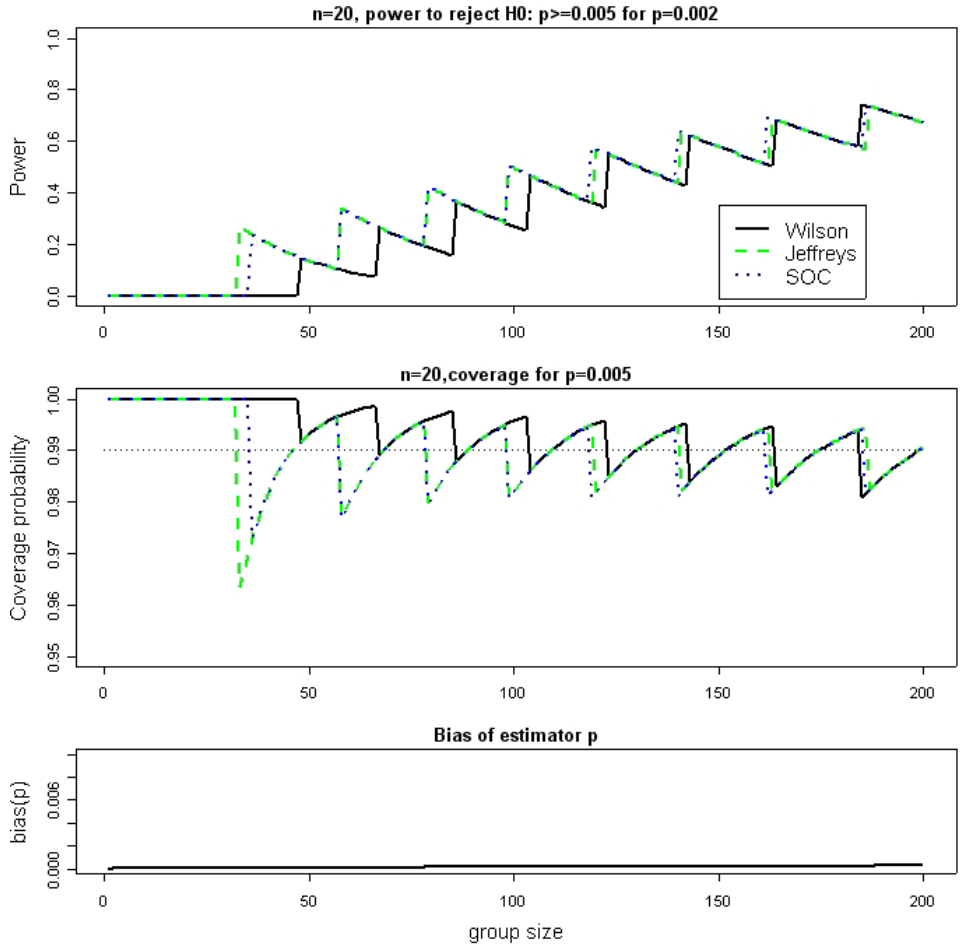

Figure 36: Upper 99%-confidence limits of Wilson, Second order corrected and Jeffreys prior interval applied in binomial group testing n=20, s=1,…,200
First plot: power to reject H0: $\pi \geq 0.005$ for π=0.002
Second plot: Coverage probability for π=0.005

8 Application

Example 1: Estimation of pathogen incidence in a natural vector population

Tedeschi et al. (2003) examined the role of psyllids (*Cacopsylla melaneura*, *Psyllidae*) for epidemiology of apple proliferation (AP). The phloem-sucking insect previously had been shown to transmit the phytoplasma causing AP. One objective was estimation of the proportion of psyllids infected with the phytoplasma. A group testing design with a group size *s*=5 was applied for this reason. The phytoplasma was detected in the single groups using an appropriate PCR assay. Although the design of their study is complicated, including different locations, repeated measurements, in two years and for different developmental stages of the vector, the authors finally sum over locations and time to estimate the general proportion of AP-infected vectors for the developmental stages.

For the over-wintered psyllids sampled in spring 2000, the following data were obtained:

Out of *n*=96 groups, each of size *s*=5, *Y*=16 groups were AP-positive.

This leads to the estimator $p = 1 - \left(1 - 16/96\right)^{\frac{1}{5}} = 0.0358$

The main objective here is estimation, not testing a special hypothesis, therefore we are interested in two-sided confidence intervals. The two-sided, 95% Clopper-Pearson-CI for the group scale estimator *t*=16/96 is

[0.0984 ; 0.2565]

which by application of $p = 1 - \left(1 - t\right)^{\frac{1}{5}}$ for transformation of the confidence bounds results in a CI for the proportion of infected psyllids:

[0.0205 ; 0.0576].

Thus the true proportion of AP infected psyllids in the over wintered natural population can be expected with 95% probability between 2.2 and 5.7 %.

For comparison, further two-sided 95% CI are given:

CI method	lower bound	upper bound	width of CI
Clopper-Pearson	0.0205	0.0576	0.0371
Blaker	0.0210	0.0567	0.0357
Agresti-Coull	0.0218	0.0571	0.0353
Wilson Score	0.0220	0.0568	0.0348
Wald	0.0186	0.0531	0.0345

The five methods do not differ to large extend in the position of their bounds or in the interval width. This is because they are based on a rather large data set containing 96 observations, where methods perform comparatively well.

Still they differ slightly: width of Clopper-Pearson is greater than that of the improved exact Blaker and the Score test derived Wilson and Agresti-Coull interval. The Wald interval is only slightly shorter than Wilson, but differs in position: the position of Wilson is similar to the exact intervals, while Wald is shifted toward 0, corresponding to its conservative lower bound and its very liberal upper bound.

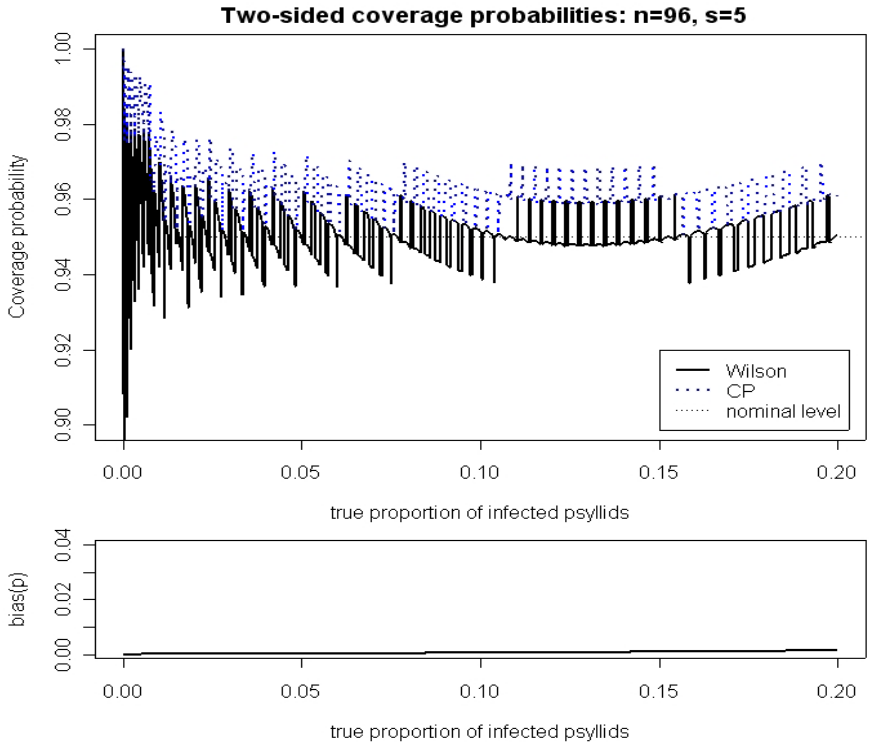

Figure 37: Coverage probability of two-sided 95% Wilson and Clopper Pearson CI and bias (p) for n=96 groups each of group size s=5 for π=0,…,0.2

Figure 37 shows that this design, although reducing the number of observation 5 times compared to simple binomial testing, is appropriate for estimation of a broad range of true infection rates. The true, unknown proportion of infected psyllids was varied from 0 to 20%. Over this range, bias was negligible. The Wilson CI performs as expected for the two-sided case: mean coverage seems close to the nominal level, and only shows some liberal spikes for proportions between 0 and 1%. The Clopper-Pearson CI is clearly conservative especially for proportions < 1%. If smaller proportions of infected psyllids shall be reliably estimated, either group size should be increased or, if this is not possible, more assays have to be performed.

Example 2: Vector transfer design

Tebbs and Bilder (2004) give an example of a vector transfer experiment performed by Ornaghi et al. (1999). 7 individuals of the planthopper *Delphacodes kuscheli* (*Macroptera*) from natural populations were placed on each of 24 healthy test plants, which were isolated from each other. After a sufficient time for infection and for establishing of the virus, each single test plant was evaluated for showing the disease or not. 3 test plants were found infected, 21 still were healthy.

The estimator of infection rate from this trial is $p = 1 - \left(1 - \frac{3}{24}\right)^{\frac{1}{7}} = 0.0189$

A two-sided 95% Wilson confidence interval for the transmission rate of a single plant hopper is [0.0063; 0.0516]. It can be concluded, that with only 5% probability, less than 0.63% or more than 5.16% of the individuals in the observed plant hopper population will transmit the virus to the plants. In other words, if the vector transfer design would be extended by including further test plants exposed to the planthoppers, 95% of the new estimators of transmission rates would range between 0.63 and 5.16 %.

Example 3: Resistance breeding: Estimation of the proportion of susceptible individuals in a breeding population

A certain pathogen resistance is inherited through a single dominant-recessive gene with the alleles R for resistance and s for susceptibility. The genotypes RR and Rs are resistant, ss is susceptible to the disease. Here also the

resistant Rs genotype inherits the susceptibility allele s to its offspring. A molecular marker exists, which allows classification of individuals as RR, Rs and ss, so populations or inbred lines for hybrid breeding can be selected for a low frequency of either Rs or ss genotypes. Objective of an experiment (Weissleder K, Fa KWS, personal communication) was the estimation of the low proportion of Rs + ss individuals. Since usually many lines or populations have to be evaluated in the breeding process and application of molecular markers is expensive, the costs for classification of a single line are limited. If group testing is applied for this problem, groups can be classified to contain either only R alleles, i.e. to contain only RR individuals, or contain only s alleles, i.e. to contain only ss individuals. If an assay detects both R and s, the group might contain any genotypes.

The following data show the results of group testing experiments on two populations. Groups each containing plant material of 5 individuals were classified using the molecular marker. The group size was limited to 5 because for larger group sizes the assay might not detect single Rs individuals in a group anymore. Larger group sizes might have lead to violation of assumption 4 and 5 to have sensitivity and specificity =1.
For the first population, out of 58 assays, 57 assays detected only R alleles, whereas in one group R and s alleles were detected and no group contained only s alleles.
For the second population, out of 60 assays, 1 detected only R alleles, 50 detected both R and s, and in 9 groups only ss genotypes were present.

Population	group size *s*	*n*	R only	R and s	s only
1	5	58	57	1	0
2	5	60	1	50	9

The primary aim is to calculate an estimator and confidence limits for the proportion of ‘not-RR’ individuals, that is Rs+ss. Groups somehow containing the s allele are counted as positive in the sense given in the notations. Of course it is also possible to estimate the proportion of ‘not-ss’ genotypes. But this is not of interest for this problem.

It has to be mentioned, that from this experiment, a CI for the proportion of RR-individuals cannot be calculated because of assumption 4. I.e. only traits can be defined as 'positive' in the sense of the notation given above, if a single positive individual in a group results in positive classification of the group. A single RR-individual can not be detected in a group containing any s-alleles: This group will be classified as containing R and s alleles and it can not be decided whether the members of the group are Rs, RR and Rs, RR and ss or further combinations.

Population 1:

Here, *Y=1* was found to be 'not-RR' among the *n*=58 groups under observation.

Estimator for proportion of not-RR individuals is $p = 1 - \left(1 - \frac{1}{58}\right)^{\frac{1}{5}} = 0.003472$

Since interest is only in small proportions of 'not-RR' individuals, one might calculate one-sided 95%-Clopper-Pearson CI with an upper bound, to take the uncertainty of estimating *p* into account. The simple binomial Clopper-Pearson CI for the group scale estimator *t*=1/58 is [0 ; 0.0792], transformation to the individual scale results in [0 ; 0.0164]. The breeder can conclude with 95% certainty, that population 1 does not contain more than 1.6% Rs and ss genotypes.

For illustration, further one-sided 95% CI are shown:

CI method	Lower bound	Upper bound
Clopper-Pearson	0	0.0164
Agresti-Coull	0	0.0164
Wilson Score	0	0.0152
Wald	0	0.0092

Population 2:

The same can be done for population 2: Among 60 groups, 59 groups were found to contain 'not-RR' individuals.

Estimator for proportion of 'not-RR' individuals is $p = 1 - \left(1 - \frac{59}{60}\right)^{\frac{1}{5}} = 0.55907$

The Clopper-Pearson CI then is [0; 0.7566], leading to the interpretation, that population 2 may contain until 75.7% Rs and ss genotypes with a probability of 95%. As obvious from the data and the estimator, the design was not appropriate for the underlying unknown proportion.

For comparison of the methods:

CI method	Lower bound	Upper bound
Clopper-Pearson	0	0.7566
Agresti-Coull	0	1
Wilson Score	0	0.6732
Wald	0	0.7029

The upper limit of the Agresti-Coull interval equals 1 because for *Y* close to or equal 1, the binomial Agresti-Coull interval on the group scale can have upper limits slightly greater than 1. This has to be corrected to 1 for transformation to the individual scale. This does not happen for Wilson or Clopper Pearson CI, which have upper bounds slightly lower than 1 for the same *Y* and *n*. Anyway, the estimation is very uncertain for all CI methods because of the probably not small proportion. This shows the limitation of group testing: the methods become insensitive if the proportion of the positive individuals becomes large.

Example 4a: Testing on GMO in an agricultural seed lot

In this experiment (Weissleder K, Fa KWS, personal communication) PCR was applied to test whether GMO were present in a seed lot. Objective is to show that the proportion of GMO in the seed lot is lower than 0.005:

$$H_0: \pi \geq 0.005 \quad \text{vs.} \quad H_1: \pi < 0.005$$

3 groups were tested, each containing *s*=3000 seeds. No group was tested GMO-positive.

In case that the assumptions come true, the following can be concluded:

The estimator of GMO in the sample is $p = 0$.

The upper Clopper-Pearson confidence limit for the proportion of GMO in the sample is [; 0.00033]. Since the upper bound is clearly smaller than the threshold 0.005 (the CI does not contain the threshold) it can be concluded that the proportion of CI is significantly smaller than 0.005 for an error probability of 5%.

For comparison, the corresponding p-value of an exact test (Hepworth, 1996) can be calculated: The probability to observe a Y favoring the alternative H_1 more than or equally as the observed Y=0 is $P(Y \leq 0 | n=3, s=3000) < 0.0001$.
From the confidence interval it can be concluded with 95% confidence that not more than 0.033% GMO are present in the seed lot. In other words, if further assays would be performed on the same seed lot, the new estimators of GMO content would lay in 95% of the cases within 0 and 0.033%.

Figure 38 shows the performance of this design for the range of π=0,…,0.005 which is of interest in GMO-testing.

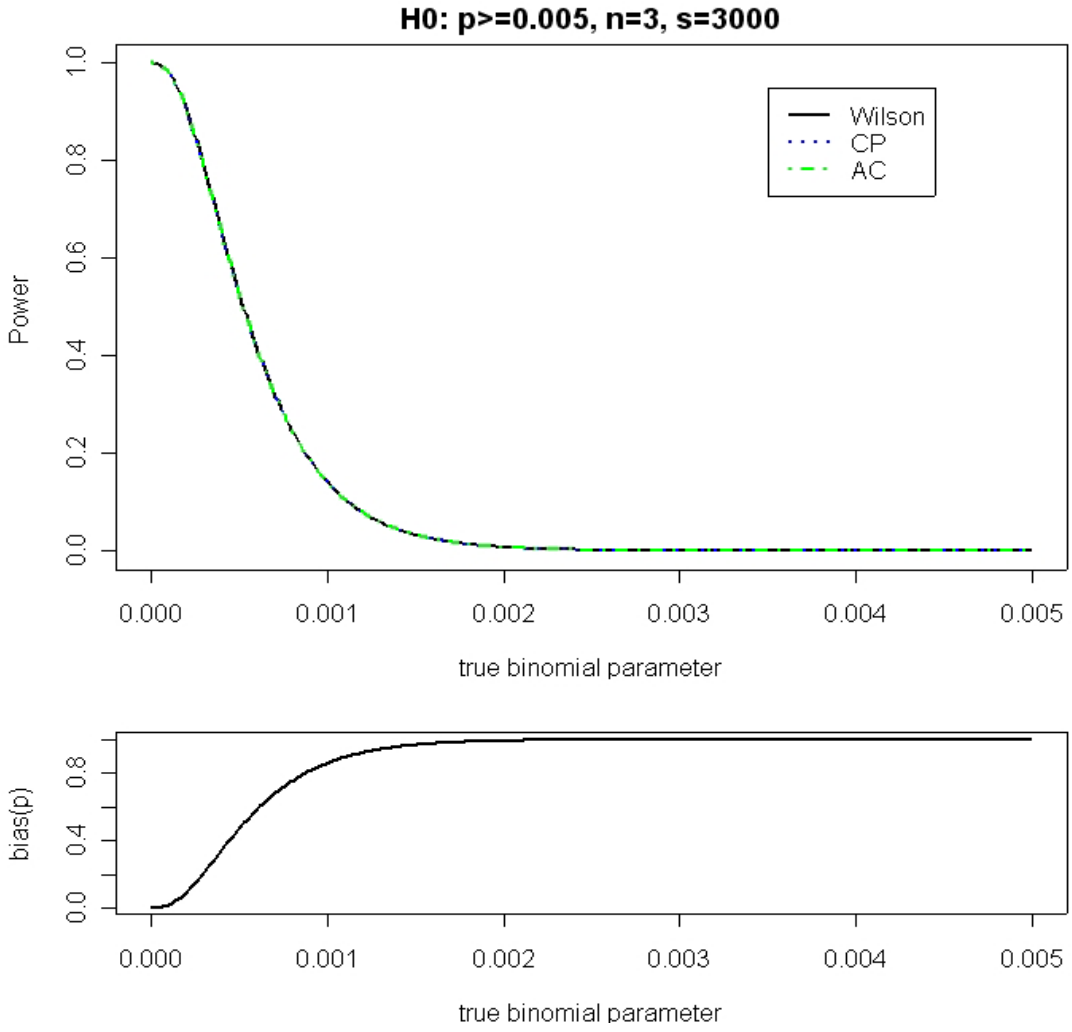

Figure 38: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wilson, Agresti-Coull and Clopper Pearson limits and bias (p) for n=3 groups each of group size s=3000 for π=0,…,0.005

This design is only appropriate for estimating very small incidences. To test against a hypothesis H_0: $\pi \geq 0.005$, it has a sufficient power only for very small

proportions (π<0.0003) of GMO. For proportions π > 0.0015, nearly every experiment will result in the outcome *Y*=3, and thus in the estimator *p*=1. This results in the large bias of the estimator and the inability to reject the null hypothesis. For this special design and hypothesis, all three methods have the same power depending on π.

Example 4b: A higher number of assays

In a second experiment the same methods were used to detect the same GMO in another seed lot, but a more assays were performed. Among 21 groups, each containing 3000 seeds, one group was found GMO-positive 20 groups were found GMO-negative.

The estimator = $1-(1-1/21)^{1/3000}$ = 0.000017 and a 95% one-sided Clopper-Pearson confidence interval is [0 ; 0.000077]. Here again the question arises, whether this experimental design was appropriate to test the hypothesis H_0: $\pi \geq$ 0.005. In the following graph shows the power of the design *n*=21, *s*=3000 in dependence for true GMO contaminations π = 0,…, 0.005.

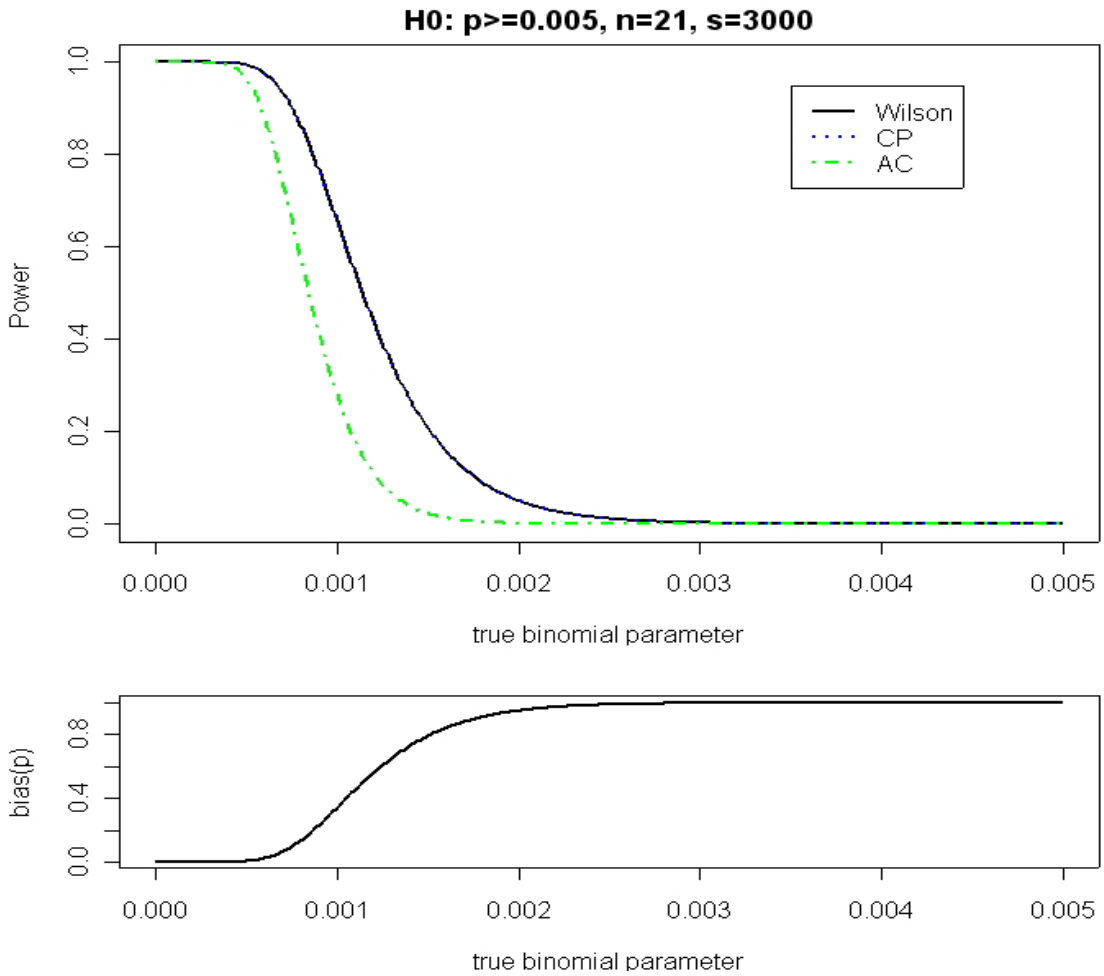

Figure 39: Power to reject H0: $\pi \geq 0.005$ using upper 95% Wilson, Agresti-Coull and Clopper-Pearson limits and bias (p) for n=21 groups each of group size s=3000 for π=0,…,0.005

As obvious from figure 39, the design is very sensitive to detect contaminations between lower than π= 0.00075. As bias increases for higher values of π, power decreases and the design becomes inappropriate for π > 0.0015. Obviously, the group size s=3000 is too high if GMO-proportions become too high although being still lower than the threshold 0.005.

If the experimenter f.e. also wants a sufficient power to reject H_0: $\pi \geq 0.005$ if the true GMO contamination in the seed lot goes until π = 0.003, smaller group sizes are optimal: Figure 40 shows the power of Wilson, Clopper-Pearson and Agresti-Coull CIs upper 95% limits for increasing group size *s*=1,…, 3000 if the true GMO contamination of the seed lot is π = 0.003.

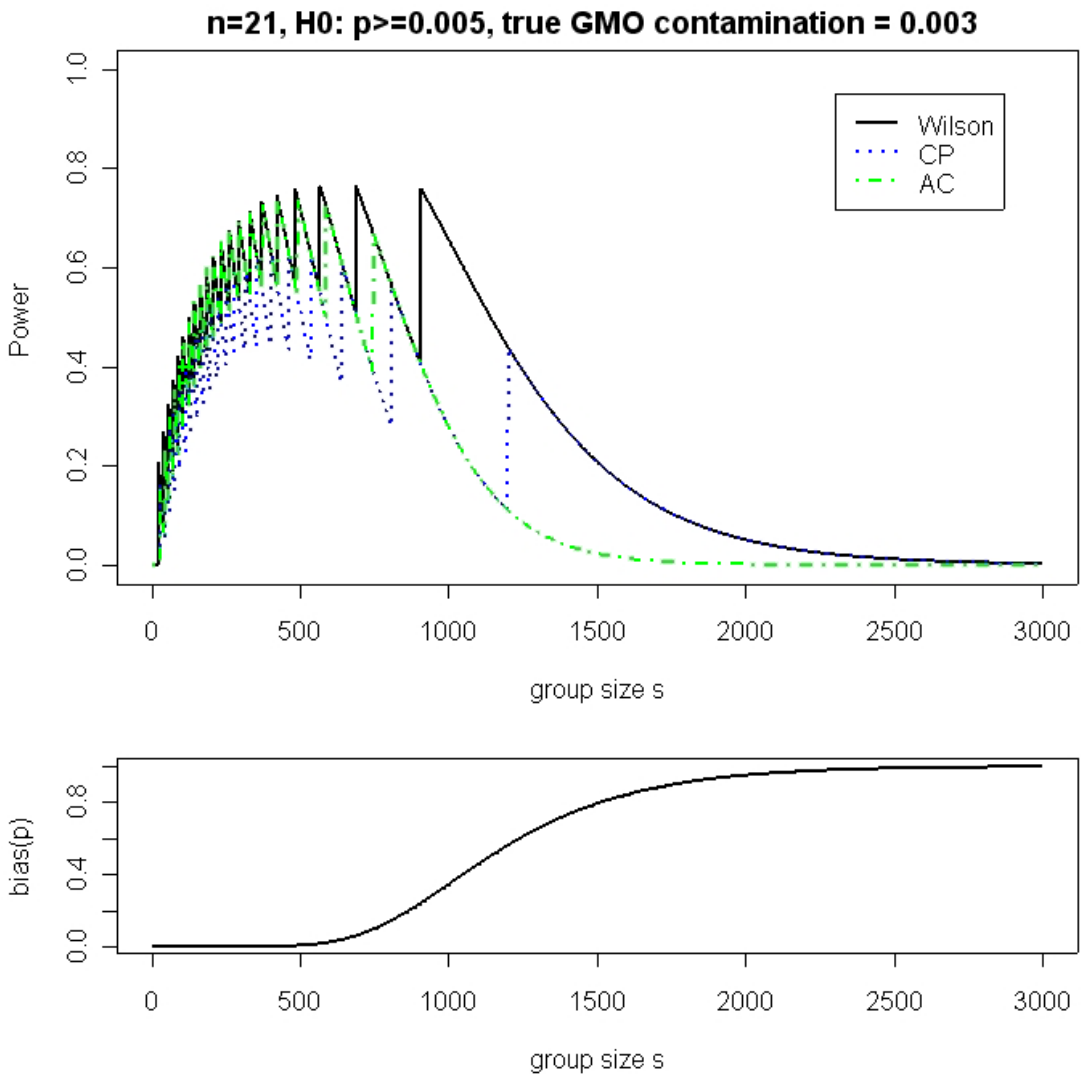

Figure 40: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wilson, Agresti-Coull and Clopper Pearson limits and bias (p) for n=21 groups and $\pi = 0.003$ for s=1,…,3000

Obviously the maximal power for this sample size *n*=21 can be achieved with group sizes between 400 and 500. Further increasing the group size results in high bias of the estimator p and decreasing power. But how does a design of *n*=21 and *s*=500 perform for other values of π=0,…,0.005 ? Power and bias(*p*) for this design are shown in figure 40.

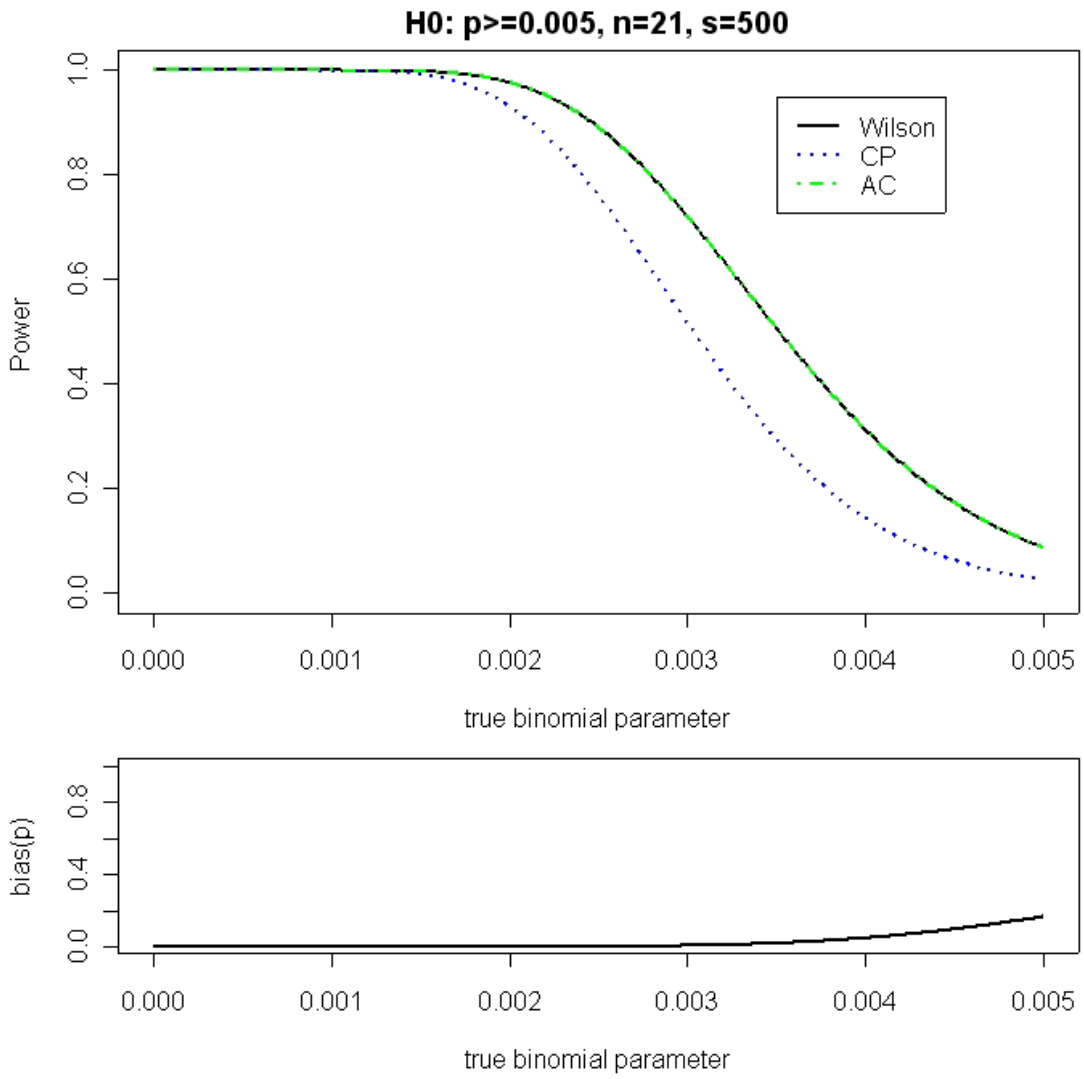

Figure 41: Power to reject H_0: $\pi \geq 0.005$ using upper 95% Wilson, Agresti-Coull and Clopper-Pearson limits and bias (p) for n=21 groups each of group size s=500 for π=0,…,0.005

For very small π, the design has the same high power as if group size *s*=3000 would have been used, but using *s*=500 also GMO contaminations of 0.3 % can still be shown with a high probability to be significantly lower than 0.5%. This again stresses the importance of careful experimental design including calculation of bias and power for the expected range of π.

9 General discussion and prospect

The usefulness of the group testing approach for estimation of small binomial proportions in case of limited assay number is described in the references (Thompson 1962, Swallow 1985, Tebbs and Bilder 2004). In this thesis, it was

shown that it can also greatly improve the power in a proof of safety, if very small proportions are considered as threshold proportions for 'unsafety' and the number of observations is limited. As mentioned already in early publications (Thompson 1962, Swallow 1987) the bias of the estimator beside others is a critical parameter to judge the goodness of a group testing design. If the type-I-error α shall be controlled in a strict and conservative way, exact intervals as the Blaker interval are recommended for the two-sided case (Tebbs and Bilder 2004, Reiczigel, 2003, Blaker, 2000) and the Clopper-Pearson interval has been shown to be appropriate for the one-sided case in this thesis.
If slightly liberal performance is acceptable, Wilson Score interval can be recommended for the two-sided (Tebbs and Bilder, 2004) and here was also found to be acceptable for the one-sided case. The recently proposed second-order corrected confidence interval (Cai, 2005) seems to be a clear improvement of the known asymptotic methods because of its symmetric coverage probability, and only moderate violation of the nominal level over the whole range of the binomial proportion. Especially if higher confidence levels (f.e. 99%) are required or if the methods are applied in group testing with the risk of choosing the group size inappropriately high, this new method is recommended instead of the Wilson interval. Because of its more complicated calculation it needs to be implemented in a software package.

One important area of application of a proof of safety using group testing is testing for GMO contamination. An alternative approach, recently proposed by the EU commission (Anonymous 2004), is the quantitative measurement of contamination in a sample. Even if quantitative methods are used for characterization, group testing might be applied after dichotomization of the continuous outcomes using a cut point (Xie et al. 2001).
Another application for group testing is epidemiology of plant, animal or human diseases. Beside the simple estimation of proportions, the comparison of two or k proportions is of high interest. Here several open problems still exist: Beside a Wald type confidence interval for the difference of two proportions proposed by Swallow (1985), no other proposals or even discussions of the performance of this method were found in the references, although CI for difference of proportions estimated from simple binomial testing are extensively discussed

(f.e. Newcombe 1998, Agresti and Caffo 2000, Zhou et al. 2004). This certainly is a problem which needs further examination. Although the two-sample comparison seems to be not solved for group testing, regression methods or methods for ordered binomial proportions for k simple binomial proportion can be transferred to the group testing approach (Tebbs and Swallow 2003a,b, Xie 2001, Hung and Swallow 2000).

10 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.

ANONYMOUS (2004): Commission Recommendation on technical guidance for sampling and detection of genetically modified organisms and materials produced from genetically modified organisms as or in products in the context of regulation (EC) 1830 (2003). Official Journal of the European Union L348.

AGRESTI A and CAFFO B (2000): Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. The American Statistician, November 2000, 54 (4): 280-288.

AGRESTI A and COULL BA (1998): Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician, May 1998, 52 (2): 119 -126.

AGRESTI A and MIN Y (2001): On small-sample confidence intervals for parameters in discrete distributions. Biometrics 57, 963-971.

BLAKER H (2000): Confidence curves and improved exact confidence intervals for discrete distributions. The Canadian Journal of Statistics 28 (4): 783-798.

BLYTH C and STILL H (1983): Binomial confidence intervals. Journal of the American Statistical Association 78, 108-116.

BOCK J (1998): Die Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte klinische Studien. Oldenbourg Verlag.

BROWN LD, CAI TT, DASGUPTA A (2001a): Interval estimation for a binomial proportion. Statistical Science 2001, Vol.16, No.2, 101-128.

BROWN LD, CAI TT, DASGUPTA A (2001b): Rejoinder: Interval estimation for a binomial proportion. Statistical Science 2001, Vol.16, No.2, 128-133.

BROWN LD, CAI TT, DASGUPTA A (2002): Confidence intervals for a binomial proportion and asymptotic expansions. The Annals of Statistics 2002, 30 (1), 160-201.

CAI, TT (2005): One-sided confidence intervals in discrete distributions. Journal of Statistical Planning and Inference 131: 63-88.

CASELLA G (1986): Refining binomial confidence intervals. Canadian Journal of Statistics 14, 113-129.

CASELLA G (2001): Comment: Interval estimation for a binomial proportion. Statistical Science 2001, Vol.16, No.2, 120-122.

EFRON B, TIBSHIRANI RJ (1993): An introduction to the bootstrap. Monographs on statistics and applied probability 57. Chapman and Hall New York.

GERLING D (2002): Biostatistische Methoden zur Schätzung des Stichprobenumfangs bei der Qualitätskontrolle von Saatgut. Diplomarbeit am Lehrgebiet Bioinformatik, Fachbereich Gartenbau, Universität Hannover.

GIOVANNINI T, CONCILIO L (2002): PCR detection of genetically modified organisms: A review. Starch/Stärke 54: 321-327.

GU W, LAMPMAN R, NOVAK RJ (2004): Assessment of arbovirus vector infection rates using variable size pooling. Medical and Veterinary Entomology 18, 200-204.

HEPWORTH G (1996): Exact confidence intervals for proportions estimated by group testing. Biometrics 52, 1134-1146.

HEPWORTH G (2004): Mid-p confidence intervals based on the likelihood ratio for proportions estimated by group testing. Aust. N. Z. J. Stat. 46(3), 391-405.

HOLST-JENSEN A, RONNING SB, LOVSETH A, BERDAL KG (2003): PCR technology for screening and quantification of genetically modified organisms (GMOs). Analytical and Bioanalytical Chemistry 375: 985-993.

HUNG MC AND SWALLOW WH (1999): Robustness of group testing in the estimation of proportions. Biometrics 55, 231-237.

HUNG MC AND SWALLOW WH (2000): Use of group testing in tests of hypothesis for classification or quantitative covariables. Biometrics 56, 204-212.

JANKIEWICZ A, BROLL H, ZAGON J (1999): The official method for the detection of genetically modified soybeans (German Food Act LMBG § 35): a semi-quantitative study of sensitivity limits with glyphosate-tolerant soybeans (Roundup Ready) and insect-resistant Bt maize (Maximizer). European Food Research and Technology 209 (2): 77-82.

KOTZ S AND JOHNSON NL (ed.) (1985): Encyclopedia of statistical sciences. John Wiley and Sons New York Chichester Brisbane Toronto Singapore.

NEWCOMBE RG (1998): Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine 17: 873-890.

ORNAGHI J, MARCH G, BOITO G, MARINELLI A, BEVIACQUA J, GIUGGIA J, LENARDON S (1999): Infectivity in natural populations of Delphacodes kuscheli vector of 'Mal Rio Cuarto' Virus. Maydica, 44, 219-223.

PIEGORSCH WW (2004): Sample sizes for improved binomial confidence intervals. Computational Statistics and Data Analysis 46, 309-316.

REICZIGEL J (2003): Confidence intervals for the binomial parameter: some new considerations. Statistics in Medicine 22: 611-621.

REMUND KM, DIXON DA, WRIGHT DL, HOLDEN LR (2001). Statistical considerations on seed purity testing on transgenic traits. Seed Sci Res 11: 101-119.

SACHS L (1991): Angewandte Statistik, 7. edition. Springer-Verlag. Berlin Heidelberg New York.

SANTNER TJ, DUFFY DE (1989): The statistical analysis of discrete data. Springer Verlag New York Berlin Heidelberg.

STERNE TE (1954): Some remarks on confidence or fiducial limits. Biometrika 41: 275-278.

SWALLOW WH, 1985: Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology Vol.75, N.8, 882-889.

TEBBS JM & BILDER CR, 2004: Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. Journal of Agricultural, Biological and Environmental Statistics, Vol.9, N.1, 75-90.

TEBBS JM & SWALLOW WH, 2003a: More powerful likelihood ratio tests for isotonic binomial proportions. Biometrical Journal 45 (5): 618-630.

TEBBS JM & SWALLOW WH, 2003b: Estimating ordered binomial proportions with the use of group testing. Biometrika 90 (2): 471-477.

TEDESCHI R, VISENTIN C, ALMA A, BOSCO D (2003): Epidemiology of apple proliferation (AP) in northwestern Italy: evaluation of the frequency of AP-

positive psyllids in naturally infected populations of *Cacopsylla melanoneura* (*Homoptera: Psyllidae*). Annals of applied Biology 142: 285-290.

THOMPSON KH (1962): Estimation of the proportion of vectors in a natural population of insects. Biometrics 18: 568-578.

WALCOTT RR (2003): Detection of seedborn pathogens. HortTechnology 13 (1) : 40-47.

WILSON EB (1927): Probable inference, the law of succession, and statistical inference. Journal of the American statistical association 22: 209-212.

XIE M (2001): Regression analysis of group testing samples. Statistics in Medicine 20: 1957-1969.

XIE M, TATSUOKA K, SACKS J, YOUNG S (2001): Group testing with blockers and synergism. Journal of the American Statistical Association 96, 92-102.

ZHOU XH, TSAO M, QIN G (2004): New intervals for the difference between two independent binomial proportions. Journal of Statistical Planning and Inference 123: 97-115.

11 Annex: R code

NOTATION IN THE R-CODE

n	number of assays or groups or observations
s	number of units in the groups
Y	number of positive observations, i.e. observed number of positive groups (group testing) or positive individuals (simple binomial)
p.tr	true proportion of positive individual units in the population
p.hyp	threshold proportion in the hypothesis
conf.level	1-alpha
alternative	direction of the alternative hypothesis:

"less" means p.tr < p.hyp only upper bound with each 1-α
"greater" means p.tr > p.hyp only lower bound with each 1-α
"two.sided" means p.tr != p.hyp upper and lower bound with each 1-α/2
interval bounds greater 1 or smaller 0 are set 1 or 0 !! otherwise: NaN after transformation to the group scale for t>1

11.1 R code for binomial group testing

The examples for usage of the functions and the plots can be carried out by transferring the Rcode given below (marked blue) to the GUI of R, f.e. by 'copy and paste'. No packages additional to the R version 2.0.1 needed.

1) simple functions for the calculations of the single interval types are given
2) the basic functions for calculation of power, coverage and bias are given, which can easily be modified to calculate interval length, etc. ; these need the functions under 1) present in the R working space
3) gives functions to vary the design parameters n, s, p.tr, p.hyp in a standardized manner, which can easily be used in plots including one example of a plot; these functions need the functions under 1) and 2) to be present in the working space
4) gives short functions for approximate power and sample size calculation for the group testing Score test and Wilson CI
5) gives a function for simulation of coverage in the Wilson, AgrestiCoull, Wald, add-4 and Clopper-Pearson CI, only for comparison with closed calculation; needs the functions under 1) to be present in the working space

```
############################################################
#                                                          #
#  1) CONFIDENCE INTERVALS FOR BINOMIAL GROUP TESTING      #
#                                                          #
############################################################

############################
#### wilson CI for bgt ####
############################

bgt.wilson<-function(n,Y,s,conf.level=0.95,alternative="two.sided")
{
alpha=1-conf.level
th=Y/n
est.int=(Y+(qnorm(1-alpha/2)^2)/2)/(n+(qnorm(1-alpha/2))^2)
est.int1s=(Y+(qnorm(1-alpha)^2)/2)/(n+(qnorm(1-alpha))^2)

if(alternative =="two.sided"){
    w.se=((qnorm(1-alpha/2))*sqrt(n*th*(1-th)+(qnorm(1-alpha/2)^2)/4))/(n+qnorm(1-
alpha/2)^2)
    KI.int.l=est.int-w.se
    KI.int.u=est.int+w.se
```

```
        if (KI.int.u>1){KI.int.u=1}
        if (KI.int.l<0){KI.int.l=0}
    KI=c( 1-(1-KI.int.l)^(1/s), 1-(1-KI.int.u)^(1/s) )
}
else{if(alternative=="less"){
    w.se=((qnorm(1-alpha))*sqrt(n*th*(1-th)+(qnorm(1-alpha)^2)/4))/(n+qnorm(1-alpha)^2)
    KI.int.u=est.int1s+w.se
        if (KI.int.u>1){KI.int.u=1}
    KI=c( 0, 1-(1-KI.int.u)^(1/s) )
}
else{if(alternative=="greater"){
    w.se=((qnorm(1-alpha))*sqrt(n*th*(1-th)+(qnorm(1-alpha)^2)/4))/(n+qnorm(1-alpha)^2)
    KI.int.l=est.int1s-w.se
        if (KI.int.l<0){KI.int.l=0}
    KI=c( 1-(1-KI.int.l)^(1/s), 1 )
}
else{stop("argument alternative misspecified")}}}
estimate = 1-(1-th)^(1/s)
list(conf.int = KI,estimate=estimate) }


# # # Examples for usage:
# Compare Tebbs and Bilder(2004),p.86
bgt.wilson(n=24,Y=3,s=7)
bgt.wilson(n=24,Y=3,s=7,conf.level=0.95,alternative="two.sided")
# one.sided
bgt.wilson(n=24,Y=3,s=7,conf.level=0.975,alternative="less")

################################
#### Agresti-Coull-CI for bgt ###
################################

bgt.AC<-function(n, Y, s, conf.level=0.95, alternative="two.sided") {
alpha=1-conf.level
est.int=(Y+(qnorm(1-alpha/2)^2)/2)/(n+(qnorm(1-alpha/2))^2)
est.int1s=(Y+(qnorm(1-alpha)^2)/2)/(n+(qnorm(1-alpha))^2)

if(alternative =="two.sided"){
        AC.se=(qnorm(1-alpha/2))*sqrt((est.int*(1-est.int))/(n+(qnorm(1-alpha/2))^2))
        KI.int.l=est.int-AC.se
        KI.int.u=est.int+AC.se
         if (KI.int.u>1){KI.int.u=1}
         if (KI.int.l<0){KI.int.l=0}
        KI=c(1-(1-KI.int.l)^(1/s),1-(1-KI.int.u)^(1/s))
}
else{if(alternative=="less"){
        AC.se=(qnorm(1-alpha))*sqrt((est.int1s*(1-est.int1s))/(n+(qnorm(1-alpha))^2))
        KI.int.u=est.int1s+AC.se
         if (KI.int.u>1){KI.int.u=1}
        KI=c(0, 1-(1-KI.int.u)^(1/s))
}
else{if(alternative=="greater"){
        AC.se=(qnorm(1-alpha))*sqrt((est.int1s*(1-est.int1s))/(n+(qnorm(1-alpha))^2))
        KI.int.l=est.int1s-AC.se
         if (KI.int.l<0){KI.int.l=0}
        KI=c(1-(1-KI.int.l)^(1/s),1)
}
else{stop("argument alternative misspecified")}}}

estimate=1-(1-Y/n)^(1/s)
list( conf.int=KI,estimate=estimate )
}

# # # Example for usage:
bgt.AC(n=24,Y=3,s=7)

#############################################
# Add-4 for binomial group testing          #
# for alpha=0.05 and two-sided only !       #
#############################################

bgt.add4<-function(n, Y, s, conf.level=0.95, alternative="two.sided") {
alpha=1-conf.level
t=Y/n
est.int=(Y+2)/(n+4)

if(alternative =="two.sided"){
```

```
        add.se=(qnorm(1-alpha/2))*sqrt((est.int*(1-est.int))/(n+4))
        KI.int.l=est.int-add.se
        KI.int.u=est.int+add.se
         if (KI.int.u>1){KI.int.u=1}
         if (KI.int.l<0){KI.int.l=0}
        KI=c(1-(1-KI.int.l)^(1/s),1-(1-KI.int.u)^(1/s))
}

else{if(alternative=="less"){
        add.se=(qnorm(1-alpha))*sqrt((est.int*(1-est.int))/(n+4))
        KI.int.u=est.int+add.se
         if (KI.int.u>1){KI.int.u=1}
        KI=c(0, 1-(1-KI.int.u)^(1/s))
}

else{if(alternative=="greater"){
        add.se=(qnorm(1-alpha))*sqrt((est.int*(1-est.int))/(n+4))
        KI.int.l=est.int-add.se
         if (KI.int.l<0){KI.int.l=0}
        KI=c(1-(1-KI.int.l)^(1/s),1)
}
else{stop("argument alternative misspecified")}}}

estimate = 1-(1-t)^(1/s)
list( conf.int=KI,estimate=estimate )
}
bgt.add4(n=24,Y=3,s=7)


##############################
### Wald Intervall for bgt ###
##############################

bgt.wald<- function (n, Y, s, conf.level=0.95, alternative="two.sided")

{
if(Y>n) {stop("number of positive tests Y can not be greater than number of tests n")}
th=Y/n
esti=1-(1-th)^(1/s)
var.esti=(1-(1-esti)^s)/(n*(s^2)*(1-esti)^(s-2))
alpha=1-conf.level

if(alternative=="two.sided"){
    snquant=qnorm(p=1-alpha/2,mean=0,sd=1,lower.tail=TRUE)
    KI=c(esti-snquant*sqrt(var.esti),esti+snquant*sqrt(var.esti))
}
else{if (alternative=="less"){
    snquant=qnorm(p=1-alpha,mean=0,sd=1,lower.tail=TRUE)
    KI=c(-Inf,esti+snquant*sqrt(var.esti))
}
else {if (alternative=="greater"){
    snquant=qnorm(p=1-alpha,mean=0,sd=1,lower.tail=TRUE)
    KI=c(esti-snquant*sqrt(var.esti),Inf)
}
else {stop("argument alternative mis-specified")}}}
list( conf.int=KI,estimate=esti )
}

# # # Example for usage:
# Compare Tebbs and Bilder(2004),p.86
bgt.wald(n=24,Y=3,s=7)

######################
#  Clopper-Pearson  #
######################

bgt.CP<-function(n,s,Y, conf.level=0.95, alternative="two.sided")
{
lower<-0
upper<-1
if(alternative=="two.sided")
{
 if(Y!=0)
   {lower<-qbeta((1-conf.level)/2, Y, n-Y+1)}

 if(Y!=n)
   {upper<-qbeta(1-(1-conf.level)/2, Y+1, n-Y)}
```

```
}

if(alternative=="less")
{
 if(Y!=n)
   {upper<-qbeta(1-(1-conf.level), Y+1, n-Y)}
}

if(alternative=="greater")
{
 if(Y!=0)
   {lower<-qbeta((1-conf.level), Y, n-Y+1)}
}

estimate=1-(1-Y/n)^(1/s)
KI=c(1-(1-lower)^(1/s),1-(1-upper)^(1/s))

list(conf.int=KI, estimate=estimate)
}

# # # example for usage
# Compare Tebbs and Bilder(2004),p.86
bgt.CP(n=24,s=7, Y=3)

############################
# Second Order Corrected #
############################

bgt.SOC<-function(n,Y,s,conf.level=0.95,alternative="two.sided")
{
esti<-Y/n
kappa<-qnorm(conf.level)
eta<-(kappa^2)/3 + 1/6
gamma1<-((13/18)*kappa^2 + 17/18)*(-1)
gamma2<-(kappa^2)/18 + 7/36

midpo<-(Y+eta)/(n+2*eta)

if(alternative=="less")
  {upper = midpo + kappa * sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) +
gamma2)/n)/sqrt(n)
CI=c( 0 ,upper)
  if(Y==n||upper>1){CI=c(0,1)}
  else{ CI=c( 0 ,upper)}
 }

if(alternative=="greater")
  {CI=c( midpo - kappa*sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) + gamma2)/n)/sqrt(n) ,
1)
  if(Y==0){CI=c(0,1)} }

if (alternative=="two.sided")
{
kappa<-qnorm(1-(1-conf.level)/2)
eta<-(kappa^2)/3 + 1/6
gamma1<-((13/18)*kappa^2 + 17/18)*(-1)
gamma2<-(kappa^2)/18 + 7/36

lower= midpo - kappa*sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) + gamma2)/n)/sqrt(n)
upper= midpo + kappa*sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) + gamma2)/n)/sqrt(n)

if(Y==0){CI=c(0,upper)}
else{if(Y==n||upper>1){CI=c(lower,1)}
else{CI=c(lower, upper)}}
}

estimate=1-(1-Y/n)^(1/s)
conf.int=c( 1-(1-CI[1])^(1/s) ,  1-(1-CI[2])^(1/s))
list(conf.int=conf.int, estimate=estimate)
}

# usage:
bgt.SOC(Y=3, n=24, s=7, alternative ="less")

##################
# Jeffreys Prior #
##################
```

```
bgt.Jef<- function(n,Y,s,conf.level=0.95,alternative="two.sided")

{
if(alternative=="less")
 {CI=c( 0 , qbeta(p=conf.level, shape1= Y+0.5, shape2=n-Y+0.5) )}

if(alternative=="greater")
 {CI=c( qbeta(p=1-conf.level, shape1= Y+0.5, shape2=n-Y+0.5) , 1 )}

if (alternative=="two.sided")
 {CI=c( qbeta(p=(1-conf.level)/2, shape1=Y+0.5, shape2=n-Y+0.5) , qbeta(p=1-(1-
conf.level)/2, shape1=Y+0.5, shape2=n-Y+0.5) )}

estimate=1-(1-Y/n)^(1/s)
conf.int=c( 1-(1-CI[1])^(1/s) ,  1-(1-CI[2])^(1/s))
list(conf.int=conf.int, estimate=estimate)
}

# # # usage:
bgt.Jef(Y=3, n=24, s=7)

############################################################
#                                                          #
# 2) Closed calculation of power and coverage probabilities #
# for the group testing CI methods given above             #
#                                                          #
############################################################

# Indicator functions:

#############################################################################
# i)Power-calculation                                                       #
#                                                                           #
# P.ind calculates, whether a method rejects H0 for a given setting of n,Y,s,alpha #
#                                                                           #
#############################################################################

P.Ind<-function(n,Y,s,p.hyp,conf.level,method,alternative){

if(method=="wilson"){
        KI.wilson<-bgt.wilson(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.wilson[[1]]>=p.hyp||KI.wilson[[2]]<=p.hyp){dec=1}
        else{dec=0}
        dec}

else{if(method=="AC"){
        KI.AC<-bgt.AC(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.AC[[1]]>=p.hyp||KI.AC[[2]]<=p.hyp){dec=1}
        else{dec=0}
        dec}

else{if(method=="wald"){
        KI.wald<-bgt.wald(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.wald[[1]]>=p.hyp||KI.wald[[2]]<=p.hyp){dec=1}
        else{dec=0}
        dec}

else{if(method=="CP"){
        KI.CP<-bgt.CP(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.CP[[1]]>=p.hyp||KI.CP[[2]]<=p.hyp){dec=1}
        else{dec=0}
        dec}

else{if(method=="SOC"){
        KI.SOC<-bgt.SOC(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.SOC[[1]]>=p.hyp||KI.SOC[[2]]<=p.hyp){dec=1}
        else{dec=0}
        dec}

else{if(method=="Jef"){
```

```
        KI.Jef<-bgt.Jef(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.Jef[[1]]>=p.hyp||KI.Jef[[2]]<=p.hyp){dec=1}
        else{dec=0}
        dec}

else{stop("argument method mis-specified")}}}}}}}
}

#############################################################################
#                                                                           #
# ii) Calculation of coverage probability                                   #
# indicator function C.Ind calculates whether the CI contains the true      #
# value for a given set of Y,n,s,                                           #
#                                                                           #
#############################################################################

C.Ind<-function(n,Y,s,p.tr,conf.level,method,alternative){

if(method=="wilson"){
        KI.wilson<-bgt.wilson(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.wilson[[1]]<=p.tr && KI.wilson[[2]]>=p.tr){cov=1}
        else{cov=0}
        cov}

else{if(method=="AC"){
        KI.AC<-bgt.AC(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.AC[[1]]<=p.tr && KI.AC[[2]]>=p.tr){cov=1}
        else{cov=0}
        cov}

else{if(method=="wald"){
        KI.wald<-bgt.wald(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.wald[[1]]<=p.tr && KI.wald[[2]]>=p.tr){cov=1}
        else{cov=0}
        cov}

else{if(method=="CP"){
        KI.CP<-bgt.CP(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.CP[[1]]<=p.tr && KI.CP[[2]]>=p.tr){cov=1}
        else{cov=0}
        cov}

else{if(method=="SOC"){
        KI.SOC<-bgt.SOC(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.SOC[[1]]<=p.tr && KI.SOC[[2]]>=p.tr){cov=1}
        else{cov=0}
        cov}

else{if(method=="Jef"){
        KI.Jef<-bgt.Jef(n=n,Y=Y,s=s,conf.level=conf.level,
alternative=alternative)$conf.int
        if(KI.Jef[[1]]<=p.tr && KI.Jef[[2]]>=p.tr){cov=1}
        else{cov=0}
        cov}

else{stop("argument method mis-specified")}}}}}}}
}

#########################################################
#                                                       #
# iii) Calculation of the probability of realisation    #
# of a certain Y for a given set of n,s,p.tr            #
#                                                       #
#########################################################

CI.prob<-function(n,Y,s,p.tr){
CI.pr<-(choose(n,Y)) * ((1-(1-p.tr)^s)^Y) * (1-p.tr)^(s*(n-Y))
CI.pr
}

# can be used until n=1029, Y=514
```

```
# Since numbers greater or equal choose(1030,515) cannot be represented in R anymore
# (i.e. 1e+309=Inf), and numbers x smaller than x=1e-323 are represented by 0, thus
# result in log(x)=-Inf, the following function might be used instead of CI.prob(),
# here lchoose() calculates the natural logarithm of the binomial coefficient

CI.prob2<-function(n,Y,s,p.tr){
CI.pr<-exp( lchoose(n,Y) + Y*log(1-(1-p.tr)^s) + s*(n-Y)*log(1-p.tr) )
CI.pr
}
#######################################
#                                     #
#  iv) SYNTHESIS of i), ii) and iii)  #
#                                     #
#######################################

#########################################################################
#  Basic function 1 : power and bias of a certain design and CI method  #
#########################################################################

bgt.power<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided"){

pow.ex=0
expected=0
for(Y in 0:n)
  {
  temp=CI.prob(n=n,Y=Y,s=s,p.tr=p.tr)
  pow.ex  <-  pow.ex+(P.Ind(n=n,Y=Y,s=s,p.hyp=p.hyp,conf.level=conf.level,method=method,
alternative = alternative) * temp)
  expected=expected+(1-(1-Y/n)^(1/s))*temp
  }
bias=expected-p.tr

list(power=pow.ex,
bias=bias)
}
# usage:
bgt.power(n=25,s=18, p.tr=0.05, p.hyp=0.08)

# Bias: Compare Swallow, 1985, table 1: N=25, k=18, p=0.05

#########################################################################
#  Basic function 2 : coverage and bias of a certain design and CI method  #
#########################################################################

bgt.cover<-function(n,    s,    p.tr,    p.hyp,    conf.level=0.95,    method="wilson",
alternative="two.sided"){

cov.ex=0
expected=0
for(Y in 0:n)
  {
  temp=CI.prob(n=n,Y=Y,s=s,p.tr=p.tr)
  cov.ex   <-   cov.ex+(C.Ind(n=n,Y=Y,s=s,p.tr=p.tr,conf.level=conf.level,method=method,
alternative = alternative) * temp)
  expected=expected+(1-(1-Y/n)^(1/s))*temp
  }
bias=expected-p.tr

list(cover=cov.ex,
bias=bias)
}

# usage:
bgt.cover(n=25,s=18, p.tr=0.05, p.hyp=0.08)

#####################################################################
#                                                                   #
# 3) VECTORIZATIONS of the two Basic functions for variation        #
# of the different parameters and a defined output in               #
# vectors for plotting of the graphs                                #
#                                                                   #
#####################################################################

#############################################################
# Vary the group size s, other parameters fixed             #
# Input of s as a vector                                    #
```

```
############################################################

# Power

bgts.power<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
pow.s<-numeric(length=length(s))
bias.s<-numeric(length=length(s))

 for(i in 1:length(s))
  {
  temp<-bgt.power(n=n,s=s[i],p.tr=p.tr,p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  pow.s[i]<-temp$power
  bias.s[i]<-temp$bias
  }
list(group.size=s,power=pow.s,bias=bias.s)
}

# usage:
test<-bgts.power(n=20, s=1:100, p.tr=0.02,p.hyp=0.03, alternative="less",
method="wilson")
# a simple plot
layout(mat=matrix(1:2, ncol=1))
plot(x=test$group.size, y=test$power, type="l")
plot(x=test$group.size, y=test$bias, type="l")

# Coverage

bgts.cover<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
cov.s<-numeric(length=length(s))
bias.s<-numeric(length=length(s))

 for(i in 1:length(s))
  {
  temp<-bgt.cover(n=n,s=s[i],p.tr=p.tr,p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  cov.s[i]<-temp$cover
  bias.s[i]<-temp$bias
  }

list(group.size=s,cover=cov.s,bias=bias.s)
}
#usage:
test1<-bgts.cover(n=20, s=1:100, p.tr=0.02,p.hyp=0.03,method="wilson",
alternative="less")
# a simple plot
layout(mat=matrix(1:2, ncol=1))
plot(x=test1$group.size, y=test1$cover, type="l")
plot(x=test1$group.size, y=test1$bias, type="l")

############################################################
# Vary the number of assays n, other parameters fixed      #
# Input of n as a vector                                   #
############################################################

# Power

bgtn.power<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{

pow.n<-numeric(length=length(n))
bias.n<-numeric(length=length(n))

 for(i in 1:length(n))
  {
  temp<-bgt.power(n=n[i],s=s,p.tr=p.tr,p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  pow.n[i]<-temp$power
  bias.n[i]<-temp$bias
  }
list(n=n,power=pow.n,bias=bias.n)
}
```

```
#usage:
bgtn.power(n=20:60,s=10,p.tr=0.001,p.hyp=0.005)

# Coverage

bgtn.cover<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
cov.n<-numeric(length=length(n))
bias.n<-numeric(length=length(n))

 for(i in 1:length(n))
  {
  temp<-bgt.cover(n=n[i],s=s,p.tr=p.tr,p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  cov.n[i]<-temp$cover
  bias.n[i]<-temp$bias
  }

list(n=n,cover=cov.n,bias=bias.n)
}
#usage:
bgtn.cover(n=20:60,s=10,p.tr=0.001,p.hyp=0.005)

#####################################################################
# Vary the true binomial proportion p.tr, other parameters fixed    #
# Input of p.tr as a vector                                         #
#####################################################################

# Power

bgtp.power<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
pow.p<-numeric(length=length(p.tr))
bias.p<-numeric(length=length(p.tr))

 for(i in 1:length(p.tr))
  {

  temp<-bgt.power(n=n,s=s,p.tr=p.tr[i],p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  pow.p[i]<-temp$power
  bias.p[i]<-temp$bias
  }

list(p.tr=p.tr,power=pow.p,bias=bias.p)
}

# usage:
test<-bgtp.power(n=100, s=10, p.tr=seq(0,0.005,0.0001), p.hyp=0.005, alternative="less",
method="SOC")
plot(x=test$p.tr,y=test$power, type="l")

# Coverage

bgtp.cover<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
cov.p<-numeric(length=length(p.tr))
bias.p<-numeric(length=length(p.tr))

 for(i in 1:length(p.tr))
  {
    temp<-bgt.cover(n=n,s=s,p.tr=p.tr[i],p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  cov.p[i]<-temp$cover
  bias.p[i]<-temp$bias
  }
list(p.tr=p.tr,cover=cov.p,bias=bias.p)
}

#usage:
test<-bgtp.cover(n=100, s=10, p.tr=seq(0,0.005,0.0001), p.hyp=0.005, alternative="less",
method="SOC")
plot(x=test$p.tr, y=test$cover, type="l")
```

```
##############################################
#                                            #
# Vary sets of n*s, other parameters fixed   #
#                                            #
##############################################

# # # Power

bgtns.power<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
if(length(n)!=length(s)){stop("vectors n and s must have exactly the same length")}

pow.ns<-numeric(length=length(n))
bias.ns<-numeric(length=length(n))
ns<-numeric(length=length(n))

 for(i in 1:length(n))
  {

  temp<-bgt.power(n=n[i],s=s[i],p.tr=p.tr,p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  pow.ns[i]<-temp$power
  bias.ns[i]<-temp$bias
  ns[i]<-n[i]*s[i]
  }

list(n=n, s=s, ns=ns, power=pow.ns,bias=bias.ns)
}
# usage:
bgtns.power(n=c(80,40,20,10),s=c(10,20,40,80), p.tr=0.005, p.hyp=0.01)

# # # # Coverage

bgtns.cover<-function(n, s, p.tr, p.hyp, conf.level=0.95, method="wilson",
alternative="two.sided")
{
if(length(n)!=length(s)){stop("vectors n and s must have exactly the same length")}
cov.ns<-numeric(length=length(n))
bias.ns<-numeric(length=length(n))
ns<-numeric(length=length(n))

 for(i in 1:length(n))
  {

  temp<-bgt.cover(n=n[i],s=s[i],p.tr=p.tr,p.hyp=p.hyp,method=method,alternative =
alternative, conf.level=conf.level)
  cov.ns[i]<-temp$cover
  bias.ns[i]<-temp$bias
  ns[i]<-n[i]*s[i]
  }
list(n=n,s=s,ns=ns,cover=cov.ns,bias=bias.ns)
}
# usage:
bgtns.cover(n=c(80,40,20,10),s=c(10,20,40,80), p.tr=0.005, p.hyp=0.01)

############################################################
#                                                          #
# Example of a plot as used in Comparisons of the methods  #
#                                                          #
############################################################

# Calculating the exact coverages for each the different CI methods and the bias
(implicit, in all function values)
#  input of group size s is a vector 1:200, other parameters fixed

wi.n10<-bgts.cover(n=10,s=c(1:200),p.tr=0.0025, p.hyp=0.05, conf.level=0.95,
alternative="two.sided", method="wilson")

wa.n10<-bgts.cover(n=10,s=c(1:200),p.tr=0.0025, p.hyp=0.05, conf.level=0.95,
alternative="two.sided", method="wald")

ac.n10<-bgts.cover(n=10,s=c(1:200),p.tr=0.0025, p.hyp=0.05, conf.level=0.95,
alternative="two.sided", method="AC")
```

```
cp.n10<-bgts.cover(n=10,s=c(1:200),p.tr=0.0025, p.hyp=0.05, conf.level=0.95,
alternative="two.sided", method="CP")

# cutting the graphical device into pieces

layout(mat=matrix(1:2, ncol=1), heights=c(6,3))

# setting some parameters for the margins of the plotting areas
par(mar=c(3,4,2,1), oma=c(0,0,0,0))

#  first main plot (Coverage Wilson dependent on s

plot(y=wi.n10$cover, x=wi.n10$group.size, type="l",ylim=c(0.7,1),main="n=10, true
binomial parameter=0.0025",ylab="Coverage probability",xlab="group size s",
lwd=2,col="black", lty=1 )

# adding a line for the nominal level, and lines for coverages of other CI methods to
the first plot

lines(y=c(0.95,0.95),x=c(1,200),lty=3,lwd=1)

lines(y=wa.n10$cover, x=wa.n10$group.size, lwd=2, col="red", lty=2)
lines(y=cp.n10$cover, x=cp.n10$group.size, lwd=2, col="blue", lty=3)
lines(y=ac.n10$cover, x=ac.n10$group.size, lwd=2, col="green", lty=4)

# adding a legend to the first plot

legend(x=135, y=0.82, legend=c("Wilson","Wald","CP","AC","nominal level"),
lwd=c(2,2,2,2,1),lty=c(1,2,3,4,3),col=c("black","red","blue","green","black"))

# second plot with the bias of estimator after changing some parameters

par(mar=c(5,4,1,1))
plot(y=wi.n10$bias, x=wi.n10$group.size, ylim=c(0,0.005), type="l", lwd=2,
ylab="bias(p)", xlab="group size s")


#############################################################################
#  4) approximate sample size and power for the  group testin Wilson CI/ Score test #
#############################################################################

##############################
#  approximate sample size  #
##############################

nbgtWilson<-function(p.tr, p.hyp, s, conf.level=0.95, power=0.8)
{
t.hyp=1-(1-p.hyp)^s
t.tr=1-(1-p.tr)^s
nasy<-( t.hyp*(1-t.hyp) * (qnorm(conf.level) + qnorm(power) )^2 ) / (t.tr-t.hyp)^2
n<-round(nasy)
n
}
# # # usage:
nbgtWilson(p.tr=0.03, p.hyp=0.05,s=3)


######################
#  approximate power #
######################

powerbgtWiapprox<-function(p.tr, p.hyp, n, s, conf.level=0.95)
{
t.hyp=1-(1-p.hyp)^s
t.tr=1-(1-p.tr)^s
powerquant=sqrt( (n*(t.tr-t.hyp)^2 ) / ( t.hyp*(1-t.hyp)) ) - qnorm(conf.level)
power=pnorm(powerquant)
power
}
# # # usage:
powerbgtWiapprox(p.tr=0.03, p.hyp=0.05, n=247, s=3, conf.level=0.95)


#############################################################################
#                                                                           #
# 5) SIMULATION of COVERAGE and POWER for CI methods, all in one function   #
# could be programmed much shorter, only for comparison with closed calculation    #
```

```
##############################################################################


bgt.sim.p<-function(n, s, n.sim, conf.level=0.95, alternative="two.sided", p.tr, p.hyp){

simu<-function(){

data<-rbinom(n=n,size=s,prob=p.tr)
    n.t=0
    for(i in data){
    if (i==0){n.t=n.t+1}
    else{n.t=n.t+0}
    }
# Wald #

int.wa<-bgt.wald(n=n,Y=n-n.t,s=s,conf.level=conf.level,
alternative=alternative)$conf.int

if (alternative=="two.sided"){

    if (int.wa[[1]]>p.hyp || int.wa[[2]]<p.hyp){dec.wa=1}
    else{dec.wa=0}
    if (int.wa[[1]]>p.tr || int.wa[[2]]<p.tr){cov.wa=0}
    else{cov.wa=1}
}
else{if(alternative=="greater"){

    if (int.wa[[1]]>p.hyp){dec.wa=1}
    else{dec.wa=0}
    if (int.wa[[1]]>p.tr){cov.wa=0}
    else{cov.wa=1}
}
else{if(alternative=="less"){

    if (int.wa[[2]]<p.hyp){dec.wa=1}
    else{dec.wa=0}
    if (int.wa[[2]]<p.tr){cov.wa=0}
    else{cov.wa=1}
} else{stop("alternative mis-specified")}}}


# Wilson #

int.wi<-bgt.wilson(n=n,Y=n-n.t,s=s,conf.level=conf.level,
alternative=alternative)$conf.int

if (alternative=="two.sided"){

    if (int.wi[[1]]>p.hyp || int.wi[[2]]<p.hyp){dec.wi=1}
    else{dec.wi=0}
    if (int.wi[[1]]>p.tr || int.wi[[2]]<p.tr){cov.wi=0}
    else{cov.wi=1}
}
else{if(alternative=="greater"){

    if (int.wi[[1]]>p.hyp){dec.wi=1}
    else{dec.wi=0}
    if (int.wi[[1]]>p.tr){cov.wi=0}
    else{cov.wi=1}
}
else{if(alternative=="less"){

    if (int.wi[[2]]<p.hyp){dec.wi=1}
    else{dec.wi=0}
    if (int.wi[[2]]<p.tr){cov.wi=0}
    else{cov.wi=1}
} else{stop("alternative mis-specified")}}}

# Agresti-Coull #

int.AC<-bgt.AC(n=n,Y=n-n.t,s=s,conf.level=conf.level, alternative=alternative)$conf.int

if (alternative=="two.sided"){

    if (int.AC[[1]]>p.hyp || int.AC[[2]]<p.hyp){dec.AC=1}
    else{dec.AC=0}
    if (int.AC[[1]]>p.tr || int.AC[[2]]<p.tr){cov.AC=0}
```

```
    else{cov.AC=1}
}
else{if(alternative=="greater"){

    if (int.AC[[1]]>p.hyp){dec.AC=1}
    else{dec.AC=0}
    if (int.AC[[1]]>p.tr){cov.AC=0}
    else{cov.AC=1}
}
else{if(alternative=="less"){

    if (int.AC[[2]]<p.hyp){dec.AC=1}
    else{dec.AC=0}
    if (int.AC[[2]]<p.tr){cov.AC=0}
    else{cov.AC=1}
} else{stop("alternative mis-specified")}}}

# Add 4 #

int.a4<-bgt.add4(n=n,Y=n-n.t,s=s,conf.level=conf.level,
alternative=alternative)$conf.int

if (alternative=="two.sided"){

    if (int.a4[[1]]>p.hyp || int.a4[[2]]<p.hyp){dec.a4=1}
    else{dec.a4=0}
    if (int.a4[[1]]>p.tr || int.a4[[2]]<p.tr){cov.a4=0}
    else{cov.a4=1}
}
else{if(alternative=="greater"){

    if (int.a4[[1]]>p.hyp){dec.a4=1}
    else{dec.a4=0}
    if (int.a4[[1]]>p.tr){cov.a4=0}
    else{cov.a4=1}
}
else{if(alternative=="less"){

    if (int.a4[[2]]<p.hyp){dec.a4=1}
    else{dec.a4=0}
    if (int.a4[[2]]<p.tr){cov.a4=0}
    else{cov.a4=1}
} else{stop("alternative mis-specified")}}}

# Clopper-Pearson #

int.CP<-bgt.CP(n=n,Y=n-n.t,s=s,conf.level=conf.level, alternative=alternative)$conf.int

if (alternative=="two.sided"){

    if (int.CP[[1]]>p.hyp || int.CP[[2]]<p.hyp){dec.CP=1}
    else{dec.CP=0}
    if (int.CP[[1]]>p.tr || int.CP[[2]]<p.tr){cov.CP=0}
    else{cov.CP=1}
}
else{if(alternative=="greater"){

    if (int.CP[[1]]>p.hyp){dec.CP=1}
    else{dec.CP=0}
    if (int.CP[[1]]>p.tr){cov.CP=0}
    else{cov.CP=1}
}
else{if(alternative=="less"){

    if (int.CP[[2]]<p.hyp){dec.CP=1}
    else{dec.CP=0}
    if (int.CP[[2]]<p.tr){cov.CP=0}
    else{cov.CP=1}
} else{stop("alternative mis-specified")}}}

c( dec.wald=dec.wa,cov.wald=cov.wa,
 dec.wilson=dec.wi,cov.wilson=cov.wi,
 dec.AC=dec.AC,cov.AC=cov.AC,
 dec.a4=dec.a4,cov.a4=cov.a4,
 dec.CP=dec.CP,cov.CP=cov.CP )
}
```

```
simul<-replicate (n=n.sim,simu())

pow.wald= sum(simul[1,1:n.sim])/n.sim
pow.wilson= sum(simul[3,1:n.sim])/n.sim
pow.AC= sum(simul[5,1:n.sim])/n.sim
pow.add4=sum(simul[7,1:n.sim])/n.sim
pow.CP= sum(simul[9,1:n.sim])/n.sim

cov.wald= sum(simul[2,1:n.sim])/n.sim
cov.wilson= sum(simul[4,1:n.sim])/n.sim
cov.AC= sum(simul[6,1:n.sim])/n.sim
cov.add4=sum(simul[8,1:n.sim])/n.sim
cov.CP= sum(simul[10,1:n.sim])/n.sim

list(
pow.wald=pow.wald, cov.wald=cov.wald,
pow.wilson=pow.wilson, cov.wilson=cov.wilson,
pow.AC=pow.AC, cov.AC=cov.AC,
pow.add4=pow.add4, cov.add4=cov.add4,
pow.CP=pow.CP, cov.CP=cov.CP)
}

# usage
bgt.sim.p(n.sim=1000, n=20, s=5, p.tr=0.2, p.hyp=0.1, alternative="two.sided")
```

11.2 Resampling interval for binomial group testing

```
bgt.res<-function(Y, n, s, n.sim=1000,conf.level=0.95)
{
p.obs=1-(1-Y/n)^(1/s)

## create an experiment using the estimated probability p.obs
## n.t is the number of positive groups in one group testing experiment
exp.step<-function(p.obs,n,s)
{
data=rbinom(n=n, size=s, prob=p.obs)
    n.t=0
    for(i in data){
    if (i==0){n.t=n.t+0}
    else{n.t=n.t+1}
    }
  exp.p.bgt=1-(1-n.t/n)^(1/s)
 # estimator for probability pi(of positive individuals) in this experiment
  exp.t=n.t/n
 # estimator of probability theta of a positive group in this experiment
c(exp.p.bgt=exp.p.bgt, exp.t=exp.t)
}
## perform n.sim experiments:
sim.step = replicate(n=n.sim, exp.step(p.obs=p.obs, n=n,s=s))
## sort the outcomes for building the empirical distribution
p.i.bgt = sort(sim.step[1,1:n.sim]) # distribution of ind. scale estimators
t.i = sort(sim.step[2,1:n.sim]) # distribution of group scale estimators
alpha=1-conf.level
## the quantiles of a distribution with n.sim elements
quant.u = as.integer((alpha/2)*n.sim)
quant.o = as.integer((1-alpha/2)*n.sim)+1

## the alpha/2 and 1-alpha/2 quantiles are chosen from the distributions, resulting in
## 1-alpha confidence intervals

CI.bgt=c(p.i.bgt[quant.u],p.i.bgt[quant.o])
 # CI from the distribution of ind. scale estimators
CI.t=c(t.i[quant.u],t.i[quant.o])
 # CI from the distribution of group scale estimators
CI.t.transf=c( 1-(1-CI.t[1])^(1/s), 1-(1-CI.t[2])^(1/s) )
 # CI for the indiv.estimator, transformed from the group scale CI.t

list(CI.res.bgt=CI.bgt, CI.transformed.from.Y.by.n=CI.t.transf)
}
# # # the same, can be tried for other examples
bgt.res(Y=3,n=24,s=7,n.sim=10000)

############################################################
# Monotony:                                                #
```

```
# Compare the outcome of ordering on different scales       #
#                                                           #
#############################################################

## use the function above to compare ordering on different scales:

bgt.res.Y<-function(n, s, n.sim=1000)
{
Y=0:n
diff.l<-numeric(length=length(Y))
diff.u<-numeric(length=length(Y))

# build the differences between the bounds of the two interval types
# for each possible value of Y

for(i in 1:length(Y))
  {
   int<-bgt.res(Y=Y[i],n=n,s=s,n.sim=n.sim)
   diff.l[i] = int$CI.res.bgt[1] - int$CI.transformed.from.Y.by.n[1]
   diff.u[i] = int$CI.res.bgt[2] - int$CI.transformed.from.Y.by.n[2]
  }
list( diff.lower=diff.l,diff.upper=diff.u)
}
bgt.res.Y(n=20, s=5, n.sim=1000)
# difference is always 0
```

11.3 R code for simple binomial testing

1) gives the functions for calculation of the used interval methods
2) gives indicator functions, probability and closed calculation for the binomial methods; needs the functions under 1) to be present in the R working space

```
#############################################
#                                           #
# Confidence interval for binomial testing  #
# used in the comparisons and illustrations #
#                                           #
#############################################

# 1) simple functions for CI calculation

########
# Wald #
########

wald<-function(Y, n, conf.level=0.95, alternative="two.sided")
{
alpha=1-conf.level
est=Y/n
z1s=qnorm(conf.level)
z2s=qnorm(1-alpha/2)

if(alternative=="two.sided"){
KI=c(est-z2s*sqrt(est*(1-est)/(n)),
     est+z2s*sqrt(est*(1-est)/(n)) )
}
else{if (alternative=="less"){
KI=c( 0 , est+z1s*sqrt(est*(1-est)/(n)) )
}
else{if(alternative=="greater"){
KI=c(est-z1s*sqrt(est*(1-est)/(n)), 1 )
}
else {stop("alternative mis-specified")}}}
conf.int=KI
conf.int
}

#usage:
wald(Y=0, n=20, alternative="less")

##########
# Wilson #
##########
```

```
wilson<-function(n,Y,conf.level=0.95,alternative="two.sided") {
alpha=1-conf.level
t=Y/n
if(alternative =="two.sided"){
        est.int=(Y+(qnorm(1-alpha/2)^2)/2)/(n+(qnorm(1-alpha/2))^2)
        w.se=((qnorm(1-alpha/2))*sqrt(n*t*(1-t)+(qnorm(1-alpha/2)^2)/4))/(n+qnorm(1-
alpha/2)^2)
        KI=c( est.int-w.se, est.int+w.se )
KI}
else{if(alternative=="less"){
        est.int=(Y+(qnorm(1-alpha)^2)/2)/(n+(qnorm(1-alpha))^2)
        w.se=((qnorm(1-alpha))*sqrt(n*t*(1-t)+(qnorm(1-alpha)^2)/4))/(n+qnorm(1-
alpha)^2)
        KI=c( 0, est.int+w.se )
}
else{if(alternative=="greater"){
        est.int=(Y+(qnorm(1-alpha)^2)/2)/(n+(qnorm(1-alpha))^2)
        w.se=((qnorm(1-alpha))*sqrt(n*t*(1-t)+(qnorm(1-alpha)^2)/4))/(n+qnorm(1-
alpha)^2)
        KI=c( est.int-w.se , 1 )
}
else{stop("argument alternative misspecified")}}}

conf.int = KI
conf.int
}

# usage:
wilson(Y=0, n=20, alternative="less")

##################
# Agresti-Coull  #
##################

AC<-function(Y, n, conf.level=0.95, alternative="two.sided")
{
alpha=1-conf.level
est=Y/n
z1s=qnorm(conf.level)
z2s=qnorm(1-alpha/2)

esti1s=(Y+(z1s^2)/2)/(n+z1s^2)
esti2s=(Y+(z2s^2)/2)/(n+z2s^2)

ni1s=n+z1s^2
ni2s=n+z2s^2

if(alternative=="two.sided"){

KI=c(esti2s-z2s*sqrt(esti2s*(1-esti2s)/(ni2s)),
     esti2s+z2s*sqrt(esti2s*(1-esti2s)/(ni2s)) )
}
else{if (alternative=="less"){
KI=c( 0 , esti1s+z1s*sqrt(esti1s*(1-esti1s)/(ni1s)) )
}

else{if(alternative=="greater"){
KI=c(esti1s-z1s*sqrt(esti1s*(1-esti1s)/(ni1s)), 1 )
}
else {stop("alternative mis-specified")}}}

conf.int=KI
conf.int
}

# usage:
AC(Y=0, n=20, alternative="less")

#############################################
# Blaker interval                           #
# derived from S code given in Blaker(2000)  #
#############################################

Blaker<-function (n,Y,conf.level=0.95, tolerance=1e-04, alternative="two.sided")
{
acceptbin <- function(Y,n,p)
```

```
{
  p1 = 1-pbinom(Y-1, n, p)
  p2 = pbinom(Y, n, p)
  a1 = p1 + pbinom( qbinom(p1,n,p)-1, n, p )
  a2 = p2+1-pbinom( qbinom(1-p2,n,p), n, p )
  return(min(a1,a2))
}

lower<-0
upper<-1

if(Y!=0)
  {lower<-qbeta((1-conf.level)/2, Y, n-Y+1)
    {while(acceptbin(Y,n,lower+tolerance)<(1-conf.level))
    lower=lower+tolerance}
  }

if(Y!=n)
  {upper<-qbeta(1-(1-conf.level)/2, Y+1, n-Y)
    {while(acceptbin(Y,n,upper-tolerance)<(1-conf.level))
    upper=upper-tolerance}
  }
conf.int=c(lower, upper)
conf.int
}

# usage: Compare to Clopper-Pearson below
Blaker(Y=3,n=10, conf.level=0.95)

###################
# Clopper-Pearson #
###################

binCP<-function(n, Y, conf.level=0.95, alternative="two.sided")
{
lower<-0
upper<-1
if(alternative=="two.sided")
{
 if(Y!=0)
   {lower<-qbeta((1-conf.level)/2, Y, n-Y+1)}
 if(Y!=n)
   {upper<-qbeta(1-(1-conf.level)/2, Y+1, n-Y)}
}
if(alternative=="less")
{
 if(Y!=n)
   {upper<-qbeta(1-(1-conf.level), Y+1, n-Y)}
}
if(alternative=="greater")
{
 if(Y!=0)
   {lower<-qbeta((1-conf.level), Y, n-Y+1)}
}
estimate=Y/n
conf.int=c(lower,upper)
conf.int
}

# usage:
CP(Y=3,n=10, conf.level=0.95)

###########################
# Second order corrected #
###########################

SOC<-function(n,Y,conf.level=0.95,alternative="two.sided")

{
esti<-Y/n
kappa<-qnorm(conf.level)
eta<-(kappa^2)/3 + 1/6
gamma1<-((13/18)*kappa^2 + 17/18)*(-1)
gamma2<-(kappa^2)/18 + 7/36

midpo<-(Y+eta)/(n+2*eta)
```

```
if(alternative=="less")
  {CI=c( 0 , midpo + kappa * sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) +
gamma2)/n)/sqrt(n) )}

if(alternative=="greater")
  {CI=c( midpo - kappa*sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) + gamma2)/n)/sqrt(n) ,
1)}

if (alternative=="two.sided")
{
kappa<-qnorm(1-(1-conf.level)/2)
eta<-(kappa^2)/3 + 1/6
gamma1<-((13/18)*kappa^2 + 17/18)*(-1)
gamma2<-(kappa^2)/18 + 7/36

CI=c( midpo - kappa*sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) + gamma2)/n)/sqrt(n) ,
 midpo + kappa*sqrt(esti*(1-esti) + (gamma1*esti*(1-esti) + gamma2)/n)/sqrt(n) )
}
CI
}

# usage:
SOC(n=20, Y=0, conf.level=0.95, alternative="less")

###################
# Jeffreys prior #
###################

Jef<- function(n,Y,conf.level=0.95,alternative="two.sided")
{
if(alternative=="less")
 {CI=c( 0 , qbeta(p=conf.level, shape1= Y+0.5, shape2=n-Y+0.5) )}

if(alternative=="greater")
 {CI=c( qbeta(p=1-conf.level, shape1= Y+0.5, shape2=n-Y+0.5) , 1 )}

if (alternative=="two.sided")
 {CI=c( qbeta(p=(1-conf.level)/2, shape1=Y+0.5, shape2=n-Y+0.5) , qbeta(p=1-(1-
conf.level)/2, shape1=Y+0.5, shape2=n-Y+0.5) )}

conf.int=CI
conf.int
}

# usage:
Jef(n=20, Y=0, conf.level=0.95, alternative="less")


##############################
#                            #
# 2) closed calculations     #
#                            #
##############################

#####################################
# Probability to observe a certain Y #
#####################################

Y.prob<-function(Y,n,p)
{
Y.p = choose(n,Y)*(p^Y)*(1-p)^(n-Y)
Y.p
}
Y.prob(Y=0,n=20,p=0.2)

# can be used until n=1029, Y=514
# Since numbers greater or equal choose(1030,515) cannot be represented in R anymore
# (i.e. 1e+309=Inf), and numbers x smaller than x=1e-323 are represented by 0, thus
# result in log(x)=-Inf, the following function might be used instead:
Y.prob2<-function(Y,n,p)
{
Y.p = exp( lchoose(n,Y) + Y*log(p) + (n-Y)*log(1-p) )
Y.p
}
Y.prob2(Y=0,n=20,p=0.2)

# This function gives exactly the same results as Y.prob() for n<1029,
```

```
# gives reasonable results for at least n<50000 and might be applied within the margins
# of representation of numbers given above for the single terms.
# Additionally it needs only 20% of the calculation time as Y.prob


################################
# Indikator function for Power #
################################

pInd.bin<-function(Y,n,p.hyp,conf.level,alternative,method)
{
if(method=="wald"){int=wald(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="wilson"){int=wilson(Y=Y,n=n, conf.level=conf.level,
alternative=alternative)}
if(method=="AC"){int=AC(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="CP"){int=CP(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="SOC"){int=SOC(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="Jef"){int=Jef(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}

if(int[1]<=p.hyp && int[2]>=p.hyp){pow=0}
else{pow=1}
pow
}

###################################
# Indikator funktion for Coverage  #
###################################

cInd.bin<-function(Y,n,p.tr,conf.level,alternative,method)
{
if(method=="wald"){int=wald(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="wilson"){int=wilson(Y=Y,n=n, conf.level=conf.level,
alternative=alternative)}
if(method=="AC"){int=AC(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="CP"){int=CP(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="SOC"){int=SOC(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}
if(method=="Jef"){int=Jef(Y=Y,n=n, conf.level=conf.level, alternative=alternative)}

if(int[1]<=p.tr && int[2]>=p.tr){cov=1}
else{cov=0}
cov
}

########################################
# Synthesis: closed calculation of POWER #
########################################

pow.bin<-function(n, p.tr, p.hyp, conf.level=0.95, alternative="two.sided",
method="wilson")
{
power=0
for(Y in 0:n)
  {power = power +
pInd.bin(Y=Y,n=n,p.hyp=p.hyp,conf.level=conf.level,alternative=alternative,method=method
) * Y.prob2(Y=Y,n=n,p=p.tr)}
power
}

#usage:
pow.bin(n=100,p.tr=0.001,p.hyp=0.005, alternative="less")
# for n=538, upper wilson(Y=0) bound sill includes 0.005
# for n=539 not anymore
pow.bin(n=538,p.tr=0.001,p.hyp=0.005, alternative="less", method="wilson")
pow.bin(n=539,p.tr=0.001,p.hyp=0.005, alternative="less", method="wilson")

#################################
# closed calculation of COVERAGE #
#################################

cov.bin<-function(n, p.tr, conf.level=0.95, alternative="two.sided", method="wilson")
{
cover=0
for(Y in 0:n)
  {cover = cover +
cInd.bin(Y=Y,n=n,p.tr=p.tr,conf.level=conf.level,alternative=alternative,method=method)
* Y.prob2(Y=Y,n=n,p=p.tr)}
cover
```

```
}
# usage:
cov.bin(n=100,p.tr=0.005, alternative="less")
cov.bin(n=538,p.tr=0.005, alternative="less", method="wilson")
cov.bin(n=539,p.tr=0.005, alternative="less", method="wilson")

########################################
# Vectorization for the true proportion  #
########################################
# Power
pow.binp.tr<- function(n, p.tr, p.hyp, conf.level=0.95, alternative="two.sided",
method="wilson")
{
power=numeric(length=length(p.tr))
for(i in 1:length(p.tr))
  {
  power[i]=pow.bin(n=n,p.tr=p.tr[i],
p.hyp=p.hyp,conf.level=conf.level,alternative=alternative,method=method)
  }
list(power=power, p.tr=p.tr)
}

# Coverage
cov.binp.tr<- function(n, p.tr, conf.level=0.95, alternative="two.sided",
method="wilson")
{
cover=numeric(length=length(p.tr))
for(i in 1:length(p.tr))
  {
cover[i]=cov.bin(n=n,p.tr=p.tr[i],conf.level=conf.level,alternative=alternative,method=m
ethod)
  }
list(cover=cover, p.tr=p.tr)
}

# Example: Coverage probability of two-sided Wilson Score interval for n=50,
# Compare Brown et al. 2001, Figure 5

test<-cov.binp.tr(n=50, p.tr=seq(0,1,0.0005), method="wilson", alternative="two.sided")
plot(x=test$p.tr,y=test$cover, type="l", ylab="Coverage Probability",
xlab="p",ylim=c(0.86,1), main= "Wilson Interval")
lines (y=c(0.95, 0.95), x=c(0,1),lty=3)

# Example: Coverage probabilities of upper 99%-limits of Wald, Wilson, Jeffreys prior
# and Second order corrected CI for n=30, Compare with Cai (2005), Fig.4

wa30<-cov.binp.tr(n=30, p.tr=seq(0,1,0.0005), method="wald",conf.level=0.99,
alternative="less")
wi30<-cov.binp.tr(n=30, p.tr=seq(0,1,0.0005), method="wilson", conf.level=0.99,
alternative="less")
soc30<-cov.binp.tr(n=30, p.tr=seq(0,1,0.0005), method="SOC", conf.level=0.99,
alternative="less")
jef30<-cov.binp.tr(n=30, p.tr=seq(0,1,0.0005), method="Jef", conf.level=0.99,
alternative="less")

layout(mat=matrix(1:4, ncol=2, byrow=TRUE))
par(mar=c(5,4,2,1),oma=c(0,0,0,0))

plot(x=wa30$p.tr,y=wa30 $cover, type="l", ylab="Coverage Probability",
xlab="p",ylim=c(0.94,1), main= "Wald Interval")
lines (y=c(0.99, 0.99), x=c(0,1),lty=3)

plot(x=wi30$p.tr,y=wi30 $cover, type="l", ylab="Coverage Probability",
xlab="p",ylim=c(0.94,1), main= "Score Interval")
lines (y=c(0.99, 0.99), x=c(0,1),lty=3)

plot(x=jef30$p.tr,y=jef30 $cover, type="l", ylab="Coverage Probability",
xlab="p",ylim=c(0.94,1), main= "Jeffreys Interval")
lines (y=c(0.99, 0.99), x=c(0,1),lty=3)

plot(x=soc30$p.tr,y=soc30 $cover, type="l", ylab="Coverage Probability",
xlab="p",ylim=c(0.94,1), main= "Second-Order Corrected Interval")
lines (y=c(0.99, 0.99), x=c(0,1),lty=3)
```

Hiermit versichere ich an Eides Statt, diese Diplomarbeit nur mit Hilfe der aufgeführten Quellen und Hilfsmittel erstellt zu haben. Die Arbeit wurde keiner anderen Prüfungskommission vorgelegt.

Hannover, 11.03.2005
Frank Schaarschmidt