NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Comparing positive predictive values for small samples with application to gene ontology testing

by

Clara-Cecilie Günther, Øyvind Bakke and
Mette Langaas

PREPRINT STATISTICS NO. 3/2009
DEPARTMENT OF MATHEMATICAL SCIENCES

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2009/S3-2009.pdf

Mette Langaas has homepage: http://www.math.ntnu.no/∼mettela
E-mail: Mette.Langaas@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and
Technology, N-7491 Trondheim, Norway.

Comparing positive predictive values for small samples with application to gene ontology testing

Clara-Cecilie Günther, Øyvind Bakke and Mette Langaas
Department of Mathematical Sciences.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

March 2009

Summary

Motivated by the challenge of detecting Gene Ontology (GO) categories which are over-represented or depleted when comparing biological findings represented by two over-lapping lists of genes, we examine the performance of different statistical tests. One key feature with this type of data is that the sample size at each GO category often is small and thus large sample asymptotic tests are not suitable. We look at four different test statistics in combination with parametric bootstrapping, and compare the methods with their asymptotic alternatives. We find that the choice of test statistic influence which GO categories are found to be significant, and all tests under study perform increasingly conservative as the sample size decreases. We observe that this problem is statistically the same as comparing the positive predictive values of two diagnostic tests.

1 Introduction

In some biological experiments the aim is, e.g. by using DNA microarrays, to discover genes that are differentially expressed between two or more conditions. The conditions may be defined by the presence or absence of a disease or by different treatments like diets, drugs or amount of physical exercise. As an example we consider a situation where the relationship between inborn aerobic capacity and cardiac gene expression in rats was studied, Bye, Langaas, Høydahl, Kemi, Heinrich, Koch, Britton, Najjar, Ellingsen and Wisløff (2008). The rats were born with either high running capacity (HCR) or low running capacity (LCR), and half of the rats were trained, while the others remained sedentary. Thus there were four groups of rats, LCR trained, LCR sedentary, HCR trained and HCR sedentary. Several comparisons were done, and the comparison of the gene expression for the sedentary HCR rats with the gene expression for the sedentary LCR rats resulted in a list of 1540 differentially expressed genes between these two groups.

However, since such lists contains only single genes, i.e. without information about potential connections to the other genes on the list, it can be challenging to interpret the biological meaning of the results. What may be more interesting for interpretation purposes is the biological pathways that are active in the conditions under study. To do this, groups of genes instead of single genes are considered. In this paper we consider groups of genes selected from a predefined set using the Gene Ontology (GO) vocabulary, The Gene Ontology Consortium (2000). GO is a vocabulary that classifies genes into the three main categories: biological process, molecular function and cellular component and their subcategories.

Given the list of differentially expressed genes from the experiment and the list of all genes present on the microarray chip, called the master list, the biologist wants to know whether certain gene classes are over-represented or depleted in the list of differentially expressed genes compared to the master list. In the rat example, we are interested in knowing if the number of genes related to aerobic capacity among those differentially expressed between the sedentary HCR rats and the sedentary LCR rats is higher than what we would expect by chance if we compare it to the master list. The list of differentially expressed genes is contained in the master list, and the statistical hypothesis problem is to test whether two binomial proportions are equal. Common approaches are Pearson's asymptotic χ^2-test and Fisher's exact test for large and small samples, respectively.

If there are more than two conditions in the experiment, several comparisons can be done which may each result in a list of differentially expressed genes between the conditions being compared. Then we would like to see whether some specific gene classes of interest are over-represented or depleted on one of the lists compared to one of the others. The two lists may either be mutually exclusive or partly overlapping. If they are mutually exclusive the problem reduces to test whether two binomial proportions are equal as for the master list problem and the same approaches can be used. We will consider only the situation of overlapping gene lists. In the rat example, we want to compare the list of differentially expressed genes between trained HCR and LCR rats to the list of differentially expressed genes between trained HCR rats and sedentary LCR rats.

Comparing two overlapping gene lists in terms of over-represented or depleted gene classes is statistically the same situation as comparing the positive predictive values for two diagnostic tests and several hypothesis tests for this situation can be found in the literature, see Leisenring, Alonzo and Pepe (2000), Wang, Davis and Soong (2006) and Moskowitz and Pepe (2006). Günther, Bakke, Lydersen and Langaas (2008) presented a likelihood ratio test and a restricted difference test and compared them to the other existing tests. Simulation experiments showed that for smaller sample sizes these tests did not preserve their test size. When comparing gene lists, the actual sample size is the number of genes associated with each of the three main GO-categories, not the number of genes on the microarray chip, nor the total number of genes on the lists. This number is usually quite small and large sample tests are not a suitable approach. Instead small sample tests should be applied.

In this paper we evaluate small sample tests for comparing two overlapping gene lists, i.e. to test whether the probabilities that a randomly chosen gene belongs to a specific gene class are equal for the two lists. We first describe the assumed model and define the null and alternative hypotheses in Section 2, and then present the test statistics and how to calculate the p-values in Section 3. A simulation study is conducted to assess the method and described in Section 4 and in Section 5 an example in which data from the literature is given. A short discussion is given in Section 6 before we end with the conclusions in Section 7.

2 Model and data

We assume that we have two lists of genes, list A and list B. For each gene we are interested in comparing the probability that it belongs to a certain gene class D given that it is on list A, with the probability that it belongs to D given that it is on list B. Our null hypothesis is that these two probabilities are equal. By defining the three events

- D: The gene belongs to gene class D.
- A: The gene is present on gene list A.
- B: The gene is present on gene list B.

we can express the null hypothesis as

$$H_0\colon P(D \mid A) = P(D \mid B). \tag{1}$$

Statistically, this is the same problem as testing equality of the positive predictive values of two diagnostic tests for the same disease. Two diagnostic tests with binary outcomes, i.e. positive or negative, are applied to each subject in the study. The positive predictive value (PPV) is defined as the probability that the subject has the disease of study given that the test is positive. If we let event D be that the subject has the disease, A the event that the outcome of test A is positive and B the event that the outcome of test B is positive, then the positive predictive value of test A is $\text{PPV}_A = P(D \mid A)$, and the positive predictive value of test B is $\text{PPV}_B = P(D \mid B)$. Our null hypothesis is that the two positive predictive values are equal, i.e. $H_0\colon P(D \mid A) = P(D \mid B)$ as in (1).

The Venn diagram in Figure 1 shows the six mutually exclusive events defined by A, B and D. We only look at the restricted sample space, i.e. $A \cup B$, and thereby only the part of D that intersects $A \cup B$. Let A^*, B^* and D^* be the complementary events of A, B and D respectively. Günther et al. (2008) argue that when comparing positive predictive values it suffices to consider only the subjects with at least one positive test result, which equals the set $A \cup B$. In the GO setting the number of genes belonging to the GO category D that are not present on any of the lists, i.e. the event $(A^* \cup B^*) \cap D$, is unknown as is the number of genes not present on the lists that do not belong to the GO category D, therefore the part of D that intersects with $A^* \cup B^*$ is not included.

To each of the six events in the Venn diagram there corresponds the probability q_i that event i occurs, $i = 1, ..., 6$. The sum of these probabilities is one, i.e. $\sum_{i=1}^{6} q_i = 1$. Associated with each event is also a random variable N_i, $i = 1, \ldots, 6$, N_i being the number of times event i occurs. We consider one main category at a time, such that in total there are $N = \sum_{i=1}^{6} N_i$ unique genes on the two lists associated with either biological process, cellular component or molecular function. The number N will typically change between the three main categories. Given N, the random variables $N_1, N_2, \ldots, N_6$ are multinomially distributed with parameters N and $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6)$. The joint probability function of $N_1, N_2, \ldots, N_6$ is

$$P\left(\bigcap_{i=1}^{6} (N_i = n_i)\right) = N! \prod_{i=1}^{6} \frac{q_i^{n_i}}{n_i!}. \tag{2}$$

The expected value of $\boldsymbol{N} = (N_1, N_2, N_3, N_4, N_5, N_6)$ is $\boldsymbol{\mu} = \text{E}(\boldsymbol{N}) = N \cdot \boldsymbol{q}$ and the covariance matrix is $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{N}) = N(\text{Diag}(\boldsymbol{q}) - \boldsymbol{q}^T\boldsymbol{q})$, Johnson, Kotz and Balakrishan (1997). We do not assume that a random gene's presence on list A is independent on its presence on list B and of whether it belongs to GO category D. This is implicitly handled by the multinomial model, since each gene yields only one observation of one of the six mutually exclusive events. We do however assume that the genes are sampled independently of each other and we will comment this further in Section 6.

Throughout this work, we assume that only N is fixed and that $\boldsymbol{N}$ are realisations of multinomial samples. Other sampling schemes are possible as well, for instance by fixing N_D, N_A and N_B, the number of genes belonging to gene class D, are present on list A and are present on list B respectively, and sampling $\boldsymbol{N}$ independently from three binomial distributions. In this report, we will not consider these approaches.

The probabilities $P(D|A)$ and $P(D|B)$ can be expressed in terms of the parameters $\boldsymbol{q}$ since

$$P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{q_4 + q_5}{q_1 + q_2 + q_4 + q_5}$$

and

$$P(D|B) = \frac{P(D \cap B)}{P(B)} = \frac{q_4 + q_6}{q_1 + q_3 + q_4 + q_6}.$$

Thus, the null hypothesis can be written

$$H_0 : \ \delta = \frac{q_4 + q_5}{q_1 + q_2 + q_4 + q_5} - \frac{q_4 + q_6}{q_1 + q_3 + q_4 + q_6} = 0. \tag{3}$$

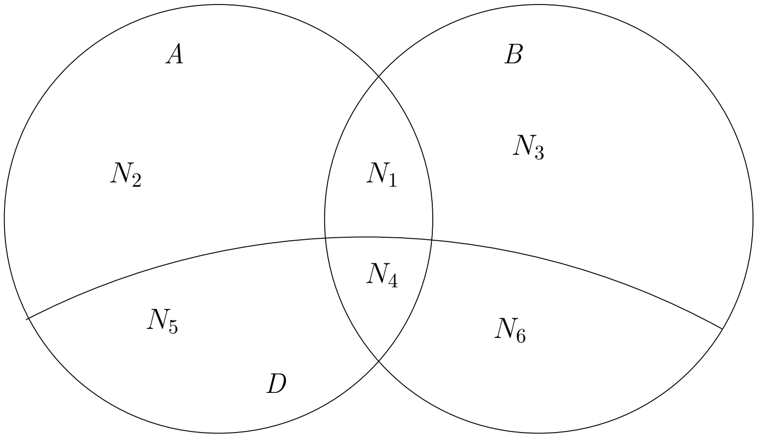


FIGURE 1: Venn diagram for the events D, A and B showing which events the random variables $N_1, \ldots, N_6$ correspond to.

There are several possible alternative hypotheses. If we are interested in whether there is an enrichment or depletion of genes belonging to gene class D on list A compared to list B, we have the two-sided alternative

$$H_1 : \ P(D|A) \neq P(D|B). \tag{4}$$

If we are interested in testing only whether there is an enrichment of genes belonging to gene class D on list A compared to list B, we have the one-sided alternative

$$H_1 : \ P(D|A) > P(D|B). \tag{5}$$

When testing whether there is a depletion of genes belonging to gene class D on list A compared to list B, the alternative hypothesis is

$$H_1 : \ P(D|A) < P(D|B). \tag{6}$$

In this work we will focus on the two sided alternative. We observe data $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6)$ which are realizations of $\boldsymbol{N} = (N_1, N_2, N_3, N_4, N_5, N_6)$ and can be represented in a table as shown in Table 1.

Event	D^*	D
$A \cap B$	n_1	n_4
$A \cap B^*$	n_2	n_5
$A^* \cap B$	n_3	n_6

TABLE 1: The observed data classified by the events A, B and D.

3 METHOD

In this section we present the test statistics we considered to test whether the probability of a gene belonging to gene class D given that it is present on list A is equal to the probability of a gene belonging to gene class D given that it is present on list B. We also describe how to calculate the p-values.

3.1 TEST STATISTICS

To test the null hypothesis (3), we consider four test statistics: a likelihood ratio test statistic, a score test statistic and two difference test statistics. They have all been shown to be asymptotically χ^2_1 distributed when the sample size is large, Casella and Berger (2002), Leisenring et al. (2000), Wang et al. (2006), but here we will use parametric bootstrapping to approximate their distribution under the null hypothesis for small samples. We describe the test statistics briefly, more details can be found in Günther et al. (2008).

The likelihood ratio test statistic is

$$T_{\text{LR}} = -2 \cdot \log(\lambda(\boldsymbol{N}))$$

where $\lambda(\boldsymbol{N})$ is the maximum likelihood of a multinomial sample under the null hypothesis divided by the general maximum likelihood of the multinomial sample. Let $\boldsymbol{Q}$ denote the parameter space for $\boldsymbol{q}$ and $\boldsymbol{Q}_0$ the subspace of $\boldsymbol{Q}$ in which $\boldsymbol{q}$ satisfy the constraint given by the null hypothesis (3). Then,

$$\lambda(\boldsymbol{n}) = \frac{\sup_{\boldsymbol{Q}_0} L(\boldsymbol{q}|\boldsymbol{n})}{\sup_{\boldsymbol{Q}} L(\boldsymbol{q}|\boldsymbol{n})}.$$

Let $\tilde{q}_i,\ i = 1, \ldots, 6$, be the restricted maximum likelihood estimates, that is, the maximum likelihood estimates under H_0, and let $\hat{q}_i$, $i = 1, \ldots, 6$ be the unrestricted general maximum likelihood estimates for the multinomial distribution, i.e. $\hat{q}_i = n_i/N$. Inserting these estimates in the log-likelihood function for the multinomial distribution leads to the test statistic

$$T_{\text{LR}} = -2 \left(\sum_{i=1}^{6} n_i \cdot (\log \tilde{q}_i - \log \hat{q}_i) \right). \tag{7}$$

Note that $\tilde{q}_i$, $i = 1, \ldots, 6$, cannot be written in any comprehensible closed form, but can be found using an optimization routine or analytically by solving a Lagrangian system of equations. We do the latter using Maple 12, for details see Günther et al. (2008).

The difference tests are based on the estimator $g(\boldsymbol{N})$ for the difference δ in (3),

$$g(\boldsymbol{N}) = \frac{N_4 + N_5}{N_1 + N_2 + N_4 + N_5} - \frac{N_4 + N_6}{N_1 + N_3 + N_4 + N_6} \tag{8}$$

and the test statistic is derived by subtracting the expectation of $g(\boldsymbol{N})$ and dividing by its approximate standard deviation, which is found by taking the variance of the first order Taylor expansion of $g(\boldsymbol{N})$. Let $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{N})$ and $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{N})$ as defined in Section 2. This yields

$$T_{\mathrm{g}} = \frac{(g(\boldsymbol{N}) - g(\boldsymbol{\mu}))^2}{G^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\, G(\boldsymbol{\mu})} \tag{9}$$

where G is a vector containing the first order partial derivatives of $g(\boldsymbol{N})$ with respect to the components of $\boldsymbol{N}$ and G^T is the transpose of G. $G(\boldsymbol{\mu})$ is G with $\boldsymbol{\mu}$ inserted for $\boldsymbol{N}$.

Under the null hypothesis $g(\boldsymbol{\mu}) = 0$. $G(\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}$ depend on the unknown parameters $\boldsymbol{q}$ which must be estimated when calculating the test statistic. We can either use the unrestricted maximum likelihood estimates $\hat{\boldsymbol{q}}$ for the multinomial distribution or the restricted maximum likelihood estimates $\tilde{\boldsymbol{q}}$ under H_0. In the first case we refer to the test as the *unrestricted* difference test (uDT) and denote the test statistic T_{uDT} and in the second case we refer to the test as the *restricted* difference test (rDT) and denote the test statistic T_{rDT}.

Leisenring et al. (2000) presented a score test, which we denote the LAP test, for testing equivalence of positive predictive values of two diagnostic tests, based on generalized estimating equations. They define three indicator variables. First D_{ij} indicates the disease status of subject i for diagnostic test j, i.e. $D_{ij} = 0$ if the subject does not have the disease and $D_{ij} = 1$ if it does have the disease. Z_{ij} indicates which test is used, it is 0 for test A and 1 for test B. X_{ij} indicates the test result, it is 0 if the test is negative and 1 if it is positive. Then the positive predictive value for test A can be written $\mathrm{PPV}_A = P(D_{ij} = 1 \mid Z_{ij} = 0, X_{ij} = 1)$ and the positive predictive value for test B is $\mathrm{PPV}_B = P(D_{ij} = 1 \mid Z_{ij} = 1, X_{ij} = 1)$. Leisenring et al. (2000) fit the generalized linear model

$$\mathrm{logit}(P(D_{ij} = 1 \mid Z_{ij}, X_{ij} = 1)) = \alpha_P + \beta_P Z_{ij},$$

and test whether $\beta_P = 0$ which is equivalent to testing whether $\mathrm{PPV}_A = \mathrm{PPV}_B$. We translate the test to the GO situation and in our notation the test statistic can be written as

$$T_{\mathrm{LAP}} = \frac{((N_1 + N_2 + N_4 + N_5)(N_4 + N_6) - (N_1 + N_3 + N_4 + N_6)(N_4 + N_5))^2}{f(N_1, N_2, N_3, N_4, N_5, N_6)}, \tag{10}$$

where

$$
\begin{aligned}
&f(N_1, N_2, N_3, N_4, N_5, N_6)\\
&\qquad = N_1(N_2 - N_3 + N_5 - N_6)^2 \left(\frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2\\
&\qquad + N_2(N_1 + N_3 + N_4 + N_6)^2 \left(\frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2\\
&\qquad + N_3(N_1 + N_2 + N_4 + N_5)^2 \left(\frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2\\
&\qquad + N_4(N_2 - N_3 + N_5 - N_6)^2 \left(1 - \frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2\\
&\qquad + N_5(N_1 + N_3 + N_4 + N_6)^2 \left(1 - \frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2\\
&\qquad + N_6(N_1 + N_2 + N_4 + N_5)^2 \left(1 - \frac{2N_4 + N_5 + N_6}{2N_1 + N_2 + N_3 + 2N_4 + N_5 + N_6} \right)^2 .
\end{aligned}
$$

The numerator can be found from by setting the difference in (8) equal to 0 and rearranging the terms. In the denominator, the number of genes that do not belong to GO category D, N_1, N_2 and N_3, are each multiplied by the proportion of genes that belong to the category D and in this proportion, the genes that are present on both lists, N_1 and N_4, are given double weight. The number of genes that belong to GO category D, N_4, N_5 and N_6, are each multiplied by the proportion of genes that do not belong to the gene class D where the genes that are present on both lists are given double weight.

3.2 Calculation of p-values

We will use parametric bootstrapping to approximate the distribution of the test statistics under the null hypothesis and find approximate p-values. The test statistic of interest, is either T_{LAP}, T_{LR}, T_{uDT} or T_{rDT}. To calculate the p-values we use the following algorithm:

1. For a given sample of size N, find the maximum likelihood estimates of the parameters under H_0, $\tilde{\boldsymbol{q}}$, and calculate the test statistic t for this sample, denoted t_s.

2. Draw B samples from the multinomial distribution with parameters N and $\tilde{\boldsymbol{q}}$.

3. Calculate the test statistic t_k for each of these samples, $1 \leq k \leq B$.

4. The p-value is given as $\sum_{k=1}^{B} I(t_k \geq t_s)$, where $I(t_k \geq t_s) = \begin{cases} 1 & \text{if } t_k \geq t_s \\ 0 & \text{if } t_k < t_s \end{cases}$, thus the p-value is the proportion of simulated test statistics greater than or equal to the given test statistics.

4 Assessment of method

To assess the performance of the four tests in terms of test size, we perform a simulation study. The test size is the probability of making a type I error, i.e. for rejecting H_0 when H_0 is true. We consider different sample sizes to evaluate the effect of N on the test size, and we also use several parameter values of $\boldsymbol{q}$ in the multinomial distribution to explore different areas of the null hypothesis. All analysis are performed using the R language, R Development Core Team (2008), except finding the maximum likelihood estimates under H_0 which is done using Maple 12.

4.1 Simulation algorithm

Given $\boldsymbol{q}$ and N, we draw M datasets from the multinomial distribution with parameters $\boldsymbol{q}$ and N. For each of these datasets we find the p-value using parametric bootstrapping as described in Section 3.2.

4.2 Cases under study

The data sets are generated from the parameters $\boldsymbol{q}$ in the multinomial distribution and we choose six cases of parameters, given in Table 2 and depicted in Figure 2.

Case	q_1	q_2	q_3	q_4	q_5	q_6
1	0.068	0.135	0.135	0.527	0.068	0.068
2	0.043	0.130	0.130	0.348	0.174	0.174
3	0.267	0.267	0.267	0.067	0.067	0.067
4	0.300	0.267	0.267	0.033	0.067	0.067
5	0.400	0.200	0.200	0.100	0.050	0.050
6	0.450	0.200	0.200	0.050	0.050	0.050

TABLE 2: Specification of parameters in the simulation study.

The parameters in case 1 and 2 are motivated by the setting for diagnostic tests and chosen as described in the multinomial simulation experiment of Günther et al. (2008). In case 3–6, we first set the probabilities $o_1 = P(A \cap B)$, $o_2 = P(A \cap B^*)$ and $o_3 = P(A^* \cap B)$ and then $p_1 = P(D|A \cap B)$, $p_2 = P(D|A \cap B^*)$ and $p_3 = P(D|A^* \cap B)$. From these probabilities $\boldsymbol{q}$ are calculated as follows,

$$q_i = \begin{cases} o_i(1-p_i) & i = 1,2,3 \\ o_i p_i & i = 4,5,6. \end{cases}$$

In case 3 $o_1 = o_2 = o_3 = 1/3$ and $p_1 = p_2 = p_3 = 1/5$. In case 4 $o_1 = o_2 = o_3 = 1/3$ and $p_1 = 1/10$ while $p_2 = p_3 = 2/10$. The probabilities in case 5 are $o_1 = 1/2$, $o_3 = o_4 = 1/4$ and $p_1 = p_2 = p_3 = 1/5$. Finally, in case 6, $o_1 = 1/2$, $o_2 = o_3 = 1/4$, $p_1 = 1/10$ and $p_2 = p_3 = 2/10$.

The remaining parameter in the multinomial distribution, N, must also be chosen and since we are considering small sample sizes, we use $N = 10, 15, 20$ and 25. For each of the values of N all the cases given in Table 2 are run. In each of the six cases we draw $M = 10000$ samples and for each of these samples we draw $B = 10000$ bootstrap samples.

4.3 Results

The test size is estimated as the proportion of p-values being less than or equal to the chosen nominal level α. Let W be a random variable counting the number of p-values smaller than or equal to α.

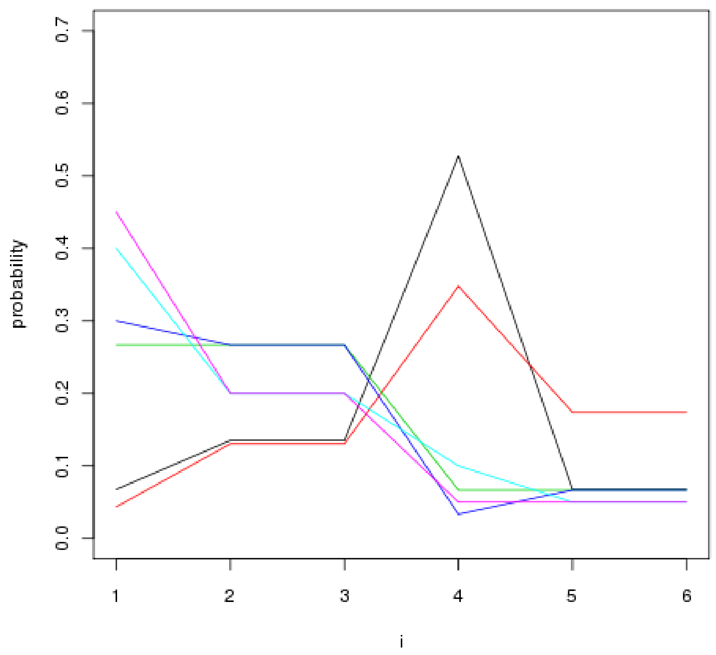


FIGURE 2: Values of q_i, $i = 1, \dots, 6$, black: case 1, red: case 2, green: case 3, dark blue: case 4, turquoise: case 5, cyan: case 6.

Then W is binomially distributed with size M, the number of p-values generated, and probability α. The estimate of the test size of the test, $\hat{\alpha}$ is then

$$\hat{\alpha} = \frac{W}{M}. \tag{11}$$

We say that the test preserves its test size if $\hat{\alpha} \leq \alpha$. The smaller $\hat{\alpha}$ is, while less than α, the more conservative the test is. If $\hat{\alpha} > \alpha$ the test does not preserve its test size and we say that it is too optimistic. We choose $\alpha = 0.05$ and calculate $\hat{\alpha}$ for the four test statistics, six cases and four values of N.

Table 3 show the estimated test size for all the combinations of $\boldsymbol{q}$, N and test statistic. There is one table for each of the six cases. The likelihood ratio test has the largest test size in all the cases and for all values of N except for $N = 20$ and $N = 25$ in case 1 and $N = 25$ in case 2. The unrestricted difference test has the smallest test size in all the cases and for all values of N except for $N = 20$ and $N = 25$ in case 1 and $N = 25$ in case 2, which are the exceptions when the likelihood ratio test has the smallest test size. The test size of the LAP test and the unrestricted difference test lies somewhere in between, which one is the largest varies.

When $N = 10$, all the tests are conservative for all the cases, but when N increases, the test size also increases and in case 1 and 2 the tests do not preserve their test size for $N \geq 15$. This also happens in case 3 for $N \geq 20$ for the likelihood ratio test and for $N = 25$ for the restricted difference test. In case 4, 5 and 6, the tests are conservative for all values of N except the likelihood ratio test which test

(a) Case 1

N	LAP	LRT	rDT	uDT
10	0.042	0.046	0.039	0.035
15	0.059	0.061	0.060	0.057
20	0.063	0.061	0.062	0.062
25	0.055	0.052	0.054	0.055

(b) Case 2

N	LAP	LRT	rDT	uDT
10	0.038	0.043	0.040	0.018
15	0.056	0.058	0.057	0.047
20	0.056	0.057	0.057	0.054
25	0.058	0.057	0.058	0.057

(c) Case 3

N	LAP	LRT	rDT	uDT
10	0.014	0.028	0.026	0.007
15	0.029	0.044	0.041	0.024
20	0.041	0.055	0.051	0.038
25	0.047	0.056	0.053	0.046

(d) Case 4

N	LAP	LRT	rDT	uDT
10	0.012	0.023	0.021	0.004
15	0.026	0.041	0.036	0.020
20	0.035	0.047	0.044	0.029
25	0.044	0.051	0.047	0.041

(e) Case 5

N	LAP	LRT	rDT	uDT
10	0.010	0.022	0.020	0.008
15	0.021	0.037	0.033	0.020
20	0.032	0.042	0.039	0.031
25	0.040	0.049	0.046	0.040

(f) Case 6

N	LAP	LRT	rDT	uDT
10	0.007	0.013	0.012	0.004
15	0.014	0.024	0.022	0.011
20	0.029	0.037	0.034	0.026
25	0.039	0.045	0.041	0.037

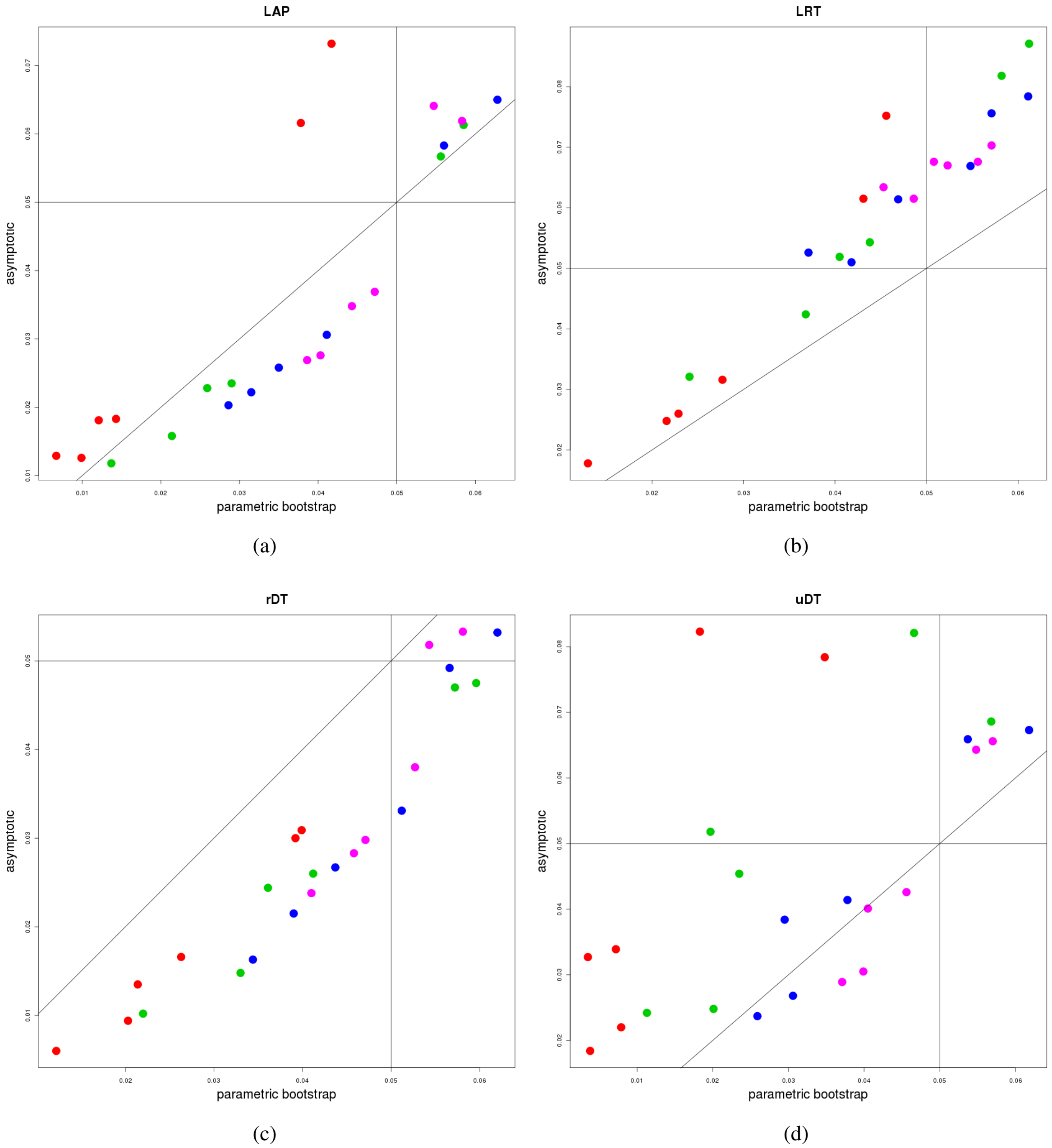
TABLE 3: Estimated test size, $\hat{\alpha}$, for $\alpha = 0.05$. LAP denotes the LAP test, LRT the likelihood ratio test and uDT and rDT denote the unrestricted and restricted difference test respectively.

size is 0.050 for $N = 25$ in case 4 and 6.

Figure 3 shows the estimated test size for the asymptotic methods plotted against the estimated test size for the parametric bootstrap methods, there is one plot for each method for $\alpha = 0.05$. If the points lie above the diagonal line, the test size of the asymptotic test is higher than the test size of the parametric bootstrap test, and lower if the points are below the line. If the points lie above the horizontal line the test size for the asymptotic test is greater than $\alpha = 0.05$ and smaller if they lie below the line. Similarly, for the points that lie to the right of the vertical line, the test size for the parametric bootstrap test is higher than 0.05 and it is lower than 0.05 if they lie to the left of this line.

We note in particular that for all the cases and for all values of N, the test size for the parametric bootstrap restricted difference test is greater than the test size for the large sample restricted difference test. For the likelihood ratio test, the opposite is true, the test size of the asymptotic likelihood ratio test is greater than the test size of the parametric bootstrap likelihood ratio test. This indicates that the parametric bootstrap test is an improvement compared to the asymptotic likelihood ratio test for small samples. However, the asymptotic likelihood ratio test does not preserve its test size in 15 of the 24 combinations of N and $\boldsymbol{q}$ and in six of those the parametric bootstrap test is still too optimistic. To use the parametric bootstrap restricted difference test does not yield an improvement compared to using the asymptotic restricted difference test.

Figure 4 shows the observed level, i.e. the test size, of the tests plotted against the nominal level for $N = 10, 15, 20, 25$ in case 3 for a chosen nominal level in the range from 0 to 0.10. We see that the test size increases when N increases and also that the tests yield more similar results for the higher values of N. The unrestricted difference test and the LAP test preserve the test size in all the cases while the likelihood ratio test is too optimistic when $N = 20$ or 25.



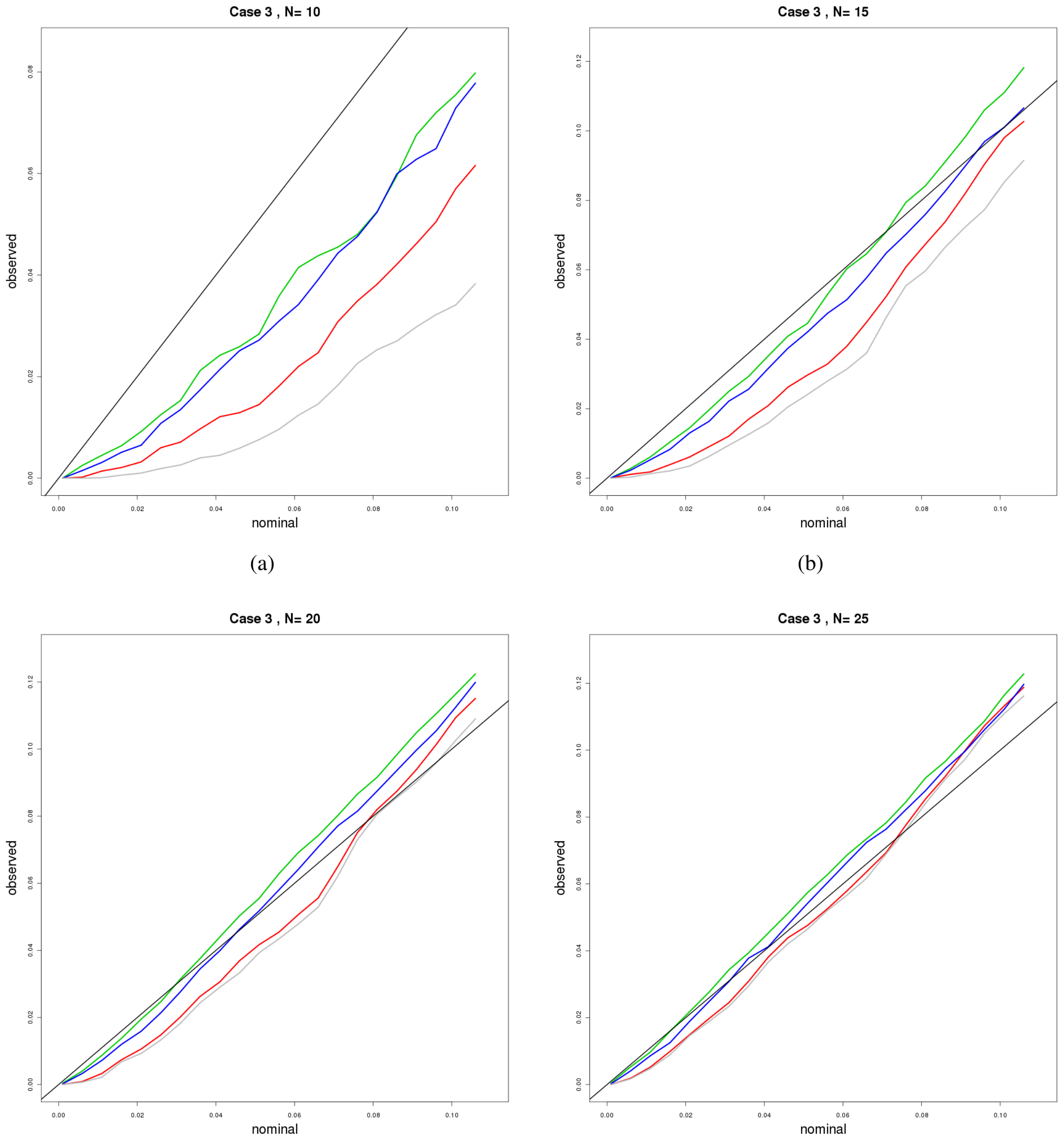
(a)

(b)

(c)

(d)

FIGURE 3: Estimated test size for the asymptotic versus the small sample tests, for different values of N: Red=10, green=15, dark blue=20, cyan=25.



(a)

(b)

(c)

(d)

FIGURE 4: Observed level versus nominal level for (a) $N = 10$, (b) $N = 15$, (c) $N = 20$, (d) $N = 25$, green = LRT, red = LAP, dark blue = rDT, grey = uDT.

5 Example from Gene Ontology

As an example of how the tests perform on a data set from literature, we use part of the data presented by Bye et al. (2008). To estimate the effect of running capacity of the trained rats, we compare the gene expression for the HCR trained rats with the LCR trained rats. This gives us a list of differentially expressed genes between these two groups, we call this list A. It may be of interest to estimate the joint effect of training and inbread running capacity by comparing the trained HCR rats versus the sedentary LCR rats. This gives us another list of differentially expressed genes which we call list B. To determine which genes are differentially expressed a cut-off must be chosen. For each gene, a p-value and an adjusted p-value are calculated. The adjusted p-values are adjusted using the Benjamini-Hochberg step-up procedure to control the false discovery rate (FDR), Benjamini and Hochberg (1995). The cut-off is chosen such that all the genes that have a p-value smaller than or equal to this value are said to be differentially expressed. We will use two different cut-offs and first we choose an FDR cut-off of 0.025 for both lists, which yields 12 genes on list A and 24 genes on list B. These lists are submitted to eGOn, Beisvåg, Jünge, Bergum, Jølsum, Lydersen, Günther, Ramampiaro, Langaas, Sandvik and Lægreid (2006), a web-based tool that automatically translates the lists to GO categories. We are interested in genes annotated to the main category molecular function. There are three genes from the first list and nine genes from the second list annotated to this category. Of these genes two are on both list A and B and therefore there are $N = 10$ unique genes on the two lists associated with molecular function.

The GO tree has several levels corresponding to the hierarchy of the GO categories. One gene can belong to more than one GO category, and given that it belongs to a subcategory it will also belong to the parent categories of this subcategory on the upper levels. After submitting the lists, one has to choose which main category to consider, i.e. either molecular function, biological process or cellular component. Level 1 is the main category itself with no subcategories, e.g. molecular function. The higher level number is chosen, the more subcategories are included, and they are all subcategories of the chosen main category. We choose to display the GO tree at level 3 for the main category molecular function and Table 4 shows the 11 GO categories that are represented on the lists, i.e. the categories which the genes on the lists belong to. A hypothesis test is performed for each category, testing whether it is over-represented or depleted on one of the lists compared to the other list.

If we use an FDR cut-off of 0.05 on differential expression instead we get two lists of 30 and 63 genes, 42 of these genes can be classified under the main category molecular function. Within this category, seven genes are present on both lists, seven genes are present only on list A and 21 genes are present only on list B, yielding $N = 35$ unique genes. Table 5 shows the GO categories for these genes along with their p-values.

Table 4 and 5 both include a column with the p-value calculated by eGOn. These p-values are calculated using the asymptotic LAP test. We calculate the p-values for the three other tests, i.e. the likelihood ratio test and the restricted and unrestricted difference test using parametric bootstrapping and compare them to the asymptotic p-values for all four tests. When performing the bootstrapping we draw $B = 10000$ bootstrap samples for each GO category.

Table 6 shows the results for the FDR cut-off of 0.025. The GO category ion binding (GO:0043167) is significant when using either the parametric bootstrap or asymptotic likelihood ratio or restricted difference test. It is also significant when using the asymptotic unrestricted difference test, while it is not significant when using the parametric bootstrap or asymptotic LAP test or the asymptotic unrestricted difference test. None of the other GO categories are significant for any of the tests.

GO identifier	Name	p-value	n_4	n_5	n_6
GO:0005488	binding	0.157	2	1	5
GO:0030246	carbohydrate binding	0.273	1	0	0
GO:0043167	ion binding	0.077	1	1	0
GO:0008289	lipid binding	0.279	0	1	0
GO:0003676	nucleic acid binding	0.317	0	0	1
GO:0005515	protein binding	0.705	2	0	5
GO:0046906	tetrapyrrole binding	0.279	0	1	0
GO:0003824	catalytic activity	0.245	1	1	2
GO:0016787	hydrolase activity	0.273	1	0	0
GO:0016874	ligase activity	0.317	0	0	1
GO:0016491	oxidoreductase activity	0.46	0	1	1

TABLE 4: GO categories within molecular function with their corresponding p-values and number of genes on the lists.

When considering only the parametric bootstrap tests, in general the likelihood ratio test and restricted difference test give similar p-values which in some cases are smaller than the p-values for the LAP test and the unrestricted difference test. One example is the GO category lipid binding (GO:0008289) where the p-values are 0.160, 0.065, 0.065 and 0.220 for the LAP, likelihood ratio, restricted difference and unrestricted difference tests respectively. Even though lipid binding is not significant for any of these tests, it is not far from being significant for the LRT and rDT tests which is not the case for the LAP and unrestricted difference tests. Together with the example ion binding, this indicates that a GO category may be declared significant more often for the likelihood ratio test and restricted difference tests than with the LAP and unrestricted difference tests. This coincide with the findings in the simulation experiments where the estimated test size in several cases were higher for the likelihood ratio and restricted difference tests than for the LAP and uDT tests.

Table 7 shows the results for the FDR cut-off of 0.05. The GO category catalytic activity (GO:0003824) is significant with a p-value <0.05 for all the tests, both the parametric bootstrap and asymptotic tests. The GO category hydrolase activity (GO:0016787) is significant when using the parametric bootstrap LRT or rDT tests and when using the asymptotic LRT, rDT and uDT tests. We see the same for the GO category substrate-specific transporter activity (GO:0022892), except that it is not significant using the asymptotic uDT test. We note that the category ion binding which is significant when we use an FDR cut-off of 0.025 is not significant now. In general, for the parametric bootstrap tests, the LRT and rDT tests yield similar p-values that are often smaller than the p-values for the LAP and uDT tests. The difference between the p-values can be quite large and for the GO term lipid binding (GO:0008289) the p-values are 0.179, 0.0345, 0.036 and 0.188 for the LAP, likelihood ratio, restricted difference and urestricted difference tests respectively. This example shows that the choice of test statistic is critical when finding GO categories that are significantly over-represented or depleted in one gene list compared to the other list. The differences between the parametric bootstrap and asymptotic tests do not follow a clear pattern, for some GO categories the parametric bootstrap p-values are smaller, for other GO categories they are greater.

The GO category chromatin binding (GO:0003682) was found to be significantly over-represented on the list of differentially expressed genes between HCR and LCR sedentary rats, using an FDR cut-off

GO identifier	Name	p-value	n_4	n_5	n_6
GO:0005488	binding	0.651	5	7	18
GO:0030246	carbohydrate binding	0.325	1	0	0
GO:0003682	chromatin binding	0.313	0	0	1
GO:0043167	ion binding	0.13	3	1	1
GO:0008289	lipid binding	0.161	1	1	0
GO:0003676	nucleic acid binding	0.485	1	1	5
GO:0000166	nucleotide binding	0.518	1	2	3
GO:0005515	protein binding	0.544	4	6	14
GO:0046906	tetrapyrrole bindin	0.308	0	1	0
GO:0003824	catalytic activity	0.016	5	5	6
GO:0016787	hydrolase activity	0.059	1	4	2
GO:0016874	ligase activity	0.56	1	0	2
GO:0016829	lyase activity	0.325	1	0	0
GO:0016491	oxidoreductase activity	0.226	2	1	1
GO:0016740	transferase activity	1	1	0	1
GO:0030234	enzyme regulator activity	0.325	1	0	0
GO:0030695	GTPase regulator activity	0.325	1	0	0
GO:0060089	molecular transducer activity	0.325	1	0	0
GO:0004871	signal transducer activity	0.325	1	0	0
GO:0005198	structural molecule activity	1	1	0	1
GO:0005201	extracellular matrix structural constituent	0.325	1	0	0
GO:0008307	structural constituent of muscle	0.313	0	0	1
GO:0030528	transcription regulator activity	0.388	1	1	1
GO:0003702	RNA polymerase II transcription factor activity	0.325	1	0	0
GO:0003700	transcription factor activity	0.388	1	1	1
GO:0016564	transcription repressor activity	0.325	1	0	0
GO:0005215	transporter activity	0.168	1	2	1
GO:0022892	substrate-specific transporter activity	0.073	1	2	0
GO:0022857	transmembrane transporter activity	0.168	1	2	1

TABLE 5: GO categories within molecular function with their corresponding p-values and number of genes on the lists.

	Parametric bootstrap				Asymptotic			
GO identifier	LAP	LRT	rDT	uDT	LAP	LRT	rDT	uDT
GO:0005488	0.198	0.315	0.344	0.204	0.157	0.249	0.354	0.109
GO:0030246	0.162	0.072	0.067	0.179	0.273	0.127	0.132	0.257
GO:0043167	0.054	0.004	0.007	0.072	0.077	0.010	0.012	0.021
GO:0008289	0.160	0.065	0.065	0.220	0.279	0.074	0.065	0.221
GO:0003676	0.383	0.371	0.401	0.446	0.317	0.433	0.540	0.289
GO:0005515	0.845	0.803	0.817	0.805	0.705	0.687	0.683	0.699
GO:0046906	0.146	0.062	0.062	0.206	0.279	0.074	0.065	0.221
GO:0003824	0.338	0.263	0.219	0.448	0.245	0.207	0.213	0.194
GO:0016787	0.164	0.073	0.065	0.181	0.273	0.127	0.132	0.257
GO:0016874	0.390	0.375	0.402	0.454	0.317	0.433	0.540	0.289
GO:0016491	0.678	0.454	0.299	0.714	0.460	0.376	0.353	0.431

TABLE 6: Parametric bootstrap and asymptotic p-values for the subcategories of molecular function.

of 0.05, compared to the list of all genes by Bye et al. (2008). Another GO category, nucleic acid binding (GO:0003676) was significantly over-represented on the list of genes that were significantly more expressed for HCR rats than for LCR rats compared to the list of genes that were significantly more expressed for LCR rats than for HCR rats, Bye et al. (2008). None of these GO-categories are over-represented in our two lists, but since we are not comparing the gene expression between sedentary HCR and LCR rats it is not surprising.

Instead of comparing the comparison of gene expression for the HCR trained rats and the LCR trained rats to the comparison of trained HCR rats and sedentary LCR rats, we could have compared the gene expression of trained LCR rats with the gene expression of the sedentary LCR rats directly. This is done in Bye et al. (2008). The list of differentially expressed genes is then submitted to eGOn and compared to the master list. However, with an FDR cut-off of 0.05 the comparison results in only one gene on the list and this gene is not annotated to any GO category.

With the first FDR cut-off of 0.025, we compared the two lists at 11 GO categories and with the second FDR cut-off at 0.05 we compared the lists at 29 GO categories. The problem thus involves multiple testing, and the p-values should be adjusted accordingly. This has not been done when comparing the methods and the p-values in Table 6 and 7 are therefore unadjusted.

6 DISCUSSION

To obtain list of differentially expressed genes, a cut-off on the differential expression must be set. The lists can then be submitted to a GO tool, e.g. eGOn, to discover GO categories that are over-represented or depleted. This approach has been criticised, see Goeman and Bühlman (2007) for an overview. Firstly, it is not clear where the cut-off should be set and secondly, one may argue that all the data should be used. Other proposed methods address this problem by either using all the p-values from the experiment or use raw expression data instead of p-values, see Goeman and Bühlman (2007).

The statistical tests in this report all treat the genes as the sampling units and are based on the assumption that the genes on the lists act independently under the null hypothesis. Statistically, it would

	Parametric bootstrap				Asymptotic			
GO identifier	LAP	LRT	rDT	uDT	LAP	LRT	rDT	uDT
GO:0005488	0.670	0.672	0.676	0.667	0.651	0.652	0.657	0.649
GO:0030246	0.369	0.180	0.171	0.371	0.325	0.240	0.308	0.326
GO:0003682	0.382	0.393	0.407	0.403	0.313	0.363	0.472	0.309
GO:0043167	0.124	0.113	0.075	0.128	0.130	0.095	0.091	0.127
GO:0008289	0.179	0.034	0.036	0.188	0.161	0.053	0.075	0.152
GO:0003676	0.511	0.528	0.535	0.512	0.485	0.500	0.510	0.483
GO:0000166	0.544	0.528	0.516	0.544	0.518	0.498	0.491	0.515
GO:0005515	0.561	0.564	0.568	0.559	0.544	0.547	0.552	0.541
GO:0046906	0.182	0.091	0.082	0.190	0.308	0.132	0.151	0.299
GO:0003824	0.017	0.017	0.016	0.020	0.016	0.015	0.016	0.011
GO:0016787	0.093	0.043	0.022	0.094	0.059	0.029	0.027	0.046
GO:0016874	0.606	0.612	0.609	0.606	0.56	0.559	0.568	0.557
GO:0016829	0.378	0.175	0.167	0.38	0.325	0.240	0.308	0.326
GO:0016491	0.262	0.227	0.166	0.262	0.226	0.180	0.175	0.222
GO:0016740	1.000	0.871	1.000	1.000	1.000	1.000	1.000	1.000
GO:0030234	0.372	0.182	0.172	0.375	0.325	0.240	0.308	0.326
GO:0030695	0.368	0.18	0.171	0.371	0.325	0.240	0.308	0.326
GO:0060089	0.374	0.178	0.168	0.378	0.325	0.240	0.308	0.326
GO:0004871	0.371	0.176	0.166	0.373	0.325	0.240	0.308	0.326
GO:0005198	1.000	0.876	1.000	1.000	1.000	1.000	1.000	1.000
GO:0005201	0.366	0.173	0.164	0.37	0.325	0.240	0.308	0.326
GO:0008307	0.377	0.382	0.393	0.394	0.313	0.363	0.472	0.309
GO:0030528	0.501	0.422	0.349	0.501	0.388	0.340	0.33	0.384
GO:0003702	0.364	0.180	0.168	0.366	0.325	0.240	0.308	0.326
GO:0003700	0.494	0.418	0.346	0.493	0.388	0.340	0.33	0.384
GO:0016564	0.373	0.180	0.169	0.376	0.325	0.240	0.308	0.326
GO:0005215	0.236	0.134	0.079	0.237	0.168	0.107	0.101	0.157
GO:0022892	0.064	0.008	0.009	0.055	0.073	0.012	0.019	0.062
GO:0022857	0.234	0.132	0.081	0.234	0.168	0.107	0.101	0.157

TABLE 7: Parametric bootstrap and asymptotic p-values for the subcategories of molecular function.

be more intuitive to use the subjects as the sampling units, as discussed by Goeman and Bühlman (2007). Indeed, when testing for equality of the positive predictive values of two diagnostic tests, the observational unit is the individual and the assumption of independence of test results between individuals is in most cases not seen to be problematic. But in the gene class setting the assumption does not hold, because genes act together in pathways and genes that are functionally related can be strongly correlated. If the gene expression measurements are correlated, the p-values tend to be positively correlated, see Goeman and Bühlman (2007). A possible extension of the methods developed in this report could be to look at different dependence structures between the observational units.

We have considered test statistics designed for comparing positive predictive values for diagnostic tests which translates to comparing association with GO categories for overlapping gene lists. Other possible approaches to handle overlapping gene lists include deleting the genes that are on both lists from each list or simply ignore the fact that there are genes that are on both lists and treat them as mutually exclusive lists. The deletion approach is implemented in the GO tool FatiGO, Al Shahrour, Diaz Uriarte and Dopazo (2004), in which Fisher's exact test is implemented. In the ignore approach Fisher's exact test or Pearson χ^2 test can be used. The asymptotic LAP test is implemented in eGOn which then handles the problem of overlapping gene lists more correctly than other GO-tools by not deleting the genes that are on both lists or ignore that the genes are overlapping.

In the simulation experiments in Section 4 and the example using data from the literature in Section 5, we see that the likelihood ratio and restricted difference tests yields similar results which differ from the LAP and unrestricted difference tests. The likelihood ratio and restricted difference test both use the maximum likelihood estimates for the parameters under the null hypothesis in addition to the general maximum likelihood estimates. The LAP and unrestricted difference also yield similar test results and these test statistics are functions of the observed data $\boldsymbol{n}$ and thereby the general maximum likelihood estimates only, and are thus not influenced by the maximum likelihood estimates under the null.

When considering small sample sizes, one or more of the cells in Table 1 have often zero counts which leads to non-computable test statistics for the LAP and difference tests. In these cases, we set the test statistics to 0 if the numerator is 0, implying that the null hypothesis will never be rejected for such tables. If only the denominator is 0, the test statistic is disregarded. For $N = 10$ there are 18 out of 3003 possible tables for which this will happen, while if $N = 15$ it will happen for 28 out of 15504 tables, for $N = 20$ for 38 of 53130 tables and for $N = 25$ for 48 of 142506 outcomes. For the likelihood ratio test statistic zero counts does not represent a problem, it can always be calculated. If any of the counts are 0, the summation term in (7) is also 0.

While the methodology in the present paper does not rely on asymptotic results, it is still approximative in the sense that it relies on simulations. Another shortcoming is that the test size is not preserved in general. Both these issues will be addressed in a forthcoming paper, Günther, Bakke, Rue and Langaas (2009), where enumeration rather than simulation will be applied to the testing method of the present paper and where it will be modified to yield p-values that preserve the test size. The test size and power will be calculated exactly.

7 Conclusions

In this report we look at the problem of testing the null hypothesis given in (3) when the sample size is small. The large sample tests using asymptotic distributions do not preserve their test size in this case and therefore small sample tests are needed. We suggest using parametric bootstrapping to approximate the distribution and to calculate the p-values. The likelihood ratio test and the restricted difference test are both functions of the maximum likelihood estimates $\boldsymbol{q}$ under the null hypothesis which may be difficult to find because of local optima. Especially zero counts causes problems, but our method that analytically solves the system of equations handles these problems well.

The simulation experiments show, at least based on the present six cases, that the small sample likelihood ratio test yields a smaller test size than the large sample likelihood ratio test, while for the restricted difference test the large sample test yields the smallest test size and is still conservative, thus for this test there was no improvement.

For testing whether there is a difference in enrichment or depletion of genes belonging to a certain GO category between two list of genes from a microarray experiment, there are several test statistics to choose from, and depending on the sample size, one can use either parametric bootstrapping or the asymptotic χ^2_1 distribution to calculate p-values. The choice of test statistic can influence which GO categories that are found to be significant and because the small sample parametric bootstrap likelihood ratio and restricted difference tests are more optimistic than the small sample parametric bootstrap LAP and unrestricted difference tests, the first two will yield more significant GO categories than the other two. The smaller the sample size is, the more conservative all the tests are which means they will not reject the null hypothesis even when it is not true., i.e. the tests will not discover gene classes that are over-represented or depleted on one list compared to the other list. Therefore, parametric bootstrapping does not seem to be an optimal solution and a better approach would probably be to use an exact small sample test that preserves its test size without being conservative, which will be investigated further.

REFERENCES

Al Shahrour, F., Diaz Uriarte, R. and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, *Bioinformatics* 20(4): 578–580.

Beisvåg, V., Jünge, F. K., Bergum, H., Jølsum, L., Lydersen, S., Günther, C.-C., Ramampiaro, H., Langaas, M., Sandvik, A. K. and Lægreid, A. (2006). GeneTools - application for functional annotation and statistical hypothesis testing, *BMC Bioinformatics* 7(470).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society* 57: 289–300.

Bye, A., Langaas, M., Høydahl, M. A., Kemi, O. J., Heinrich, G., Koch, L. G., Britton, S. L., Najjar, S. M., Ellingsen, Ø. and Wisløff, U. (2008). Aerobic capacity-dependent differences in cardiac gene expression., *Physiol Genomics* 33: 100–109.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, second edn, Duxbury, chapter 8.

Goeman, J. J. and Bühlman, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues., *Bioinformatics* 23(8): 980–987.

Günther, C.-C., Bakke, Ø., Lydersen, S. and Langaas, M. (2008). Comparison of predictive values from two diagnostic tests in large samples. Preprint Statistics No. 9, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Günther, C.-C., Bakke, Ø., Rue, H. and Langaas, M. (2009). Statistical hypothesis testing for categorical data using enumeration in the presence of nuisance parameters. Preprint Statistics No. 4, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Johnson, N. L., Kotz, S. and Balakrishan, N. (1997). *Discrete multivariate distributions*, Wiley series in probability and statistics, chapter 35.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* 56: 345–351.

Moskowitz, C. S. and Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs, *Clinical Trials* 3: 272–279.

R Development Core Team (2008). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
`http://www.r-project.org`

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology, *Nature Genetics* 25: 25–29.

Wang, W., Davis, C. S. and Soong, S.-J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares, *Statistics in Medicine* 25: 2215–2229.