

A Simulation Study of Sequential Testing Approaches in Genetics

Andres Azuero, David T. Redden, University of Alabama at Birmingham - Department of Biostatistics, Birmingham, AL

ABSTRACT

A common issue in case-control genetic association studies is that a large number of markers may be tested in the same case-control group, increasing the False Discovery Rate (FDR) due to multiple comparisons. Sobell (1993) proposed a three-staged sequential testing procedure where a large number of markers are screened by testing consecutively on three independent case-control subgroups. Sham (1994) argued that the use of independent subgroups at each stage, instead of accumulating data, decreases power. Thus, a sequential scheme was suggested, where at each stage the subsamples of previous stages are combined in order to maintain power. It was also suggested that sequential procedures based on Wald (1945) be applied to the problem. This study uses SAS to carry out the sequential procedures proposed by Sobell, Sham, and Wald's Sequential Probability Ratio Test (SPRT), and compares their properties in regards to power, percentage of genotyping, and FDR. The simulation results suggest that Wald's SPRT is the most efficient sequential scheme among those compared.

INTRODUCTION

A common problem in case-control genetic association studies is that a large number of markers are tested in the same case-control cohort. For instance, let us consider a cohort composed of two groups (case/control) with 900 subjects each, and 1000 gene expressions per subject. The objective is to find true associations between a disease and any of the gene expressions. From the point of view of statistics, these data bring about three problems:

- 1) Multiple testing: since we have 1000 gene expressions (also referred to as 'outcomes' throughout this paper) to test on the same cohort, if we don't account for multiplicity, we may increase the Type I error rate, and the observed False Discovery Rate (FDR), which is defined as the number of true null hypothesis rejected divided by the total number of hypothesis rejected.
- 2) Common multiple testing correction methods such as Bonferroni's correction, greatly decrease power, unless the sample size is increased
- 3) Large samples needed to maintain power after correction for multiplicity translate into higher genotyping costs. Ideally, we would like a testing scheme that maximizes power while minimizing FDR and genotyping costs.

SEQUENTIAL TESTING

In the past decade, sequential testing has been proposed as a procedure to address the problems mentioned above. The main characteristic of a sequential testing scheme is that the sample is divided into subsamples to be used at each stage of testing, and at the end of each stage, for a single hypothesis, one of the following three decisions is to be made:

- 1) Reject H_0
- 2) Fail to Reject (FTR) H_0
- 3) Inconclusive, test on next stage

We should note that in an "open" sequential testing procedure, if the maximum sample size is predetermined by a factor such as cost, it can be possible that at the last stage we might not be able to either reject or fail to reject a hypothesis.

Within this framework, Sobell-Schaid (1993,1994) proposed a sequential scheme where a sample size of 900 subjects per group is divided into 2 or 3 independent sub-samples to be used in 2 or 3 sequential stages, respectively. Each of the 1000 null hypotheses is of no association between disease and gene expression. In order to control for multiplicity, a 'step-wise' correction of Type I error was proposed in which, at each stage, the significance level, α , of each individual hypothesis test is divided by the number of hypotheses tested so far plus one. Thus, with an overall α of 0.05, the first hypothesis regarding the first gene expression is tested at the 0.05 significance level, while the 1000th hypothesis is tested at the 0.05/1000 significance level. We should note that with this proposed correction method, the hypotheses tested at the beginning have a less stringent significance level than the ones tested at the end, and thus rejection of a hypothesis depends also on the order in which the hypotheses are tested. A gene expression is declared significant if its t-test p-value is smaller than its significance level; it's declared suggestive if its p-value is between its significance level and 0.1; and it's declared not significant if its p-value is larger than 0.1. The suggestive outcomes from the current stage are then re-tested on the following stage.

As a commentary to Sobell's paper, Sham (1994) proposed a 'joint' approach. Instead of using independent subsamples at each stage, it was proposed to accumulate and use data from all prior stages. Power should remain fixed across all stages, and at each stage α may be decreased as data accumulates. Sham also recommended applying Wald's Sequential Probability Ratio Test (SPRT) to the problem.

More recently, Skol (2006) simulating a related problem but at a larger scale (genome-wide) and while not considering the problem of multiple testing, recommended a 2-stage joint approach.

WALD'S SPRT

Wald's SPRT (1945) is a 'joint', subject by subject, Likelihood Ratio Test (LRT). In this approach, each subject constitutes a 'stage'. The hypothesis test is of the form:

H_0 : the distribution of the sample $\underline{X}$ is $p_0(\underline{x})$, vs. H_A : the distribution of $\underline{X}$ is $p_1(\underline{x})$

Then, at each stage (at the m-th stage), calculate

$$\frac{p_{1m}}{p_{0m}} = \frac{p_{H1}(x_1, \dots, x_m)}{p_{H0}(x_1, \dots, x_m)}$$

If $p_{1m} = p_{0m} = 0$ then p_{1m} / p_{0m} is defined to be equal to 1.

Reject H_0 (conclude H_A) if:

$$\frac{p_{1m}}{p_{0m}} \geq A(\alpha, \beta); \quad A(\alpha, \beta) = \frac{1 - \beta}{\alpha}$$

FTR H_0 if:

$$\frac{p_{1m}}{p_{0m}} \leq B(\alpha, \beta); \quad B(\alpha, \beta) = \frac{\beta}{1 - \alpha}$$

Take an additional observation if:

$$B(\alpha, \beta) < \frac{p_{1m}}{p_{0m}} < A(\alpha, \beta), \text{ or, } \log(B) < \log \frac{p_{1m}}{p_{0m}} < \log(A)$$

The quantities A and B are limiting values, i.e.

$$B \geq \frac{\beta}{1 - \alpha}; \quad A \leq \frac{1 - \beta}{\alpha}$$

Our study aims to compare the properties of the sequential procedures proposed by Sobell-Schaid, Sham, and Wald's SPRT when applied to the problem of testing a large number of markers in the same case-control cohort, in regards to Power, FDR, and genotyping expenses; and to suggest improvements to the methods, if possible.

METHODS

The simulation was carried out using SAS v.9.1.

THE DATASET

The basic simulated data set consists of the following:

Two groups (case/control), 900 subjects per group. For the Control group, each subject has 1000 gene expressions or outcomes, modeled as independent identically distributed (iid) Standard Normal variables i.e. $N(0,1)$. For the Case group, each subject has 950 gene expressions modeled as iid $N(0,1)$, and 50 gene expressions modeled as iid $N(0.2,1)$ per subject. So we have 50 true differences (of 20% of a standard deviation), out of 1000, by design.

The basic dataset is made in two DATA steps using do - end loops and functions Call Streaminit (seed), and Rand ('Normal'). An example of the algorithm using SAS code is the following:

```
*creates two groups, each of 900 subjects;
data subjects;
    do group = 1 to 2;
        do i = 1 to 900;
            output;
        end;
    end;
run;

*creates 1000 outcomes for each subject;
*group 1 has all 1000 outcomes iid N(0,1);
*group 2 has first 950 outcomes N(0,1) and last 50 outcomes iid N(0.2,1);
data subjects2;
call streaminit(11131973);
set subjects;
array outcomes (*) o0001-o1000;
```

```

do j = 1 to 1000;
  if j <= 950 then do;
    outcomes(j) = rand('normal');
  end;
  if j > 950 then do;
    if group = 1 then outcomes(j) = rand('normal');
    if group = 2 then outcomes(j) = .2+rand('normal');
  end;
end;
run;

```

TESTING METHODS COMPARED

The following six methods were compared:

- 1) T-tests using the full dataset (uncorrected for multiple testing, and Bonferroni corrected)
 - 2) T-tests using Sobell-Schaid proposed sequential method, with 2, and 3 stages
 - 3) T-tests using Sham proposed sequential method, 2 to 5 stages
- (As mentioned in the Sequential Testing section above, the multiple correction method proposed by Sobell-Schaid depends on the order in which the gene expressions are tested. So, trying to use a correction method not dependent of the testing order we also run Sobell-Schaid and Sham methods, but ranking the outcomes by the absolute value of the mean difference between groups before obtaining the t-values from the t-tests and then applying a 'step-wise' correction. A larger mean difference is likely to give a larger t-value (and thus a smaller p-value), so this ranking approach is similar to Holm's correction, which ranks the p-values after the tests have been carried out.)
- 4) T-tests using Sobell-Schaid method, 'ranked', 2, and 3 stages
 - 5) T-tests using Sham's method, 'ranked', 2 to 5 stages
 - 6) Wald's SPRT uncorrected, and Bonferroni Corrected, using a $N(0,1)$ probability distribution function.

FULL DATASET APPROACH

This is the easiest case. We use PROC TTEST to test the 1000 outcomes by group. It helps to output the p-values from the tests to a data set, by means of an ODS statement, in order to select those outcomes that are significant according to different multiple correction methods. We are not interested in other output produced by the procedure, so we have to turn the regular listing output off. An example SAS code is the following:

```

ods listing close;
ods output ttests=p_alldata;
proc ttest data = subjects2;
  class group;
  var 00001-01000;
run;
ods listing;

```

The resulting dataset has, among others, two variables that correspond to the outcome names and the t-test p-values. For the uncorrected method, we select the outcomes whose test p-value is smaller than 0.05. For the Bonferroni corrected method, we select the outcomes whose test p-value is smaller than 0.05/1000, since we are testing 1000 outcomes.

SOBELL-SCHAIID'S APPROACH

Since we are going to have 2 or 3 independent stages of testing, we need to create another index variable in the first loop of the dataset-creating algorithm. This index is simply used to divide the dataset into 2 or 3 subsets, corresponding to 2 or 3 independent testing "stages". An example SAS code for 2 stages is the following:

```

data Sobell2stage;
  do group = 1 to 2;
    do i = 1 to 900;
      if i<=450 then sob_stage=1;
      if i>450 then sob_stage=2;
      output;
    end;
  end;
run;

```

The second loop remains the same as the one shown above in the Dataset section. Once we have created the complete dataset, we want to select the outcomes that are significant in the first stage. To this end we run PROC TTEST, but to use only the observations corresponding to first stage we add the line:

```

where sob_stage = 1;

```

to the body of the PROC TTEST statement. We output the p-values to a dataset using an ODS statement (similar to the code shown above in the Full Dataset Method paragraph). From this resulting dataset that contains variables corresponding to the outcome names and t-test p-values, the next step is to determine what outcomes are significant, suggestive and not significant, considering a 'step-wise' correction method. For the original approach

suggested by Sobell and Schaid, we first need to add a variable containing the 'step-wise' significance levels for each of the outcomes (i.e. for each of the hypotheses tested): for the first outcome (first hypothesis), the significance level is 0.05, for the nth outcome (nth hypothesis), the significance level is 0.05/n. This is done with a short DATA step:

```
data stage1;
set ttests1;
hypnum=_N_;
alpha=0.05/hypnum;
run;
```

We now can print the names of the significant outcomes from this stage, discard the not significant outcomes and continue testing the suggestive outcomes only, in the next stage. Clearly, the significance level depends on the order the outcomes are tested.

For the ranked approach, we need first to sort the p-values dataset by the absolute value of the mean difference between case and control groups. To obtain these differences, we need to add another ODS statement before PROC TTEST:

```
ods output ttests= p_sobstage1;
ods output Statistics=differences1;
proc ttest data = Sobell2stage2;
(...)
```

The second resulting dataset contains, among others, a variable with the names of the outcomes and a variable with the mean difference between groups for each outcome. We need the absolute value of these differences, which can be easily calculated in a DATA step using the `abs(variable_name)` function. Now after some cleaning of the second resulting dataset we merge it with the p-values resulting set, by outcome name. Next we sort descending the combined dataset by absolute value of mean difference, using PROC SORT, and then we can add the 'step-wise' significance levels as shown above. We now can print the names of the significant outcomes from this stage, discard the not significant outcomes and continue testing the suggestive outcomes only, in the next stage.

For each subsequent stage, we need to start with a dataset that contains only the suggestive outcomes from the prior stage and the subjects allocated to that stage. To obtain this dataset, we start with two datasets: the first one contains two variables, one with the name of the suggestive outcomes from the prior stage and another with the p-values from that prior stage. The other dataset is simply a subset of the main dataset, but with the subjects allocated only to that stage. These two datasets are easily done with two DATA steps. The algorithm is the following:

- 1) Transpose the subjects' dataset so that now the outcomes are the observations and the subjects are the variables (PROC TRANSPOSE)
- 2) Merge the transposed dataset with the prior suggestive outcomes dataset, by outcome (DATA step).
- 3) Delete the 'observations' (i.e. the outcomes) whose p-value is missing, since this means that they are not part of the prior suggestive outcomes (DATA step)
- 4) Transpose the merged dataset so that the outcomes become variables again and the subjects become the observations (PROC TRANSPOSE)

Now we have a dataset with suggestive outcomes from prior stage and subjects allocated only to that stage, and we can run PROC TTEST again, and select significant, suggestive, and not significant outcomes.

SHAM'S APPROACH

The programming required to apply Sham's suggested method is very similar to that of Sobell-Schaid's suggested approach. However, there are two main differences. First, each stage uses the subjects from the prior stages, as well as the subjects allocated to that particular stage; these combined datasets for each stage can be easily obtained in a DATA step. Second, since the desired statistical power is to remain fixed across stages, the overall significance level at each stage is different because the sample size increases at each stage. The significance level can be calculated using PROC POWER. For instance, for a desired 90% power to detect a 20% mean difference between two Normally distributed variables, and a combined sample size of 300 subjects per group, the SAS code to obtain the significance level is the following:

```
Proc power;
TWOSAMPLEMEANS
MEANDIFF=0.2
STDDEV=1
NPERGROUP=300
POWER=0.9
ALPHA=
;
run;
```

In this example the overall significance level is 0.243.

After running PROC TTEST for a stage, and once we have calculated the overall significance level for that stage, we can add the individual 'step-wise' significance levels for each hypothesis by means of a DATA step as shown in the Sobell-Schaid's section above. From there we can select significant, suggestive, and not significant outcomes in the following way: an outcome is declared significant if its t-test p-value is less than its respective significance level; an

outcome is declared suggestive if its p-value is between its individual significance level and the overall significance level for that stage; and an outcome is declared not significant if its p-value is larger than the overall significance level for that stage. The 'ranked' approach is programmed similarly as shown above in the Sobell-Schaid's Approach section.

WALD'S SPRT

Since Wald's SPRT is a likelihood ratio test (i.e. we calculate the ratio of the likelihood function of a sample under the null and under the alternative hypotheses) we only need the case group. To facilitate calculations, we use the natural logarithm of the ratio of likelihood functions. Each null hypothesis is that the sample is iid $N(0,1)$, the alternative is that the sample is iid $N(0.2,1)$. For simulation purposes, the algorithm that creates the basic dataset needs to be modified, to include for each observation, the values of the Null and Alternative Normal probability distribution functions, their natural logarithms, and the natural logarithms of the null and alternative likelihood functions of that observation and all observations prior to it. As with the basic dataset, the first step is a short loop that creates the 900 subjects per group. The second loop now includes calculating the natural logarithm of the value of the Normal probability function under the null and under the alternative hypotheses, as follows:

```
data wald2_1;
call streaminit(11131973);
set wald;
array outcomes (*) o0001-o1000;
array LpdfNull (*) LpN0001 - LpN1000;
array LpdfAlt (*) LPA0001 - LPA1000;
do j = 1 to 1000;
    if j <= 950 then do;
        outcomes(j) = rand('normal');
    end;
    if j > 950 then do;
        if group = 1 then outcomes(j) = rand('normal');
        if group = 2 then outcomes(j) = .2+rand('normal');
    end;
    LpdfNull(j)=log(PDF('NORMAL',outcomes(j),0,1));
    LpdfAlt(j)=log(PDF('NORMAL',outcomes(j),0.2,1));
end;
drop j;
run;
```

Next we delete the control observations, since they are not needed (note that it's necessary to create them so that the outcome values remain unchanged compared to the original basic dataset used in the other approaches).

The next step is to calculate for each observation, the natural logarithm of the likelihoods of that observation and all observations prior to it. Since we are using logarithms, the $\log(\text{likelihood})$ function is a sum, as opposed as the likelihood function, which is a product. And the likelihood ratio becomes a subtraction of $\log(\text{likelihoods})$, i.e:

$$\log \frac{p_{1m}}{p_{0m}} = \log(p_{1m}) - \log(p_{0m})$$

In order to calculate the $\log(\text{likelihoods})$, and the $\log(\text{likelihood})$ 'ratio' up to an observation, we use a DATA step with RETAIN statement, as follows:

```
data loglikes;
set wald2;
array LLN (*) LLN0001 - LLN1000;
array LLA (*) LLA0001 - LLA1000;
array LpdfNull (*) LpN0001 - LpN1000;
array LpdfAlt (*) LPA0001 - LPA1000;
array ratios (*) ratio0001 - ratio1000;
do j=1 to 1000;
    Retain LLN 0 LLA 0;
    LLN(j)=Lpdfnull(j) + LLN(j);
    LLA(j)=LpdfAlt(j) + LLA(j);
    ratios(j)=LLA(j)-LLN(j);
end;
drop j;
run;
```

Next, we need to include the natural logarithms of the decision limits. This is easily done with a DATA step. For instance, for a Bonferroni-corrected significance level, the following code adds the $\log(\text{limits})$:

```
data cases;
set loglikes;
alpha=0.00005;
```

```

power=0.95;
lnA=log(power/alpha);
lnB=log((1-power)/(1-alpha));
run;

```

Now we need to determine which likelihood ratios cross the significance (or not-significance limits) and at which observation number. This can be done with indicator variables (1 if the limit is crossed, 0 if not) in a DATA statement. Ideally, after each observation we are supposed to assess significance, however for the purpose of this exercise and since we have 1000 outcomes to evaluate, we can assess significance in groups of 100 observations.

RESULTS

The following 3 tables summarize the results of the simulations in regard to observed power, observed FDR, number of suggestive or inconclusive outcomes after the final stage, and percentage of genotyping:

Method	Observed Power (%)	Observed FDR (%)
Full Uncorrected	100.0	48.98
Full Bonferroni Corrected	60.0	0.00
Sobell Schaid 2 stages	28.0	0.00
Sobell Schaid 3 stages	24.0	0.00
Sobell Schaid ranked 2 stages	68.0	5.56
Sobell Schaid ranked 3 stages	58.0	6.45
Sham 2 stages	46.0	0.00
Sham 3 stages	62.0	13.89
Sham 4 stages	66.0	17.50
Sham 5 stages	64.0	21.95
Sham ranked 2 stages	82.0	8.89
Sham ranked 3 stages	86.0	23.21
Sham ranked 4 stages	86.0	36.76
Sham ranked 5 stages	84.0	47.50
Wald's SPRT uncorrected	100.0	27.54
Wald's SPRT Bonferroni Corrected	98.0	0.00

Table 1. Simulation results by Observed Power and Observed False Discovery Rate (FDR)

Method	Inconclusive or suggestive (n)
Full Uncorrected	0
Full Bonferroni Corrected	0
Sobell Schaid 2 stages	39
Sobell Schaid 3 stages	20
Sobell Schaid ranked 2 stages	20
Sobell Schaid ranked 3 stages	6
Sham 2 stages	25
Sham 3 stages	16
Sham 4 stages	11
Sham 5 stages	12
Sham ranked 2 stages	6
Sham ranked 3 stages	2
Sham ranked 4 stages	0
Sham ranked 5 stages	0
Wald's SPRT uncorrected	0
Wald's SPRT Bonferroni Corrected	0

Table 2. Simulation results: Number of suggestive or inconclusive outcomes after the last stage of testing.

Method	Genotyping (%)
Full Uncorrected	100.0
Full Bonferroni Corrected	100.0
Sobell Schaid 2 stages	57.6
Sobell Schaid 3 stages	39.3
Sobell Schaid ranked 2 stages	56.6
Sobell Schaid ranked 3 stages	38.2
Sham 2 stages	61.6
Sham 3 stages	51.5
Sham 4 stages	46.9
Sham 5 stages	43.3
Sham ranked 2 stages	60.2
Sham ranked 3 stages	49.6
Sham ranked 4 stages	44.5
Sham ranked 5 stages	40.4
Wald's SPRT uncorrected	13.0
Wald's SPRT Bonferroni Corrected	14.5

Table 3. Simulation results by percentage of genotyping

The question now is: what is the most efficient method? Ideally we would like a method that results in high power, low FDR, low number of inconclusive outcomes after the final stage, and low genotyping. If we define Cost Efficiency (CE) as:

$$CE = \frac{Power(\%) - FDR(\%) - Inconclusives(\%)}{Genotyping(\%)}$$

We obtain results in terms of efficiency (Table 4):

Method	Cost Efficiency
Full Uncorrected	0.5
Full Bonferroni Corrected	0.6
Sobell Schaid 2 stages	0.4
Sobell Schaid 3 stages	0.6
Sobell Schaid ranked 2 stages	1.1
Sobell Schaid ranked 3 stages	1.3
Sham 2 stages	0.7
Sham 3 stages	0.9
Sham 4 stages	1.0
Sham 5 stages	0.9
Sham ranked 2 stages	1.2
Sham ranked 3 stages	1.3
Sham ranked 4 stages	1.1
Sham ranked 5 stages	0.9
Wald's SPRT uncorrected	5.6
Wald's SPRT Bonferroni Corrected	6.8

Table 4. Simulation results by Cost-Efficiency

The simulation results suggest that Wald's SPRT is by far the most efficient method of those compared.

CONCLUSION

Wald's SPRT appears to be the most efficient method. Even when used with a Bonferroni correction, the decrease in power (given the assumptions of this study) is minimal, while it corrects for the 28% FDR of the uncorrected SPRT. However, in order to apply the SPRT, distributional assumptions must be made. Future research will be conducted in order to study the robustness of the SPRT.

Other multiple test correction procedures should be considered, in particular, alpha spending functions used in clinical trials.

This study deals with a very simple model. More levels of complexity in regards to the genetic model should be added, as well as the stratification of the population the sample comes from. Finally, this study deals with only 1000 gene expressions. The properties of the test should be studied in the presence of a larger number of genes (in the order of 300,000) such as in genome-wide data.

REFERENCES

- Schaid DJ, Sommer SS.** (1994). Need to Confirm Promising Case-Control Association Studies: Reply to Sham. *Am J Med Genet*, 54, 156-157.
- Sham P.** (1994). Sequential analysis and case-control candidate gene association studies: reply to Sobell et al. *Am J Med Genet*. 54(2):154-7.
- Skol A. et al.** (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 38(2):209-13.
- Sobell JL, Heston LL, Sommer SS.** (1993). Novel association approach for determining the genetic predisposition to schizophrenia: case-control resource and testing of a candidate gene. *Am J Med Genet*. 48(1):28-35.
- Wald A.** (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, Vol. 16, No. 2 pp. 117-186.

CONTACT INFORMATION

Andres Azuero, Graduate Student, Dept. of Biostatistics
University of Alabama at Birmingham
RPHB 327
1530 3rd Ave S.
Birmingham, AL 35294-0022
Email: andreo@uab.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.