

Power calculations for 2 by m tables in one and two stage unmatched case-control designs

József Bukszár and J. C. G van den Oord



Outline

- Large-sample approximation for Pearson's statistic applied to case-control studies
- Data pooling for two stage design



Introduction

- Case-control studies are one of the most widely applied research designs in medicine.
- In its classical form, the subjects are sampled separately from case and control populations.
- Currently one of the most important tools for mapping the genetic determinants of complex human diseases
- This popularity can be explained by the many practical advantages of the classical case-control study such as low costs and the ease of collecting large samples.



Need approximation?

- Data are often arranged in $2 \times m$ tables
- Exact distribution is not possible for large n
- No other approximation for Pearson and larger tables than 2×2
- Need good distribution for
 - design (power)
 - data pooling etc

Contingency table:

$$\begin{array}{c} \left[\begin{array}{ccc} x_1 & \cdots & x_m \\ y_1 & \cdots & y_m \end{array} \right] \begin{array}{l} \leftarrow \text{ case group} \\ \leftarrow \text{ control group} \end{array} \\ \uparrow \quad \quad \uparrow \\ \text{Categories (e.g. A/A, A/a, a/a)} \end{array}$$

Total sample size: n

$$x_1 + \dots + x_m = np$$

$$y_1 + \dots + y_m = nq$$

$$p + q = 1$$

Pearson's statistic

$$P = \sum_{i=1}^m \left[\frac{\left(x_i - \frac{(x_i + y_i)pn}{n} \right)^2}{\frac{(x_i + y_i)pn}{n}} + \frac{\left(y_i - \frac{(x_i + y_i)qn}{n} \right)^2}{\frac{(x_i + y_i)qn}{n}} \right] = \sum_{i=1}^m \frac{(qx_i - py_i)^2}{pq(x_i + y_i)}$$

P follows a discrete distribution (when $x_i \Leftrightarrow X_i$ and $y_i \Leftrightarrow Y_i$)

How can the distribution of P be approximated?

Standard Approximation (SA)

Let p_i be the probability that a randomly chosen case falls into category i and q_i be the probability that a randomly chosen control falls into category i .

Perform Pearson's statistic on the contingency table

$$\begin{bmatrix} pp_1 & \cdots & pp_m \\ qq_1 & \cdots & qq_m \end{bmatrix}$$

to obtain

$$C = pq \sum_{i=1}^m \frac{(p_i - q_i)^2}{(pp_i + qq_i)}.$$

Standard approximation is the chi-square random variable with non-centrality parameter Cn and $m - 1$ degree of freedom.

Under the null hypothesis, i.e. when $p_i = q_i$ for all $i = 1, \dots, m$, SA is the (central) chi-square with $m - 1$ degree of freedom.

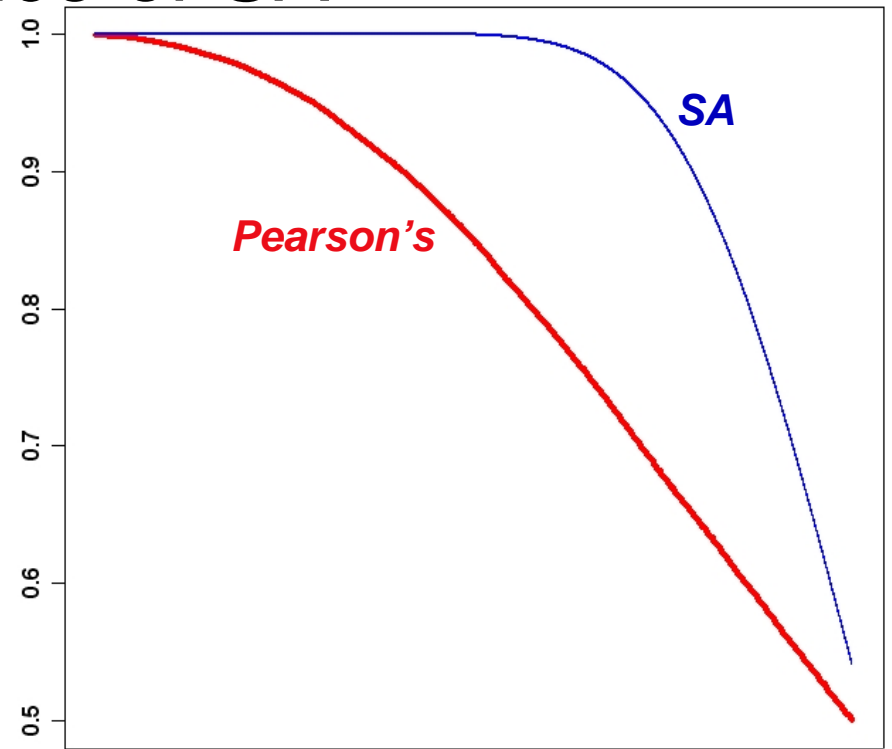
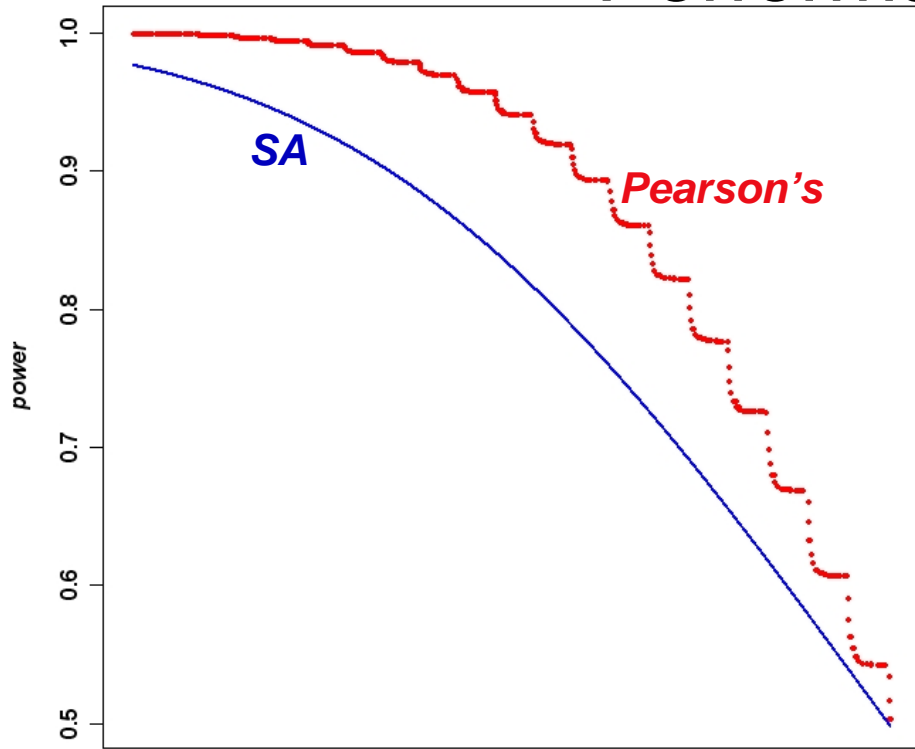


Why called standard Approximation?

The method:

- Is described in classic works on power analyses (Cohen, 1988) and categorical data analysis (Agresti, 1990),
- Appears in text books (Weir, 1996)
- Is often implemented in computer programs such as Genetic Power Calculator (Purcell, Cherny, Sham, 2003)
- Has become so well established that it is sometimes even used without reference (Akey et al. 2001; Van den Oord 1999).

Performance of SA



critical value

probability tables

critical value

$$\begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}$$

$$\begin{bmatrix} 0.46 & 0.41 & 0.13 \\ 0.49 & 0.50 & 0.01 \end{bmatrix}$$

case and control sample sizes

$$np = nq = 200$$

$$np = 100, nq = 9900$$

Asymptotic equivalent

Pearson's statistic:
$$P = \sum_{i=1}^m \frac{(qX_i - pY_i)^2}{pq(X_i + Y_i)},$$

where X_1, \dots, X_m and Y_1, \dots, Y_m are independent random variables with $M(p_1, \dots, p_m; pn)$ and $M(q_1, \dots, q_m; qn)$, respectively.

The idea is to examine the asymptotic behavior of

$$\left(\frac{qX_1 - pY_1}{\sqrt{pq(X_1 + Y_1)}}, \dots, \frac{qX_m - pY_m}{\sqrt{pq(X_m + Y_m)}} \right)$$

Theorem: Let T_n , $n = 1, 2, \dots$, be a sequence of k -dimensional statistics such that

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \Sigma)$$

Let $g : \mathbf{R}^k \rightarrow \mathbf{R}^m$ be a differentiable function.

Then
$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, g'(\theta)\Sigma g'^T(\theta)).$$

Let

$$T_n = \left(\frac{X_1}{pn}, \dots, \frac{X_m}{pn}, \frac{Y_1}{qn}, \dots, \frac{Y_m}{qn} \right) \text{ and } \theta = E(T_n) = (p_1, \dots, p_m, q_1, \dots, q_m).$$

By a suitably chosen $g : \mathbf{R}^{2m} \rightarrow \mathbf{R}^m$ we have

$$\sqrt{n}g(T_n) = \left(\frac{qX_1 - pY_1}{\sqrt{pq}(X_1 + Y_1)}, \dots, \frac{qX_m - pY_m}{\sqrt{pq}(X_m + Y_m)} \right).$$

By applying the above theorem we obtain that

$$\left(\frac{qX_1 - pY_1}{\sqrt{pq} \sqrt{X_1 - Y_1}}, \dots, \frac{qX_m - pY_m}{\sqrt{pq} \sqrt{X_m - Y_m}} \right) \stackrel{D}{\sim} N(0, J)$$

where the entries of matrix J are

$$J_{ij} = \begin{cases} \frac{1}{\sqrt{p_i p_j}} \left[\left(1 - \frac{p_i q_i}{2 p_i q_i}\right) \left(1 - \frac{p_j q_j}{2 p_j q_j}\right) \right] p_i q_j & \text{if } i \neq j \\ \frac{1}{p_i} \left[\left(1 - \frac{p_i q_i}{2 p_i q_i}\right)^2 p_i - \left(1 - \frac{p_i q_i}{2 p_i q_i}\right)^2 q_i \right] & \text{if } i = j. \end{cases}$$

Under the null hypothesis:

$$J_{ij} = \begin{cases} -\sqrt{p_i p_j} & \text{if } i \neq j \\ 1 - p_i & \text{if } i = j. \end{cases}$$

We obtained that

$$\left(\frac{qX_1 - pY_1}{\sqrt{pq(X_1 + Y_1)}}, \dots, \frac{qX_m - pY_m}{\sqrt{pq(X_m + Y_m)}} \right)$$

is asymptotically equivalent with

$$(U_1, \dots, U_m) = \mathbf{U} \sim N(\boldsymbol{\mu}, \mathbf{J})$$

where

$$\boldsymbol{\mu}^T = \sqrt{n} g(\boldsymbol{\theta}) = \left(\frac{(p_1 - q_1)\sqrt{pqn}}{\sqrt{pp_1 + qq_1}}, \dots, \frac{(p_m - q_m)\sqrt{pqn}}{\sqrt{pp_m + qq_m}} \right).$$

Pearson's statistics is asymptotically equivalent with $\sum_{i=1}^m U_i^2$.

The components U_1, \dots, U_m are dependent random variables.

Our aim is to find *independent* normal r. variables Z_1, \dots, Z_m

with

$$\sum_{i=1}^m Z_i^2 = \sum_{i=1}^m U_i^2.$$

Since J is a symmetric positive semi-definite matrix, it has an eigen decomposition

$$J = A^T D A,$$

where A is an orthogonal matrix whose rows are the eigenvectors of J , and D is a diagonal matrix with J 's eigenvalues in its diagonal.

Let

$$(Z_1, \dots, Z_m) = \mathbf{Z} = A\mathbf{U}.$$

Then

$$\sum_{i=1}^m Z_i^2 = \mathbf{Z}^T \mathbf{Z} = \mathbf{U}^T A^T A \mathbf{U} = \mathbf{U}^T \mathbf{U} = \sum_{i=1}^m U_i^2$$

and

$$\text{Cov}(\mathbf{Z}) = A \text{Cov}(\mathbf{U}) A^T = A J A^T = A A^T D A A^T = D.$$

Thus, Z_1, \dots, Z_m are independent normal random variables and their variances are the eigenvalues of J .

Cost-effective approximation

$$\text{Pearson's statistics} \sim \sum_{i=1}^m Z_i^2 \text{ (AE).}$$

The right-hand side can be approximated by the sum of squares of $m - 1$ independent normal random variables (cost-effective appr.)

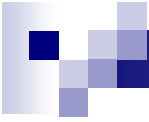
The computation cost of the cost-effective approximation (CE) is significantly less than that of AE when m is small, e.g. 2 or 3.

Furthermore we shall see that CE is almost as accurate as AE.

The idea is based on that the variance and expected value of a component of \mathbf{Z} is small.

Particularly, one of the eigenvectors of \mathbf{J} is approximately

$$\mathbf{v} \approx \left(\sqrt{p_1 p_{-1} q_1}, \dots, \sqrt{p_m p_{-m} q_m} \right) \text{ with eigenvalue close to 0.}$$



Intuitively, it can be seen as follows. Matrix J

$$J_{ij} = \begin{cases} \frac{1}{\sqrt{p_i p_{-i|q}} \sqrt{p_j p_{-j|q}}} \left[\left(1 - \frac{p_i q_i}{2 p_i p_{-i|q}}\right) \left(1 - \frac{p_j q_j}{2 p_j p_{-j|q}}\right) p q_i q_j - \left(1 - \frac{p_i q_i}{2 p_i p_{-i|q}}\right) \left(1 - \frac{p_j q_j}{2 p_j p_{-j|q}}\right) \right] q p_i p_j & \text{if } i \neq j \\ \frac{1}{p_i p_{-i|q}} \left[\left(1 - \frac{p_i q_i}{2 p_i p_{-i|q}}\right)^2 p q_i - \left(1 - \frac{p_i q_i}{2 p_i p_{-i|q}}\right)^2 q p_i - p_i \right] & \text{if } i = j. \end{cases}$$

can be well approximated by matrix G

$$G_{ij} = \begin{cases} \frac{1}{\sqrt{p_i p_{-i|q}} \sqrt{p_j p_{-j|q}}} p q_i q_j - q p_i p_j & \text{if } i \neq j \\ \frac{1}{p_i p_{-i|q}} p q_i - q p_i - p_i & \text{if } i = j. \end{cases}$$

Furthermore, one of the eigenvectors of G is

$$\mathbf{v} = \left(\sqrt{p_1 p_{-1|q}}, \dots, \sqrt{p_m p_{-m|q}} \right) \quad \text{with eigenvalue } 0.$$

We can assume that the eigenvector of J which is close to \mathbf{v} is the last row of A , that is A_m .

Then

$$E(Z_m) = E(A_m \cdot \mathbf{U}) = A_m \cdot E(\mathbf{U}) = A_m \cdot \boldsymbol{\mu} \approx \mathbf{v} \boldsymbol{\mu} = 0,$$

and the variance of Z_m is close to 0.

Therefore, Z_1^2, \dots, Z_m^2 can be well approximated by

$$Z_1^{\diamondsuit}, \dots, Z_m^{\diamondsuit},$$

where $Z_1^{\diamondsuit}, \dots, Z_m^{\diamondsuit}$ are independent normal random variables with

$$\text{Var}(Z_i^{\diamondsuit}) = \text{Var}(Z_i)$$

$$E^2(Z_i^{\diamondsuit}) = E^2(Z_i) = \frac{1}{m-1} E^2(Z_m)$$

Standard Approximation

Standard approximation can be obtained from the cost-effective approximation by setting the variances of components to 1, i.e. standard approximation is

$$Z_1''^2 + \dots + Z_{m-1}''^2,$$

where Z_1'', \dots, Z_{m-1}'' , are independent normal random variables with

$$\text{Var}(Z_i'') = 1,$$

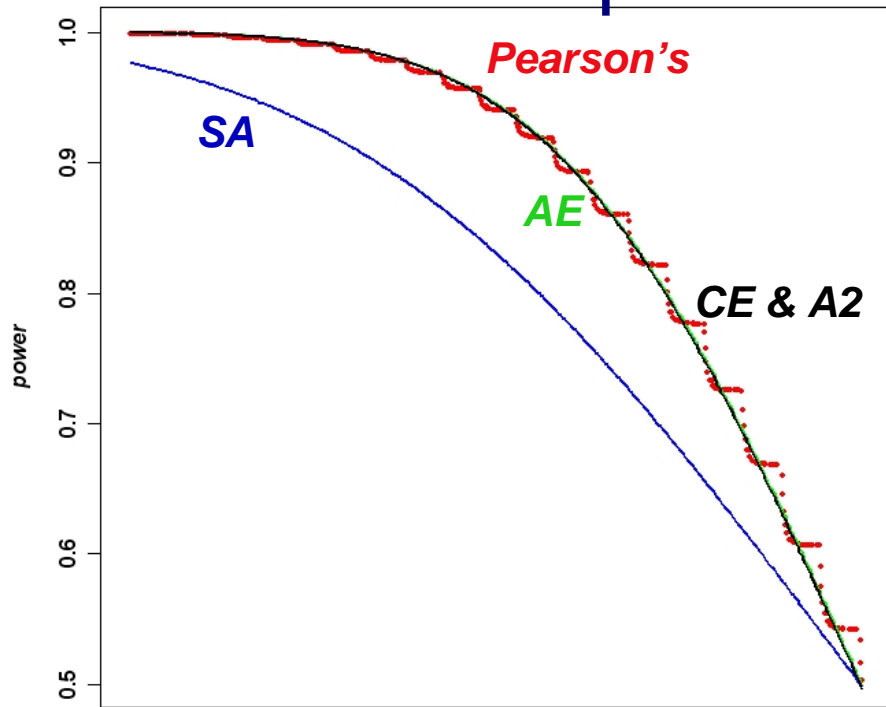
$$E^2(Z_i'') = E^2(Z_i) + \frac{1}{m-1} E^2(Z_m).$$

It is the chi-square random variable with non-centrality parameter

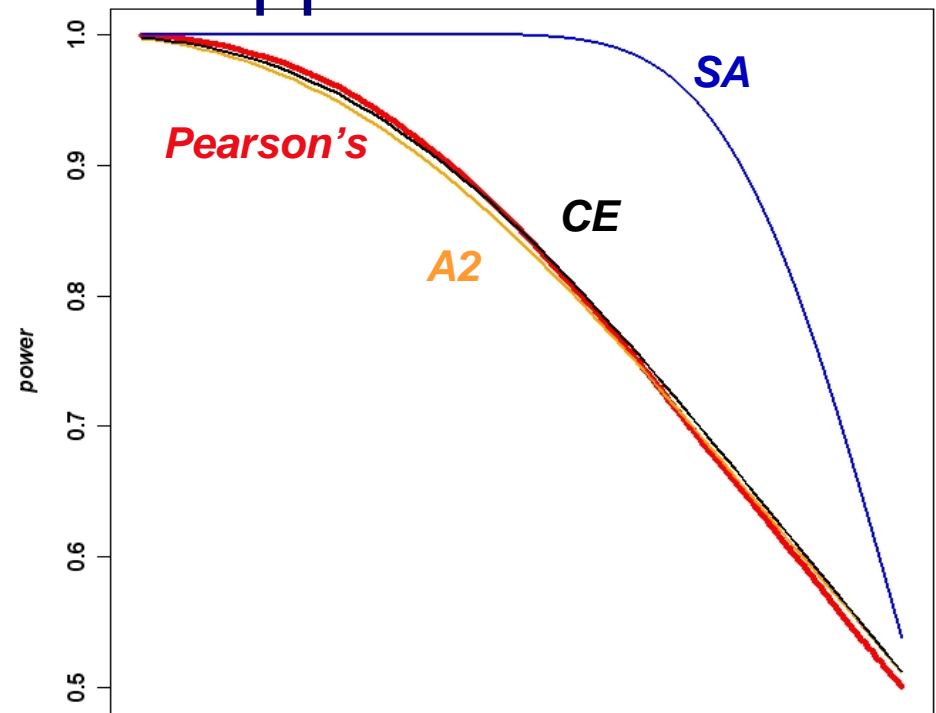
$$\sum_{i=1}^m E^2(Z_i)$$

and $m - 1$ degree of freedom.

Comparison of the approximations



critical value



critical value

probability tables

$$\begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}$$

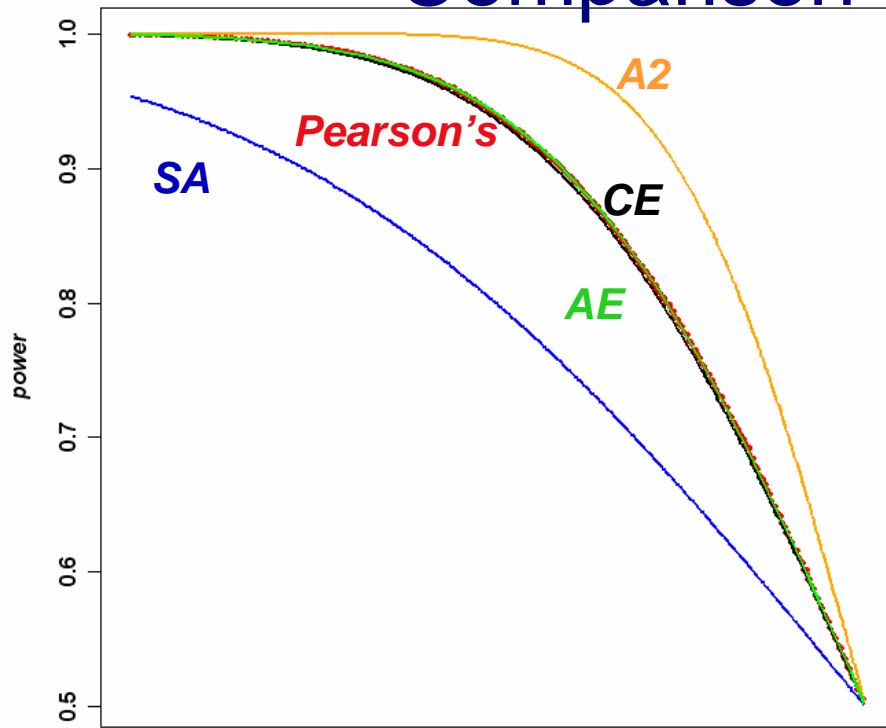
$$\begin{bmatrix} 0.46 & 0.41 & 0.13 \\ 0.49 & 0.50 & 0.01 \end{bmatrix}$$

case and control sample sizes

$$np = nq = 200$$

$$np = 100, nq = 9900$$

Comparison of the approximations



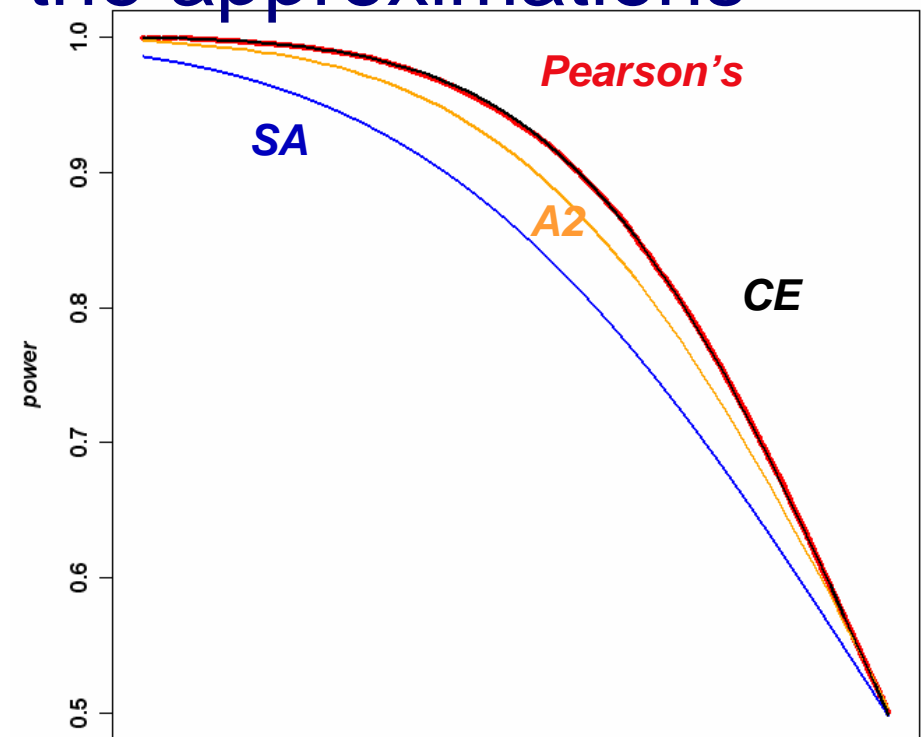
critical value

probability tables

$$\begin{bmatrix} 0.01 & 0.99 \\ 0.99 & 0.01 \end{bmatrix}$$

case and control sample sizes

$$np = 1000, nq = 10000$$



critical value

$$\begin{bmatrix} 0.44 & 0.43 & 0.13 \\ 0.04 & 0.01 & 0.95 \end{bmatrix}$$

$$np = 2000, nq = 8000$$

np, nq	q_1	.1	.3	.6	.9		
100		.0086	.0085	.0061	.0273	←	A2
		.0080	.0072	.0050	.0058	←	CE
		.0422	.0913	.0971	.0382	←	SA
		.0080	.0071	.0049	.0029	←	AE
300		.0043	.0175	.0083	.0116		
		.0035	.0031	.0039	.0052		
		.0215	.0630	.0741	.0873		
		.0035	.0031	.0019	.0028		
500		.0028	.0209	.0193	.0033		
		.0030	.0017	.0018	.0031		
		.0035	.0293	.0521	.1036		
		.0031	.0020	.0019	.0027		
900		.0071	.0112	.0162	.0079		
		.0088	.0011	.0015	.0033		
		.0424	.0475	.0343	.0129		
		.0090	.0012	.0015	.0031		

$p_1 = 0.05$

MAD values

MAD of AE and CE is 0 in the first two decimals

The same is not true for SA and A2.

Mean of the Absolute Differences (MAD) between the power calculated using the "exact" multinomial distributions versus an approximation.



Log-likelihood ratio statistic

The asymptotic equivalent as well as the cost-effective approximation can be obtained for the log-likelihood ratio statistic (in place of Pearson's statistic) analogously.

The numerical results are also very similar.

Two stage design

M markers, M_1 are causal

Stage I.

T is calculated for each marker based on n samples.
Reject if $T > c$.

Markers rejected at Stage I

Stage II.

T' is calculated for each marker based on n' samples.
Reject if $T' > c'$.

Markers believed to be causal

How to choose n , n' , c and c' for that

$$\text{FDR} \leq \alpha$$

$$\text{APW} \geq \beta ?$$

$$\text{FDR} = E \left[\frac{\# \text{ falsely rejected markers}}{\# \text{ rejected markers}} \right]$$

APW = average power

4 parameters, 2 constraints \Rightarrow
some freedom in choosing n , n' , c , c' .

We can optimize.

Two stage design & data pooling

M markers, M_1 are causal

Stage I.

T is calculated for each marker based on n samples.
Reject if $T > c$.

Markers rejected at Stage I

Stage II.

T^* is calculated for each marker based on n' samples.
Reject if $T^* > c^*$.

Markers believed to be causal

Stage I data are reused in Stage II.

That is Pearson's statistic in Stage II (T^*) is calculated on the aggregated contingency table:

$$\begin{bmatrix} x_1 & \equiv & x_1^{\diamond} & \blacklozenge & x_m & \equiv & x_m^{\diamond} \\ y_1 & \equiv & y_1^{\diamond} & \blacklozenge & y_m & \equiv & y_m^{\diamond} \end{bmatrix}.$$

T^* is no longer independent of T .

What is the distribution of T^* ?



Data pooling in multistage designs

- In the context of genetic studies multistage can reduce genotyping 50-70%
- A disadvantage is that larger samples are needed
- Pooling data from multiple stages can mitigate this effect on sample size



Consequences of data pooling

- Only markers significant at stage 1 are typed in the 2nd stage.
- As a result, traditional tests can no longer be used
- In the literature 3 approaches have been used to deal with this:
 - 1) Ignore problem (e.g. Satagopan, 2003/2004). Disadvantage inflation of type I error and incorrect power calculations, for 2x2 tables only.
 - 2) Use simulation (e.g. Clayton, 2004). Disadvantages: loss of control when designing studies, time consuming, sampling fluctuations, need good random number generator etc.
 - 3) Combine p-values or test statistics (group sequential design literature) from different stages. In principle flexible but not clear most powerful approach.

The probability that a marker is rejected in the second stage is

$$\Pr(T > c, T > c)$$

We have seen that T is approximately

$$T \approx Z_1^2 + \dots + Z_{m-1}^2 = \left(A_1 \mathbf{U} + \sqrt{n} \sqrt{v_1^2 + \frac{1}{m-1} v_m^2} \right)^2 + \dots + \left(A_{m-1} \mathbf{U} + \sqrt{n} \sqrt{v_{m-1}^2 + \frac{1}{m-1} v_m^2} \right)^2,$$

where

$$U_i = \frac{1}{\sqrt{p_i p_i + q_i q_i}} \left[\left(\frac{q_i - p_i \sqrt{p}}{2 p_i p_i + q_i q_i} \right) \frac{Y_i - q_i n}{\sqrt{q_i n}} - \left(\frac{q_i - p_i \sqrt{q}}{2 p_i p_i + q_i q_i} \right) \frac{X_i - p_i n}{\sqrt{p_i n}} \right]$$

and $\mathbf{v} = \frac{1}{\sqrt{n}} A \boldsymbol{\mu} = A \left(\frac{(p_1 - q_1) \sqrt{pq}}{\sqrt{pp_1 + qq_1}}, \dots, \frac{(p_m - q_m) \sqrt{pq}}{\sqrt{pp_m + qq_m}} \right)^T$ ← does not depend on n

Test statistic on the Stage II data only.

$$\rightarrow T' \approx Z_1'^2 + \dots + Z_{m-1}'^2 = \left(A_1 \mathbf{U}' + \sqrt{n'} \sqrt{v_1^2 + \frac{1}{m-1} v_m^2} \right)^2 + \dots + \left(A_{m-1} \mathbf{U}' + \sqrt{n'} \sqrt{v_{m-1}^2 + \frac{1}{m-1} v_m^2} \right)^2$$

Similarly for the test statistic on the pooled data $T^* \approx Z_1^{*2} + \dots + Z_{m-1}^{*2}$.

It can be seen that $Z_i^* = \frac{\sqrt{n}}{\sqrt{n+n'}} Z_i + \frac{\sqrt{n'}}{\sqrt{n+n'}} Z_i', \quad i = 1, \dots, m.$

Let $s_1 = \frac{\sqrt{n}}{\sqrt{n+n'}}$ and $s_2 = \frac{\sqrt{n'}}{\sqrt{n+n'}}$.

We have that

$$\Pr(T^* > c^*, T > c) \approx \Pr\left(\left(s_1 Z_1 + s_2 Z'_1\right)^2 + \dots + \left(s_1 Z_{m-1} + s_2 Z'_{m-1}\right)^2 > c^*, Z_1^2 + \dots + Z_{m-1}^2 > c\right)$$

all of them are independent

Equivalently

$$\Pr(T^* > c^*, T > c) \approx \Pr\left(Z_1^{*2} + \dots + Z_{m-1}^{*2} > c^*, Z_1^2 + \dots + Z_{m-1}^2 > c\right)$$

where $Cor(Z_i^*, Z_i) = s_1$.

This probability can be computed by mathematical methods.

For 2x2 tables ($m=2$):

$$\Pr\left(Z_1^{*2} > c^*, Z_1^2 > c\right) = \Pr\left(|Z_1^*| > c^*, |Z_1| > c\right) = \Pr\left(Z_1^* > c^*, Z_1 > c\right) + \Pr\left(Z_1^* > c^*, Z_1 < -c\right) + \dots$$

these can be computed by mvtnorm (R package)

Under the null hypothesis

Standard normal variables

$$\Pr\left(\frac{T - c}{c} \leq T \leq c\right) \sim \Pr\left(\frac{Z_1^2}{n} \leq \dots \leq \frac{Z_m^2}{m}\right) \leq c, Z_1^2 \leq \dots \leq \frac{Z_m^2}{m} \leq c$$

where $Cor(Z_i^*, Z_i) = s_1$.

Thus, the right-hand side depends on c , c' and $s_1 = \frac{\sqrt{n}}{\sqrt{n+n'}}$ only.

Accuracy large sample approximation for data pooling

n, n', (cell props)	Data pooling			
	Large sample approx.		Exact distribution	
	Stage II crit. value	Pow	Stage II crit. value	Pow
500,1000 .07,.002,.5,.419	14.21	.419	14.26	.420
200,1200 .07, .002,.5,.37	12.82	.945	12.80	.945
100,100 .01,.01,.5, .251	14.30	.690	14.21	.698
200,200 .01,.1,.5, .321	8.94	.936	8.95	.934

Optimal design

How to choose n , n' , c and c' for that

$$\text{FDR} \leq \alpha$$

$$\text{APW} \geq \beta$$

or something else

and the *genotyping burden* is minimal?

This leads to the following non-linear optimization problem.

$$n \text{ and } n' \text{ } \Pr(T > c | H_0) \leq p_0 \text{ and } \Pr(T > c | H_1) \geq p_1 \text{ } \min$$

Subject to

$$\text{I. } g_1(c, c^*, n, n') = \Pr(T^* > c^*, T > c | H_1) = \text{APW}$$

$$\text{II. } g_2\left(c, c^*, \frac{\sqrt{n}}{\sqrt{n+n'}}\right) = \Pr(T^* > c^*, T > c | H_0) = \frac{\text{APW}(1-p_0)}{(1/\text{FDR}-1)p_0}.$$

Method

For a given c and $r = \frac{\sqrt{n}}{\sqrt{n+n'}}$, calculate c^* by II. and n' by I. \Rightarrow genotyping burden

Unconstrained optimization on (c, r) .

What we gain by pooling data?

		Sample size	gen. burden
$p_0=0.999$ $p_1=0.5, q_1=0.4$	non-pooled	$177 + 1074 = 1251$	506.23
	pooled	$171 + 931 = 1102$	486.49
$p_0=0.999$ $p_1=0.5, q_1=0.35$	non-pooled	$77 + 469 = 546$	222.24
	pooled	$75 + 406 = 481$	213.42
$p_0=0.9999$ $p_1=0.5, q_1=0.4$	non-pooled	$80 + 1100 = 1180$	254.50
	pooled	$77 + 997 = 1074$	245.86
$p_0=0.9999$ $p_1=0.5, q_1=0.35$	non-pooled	$33 + 468 = 501$	112.52
	pooled	$32 + 435 = 467$	108.35