ON BIOMEDICAL GENOMICS

KJELL DOKSUM
JOINT WITH KAM-WAH TSUI AND FAN YANG

Department of Statistics
University of Wisconsin Madison

JUNE 10, 2011, TORONTO FIELD INSTITUTE WORKSHOP

OUTLINE

STUDIES OF ASSOCIATION BETWEEN
DISEASE AND GENETIC MARKERS

COMPARISONS BETWEEN POPULATION GENETIC
AND STATISTICAL APPROACHES

MODEL COMPARISONS
CRITERIA COMPARISONS

BACKGROUND: GENOMICS RESEARCH PROJECT

THE LARGE p SMALL n PROBLEM, WHERE
$n =$ SAMPLE SIZE AND
$p =$ No. OF PREDICTORS.

EXAMPLES INCLUDE $n = 3000$ AND $p = 500,000$.
THIS IS NOT A PROBLEM, SAY POPULATION GENETICISTS.

QUESTIONS:

(1). WHY NOT A PROBLEM? THEIR ANSWER: DUAL PCA, CORRELATION AND BONFERRONI.

(2). WHAT MODEL DO POPULATION GENETICISTS PROPOSE?

(3). FOR THEIR MODEL, WHAT ARE GOOD PROCEDURES?

BACKGROUND: GENOMICS RESEARCH PROJECT

THE DATA CONTAINS
1500 CASES (TYPE II DIABETES);
1500 CONTROLS (NOT TYPE II DIABETES);
3000 INDIVIDUALS, i.e., $n = 3000$.

QUESTION: HOW DO CASES AND CONTROLS DIFFER GENETICALLY?

FOR EACH INDIVIDUAL, WE MEASURE $p = 500,000$ GENETIC MARKERS, CALLED SNPs (SINGLE NUCLEOTIDE POLYMORPHISMS).

EACH SNP SCORE IS 0, 1 or 2. IT MEASURES HOW CLOSE AN INDIVIDUAL'S SNP AT A CERTAIN LOCATION IS TO THE SNP OF A REFERENCE GENOME AT THE SAME LOCATION.

BACKGROUND: GENOMICS RESEARCH PROJECT

GENETIC MARKER VALUES
$g_{ij} \in \{0, 1, 2\}$

AT $p = 500,000$ GENOME LOCATIONS

FOR $n = 3000$ INDIVIDUALS

$1 \leq j \leq n$, $1 \leq i \leq p$.

ASK: AT LOCATION i, IS THERE A SIGNIFICANT ASSOCIATION BETWEEN GENOTYPE AND DISEASE, $i = 1, \ldots, 500,000$?

H_{0i}: NO ASSOCIATION AT LOCATION i

BASIC TEST STATISTICS

USE TEST STATISTICS $T_i = R_i\sqrt{\frac{n-2}{1-R_i^2}}$

WHERE

$$R_i = Corr(g_{i1}, \cdots, g_{in}; d_1, \cdots, d_n), i = 1, \cdots, p.$$

AND

$$d_j = \begin{cases} 1, & \text{if disease} \\ 0, & \text{otherwise} \end{cases} \qquad j = 1, \cdots, n.$$

THE TESTS BASED ON $|T_i|$ ARE EQUIVALENT TO CHI-SQUARE-TESTS.

LET ρ_i = POPULATION CORRELATION, WE TEST

$$H_0 : \rho_i = 0 \qquad H_1 : \rho_i \neq 0, \qquad i = 1, \cdots, p.$$

USING TEST STATISTICS $|T_i|$ AND BONFERRONI CRITICAL VALUES. THEY USE CR.VAL.$= .05/500000 = 10^{-7}$.
REJECT H_0 IF $p_i < 10^{-7}$, WHERE p_i = P-VALUE FOR THE iTH SNP.

GENOTYPE STANDARDIZATION:

RECALL $g_{ij} \in \{0, 1, 2\}$. IN THE HARDY-WEINBERG GENETIC MODEL, $g_{ij} \sim Bin(2, q_i)$. HERE $E(g_{ij}) = 2q_i$ AND $SD(g_{ij}) = \sqrt{2q_i(1 - q_i)}, 1 \leq i \leq p, 1 \leq j \leq n$.

INSTEAD OF g_{ij}, USE THE STANDARDIZED SCORE

$$NEW\ g_{ij} = \frac{g_{ij} - \bar{g}_i}{\sqrt{2\hat{q}_i(1 - \hat{q}_i)}},$$

where $\bar{g}_i = (\sum_{j=1}^{n} g_{ij})/n$ and

$$\begin{aligned}
\hat{q}_i &= \frac{1 + \sum_{j=1}^{n} g_{ij}}{2 + 2n} \\
&= \text{ESTIMATE OF ALLELE FREQUENCY OF SNP i} \\
&= \text{BAYES ESTIMATE BASED ON BETA(2,2) PRIOR} \\
&\quad \text{IN HARDY-WEINBERG MODEL}
\end{aligned}$$

GENOTYPE STANDARDIZATION:

THIS TRANSFORMATION MAKES THE NEW g_{ij}'s HAVE MEANS ZERO AND APPROXIMATELY THE SAME SD's.

IT IS USED IN THE CORRELATION VERSION OF PRINCIPAL COMPONENT ANALYSIS (PCA).

IT GIVES MORE WEIGHT TO RARE SNPS.

CONFOUNDING PROBLEM

BIG PROBLEM: THERE MAY BE SPURIOUS CORRELATION BECAUSE OF GENETIC CLUSTERS.
THIS SPECIFIED MARKER g is **NOT** ASSOCIATED WITH DISEASE, THE CLUSTERS ARE DETERMINED BY MARKERS OTHER THAN THE SPECIFIED MARKER g.

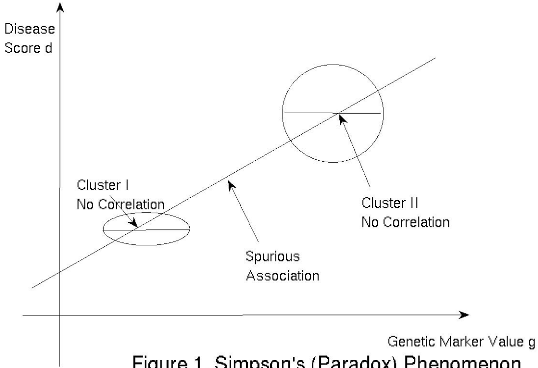


Figure 1. Simpson's (Paradox) Phenomenon

ANCESTRY STRATA

THE CLUSTERS ARE CALLED GENETIC ANCESTY STRATA.
PEOPLE WITH SIMILAR "ANCESTRY" ARE IN THE SAME STRATA.
ANCESTRY IS A CONFOUNDING VARIABLE.

HOW TO CORRECT FOR ANCESTRY AND THEREBY AVOID SPURIOUS ASSOCIATION?

ANSWER: DUAL PCA = DUAL PRINCIPAL COMPONENT ANALYSIS.

THE ALGORITHM IS CALLED "EIGENSTRAT" OR "EIGENSOFT".

DUAL PCA:

THE TRANSPOSE OF THE GENOTYPE DESIGN MATRIX IS

$$X = (g_{ij})_{p\times n} = (DESIGN\ MATRIX)^T,$$

$$\Psi_{n\times n} = p^{-1}X^TX = DUAL\ \text{``}COVARIANCE\text{''}\ MATRIX.$$

WHY NOT USE THE USUAL $(n^{-1}XX^T)_{p\times p}$?

BECAUSE COMPUTER CANNOT HANDLE A $p \times p$ MATRIX.

TREAT DATA ACROSS MARKERS AS IID, TREAT DATA ACROSS INDIVIDUALS AS VARIABLES THAT ARE NOT IID.

IS THIS A PROBLEM?
NO, BECAUSE:
RESULT: X^TX AND XX^T HAVE THE SAME NONZERO EIGENVALUES.

ANCESTRY

DEFINITION: THE ANCESTRY a_{kj} OF INDIVIDUAL j ALONG THE kTH AXIS OF ANCESTRY VARIATION IS THE jTH COORDINATE OF THE kTH EIGENVECTOR A_k OF $(X^T X)_{n \times n}$, $1 \leq k \leq K$. $\sum_j a_{kj} = 0$, $\sum_j a_{kj}^2 = 1$, $\sum_j a_{kj} a_{k'j} = 0$ WITH $k \neq k'$.
HERE K IS CHOSEN BY THE JOHNSTONE TEST, WHICH IS BASED ON THE TRACY-WIDOM DISTRIBUTION.

THE FIRST SAMPLE DUAL PRINCIPAL COMPONENT EVALUATED AT THE jTH PERSON $= \sqrt{\lambda_1} a_{1j}$,
WHERE λ_1 IS THE LARGEST EIGENVALUE OF $(X^T X)_{n \times n}$.

THUS, $a_{1j} = \frac{1}{\sqrt{\lambda_1}} * \{$LOADING FACTOR OF THE jTH INDIVIDUAL IN THE FIRST DUAL PC$\}$.
INDIVIDUALS WITH SIMILAR LOADING FACTORS ARE IN THE SAME CLUSTER.

ANCESTRY ADJUSTMENT:

THERE IS NO NEED TO USE CLUSTERS.

INSTEAD USE A CONTINUOUS ANCESTRY ADJUSTMENT.

EACH PERSON IS ASSIGNED A GENETIC ANCESTRY SCORE, WHICH IS SUBTRACTED FROM THEIR GENOTYPE g_{ij}.

ANCESTRY ADJUSTMENT:

LET $\hat{g}_{kij}$ = PREDICTED GENOTYPE AT LOCUS i FOR INDIVIDUAL j BASED ON THE ANCESTRY a_{kj}.
THEN $\hat{g}_{kij} = \gamma_{ki} a_{kj}$ WHERE

$$\begin{aligned}
\gamma_{ki} &= \sum_j a_{kj} g_{ij} \\
&= \text{DUAL } k \text{ TH PC EVALUATED AT} \\
&\quad \text{THE } i \text{ TH COLUMN OF } X = (g_{ij}) \\
&= \text{REGRESSION COEFFICIENT WHEN} \\
&\quad \text{REGRESSING } g_{i1}, \cdots, g_{in} \text{ LINEARLY ON } A_k
\end{aligned}$$

ANCESTRY ADJUSTMENT

$\hat{g}_{kij}$ = BEST LINEAR PREDICTOR OF g_{ij} BASED ON A_k.

DEFINION: THE GENOTYPE ADJUSTED FOR ANCESTRY ALONG THE kTH ANCESTRY AXIS IS

$$g_{kij} = g_{ij} - \hat{g}_{kij}$$

HOW DOES THIS COMPARE WITH THE COVENTIONAL STATICIAL PCA?

ANCESTRY ADJUSTMENT

REGRESSING SIMULTANEOUSLY ON ALL $PCA_k, 1 \leq k \leq K$, ANCESTRY SCORES

IS EQUIVALENT TO

REGRESSING ON EACH PCA_k ONE AT A TIME IN SEQUENCE,

BECAUSE THE *PCs* ARE ORTHOGONAL.

ANCESTRY ADJUSTMENT

SUPPOSE WE USE $(X^T X)_{p \times p}$
INSTEAD OF THE DUAL $(XX^T)_{n \times n}$
THEN,

$$a_{kj} = \frac{1}{\sqrt{\lambda_k}} \{(PC)_k \text{ EVALUATED FOR THE INDIVIDUAL } j\}$$

WE COULD HAVE USED SINGULAR VALUE DECOMPOSITION TO ARRIVE AT THIS POINT.

ANCESTRY ADJUSTMENT

THUS, THE ANCESTRY a_{kj} OF THE jTH INDIVIDUAL ALONG THE kTH ANCESTRY AXIS IS PROPORTIONAL TO THE kTH CONVENTIONAL PRINCIPAL COMPONENT EVALUATED AT THE GENETIC MARKER SCORES FOR THE jTH INDIVIDUAL.

WHEN ADJUSTING GENOTYPE FOR ANCESTRY BY FORMING $g_{ij} - \hat{g_{kij}}$, WE ARE ADJUSTING BY USING THE BEST LINEAR PREDICTOR $\hat{g_{kij}}$ OF GENOTYPE g_{ij} FOR THE INDIVIDUAL j BASED ON THE kTH CONVENTIONAL PRINCIPAL COMPONENT.

PHENOTYPE ADJUSTMENT

THE PHENOTYPES: $d_1, \cdots, d_n$
$d_j =$ DISEASE INDICATOR, 0 OR 1, FOR jTH INDIVIDUAL.
ALSO ADJUST FOR ANCESTRY

$$d_j \to d_j - \hat{d}_{kj} = d_{kj}$$

$\hat{d}_{kj}$ IS THE BEST LINEAR PREDICTOR OF d_j BASED ON THE ANCESTRY AXIS A_k, WHICH IS ALSO THE BEST LINEAR PREDICTOR OF A_k BASED ON THE CONVENTIONAL EIGENVECTOR B_k.

REGRESS $\{PHENOTYPE - E(GE|A)\}$ ON $\{GENOTYPE - E(PH|A)\}$.

DECISION RULE

THE FINAL STATISTICS ARE

$$T_i = R_i \sqrt{\frac{n-2}{1-R_i^2}},$$

WHERE R_i IS THE CORRELATION BETWEEN THE ANCESTRY ADJUSTED STANDARDIZED GENOTYPES g_{kij}, $1 \leq k \leq K, 1 \leq i \leq p, 1 \leq j \leq n$ AND THE ANCESTRY ADJUSTED DISEASE INDICATORS d_{kj}, $1 \leq k \leq K, 1 \leq j \leq n$.

LET p_i =P-VALUE BASED ON T_i

DECIDE THAT THE *iTH* MARKER IS ASSOCIATED WITH DISEASE IF $p_i \leq 10^{-7}, i = 1, \cdots, 500,000$.

A BIOMEDICAL GENOMICS MODEL

HIERARCHICAL MODEL:

(1). CHOOSE U $UNIFORM(0.1, 0.9)$.
OUTPUT $U = u$.

(2). CHOOSE P_1, P_2 I.I.D. $BETA(\alpha, \beta)$
WITH $\alpha = \frac{1-d}{d}u$ and $\beta = \frac{1-d}{d}(1-u)$
WHERE

$$\begin{aligned} d &= \text{GENETIC DIFFERENTIATION WITHIN A POPULATION} \\ &= F_{st} = 0.01(\text{EUROPE}) \\ \alpha &= 99u \end{aligned}$$

HERE $E(P_1|u) = E(P_2|u) = u$, $VAR(P_l|u)$ IS SMALL.

OUTPUT $(P_1, P_2) = (p_1, p_2)$.

(3)

(a) STRATA I: HARDY-WEINBERG MODEL CHOOSES CONTROL AND CASE GENOMIC DATA AS
$G_{CO} \sim BINOMIAL(2, p_1),\ g_{ij}^{I}, d_j^{I} = 0$
$G_{CA} \sim BINOMIAL(2, p_1^*),\ g_{ij}^{I}, d_j^{I} = 1$
WHERE
$p_1^* = \frac{Rp_1}{1-p_1+Rp_1}$ AND
R =RELATIVE RISK OF DISEASE$=\frac{p_1^*}{1-p_1^*}\frac{1-p_1}{p_1}$
$R = 1$ MEANS SAME RISK
OUTPUT g_{ij}^{I}, d_{ij}^{I}

(b) STRATA II: SAME, EXCEPT USE p_2.
OUTPUT g_{ij}^{II}, d_j^{II}

A BIOMEDICAL GENOMICS MODEL

(4)
GENERATE 600 CASES AND 400 CONTROLS FROM STRATA I.

GENERATE 400 CASES AND 600 CONTROLS FROM STRATA II.

OUTPUT $\{g_{ij}, d_j\}$. THE STRATA LABELS I AND II ARE DROPPED.

METHOD 1: EIGENSTRAT

LET p_i BE P-VALUE FOR iTH SNP USING $T_i = R_i \sqrt{\frac{n-2}{1-R_i^2}}$.

SELECT THE SNPs WITH $p_i < 10^{-7}$, BONFERRONI WITH $\alpha = 0.05$ AND I. JOHNSTONE "SELECTION" OF NO. OF STRATA.

$R_i =$ CORR BETWEEN iTH GENETIC MARKER AND DISEASE INDICATOR USING STANDARDIZATION AND ANCESTRY ADJUSTMENT.

METHOD 2: LOGISTIC REGRESSION

$$LOGIT\,[PROB(DISEASE|GENOTYPE)] = \alpha + \beta GENOTYPE$$

$$T_i = t - STAT = \hat{\beta}/SE(\hat{\beta})$$

HERE GENOTYPE IS STANDARDIZED AND
ANCESTY ADJUSTED.
PHENOTYPE IS NOT.

METHOD 3: EFRON'S EMPIRICAL BAYES

H_{0i}: CORR(PHENOTYPE,GENOTYPE G_i)$= 0$
H_{1i}: CORR$\neq 0$

$\pi_0 =$ PRIOR PROB OF H_{0i}. HERE $\pi_0 =$ IS CLOSE TO 1.
$p_i =$ P-VALUE FOR SOME STAT SUCH AS T_i
$Z_i = \phi^{-1}(p_i) \sim N(0,1)$ WHEN H_{0i} HOLDS
THE ESTIMATED POSTERIOR PROBABILITY OF H_{0i} IS

$$\hat{P}(H_{0i}|Z_i) = \frac{\pi_0 f_0(z_i)}{\pi_0 f_0(z_i) + (1 - \pi_0)\hat{f}_1(z_i)}$$

WHERE $f_0 = N(0,1)$ and $\hat{f}_1$ IS AN ESTIMATE OF f_1.

DECIDE H_{1i} IF $\hat{P}(H_{0i}|Z_i) \leq 0.2$
THIS PROCEDURE IS LABELLED "LOCAL FDR".

METHODS:

HOW TO ESTIMATE f_1?

ASSUME A MIXTURE MODEL:

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$$

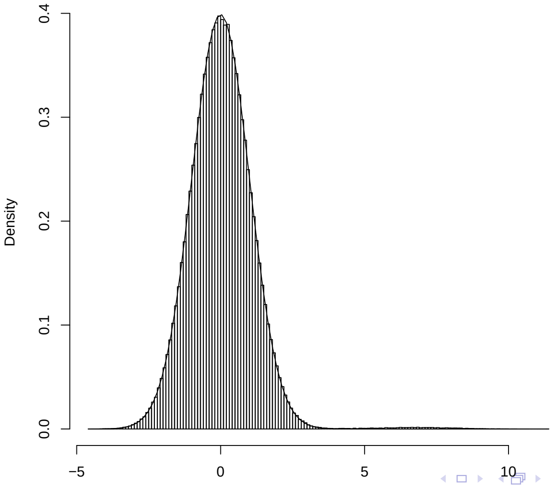
$$f_0(z) = N(0, 1)$$

$$f_1(z) = N(\mu, \sigma^2)$$

SEE GRAPH NEXT PAGE.

THE DISTRIBUTION OF ALL Z-VALUES BY EIGENSTRAT

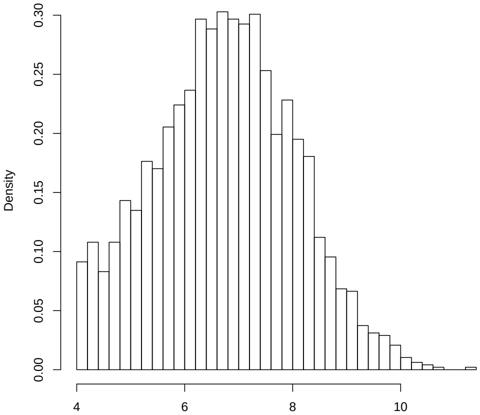
10000 SNPs, 50 relevant SNPs, 1000 persons, $R = 2$, 50 MC simulations



THE DISTRIBUTION OF Z-VALUES GREATER THAN 4, BY EIGENSTRAT

10000 SNPs, 50 relevant SNPs, 1000 persons, $R = 2$, 50 MC simulations

THIS IS AN ESTIMATE OF f_1 IN $(1 - \pi_0)f_1$



EFRON'S METHOD (EMP. BAYES) ONLY REQUIRES P-VALUES. THUS WE CAN COMBINE METHOD 3 WITH METHODS 1 AND 2.

METHOD 3A: EFRON-EIG

EFRON EMP. BAYES WITH T_i FROM EIGENSTRAT

METHOD 3B: EFRON-LOG

EFRON EMP. BAYES WITH T_i FROM LOGISTIC REGRESSION

OPTIMAL EMP. BAYES

LET $X = (g, Y)$ = ALL DATA
RECALL BAYES RULE:
DECIDE H_{1i} IF $P(H_{1i}|X) > P(H_{0i}|X)$.

CRITERIA # 1: SUN AND CAI (07 JASA)
HYBRID: BAYES-NEYMAN-PEARSON
SUBJECT TO $P(H_{0i}|X) \leq q$
MAXIMIZE $P(H_{1i}|X)$

EFRON (BOOK 2010), SUN AND CAI (07 JASA) AND OTHERS PROPOSED CLEVER WAYS OF ESTIMATING π_0, $P(H_{1i}|X)$ AND $P(H_{0i}|X)$.
ROBBINS, STEIN, EFRON-MORRIS IDEAS

CRITERIA # 2: NEYMAN-PEARSON-SPJϕTVOLL

CONSIDER ALL PROCEDURES WITH
I: $\sum_{i=1}^{p} P(H_{0i} \text{ REJECTED}|H_{0i} \text{TRUE}) = \gamma =$ "LEVEL"

AMONG THESE, TRY TO "MAXIMIZE"
II^c: $\sum_{i \in A} P(H_{0i} \text{ REJECTED}|H_{1i} \text{ TRUE})=$"POWER"
WHERE

$$\begin{aligned} A &= \{i : i\text{TH SNP RELEVANT}\} \\ &= \text{TRUE ASSOCIATION SET} \end{aligned}$$

THIS CRITERIA IS USED IN THE RECENT BIOMEDICAL GENOMICS LITERATURE.
WHAT SAMPLE SIZE IS NEEDED TO HAVE REASONABLE POWER FOR THE BIOMEDICAL MODEL?

CRITERIA # 3: FDR AND FNR

H_0: ALL H_{0i} ARE TRUE. "THERE ARE NO RELEVANT SNPS."

FDR = FALSE DISCOVERY RATE= $E_{H_0}(FDR^*)$ WHERE

$$FDR^* = \frac{\#\text{ IRRELEVANT SNPs SELECTED}}{\#\text{SELECTED SNPs}} = \frac{\#\text{FALSE DISC'S}}{\#\text{OF DISC'S}}$$

$FNR = E_A(FNR^*)$ WHERE

$$FNR^* = \frac{\#\text{ RELEVANT SNPs NOT SELECTED}}{\#\text{ SNPs NOT SELECTED}}$$

AND $A = \{i : \text{SNP } i \text{ IS RELEVANT}\}$

BENJAMINI AND HOCHBERG (1995) CONSTRUCTED A PROCEDURE WITH $FDR < q$, FOR A PREASSIGNED $q \in (0, 1)$. HERE $0/0 = 0$.

CRITERIA # 2 REVISITED:

SPJϕTVOLL (72 AMS), STOREY (07 JRSSB),
SUN AND CAI (07 JASA)

IN THE CLASS OF PROCEDURES WITH
E(#FALSE DISCOVERIES)=GAMMA
MAXIMIZE
E(#CORRECT DISCOVERIES).

THE DISCOVERY RULES LOOK LIKE:

$$\Psi_i(X) = \begin{cases} 0 & \text{DECIDE SNP } i \text{ IRRELEVANT} \\ 1 & \text{DECIDE SNP } i \text{ RELEVANT} \end{cases}$$

HERE X = ALL DATA ACROSS ALL SNPS AND DISEASE INDICATORS.

SPJϕTVOLL's THEOREM (1972)

LET $f_{01}, \ldots, f_{op}$ AND $f_1, \ldots, f_p$
GE GIVEN INTEGRABLE FUNCTIONS.

LET $S(\gamma)$ =TESTS
$\psi_1, \ldots, \psi_p$ WITH
$\sum_{i=1}^{p} \int \psi_i(x) f_{0i}(x) d\mu(x) = \gamma$ (E.G., EXPECTED NO. OF FALSE DISCOVERIES)

THEN THE TEST THAT MAXIMIZES
$\sum_{i=1}^{p} \int \psi_i(x) f_i(x) d\mu(x)$ (E.G., EXPECTED NO. OF CORRECT DISCOVERIES)
IS

$$\phi_1, \ldots, \phi_p = \{1[f_i(x) > cf_{0i}(x)] : i = 1 \ldots, p\}$$

.

SPJϕTVOLL, SPECIAL CASE (a)

TAKE

$$f_{0i} = \text{DENSITY FOR THE IRRELEVANT CASE}$$
$$f_i = \text{DENSITY FOR THE RELEVANT CASE}$$

CONSIDER TWO DIFFERENT SCENARIOS:

CASE I: ALL SNP's ARE IRRELEVANT.
CASE II: EXACTLY FIVE SNPS ARE RELEVANT.

WE CAN ASK:
IF WE CONTROL THE CASE I EXPECTED NO. OF FALSE DISCOVERIES AT γ, THAT IS SPJ LEVEL $= \gamma$,
WHAT IS THE EXPECTED NO. OF CORRECT DISCOVERIES FOR CASE II USING METHOD k? WHAT IS THE POWER?
HOW DOES METHOD k COMPARE TO THE ORACLE THAT USE SPJϕTVOLL's OPTIMAL RULE?

SPJϕTVOLL, SPECIAL CASE (b)

TAKE

$$h_{0i} = \text{DENSITY FOR THE IRRELEVANT CASE}$$

TAKE

$$f_{0i} = h_{0i}/\text{NO. OF DISCOVERIES}$$

SPJϕTVOLL γ = FDR = FALSE DISCOVERY RATE
COMPUTE (OR MAXIMIZE) CORRECT DISCOVERY RATE
FOR VARIOUS SCENARIOS.

$p = 10,000$ SNPs,
$n = 1,000$ PEOPLE,
$M = 200$ MONTE CARLO TRIALS,
$p_i < 10^{-7}$ IN EIGENSTRAT,
DATA GENERATED USING POPULATION GENETIC MODEL.

TABLE 1. CRITERIA # 3:
FDR IS THE EXPECTED VALUE OF

$$FDR^* = \frac{\#\text{ IRRELEVANT SNPs SELECTED}}{\#\text{SELECTED SNPs}}$$

FNR IS THE EXPECTED VALUE OF

$$FNR^* = \frac{\#\text{ RELEVANT SNPs NOT SELECTED}}{\#\text{ SNPs NOT SELECTED}}$$

TABLE 1. BIOMEDICAL MODEL FDR*

WHEN $R = 1$ ALL H_{0i} HOLD NO ASSOCIATION.
BH = BENJAMINI AND HOCHBERG

SET $FDR = 0.2$

	EIGENSTRAT t	LOGISTIC t
BH, $FDR = 0.2$	0.21	0.11
SPJ, $\gamma = 0.5$	0.40	0.26
SPJ, $\gamma = 1.0$	0.59	0.43
SPJ, $\gamma = 1.5$	0.75	0.56
SPJ, $\gamma = 2.0$	0.87	0.71
SPJ, $\gamma = 3.0$	0.96	0.87
EFRON BAYES ≤ 0.2	0.11	0.06

RECALL THAT FDR^* IS EITHER 0 OR 1. IN EACH TRIAL, FDR^* IS EITHER 0 OR 1.

SPJ LEVEL AND FDR LEVELS HAVE SIMILAR PROPERTIES.

RECALL: SPJ LEVEL γ MEANS CUT OFF POINT $\gamma/10000$ FOR THE *iTH* P-VALUE.

THE SMALLER γ IS, THE MORE CONSERVATIVE THE TEST IS.

TABLE 2. 100 TIMES FDR^*, FNR^* AND SPJϕTVOLL POWER WHEN $R = 1.75$, $d = 50$.

EIGENSTRAT t STATISTIC

	FDR^*	FNR^*	$POWER^*$
BH $FDR = 0.2$	19.8	0.03	94.0
SPJ, $\gamma = 0.5$	1.1	0.08	83.6
SPJ, $\gamma = 1.0$	1.2	0.07	86.5
SPJ, $\gamma = 1.5$	2.9	0.06	88.2
SPJ, $\gamma = 2.0$	4.1	0.06	89.0
SPJ, $\gamma = 3.0$	6.0	0.05	90.0
EFRON BAYES	2.0	0.06	88.1

$POWER^* = 100(ESTIMATEDPOWER)/d$
HERE FDR^*, FNR^* AND $POWER^*$ ARE IN HARMONY.
FNR^* IS WORSE IN THE CONSERVATIVE CASE AS IS THE POWER.

SPJ LEVEL AND FDR LEVEL BEHAVE THE SAME WAY.

TABLE 3. SPJ γ-LEVEL $\times 10^4$.

HERE $p = \#\text{SNP's} = 10^4$, $d = 50$, NOMINAL $\gamma = 2.5$, 200 MC TRIALS

METHOD	$R = 1.00$	$R = 1.25$	$R = 1.75$	$R = 2.00$
(1) EIGEN: $p < 2.5/10^4$	2.44	2.45	2.45	2.46
(2) LOGREG: $p < 2.5/10^4$	1.47	1.47	1.48	1.47
(3) EIGEN+EFRON	2.49	2.49	2.50	2.50
(4) LOGREG+EFRON	2.48	2.52	2.50	2.50

LESSON: SPJ γ-LEVEL CORRECT FOR (1),(3),(4)
AND STABLE AS A FUNCTION OF $R =$ ODDS RATIO.

TABLE 4. COMPARISONS OF THREE CRITERIA FOR METHODS WITH SPJ $\gamma = 2.5$

METHOD	$R = 1.25$			$R = 1.75$		
	$POWER^*$	FDR^*	$FNR^* * 100$	$POWER^*$	FDR^*	$FNR^* * 100$
(1)EIGEN: $p < 2.5/10^4$	0.071	0.40	0.65	0.90	0.51	0.53
(3)EIGEN+EFRON	0.084	0.30	0.58	0.91	0.49	0.46
(4)LOGREGR+EFRON	0.085	0.30	0.58	0.91	0.49	0.47

THE THREE CRITERIA ARE CONSISTENT. THEY FAVOR (3) AND (4). THESE ARE "ADAPTIVE COMPOUND" RULES, WHICH MEANS THEY USE DATA FROM ALL SNPS WHEN DECIDING WHETHER THE *iTH* SNP IS RELEVANT.