2. (10 points) Derive a statistic that is a multiple of the allele frequency difference which has variance 1. What is the mean of this statistic?

variance

$$p^+_A(1-p^+_A) + p^-_A(1-p^-_A) = 2\hat{p}_A(1-\hat{p}_A)$$

where $p_A = \frac{p^+_A + p^-_A}{2}$.

$$S_A \sim N(p^+_A - p^-_A,\ 2p_A(1-p_A))$$

divide by std dev:

$$\sim N\left(\frac{p^+_A - p^-_A}{\sqrt{2p_A(1-p_A)/N}},\ \frac{2p_A(1-p_A)}{2p_A(1-p_A)}\right)$$

$$\sim N\left(\frac{p^+_A - p^-_A}{\sqrt{2p_A(1-p_A)}}\sqrt{N},\ 1\right)$$

$$\lambda_A = \frac{p^+_A - p^-_A}{\sqrt{2p_A(1-p_A)}}$$

which is the non-centrality parameter.

2 Calculating Power (10 points)

We genotype SNP A with alleles {A,a}. Assume that the true case/control probabilities in the target population are 0.5 and 0.3, respectively. If we collect 400 case and 400 control individuals, given a significance threshold of 0.05, what is the power of this association study ? 800 case 800 control

Provide your answers in terms of $\Phi(x)$ and $\Phi^{-1}(x)$ or pnorm and qnorm in R.

$p^1_A = 0.5$ $p^2_A = 0.3$ $p_A = \frac{0.5+0.3}{2} = 0.4$

$\lambda_A = \frac{p^1_A - p^2_A}{\sqrt{2p_A(1-p_A)}}$ ~~N=1600~~

$\alpha = 0.05$

power = $\Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N})$

-4

4. (25 points) Now assume that we are performing an association at SNP A and while the causal mutation is at SNP B. Assume the correlation coefficient between SNPs A and B is r^2. Show power of detecting the association at SNP A by genotyping $\frac{N}{r^2}$ individuals is equal to the power of detecting the association if we genotyped SNP B with N individuals. Make sure you include all steps discussed in the lecture.

$$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}} \qquad \lambda_B = \frac{p_B^+ - p_B^-}{\sqrt{2p_B(1-p_B)}}$$

derivation

$p_A^+ = p_{AB} + p_{Ab}$ one more step -2

Define p_A^- -2

$p_B^+ p_{AB} - p_B^+ p_{Ab}$

$-p_B^+ p_{AB} + p_B^- p_{Ab}$

$p_B^- (-p_{AB} + p_{Ab})$

$p_B - p_{AB} \quad (1-p_B) - p_{Ab}$

due to conditional probability distribution

$$\lambda_A = \frac{(p_B^+ - p_B^-)(p_{AB} - p_{Ab})}{\sqrt{2p_A(1-p_A)}}$$

multiply by 1.

$$\lambda_A = \frac{(p_B^+ - p_B^-)(p_{AB} - p_{Ab})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}\sqrt{2p_B(1-p_B)}}$$

$$\lambda_A = (\lambda_B)\frac{(p_{AB} - p_{Ab})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

$$= \lambda_B \frac{\left(\frac{p_{AB}}{p_B} - \frac{p_{Ab}}{(1-p_B)}\right)\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

$$= \lambda_B \frac{\left(\frac{p_{AB}(1-p_B) - p_{Ab}p_B}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

$$= \lambda_B \frac{(p_{AB} - p_{AB}p_B - p_{Ab}p_B)}{\sqrt{p_A(1-p_A)}\sqrt{p_B(1-p_B)}}$$

$$= \frac{\lambda_B (p_{AB} - p_A p_B)}{\sqrt{p_A(1-p_A)}\sqrt{p_B(1-p_B)}} = \lambda_B\sqrt{r^2} \rightarrow$$

$p_A^+ - p_A^- = (p_B^+ - p_B^-)(p_{AB} - p_{Ab})$

$p_A^+ - p_A^- = p_{AB} - p_B^+ p_{Ab} - p_B^- p_{AB} + p_B^- p_{Ab}$

$= \underbrace{p_{AB} + p_{Ab}}_{p_A^+} \underbrace{- p_B^+ p_{Ab} - p_B^- p_{AB}}$

$\frac{p_{Ab}}{(1-p_B)} + \frac{p_{AB}}{p_B}$

$p_A^- = p_{Ab} + p_{AB}$

$\lambda_A = \lambda_B r^2 \Rightarrow \lambda_A\sqrt{N_A} = \lambda_B\sqrt{N_B}$

$\lambda_B\sqrt{r^2}\sqrt{N_A} = \lambda_B\sqrt{N_B}$

$r^2 N_A = N_B \rightarrow N_A = \frac{N_B}{r^2}$

5 Relative Risk (5 points)

Assume we are studying a rare disease with a disease prevalence rate ~ 0. Let a SNP A be a causal SNP of this disease with a relative risk of 2.0. The true population minor allele frequency of A is $P_a = 0.2$. what are the true population minor allele frequencies in the case population (p_a^+) and in the control population (p_a^-)?

$$p^+_a = \frac{\gamma(P_a)}{(\gamma-1)P_a+1} = \frac{(2)(0.2)}{0.2+1} = \frac{0.4}{1.2}$$

$$p^-_a = P_a = 0.2$$

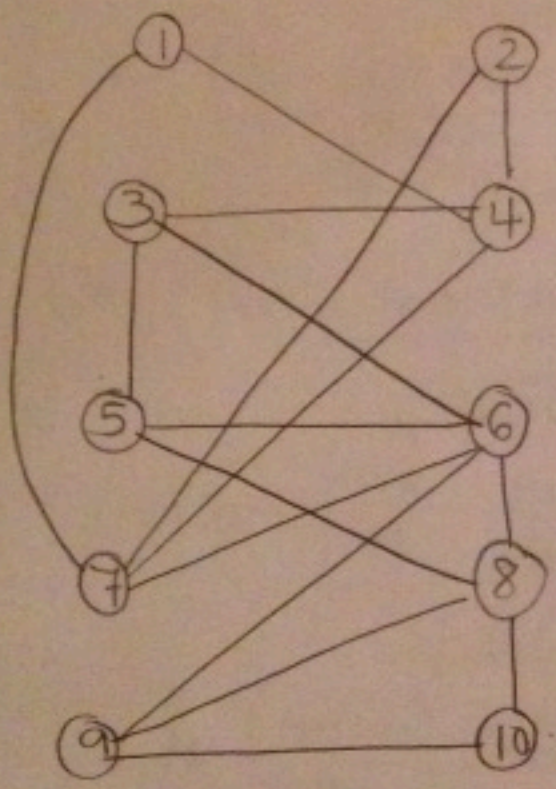
6 Tag SNP Selection (10 points)

We are given the following matrix of corelations, r, between 10 SNPs.

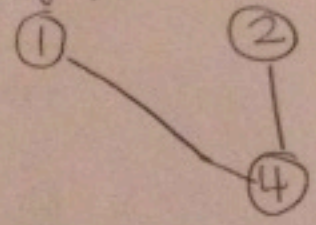
	1	2	3	4	5	6	7	8	9	10
1	1	0.1	0.2	0.8	0.2	0.2	0.9	0.2	0.1	0.2
2		1	0.5	0.95	0.2	0.1	0.9	0.1	0.2	0.1
3			1	0.9	0.8	0.75	0.5	0.5	0.3	0.2
4				1	0.1	0.5	0.85	0.6	0.3	0.2
5					1	0.75	0.6	0.75	0.6	0.5
6						1	0.9	0.8	0.85	0.3
7							1	0.5	0.6	0.4
8								1	0.95	0.75
9									1	0.8
10										1

1. Use the greedy algorithm to find a minimum set of tag SNPs with $r \geq 0.7$.
2. Is the greedy solution the optimal solution? If not, what is the optimal solution?

Please show your work for both problems by drawing graphs before and after you choose each tag SNP.



choose ⑥ tags: 6



10

choose ④ tags: 6,4

10

choose ⑩ tags 6,4,10

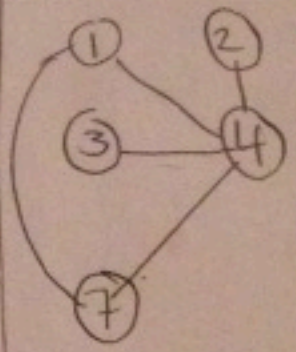
greedy

OPTIMAL:

greedy is not optimal

the optimal is {8,4}

choose ⑧ tags: 8



choose 4 tags: 8,4

-4

3 MultiSNP Power (15 points)

Assume that we collect 5 independent SNPs. 3 have minor allele frequency (MAF) of 0.4 and 2 have MAF of 0.2. Assume that relative risk of one of them is 2 (we do not know which one). Assume that we are collecting 300 case and 300 control individuals. With $\alpha = 0.05$, what is the power of this association study?

Provide your answers in terms of $\Phi(x)$ and $\Phi^{-1}(x)$ or pnorm and qnorm in R.

MAF = 0.4:

$\gamma = 2$

$p_A^+ = \frac{\gamma(0.4)}{(\gamma-1)(0.4)+1} = \frac{0.8}{1.4}$ ✓

$p_A^- = 0.4$ ✓

$p_A = \frac{p_A^+ + p_A^-}{2}$ ✓

$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}$ ✓

$power_{0.4} = \Phi(\Phi^{-1}(0.05/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(0.05/2) + \lambda_A\sqrt{N})$ ✓

where $N = \frac{1200}{600}$ -1

MAF = 0.2:

$\gamma = 2$

$p_A^+ = \frac{\gamma(0.2)}{(\gamma-1)(0.2)+1} = \frac{0.4}{1.2}$ ✓

$p_A^- = 0.2$ ✓

$p_A = \frac{p_A^+ + p_A^-}{2}$ ✓

$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}$ ✓

$\alpha_S = \frac{\alpha}{M}$ -3

$power_{0.2} = \Phi(\Phi^{-1}(0.05/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(0.05/2) + \lambda_A\sqrt{N})$

where $N = \frac{1200}{600}$

the total power $= \frac{3(power_{0.4}) + 2(power_{0.2})}{5}$ ✓

3. (25 points Graduate Student Only) Now assume that there are $N^+/2$ case individuals and $N^-/2$ control individuals in the association study. Derive a new statistic that follows the standard normal distribution. What is the power of such a study compared to a study with N individuals?

Provide your answers in terms of $\Phi(x)$ and $\Phi^{-1}(x)$ or pnorm and qnorm in R.

-1

1 Computing Association Statistics (10 points)

We genotype SNP A with alleles {A,a}. After we sample 400 case and 400 control individuals, which gives us a total of 800 case chromosomes and 800 control chromosomes, we observe 450 allele As in the cases and 400 allele As in the controls. Given $\alpha = 0.05$, we want to test whether to reject or accept the null hypothesis. Let the null hypothesis be that SNP A is not associated with the target disease. Using the test framework we learned in the class, provide an inequality test statement such that we reject the null hypothesis if the statement is true or we accept the null hypothesis if the statement is false.

Provide your answers in terms of $\Phi(x)$ and $\Phi^{-1}(x)$ or pnorm and qnorm in R.

$\hat{p}_A^+ = \frac{450}{\cancel{1600}\ 800}$ $\hat{p}_A^- = \frac{400}{\cancel{1600}\ 800}$ $\hat{p}_A = \frac{p_A^+ + p_A^-}{2}$ ✓

$S_A = \frac{p_A^+ - p_A^-}{2p_A(1-p_A)}$ ✓ ✓

S_A is significant & we can reject the null hypothesis if

$S_A < \Phi(\alpha/2)$ or $S_A > -\Phi(\alpha/2)$

-2

4 Derivation Question

1. (15 points) Given N/2 case individuals and N/2 control individuals. $\hat{p}_A^+$ and $\hat{p}_A^-$ are the observed frequencies. If the true frequencies are p_A^+ and p_A^-, show that the difference of the observed frequencies is normally distributed with mean μ and variance σ^2.

case/control chromosomes

-2

$$\hat{P}_A^+ \sim N(p_A^+, p_A^+(1-p_A^+)/N) \checkmark$$

$$\hat{P}_A^- \sim N(p_A^-, p_A^-(1-p_A^-)/N) \checkmark$$

$$\hat{P}_A^+ - \hat{P}_A^- \sim N(p_A^+ - p_A^-, (p^+(1-p_A^-) + p_A^-(1-p_A^-))/N) \checkmark$$