100/100

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. pnorm(x) implicitly takes mean 0 and variance 1.

$\hat{p}_A^+ = \frac{450}{800}$ $\quad$ $\hat{p}_A^- = \frac{400}{800}$ $\quad$ $\hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = \frac{425}{800}$

$S = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2\hat{p}_A(1-\hat{p}_A)}}\sqrt{800}$ $\quad$ $\alpha = 0.05$

if $\boxed{S > -\Phi^{-1}(\frac{\alpha}{2}) \quad \text{or} \quad S < \Phi^{-1}(\frac{\alpha}{2})}$

reject null hypothesis

10

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

$p_A^+ = 0.5 \quad p_A^- = 0.3 \quad \alpha = 0.05 \quad N = 800 \quad p_A = \frac{p_A^+ + p_A^-}{2} = 0.4$

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

$$\text{Power} = \Phi\left(\Phi^{-1}\left(\tfrac{\alpha}{2}\right) + \lambda_A\sqrt{N}\right) + \left[1 - \Phi\left(-\Phi^{-1}\left(\tfrac{\alpha}{2}\right) + \lambda_A\sqrt{N}\right)\right]$$

10

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.

$\alpha_s = \frac{\alpha}{5} = 0.01$ $\quad N = 600 \quad \gamma = 2$

for $p=0.4$: $\quad p_A^+ = \frac{\gamma p}{(\gamma-1)p+1} = \frac{0.8}{0.4+1} = \frac{0.8}{1.4} \quad p_A^- = p = 0.4 \quad p_A = \frac{p_A^+ + p_A^-}{2}$

$$\lambda_{p=0.4} = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}$$

for $p=0.2$: $\quad p_B^+ = \frac{\gamma p}{(\gamma-1)p+1} = \frac{0.4}{0.2+1} = \frac{0.4}{1.2} = 0.333 \quad p_B^- = p = 0.2 \quad p_B = \frac{p_B^+ + p_B^-}{2}$

$$\lambda_{p=0.2} = \frac{p_B^+ - p_B^-}{\sqrt{2p_B(1-p_B)}}$$

Power for $p=0.4$: $\quad P_{p=0.4} = \Phi\left(\Phi^{-1}\left(\frac{\alpha_s}{2}\right) + \lambda_{p=0.4}\sqrt{N}\right) + \left[1 - \Phi\left(-\Phi^{-1}\left(\frac{\alpha_s}{2}\right) + \lambda_{p=0.4}\sqrt{N}\right)\right]$

Power for $p=0.2$: $\quad P_{p=0.2} = \Phi\left(\Phi^{-1}\left(\frac{\alpha_s}{2}\right) + \lambda_{p=0.2}\sqrt{N}\right) + \left[1 - \Phi\left(-\Phi^{-1}\left(\frac{\alpha_s}{2}\right) + \lambda_{p=0.2}\sqrt{N}\right)\right]$

Total power: $\quad \text{Power} = \frac{3 \cdot P_{p=0.4} + 2 \cdot P_{p=0.2}}{5}$

15

$\frac{100}{100}$ good job!

4 Derivation Question (75 points)

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.

$N\hat{p}_A^+ \sim N(Np_A^+, Np_A^+(1-p_A^+))$ by normal approximation of binomial distribution

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

$$\boxed{\hat{p}_A^+ - \hat{p}_A^- \sim N(p_A^+ - p_A^-, [p_A^+(1-p_A^+) + p_A^-(1-p_A^-)]/N)}$$

Linear combination of 2 normal distributions is normal

(ex: For $N = N_A - N_B$, $\mu = \mu_A - \mu_B$ and $\sigma^2 = \sigma_A^2 + \sigma_B^2$)

15

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?

$$\hat{p}_A^+ - \hat{p}_A^- \sim N\left(p_A^+ - p_A^-, \left[p_A^+(1-p_A^+) + p_A^-(1-p_A^-)\right]/N\right).$$

Let $\hat{p}_A = \dfrac{\hat{p}_A^+ + \hat{p}_A^-}{2}$. Then $p_A^+(1-p_A^+) + p_A^-(1-p_A^-) \approx 2p_A(1-p_A)$.

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

Make $\sigma^2 = 1$:

Let $S = \dfrac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2p_A(1-p_A)}}\sqrt{N}$. Then $S \sim N\left(\dfrac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}\sqrt{N}, 1\right)$. $S \sim N(\lambda_A\sqrt{N}, 1)$

$\Rightarrow$ The mean of $S = \lambda_A\sqrt{N} = \dfrac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}\sqrt{N}$

10

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 (N/2 case and N/2 control individuals)?
Provide your answers in terms of $\Phi(x)$ and $\Phi^{-1}(x)$ or pnorm and qnorm in R.

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 (Start with the distributions of S_A and S_B and derive the relationship between λ_A and λ_B).

$S_A \sim N(\lambda_A\sqrt{N_A}, 1)$ $S_B \sim N(\lambda_B\sqrt{N_B}, 1)$. For equal power, $\lambda_A\sqrt{N_A}$ must equal $\lambda_B\sqrt{N_B}$. $N_A = \frac{N}{r^2}$ and $N_B = N$. are given, requiring:

$\Rightarrow \lambda_A\sqrt{\frac{N}{r^2}} = \lambda_B\sqrt{N} \Rightarrow \lambda_A = \lambda_B\sqrt{r^2}$. Lets prove this!

$p_A^+ = p_{AB}^+ + p_{A\neg B}^+$ (law of total prob.) and $p_A^- = p_{AB}^- + p_{A\neg B}^-$

Because we can assume the conditional probability distributions are the same for case and control:

$p_A^+ = p_B^+ p_{A|B} + (1-p_B^+)p_{A|\neg B}$. Similarly, $p_A^- = p_B^- p_{A|B} + (1-p_B^-)p_{A|\neg B}$ (by law of conditional probability)

$p_A^+ - p_A^- = p_B^+ p_{A|B} + (1-p_B^+)p_{A|\neg B} - p_B^- p_{A|B} - (1-p_B^-)p_{A|\neg B}$.

$p_A^+ - p_A^- = p_{A|B}(p_B^+ - p_B^-) - p_{A|\neg B}(p_B^+ - 1 + 1 - p_B^-) = p_{A|B}(p_B^+ - p_B^-) - p_{A|\neg B}(p_B^+ - p_B^-)$.

$p_A^+ - p_A^- = (p_B^+ - p_B^-)(p_{A|B} - p_{A|\neg B})$.

We defined $\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}$ and $\lambda_B = \frac{p_B^+ - p_B^-}{\sqrt{2p_B(1-p_B)}}$.

$$\lambda_A = \frac{(p_B^+ - p_B^-)(p_{A|B} - p_{A|\neg B})}{\sqrt{2p_A(1-p_A)}} = \frac{\lambda_B(p_{A|B} - p_{A|\neg B})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}} = \frac{\lambda_B\left(\frac{p_{AB}}{p_B} - \frac{p_{A\neg B}}{1-p_B}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

(by law of conditional probability)

$$\lambda_A = \frac{\lambda_B\left(\frac{p_{AB} - p_{AB}p_B - p_{A\neg B}p_B}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}} = \frac{\lambda_B(p_{AB} - p_B(p_{AB} + p_{A\neg B}))}{\sqrt{p_A(1-p_A)}\cdot\sqrt{p_B(1-p_B)}}$$

$$\lambda_A = \frac{\lambda_B(p_{AB} - p_A p_B)}{\sqrt{p_A(1-p_A)}\cdot\sqrt{p_B(1-p_B)}} = \lambda_B\sqrt{r^2}$$ (by definition of r^2)

25

5 Relative Risk (5 points)

Assume we are studying a rare disease with a disease prevalence rate approximately near 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^-)?

$\gamma = 2 \qquad P_A = 0.2$

$$p_A^+ = \frac{\gamma P_A}{(\gamma - 1)P_A + 1} = \frac{0.4}{0.2+1} = \frac{0.4}{1.2} = 0.333$$

$p_A^- = P_A = 0.2$ because prevalence rate ≈ 0

$p_A^+ = 0.333 \qquad p_A^- = 0.2$

5

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

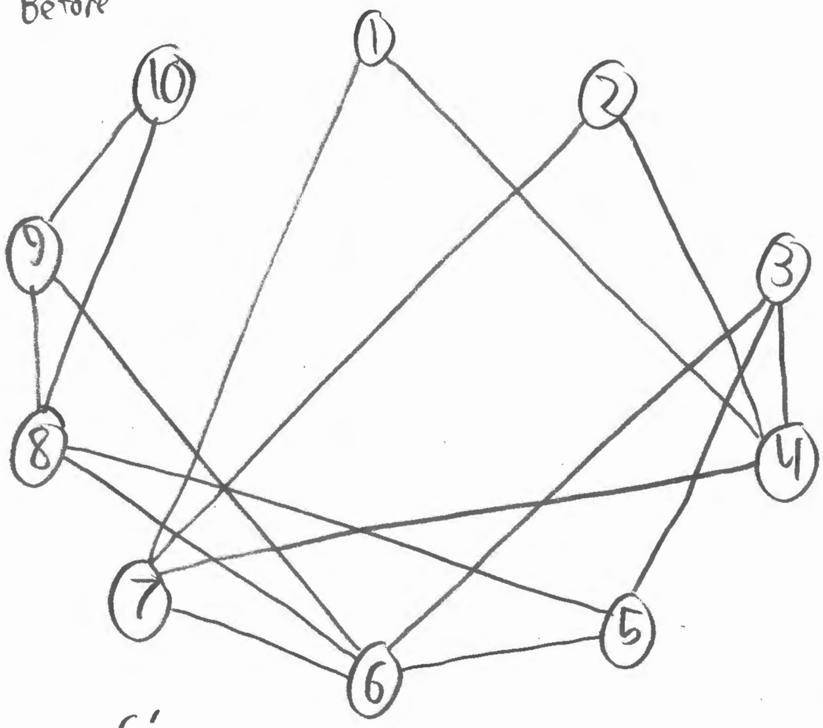
1. Use the greedy algorithm to find a minimum set of tag SNPs with $r \geq 0.7$.
2. Is the greedy solution the optimal solution? If not, what is the optimal solution?

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

Edge $\{v,w\}$ if $r_{vw} \geq 0.7$

1) Before

G



Greedy: Pick tag with most SNPs $r \geq 0.7$. Add this tag and its SNPS to new G' and remove from G. Repeat until all SNPs in G' ✓

Greedy's choices:

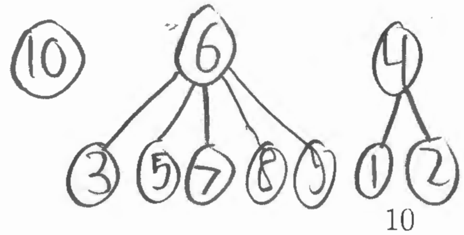
6, takes 3,5,7,8,9 ✓

4, takes 1,2

10.

G'

After:



with tags 10, 6, and 4

Optimal on back page

10