Homework 2

CS/HG 124/224

Due: April 22, 2014

- All assignments must be submitted by a hard copy to TA's office by 4 pm on the day they are due.
- Please indicate whether you are an undergraduate or a graduate student on your HW.
- Show all your work.

1 Multiple Hypothesis Testing

In class, we talked about two methods to correct for multiple hypothesis testing, Sidak and Bonferroni. Consider a multi-SNP association study where one is interested in looking for *any* SNP that is associated with a disease phenotype with a probability of 0.05 or 0.01. Compute the thresholds for association at each individual SNP if the researcher decides to consider 2, 5, 10, 100 and 1000 SNPs using both Sidak and Bonferroni corrections. Assume that the SNPs are independent.

		Significant Threshold			
		Sidak		Bonferroni	
		0.05	0.01	0.05	0.01
	2				
	5				
Number of SNPs	10				
	100				
	1000				

2 Tag SNP Selection Problem

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

	1	2	3	4	5	6	7	8	9	10
1	1	0.9	0.85	0.5	0.4	0.2	0.2	0.15	0.15	0.1
2		1	0.95	0.5	0.8	0.2	0.2	0.15	0.15	0.1
3			1	0.65	0.9	0.7	0.5	0.5	0.3	0.2
4				1	0.85	0.5	0.85	0.6	0.7	0.7
5					1	0.75	0.6	0.75	0.6	0.5
6						1	0.6	0.75	0.4	0.3
7							1	0.8	0.85	0.8
8								1	0.6	0.5
9									1	0.5
10										1

2.1 Computing Power

Assume that we collect all 10 SNPs and the minor allele frequency (MAF) of SNPs 1 to 5 is 0.3 and MAF of SNPs 6 to 10 is 0.15. Assume that the relative risk of one of them is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?

2.2 Greedy algorithm

2.2.1 Finding Tag SNPs

Use the greedy algorithm to find a minimum set of tag SNPs with a $r \geq 0.7$. Please show your work by drawing graphs before and after you choose each tag SNP.

2.2.2 Computing Power

Assume that the relative risk of one of tag SNPs in the greedy solution is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?

2.3 Optimal algorithm

2.3.1 Finding Tag SNPs

The greedy solution for finding the minimum set of tag SNPs is not the optimal solution. What is the optimal solution? Please show your work by drawing graphs before and after you choose each tag SNP.

2.3.2 Computing Power

Assume that the relative risk of one of tag SNPs in the optimal solution is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?

3 Indirect Association Study Problem

3.1 Calculating Correlation

Let's assume that we have a following reference dataset of 10 individuals representing a population such as the HapMap. What is the correlation, r, between SNP A and SNP B?

Individuals	SNP A	SNP B
Individual 1	A	A
Individual 2	a	a
Individual 3	A	A
Individual 4	A	a
Individual 5	a	a
Individual 6	A	A
Individual 7	a	A
Individual 8	A	A
Individual 9	A	a
Individual 10	A	A

3.2 Indirect Association Power

Assume the causal SNP is B, but we collect SNP A. Assume that true case probability and and true control probability are 0.4 and 0.5 respectively at SNP B. If we collect 500 case and 500 control individuals and have a significance threshold of 0.05, what is the power at SNP A? (Note : Use the correlation that you get from above question)

4 Association Study with Multiple Disease (Grad Students Only)

We know from the homework, that the most efficient association studies have the same number of cases and controls. The Wellcome Trust Case Control Consortium used 2000 cases and 3000 controls for each of their disease associations. If you use the formula from the homework, this turns out to be equivalent to an balanced case/control study with 2400 each. So in essence, they used 5000 people but only got the equivalent power of using 4800.

However, what they did was have 7 diseases where they collected 2000 cases and they used the same 3000 controls for each association study. So they effectively used the 3000 controls many times while the each cases individual was only used once. They collected a total of 7*2000+3000=17000 individuals.

Now the question is did they collect the right number of cases and controls in this kind of scenario? If not, how many should they have collected. What if there were only 3 diseases (the total number of individuals is 3*2000+3000 = 9000)? How about 10 diseases (the total number of individuals is 10*2000+3000 = 23000)?