Extra practice problems for final exam:

Note: these practice problems serve to cover material taught by me (GWAS and QTLs) that was NOT covered already in problem sets 5 and 6, exam III, and the previous practice problems I released. The remaining material is also relevant for the final exam, but you already have a variety of problems that cover it.

Good luck,
Aviv

Problem 1

You study the genetic basis of heart disease in the Framingham Heart Study. Within the cohort, you have identified 140 cases (with heart disease) and 200 controls (without heart disease).

1. You genotype the gene ApoE2, and obtain the following results:

	Cases	Controls
11	**60**	**60**
10	**50**	**50**
00	**30**	**90**
Totals	**140**	**200**

Estimate whether the ApoE2 genotype is significantly associated with heart disease (chi-squared values are below). Take P<0.01 as significant.

SOLUTION:
Null hypothesis: ApoE2 genotype is not associated with heart disease.
Alternative hypothesis: ApoE2 genotype is associated with heart disease.

We use the chi-squared test to estimate the deviation of the observed values from the expected ones given no association.

	Case-expected	Controls-expected
11	49.41176471	70.58823529
10	41.17647059	58.82352941
00	49.41176471	70.58823529

(O-E)^2 (cases)	(O-E)^2 (controls)		(O-E)^2/E (cases)	(O-E)^2/E(controls)		Sum((O-E)^2/E)
112.11	112.11		2.268907563	1.588235294		**20.03571429**
77.85	77.85		1.890756303	1.323529412		
376.82	376.82		7.62605042	5.338235294		

The chi-square value is 20.03. We compare to the critical values with **df=2** and find that this value is above the critical threshold for $p<10^{-4}$ (but below that for 10^{-5}). We therefore reject the null hypothesis and take the association as significant.

Note: the results of (O-E)^2 are always symmetric for each genotype between the cases and the controls so you can calculate just one, **but remember**: **they each need to be divided by a different expected!**

2 You now realize that for privacy reasons, the actual genotypes have been scrambled but you can still work with the allele frequencies. As a result your table is

	Cases	Controls
1	**170**	**170**
0	**110**	**230**
Totals	**280**	**400**

Estimate whether allele 1 is significantly associated with heart disease. Take P<0.01 as significant.

SOLUTION:

Null hypothesis: Allele 1 is not associated with heart disease.
Alternative hypothesis: Allele 1 is associated with heart disease.

We first calculate the OR for allele 1:
OR (Allele 1)= (170*230)/(110*170) 2.090909

To estimate significance, we compute the chi-squared statistic. In this case we can use the 'shortcut' formula:

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

chi-squared =((170*230-110*170)^2*680)/(340*340*280*400) 21.85714286

Notice, that had we calculated chi-squared as in (1) we would have reached the same value:

	Case expected	Controls expected	(O-E)^2 (cases)	(O-E)^2 (controls)	(O-E)^2/E (cases)	(O-E)^2/E(controls)	Sum((O-E)^2/E)
1	140	200	900	900	6.428571429	4.5	**21.85714286**
0	140	200	900	900	6.428571429	4.5	

The chi-square value is 21.85. We compare to the critical values with **df=1** and find that this value is above the critical threshold for $p<10^{-5}$ (but below that for 10^{-6}). We therefore reject the null hypothesis and take the association of the allele as significant.

3. What is the required level of nominal P-values from an individual test to achieve genome-wide significance of 0.01 if you test 4 million loci?

SOLUTION:

To achieve genome-wide significant of 0.01 when performing $4x10^6$ tests, we use the Bonferoni correction and accept only nominal p-values below

$0.01/4x10^6=0.25x10^{-8}$

4. Given your answer in 3, how would your significance estimate change in 1 and 2 if the results had come as part of measuring 6 million genotypes, rather than a single locus?

SOLUTION:

Neither result would be significant at the genomewide level.

5. You now study the same locus using a trio design. You have recruited 500 families with 600 heterozygous parents (i.e. in 400 trios there is only one heterozygous parent and in 100 families both parents are heterozygous). You find that 350 heterozygous parents transmitted the 1 allele and 250 transmitted the 0 allele. Estimate the significance of association of the 1 allele with heart disease.

SOLUTION:

We rely on the TDT test using the chi-squared statistic computed by:

Null hypothesis: Allele 1 is not associated with heart disease.
Alternative hypothesis: Allele 1 is associated with heart disease.

$$\chi^2_{TDT} = \frac{(n_1 - n_2)^2}{(n_1 + n_2)}$$

n_1 will be 350
n_2 will be 250

so the chi-square value will be

chi-squared=(350-250)^2/(350+250)= 16.66666667

The chi-square value is 16.66. We compare to the critical values with **df=1** and find that this value is above the critical threshold for $p<10^{-4}$ (but below that for 10^{-5}). We therefore reject the null hypothesis and take the association of the allele as significant.

p-value	df=1	df=2
0.1	2.705544	4.60517
0.01	6.634897	9.21034
10^{-3}	10.82757	13.81551
10^{-4}	15.13671	18.42068
10^{-5}	19.51142	23.02585
10^{-6}	23.92813	27.63102

Problem 2

You study the genetic basis of Crohn's disease. Your colleague at MGH has identified 200 cases (with Crohn's) and 350 controls (without Crohn's). You perform a genome-wide study of 500,000 loci.

1. The best-scoring locus, in chromosome 5, had the following results:

	Cases	Controls
11	**55**	**125**
10	**80**	**75**
00	**65**	**150**
Totals	**200**	**350**

Estimate whether the locus' genotype is significantly associated with Chron's (chi-squared values are below) at a genome-wide significance level of 0.01.

SOLUTION:

Null hypothesis: The locus' genotype is not associated with Crohn's.
Alternative hypothesis: The locus' genotype is associated with Crohn's.

We use the chi-squared test to estimate the deviation of the observed values from the expected ones given no association.

	Case-expected	Controls-expected	(O-E)^2 (cases)	(O-E)^2 (controls)	(O-E)^2/E (cases)	(O-E)^2/E (controls)	Sum((O-E)^2/E)
11	65.45454545	114.5454545	109.30	109.30	1.669823232	0.954184704	**21.69256972**
10	56.36363636	98.63636364	558.68	558.68	9.91202346	5.664013406	
00	78.18181818	136.8181818	173.76	173.76	2.222515856	1.270009061	

The chi-square value is 21.69. We compare to the critical values with **df=2** and find that this value is above the critical threshold for $p<10^{-4}$ (but below that for 10^{-5}).

However, we need to consider the 500,000 tests we did. To achieve genome-wide significant of 0.01 when performing $5x10^5$ tests, we use the Bonferoni correction and accept only nominal p-values below

$0.01/5x10^5=0.2x10^{-7}$

Our test was only significant at a nominal level between 10^{-4} and 10^{-5}. We therefore CANNOT reject the null hypothesis at a genome wide significance level of 0.01.

2. What would have happened if you only worked with alleles, not genotypes? Would allele 1 be significantly associated with Crohn's and if so at which level? Assuming that you did NOT perform multiple tests, what could be the reason to the discrepancy between your result in 1 and 2?

SOLUTION:

Null hypothesis: Allele 1 is not associated with heart disease.
Alternative hypothesis: Allele 1 is associated with heart disease.

We first derive the appropriate contingency table from the genotype information in (1):

	Cases	Controls
1	**190**	**325**
0	**210**	**375**
Totals	**400**	**700**

We now calculate the OR for allele 1:

OR (Allele 1)= (190*375)/(325*210)= 1.043956044

This suggests that the allele is not associated with risk by this measure. To estimate our confidence we perform a chi-squared test, using our shortcut formula

chi-squared =((190*375-325*210)^2*1100)/(515*585*400*700)= 0.11735837

The chi-square value is 0.11. We compare to the critical values with **df=1** and find that it is below the critical threshold for p<0.1. We therefore CANNOT reject the null hypothesis.

If we had conducted only one test each, then the genotype would have been significantly associated, but the allele would not have. This is likely because the heterozygous genotype carries much of the risk, rather than an individual allele.

p-value	df=1	df=2
0.1	2.705544	4.60517
0.01	6.634897	9.21034
10^{-3}	10.82757	13.81551
10^{-4}	15.13671	18.42068
10^{-5}	19.51142	23.02585
10^{-6}	23.92813	27.63102

Problem 3

You study abdominal bristles in flies. You cross two inbred lines of *Drosophila* one with a mean bristle number of 30 and the other with 50, both with a standard deviation of 4.

1. You obtain F1s with a mean number of 70 bristles and a standard deviation of 8. Are these results roughly consistent with an additive model of genotypic variance? Explain.

SOLUTION:

They are not. Under the additive model we expect the mean of the F1s to be approximately the average of the parental means (~40) and the standard deviation to be similar to the parental one (4), since it reflects environmental variance, not genotypic variance.

(Note: there are statistical tests that we could use, had we known the sample size, to formally estimate the significance of these differences).

2. You now realize that your lab mate has mixed the fly labels, and that in fact you have produced the F1s from a cross of two other lines: one with a mean bristle number of

60, the other with 80 and a variance of 16. Are these results roughly consistent with an additive model?

SOLUTION:

Yes, they are: the mean in F1s (70) is the average of the parental means. The standard deviation in the F1s is $\sqrt{16}=4$, same as that in the parental strains.

3. You proceed to produce F2s, which have mean bristle size of 70 and a standard deviation of 5. What are the environmental variance, genotypic variance and broad-sense heritability of bristle number?

SOLUTION:

In F1's the genotypic variance (σ^2_g) should be zero (they are genetically identical) and thus $\sigma^2_e = \sigma^2_p - 0 = 16-0 = 16$.

In the F2's, $\sigma^2_g = 25 - 16 = 9$.

The broad sense heritability $H^2=\sigma^2_g / \sigma^2_p = 9/25 = 0.36$.

4. What is the minimal number of genes that may be affecting bristle number in your fly population?

SOLUTION:

To estimate the minimal number of genes that may be affecting bristle number we compute

$$n = \frac{D^2}{8\sigma_g^2}$$

n=(80-60)^2/(8*9) 5.555555556

At least 5 or 6 genes affect bristle number.

Problem 4

You are interested in the heritability of longevity.

1. You start by studying longevity in mice. Since lab strains appear to have little variation in lifespan, you decide to conduct a large scale breeding experiment in a wild population of mice captured in Boston. In your captured population, the mean life span is 2 years, and the standard deviation is 0.5 year. You choose as your truncation point 3 years, and obtain a population with a mean of 4 years. The

offspring of the selected animals have a mean life span of 2.6 years. What is the narrow sense heritability of lifespan in your mice?

SOLUTION:

To calculate the narrow sense heritability we use:

$$h^2 = \frac{M' - M}{M^* - M}$$

In our case M=2, M'=2.6, and M*=4

Hence, h^2=(2.6-2)/(4-2)=0.3

2. You decide to repeat the selection process, choosing a new truncation of 3.5 years, obtaining a population (from the offspring) with a mean life span of 4.2 years. What is the expected mean life span of <u>their</u> offspring?

SOLUTION:

To calculate the expected life span of the offspring we use the narrow sense measure we calculated in (1)

M''=M'+ h^2(M*'-M') = 2.6+0.3*(4.2-2.6) = 3.08

The mean expected lifespan in the next generation is 3.08 years.

3. You wish to compare your estimates to ones from human studies. You write your colleague, a human geneticist, who sends you data on longevity from a study of identical twins. The correlation coefficient of longevity between identical twins is 0.3. What is the estimated broad-sense heritability of longevity according to this data?

SOLUTION:

The correlation coefficient in a quantitative trait between identical twins is approximately equal to the broad sense heritability. We therefore estimate **H^2=0.3**.

4. While you are preparing your study for publication, another study of longevity in humans is published, this one based on measures from full-siblings. The correlation coefficient reported by that study was 0.1. What is the broad-sense heritability of longevity based on your competitor's study? How can you reconcile your finding in (3) with this new one?

SOLUTION:

The correlation coefficient in a quantitative trait between full siblings is approximately equal to half the broad sense heritability. We therefore estimate **H^2=0.2**.

One possible source of a higher similarity between identical twins is that their environments are also more similar than those of full siblings. They **(1)** often share embryonic membranes, **(2)** are treated more similarly by other people, **(3)** grew up exactly at the same time, and **(4)** are always the same sex. For example, to control for the factor **(4)** we should compare to measures of heritability between siblings of the same sex. To control for **(3)** we should compare to measures of heritability between fraternal twins.