

## Homework #6

Due Thursday Feb. 28 [**NOTE unusual due date!**] at the beginning of class. Assignments turned in more than 5 minutes after the beginning of class will be penalized 10 points, with an additional 10 points every 24 hours thereafter. You may discuss the homework assignment with other students, but do not share your work. **All Python programs should be run before being turned in. Even experienced programmers can seldom write a program perfectly on the first try.**

A new syndrome has been recognized among young children, and a researcher wants to determine whether it is genetic, environmental, or a mix of the two. She surveys 20 pairs of monozygotic (identical) twins and 20 pairs of dizygotic (fraternal) twins with at least one twin affected. Concordance is 19/20 monozygotic twins and 20/20 dizygotic twins. She also surveys siblings of affected children, and finds that 10% are affected.

1. (10 points) Assuming that her data collection method was sound, is this syndrome likely to be mostly genetic, mostly environmental, or a mix? Why? *It is likely to be mostly environmental, probably related to the uterine environment. If it had a strong genetic component, we would not expect such high dizygotic twin concordance alongside such low sibling concordance, since dizygotic twins are genetically no closer than other siblings.*
2. (10 points) Describe a data-collection problem that could cast doubts on her results. *This conclusion could be invalid if she used a method for finding her twins that had a bias toward concordant twins, such as advertising for "twins with this syndrome." Similar problems arise if it is easier to diagnose the syndrome when both twins have it than when only one twin has it.*

For the following three questions, determine how many haplotypes the two siblings share IBS (identical by state) and list all possibilities for sharing IBD (identical by descent). For example, you might say "These sibs share 1 haplotype IBS, and may share 1 or 0 IBD."

1. (5 points) First sibling has AB; second has CD. *0 IBS, 0 IBD*
2. (5 points) First sibling has AA; second has AB. *1 IBS, 0 or 1 IBD*
3. (5 points) First sibling has AB; second has AC. *1 IBS, 0 or 1 IBD*
4. (5 points) First sibling has AA; second has AA. *2 IBS, 0, 1 or 2 IBD*

Briefly critique the following data collection methods. What problems might they introduce into a mapping study?

1. (10 points) Contact all members of a patients' support group for your disease and ask them to participate in your study. *Support groups are likely to be enriched for families with multiple cases and for people with more severe disease. This may lead to wrong conclusions about the family clustering of the disease. Also, it will be hard to identify an appropriate control group, since it will be hard to tell how the patients in the support group were chosen.*
2. (10 points) Survey all patients admitted with your disease from a South Seattle hospital. Use healthy UW students as a control group. *UW has a very different mix of ages and ethnicities than South Seattle, and this will lead to bias and possible false conclusions. Also, depending on the disease, using hospital admissions may bias the result if there is variation in whether or not people are admitted. For example, females with heart attacks are less often hospitalized than males, which can skew estimates of heart disease rates in the two sexes.*

Stand-alone questions:

1. (20 points) Write a Python program which accepts two siblings (with single-letter haplotypes as shown above) and scores IBS. (You do not need to score IBD.) Test it on the sample file sibs.txt and include your test results with your homework. *There were clearly a lot of ways to do this, and I gave full credit to anything that worked, even if it looked completely different from the solution below. I did mark off, however, for programs that failed on cases such as "AB AB" and "AB BA" even though those were not in the sample data set.*

```
import sys
filename = sys.argv[1]
sibfile = open(filename,"r")
sibs = sibfile.readlines()
for sibpair in sibs:
    sib1 = sibpair[0:2]
    sib2 = sibpair[3:2]
    sib2reversed = sib2[1]+sib2[0]
    # cases with 2 matches
    if sib1 == sib2 or sib1 == sib2reversed :
        answer = 2
    # cases with 1 match
    elif sib1[0] == sib2[0] or sib1[0] == sib2[1] \
        or sib1[1] == sib2[0] or sib1[1] == sib2[1] :
        answer = 1
    else :
        # cases with 0 matches
        answer = 0
    print "sib pair", sib1, sib2, "shares", answer, "IBS"
```

2. (20 points) Write a Python program which reads observed values of P(AB),P(Ab),P(aB) and P(ab), in order, from a file. Calculate the expected values for these four numbers if linkage equilibrium were present. Compute the chi-square value comparing observed (the numbers read from the file) with expected (the numbers you calculated). You do not need to find the significance level of your chi-square. Hint:

$$\chi^2 = \sum [(observed - expected)^2 / expected]$$

*This problem is badly posed: you can't do a test of this kind on proportions, only on counts. Some of you assumed that the input data were counts, others read in or assumed a total; I accepted both. The solution below assumes that the input data are counts.*

```
import sys
filename = sys.argv[1]
datafile = open(filename,"r")
data = datafile.readline()
observed = data.split()
total = 0
for index in range(0,4) :
    observed[index] = float(observed[index])
    total += observed[index]

pAB = observed[0]/total
pAb = observed[1]/total
paB = observed[2]/total
```

```
pab = observed[3]/total

pA = pAB+pAb
pa = paB+pab
pB = pAB+paB
pb = pAb+pab

expected = []
expected.append(pA*pB*total)
expected.append(pA*pb*total)
expected.append(pa*pB*total)
expected.append(pa*pb*total)

chisquare = 0.0
for (o,e) in zip(observed,expected) :
    chisquare += (o-e)**2/e
print "chi-square is",chisquare
```