STAT 530 Bioinformatics: Homework 1

Due Feb 13, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF.

Problem 1 (5 points)

Using the notation on p. 64 of the slides. Show that under cross-sectional sampling, the estimate

$$\frac{a}{a+b} - \frac{c}{c+d}$$

accurately estimates the risk difference, but under case-control sampling it does not.

Problem 2 (5 points)

Suppose the outcome Y_i, genotype X_i, and other covariates $\boldsymbol{W}_i$ truly obey the logistic model

$$\text{logit}\, P(Y_i = 1 \mid X_i, \boldsymbol{W}_i) = \beta_0 + \beta_1 X_i + \boldsymbol{\gamma}^\top \boldsymbol{W}_i.$$

Suppose you have case-control sampled data. Do you think you can use it to estimate β_1? Explain.

Problem 3 (5 points)

Install Ubuntu on your computer. Open a terminal and report the outputs of both of the following commands:

- `head /proc/cpuinfo`
- `head /proc/meminfo`

Problem 4 (5 points)

Download the PLINK example data from `http://zzz.bwh.harvard.edu/plink/hapmap1.zip`. Load the data into PLINK and look at some basic summary statistics. How many cases and controls are there, and what is the total genotyping rate?

Problem 5 (5 points)

Suppose you measure 10 subjects with genotype AA, 25 with Aa, and 81 with aa. You now want to test whether the A and a alleles are in Hardy-Weinberg equilibrium. Use R to conduct a chi-square goodness-of-fit test for this hypothesis. Report the value of the test statistic and its p-value.