

Homework exercise 4

Due date: 7 June 2016 in class

1. The effect of population structure on heritability estimation.

Assume that a population is composed of two genetically distinct sub-populations, which almost never not interbreed (say Jews and Arabs). Furthermore assume both sub-populations are exposed to exactly the same environments, and that we are investigating heritability of a trait with a pure, clean architecture: $Y = \mu + G + E$, and G is also additive, so heritability is cleanly defined as

$$H^2 = \sigma_g^2 / \sigma_y^2.$$

Note the addition of μ , the overall (sub)-population mean explicitly to the expression. We assume that the two sub-populations have means μ_1, μ_2 , not necessarily equal. However, we do assume that within each of the two subpopulations H^2 is the same.

- (a) Will heritability be the same, higher or lower in the joint population made of the union of the two populations? You can base your answer mostly on simulations, but a theoretical argument (or at least convincing intuition) is also required.
- (b) Assume we are sampling pairs of twins from the joint population and using the standard estimate we discussed in class $\hat{H}^2 = 2(r_{MZ} - r_{DZ})$. Is this an estimate of the heritability within each sub-population or in the overall population? Explain

2. The effects of case-control sampling in the liability threshold model.

We assume the normal liability threshold model, i.e., $L = G + E$, with $G \sim N(0, \sigma_g^2)$, $E \sim N(0, 1 - \sigma_g^2)$ independent, and $Y = I\{L > t\}$ for $t = \Phi^{-1}(1 - K)$.

We are interested in investigating a balanced “case-control population”, i.e., one where the weight of observations for which $L > t$ is inflated to 0.5. Our first task is to draw random samples from this population (which are case-control samples in the original population), and then investigate the marginal and joint distributions.

We will work with $K = 0.001$, $H^2 = \sigma_g^2 = 0.5$, representative values for human disease.

- (a) Draw a case control sample by drawing a 10^6 sample from the original population (G,E,L,Y sets) and keeping all cases and $K/(1 - K)$ of controls (justify this choice). Also keep a random sample of size 2000 from the original population.
- (b) The code in <http://tau.ac.il/~saharon/StatGen2014/MCMC.r> implements a simple Markov Chain Monte Carlo (Metropolis) approach for sampling from the same case control distribution. Run it to generate a sample of 2000. Write the candidate distribution and acceptance rule as explicit formulae and explain them.
- (c) (* Extra credit) Experiment with different formulations for the candidate distribution. For those familiar with the topic, you can try asymmetric ones and Hastings rule, or even better – change the sampling probability and do importance sampling. Investigate the different approaches in terms of the variance of the estimates they generate of various quantities (like $E(L)$) in the case control population.

- (d) Compare the two samples in item a. and the case-control samples you drew in a. and b. according to the following criteria:
- Empirical mean of G, L .
 - Empirical variance of G, L .
 - Empirical heritability.
 - Empirical distribution of L (use `plot(density(...))` in R).

Comment on the effects of the case control sampling on these quantities, and try to explain it intuitively.

- (e) Calculate analytically the mean of L in the case-control population using the formula that for $X \sim N(0, 1)$ we have

$$E(X|X > t) = \frac{\phi(t)}{1 - \Phi(t)}.$$

- (f) (* Extra credit) Do the same for the variance of L using the law of total variation and the appropriate conditional formulae for higher moments.