Stat 536 Homework 7

Due: 10/27/08

1. Suppose you collect the following data from 12 subpopulations.

Subpopulation	A_1A_1	A_1A_2	A_2A_2	$\hat{q}$
1	208	484	311	0.44865
2	202	513	288	0.45713
3	705	305	29	0.82531
4	114	456	473	0.32790
5	772	207	10	0.88524
6	95	446	471	0.31423
7	791	252	28	0.85621
8	520	358	82	0.72813
9	279	489	229	0.52508
10	395	465	165	0.61220
11	686	249	23	0.84603
12	434	459	127	0.65049
$\hat{p} = 0.62232$				

where $\hat{q}$ and $\hat{p}$ are explained below. Use this data to estimate F_{IS}, F_{ST}, and F_{IT} for the population, following these steps.

(a) (10 pts) Assume that q across subpopulations are distributed as independent truncated-normal random variables as discussed in class. Specify the distributions f and g in the following formula for the likelihood of the table data $n = \{n_{ij} : i = 1, 2, \cdots, 12, j = 1, 2, 3\}$.

$$L(n \mid F_{IS}, p, c) = \int_q g(n; F_{IS}, q) f(q; p, c) dq \tag{1}$$

where g is the density of n and involves parameters F_{IS} and q and f is the density of q, involving parameters p and c.

(b) (20 pts) The preceding likelihood derives from the law of total probability, where the integration is over all the possible outcomes of the subpopulation allele frequencies q. In general, it is difficult to compute the integral in eq. (1), however, there are a lot of data in the table, so $q = (q_1, \ldots, q_{12})$ and p may be well-estimated by the sample proportions $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_{12})$ and $\hat{p}$ (shown in table). Under this assumption, the log likelihood becomes approximately

$$\ln L(n \mid F_{IS}, c) \approx \ln g(n; F_{IS}, \hat{q}) + \ln f(\hat{q}; \hat{p}, c) \tag{2}$$

now a function of only two unknown population parameters (F_{IS}, c). Write a function to compute the right-hand side of eq. (2), and use the `R` function `optim` to find the MLEs $\hat{F}_{IS}$ and $\hat{c}$.

(c) (10 pts) Convert your MLEs into estimates of the three F statistics. Recall that F_{ST} is related to the variance of the distribution of q. To obtain the variance of the truncated normal, you may use a numerical approach.