Stat 536 Homework 3

Due: 9/22/08

1. Consider two estimators for within-person gene correlation f: (1) MLE $\hat{f}$ and (2) MOM $\tilde{f}$. In this problem, you study possible differences between these two estimators in the two-allele case.

   (a) Using a rough argument, find a choice of allele frequency p and f for which you would have 80% power to reject $H_0 : f = 0$ given a sample of size $n = 100$. [Hint: Use the results for D_1 and your choice of allele frequencies, then convert from D_1 to f and assume that the power argument transfers to a test of $H_0 : f = 0$.]

   (b) Generate multiple simulated datasets using the p and f identified in part (a). Each dataset is like one random sample from a hypothetical population at HW disequilibrium. For each dataset, estimate $\hat{f}$ and $\tilde{f}$, and use these estimates to answer the question whether one estimator is better than the other. Recall that we prefer unbiased estimators and those with smaller variance.

   [Hint: You can generate simulated data with `R` code

   ```
   n.ij <- rmultinom(n=1, size=100, prob=c(P.AA,P.Aa,P.aa))
   ```

   after computing the population genotype proportions `P.AA, P.Aa`, and `P.aa` according to the selected model.]

   Solution:

   (a) Let's choose $p = 0.2$. Then, the disequilibrium D detectable at 80% power with sample size $n = 100$ is given as the solution to

   $$\nu = \frac{nD^2}{p^2(1-p)^2}$$

   is obtained as

   ```
   > qchisq(0.80, df=1, ncp=qchisq(1-0.05, df=1))
   [1] 7.848899
   ```

   Therefore,

   $$D = p(1-p)\sqrt{\frac{\nu}{n}} = 0.0448$$

   To convert to correlation f that we will use in simulation, we use,

   $$P_{11} = p^2 + D = p^2 + p(1-p)f$$

   leading to

   $$f = \frac{D}{p(1-p)} = \sqrt{\frac{\nu}{n}} = 0.28$$

(b) The MLE estimator of f , $\hat{f}$ can be obtained using Bailey's method. The two allele case of the model has two unknowns, p_1 and f and two bits of information.

$$P_{11} = p_1^2 + p_1(1-p_1)f$$
$$P_{12} = 2p_1(1-p_1)(1-f)$$

We use these to compute Expectations

$$E[n_{11}] = n[\hat{p_1}^2 + \hat{p_1}(1-\hat{p_1})\hat{f}] = n_{11} \quad (1)$$
$$E[n_{12}] = n[2\hat{p_1}(1-\hat{p_1})(1-\hat{f})] = n_{12} \quad (2)$$

Rearranging (1) and (2) we get,

$$\hat{f} = \frac{4nn_{11} - (2n_{11}+n_{12})^2}{(2n_{11}+n_{12})(2n-2n_{11}-n_{12})}$$

The MOM estimator can be computed by defining for the two allele case,

$$F = \sum E((\tilde{p_u^2})) = (\frac{2n_{11}+n_{12}}{n})^2 + (\frac{2n_{22}+n_{12}}{n})^2$$
$$G = \sum E(\tilde{P}_{uu}) = \frac{n_{11}+n_{22}}{n}$$

These can now be used to express $\tilde{f}$ explicitly in terms of the counts.

$$\tilde{f} = \frac{G - F - \frac{1}{2n} - \frac{1}{2n}G}{1 - F - \frac{1}{2n} + \frac{1}{2n}G}$$

See Linked Code.

```
[1] "Average mom: $0.276656696073421"
[1] "Average mle: $0.272077365978802"
[1] "mom bias: -$0.00334330392657889"
[1] "mle bias: -$0.00792263402119842"
[1] "Variance mom: 0.0134058977644156"
[1] "Variance mle: 0.0134794727111743"
```

Furthermore, a two-sample t-test verifies that the dirence in bias is significant:

```
print(t.test(f.mom, f.mle))
Welch Two Sample t-test
data: f.mom and f.mle
t = 2.7928, df = 19997.85, p-value = 0.00523
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
$0.00136543 0.00779323
sample estimates:
mean of x mean of y
$0.2766567 0.2720774
```

An F-test of the variances does not detect a difference:

```
print(var.test(f.mom, f.mle))
F test to compare two variances
data: f.mom and f.mle
F = 0.9945, num df = 9999, denom df = 9999, p-value = 0.7843
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
$0.9563058 1.0343047
sample estimates:
ratio of variances
$0.994541
```

We conclude that the MLE is significantly more biased than the MOM, so the latter should be preferred. Other settings may lead to different conclusions

2. Follow the WinBUGS tutorial (linked separately) to estimate the additive HW disequilibrium parameters D_{uv} for the data below.

Genotype	AA	BB	CC	DD	AB	AC	AD	BC	BD	CD
Count n_{xy}	101	430	329	568	103	214	74	99	65	402

   (a) Report the posterior mean and 95% credible sets for each parameter D_{uv}.

   (b) Retool the above example to estimate f_{uv} and provide posterior means and credible sets for these parameters.

   (c) Compare the full model with distinct f_{uv} to the model with $f_{uv} = f$ for all u, v using the DIC.

   (d) Write a short discussion about this data analysis, answering the following questions.

      i. Is there evidence to reject HWE? What about the hypothesis of homogeneity, i.e. $f_{uv} = f$?

      ii. How would you construct a likelihood ratio test for homogeneity? What is the asymptotic sampling distribution?

      iii. Could you have computed an exact probability for testing HWE? How would you compute it, using what tools? Could exact tests be used to test homogeneity?

      iv. Could you have tested the homogeneity of disequilibrium using the D_{uv} model? If so, write down the constrained model.

Solution:

   (a) See linked WinBUGS code.

node	mean	sd	2.5%	median	97.5%
$D.ab$	0.007831	0.002066	0.00388	0.007822	0.01187
$D.ac$	−0.009001	0.002479	−0.01389	−0.008998	−0.004127
$D.ad$	0.02812	0.002107	0.0241	0.02809	0.03232
$D.bc$	0.04725	0.002658	0.04212	0.04725	0.05245
$D.bd$	0.0693	0.002817	0.06388	0.0693	0.07493
$D.cd$	0.01675	0.003763	0.009313	0.01678	0.02418
$p.a$	0.1248	0.005356	0.1147	0.1247	0.1355
$p.b$	0.2369	0.007985	0.2216	0.2367	0.2526
$p.c$	0.2876	0.007424	0.2734	0.2876	0.3023
$p.d$	0.3508	0.008588	0.3338	0.3509	0.3677

(b) See linked WinBUGS code for model with f_{uv}.

node	mean	sd	2.5%	median	97.5%
$f.aa$	0.2464	0.02734	0.1935	0.2461	0.3001
$f.ab$	0.2628	0.06643	0.1284	0.2644	0.3876
$f.ac$	−0.2513	0.07177	−0.3958	−0.2503	−0.115
$f.ad$	0.6423	0.03893	0.5625	0.6434	0.7146
$f.bb$	0.6878	0.0175	0.6525	0.688	0.7212
$f.bc$	0.693	0.02894	0.6338	0.6935	0.7474
$f.bd$	0.8339	0.0196	0.7936	0.8347	0.8699
$f.cc$	0.2686	0.02106	0.227	0.2685	0.3093
$f.cd$	0.1671	0.03555	0.09628	0.1672	0.236
$f.dd$	0.5013	0.01857	0.4642	0.5014	0.5375
$p.a$	0.1245	0.005364	0.1142	0.1244	0.1353
$p.b$	0.2362	0.008047	0.2206	0.2361	0.2522
$p.c$	0.2878	0.007409	0.2734	0.2878	0.3026
$p.d$	0.3515	0.008433	0.335	0.3515	0.3682

(c) See linked WinBUGS code for reduce model with f. The following estimates are obtained for the model parameters in this reduced model.

node	mean	sd	2.5%	median	97.5%
f	0.445	0.01399	0.4175	0.4451	0.4718
$p.a$	0.1367	0.005715	0.1258	0.1367	0.148
$p.b$	0.2123	0.007001	0.1987	0.2122	0.2261
$p.c$	0.3059	0.007858	0.2906	0.306	0.3214
$p.d$	0.3451	0.008084	0.3294	0.345	0.3611

Comparing DIC values for model from part (b) (Full model) and this model (Reduced model), we conclude that the full model is a better fit since it has the lower DIC value.

Model	DBar	Dhat	pD	DIC
Full	69.666	60.700	8.966	78.631
Reduced	507.221	503.293	3.927	511.148

(d) (i) From the model in in part (a) we see that all the D_{uv} values are non-zero and none of the credible intervals include zero. This gives us evidence to reject HWE. The hypothesis of homogeneity is also rejected since the full model in part (b) does significantly better than the reduced model in part (d).

(ii) The likelihood ratio test for testing homogeneity can be constructed by writing out the ratio of likelihoods for the full and reduced models in parts (b) and (c) respectively.

$$\lambda = \frac{lnL(n_{uv}, p_u, f_{uv})}{lnL(n_{uv}, p_u, f)}$$

Asymptotically,

$$-2ln\lambda \tilde{} \chi^2_5$$

Since the difference in the number of parameters in the two models in 5.

(iii) Yes an exact probability for testing HWE could have been computed by rearranging the genotype frequencies while maintaining allele frequencies. GENEPOP could have been used for this since manual computations would become quickly overwhelming. Additionally, approximate exact tests could have been carried out that use Monte Carlo Simulation Methods to compute exact probabilities for the current dataset.
It is not obvious how to construct an exact test for testing homogeneity. The null hypothesis involves genotype proportions with non-zero parameter f, so the alleles are already not independently distributed between the genotypes. What conditional distribution would apply?

(iv) It is no accident that we did not talk about hypothesis testing of this sort for the additive disequilibrium model. There is no constant D that could be substituted in for all D_{uv}, $u \neq v$. The problem is that the size of additive disequilibrium depends on the allele frequencies, and these vary for different combinations of u and v. So, no, I would not know how to test the homogeneity assumption with the D model.