Stat 536 HW 2 solutions

1. *Consider a single locus with two alleles. Suppose p_f is the frequency of the first allele in females and p_m is the frequency in males. How long does it take to reach HWE and what is the equilibrium allele frequency at HWE?*
   Solution:
   At generation t= 0 the frequencies of the alleles A and a are p_f and $1 - p_f$ respectively in females and p_m and $1 - pm$ respectively in males. In generation t=1, produced by mating of individuals from these two populations, the genotype frequencies are:

Maternal	Paternal	Genotype	Frequency
A	A	AA	$p_f p_m$
A	a	Aa	$p_f(1-p_m)$
a	A	aA	$(1-p_f)p_m$
a	a	aa	$(1-p_f)(1-p_m)$

   Assuming that the locus is autosomal alongwith all the other HWE assumptions, this gives rise to a allele frequencies P_A and P_a for alleles A and a respectively, at generation t= 1 of:

$$P_A = p_f p_m + 0.5(p_f(1-p_m) + (1-p_f)p_m) = 0.5(p_f + p_m)$$
$$P_a = (1-p_f)(1-p_m) + 0.5(p_f(1-p_m) + (1-p_f)p_m) = 1 - 0.5(p_f + p_m) = 1 - P_A$$

   This is the equilibrium frequency. It can be further verified that the frequencies remain constant at generation t= 2 and later.

2. *Now, suppose that the locus in question* 1 *is on the X chromosome. Females have two X chromosomes and males have one. A daughter receives one of her mom's X chromosomes at random and her dad's only X chromosome. A son receives one of his mom's X chromosomes at random and his dad's Y chromosome. Determine whether the difference in allele frequencies decreases after one generation if the initial allele frequencies in the sexes are $p_f \neq p_m$ but HWE assumptions apply thereafter? (Note, p_m means the frequency of the allele on the X chromosomes in males.) Does the difference decrease to zero in one generation as in Question 1? Will the allele and genotype frequencies stay constant if $p_f = p_m$? (Please note that autosomal loci are loci that are not on the X or Y chromosome. Sex-linked loci are on the X chromosome.)*
   Solution:
   We will change notation slightly so that p_{f_0} will now denote the allele fequency of allele A in females in generation 0 and p_{m_0} in males.Assume that sons and daughters are equally likely. Now consider a mating table like that in answer 1.

Maternal	Paternal	Daughter Genotype	Daughter Frequency
A	A	AA	$p_{f_0}p_{m_0}$
A	a	Aa	$p_{f_0}(1-p_{m_0})$
a	A	aA	$(1-p_{f_0})p_{m_0}$
a	a	aa	$(1-p_{f_0})(1-p_{m_0})$
Maternal	Paternal	Son Genotype	Son Frequency
A	A	A	p_{f_0}
A	a	A	p_{f_0}
a	A	a	$(1-p_{f_0})$
a	a	a	$(1-p_{f_0})$

Just like before the allele fequencies in the daughters (i.e females in generation 1) is going to be:

$$p_{f_1} = 0.5(p_{f_0} + p_{m_0})$$

In sons (i.e. generation 1 males) the frequencies are:

$$p_{m_1} = p_{f_0}$$

Similarly it can be shown that at generation 2 the corresponding frequencies are:

$$p_{f_2} = 0.5(p_{f_1} + p_{m_1}) = 0.5(0.5(p_{f_0} + p_{m_0}) + p_{f_0}) = 0.75(p_{f_0}) + 0.25(p_{m_0})$$
$$p_{m_2} = p_{f_1} = 0.5(p_{f_0} + p_{m_0})$$

Notice how the male frequencies "chase" the female frequencies with the allele frequency in males at generation n being equal to that in females in generation *n-1* The difference between male and female allele frequencies reduces exponentially with each generation.

$$D_0 = |p_{f_0} - p_{m_0})|$$
$$D_1 = |p_{f_1} - p_{m_1})| = 0.5(|p_{m_0} - p_{f_0}|)$$
$$D_2 = |p_{f_2} - p_{m_2})| = 0.25(|p_{m_0} - p_{f_0}|)$$
$$D_t = (0.5)^t(|p_{m_0} - p_{f_0}|)$$

Over a large number of generations this difference will approach zero but not in the first generation, as in the autosomal locus of question 1. When $p_{m_0} = p_{f_0}$ this difference will stay constant, as will the allele frequencies.

3. *When the allele frequencies at an autosomal locus differ by δ in the sexes, it can be shown that the genotype frequencies in the next generation diverge from HWE according to the fo llowing equations.*

$$\begin{aligned} P_{11} &= p_1^2 - \frac{\delta^2}{4} \\ P_{12} &= 2p_1p_2 + \frac{\delta^2}{2} \\ P_{22} &= p_2^2 - \frac{\delta^2}{4} \end{aligned} \tag{1}$$

(a) *Derive formulae for the maximum likelihood estimates for $p_1; p_2, and\delta$ and compute $\hat{p_1}; \hat{p_2}; and\hat{\delta}$ for the data in the table below.*

A_1A_1	A_1A_2	A_2A_2
1	*17*	*82*

Solution:

Use Bailey's method and apply genotype frequencies from eqn (1)

$$\begin{aligned} &E(n_{11}) = n(P_{11}) = n_{11} \\ &=> \hat{p_1^2} - \frac{\hat{\delta^2}}{4} = \frac{n_{11}}{n} \end{aligned} \tag{2}$$

$$\begin{aligned} &E(n_{12}) = n(P_{12}) = n_{12} \\ &=> 2\hat{p_1}(1-\hat{p_1}) + \frac{\hat{\delta^2}}{2} = \frac{n_{12}}{n} \end{aligned} \tag{3}$$

Adding 2 times (2) to (3)

$$\hat{p_1} = \frac{(n_{12} + 2n_{11})}{2n}$$

Similarly,

$$\hat{p_2} = \frac{(n_{12} + 2n_{22})}{2n}$$

Substituting value of $\hat{p_2}$ in equation for P_{22} and solving for $\hat{\delta}$ we get ,

$$\hat{\delta} = 2\sqrt{\frac{(n_{12} + 2n_{22})^2}{4n^2} - \frac{n_{22}}{n}}$$

We now substitute values for n_{11}, n_{12} and n_{22} from the table

$$\hat{p_1} = 0.095$$
$$\hat{p_2} = 0.905$$
$$\hat{\delta} = 2\sqrt{-0.000975}$$

Notice that the value for $\hat{\delta}$ is not possible. This indicates that the model does not fit the data well. In fact the legal value of $\hat{\delta}$ that maximizes the Likelihood is 0.

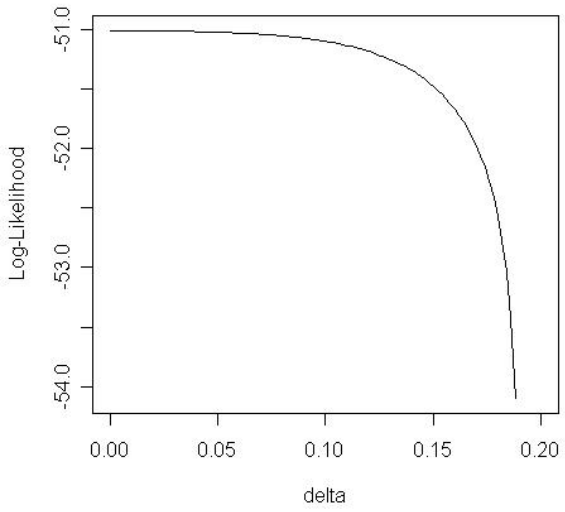

Figure 1: Profile Log Likelihood of delta

(b) *Estimate the variance of these estimates using the information matrix.*
Solution:
With $\hat{\delta} = 0$ the above equations reduce to a standard HWE set.Refer notes for variance computations. Because there was no evidence supporting the model with δ, we will assume HWE model and use the information "matrix" to obtain the variance of $\hat{p}$. Let C be the multinomial constant and $C' = C + n_{12} \ln 2$, then

$$\begin{aligned}
\ln L &= \ln C + n_{11} \ln p_1^2 + n_{12} \ln 2p_1(1-p_1) + n_{22} \ln(1-p_1)^2 \\
&= \ln C' + 2n_{11} \ln p_1 + n_{12} \ln p_1 + n_{12} \ln(1-p_1) + 2n_{22} \ln(1-p_1) \\
\frac{\partial \ln L}{\partial p_1} &= \frac{2n_{11}}{p_1} + \frac{n_{12}}{p_1} - \frac{n_{12}}{1-p_1} - \frac{2n_{22}}{1-p_1} \\
\frac{\partial^2 \ln L}{\partial p_1^2} &= -\frac{2n_{11}}{p_1^2} - \frac{n_{12}}{p_1^2} - \frac{n_{12}}{(1-p_1)^2} - \frac{2n_{22}}{(1-p_1)^2}
\end{aligned}$$

$$
\begin{aligned}
E\left[-\frac{\partial^2 \ln L}{\partial p_1^2}\right] &= \frac{2np_1^2}{p_1^2} + \frac{2np_1(1-p_1)}{p_1^2} + \frac{2np_1(1-p_1)}{(1-p_1)^2} + \frac{2n(1-p_1)^2}{(1-p_1)^2} \\
&= \frac{2np_1}{p_1^2} + \frac{2n(1-p_1)}{(1-p_1)^2} \\
&= \frac{2n}{p_1(1-p_1)} \\
\mathrm{Var}(\hat{p}_1) &= \frac{p_1(1-p_1)}{2n} \approx 0.000429875
\end{aligned}
$$

which, lo and behold, is the formula for variance we derived via other means in class.

Computing analytic variance for the full model is substantially more complicated. Let's write p for p_1 and note $(p_1^2 - \delta^2/4) = (p_1 - \delta/2)(p_1 + \delta/2)$ and $[(1-p_1)^2 - \delta^2/4] = (1 - p_1 - \delta/2)(1 - p_1 + \delta/2)$, Then the negative second derivatives are

$$
\begin{aligned}
-\frac{\partial^2 \ln L}{\partial p^2} &= \frac{n_{11}}{(p-\delta/2)^2} + \frac{n_{11}}{(p+\delta/2)^2} + \frac{4n_{12}}{2p(1-p)+\delta^2/2} + \frac{n_{12}[2(1-2p)]^2}{[2p(1-p)+\delta^2/2]^2} \\
&\quad + \frac{n_{22}}{(1-p-\delta/2)^2} + \frac{n_{22}}{(1-p+\delta/2)^2} \\
-\frac{\partial^2 \ln L}{\partial p \partial \delta} &= -\frac{n_{11}/2}{(p-\delta/2)^2} + \frac{n_{11}/2}{(p+\delta/2)^2} + \frac{2(1-2p)\delta n_{12}}{(2p(1-p)+\delta^2/2)^2} \\
&\quad + \frac{n_{22}/2}{(1-p-\delta/2)^2} - \frac{n_{22}/2}{(1-p+\delta/2)^2} \\
-\frac{\partial^2 \ln L}{\partial^2 \delta} &= \frac{n_{11}/4}{(p-\delta/2)^2} + \frac{n_{11}/4}{(p+\delta/2)^2} - \frac{n_{12}}{2p(1-p)+\delta^2/2} + \frac{\delta^2 n_{12}}{[2p(1-p)+\delta^2/2]^2} \\
&\quad + \frac{n_{22}/4}{(1-p-\delta/2)^2} + \frac{n_{22}/4}{(1-p+\delta/2)^2}
\end{aligned}
$$

Taking the expectation amounts to substituting $nP_{11}, nP_{12}, nP_{22}$, where the population proportions are given by eq. (2), for n_{11}, n_{12}, n_{22} in the above equations.

$$
\begin{aligned}
E\left[-\frac{\partial^2 \ln L}{\partial p^2}\right] &= \frac{n(p+\delta/2)}{p-\delta/2} + \frac{n(p-\delta/2)}{p+\delta/2} + 4n + \frac{n[2(1-2p)]^2}{2p(1-p)+\delta^2/2} \\
&\quad + \frac{n(1-p+\delta/2)}{1-p-\delta/2} + \frac{n(1-p-\delta/2)}{1-p+\delta/2} \\
E\left[-\frac{\partial^2 \ln L}{\partial p \partial \delta}\right] &= -\frac{n(p+\delta/2)/2}{(p-\delta/2)} + \frac{n(p-\delta/2)/2}{(p+\delta/2)} + \frac{2(1-2p)\delta n}{2p(1-p)+\delta^2/2} \\
&\quad + \frac{n(1-p+\delta/2)/2}{1-p-\delta/2} - \frac{n(1-p-\delta/2)/2}{1-p+\delta/2} \\
E\left[-\frac{\partial^2 \ln L}{\partial^2 \delta}\right] &= \frac{n(p+\delta/2)/4}{p-\delta/2} + \frac{n(p-\delta/2)/4}{p+\delta/2} - n + \frac{\delta^2 n}{2p(1-p)+\delta^2/2} \\
&\quad + \frac{n(1-p+\delta/2)/4}{1-p-\delta/2} + \frac{n(1-p-\delta/2)/4}{1-p+\delta/2}
\end{aligned}
$$

At this point I give up trying to do more analytically and turn to R to be my calculator. First, I substitute $\hat{p} = 0.095, \hat{\delta} = 0$ for p, δ and invert a 2×2 matrix. Recall

$$
\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}
$$

If I store the $E\left[-\frac{\partial^2}{\partial(p,\delta)^2}\right]$ matrix in `info`, then `R` tells us the variance matrix is

```
var <- matrix(nrow=2,ncol=2)
det <- info[1,1]*info[2,2] - info[1,2]^2
var[1,1] <- info[2,2]/det
var[2,2] <- info[1,1]/det
var[1,2] <- var[2,1] <- info[1,2]/det
print(var)
```

```
             [,1]         [,2]
[1,] 0.000429875  0.000000e+00
[2,] 0.0000000000 7.036874e+13
```

It should be no surprise that there is no information about δ in this data, especially since we were already forced to set $\hat{\delta} = 0$ based on common sense.

(c) *Find 95% confidence intervals for p_1.*
Solution:
Using the variance computed using HWE, we have

```
p.hat.var <- p.hat*(1-p.hat)/2/sum(n)
print(paste("Variance under HWE:", p.hat.var))
print(paste("95% CI:", p.hat + qnorm(0.025)*sqrt(p.hat.var), ",", p.hat + qnorm(0.975)*sqrt(p.hat.var)))
"95% CI: 0.0543632295753478 , 0.135636770424652"
```

Or using the variance computed from the full model, we have

```
print(paste("95% CI:", p.hat + qnorm(0.025)*sqrt(var[1,1]), ",", p.hat + qnorm(0.975)*sqrt(var[1,1])))
[1] "95% CI: 0.0448311476238862 , 0.145168852376114"
```

[The full code for solving this problem is linked separately in file hw2p3.R]

4. *Consider this blood type data taken from a sample of Kuwaiti individuals.*

A	B	AB	O
29	*23*	*14*	*35*

   (a) Write an EM algorithm to estimate the allele frequencies $\hat{p}_A, \hat{p}_B$, and $\hat{p}_O$.

   (b) Use bootstrap to estimate the variance of these estimates.

   (c) Do the allele frequencies in this population differ from those in Korea, where $\hat{p}_A = 0.231, \hat{p}_B = 0.209$, and $\hat{p}_O = 0.560$ were estimated based on a sample of size $n = 253$?

Solution:
See linked code for all solutions.File hw2p4.R.