Solution of assignment 11, ST2304

Problem 1

1. Simulating 1000 realisations of Poisson distributed variables with d equal to 0.1, 0.5 and 5 we find that this gives recombination probabilities of 0.093, 0.331 and 0.5, respectively.

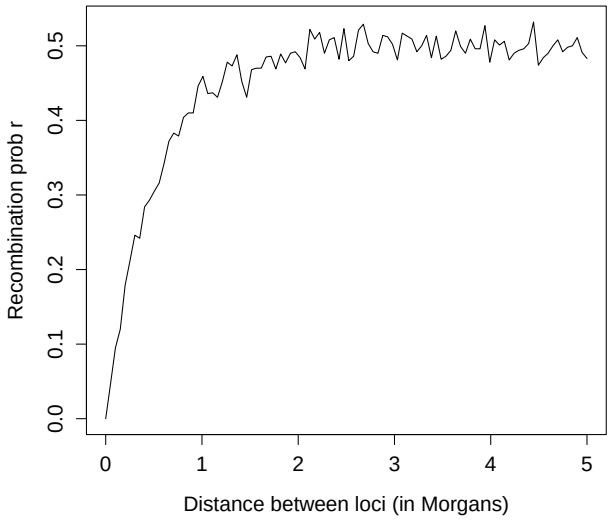
```
recombprob <- function(d,n=1000) {
  n.odd <- 0
  for(i in 1:n) {
    X<-rpois(1,r)
    if (X%%2==1) { # if the
      n.odd <- n.odd + 1
    }
  }
  n.odd/n
}
> recombprob(0.1)
[1] 0.093
> recombprob(0.5)
[1] 0.331
> recombprob(5)
[1] 0.495
```

2. Computing the recombination probability for a very large value of d (the expected number of crossing over events) suggest that the recombination probability goes towards 0.5.

```
> recombprob(50,n=100000)
[1] 0.49909
```

A graph showing the relationship can be made as follows:

```
dd <- seq(0,5,len=100)
rr <- NULL
for (i in 1:length(dd))
   rr[i] <- recombprob(dd[i])
plot(dd,rr,type="l",xlab="Distance between loci (in Morgans)",ylab="Recombination p
```



This is know as Haldane's mapping function $r = \frac{1}{2}(1 - e^{-2d})$.

Problem 2

1. This mean that the upper and lower boundary A and B of the interval are stochastic variables which includes the unknown parameter with probability $P(A < \sigma^2 < B) = 0.95$.

2. We can verify this for different values of σ^2, μ and n as follows.

```
coverage <- function(mu,sigma2,n,alpha=.05,m=1000) {
  n.hits <- 0
  q.upper <- qchisq(1-alpha/2,df=n-1)
  q.lower <- qchisq(alpha/2,df=n-1)
  for (i in 1:m) {
     x <- rnorm(n,mean=mu,sd=sqrt(sigma2))
     s <- var(x)
     ci <- c(s*(n-1)/q.upper, s*(n-1)/q.lower)
     if (ci[1]<sigma2 & ci[2]>sigma2)
        n.hits <- n.hits + 1
  }
  n.hits/m
}
> coverage(0,1,10)
[1] 0.953
> coverage(0,1,10)
[1] 0.956
> coverage(10,100,10)
[1] 0.962
> coverage(-10,2,10)
[1] 0.957
> coverage(-10,2,1000)
[1] 0.952
```

Problem 3

1. Under the full model, all n p_i's are free parameters (no relationship $p_i = q\phi(\beta_0 + \beta_1 \text{time}_i)$ is imposed) and the MLEs are $\hat{p}_i = x_i/n$ which can be computed as follows in R.

```
> phat <- x/n
> phat
 [1] 0.0000000 0.0000000 0.0000000 0.0000000 0.1875000 0.1190476 0.2000000
 [8] 0.1851852 0.4000000 0.3181818 0.2857143 0.4615385 0.0000000 0.5000000
[15] 0.6250000 0.8055556 0.7272727 0.6666667 0.6551724 0.6969697 0.8214286
[22] 0.8571429 0.9333333 0.8000000 0.9166667 0.7826087 0.7857143 0.7826087
[29] 0.8461538 1.0000000 0.8000000 0.9285714 0.6666667 1.0000000 0.7500000
[36] 0.9000000 0.9000000 0.7777778 0.7500000 1.0000000 0.8571429 1.0000000
[43] 1.0000000 1.0000000 0.5000000 1.0000000 0.0000000 1.0000000 1.0000000
```

2. The maximum log likelihood under the full model is the log likelihood at the point $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$ in the parameter space. At this point the log likelihood $\ln L(p_1, p_2, \ldots, p_n) = \sum \ln f(x_i)$ is

```
> sum(dbinom(x,size=n,prob=phat,log=T))
[1] -47.56002
```

3. From the solution to assignment 10, the maximum log likelihood of the model $p_i = q\phi(\beta_0 + \beta_1 \text{time}_i)$ is -68.21 (the maximum negative log likelihood is in the `$value` component of the list returned by `optim`).

4. The observed deviance is two times the difference between the maximum log likelihoods, that is,

```
> 2*((-47.56)-(-68.21))
[1] 41.3
```

5. Under the null hypothesis that the fitted model is correct the deviance D is chi-square distributed with $n - p = 49 - 3 = 46$ degrees of freedom. We reject this null hypothesis if D is larger than the upper 0.05-quantile of the chi-square distribution,

```
> qchisq(.05,df=46,lower=F)
[1] 62.82962
```

that is, $\chi^2_{46} = 62.83$ so we can not reject the hypothesis that the model is correct. The P-value becomes

```
> pchisq(41.3,df=46,lower=F)
[1] 0.6691562
```

6. The expected value of a chi-square distributed variable is equal to it's degrees of freedom, that is, in our case 46. The fact that the observed value of D is slightly smaller than this indicates that the is some (statistically non-significant) under-dispersion in the data.