Part B

Quantitative Genetics

1 Genetic constitution of a population

Exercise 1.
1. Load 1000Genomes.RData
2. Calculate allele frequencies for a subset of SNPs in various populations
3. Visualize those alleles frequencies
4. Test for Hardy-Weinberg Equilibrium within populations and across populations

```
> load("Data/PartB/1000Genomes-MHC.RData")
> dim(genotypes)
```

```
[1] 26827  2509
```

```
> dim(pops)
```

```
[1] 2504    6
```

```
> #look at the data
> #genotypes[1:5,1:10]
>
> geno = genotypes[,-c(1:5)]
> rownames(geno) = genotypes$SNP
> geno=t(geno) # to put the individuals in rows
> #geno[1:5,1:5]
>
> n = ncol(geno)
> p = nrow(geno)
```

Question. Calculate allele frequencies for a subset of SNPs in various populations

Choice of a population

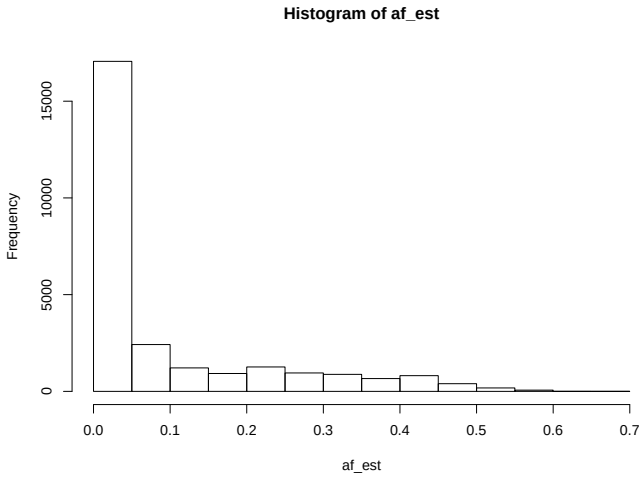
```
> my_pop = "CLM"
> my_SNP = geno[which(pops$Population.code == my_pop),]
> my_p = ncol(my_SNP) #number of SNPs
> my_n = nrow(my_SNP) #number of samples
```

For a haploid genome coded in 0 and 1, this estimate would just be the mean across individuals, but since we have a diploid genome, it is the mean divided by two.

```
> af_est = colMeans(my_SNP)/2
```

Question. Visualize those alleles frequencies

```
> hist(af_est)
```



Are these estimates of the minor allele frequency?

Question. Testing for Hardy-Weinberg Equilibrium within populations and across populations

There is a known relationship between genotypes and allele frequencies (as was seen in the Lecture 'Genetic constitution of a population'). If we have p the frequency of allele A, q the allele frequency of allele a, then under Hardy-Weinberg equilibrium we have several equalities:

$$p+q=1 \qquad (p+q)^2=1 \qquad p^2+2pq+q^2=1$$
$$p_{AA}=p^2 \qquad p_{Aa}=2pq=2p(1-p) \qquad p_{aa}=q^2$$

where p_{AA} is the frequency of the genotype AA.

In this step, we are testing for a deviation from HWE on a SNP per SNP basis, based on a χ^2 with one degree of freedom. A p-value is obtained for each SNP and a threshold of 10^{-6} is generally used (each SNP with a p-value lower than the threshold is removed).

Let us focus on a single SNP for which we have the observed counts of each genotype; let's call the alleles 'A' and 'a' (the real information is in 'genotypes')

```
> SNP_j = my_SNP[,162]
> table(SNP_j) #observed counts
```

```
SNP_j
 0  1  2
43 34 17
```

We have already calculated the frequency of each allele with 'af_est'.

```
> pa=af_est[162]
> pa
```

```
rs9468334
0.3617021
```

From these observed counts, we calculate the expected counts that we should observe if the SNP was under HWE.

$$E(n_{AA}) = p_{AA} \times N_{total} = p_A^2 \times N_{total} = (1 - p_a)^2 \times N_{total} = 38.30$$
$$E(n_{aa}) = p_{aa} \times N_{total} = p_a^2 \times N_{total} = 12.30$$
$$E(n_{Aa}) = p_{Aa} \times N_{total} = 2p_A p_a \times N_{total} = 43.40$$

We can create a table summarising the observed and expected counts (Table 1).

	0	1	2
Observed	43.000	34.000	17.000
Expected	38.300	43.400	12.300

Table 1: *Observed and expected counts*

We then calculate the chi-square statistics

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

where O_i are the observed values and E_i are the expected values. We finally compare the χ^2 to a chi-square distribution with one degree of freedom, and decide whether the SNP is deviating from HWE.

```
> chi2 = (43-38.30)^2/38.30+(34-43.40)^2/43.40+(17-12.300)^2/12.300
> chi2
```

```
[1] 4.408642
```

We can obtain a p-value for the test, using the *pchisq* function as follows

```
> 1-pchisq(chi2, df =  1)
```

```
[1] 0.03575729
```

If we are looking at only a single SNP (and thus performing a single test), we would compare the p-value to an arbitrary threshold of 0.01 and we would conclude that we cannot reject the null hypothesis and that this SNP is under HWE. However, if we are performing thousands of independent tests, we need to account for multiple testing. Usually, the p-value is not compared to the original threshold of 0.01 but to 10^{-6}.

We now perform the same test over all the SNPs.

```
> # expected counts if at the Hardy-Weimber equilibrium
> E_n_Aa = 2 * af_est * (1-af_est) *my_n # 2p(1-p) *N= 2pq *N
> E_n_AA = (1-af_est)^2 *my_n             # p^2 *N
> E_n_aa = af_est^2 *my_n                 # q^2 *N
```

```
> # observed counts
> O_n_Aa = apply(my_SNP,2,function(x){sum(x == 1)}) # 2pq *N
> O_n_AA = apply(my_SNP,2,function(x){sum(x == 0)}) # p^2 *N
> O_n_aa = apply(my_SNP,2,function(x){sum(x == 2)}) # q^2 *N
```

We compare the number of events between expected and observed, this follows a Chi square with one degree of freedom:

```
> chi2 =
+   (O_n_AA - E_n_AA)^2/(E_n_AA) +
+   (O_n_Aa - E_n_Aa)^2/(E_n_Aa) +
+   (O_n_aa- E_n_aa)^2/(E_n_aa)
> hist(chi2)
```

Number of SNPs that deviate from the Hardy-Weimber equilibrium, without correcting for multiple testing

```
> sum(chi2 > qchisq(0.99,1), na.rm = T)
```

```
[1] 1550
```

Number of SNPs that deviate from the Hardy-Weimber equilibrium, before and after correcting for multiple testing

```
> pvalues=1-pchisq(chi2,1)
> sum(pvalues<0.01,na.rm=TRUE)
```

```
[1] 1550
```

```
> sum(pvalues<1e-6,na.rm=TRUE)
```

```
[1] 173
```

Question. Try with different populations and across populations

2 Estimation of Genetic Variance

Exercise 2.

1. Simulate $m = 500$ SNPs for $N = 400$ individuals. Choose the allele frequencies as coming from a uniform distribution between 0 and 0.5, with a seed of 6155 (for reproducibility).
2. Random mate to generate offspring
3. Simulate effects for each SNP given a certain heritability h^2 and breeding values (A) and environmental values (E)
4. Simulate a phenotypes based on SNP data: $P = A + E$
5. Use parent-offspring regression to estimate heritability h^2

Question. Simulate $m = 500$ SNPs for $N = 400$ individuals. Choose the allele frequencies as coming from a uniform distribution between 0 and 0.5, with a seed of 6155 (for reproducibility).

Let's assume that the minor allele frequency of our m SNPs come from a uniform distribution between 00 and 0.50.

```
> set.seed(6155) # for reproducibility
> m = 500                                  # number of SNPs
> maf = runif(m, 0, .5)                     # random MAF for each SNP
```

We can then draw genotypes for one person for each SNP.

```
> x012 = rbinom(m, 2, maf)
```

We want genotypes for N individuals, so let's replicate this N times.

```
> N = 400                                     # number of individuals
> x012 = t(replicate(N, rbinom(m, 2, maf)))  # n x m genotype matrix
```

Check that all the SNPs are polymorphic, otherwise we need to discard them.

```
> polymorphic = apply(x012, 2, var) == 0
> sum(polymorphic)
```

```
[1] 0
```

```
> x012[1:5, 1:10]
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    0    0    0    2    0    0    1    0    1     0
[2,]    0    0    0    1    0    0    1    0    0     1
[3,]    0    0    0    0    0    0    1    1    0     2
[4,]    0    0    0    0    0    0    1    0    1     0
[5,]    0    1    1    1    0    1    2    0    2     1
```

Question. Random mate to generate offspring

```
> # if dad is AA and mom is AA => dad gives A, mom gives A =>kid is AA
> # if dad is aa and mom is aa => dad gives a, mom gives a =>kid is aa
> # if dad (or mom) is Aa, then probability 0.5 of giving a
>
>
> # Given a parent individual, get one of their Alleles in the gamete
> getGamete <- function(indiv) {
+   indiv_gives = indiv/2 # creates a vector of 0, or 1 . 0:a  and 1:A
+
+   ind_1 = which(indiv_gives == 0.5) #we isolate the 1s (Aa) in the parent
+   #if Parent is Aa, the gamete is one binomial trial with probability 0.5
+   if (length(ind_1) > 0)
+     indiv_gives[ind_1] = rbinom(length(ind_1), size=1, 0.5)
+     # gives 0 for a or 1 for A
+
+   return(indiv_gives) # vector of 0s and 1s
+ }
> # Two parental Gametes combine to form a zygote
```

```
> combineGametes <- function(momGam, dadGam) {
+   kid = (momGam + dadGam)
+   # 0(a) + 0(a) = 0 (aa)
+   # 0(a) + 1(A) = 1 (aA)
+   # 1(A) + 1(A) = 2 (AA)
+
+   kid
+ }
> # Given two individuals, get an offspring for next generation
> getOffspring <- function(mom, dad) {
+   momGam <- getGamete(mom)
+   dadGam <- getGamete(dad)
+   return(combineGametes(momGam, dadGam))
+ }
```

Now we can simulate a offspring based on two parents in $x012$.

```
> kid012 = getOffspring(x012[1, ], x012[2, ])
```

If we want to create $N = 400$ offsprings

```
> kid012 = NULL # where we will store the offsprings
> pedigree = matrix(0, nrow = N, ncol = N) # we want to record who are the parents
> for(i in 1:N){
+   # randomly pick two parents
+   parents = sample (1:N,2)
+   # generate offspring from these two parents
+   kid = getOffspring(x012[parents[1], ], x012[parents[2], ])
+
+   #
+   pedigree[i, parents] = 1
+
+   # store the new offspring with the other ones
+   kid012 = rbind(kid012, kid)
+ }
```

```
> kid012[1:5, 1:10]
```

```
    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
kid    0    0    0    1    0    0    0    0    2     0
kid    0    0    1    0    0    0    0    0    0     1
kid    0    0    1    2    1    1    0    0    1     0
kid    0    0    1    1    0    0    0    0    0     0
kid    0    0    0    1    0    0    0    0    0     1
```

```
> # who are the parents of offspring 241?
> which(pedigree[241,] == 1)
```

```
[1] 167 340
```

```
> # check that each offspring has only 2 parents
> sum(apply(pedigree,1,sum)==2) # should be N
```

```
[1] 400
```

Question. Simulate effects for each SNP given a certain heritability h^2 and breeding values (A) and environmental values (E). Then, simulate a phenotype based on SNP data: $P = A + E$

Here we will assume that all SNPs have an effect and that their effect size follows a normal distribution.

We first scale the data

```
> scale_x012 = scale(x012)
> scale_x012[which(is.na(scale_x012))] = 0
> scale_kid012 = scale(kid012)
> scale_kid012[which(is.na(scale_kid012))] = 0
```

```
> h2 = 0.7
> beta = rnorm(m, 0, sqrt(h2/m))
> g_parent = scale_x012 %*% beta
> e_parent = rnorm(N, 0, sqrt(1-h2))
> # phenotype parent
> y_parent = g_parent + e_parent
> var(y_parent) # should be around 1
```

```
          [,1]
[1,] 1.042881
```

Phenotype for kids

```
> g_kid = scale_kid012 %*% beta
> e_kid = rnorm(N, 0, sqrt(1-h2))
> # phenotype for kid
> y_kid = g_kid + e_kid
> var(y_kid) # should be around 1
```

```
         [,1]
[1,] 1.02187
```

Question. Use parent-offspring regression to estimate heritability h^2

Find the midparent value for each kid

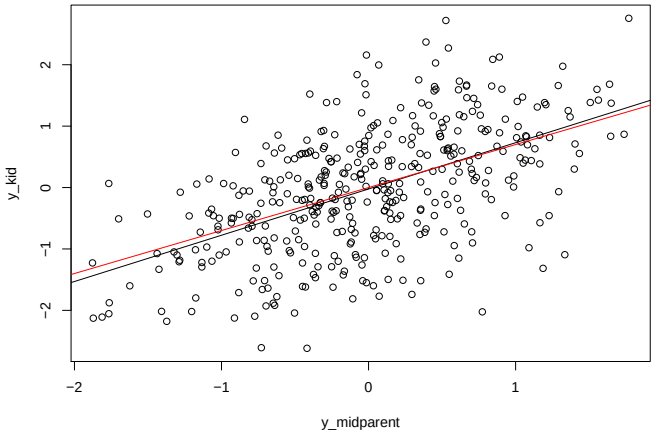
```
> y_midparent = vector(length=N)
> for (i in 1:N){
+   parents = which(pedigree[i,] == 1)
+   y_midparent[i] = mean(y_parent[parents])
+ }
>
```

```
> plot(y_midparent, y_kid)
> reg = lm(y_kid~y_midparent)
> reg
```

```
Call:
lm(formula = y_kid ~ y_midparent)

Coefficients:
(Intercept)  y_midparent
   -0.02336      0.75342
```

```
> abline(reg)
> abline(0,h2, col="red")
```



Exercise 3. Using real twin data, answer the following questions.

1. What is the phenotypic correlation for both height and BMI for the dizygotic (DZ) and monozygotic (MZ) twins?
2. Is the phenotypic correlation for height and BMI the same for both males and females?
3. What is the estimate of the heritability for height and BMI given we have estimated these correlations?
4. After adjusting for sex do we get the same estimate for the heritability for height and BMI?
5. Is there evidence for common environment effects?

The classic twin study begins from assessing the variance of a phenotype in a population, and attempts to estimate how much of this is due to: genetic effects (heritability); shared environment - events that happen to both twins, affecting them in the same way; unshared, or unique, environment - events that occur to one twin but not the other, or events that affect either twin in a different way. These three components can be modelled as a composition of additive genetics (A), common environment or family effect (C), and unique environment (E), or the ACE model. It is also possible to examine non-additive genetics effects (for example, dominance and the ADE model). The ACE model indicates what proportion of variance in a trait is heritable, versus the proportion due to shared environment or un-shared environment.

We have

$$y_{ij} = \mu + c_i + a_{ij} + e_{ij}$$

where y_{ij} is the phenotypic measurement for the jth individual in the ith family. We have that

$$\begin{aligned} \text{Var}(\boldsymbol{y}) &= \sigma_c^2 + \sigma_a^2 + \sigma_e^2, \text{ and} \\ \text{Cov}(y_{ij}, y_{ik}) &= \sigma_c^2 + \pi_{a(jk)}\sigma_a^2, \end{aligned}$$

where $\pi_{a(jk)}$ is the expected (in this case) additive coefficient of relationship between individuals j and k in the same family. Therefore, for monozygotic and dizygotic twins

$$\begin{aligned} \text{Cov}_{MZ}(y_{ij}, y_{ik}) &= \sigma_c^2 + \sigma_a^2, \\ \text{Cov}_{DZ}(y_{ij}, y_{ik}) &= \sigma_c^2 + \frac{\sigma_a^2}{2}, \end{aligned}$$

which implies

$$\begin{aligned} \text{Cov}_{MZ}(y_{ij}, y_{ik}) - \text{Cov}_{DZ}(y_{ij}, y_{ik}) &= \sigma_c^2 + \sigma_a^2 - \sigma_c^2 - \frac{\sigma_a^2}{2} \\ \widehat{\sigma}_a^2 &= 2[\text{Cov}_{MZ}(y_{ij}, y_{ik}) - \text{Cov}_{DZ}(y_{ij}, y_{ik})] \end{aligned}$$

and the contribution from C can be reconstructed as $\widehat{\sigma}_c^2 = \text{Cov}_{MZ}(y_{ij}, y_{ik}) - \widehat{\sigma}_a^2$.

(See Falconer's formula)

Question. What is the phenotypic correlation for both height and BMI for the dizygotic (DZ) and monozygotic (MZ) twins?

We first read the data

```
> twin.data <- read.table("Data/PartB/twin_height_bmi.txt", header = TRUE)
> dim(twin.data)
```

```
[1] 5432    7
```

```
> head(twin.data)
```

```
   dob  ht_t1   bmi_t1  ht_t2   bmi_t2 sex twin
1 1902 165.10 25.86451 173.00 24.05693   1   MZ
2 1903 175.26 22.74054 143.00 33.74248   1   MZ
3 1910 164.00 28.62879 165.00 27.54821   1   MZ
4 1906 167.64 21.46733 167.00 18.28678   1   MZ
5 1911 168.00 23.03005 167.64 19.94806   1   MZ
6 1912 172.72 25.15385 166.00 28.30600   1   MZ
```

We isolate the MZ and DZ

```
> table(twin.data$twin)
```

```
  DZ   MZ
2224 3208
```

```
> MZ = which(twin.data$twin == "MZ")
> DZ = which(twin.data$twin == "DZ")
```

We then calculate the phenotypic correlation for height

```
> cor(twin.data$ht_t1[MZ],twin.data$ht_t2[MZ])
```

```
[1] 0.9186526
```

```
> cor(twin.data$ht_t1[DZ],twin.data$ht_t2[DZ])
```

```
[1] 0.7507025
```

and for BMI

```
> cor(twin.data$bmi_t1[MZ],twin.data$bmi_t2[MZ])
```

```
[1] 0.6986938
```

```
> cor(twin.data$bmi_t1[DZ],twin.data$bmi_t2[DZ])
```

```
[1] 0.351685
```

Height

Question. Is the phenotypic correlation for height the same for both males and females?

```
> cor(twin.data$ht_t1[MZ][twin.data$sex[MZ] == 1],twin.data$ht_t2[MZ][twin.data$sex[MZ] == 1])
```

```
[1] 0.8465659
```

```
> cor(twin.data$ht_t1[MZ][twin.data$sex[MZ] == 2],twin.data$ht_t2[MZ][twin.data$sex[MZ] == 2])
```

```
[1] 0.8352737
```

```
> cor(twin.data$ht_t1[DZ][twin.data$sex[DZ] == 1],twin.data$ht_t2[DZ][twin.data$sex[DZ] == 1])
```

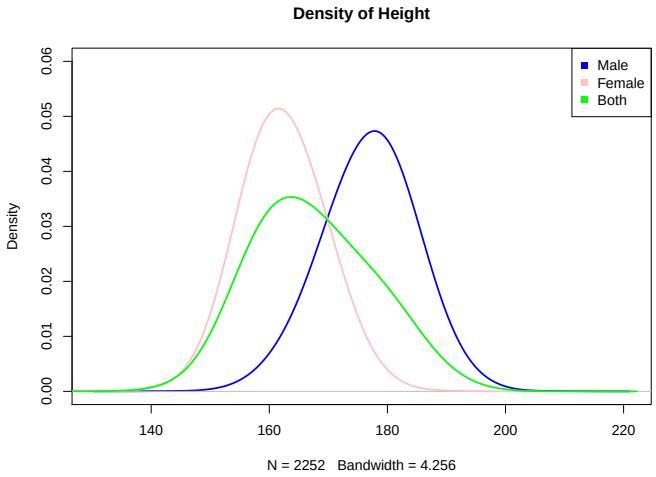
```
[1] 0.462717
```

```
> cor(twin.data$ht_t1[DZ][twin.data$sex[DZ] == 2],twin.data$ht_t2[DZ][twin.data$sex[DZ] == 2])
```

```
[1] 0.4844409
```

We plot the density of 'height' for monozygote males, females and for the combined data.

```
> plot(density(c(twin.data$ht_t1[MZ][twin.data$sex[MZ] == 1],twin.data$ht_t2[MZ][twin.data$sex[MZ] == 1]),
+       adjust = 3), col = "Blue", lwd = 2, ylim = c(0, 0.06), main = "Density of Height")
> lines(density(c(twin.data$ht_t1[MZ][twin.data$sex[MZ] == 2],twin.data$ht_t2[MZ][twin.data$sex[MZ] == 2]),
+       adjust = 3), col = "Pink", lwd = 2)
> lines(density(c(twin.data$ht_t1[MZ], twin.data$ht_t2[MZ]),
+       adjust = 3), col = "Green", lwd = 2)
> legend("topright",legend = c("Male", "Female", "Both"), col = c("Blue", "Pink", "Green"), pch=15)
```



Question. What is the estimate of the heritability for height given we have estimated these correlations?

```
> 2*(cor(twin.data$ht_t1[MZ],twin.data$ht_t2[MZ]) - cor(twin.data$ht_t1[DZ],twin.data$ht_t2[DZ]))
```

```
[1] 0.3359002
```

Question. After adjusting for sex do we get the same estimate for the heritability for height and BMI?

```
> reg.MZ_ht = lm(twin.data$ht_t2[MZ] ~ twin.data$sex[MZ] + twin.data$ht_t1[MZ])
> reg.DZ_ht = lm(twin.data$ht_t2[DZ] ~ twin.data$sex[DZ] + twin.data$ht_t1[DZ])
> A_ht = 2*(coef(reg.MZ_ht)[3] - coef(reg.DZ_ht)[3])
> A_ht
```

```
twin.data$ht_t1[MZ]
          0.7215121
```

Question. Is there evidence for common environment effects?

We calculate $\widehat{\sigma}_c^2$ and $\widehat{\sigma}_e^2$.

```
> C_ht = coef(reg.MZ_ht)[3] - A_ht
> E_ht = 1 - (A_ht + C_ht)
```

BMI

Question. Is the phenotypic correlation for BMI the same for both males and females?

```
> cor(twin.data$bmi_t1[MZ][twin.data$sex[MZ] == 1],twin.data$bmi_t2[MZ][twin.data$sex[MZ] == 1])
> cor(twin.data$bmi_t1[MZ][twin.data$sex[MZ] == 2],twin.data$bmi_t2[MZ][twin.data$sex[MZ] == 2])
> cor(twin.data$bmi_t1[DZ][twin.data$sex[DZ] == 1],twin.data$bmi_t2[DZ][twin.data$sex[DZ] == 1])
> cor(twin.data$bmi_t1[DZ][twin.data$sex[DZ] == 2],twin.data$bmi_t2[DZ][twin.data$sex[DZ] == 2])
```

Question. What is the estimate of the heritability for BMI given we have estimated these correlations?

```
> 2*(cor(twin.data$bmi_t1[MZ],twin.data$bmi_t2[MZ]) - cor(twin.data$bmi_t1[DZ],twin.data$bmi_t2[DZ]))
```

Question. After adjusting for sex do we get the same estimate for the heritability for height and BMI?

```
> reg.MZ_bmi = lm(twin.data$bmi_t2[MZ] ~ twin.data$sex[MZ] + twin.data$bmi_t1[MZ])
> reg.DZ_bmi = lm(twin.data$bmi_t2[DZ] ~ twin.data$sex[DZ] + twin.data$bmi_t1[DZ])
> A_bmi = 2*(coef(reg.MZ_bmi)[3] - coef(reg.DZ_bmi)[3])
> A_bmi
```

Question. Is there evidence for common environment effects?

We calculate $\widehat{\sigma}_c^2$ and $\widehat{\sigma}_e^2$.

```
> C_bmi = coef(reg.MZ_bmi)[3] - A_bmi
> E_ht = 1 - (A_bmi + C_bmi)
```

```
Writing to file Additional Files For Students/Rcode/PartB - Quantitative Genetics.R
```