Practical 01 SG: Descriptive analysis of genetic markers

Carlos Moyano & Kleber Reyes

2022-11-07

SNP dataset

Questions about SNP dataset

3. How many variants are there in this database? What percentage of the data is missing? How many individuals in the database are males and how many are females?

Our sample of data has 20659 variants (number of columns of our genetic subset), as for missing data, value is very low taking into account that is common to reach to values around 10%, in our case is less than 1% (0.2%). Finally we have checked the number of males and females. From here on, let's assume values 1 is male and value 2 is female. The dataset appears to be well distributed since we have distribution close to 50-50.

```
variants <- ncol(geneticData); variants # variants in the database
```

```
## [1] 20649
```

```
individuals <- nrow(geneticData)
perc.mis <- 100*sum(is.na(geneticData))/(variants*individuals); perc.mis # 0.1987%
```

```
## [1] 0.1986518
```

```
male.ind <- length(which(individualData$SEX == 1))
female.ind <- length(which(individualData$SEX == 2))

perc.male <- 100* male.ind / individuals
perc.female <- 100* female.ind / individuals
perc.male; perc.female # 51.96% male - 48.04% female
```

```
## [1] 51.96078
```

```
## [1] 48.03922
```

4. Calculate the percentage of monomorphic variants (AA or BB). Exclude all monomorphics from the database for all posterior computations of the practical. How many variants do remain in your database?

In order to be efficient, we have calculated first the polymorphic variants (columns that contains AB or BA) that is represented with value 1. If a variant contains a polymorphic we can include it in our final dataset. The rest is used to calculate the percentage of monomorphic variants. The percentage of monomorphic variants is 11.5%, that is, less than a quarter of our dataset. The number of variants that we remain are the number of polymorphic variants previously calculated, 18274.

```
cols <- which(colSums(geneticData == 1, na.rm = TRUE) > 0) # Non monomorphic (contains AB)
variants.poly <-length(cols); variants.poly
```

```
## [1] 18274
```

```
variants.mono <- variants-variants.poly
perc.mono <- 100*variants.mono/variants; perc.mono
```

```
## [1] 11.50177
```

```
geneticData.poly <- geneticData[, cols]
```

5. Report the genotype counts and the minor allele count of polymorphism rs8138488_C, and calculate the MAF (Minor Allele Frequency) of this variant.

In order to use the genotype function from the genetics package we must recode the variant to transform our numeric codes into a character pairs. Once *rs8138488_C* has been transformed we can use the genotype function without any problem. Genotype is able to calculate the required information. *genotype.freq* returns us a table with count and proportions of the 3 kinds of genotypes we can found. The minor allele count (MAC) and MAF is alelle "B" with 75 appearances and a frequency of 36.76%

```
rs8138488_C <- dplyr::recode(geneticData.poly[, "rs8138488_C"], `0`="AA", `1`="AB", `2`="BB")
rs8138488_C.g <- genotype(rs8138488_C,sep="")
rs8138488_C.g.summary <- summary(rs8138488_C.g)
rs8138488_C.g.summary$genotype.freq
```

```
##     Count Proportion
## A/A    41  0.4019608
## A/B    47  0.4607843
## B/B    14  0.1372549
```

```
rs8138488_C.g.summary$allele.freq
```

```
##   Count Proportion
## A   129  0.6323529
## B    75  0.3676471
```

```
MAC = min(rs8138488_C.g.summary$allele.freq[,"Count"]); MAC
```

```
## [1] 75
```

```
MAF = 100*min(rs8138488_C.g.summary$allele.freq[,"Proportion"]); MAF
```

```
## [1] 36.76471
```

6. Compute the minor allele frequencies (MAF) for all markers, and make a histogram of it. Does the MAF follow a uniform distribution? What percentage of the markers have a MAF below 0.05? And below 0.01? Can you explain the observed pattern?

With the aim of answer the next questions, we had to recode all the genetic data to be able to use genotype().

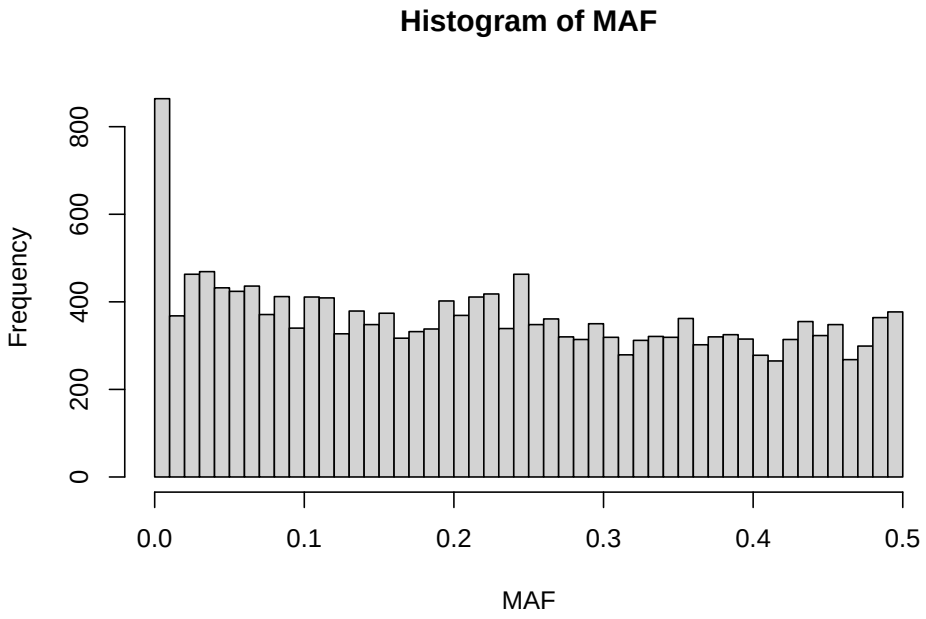
```
for (i in 1:variants.poly) {
  geneticData.poly[, i] <- dplyr::recode(geneticData.poly[, i], `0`="AA", `1`="AB", `2`="BB")
}
```

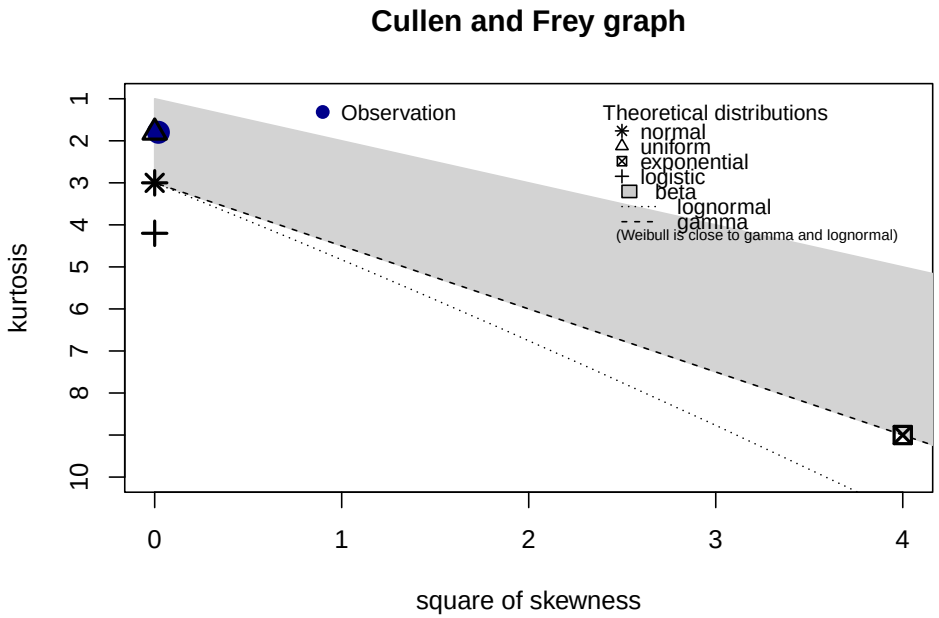
We had to populate 3 arrays in order to calculate this and the following questions, to avoid redundant operations we take advantage of the same loop to calculate 3 things, the MAF, the observed heterozygosity (Ho) and the expected heterozygosity (He) of each variation. We have calculate the MAF in the same way as the previous exercise. To calculate 'Ho' and 'He' we have used the formulas used in the class slides. The *genotype.freq* table is required to our calculations and we have overlooked the *Hu* value given by genotype(). So, Ho is the frequency of the AB genotype and He is equal to 1 - sum(AA and BB genotypes).

```
maf.list <- vector(mode="numeric", length=variants.poly)

ho.list <- vector(mode="numeric", length=variants.poly)
he.list <- vector(mode="numeric", length=variants.poly)

for (i in 1:variants.poly) {
  variant.g <- genotype(geneticData.poly[, i],sep="")
  variant.g.summary <- summary(variant.g)

  # MAF
  maf.list[i] <- min(variant.g.summary$allele.freq[,"Proportion"], na.rm = T)

  gen.freq <- variant.g.summary$genotype.freq
  # Ho
  if ("A/B" %in% rownames(gen.freq)) {
    ho.list[i] <- gen.freq["A/B",2]
  } else if ("B/A" %in% rownames(gen.freq)) {
    ho.list[i] <- gen.freq["B/A",2]
  }

  # He
  sum.p2 <- 0
  if ("A/A" %in% rownames(gen.freq)) {
    sum.p2 <- sum.p2 + gen.freq["A/A",2]
  }
  if ("B/B" %in% rownames(gen.freq)) {
    sum.p2 <- sum.p2 + gen.freq["B/B",2]
  }

  he.list[i] <- 1- sum.p2
}
```

As for this exercise we have to create an histogram of the MAF values obtained. When we set breaks to 50 we can realize that values between 0 and 0.01 are the most common, but in general, the distribution seems to be uniform.

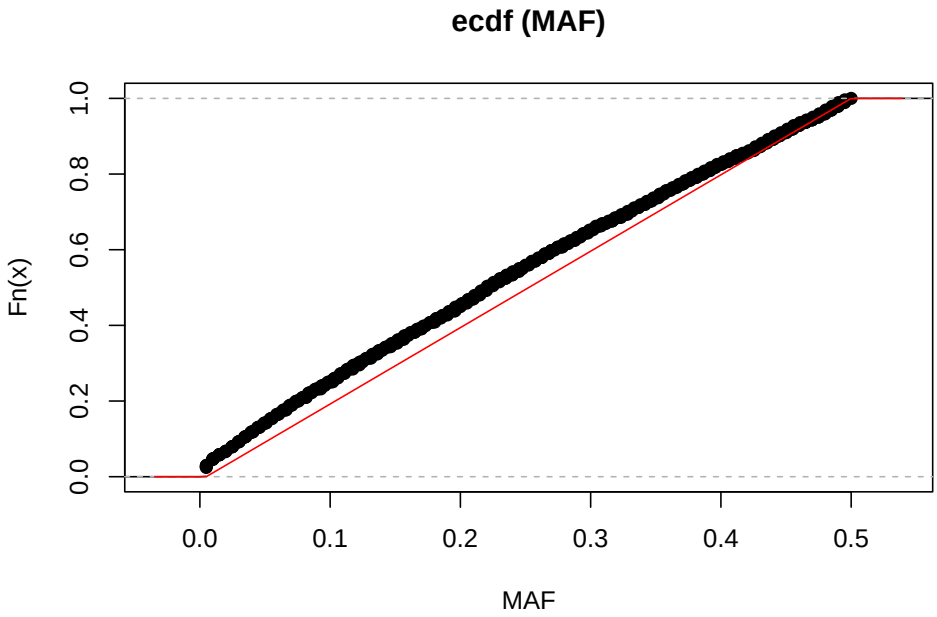
Additional plots have been made in order to reinforce our first impressions. The Cullen and Frey graph indicates that it is practically uniform. Same conclusions with ecdf. In all cases, uniform distribution is not perfect due to the observed deviation of the small values.






```
## summary statistics
```

```
## ------
## min:  0.004901961   max:  0.5
## median:  0.2205882
## mean:  0.2309362
## estimated sd:  0.1474513
## estimated skewness:  0.1407874
## estimated kurtosis:  1.797766
```



As we have seen in the previous plots, the distribution is close to be uniform, but it has some deviations. The percentage of markers below 0.05 is 14.18% and and the percentage of markers below 0.01 is 4.68%.

Since range of values are between 0 and 0.5 we expect that an increase of 0.01 is equivalent to and increment of 2% of the data, but in our case the first 0.01 increase 4.7% and an increase of 0.05 (that should be 10%) is 14.18%. This entails that rare variants (less than 0.05) are more common than the other ones.

```
values.under.005 <- length(which(maf.list < 0.05)); values.under.005
```

```
## [1] 2592
```

```
values.under.001 <- length(which(maf.list < 0.01)); values.under.001
```

```
## [1] 856
```

```
maf.005 <- 100 * values.under.005 / variants.poly; maf.005
```

```
## [1] 14.18409
```

```r
maf.001 <- 100 * values.under.001 / variants.poly; maf.001
```

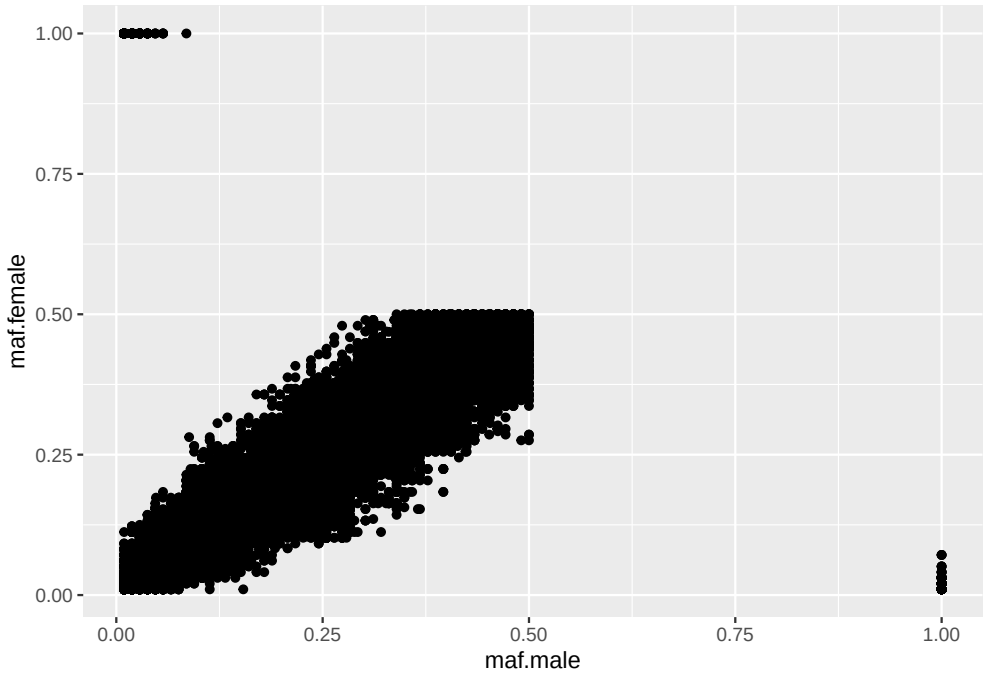
```
## [1] 4.684251
```

7. Calculate the minor allele frequency for males and for females and present a scatterplot of these variables. What do you observe? Calculate and report their correlation coefficient.

We have done the same as the question 6 but now splitting the dataframe into male individuals and female individuals.

```r
geneticData.poly.male <- geneticData.poly[ which( individualData$SEX == 1), ]
geneticData.poly.female <- geneticData.poly[which( individualData$SEX == 2), ]

maf.male <- vector(mode="numeric", length=variants.poly)
maf.female <- vector(mode="numeric", length=variants.poly)

for (i in 1:variants.poly) {
  variant.g <- genotype(geneticData.poly.male[, i],sep="")
  variant.g.summary <- summary(variant.g)
  maf.male[i] = min(variant.g.summary$allele.freq[,"Proportion"], na.rm = T)

}

for (i in 1:variants.poly) {
  variant.g <- genotype(geneticData.poly.female[, i],sep="")
  variant.g.summary <- summary(variant.g)
  maf.female[i] = min(variant.g.summary$allele.freq[,"Proportion"], na.rm = T)

}
```

From the scatterplot we can realize that for the majority of markers, the values are more or less the same for male and female. There are some exceptions in which while for a group the value is 1 for the other values is low. Since the value is calculated using the minimum a value of 1 implies that there is not a second allele for this variant, this means that for this variantions, sex could be important.



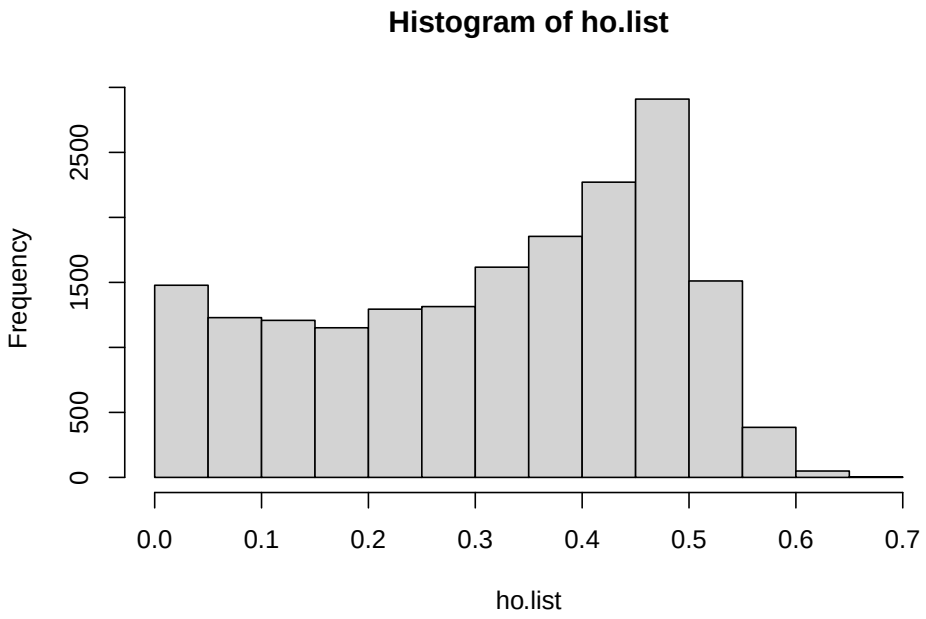
We have calculated the correlation using spearman since distributions are not normal. There is a high positive correlation (0.71) that implies that there is not a great difference between males and females.

```
cor(maf.df, method="spearman")
```

```
##             maf.male maf.female
## maf.male   1.0000000  0.7073359
## maf.female 0.7073359  1.0000000
```

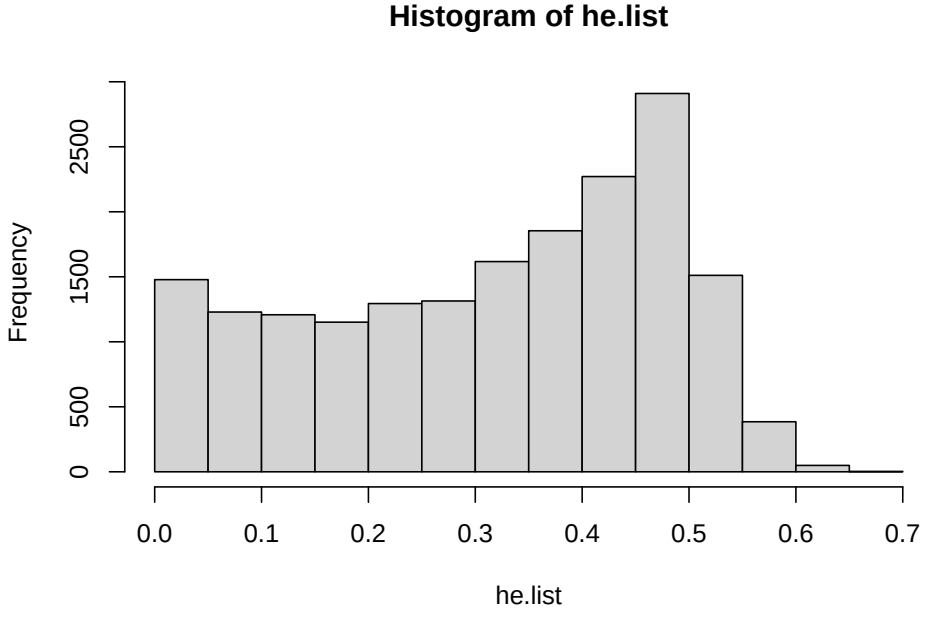
8. Calculate the observed heterozygosity (Ho), and make a histogram of it. What is, theoretically, the range of variation of this statistic?

Observed heterozygosity has been calculated previously. Theoretically, the observed heterozygosity range is from 0 to 1, but in the histogram we can see that the range is from 0 to 0.7.



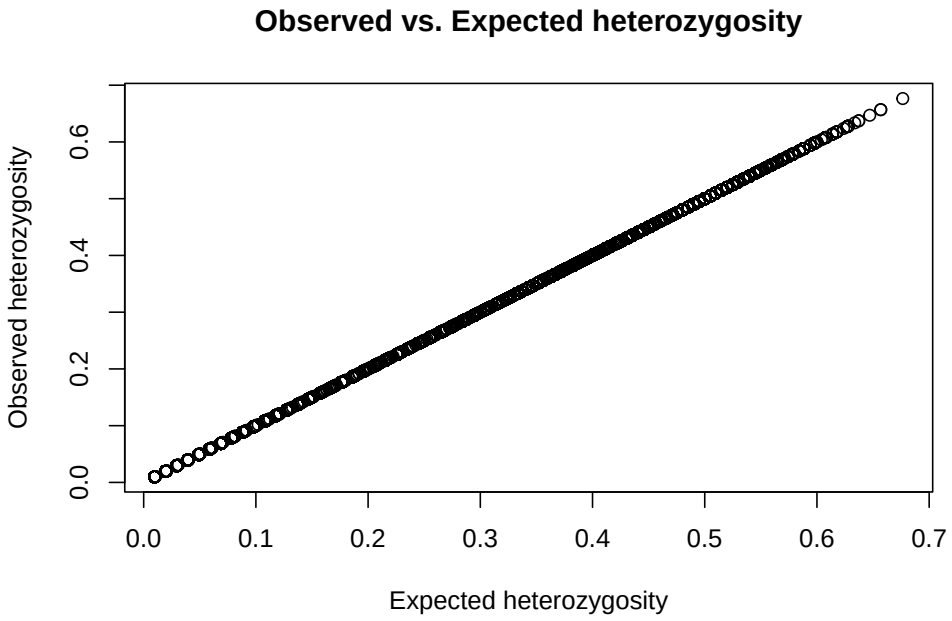
9. Compute for each marker its expected heterozygosity (He), where the expected heterozygosity for a bi-allelic marker is defined as 1 − E(from i=1 to k) pi^2 , where pi is the frequency of the ith allele. Make a histogram of the expected heterozygosity. What is, theoretically, the range of variation of this statistic? What is the average of He for this database?

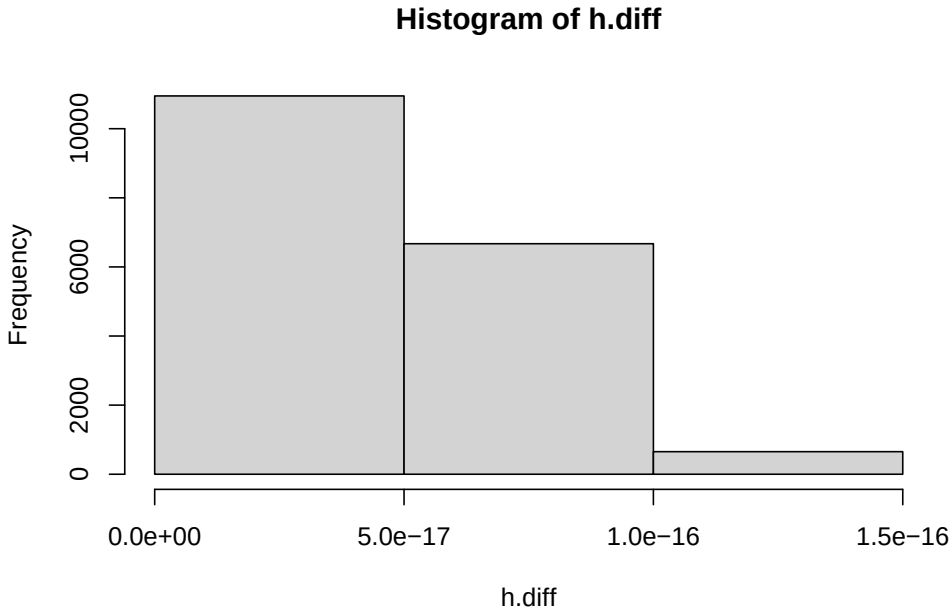
Expected heterozygosity has been calculated previously. Theoretically, the expected heterozygosity range is from 0 to 1, but in the histogram we can see that the range is from 0 to 0.7. The mean He is 0.314.



```
## [1] 0.314699
```

Since histograms apparently are the same we can calculate the absolute difference. We can see that there is a slighly difference between the 2 values.





STR dataset

Questions about STR dataset

2. How many individuals and how many STRs contains the database?

```
X <- NistSTRs
n <- nrow(X) # number of individuals
p <- ncol(X)/2 # number of STRs
n
```

```
## [1] 361
```

```
p
```

```
## [1] 29
```

There are 361 individuals and 29 STRs.

3. Write a function that determines the number of alleles for a STR. Determine the number of alleles for each STR in the database. Compute basic descriptive statistics of the number of alleles (mean, standard deviation, median, minimum, maximum).

```
# Function that determines the number of alleles for a STR.
n.alleles <- function(X, str.index) {
  allele.1 <- as.list(X[,str.index])
  allele.2 <- as.list(X[,(str.index+1)])
  return(length(table(unlist(c(allele.1, allele.2))))) # number of alleles
}

n.alleles.per.str.list <- list()
str.index <- 1
for (str.num in 1:p) {
  n.alleles.per.str.list  <- append(n.alleles.per.str.list, n.alleles(X, str.index))
  str.index <- str.index + 2
}
n.alleles.per.str <- unlist(n.alleles.per.str.list)

# Basic descriptive statistics of the number of alleles
mean(n.alleles.per.str)
```

```
## [1] 11.86207
```

```
sd(n.alleles.per.str)
```

```
## [1] 6.226236
```

```
median(n.alleles.per.str)
```

```
## [1] 10
```

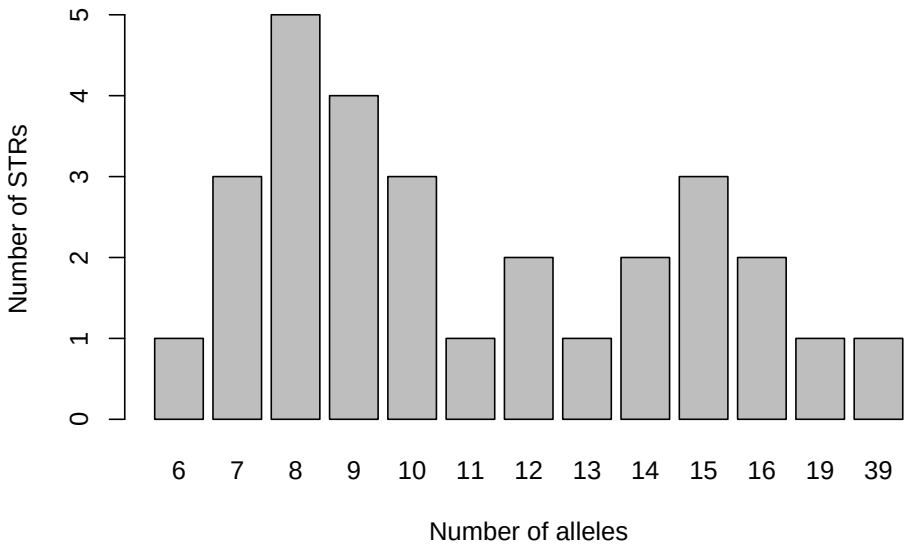
```
max(n.alleles.per.str)
```

```
## [1] 39
```

```
min(n.alleles.per.str)
```

```
## [1] 6
```

4. Make a table with the number of STRs for a given number of alleles and present a barplot of the number STRs in each category. What is the most common number of alleles for an STR?



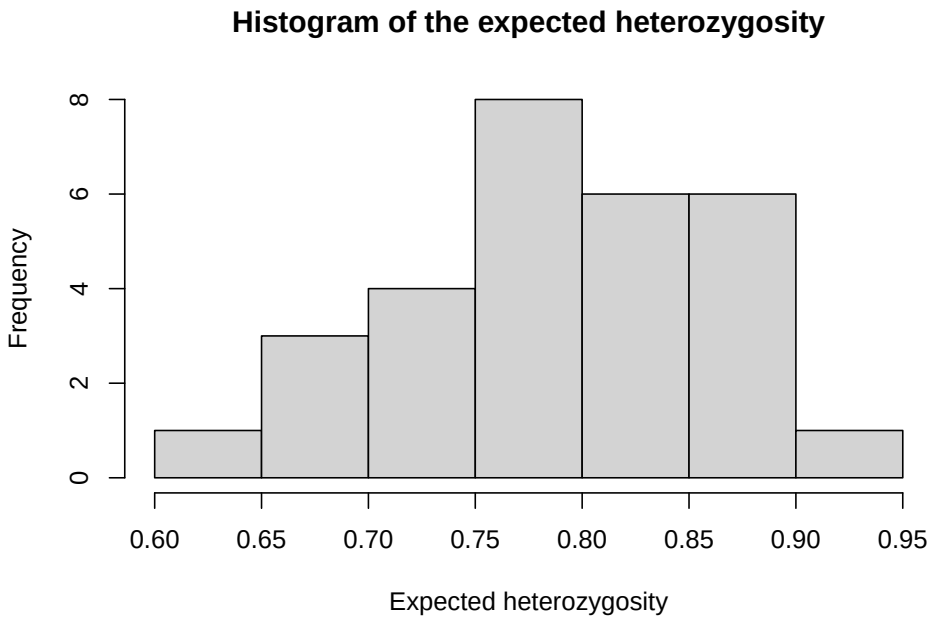
The most common number of alleles for an STR is 8.

5. Compute the expected heterozygosity for each STR. Make a histogram of the expected heterozygosity over all STRS. Compute the average expected heterozygosity over all STRs.

```
exp.heter <- function(X, str.index) {
  allele.1 <- as.list(X[,str.index])
  allele.2 <- as.list(X[,(str.index+1)])
  t <- table(unlist(c(allele.1, allele.2)))
  sum.t <- sum(unname(t)) # we sum the counts
```

```
  exp.heter <- round(1 - sum(sapply(unname(t), function(x) (x / sum.t)^2 )), 3)
  return(exp.heter) # expected heterozygosity formula
}

exp.heter.per.str.list <- list()
str.index <- 1
for (str.num in 1:p) {
  exp.heter.per.str.list  <- append(exp.heter.per.str.list, exp.heter(X, str.index))
  str.index <- str.index + 2
}
exp.heter.per.str <- unlist(exp.heter.per.str.list)

round(mean(exp.heter.per.str), 3) # average expected heterozygosity over all STRs
```

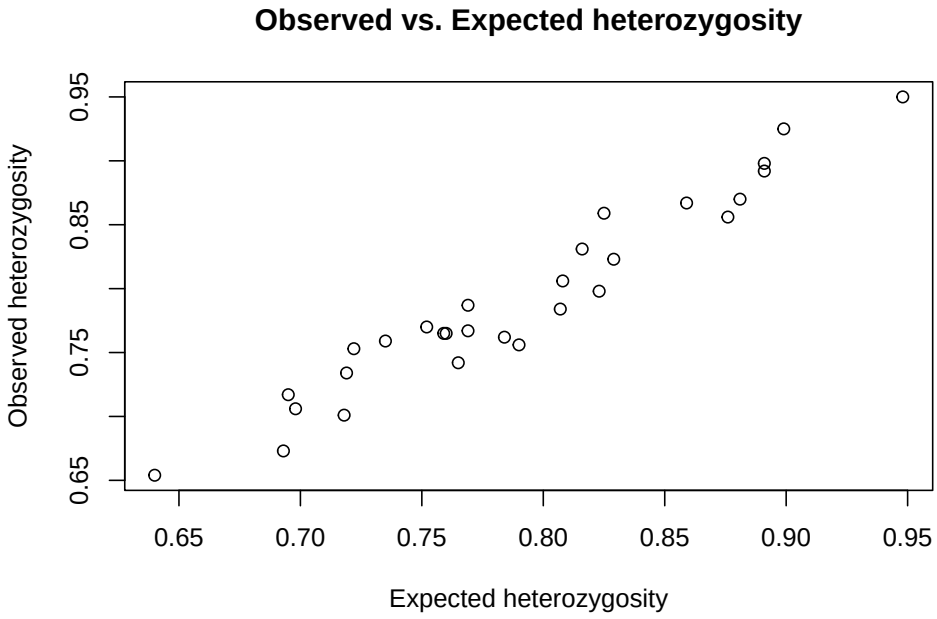
```
## [1] 0.79
```



6. Calculate also the observed heterozygosity for each STR. Plot observed against expected heterozygosity, using all STRs. What do you observe? (Ho = fAB)

```
obs.heter <- function(X, str.index) {

  allele.1 <- X[,str.index]
  allele.2 <- X[,str.index+1]
  allele.1n <- pmin(allele.1,allele.2)
  allele.2n <- pmax(allele.1,allele.2)

  index_different <- allele.1n != allele.2n
```

```
  individuals_heter <- paste(allele.1n[index_different], allele.2n[index_different],sep="/")
  individuals_heter

  individuals <- paste(allele.1n, allele.2n,sep="/")
  g.counts.sum <- sum(table(individuals))

  g.heter.counts.sum <- sum(table(individuals_heter))
  g.heter.counts.sum

  Ho <- round(g.heter.counts.sum / g.counts.sum, 3)

  return(Ho)
}

obs.heter.per.str.list <- list()
str.index <- 1
for (str.num in 1:p) {
  obs.heter.per.str.list  <- append(obs.heter.per.str.list, obs.heter(X, str.index))
  str.index <- str.index + 2
}
obs.heter.per.str <- unlist(obs.heter.per.str.list)
```



The plot above show a clear linear relationship between the Ho and the He.

7. Compare, overall, the results you obtained for the SNP database with those you obtained for the STR database. What differences do you observe between these two types of genetic markers?

In both datasets we have a intense linear relationship between Ho and He, maybe stronger in the case of the Italian sample. We have noted that the Caucasian ancestry sample has an average He over all STRs almost

3 times higher than in the case of the italian sample (0.79 vs. 0.314). This means that, in general, it's more likely for the genetic variables associated to chromosome 21 to come from the same parent.