## Exercise

# Population Genetics

Task 1

From a case-control study, the following genotype counts have been observed for a SNP with alleles A and B:

Cases: nobs(AA) = 159 nobs(AB) = 122 nobs(BB) = 19

Controls: nobs(AA) = 120 nobs(AB) = 139 nob (BB) = 41

Test this SNP for deviation from Hardy-Weinberg equilibrium, calculate the *FST* statistic in controls, test for association under a genotypic and under an allelic risk model and calculate the odds ratios for the SNP genotypes and alleles, respectively.

1. Testing for deviation from Hardy-Weinberg equilibrium (HWE) in controls

1. Calculate the genotype and allele frequencies in controls:

n = nobs(AA) + nobs(AB) + nobs(BB) =

fobs(AA) = nobs(AA) / n =

fobs(AB) = nobs(AB) / n =

fobs(BB) = nobs(BB) / n =

f(A) = [2×nobs(AA)+nobs(AB)] / 2n =

f(B) = [2×nobs(BB)+nobs(AB)] / 2n =

2. Calculate the expected genotypic counts under the null hypothesis of HWE:

nexp(AA) = n × [ f(A)×f(A) ] =

nexp(AB) = n × [ 2×f(A)×f(B) ] =

nexp(BB) = n × [ f(B)×f(B) ] =

3. Arrange observed and expected genotype counts in a 2×3 table and calculate the chi-square statistic:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AA** | **AB** | **BB** |
| Observed (nobs) | 120 | 139 | 41 |
| Expected (nexp) |  |  |  |

X2 = [nobs(AA)- nexp(AA)]2 / nexp(AA) + [nobs(AB)- nexp(AB)]2 / nexp(AB)  
 + [nobs(BB)- nexp(BB) ]2 / nexp(BB)

= + +

=

4. Obtain the corresponding *P*-value from a *one*-df χ2 distribution:

Quantiles of the 1-df χ2 distribution using R:

pchisq( <<*QUANTILE>>*, df=1, ncp=0, lower.tail=F)

*p* =

## Questions

1. Is there statistical evidence at the 0.05 level that the marker is not in HWE?

2. The reported genotype counts were observed in controls only. Would it be beneficial to merge the control genotype counts with those from the cases to test HWE testing, since it would increase the sample size and power for this test? Give a reason for your answer.

2. Calculating the *FST* statistic in controls

The *FST* statistic can be calculated by the formula given below (introduced in the lecture on population genetics). Use the control frequencies calculated in Exercise 1 on HWE testing.

*FST* = [ fobs(AA) – f(A)×f(A)] / [ f(A) – f(A)×f(A) ]

*FST* = /

*FST* =

## Question

3. How do you interpret this value of *FST* with regard to the sample (see the lecture for an interpretation of the value)?

Task 2

Perform a simulation study on genetic drift. In particular, assume a SNP with two alleles of frequency 50% at generation 0, population sizes of 10, 100, 1000 and 10,000 chromosomes, respectively, and 50 generations. For each population size, perform 100 replications. Graph the allele frequencies by generation for all 100 replications of size 1000 within a single graph. Summarize replication results by generating a histogram and calculate summary statistics for the allele frequencies after 50 generations for each of the three population sizes.

The required R script is placed in the file **popgen\_drift.q**. Open this file in a text editor and copy-paste the commands in the R console (you can also use the source command in R). Inspect the resulting graphs.

**popgen\_drift.q**

N.s = c(10, 100, 1000, 10000)

n.gen = 50

n.rep = 100

# === Simulate genetic drift === #

for (n in N.s) {

freqs = matrix(NA, ncol=n.gen+1, nrow=n.rep)

for (i in 1:n.rep) {

alleles = c(rep(0,n/2), rep(1,n/2))

freqs[i,1] = sum(alleles==1) / length(alleles)

for (j in 1:n.gen) {

alleles = sample (alleles, length(alleles), replace=T)

freqs[i,j+1] = sum(alleles==1) / length(alleles)

}

}

assign (paste("freqs.n", n, sep=""), freqs)

}

summary(freqs.n10 [,n.gen+1])

summary(freqs.n100 [,n.gen+1])

summary(freqs.n1000 [,n.gen+1])

summary(freqs.n10000[,n.gen+1])

# === Graph allele frequency changes === #

pdf("drift\_plot.pdf", paper="special", height=4\*2, width=4\*2, onefile=F)

split.screen(c(2,2))

screen(1)

plot(x=0, y=0, type="n", xlim=c(0,n.gen), ylim=c(0,1), xlab="Generation", ylab="Allele frequency", main="N=10")

lines (c(0,n.gen), rep(0.5, 2), lty=3, col="#AAAAAA")

freqs = get(paste("freqs.n", 10 , sep=""))

for (i in 1:n.rep) {

lines(0:n.gen,freqs[i,], col="#44AAAA")

}

screen(2)

plot(x=0, y=0, type="n", xlim=c(0,n.gen), ylim=c(0,1), xlab="Generation", ylab="Allele frequency", main="N=100")

lines (c(0,n.gen), rep(0.5, 2), lty=3, col="#AAAAAA")

freqs = get(paste("freqs.n", 100 , sep=""))

for (i in 1:n.rep) {

lines(0:n.gen,freqs[i,], col="#44AAAA")

}

screen(3)

plot(x=0, y=0, type="n", xlim=c(0,n.gen), ylim=c(0,1), xlab="Generation", ylab="Allele frequency", main="N=1000")

lines (c(0,n.gen), rep(0.5, 2), lty=3, col="#AAAAAA")

freqs = get(paste("freqs.n", 1000 , sep=""))

for (i in 1:n.rep) {

lines(0:n.gen,freqs[i,], col="#44AAAA")

}

screen(4)

plot(x=0, y=0, type="n", xlim=c(0,n.gen), ylim=c(0,1), xlab="Generation", ylab="Allele frequency", main="N=10000")

lines (c(0,n.gen), rep(0.5, 2), lty=3, col="#AAAAAA")

freqs = get(paste("freqs.n", 10000, sep=""))

for (i in 1:n.rep) {

lines(0:n.gen,freqs[i,], col="#44AAAA")

}

close.screen(all.screens=T)

dev.off()

# === Graph allele frequency after 50 generations === #

pdf("drift\_hist.pdf", paper="special", height=4\*2, width=4\*2, onefile=F)

split.screen(c(2,2))

screen(1)

hist(freqs.n10 [,n.gen+1], main="N=10", xlab="Allele frequency", xlim=c(0,1))

screen(2)

hist(freqs.n100 [,n.gen+1], main="N=100", xlab="Allele frequency", xlim=c(0,1))

screen(3)

hist(freqs.n1000 [,n.gen+1], main="N=1000", xlab="Allele frequency", xlim=c(0,1))

screen(4)

hist(freqs.n10000[,n.gen+1], main="N=10000", xlab="Allele frequency", xlim=c(0,1))

close.screen(all.screens=T)

dev.off()

Task 3

Explore the impact of the strength of selection on the change in allele frequency. To this end, assume selection coefficients s of 0.001, 0.01 and 0.1, respectively, a heterozygote effect of h=0.5 and a deleterious allele frequency of 50% at generation 0. Calculate the allele frequencies within the next 100 generations and graph them by generation. Give the allele frequencies after 100 generations for each of the three coefficients.

fs=0.001(100) = fs=0.01(100) = fs=0.1(100) = .

The required R script is placed in the file **popgen\_selection.q**. Open this file in a text editor and copy-paste the commands in the R console (you can also use the source command in R). Inspect the resulting graph.

**popgen\_selection.q**

s.s = c(0.001, 0.01, 0.1)

h = 0.5

n.gen = 100

# === Calculate allele frequency changes === #

for (s in s.s) {

freqs = rep(NA, n.gen+1)

af = 0.5

freqs[1] = 1-af

for (j in 1:n.gen) {

omega = 1 - 2\*af\*(1-af)\*h\*s - (1-af)\*(1-af)\*s

f.het = (1-h\*s)\*2\*af\*(1-af)/omega

f.hom = af\*af/omega

af = f.hom + f.het/2

freqs[j+1] = 1-af

}

assign (paste("freqs.s", s, sep=""), freqs)

}

# === Report allele frequencies after 100 generations === #

for (s in s.s) {

cat("s=") ; cat(s) ; cat(": ")

freqs = get(paste("freqs.s", s, sep=""))

cat(freqs[n.gen+1]) ; cat("\n")

}

# === Graph allele frequency changes === #

pdf("selection\_plot.pdf", paper="special", height=4\*2, width=4\*2, onefile=F)

plot(x=0, y=0, type="n", xlim=c(0,n.gen), ylim=c(0,1), xlab="Generation", ylab="Allele frequency")

lines (c(0,n.gen), rep(0.5, 2), lty=3, col="#AAAAAA")

for (s in s.s) {

freqs = get(paste("freqs.s", s, sep=""))

lines(0:n.gen,freqs, col="#44AAAA")

}

dev.off()

## Answers

# Population Genetics

Task

From a case-control study, the following genotype counts have been observed for a SNP with alleles A and B:

Cases: nobs(AA) = 159 nobs(AB) = 122 nobs(BB) = 19

Controls: nobs(AA) = 120 nobs(AB) = 139 nob (BB) = 41

Test this SNP for deviation from Hardy-Weinberg equilibrium, calculate the *FST* statistic in controls, test for association under a genotypic and under an allelic risk model and calculate the odds ratios for the SNP genotypes and alleles, respectively.

1. Testing for deviation from Hardy-Weinberg equilibrium (HWE) in controls

1. Calculate the genotype and allele frequencies in controls:

n = nobs(AA) + nobs(AB) + nobs(BB) = **300**

fobs(AA) = nobs(AA) / n = **0.400**

fobs(AB) = nobs(AB) / n = **0.463**

fobs(BB) = nobs(BB) / n = **0.137**

f(A) = [2×nobs(AA)+nobs(AB)] / 2n = **0.632**

f(B) = [2×nobs(BB)+nobs(AB)] / 2n = **0.368**

2. Calculate the expected genotypic counts under the null hypothesis of HWE:

nexp(AA) = n × [ f(A)×f(A) ] = **119.7**

nexp(AB) = n × [ 2\*f(A)×f(B) ] = **139.6**

nexp(BB) = n × [ f(B)×f(B) ] = **40.7**

3. Arrange observed and expected genotype counts in a 2×3 table and calculate the chi-square statistic:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AA** | **AB** | **BB** |
| Observed (nobs) | 120 | 139 | 41 |
| Expected (nexp) | **119.7** | **139.6** | **40.7** |

X2 = [nobs(AA)- nexp(AA)]2 / nexp(AA) + [nobs(AB)- nexp(AB)]2 / nexp(AB)  
 + [nobs(BB)- nexp(BB) ]2 / nexp(BB)

= **0.00075** + **0.00258** +**0.00221**

= **0.00554**

4. Obtain the corresponding *P*-value from a *one*-df χ2 distribution

Quantiles of the 1-df χ2 distribution using R:

pchisq(0.00554, df=1, ncp=0, lower.tail=F)

[1] 0.9406673

*p* = **0.94**

## Questions

1. Is there statistical evidence that the marker is not in HWE?

No, the deviation from HWE is not statistically significant.

2. The reported genotype counts were observed in controls only. Would it be beneficial also to use the case genotypes for HWE testing, since it would increase the sample size for these tests? Give the reason for your answer.

No! Cases and controls have been sampled retrospectively and separately. Mixing cases and controls may cause a bias in the frequency estimates since cases are very likely oversampled, compared to their frequency in the population (unless the disease has a very high prevalence). Additionally, HWE is to be expected in the case cohort only under a multiplicative risk model.

HWE should always be tested separately in cases and controls. Inferences of potential genotyping errors by deviations from HWE should be made cautiously. For example, a deviation from HWE in cases might reflect not a genotyping error, but a genuine genetic effect. Removing that SNP because of its low HWE *P*-value in cases would likely reduce the power of the study!

2. Calculating the *FST* statistic in controls

The *FST* statistic can be calculated by the formula given below (introduced in the lecture on population genetics). Use the control frequencies calculated in Exercise 1 on HWE testing.

*FST* = [ fobs(AA) – f(A)×f(A)] / [ f(A) – f(A)×f(A) ]

*FST* = **0.00010** / **0.23266**

*FST* = **0.00429**

## Question

3. How do you interpret this value of *FST* with regard to the sample (see lecture)?

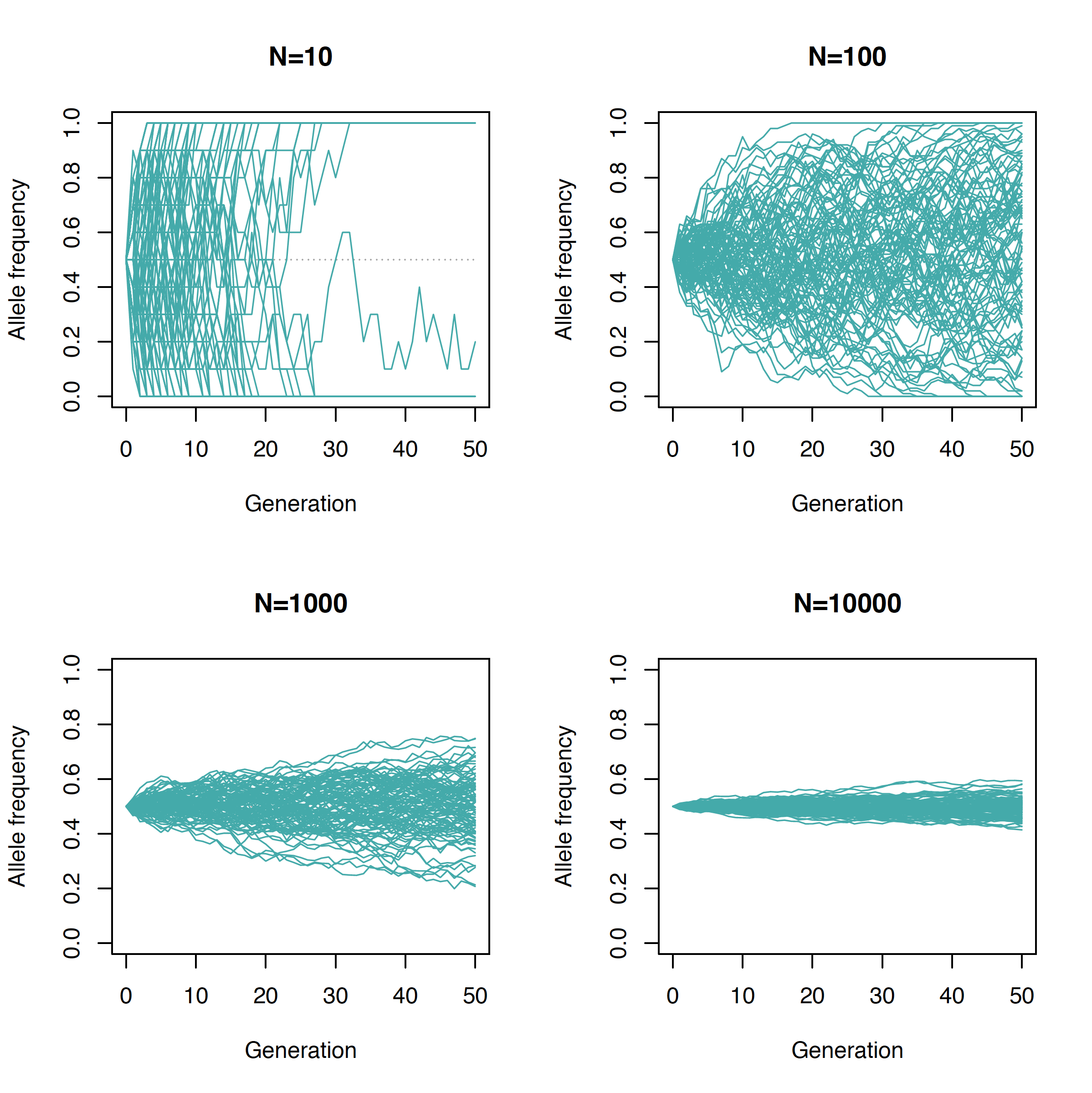
There is a slight deficit of heterozygous genotypes for the investigated marker in the control cohort.

Task 2

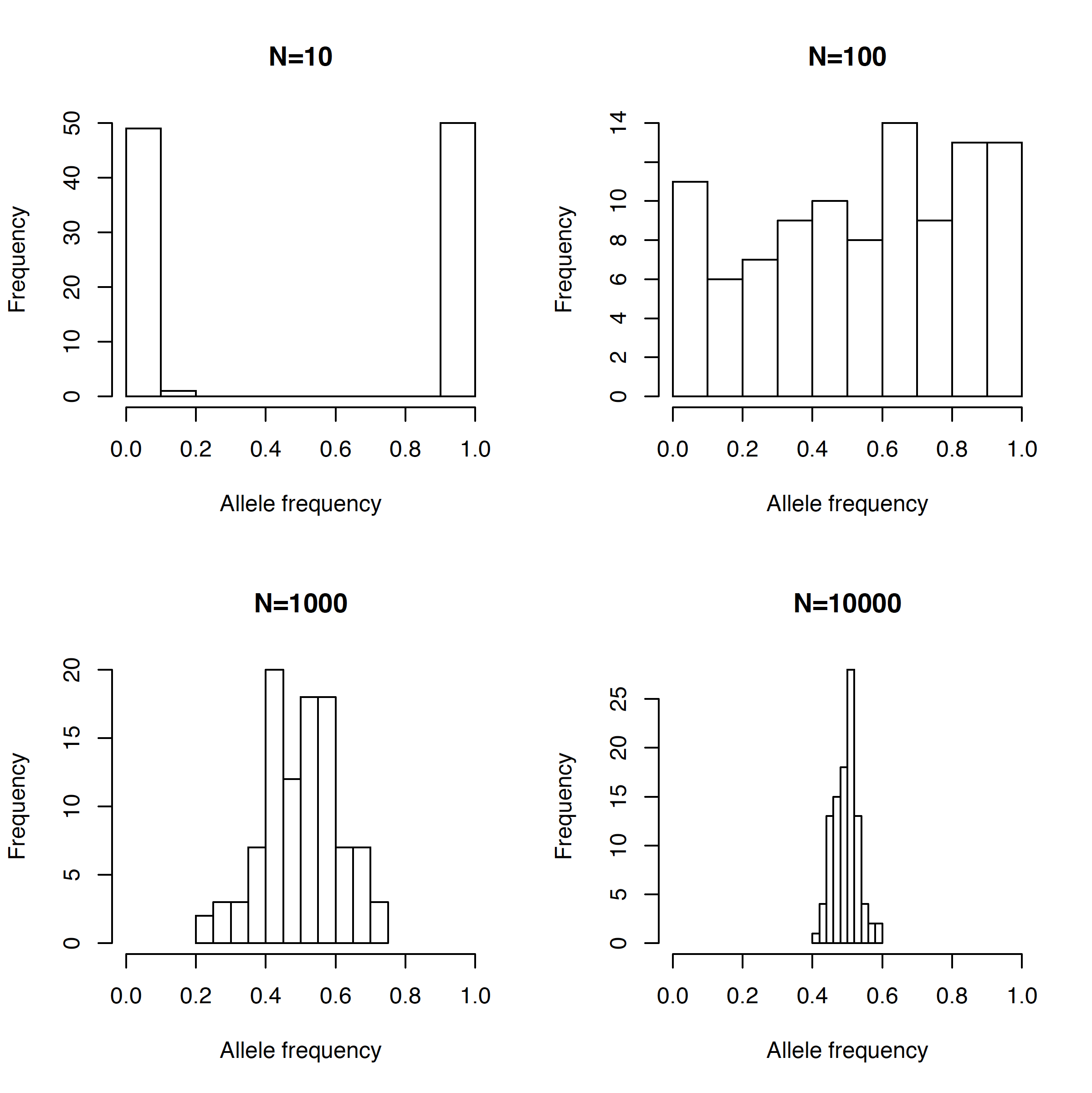
Perform a simulation study on genetic drift. In particular, assume a SNP with two alleles of frequency 50% at generation 0, population sizes of 10, 100, 1000 and 10,000 chromosomes, respectively, and 50 generations. For each population size, perform 100 replications. Graph the allele frequencies by generation for all 100 replications of size 1000 within a single graph. Summarize replication results by generating a histogram and calculate summary statistics for the allele frequencies after 50 generations for each of the three population sizes.

The required R script is placed in the file **popgen\_drift.q**. Open this file in a text editor and copy-paste the commands in the R console (you can also use the source command in R). Inspect the resulting graphs.

**Allele frequency changes**



**Allele frequencies after 50 generations**



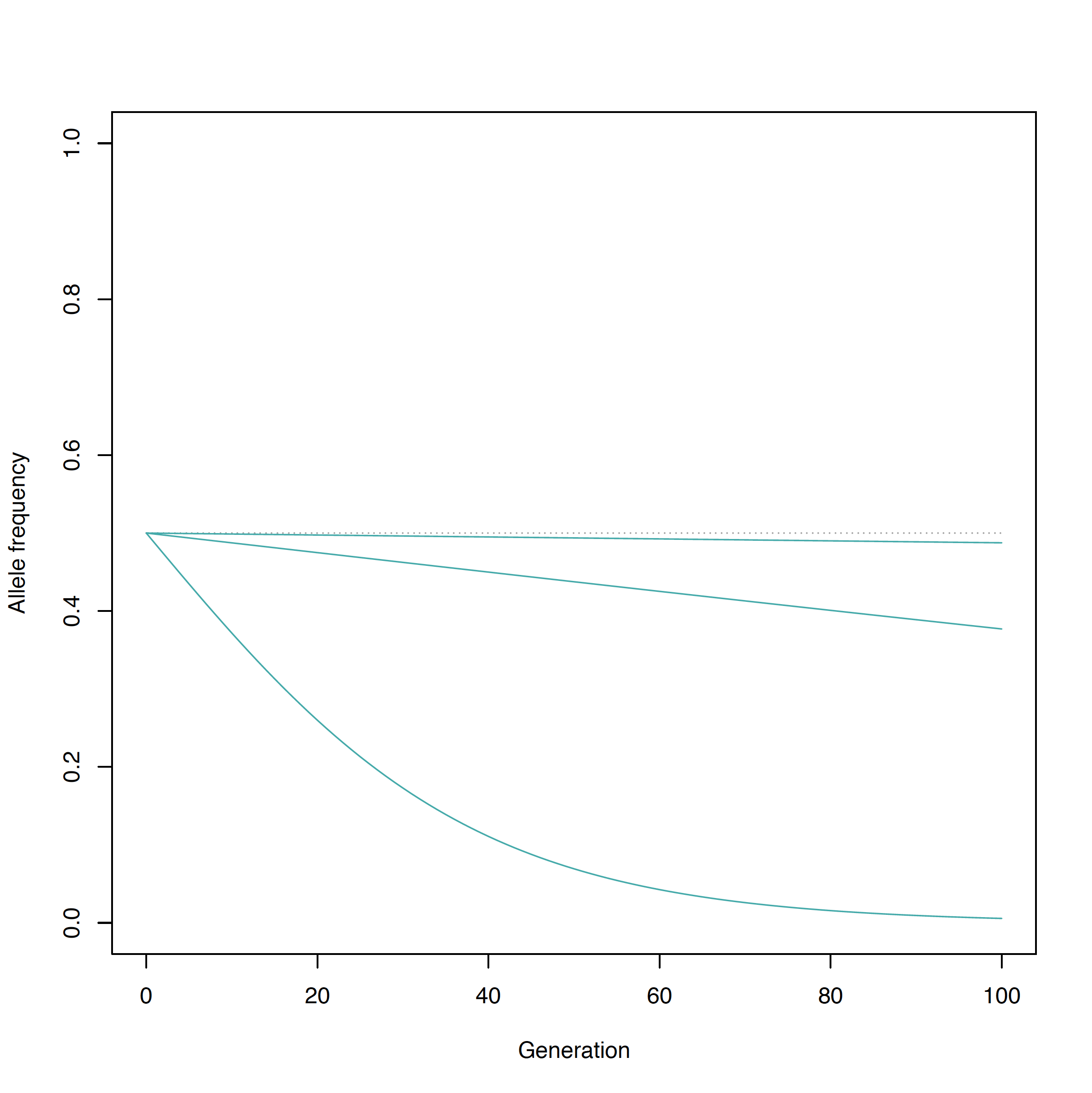
Task 3

Explore the impact of the strength of selection on the change in allele frequency. To this end, assume selection coefficients s of 0.001, 0.01 and 0.1, respectively, a heterozygote effect of h=0.5 and a deleterious allele frequency of 50% at generation 0. Calculate the allele frequencies within the next 100 generations and graph them by generation. Give the allele frequencies after 100 generations for each of the three coefficients.

fs=0.001(100) = fs=0.01(100) = fs=0.1(100) = .

The required R script is placed in the file **popgen\_selection.q**. Open this file in a text editor and copy-paste the commands in the R console (you can also use the source command in R). Inspect the resulting graph.

**Allele frequency changes**



**Allele frequencies after 100 generations**

fs=0.001(100) = 0.4875 fs=0.01(100) = 0.3770 fs=0.1(100) = 0.0057