Homework 1

AUTHOR
Benjamin Zeiger

Question 1

Determine how many different genotypes are present at a locus with n alleles that differ by state. You already know that there is one genotype at a locus with one allele and three genotypes at a locus with two alleles. Continue this with three, four, and more alleles until you derive the general case. Then create a graph showing how the number of genotypes varies by the number of alleles at a diploid locus.

Answer

In a one-allele, diploid gene, there is only one option: AA.

In a two-allele, diploid gene, there will be three options: AA, plus two new ones: BB and AB.

In a three-allele gene, there are now six genotypes: all the previous options, plus three new ones: AC, BC, and CC

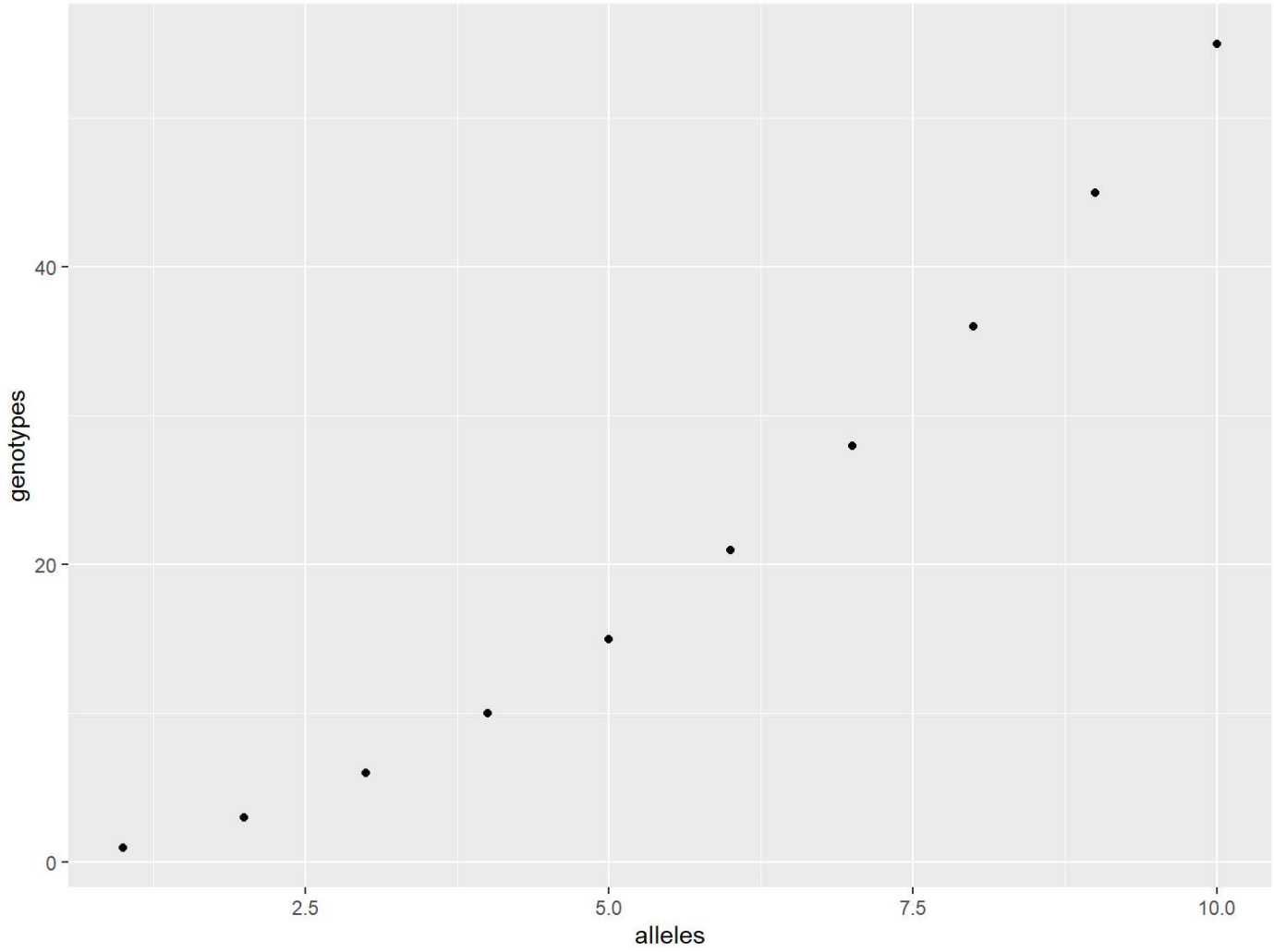
The number of possible genotypes increase following the triangle numbers (1, 3, 6, 10), so the number of possible genotypes in a system of *n* alleles is equal to the sum of all integers between 1 and *n*. This sum is also equal to $n(n+1)/2$, which makes it easier to generate a vector, but I forgot how to derive that from the summation, so I did have to reference the wikipedia on "Triangular Numbers" to get here.

Graphed below, we see the number of possible genotypes increases exponentially as the number of alleles increases:

```
library(tidyverse)

alleles = seq(1,10,1)
genotypes = alleles*(alleles+1)/2

ggplot(mapping = aes(x = alleles, y = genotypes)) +
  geom_point()
```



Question 2

Graph the frequencies of the heterozygotes as a function of the allele frequencies ‘p’ & ‘q’ in a three ALLELE Hardy Weinberg Equilibrium system. At what allele frequencies are the frequency of heterozygotes maximized?

Answer

If we allow ‘p’ to represent the frequency of allele ‘A,’ and ‘q’ to represent frequency of allele ‘B,’ then our 3rd value will be ‘r’ and will represent the frequency of allele ‘C.’

In order to represent this in terms of ‘p’ and ‘q’ we will let r = 1 - p - q.

The frequencies of each genotype should then be:

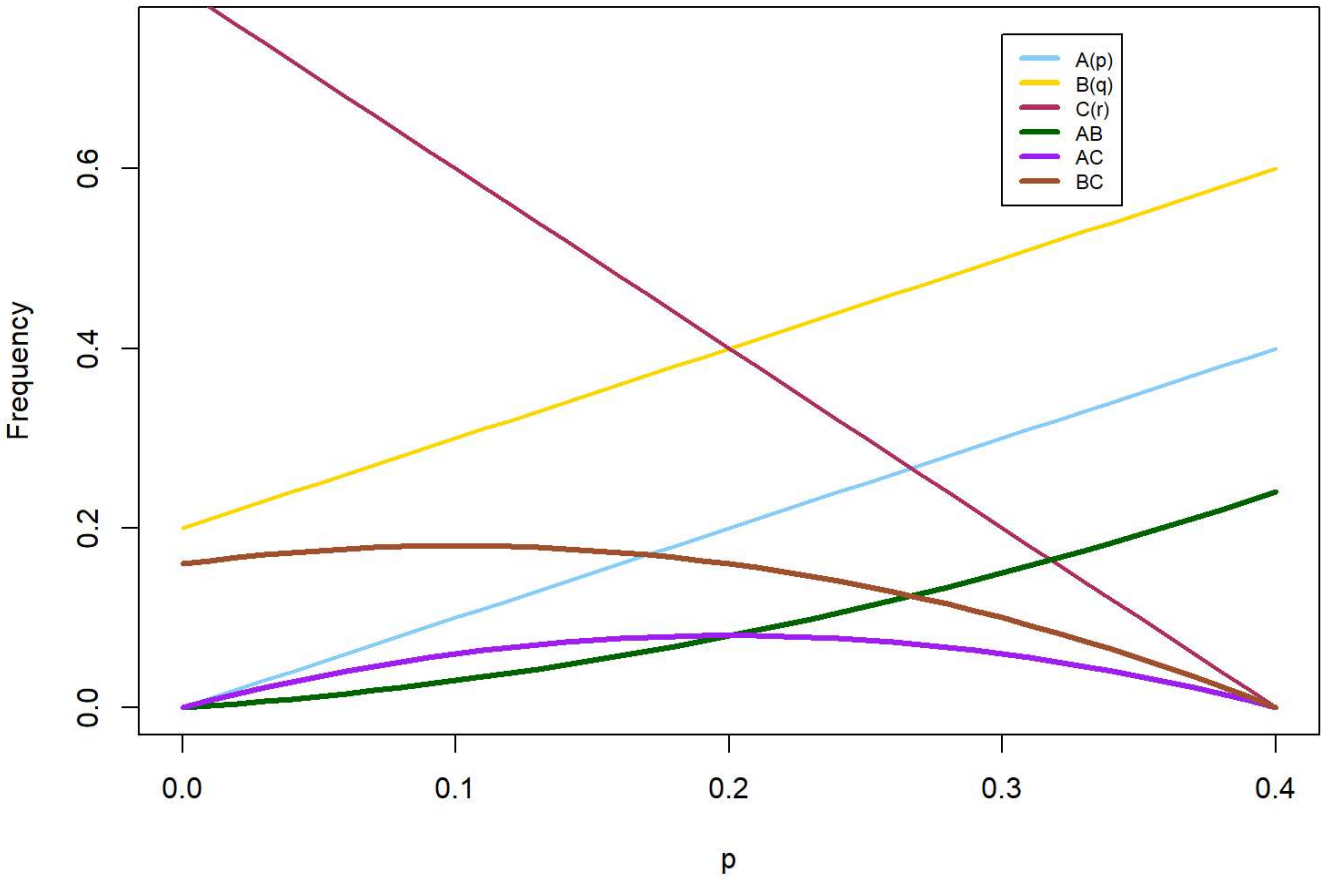
- $f_{AA} = p^2$
- $f_{BB} = q^2$

- $f_{CC} = (1-p-q)^2$
- $f_{AB} = pq$
- $f_{AC} = p * (1 - p - q) = p - p^2 - pq$
- $f_{BC} = q * (1 - p - q) = q - pq - q^2$

```
vec_p = seq(0, 0.4, 0.01)
vec_q = seq(0.2, 0.6, 0.01)
vec_r = 1 - vec_p - vec_q
vec_x = seq(0, 50, 1)

vec_fab = vec_p * vec_q
vec_fac = vec_p - vec_p^2 - vec_fab
vec_fbc = vec_q - vec_q^2 - vec_fab

plot(vec_p, vec_p, type='l',xlab="p",ylab="Frequency",
     ylim = c(0,0.75), col = "lightskyblue", lwd=2)
lines(vec_p, vec_q, col="gold",lwd=2)
lines(vec_p, vec_r, col="maroon",lwd=2)
lines(vec_p, vec_fab, col = "darkgreen", lwd = 3)
lines(vec_p, vec_fac, col = "purple", lwd = 3)
lines(vec_p, vec_fbc, col = "sienna", lwd = 3)
legend(0.3,0.75,legend=c( "A(p)","B(q)", "C(r)","AB","AC","BC"),col=c("lightskyblue","gold",
         "maroon","darkgreen","purple","sienna"),lwd=c(3,3,3),cex=0.7)
```



Each genotype's frequency will be maximized when frequency of each of its alleles is 0.5, (e.g. when A and B are both 0.5, AB will make up 0.25 of the present genotypes). Notably, it seems like in this equilibrium, none of the heterozygotes can have a frequency higher than 0.25 at any given time.

Question 3

In a sample of 1,617 Spanish Basques, the numbers of A, B, O, and AB blood types observed were 724, 110, 763, and 20, respectively. These blood groups are due to three alleles, IA, IB, and IO, with IAIAand IAIO having blood group A; IBIB and IBIO having blood group B, IAIB having blood group AB, and IOIOhaving blood group O. The best estimates of the alleles frequencies in Basque sample are 0.2661 for IA, 0.0411 for IB, and 0.6928 for IO. Calculate the expected numbers of the four blood group phenotypes and carry out a chi-square test for HWE.

Answer

If the frequencies of IA, IB, and IO are represented by 'p,' 'q,' and 'r,' respectively, then the expected frequencies of each blood group would be:

- $A = p^2 + pr$

- B = q^2 + qr
- O = r^2
- AB = pq

And the expected *values* of the blood groups would be equal to these frequencies multiplied by the sample size.

```
p = 0.2661
q = 0.0441
r = 0.6928
N = 1617

f_A = p^2 + (p*r) #frequency of IAIA + frequency of IAIO
f_B = q^2 + (q*r) #frequency of IBIB + frequency of IBIO
f_O = r^2 #frequency of IOIO
f_AB = p*q #frequency of IAIB

basque_chart <- data.frame(phenotype = c("A", "B", "O", "AB"),
                           observed = c(724, 110, 763, 20),
                           expected = c(f_A*N, f_B*N, f_O*N, f_AB*N))

basque_chart$O_E = basque_chart$observed - basque_chart$expected
basque_chart$error = ((basque_chart$O_E)^2 / basque_chart$expected)

test_stat = sum(basque_chart$error)

1 - pchisq(test_stat, 1)
```

```
[1] 0
```

I believe degrees of freedom should be 1 in this case, since there are 2 allele frequencies 'estimated' here ('p' and 'q' would be the "estimated parameters" and the value of 'r' is then assumed from the estimated values?). A chi-square test using a test statistic of 298.1 and a df of 1 results in a p-value of 0, so we reject the null hypothesis. There is sufficient evidence that the population is *not* in HWE.