Practicals in Quantiative Genetic Analyses

Rasmus Bak Stephansen & Peter Sørensen

2022-03-08

Introduction

In these practicals we will be analysing quantitative traits observed in a mice population. The mouse data consist of phenotypes for traits related to growth and obesity (e.g. body weight, glucose levels in blood), pedigree information, and genetic marker data. The practicals will be a mix of theoretical and practical exercises in R that are used for illustrating/applying the theory presented in the lectures and corresponding notes.

- Practical 1: Use R for Analysing Quantitative Traits
- Practical 2: Basic Quantitative Genetics illustrated in the mouse data
- Practical 3: Estimation of Genetic Parameters for traits in the mouse data
- Practical 4: Estimation of Breeding Values for traits in the mouse data
- Practical 5: Estimation of Genomic Breeding Values for traits in the mouse data

Mouse data

The **M16 mouse** was established as an outbred animal model of early onset polygenic obesity and diabesity. This was done by selection for 3- to 6-week weight gain for 27 generations from an outbred ICR base population. Breeding criterion was within-litter selection for the male and female with the largest weight gain from 3 to 6 weeks of age. An ICR control line was maintained in parallel, with random mating from generation to generation but maintaining a similar effective population size. Mice from the M16 line ar elarger at all ages and have increased body fat percentage, fat cell size, fat cell numbers, and organ weights when compared with ICR. Mice from the M16 line are larger at all ages and have increased body fat percentage, fat cell size, fat cell numbers, and organ weights when compared with ICR. These mice also exhibit hyperphagia, accompanied by moderate obesity, and are hyperglycemic, hyperinsulinemic, and hypercholesterolemic.

The **ICR mouse** is a strain of albino mice originating in SWISS and selected by Dr. Hauschka to create a fertile mouse line. Because mice of this strain have been sent to various places from the Institute of Cancer Research in the USA, the strain was named ICR after the initial letters of the institute. Mice of this strain are relatively large albinos with a gentle nature that grow well. The ICR mouse is a general-purpose model used for studies in a wide range of fields including toxicity, pharmacology, drug efficacy, and immunology.

A large **F2 population** (n=1181) was **established by crossing the M16 and ICR lines** (for a recent description of relevant phenotypes in the parental lines, see https://onlinelibrary.wiley.com/doi/epdf/10.1038/oby.2004.176). Twelve F1 families resulted from six pair matings of M16 males x ICR females and six pair matings of the reciprocal cross. A total of 55 F1 dams were mated to 11 F1 sires in sets of five F1 full sisters mated to the same F1 sire. These same specific matings were repeated in three consecutive replicates. Thus, the F2 population consisted of 55 full-sib families of up to 24 individuals each and 11 sire families families of up to 120 individuals each. Actual numbers of mice within families varied slightly due to a small number of failed pregnancies. All litters were standardized at birth to eight pups, with approximately equal representation of males and females.

More information about the mouse data can be found in the following publications:

https://onlinelibrary.wiley.com/doi/epdf/10.1038/oby.2004.176

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1449794/

Practical 1: Use R for Analysing Quantitative Traits

Time schedule of practical session 1:

Time	Activity
11:15	Questions to lecture and multiple-choice questions
11:25	Introduction to R-studio
11:45	Assignments to groups – work with exercises
12:00	Break
12:35	Prepare final words of exercises in each group
12:45	Present final words
12:55	Repeat multiple-choice questions
13:00	End of practical session 1

Introduction:

In this practical we use R for explorative data analyses of two quantitative traits, body weight and blood glucose levels, observed in the F2 mouse population. These explorative data analyses includes computation of basic descriptive statistics such as mean, and variance used to describe each of these traits. Distribution plots (e.g., histogram) will be used to visualize whether the trait phenotypes follow a normal distribution. Boxplots will be used to spot potential effects of explanatory variables. Furthermore relationships between traits and variables will be characterized in terms of correlations and linear relationships.

Let's get started to explore our mouse data

One of the first thing to do is to explore the data used in the analysis. The goal is to understand the variables, how many records the data set contains, how many missing values, what is the variable structure, what are the variable relationships and more. Several commands/functions will be used. To read more about a specific function (e.g., `str`) write `?str`.

The mouse data set can be loaded using the following command:

```
mouse <- readRDS(url("https://github.com/psoerensen/bgcourse/raw/main/data/mouse.rds"))
```

Question 1: How many observations and which variables do we have in the data set? To get a fast overview of the data set you are working with you can use the `str` function:

Answer:

```
str(mouse)
```

```
## 'data.frame':    1177 obs. of  6 variables:
##  $ sire: Factor w/ 11 levels "25","28","34",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ dam : Factor w/ 55 levels "26","27","29",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ sex : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 2 2 2 1 1 ...
##  $ reps: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 2 2 ...
```

3

```
##  $ Gl  : num  187 136 115 125 112 190 169 159 111 89 ...
##  $ BW  : num  36.6 33.3 42.1 37.1 38.4 ...
```

The two quantitative traits we will be analysing are glucose levels in the blood (Gl) and body weight (BW) measured in the mice at 8 weeks of age. A more detailed view of the two quantitative traits in the `data.frame` is provided by the `summary` function:

```
summary(mouse[,5:6])
```

```
##       Gl              BW
##  Min.   : 65.0   Min.   :23.04
##  1st Qu.:121.0   1st Qu.:34.06
##  Median :139.0   Median :38.32
##  Mean   :144.2   Mean   :38.72
##  3rd Qu.:164.0   3rd Qu.:43.40
##  Max.   :292.0   Max.   :60.28
```

Question 2: What is the mean and variance of body weight and blood glucose levels? Use the `mean` and `var` functions to compute the mean and variance two traits:

Answer:

```
weight <- mouse[,"BW"]
glucose <- mouse[,"Gl"]
mean(weight)
```

```
## [1] 38.72392
```

```
mean(glucose)
```

```
## [1] 144.2234
```
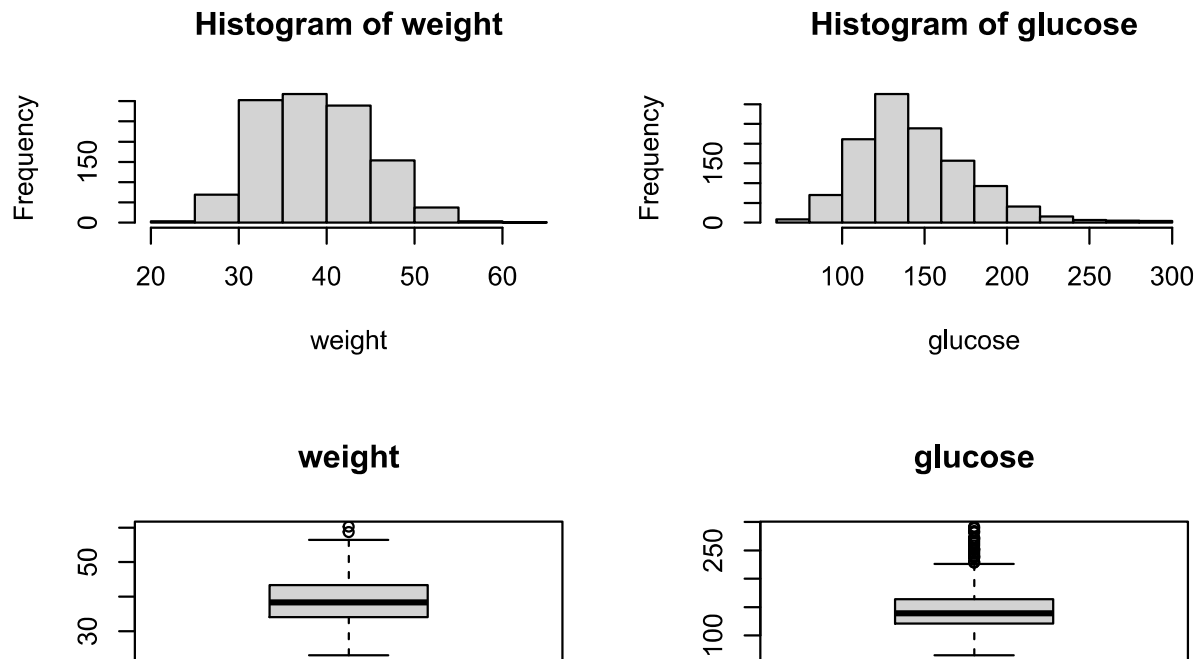
```
var(weight)
```

```
## [1] 37.84458
```

```
var(glucose)
```

```
## [1] 1150.66
```

Question 3: How are the phenotypes of weight and glucose distributed? Use the `histogram` and `boxplot` functions to visualize the distribution the two traits:

Answer:

```
layout(matrix(1:4,ncol=2,byrow=TRUE))
hist(weight)
hist(glucose)
boxplot(weight, main="weight")
boxplot(glucose, main="glucose")
```


Question 4: Are the phenotypes of weight and glucose normally distributed? Use the `qqnorm` function to create a quantile-quantile (QQ) plot of the trait values. Use the qqline function to add a line to a "theoretical," by default normal, quantile-quantile plot:

Answer:

```
layout(matrix(1:2,ncol=2))
qqnorm(weight)
qqline(weight)
qqnorm(glucose)
qqline(glucose)
```


Normal Q-Q Plot   **Normal Q-Q Plot**
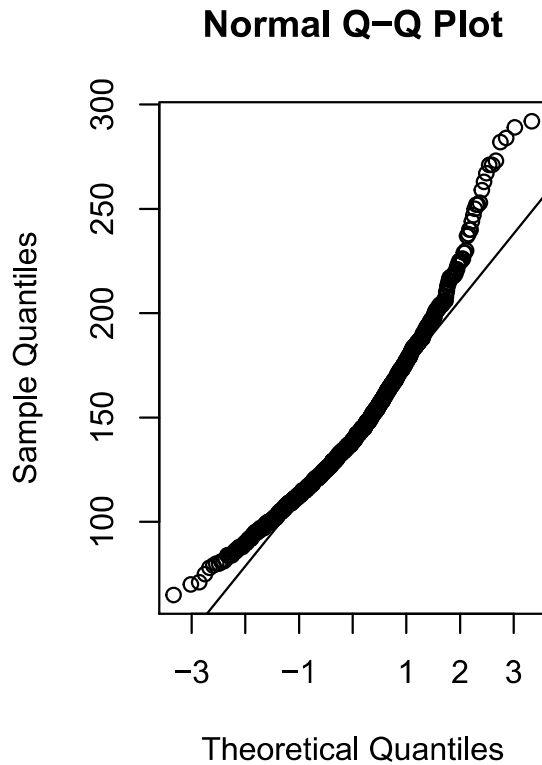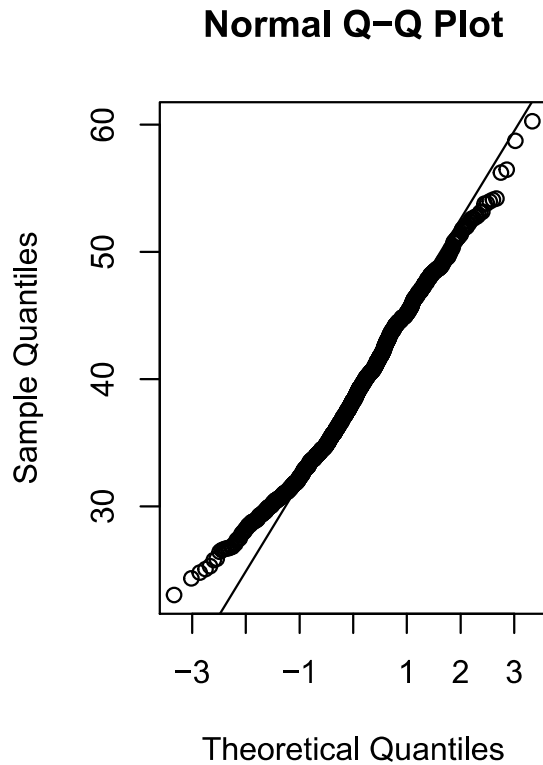
Question 5: Is there a relationship between the phenotypes of weight and glucose? Make a scatter plot of the the 2 traits using the `plot` function. Compute the correlation using the `cor` function and perform a statistical test to assess the significance of correlation between values of weight and glucose using the `cor.test` function:

Answer:

```
plot(weight,glucose)
```
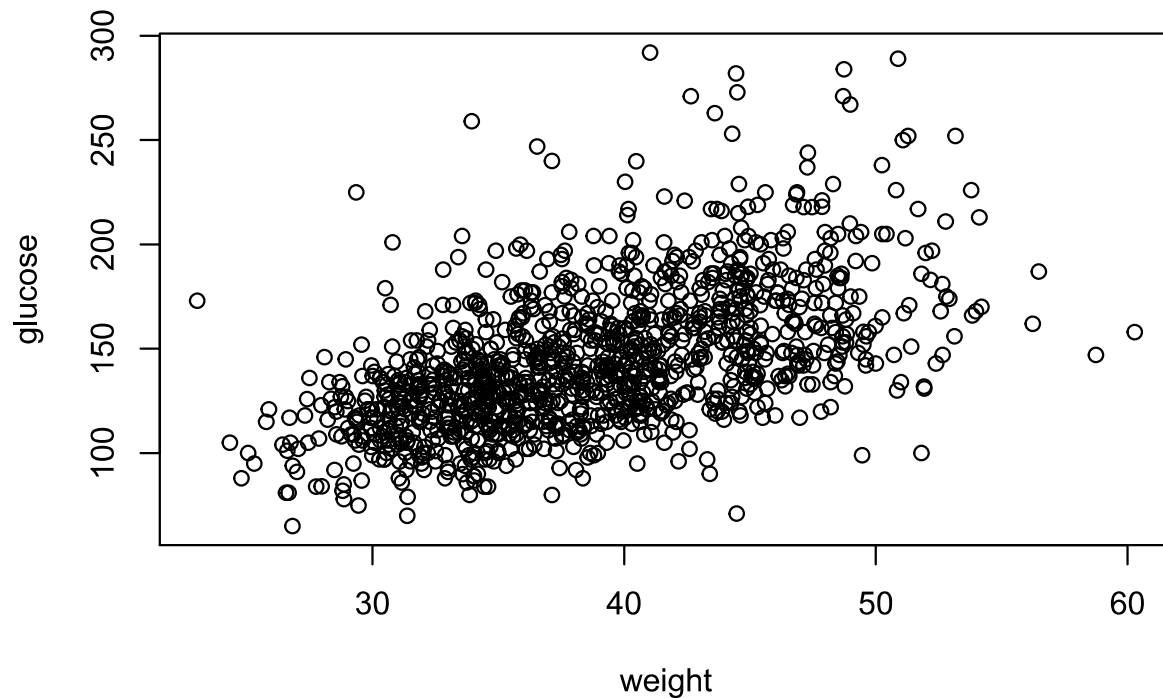
```
cor(weight,glucose)
```

```
## [1] 0.5440533
```

```
cor.test(weight,glucose)
```

```
##
##  Pearson's product-moment correlation
##
## data:  weight and glucose
## t = 22.227, df = 1175, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5025357 0.5830674
## sample estimates:
##       cor
## 0.5440533
```

Let us explore the family structure. Use the `table` function to determine the family size for sires and dams:

```
table(mouse$sire)
```

```
##
##  25  28  34  40  51  63  69  72  78  79  85
## 115 114 107  95 110 103 103 119 119 118  74
```

```
table(mouse$dam)
```

```
##
## 26 27 29 30 31 32 33 35 36 37 38 39 41 42 43 44 45 46 47 48 49 50 52 53 54 55
## 24 16 23 24 24 24 24 24 23 16 16 24 23 16 24 24 16 16 24 16 23 16 16 15 16 24
## 56 57 58 59 60 61 62 64 65 66 67 68 70 71 73 74 75 76 77 80 81 82 83 84 86 87
## 16 23 24 21 24 21 24 24 24 23 22 22 15 19 23 24 23 24 24 24 23 24 24 24 16 23
## 88 89 90
## 24 24 20
```

Question 6: What are the min and max family size? Use the `table` and `min` or `max` functions to determine the min/max family size for sires and dams:

Answer:

```
min(table(mouse$sire))
```

```
## [1] 74
```

```
max(table(mouse$sire))
```

```
## [1] 119
```
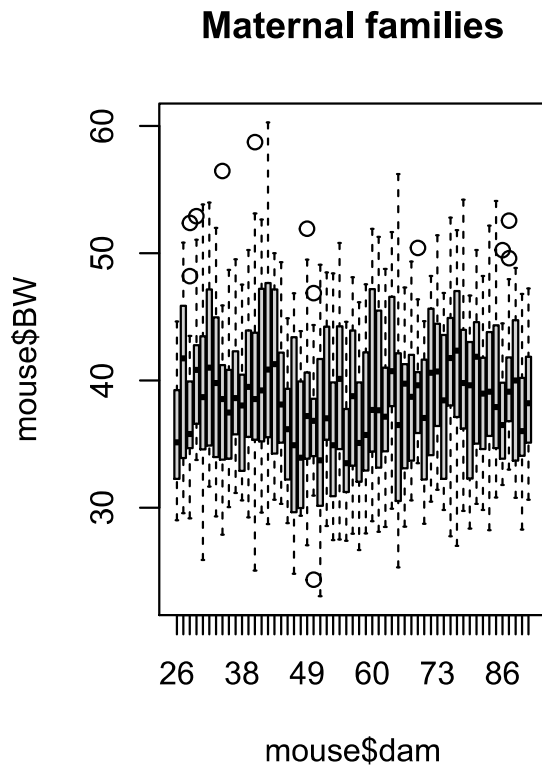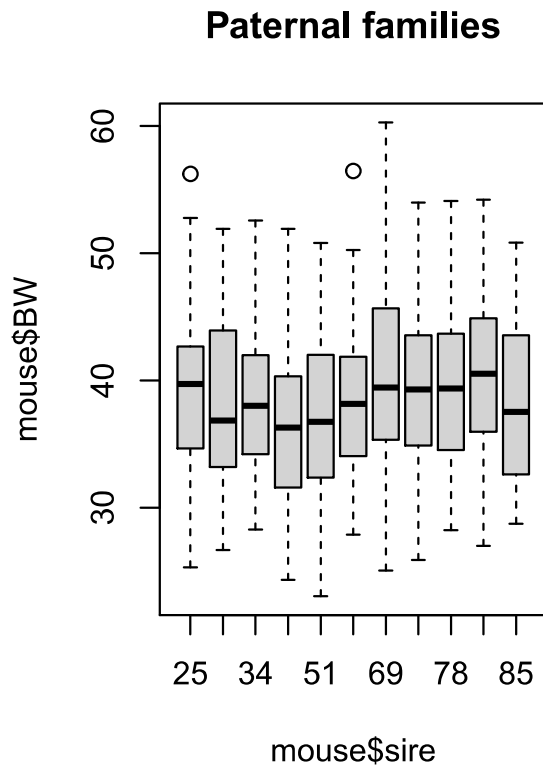
```
min(table(mouse$dam))
```

```
## [1] 15
```

```
max(table(mouse$dam))
```

```
## [1] 24
```

Question 7: Does family influence the traits? Use the `boxplot` function to visualize the potential effect of family on the two traits:

Answer:

```
layout(matrix(1:2,ncol=2))
boxplot(mouse$BW~mouse$sire, main="Paternal families")
boxplot(mouse$BW~mouse$dam, main="Maternal families")
```

Paternal families   **Maternal families**

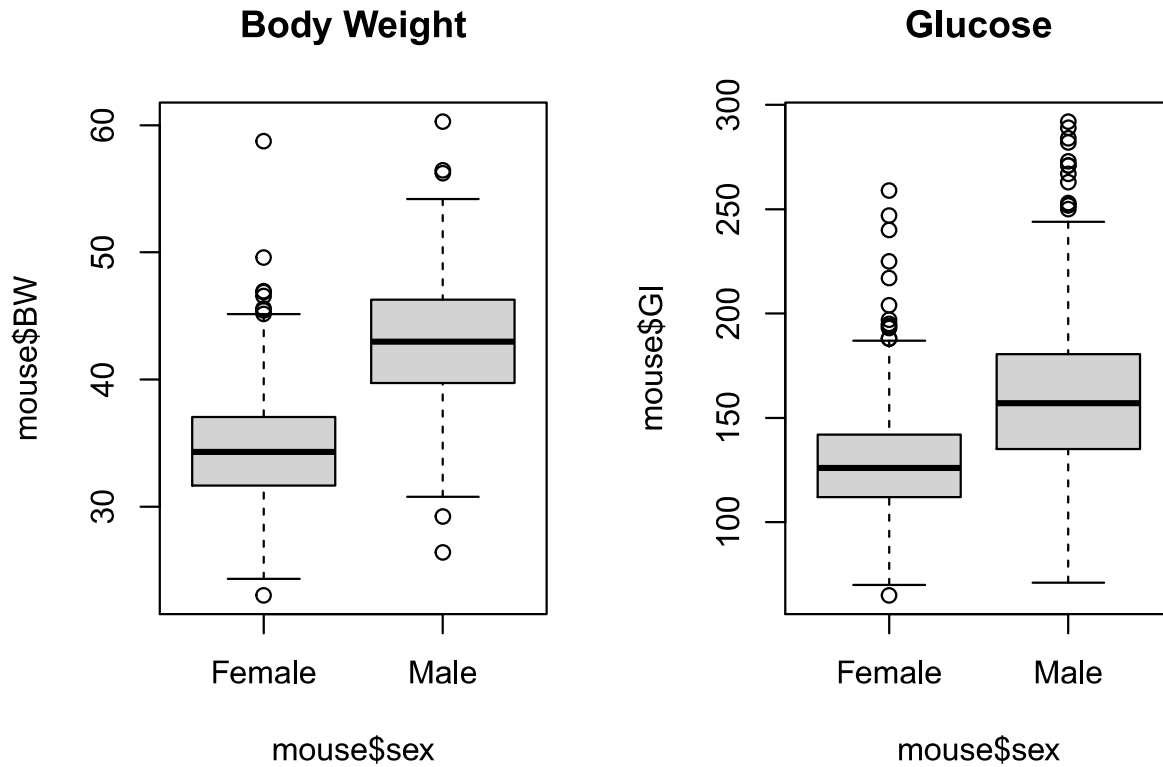Question 8: How many males and females?

Answer:

```
table(mouse$sex)
```

```
##
## Female    Male
##    589     588
```

Question 9: Does gender influence the traits? Use the `boxplot` function to visualize the potential effect of gender on the two traits:

Answer:

```
layout(matrix(1:2,ncol=2))
boxplot(mouse$BW~mouse$sex, main="Body Weight")
boxplot(mouse$Gl~mouse$sex, main="Glucose")
```

Body Weight

Glucose

Question 10: How many observations in each replicate?
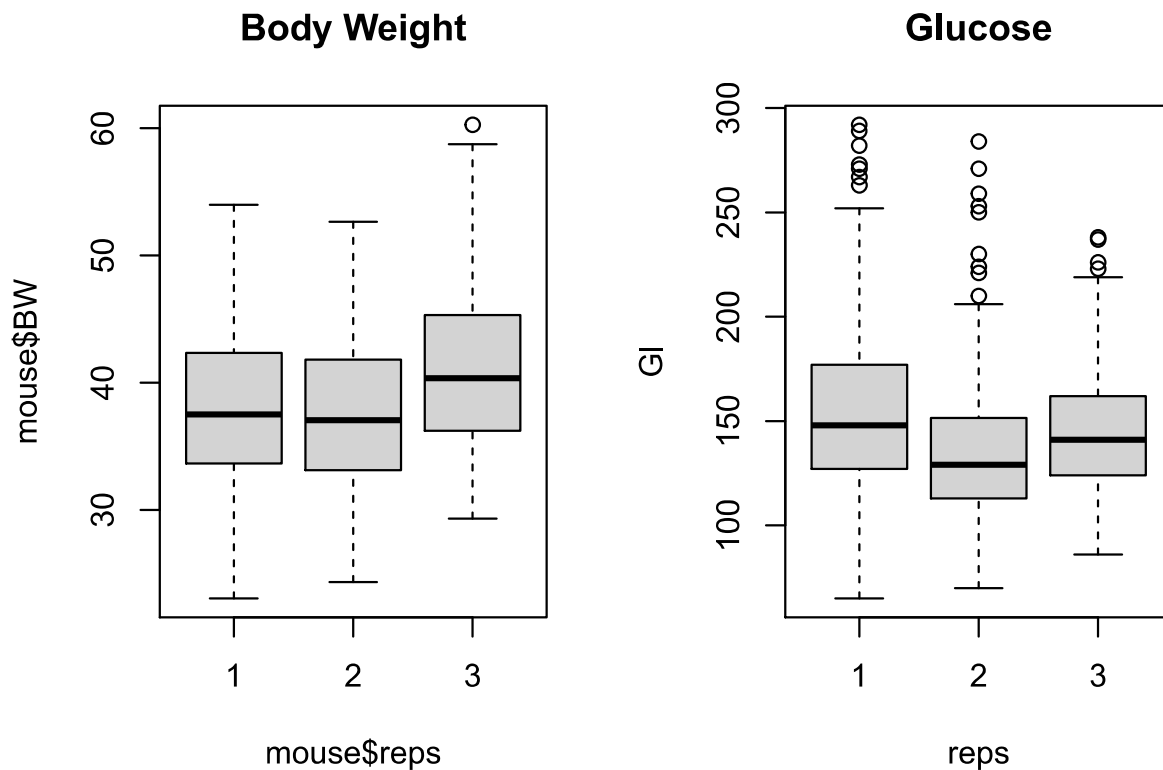
Answer:

```
table(mouse$reps)
```

```
##
##   1   2   3
## 415 427 335
```

Question 11: Does replicate influence the phenotype of weight and glucose? Use the `boxplot` function to visualize the potential effect of replicate on the two traits:

Answer:

```
layout(matrix(1:2,ncol=2))
boxplot(mouse$BW~mouse$reps, main="Body Weight")
boxplot(Gl~reps, main="Glucose", data=mouse)
```


The exploratory data analysis is the process of analyzing and visualizing the data to get a better under-standing of the data. It is not a formal statistical test.

Which factors should we include in the statistical model? To best answer these question we can fit a linear model that include these factors (sire, dam, sex, reps) in the model. This can be done using the `lm` function:

```
fit <- lm(BW~sire+dam+sex+reps, data=mouse)
```

To test the effect of the variables in the model use the `anova` function on the `fit` object from the `lm` function:

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: BW
##              Df  Sum Sq Mean Sq  F value      Pr(>F)
## sire         10  1536.6   153.7   9.2514  7.705e-15 ***
```

```
## dam          44  2020.9    45.9    2.7652 1.238e-08 ***
## sex           1 20637.7 20637.7 1242.5000 < 2.2e-16 ***
## reps          2  1723.6   861.8   51.8858 < 2.2e-16 ***
## Residuals 1119 18586.4    16.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 12: Do genetic factors influence the traits? Look at the output of the `anova` function.

Answer:

Some useful links explaining how to use R for basic concepts in statistics:

http://www.r-tutor.com/elementary-statistics

http://www.r-tutor.com/elementary-statistics/probability-distributions/normal-distribution

http://www.r-tutor.com/elementary-statistics/numerical-measures

http://www.r-tutor.com/elementary-statistics/numerical-measures/mean

http://www.r-tutor.com/elementary-statistics/numerical-measures/variance

http://www.r-tutor.com/elementary-statistics/numerical-measures/standard-deviation

http://www.r-tutor.com/elementary-statistics/numerical-measures/covariance

http://www.r-tutor.com/elementary-statistics/numerical-measures/correlation-coefficient

http://www.r-tutor.com/elementary-statistics/simple-linear-regression

http://www.r-tutor.com/elementary-statistics/multiple-linear-regression

http://www.r-tutor.com/elementary-statistics/analysis-variance

https://antoinesoetewey.shinyapps.io/statistics-202/

Practical 2: Basic Quantitative Genetics

Time schedule of practical session 2:

11:15	Question to lectures, multiple-choice question and follow up on previous multiple choice questions
11:30	Today´s exercise and assignment to groups
12:00	15 minutes break
12:30	Go through excercises using final word
12:50	Repeat multiple-choice questions
13:00	End of practical session 2

Introduction:

In this practical we use R for explorative data analyses of two quantitative traits, body weight and blood glucose levels, observed in the F2 mouse population. We will be characterizing and investigating the potential effects of a single marker locus. This includes computation of allele and genotype frequencies, evaluating different genetic models, and estimation of the breeding values and genetic variances for the single marker locus.

Furthermore you may also want to explore these `shinyapps` that could help understand some of the basic concepts of quantitative genetics:

https://neyhartj.shinyapps.io/qgshiny/

https://shiny.cnsgenomics.com/Falconer2/

Let's continue explore our mouse data

The mouse data set can be loaded using the following command:

```r
mouse <- readRDS(url("https://github.com/psoerensen/bgcourse/raw/main/data/mouseqtl.rds"))
```

Question 1: How many observations and which variables do we have in the data set? To get a fast overview of the data set you are working with you can use the `str` function:

Answer:

```r
str(mouse)
```

```
## 'data.frame':    1177 obs. of  8 variables:
##  $ sire : Factor w/ 11 levels "25","28","34",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ dam  : Factor w/ 55 levels "26","27","29",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ sex  : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 2 2 2 1 1 ...
##  $ reps : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Gl   : num  187 136 115 125 112 190 169 159 111 89 ...
##  $ BW   : num  36.6 33.3 42.1 37.1 38.4 ...
##  $ M227 : Factor w/ 3 levels "AA","AB","BB": 2 1 2 2 2 1 2 1 2 2 ...
##  $ M1139: Factor w/ 3 levels "AA","AB","BB": 3 NA 1 1 1 2 3 3 2 2 ...
```

Question 2: How many observations do the two marker variables have in each genotype class?
Use the `table` function to explore the two marker variables:

Answer:

```
table(mouse$M227)
```

```
##
##  AA  AB  BB
## 493 536 145
```

Question 2: What are the genotype and allele frequencies for M227? Include the allele and
genotype frequencies for M227 in the following table:

Variable	M227
f_{AA}	
f_{AB}	
f_{BB}	
f_A	
f_B	

```
freq_genotypes <- table(mouse$M227)/sum(table(mouse$M227))
fA <- sum(table(mouse$M227)*c(2,1,0))/(sum(table(mouse$M227))*2)
fB <- sum(table(mouse$M227)*c(0,1,2))/(sum(table(mouse$M227))*2)
freq_alleles <- c(fA,fB)
names(freq_alleles) <- c("A","B")
layout(matrix(1:2,nrow=2))
barplot(freq_genotypes, main="Genotype Frequencies")
barplot(freq_alleles, main="Allele Frequencies")
```

Genotype Frequencies
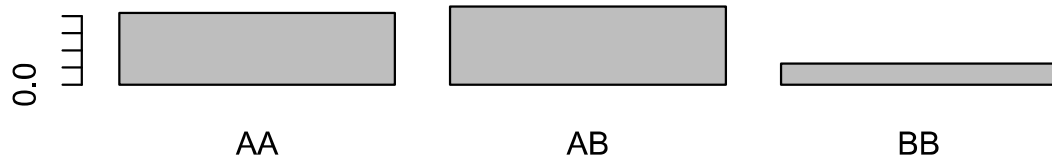
Allele Frequencies
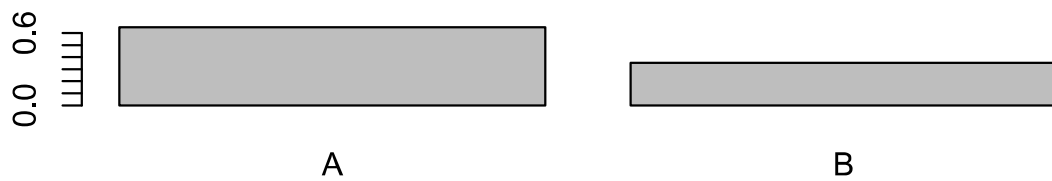

```
freq_genotypes
```

```
##
##        AA        AB        BB
## 0.4199319 0.4565588 0.1235094
```

```
freq_alleles
```

```
##        A        B
## 0.6482112 0.3517888
```

Question 3: Does the marker variable M227 potentially influence body weight and glucose?
Use the `boxplot` function to visualize the potential effect of the marker variable M227 on the two traits:

Answer:

```
layout(matrix(1:2,ncol=2))
boxplot(mouse$BW~mouse$M227, main="M227 - Body Weight")
boxplot(mouse$Gl~mouse$M227, main="M227 - Glucose")
```

M227 − Body Weight

M227 − Glucose
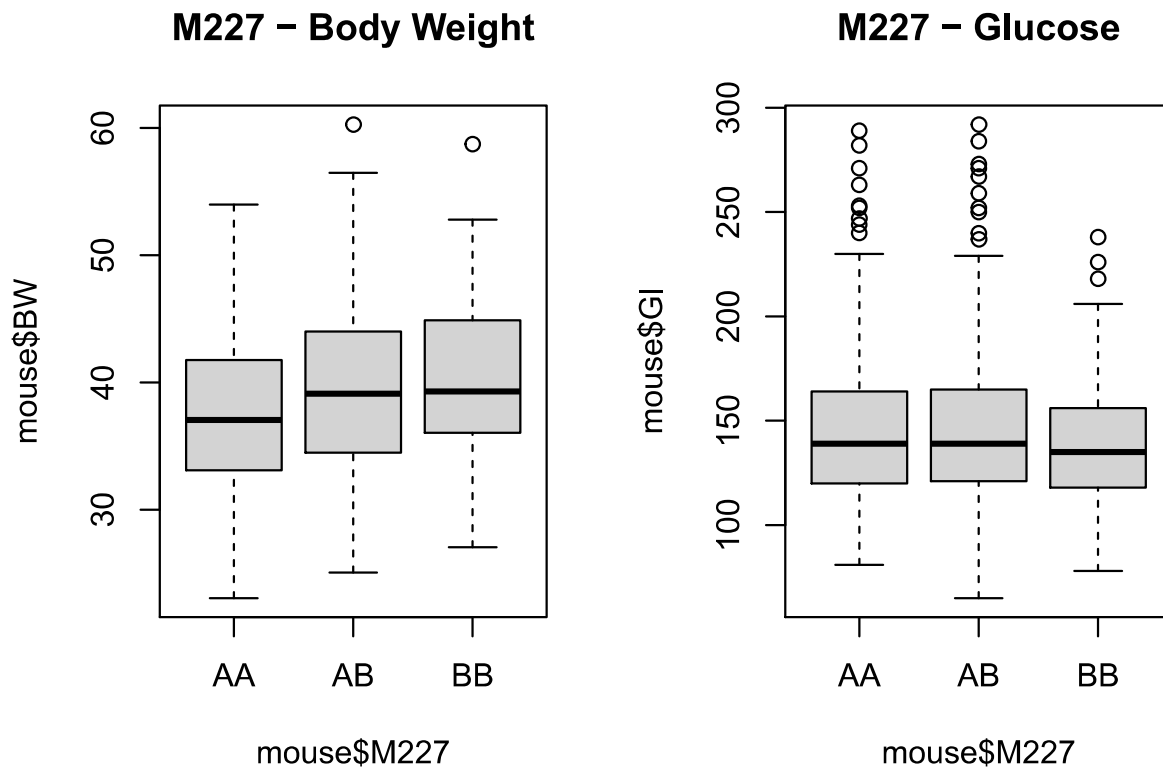
To best answer these question we can fit a linear model that also include the effect of the marker variable in addition to sex and reps. This can be done using the `lm` function:

```
fit <- lm(BW~ sex + reps + M227, data=mouse)
```

To test the effect of the variables in the model use the `anova` function on the `fit` object from the `lm` function:

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: BW
##             Df  Sum Sq Mean Sq  F value     Pr(>F)
## sex          1 20542.9 20542.9 1203.352 < 2.2e-16 ***
## reps         2  2195.9  1097.9   64.315 < 2.2e-16 ***
## M227         2  1660.3   830.1   48.627 < 2.2e-16 ***
## Residuals 1168 19939.3    17.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 4: Based on the linear model results do marker variable M227 influence body weight?

Answer:

The additive effect is modeled by a variable, **add**, with levels that is coded as -1, 0, and 1 (corresponding to -a, 0, a) for the genotypes AA, AB, and BB. The following lines of R code create a the **add** variable, fit the linear model and test the effects:

```
alleles <- c(-1,0,1)
names(alleles) <- c("AA","AB","BB")
mouse$add <- alleles[mouse$M227]
fit <- lm(BW~ sex + reps + add, data=mouse)
summary(fit)
```

```
##
## Call:
## lm(formula = BW ~ sex + reps + add, data = mouse)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -15.9743  -2.6780  -0.0483   2.5625  19.7455
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.3598     0.2396 143.399   <2e-16 ***
## sexMale        8.4133     0.2415  34.838   <2e-16 ***
## reps2         -0.3787     0.2852  -1.328    0.184
## reps3          2.8966     0.3043   9.518   <2e-16 ***
## add            1.7381     0.1790   9.713   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.135 on 1169 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5492, Adjusted R-squared:  0.5477
## F-statistic: 356.1 on 4 and 1169 DF,  p-value: < 2.2e-16
```

The **summary(fit)** command produced

- parameter estimates (or Coefficients) $\widehat{\mu}$ and $\widehat{\beta}$,
- their standard errors (SE) (estimates for square root of the sampling variance of the parameter estimates),
- t-statistic (estimate/SE) and
- P-value under the null hypothesis that the parameter is 0 and errors are uncorrelated and have distribution $N(0, \sigma^2)$.

Under the assumptions of linear model, sampling distribution of t-statistic is t-distribution and hence q% confidence intervals are determined as $\widehat{\beta} \pm a \times \text{SE}$, where a is the q/2% quantile of t-distribution with $n - 2$ degrees of freedom. To get a confidence interval use the **confint** function:

```
confint(fit,parm="add")
```

```
##        2.5 %   97.5 %
## add 1.387014 2.089222
```

The regression coefficient for the variable `add` is 1.74. The coefficient corresponds to the allele substitution effect (α). Previously we have estimated allele and genotype frequencies for M227. The following table summarizes all genotypic values, all breeding values and the dominance deviations.

Genotyp $A_i A_j$	Genotypic value GV_{ij}	Breeding Value BV_{ij}	Dominance Deviation D_{ij}
$A_1 A_1$	a	$2q\alpha$	$-2q^2 d$
$A_1 A_2$	d	$(q-p)\alpha$	$2pqd$
$A_2 A_2$	$-a$	$-2p\alpha$	$-2p^2 d$

Question 5: What are the breeding values for body weight based on the M227 locus?

Answer:

```r
alpha <- -fit$coefficients["add"]
BV_AA <- 2*fA*alpha
BV_AB <- (fA-fB)*alpha
BV_BB <- -2*fA*alpha
BV <- c(BV_AA,BV_AB,BV_BB)
names(BV) <- c("AA","AB","BB")
barplot(BV, main="Breeding Values")
```

Breeding Values
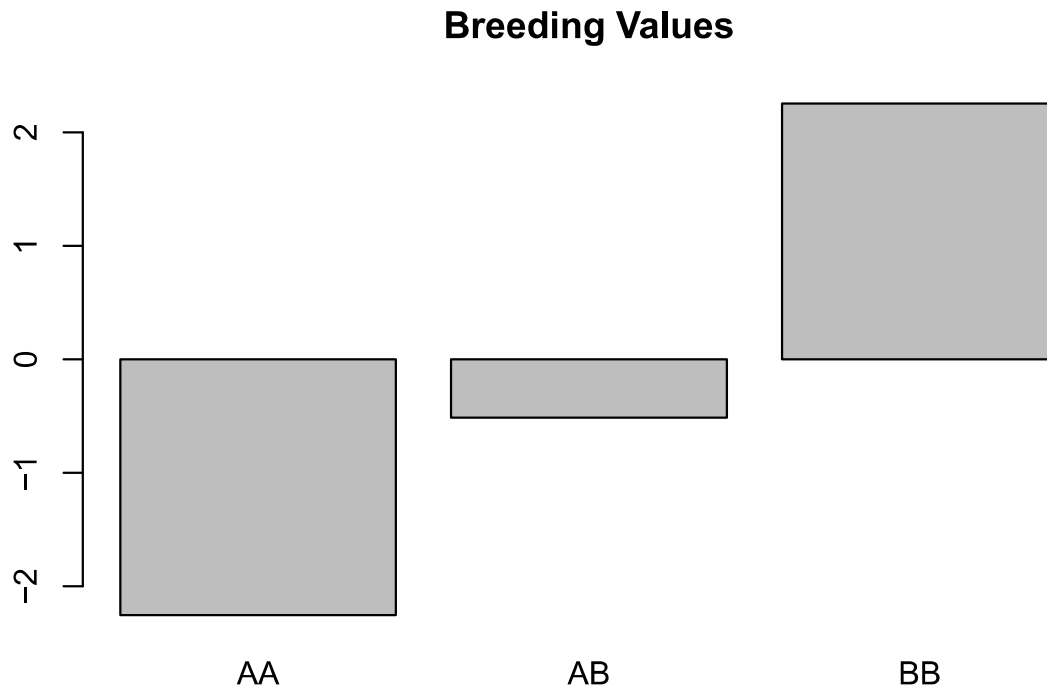
Now we want to compute the genetic variance associated with marker M227. The formula below shows that genetic variance for a single locus model σ_G^2 consists of two components. The first component σ_A^2 is called the **genetic additive variance** and the second component σ_D^2 is termed **dominance variance**. Here σ_A^2 corresponds to the variance of the breeding values. The variance of breeding values is also called the additive genetic variance, because as we have already seen the breeding values are additive in the number of favorable alleles. In populations where there is no additive genetic variance, individuals all have the same breeding value. Therefore, they will produce offspring with the same expected advantage (zero), and selection cannot generate any improvement over generations. Because σ_D^2 corresponds to the variance of the dominance deviation effects it is called dominance variance.
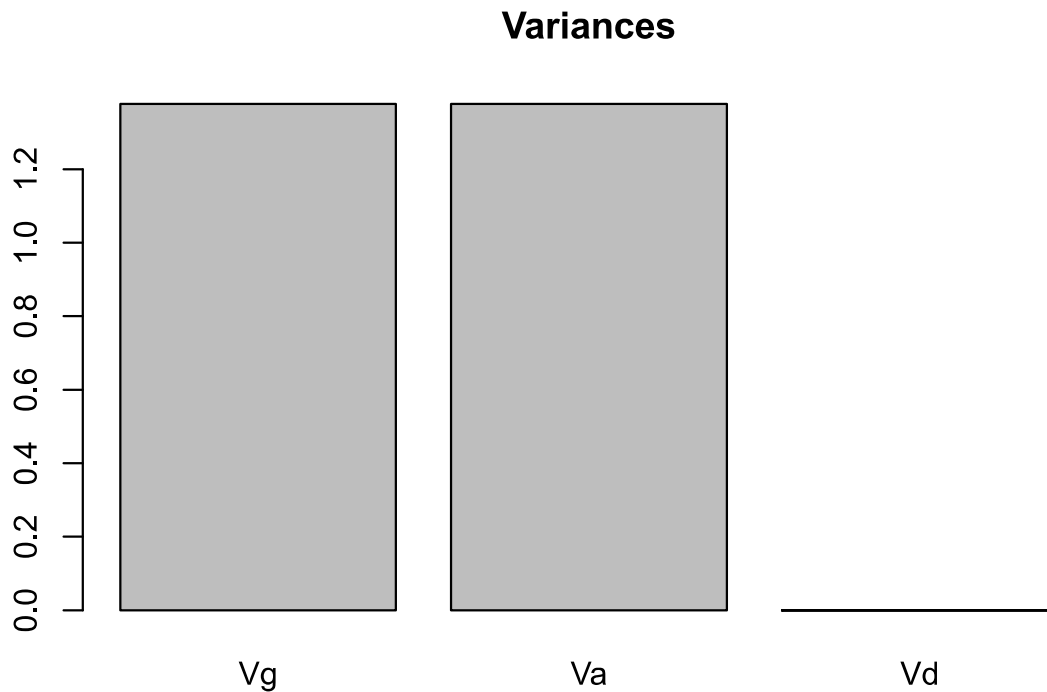
$$\sigma_G^2 = 2pq\alpha^2 + (2pqd)^2$$
$$= \sigma_A^2 + \sigma_D^2$$

Question 6: What is the additive genetic variance associated with M227 for body weight?

Answer:

```
alpha <- fit$coefficients["add"]
d <- 0
Va <- 2*fA*fB*alpha^2
Vd <-  (2*fA*fB*d)^2
```

```
Vg <- Va + Vd
V <- c(Vg,Va,Vd)
names(V) <- c("Vg","Va","Vd")
barplot(V, main="Variances")
```

Variances

Question 7: Should you have considered other factors in the linear model specified above?

Answer:

Now we will fit the full genetic model to locus M227 including both additive and dominance effects. The additive effect is modeled as previously shown by a variable `add` that is coded as -1, 0, and 1 (corresponding to -a, 0, a) for the genotypes AA, AB, and BB. The dominance effect is modeled by a variable `dom` that is coded as 0, 1, and 0 (corresponding to 0,d,0) for the genotypes AA, AB, and BB. The corresponding R code is shown below:

```
alleles <- c(-1,0,1)
names(alleles) <- c("AA","AB","BB")
mouse$add <- alleles[mouse$M227]
mouse$dom <- as.numeric(mouse$add==1)
fit <- lm(BW~sex + reps + add+dom, data=mouse)
summary(fit)
```

```
##
## Call:
## lm(formula = BW ~ sex + reps + add + dom, data = mouse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1773  -2.7642  -0.0437   2.5549  20.1121
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.5549     0.2666 129.635  < 2e-16 ***
## sexMale       8.4130     0.2413  34.863  < 2e-16 ***
## reps2        -0.3706     0.2850  -1.300   0.1937
## reps3         2.9062     0.3041   9.555  < 2e-16 ***
## add           2.0479     0.2580   7.937 4.82e-15 ***
## dom          -0.8811     0.5290  -1.665   0.0961 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.132 on 1168 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5503, Adjusted R-squared:  0.5484
## F-statistic: 285.8 on 5 and 1168 DF,  p-value: < 2.2e-16
```

```
confint(fit,parm="add")
```

```
##        2.5 %   97.5 %
## add 1.541656 2.554078
```

```
confint(fit,parm="dom")
```

```
##         2.5 %    97.5 %
## dom -1.919038 0.1569128
```

The results from the linear model analysis suggest that only the additive genetic effect, add, is significantly different from 0. However in the following exercise we will be using the both the additive effect (add) and dominance effect (dom) estimated for locus M227, and the frequency of the positive allele (p) to explore the effect of changes in allele frequency.

Use the following shinyapp, https://shiny.cnsgenomics.com/Falconer2/, to understand the relationship between allelic substitution effect (α) and additive gene action (a), dominance gene action (d), and allele frequency (p).

Question 8: Use the estimated gene actions (Question 7) and the estimated allele frequency (Question 2) to obtain the predicted allelic substitution effect? Use rounded values if necessary.

Answer:

Question 9: Does the value of α match the estimate of the (marginal) additive effect from Question 5?

Answer:

Question 10: How does α depend on a larger dominance gene action d (e.g., maximum value, 10)?

Answer:

Question 11: How does α depend on a different allele frequency p (e.g., 0.95)?

Answer:

Question 12: Under that new value of p, how does α depend on d (e.g., from the initial value of d to the maximum value)?

Answer: