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:

```
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:

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:

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	

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:

To best answer these question we can fit a linear model that also include the effect of the marker variable. This can be done using the `lm` function:

```
fit <- lm(BW~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)
## M227         2   1220  610.10  16.569 8.019e-08 ***
## Residuals 1171  43118   36.82
## ---
## 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~add, data=mouse)
summary(fit)
```

```
##
## Call:
## lm(formula = BW ~ add, data = mouse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.643  -0.426   4.516  21.112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.1683     0.1934 202.487  < 2e-16 ***
## add           1.4845     0.2624   5.657 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.068 on 1172 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.02658,    Adjusted R-squared:  0.02575
## F-statistic: 32.01 on 1 and 1172 DF,  p-value: 1.929e-08
```

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 0.9696885 1.999335
```

The regression coefficient for the variable `add` is 1.48. 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_iA_j	Genotypic value GV_{ij}	Breeding Value BV_{ij}	Dominance Deviation D_{ij}
A_1A_1	a	$2q\alpha$	$-2q^2d$
A_1A_2	d	$(q-p)\alpha$	$2pqd$
A_2A_2	$-a$	$-2p\alpha$	$-2p^2d$

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

Answer:

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^2_G consists of two components. The first component σ^2_A is called the **genetic additive variance** and the second component σ^2_D is termed **dominance variance**. Here σ^2_A 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^2_D corresponds to the variance of the dominance deviation effects it is called dominance variance.

$$\begin{aligned}\sigma^2_G &= 2pq\alpha^2 + (2pqd)^2 \\ &= \sigma^2_A + \sigma^2_D\end{aligned}$$

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

Answer:

Question 7: Should 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~add+dom, data=mouse)
summary(fit)
```

```
##
## Call:
## lm(formula = BW ~ add + dom, data = mouse)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -14.5417  -4.6222  -0.4561   4.5072  20.9239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.3561     0.2621 150.156  < 2e-16 ***
## add           1.7744     0.3787   4.686 3.11e-06 ***
## dom          -0.8249     0.7768  -1.062    0.289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.068 on 1171 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.02752,    Adjusted R-squared:  0.02586
## F-statistic: 16.57 on 2 and 1171 DF,  p-value: 8.019e-08
```

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

```
##        2.5 %   97.5 %
## add 1.031495 2.517364
```

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

```
##         2.5 %    97.5 %
## dom -2.348952 0.6992021
```

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 allele frequency of the positive allele (B) 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 6) 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: