Class 12 Homework

Zainab Fatima (PID: A16880407)

2025-02-15

Table of contents

Determining Sample Size and Median Expression Levels

> Question 13. Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

Answer: The sample size for each phenotype is 108 samples for A/A, 233 for A/G, and 121 for G/G. The median expression levels for each genotype is 31.25 for A/A, 25.06 for A/G, and 20.07 for G/G. The code is listed below.

How many samples do we have?

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
head(expr)
```

```
  sample geno      exp
1 HG00367  A/G 28.96038
2 NA20768  A/G 20.24449
3 HG00361  A/A 31.32628
4 HG00135  A/A 34.11169
5 NA18870  G/G 18.25141
6 NA11993  A/A 32.89721
```
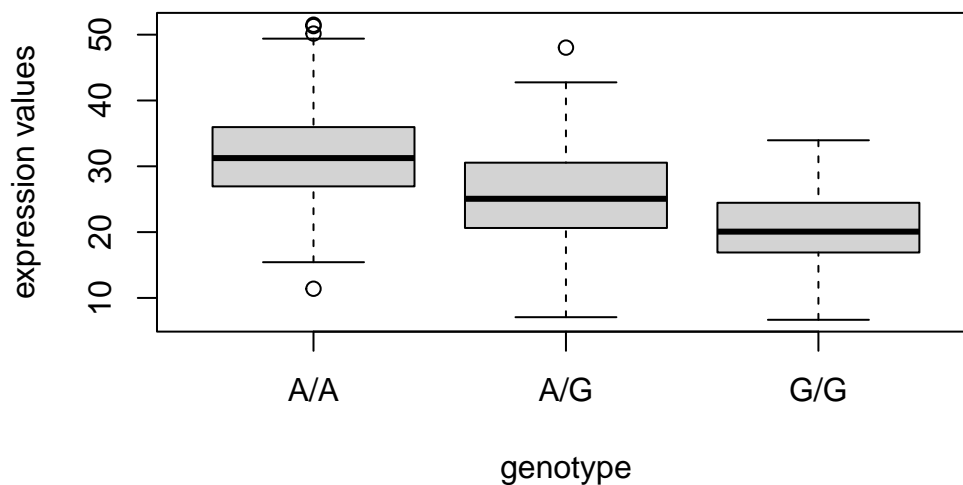
```
nrow(expr)
```

```
[1] 462
```

How many of each sample do we have?

```
table(expr$geno)
```

```
A/A A/G G/G
108 233 121
```

To find the median of each expression level, let's make a boxplot.

```
boxplot(exp ~ geno, data = expr, xlab = "genotype", ylab = "expression values")
```


The function `tapply()` is used to apply a function over subsets of a vector, we can use this function to find the median value of each genotype.

```
tapply(expr$exp, expr$geno, median)
```

```
     A/A      A/G      G/G
31.24847 25.06486 20.07363
```

The median values of each genotype are 31.24 for A/A, 25.06 for A/G, and 20.07 for G/G.

Generating Boxplot

Q14. Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

Boxplot drawn below using ggplot. I can infer that expression value between A/A and G/G is different and that A/A SNP promotes more expression compared to G/G. This is because of the location of the "notches" in the boxplots and the medians of the graphs show that the level of A/A is higher than G/G. The SNP effects the expression of ORMDL3 based on the different levels of each group (A/A is highest expression level, A/G is medium, and G/G is the lowest expression level).

```
library(ggplot2)
graph_of_exp <- ggplot(expr) + aes(geno, exp, fill = geno) +
  geom_boxplot(notch = TRUE) +
  labs(title="Gene Expression by Genotype",
       x ="Genotype", y = "Expression Level", fill = "Genotype")

graph_of_exp
```