Lab 10: Comparing Multiple Means—One-way and two-way ANOVA tests

Introduction and Objective:

We learned how to calculate P values to help us making crisp decisions by using t-test analyses if we want to compare the means of two groups of values. If we want to compare the means of values from more than 2 groups, as we discussed in the lecture, we need to perform ANOVA tests.

In today's lab, we will learn how to use R to perform both one-way and two-way ANOVA tests. You will see it is much easier to do the F tests and calculate the P values using R than doing the step-by-step calculations by hand as we did in the lecture.

The following files you should have received and are needed for today's class:

1) DAC_VC.csv
2) PlaceboAnd3Drugs.csv
3) DACVC_Time.csv
4) MiceBodyWeight.csv

All the R command lines are in **bold**; all the notes are following a #; all the R results directly follow the R codes/command lines and are not in bold or following a #.

```
setwd("C:/R")
```

I. One-way ANOVA test on 3 groups:

```
#We use the example that has been discussed in the lecture:
# A colon cancer cell line was treated with decitabine (DAC) and vitamin C (VC). The concentration of DAC was always 1 nM. However, there were three different concentrations of VC. Each treatment had 6 biological replicates. Reactivation of an epigenetically silenced gene was measured by qPCR. Delta CT was reported in the table that summarized the results.
```

DAC_VC1	DAC_VC2	DAC_VC3
4	6	11

3	8	12
4	11	13
5	9	7
6	8	8
8	12	9

Challenge Question 1: What could be your null hypothesis? (The answer is at the end of the notes.)

Challenge Question 2: What could be the alpha? (The answer is at the end of the notes.)

# Before we can perform the ANOVA test, we have to transform our data.

# If you want to change the original table to one that looks like the following, what will you do?

GeneExpression	Treatments
4	DAC_VC1
3	DAC_VC1
4	DAC_VC1
5	DAC_VC1
6	DAC_VC1
8	DAC_VC1
6	DAC_VC2
8	DAC_VC2
11	DAC_VC2
9	DAC_VC2
8	DAC_VC2
12	DAC_VC2
11	DAC_VC3
12	DAC_VC3
13	DAC_VC3
7	DAC_VC3
8	DAC_VC3
9	DAC_VC3

# You can write the following codes:

```
a<-read.csv("DAC_VC.csv")
View(a)
dim(a)
```

```
[1] 6 3
DAC_VC1<-a[,1]
DAC_VC1
[1] 4 3 4 5 6 8
DAC_VC2<-a[,2]
DAC_VC2
[1]  6  8 11  9  8 12
DAC_VC3<-a[,3]
DAC_VC3
[1] 11 12 13  7  8  9
GeneExpression<-c(DAC_VC1,DAC_VC2,DAC_VC3)
GeneExpression
 [1]  4  3  4  5  6  8  6  8 11  9  8 12 11 12 13  7  8  9
Treatments<-c(rep("DAC_VC1",6),rep("DAC_VC2",6),rep("DAC_VC3",6))
    # The rep() is a very useful function. It makes “repeats” of whatever you write in the brackets and right before the comma, and repeat it for the number of times as you indicate right after the comma.

    # Then, we use our old friend “data.frame”:

DV<-data.frame(GeneExpression,Treatments)
DV
   GeneExpression Treatments
1               4    DAC_VC1
2               3    DAC_VC1
3               4    DAC_VC1
4               5    DAC_VC1
5               6    DAC_VC1
6               8    DAC_VC1
7               6    DAC_VC2
8               8    DAC_VC2
9              11    DAC_VC2
10              9    DAC_VC2
11              8    DAC_VC2
```

```
12          12   DAC_VC2
13          11   DAC_VC3
14          12   DAC_VC3
15          13   DAC_VC3
16           7   DAC_VC3
17           8   DAC_VC3
18           9   DAC_VC3
```

# If you want, you can create a CSV file. And if you want to remove the row numbers given by R, you can write the following command line:

```
write.csv(DV,"DV.csv", row.names=FALSE)
```

# As we learned in Lab5 if you don't add "row.names=FALSE" or you add "row.names=TRUE", R will automatically add numerical row numbers at the beginning of each row.

# Next, we can perform a one-way ANOVA:

```
DVAOV<-aov(GeneExpression~Treatments, data=DV)
```

# The R built-in function, aov(), allows us to perform the ANOVA tests. Before the tilde (~) is the "response", after the tilde is the "factor" that you are interested in. Since the treatment (the different concentrations of Vitamin C) is the factor and the only factor we are interested in, the object "Treatments" is put after the tilde.

# If you type:

```
DVAOV
```

# You will get:

```
Call:
   aov(formula = GeneExpression ~ Treatments, data = DV)

Terms:
                Treatments Residuals
Sum of Squares          84        68
Deg. of Freedom          2        15

Residual standard error: 2.129163
Estimated effects may be unbalanced
```

# Besides the degrees of freedom, the sum of squared differences, if you want to see the mean squares, the F-statistics/F-ratios, and the P values, you have to use the function summary():

summary(DVAOV)

```
           Df Sum Sq Mean Sq F value Pr(>F)
Treatments   2     84   42.00   9.265 0.0024 **
Residuals   15     68    4.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Challenge Question 3: What is your conclusion from the one-way ANOVA test? (The answer is at the end of the notes.)

II. ANOVA test on more than 3 groups with different numbers of observations:

# The example we tested above only had 3 groups. As we discussed in the lecture, the ANOVA test can be performed on data that have more than 3 groups.

# In addition, since ANOVA does not require all samples have the same number of observations, we can perform an ANOVA test on data as in the following example:

# A pharmaceutical company tested 3 drugs together with a placebo to see whether the putative drugs have an anti-cancer function. The experiments were done in the same kinds of mice. Before the study, the mice had tumor xenografts of the same size. After the treatments, the tumor size (in mm^3) was reported in the following table:

Placebo	DrugA	DrugB	DrugC
620	321	568	623
733	367	562	489
665	310	658	456
692	289	632	398
638	297	523	527
712	378	489	556
701	432	386	423
682	356	612	439
633	286		412
678			432

```
# We can perform a one-way ANOVA. After you proposed your H0 and set the alpha, we can write:
b<-read.csv("PlaceboAnd3Drugs.csv")
View(b)
dim(b)
[1] 10  4
b
   Placebo DrugA DrugB DrugC
1      620   321   568   623
2      733   367   562   489
3      665   310   658   456
4      692   289   632   398
5      638   297   523   527
6      712   378   489   556
7      701   432   386   423
8      682   356   612   439
9      633   286    NA   412
10     678    NA    NA   432

# You may notice that there are 3 “NA”s.
# “NA” (not available) is for missing data (data that is not available). It is used for both character and numeric data.
# You don’t have to remove them. It won’t affect your ANOVA test (you will see this later).

Placebo<-b[,1]
Placebo
 [1] 620 733 665 692 638 712 701 682 633 678
DrugA<-b[,2]
DrugA
 [1] 321 367 310 289 297 378 432 356 286  NA
DrugB<-b[,3]
DrugB
 [1] 568 562 658 632 523 489 386 612  NA  NA
```

```
DrugC<-b[,4]
DrugC
 [1] 623 489 456 398 527 556 423 439 412 432
TumorSize<-c(Placebo,DrugA,DrugB,DrugC)
TumorSize
 [1] 620 733 665 692 638 712 701 682 633 678 321 367 310 289 297 378 432 356 286
[20]  NA 568 562 658 632 523 489 386 612  NA  NA 623 489 456 398 527 556 423 439
[39] 412 432
Treatments<-c(rep("Placebo",10),rep("DrugA",10),rep("DrugB",10),rep("DrugC",10))
Treatments
 [1] "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo"
 [8] "Placebo" "Placebo" "Placebo" "DrugA"   "DrugA"   "DrugA"   "DrugA"  
[15] "DrugA"   "DrugA"   "DrugA"   "DrugA"   "DrugA"   "DrugA"   "DrugB"  
[22] "DrugB"   "DrugB"   "DrugB"   "DrugB"   "DrugB"   "DrugB"   "DrugB"  
[29] "DrugB"   "DrugB"   "DrugC"   "DrugC"   "DrugC"   "DrugC"   "DrugC"  
[36] "DrugC"   "DrugC"   "DrugC"   "DrugC"   "DrugC"  

PD<-data.frame(TumorSize,Treatments)
PD
   TumorSize Treatments
1        620    Placebo
2        733    Placebo
3        665    Placebo
4        692    Placebo
5        638    Placebo
6        712    Placebo
7        701    Placebo
8        682    Placebo
9        633    Placebo
10       678    Placebo
11       321      DrugA
12       367      DrugA
13       310      DrugA
14       289      DrugA
15       297      DrugA
16       378      DrugA
17       432      DrugA
18       356      DrugA
```

```
19      286     DrugA
20       NA      DrugA
21      568     DrugB
22      562     DrugB
23      658     DrugB
24      632     DrugB
25      523     DrugB
26      489     DrugB
27      386     DrugB
28      612     DrugB
29       NA      DrugB
30       NA      DrugB
31      623     DrugC
32      489     DrugC
33      456     DrugC
34      398     DrugC
35      527     DrugC
36      556     DrugC
37      423     DrugC
38      439     DrugC
39      412     DrugC
40      432     DrugC
```

PDAOV<-aov(TumorSize~Treatments,data=PD)

PDAOV

```
Call:
   aov(formula = TumorSize ~ Treatments, data = PD)

Terms:
                Treatments Residuals
Sum of Squares    568812.6  132872.4
Deg. of Freedom          3        33

Residual standard error: 63.45421
Estimated effects may be unbalanced
3 observations deleted due to missingness
```

summary(PDAOV)

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Treatments   3 568813  189604   47.09 5.05e-12 ***
```

```
Residuals   33 132872    4026
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
3 observations deleted due to missingness
```

# You can see for the “Residuals” (intra-groups), the df is 33. It tells you that R did not count the three NAs when it did the test. Also, at the bottom, it says “3 observations deleted due to missingness”. So, if you have uneven numbers of observations for different groups, you can still perform the ANOVA test, and don’t have to remove the “NA”s.

III. Two-way ANOVA

# Let’s use the same example we used in a lecture to perform a Two-way ANOVA test using R:

# A colon cancer cell line was treated with decitabine (DAC) and vitamin C (VC). The concentration of DAC was always 1 nM. There were three different concentrations of VC.

# In addition, we would like to know whether to treat the cells for a longer time would have more gene expression.

# So, we treated the cells for two different lengths of time: 12hours and 24 hours.

# Each treatment had 4 biological replicates. Reactivation of an epigenetically silenced gene was measured by qPCR. The delta CT was reported in the table summarizing the results.

	DAC_VC1	DAC_VC2	DAC_VC3
12h	4	7	10
12h	5	9	12
12h	6	8	11
12h	5	12	9
24h	6	13	12
24h	6	15	13
24h	4	12	10
24h	4	12	13

# First, we need to transform the data into the following format:

TimeLength	Treatments	GeneExpression
12h	DAC_VC1	4

12h	DAC_VC1	5
12h	DAC_VC1	6
12h	DAC_VC1	5
24h	DAC_VC1	6
24h	DAC_VC1	6
24h	DAC_VC1	4
24h	DAC_VC1	4
12h	DAC_VC2	7
12h	DAC_VC2	9
12h	DAC_VC2	8
12h	DAC_VC2	12
24h	DAC_VC2	13
24h	DAC_VC2	15
24h	DAC_VC2	12
24h	DAC_VC2	12
12h	DAC_VC3	10
12h	DAC_VC3	12
12h	DAC_VC3	11
12h	DAC_VC3	9
24h	DAC_VC3	12
24h	DAC_VC3	13
24h	DAC_VC3	10
24h	DAC_VC3	13

```
c<-read.csv("DACVC_Time.csv")
View(c)
dim(c)
[1] 8 4

DAC_VC1<-c[1:8,2]
DAC_VC1
[1] 4 5 6 5 6 6 4 4


DAC_VC2<-c[1:8,3]
DAC_VC2
[1]  7  9  8 12 13 15 12 12


DAC_VC3<-c[1:8,4]
```

```
DAC_VC3
[1] 10 12 11  9 12 13 10 13
GeneExpression<-c(DAC_VC1,DAC_VC2,DAC_VC3)
GeneExpression
 [1]  4  5  6  5  6  6  4  4  7  9  8 12 13 15 12 12 10 12 11  9 12 13 10 13


TimeLength<-
c(rep("12h",4),rep("24h",4),rep("12h",4),rep("24h",4),rep("12h",4),rep("24h",4))
TimeLength
 [1] "12h" "12h" "12h" "12h" "24h" "24h" "24h" "24h" "12h" "12h" "12h" "12h"
[13] "24h" "24h" "24h" "24h" "12h" "12h" "12h" "12h" "24h" "24h" "24h" "24h"

Treatments<-c(rep("DAC_VC1",8),rep("DAC_VC2",8),rep("DAC_VC3",8))
Treatments
 [1] "DAC_VC1" "DAC_VC1" "DAC_VC1" "DAC_VC1" "DAC_VC1" "DAC_VC1"
          "DAC_VC1"
 [8] "DAC_VC1" "DAC_VC2" "DAC_VC2" "DAC_VC2" "DAC_VC2" "DAC_VC2"
"DAC_VC2"
[15] "DAC_VC2" "DAC_VC2" "DAC_VC3" "DAC_VC3" "DAC_VC3" "DAC_VC3"
"DAC_VC3"
[22] "DAC_VC3" "DAC_VC3" "DAC_VC3"

DACVCTime<-data.frame(TimeLength,Treatments,GeneExpression)
DACVCTime
   TimeLength Treatments GeneExpression
1        12h    DAC_VC1              4
2        12h    DAC_VC1              5
3        12h    DAC_VC1              6
4        12h    DAC_VC1              5
5        24h    DAC_VC1              6
6        24h    DAC_VC1              6
7        24h    DAC_VC1              4
8        24h    DAC_VC1              4
9        12h    DAC_VC2              7
```

```
10    12h  DAC_VC2       9
11    12h  DAC_VC2       8
12    12h  DAC_VC2      12
13    24h  DAC_VC2      13
14    24h  DAC_VC2      15
15    24h  DAC_VC2      12
16    24h  DAC_VC2      12
17    12h  DAC_VC3      10
18    12h  DAC_VC3      12
19    12h  DAC_VC3      11
20    12h  DAC_VC3       9
21    24h  DAC_VC3      12
22    24h  DAC_VC3      13
23    24h  DAC_VC3      10
24    24h  DAC_VC3      13
```

```
write.csv(DACVCTime,"DACVCTime.csv", row.names=FALSE)
aov2<-aov(GeneExpression~TimeLength*Treatments,data=DACVCTime)
```

```
    # There are 2 factors (the treatment and the length of treatment). The “*” is used to telling R
that we are interested in both factors.
```

```
aov2
```

```
Call:
   aov(formula = GeneExpression ~ TimeLength * Treatments, data = DACVCTime)

Terms:
                TimeLength Treatments TimeLength:Treatments Residuals
Sum of Squares    20.16667  200.33333              16.33333  37.00000
Deg. of Freedom          1          2                     2        18

Residual standard error: 1.433721
Estimated effects may be unbalanced
```

summary(aov2)

```
                      Df Sum Sq Mean Sq F value   Pr(>F)    
TimeLength             1  20.17   20.17   9.811  0.00576 ** 
Treatments             2 200.33  100.17  48.730 5.44e-08 ***
TimeLength:Treatments  2  16.33    8.17   3.973  0.03722 *  
Residuals             18  37.00    2.06                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# As you can see from the results, both the differences between the delta CTs from cells received different treatment and different lengths of treatment are statistically significant.

Answers to the Challenge Questions:

Challenge Question 1:

H_0: There is no difference between the average delta CTs from cells that received different treatments. Any observed differences would result from coincidence, sampling, or experimental errors.

Challenge Question 2: Set the alpha at 0.05. This is the significance level traditionally being used.

Challenge Question 3: Since the P-value is less than the alpha, the null hypothesis is rejected. Thus, the difference between the average delta CTs from cells that received different treatments is statistically significant.

Groupwork Assignment 8 (Part 1):

You are studying the relationships between diets and body weight, as well as sex. You fed both male and female mice with 4 different diets for 2 months. These mice had the same genetic background, at the same age, and were housed in the same room, had the same body weight before the study. You measured their body weight after 2 months, recorded the data (in gram) in the file “MiceBodyWeight.csv”, which looks like:

	DietA	DietB	DietC	DietD
Male	20	21	25.5	23
Male	21	21	24	24.5
Male	23	22	25	24
Female	19	20	22	25
Female	19.5	19.8	23	25.5
Female	19	20.5	22.5	26

If you want to find out whether gender and diets have an effect on the body weight, and at this stage of the study you might not be interested in whether sex interacts with diet:

1) What kind of statistical test would you perform? (1pt)

2) What are your null hypothesis/hypotheses? (1pt)

3) Show your R commands and the results. (7pts)

4) What are your conclusions? (1pt)