Applied Psycholinguistics

Analysis of Variance

Hyunah Ahn

November 16, 2020

Since we started discussing inferential statistics, we have been focusing on comparing means, either comparing a group mean to a population mean or comparing the means of two groups. But we haven't learned what to do when we have to analyze statistics of more than two groups. Here comes ANOVA, a.k.a. the Analysis of Variance, to our rescue. First things fist, we will learn about One-way ANOVA.

1 One-way ANOVA

Remember how we calculated variance? In calculating variance the sum of squares was a very important concept. We were interested in subtracting each observed data point from the group mean so as to calculate the extent to which data were scattered. ANOVA deals with the 'variance' explained by factors that you are interested in and that not explained by any known factors. Let's first load and plot data.

```
oneway ← read.csv("C:/Users/hyuna/OneDrive/Documents/01SNU/03GraduateSeminar/crawley/data/
    oneway.csv")

attach(oneway)

oneway
```

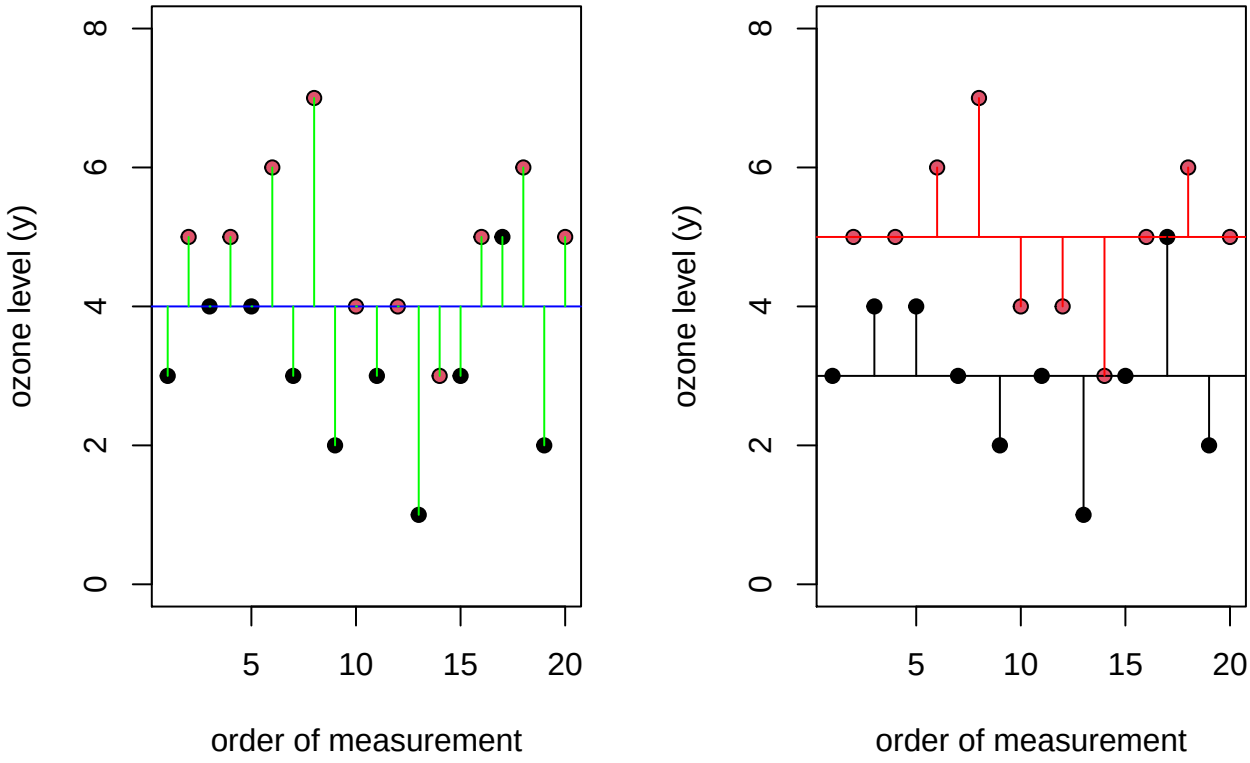
```
   ozone garden
1      3      A
2      5      B
3      4      A
4      5      B
5      4      A
6      6      B
7      3      A
8      7      B
9      2      A
10     4      B
11     3      A
12     4      B
13     1      A
14     3      B
15     3      A
16     5      B
17     5      A
18     6      B
19     2      A
20     5      B
```

As you can see, I just loaded the data file named 'oneway.csv' and attached it. The data has two columns, 'ozone' and 'garden.' On the leftmost side of the table is the row number. You can easily see that there are twenty data points in total. What you can also see is that the ozone measurements were made for gardens A and B. Instead of plotting y in terms of x, for now, we will plot the ozone measurements in the order measured.

```
par(mfrow=c(1,2))
plot(1:20, ozone, ylim = c(0, 8), ylab = "ozone level (y)", xlab = "order of measurement", pch=21,
     bg=as.factor(garden))
abline(h=mean(ozone), col="blue")
for(i in 1:20) lines(c(i,i), c(mean(ozone), ozone[i]), col="green")
```

```
plot(1:20, ozone, ylim = c(0, 8), ylab = "ozone level (y)", xlab = "order of measurement", pch=21,
     bg=as.factor(garden))
abline(h=mean(ozone[garden=="A"]))
abline(h=mean(ozone[garden=="B"]), col="red")

index←1:length(ozone)
for(i in 1:length(index)){
  if (garden[i]=="A")
    lines(c(index[i], index[i]), c(mean(ozone[garden=="A"]), ozone[i]))
  else
    lines(c(index[i], index[i]), c(mean(ozone[garden=="B"]), ozone[i]), col="red")
}
```



On the left-hand side, each data point indicates the twenty data points from the table above. The dots are marked with black and red for data points from Garden A and Garden B, respectively. The blue horizontal line is the mean value of the entire ozone data. The green lines are the distances between the mean and each data point. As you would probably remember, you can get the sum of squares by squaring the distances indicated by the green lines and summing all of them. Since this calculates the sum of squares of the entire data, we will call it the total sum of squares, SSY, which indicates the extent to which the overall data are scattered.

The plot on the right-hand side shows two horizontal lines instead of one. The two horizontal lines show that the mean ozone level of Garden A is lower than that of Garden B. But as students of statistics, let's ask ourselves this question: Is the difference statistically different?

If the two means are exactly the same, the horizontal lines will appear in the same place as the green line on the left side. But they are placed apart and we want to know if the distance is far enough for us to declare the two groups have statistically significantly different means.

In terms of the entire data, we can calculate the total sum of squares(SSY) by subtracting each data point from

the grand mean. But we can calculate the variance of data in separate groups. That is, data points in Garden A and those in Garden B are somewhat scattered in their respective groups. We can calculate how scattered or clustered data are in their respective groups.

Now, instead of measuring the distance between each data point and the grand mean (the mean of all data), we can subtract each data point from its corresponding group mean. That is, distances will be measured (1) between the red dots and the red horizontal line and (2) between the black dots and the black horizontal line. If we compare the total sum of squares (SSY) to the sum of squares calculated this way on the right-hand plot, which do you think will be longer? We call the latter error sum of squares (SSE) and it can not be larger than SSY.

$$SSY \geq SSE$$

In fact, what remains after SSE is subtracted from SSY is the variance explained by the group's difference. Since we called the variance of the entire data the total sum of squares (SSY) and the variance within each group the error sum of squares (SSE), the variance created by the group difference can be called the treatment sum of squares (SSA). The total sum of squares (SSY) is the sum of the treatment sum of squares (SSA) and the error sum of squares (SSE).

$$SSY = SSA + SSE$$

From this, we can calculate the variance explained by the treatement (or the factor, in this case, 'Which garden?').

$$SSA = SSY - SSE$$

Let's try and calculate SSY (Table 1) and SSE (Table 2) by hand.

Did you finish your calculations? You will be provided an excel file along with this handout and using functions in excel will let you quickly calculate the sum of squares. Or you can use the functions below.

```
SSY <- sum((ozone-mean(ozone))^2)
SSY
```

```
[1] 44
```

```
sum((ozone[garden=="A"]- mean(ozone[garden=="A"]))^2)
```

```
[1] 12
```

```
sum((ozone[garden=="B"]- mean(ozone[garden=="B"]))^2)
```

```
[1] 12
```

Now we know the total sum of squares and the error sum of squares, we can calculate the treatment sum of squares.

$$SSA = SSY - SSE = 44 - 24 = 20$$

When we use ANOVA, our main aim is to see if the factor (or treatment or group) explains a large enough amount of variance in the total variance. In this case, we are interested in the difference by garden (whether observations were made in Garden A or in Garden B). If the variance of your data is explained more by error than by the factor of interest, then, maybe it is not a meaningful factor. Calculate the ratio of variance explained by the factor to that unexplained (error variance).

Do you remember that variance is calculated by dividing the sum of squares by the degree of freedom? The factor degree of freedom is calculated by subtracting 1 from the number of levels in the factor. Since the factor 'garden' has two levels (A and B), the degree of freedom is 1. The error degree of freedom is calculated by subtracting the number of levels in the factor from the total number of data points (20 - 2 = 18). Mean square column shows the variances and is calculated by dividing the sum of squares by the degree of freedom. Since the treatment sum of squares is 20, and the degree of freedom is 1, the treatment mean square is 20. The mean square of the error is 1.333 (= 24/18). The F ratio is calculated by dividing the treatment mean square by the mean square of the error. The calculated F ratio is 15.0 (= 20/1.333).

Table 1: Calculating the total sum of squares (SSY)

Order	y	$\bar{y}$	$y - \bar{y}$	$(y - \bar{y})^2$
1	3	4		
2	5	4		
3	4	4		
4	5	4		
5	4	4		
6	6	4		
7	3	4		
8	7	4		
9	2	4		
10	4	4		
11	3	4		
12	4	4		
13	1	4		
14	3	4		
15	3	4		
16	5	4		
17	5	4		
18	6	4		
19	2	4		
20	5	4		
			Total	

Table 2: Calculating the error sum of squares (SSE)

Order	y_A	$\bar{y}_A$	$y - \bar{y}_A$	$(y - \bar{y}_A)^2$	y_B	$\bar{y}_B$	$y - \bar{y}_B$	$(y - \bar{y}_B)^2$
1	3	3			5	5		
2	4	3			5	5		
3	4	3			6	5		
4	3	3			7	5		
5	2	3			4	5		
6	3	3			4	5		
7	1	3			3	5		
8	3	3			5	5		
9	5	3			6	5		
10	2	3			5	5		
			Garden A				Garden B	

Table 3: ANOVA table

Source	Sum of squares	Degree of freedom	Mean square	F ratio
Garden	20	1	20	15
Error	24	18	$s^2 = 1.33$	
Total	44	19		

Now, with the F ratio of 15, we'd like to know if the value is above the 95% certainty level. We can use the F distribution table (in pdf or in hard copies) or we can simply use the r function to find the critical value at 95% certainty.

```
qf(0.95, 1, 18)
```

```
[1] 4.413873
```

The critical value is 4.41 and the F ratio we have is 15. We can safely say that the variance explained by the factor is much larger than the variance caused by error.

When you're done with calcualting the F ratio, you can also calculate its p-value by using the funciton below.

```
1-pf(15, 1, 18)
```

```
[1] 0.001114539
```

We have gone through a lot of steps to calculate the F ratio and the p value. But please rest assured that you don't have to do all these calculations every time you want to see if two or more group means are statistically significantly different. Use the `aov()` funciton.

```
summary(aov(ozone~garden))
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
garden       1     20  20.000      15 0.00111 **
Residuals   18     24   1.333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(aov(ozone~garden))
```

```

Call:
aov(formula = ozone ~ garden)

Residuals:
   Min     1Q Median     3Q    Max
    -2     -1      0      1      2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0000     0.3651   8.216 1.67e-07 ***
gardenB       2.0000     0.5164   3.873  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.155 on 18 degrees of freedom
Multiple R^2:  0.4545,    Adjusted R^2:  0.4242
F-statistic:    15 on 1 and 18 DF,  p-value: 0.001115
```

As we have seen above, the mean ozone level for Garden A is 3 and that for Garden B is 5. The regression coefficients table shows that the Intercept is 3 and it moves 2 units with gardenB. This means that the mean ozone level of garden A is set to be the intercept and when the garden factor changes to B, the mean value increases 2 units. You can get the t value using `t.test()` as well.

```
t.test(ozone~garden)
```

```
        Welch Two Sample t-test

data:  ozone by garden
t = -3.873, df = 18, p-value = 0.001115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0849115 -0.9150885
sample estimates:
mean in group A mean in group B
              3               5
```

When you calculate the F value and the t value for the mean difference of only two groups, you can square the t value to get F.

$$t^2 = F$$

The t value of (-)3.873 is seen in the outputs of both `t.test()` and the linear regression. If you square (-)3.873, the output is 15.00013.

The procedure above was to explain to you how things work in ANOVA but we don't use ANOVA to compare two sample means. For two sample means' comparison, t.test is simpler. Let's bring in another file with three levels in a factor so we can compare three sample means.

```
growth ← read.csv("C:/Users/hyuna/OneDrive/Documents/01SNU/03GraduateSeminar/crawley/data/
    growth.csv")

head(growth)
```

```
  supplement  diet     gain
1  supergain wheat 17.37125
2  supergain wheat 16.81489
3  supergain wheat 18.08184
4  supergain wheat 15.78175
5    control wheat 17.70656
6    control wheat 18.22717
```

```
nrow(growth)
```

```
[1] 48
```

```
attach(growth)
tapply(gain, list(supplement, diet), mean)
```

```
            barley     oats    wheat
agrimore  26.34848 23.29838 19.63907
control   23.29665 20.49366 17.40552
supergain 22.46612 19.66300 17.01243
supersupp 25.57530 21.86023 19.66834
```

As you can see from above, there are three columns in the data frame: supplement, diet, and gain. Also, there are 48 observations. The data frame shows the growth level of farm animals by two factors: supplement and diet. The supplement factor has four levels: agrimore, control, supergain, and supersupp. The diet factor has three levels: barley, oats, wheat. The table above summarizes the mean values in each of the twelve diffrent conditions.

Of course, we can calculate the mean values of growth by supplement only.

```
tapply(gain, supplement, mean)
```

```
 agrimore   control supergain supersupp
 23.09531  20.39861  19.71385  22.36796
```

OR we can do that by diet.

```
tapply(gain, diet, mean)
```

```
  barley     oats    wheat
24.42164 21.32882 18.43134
```

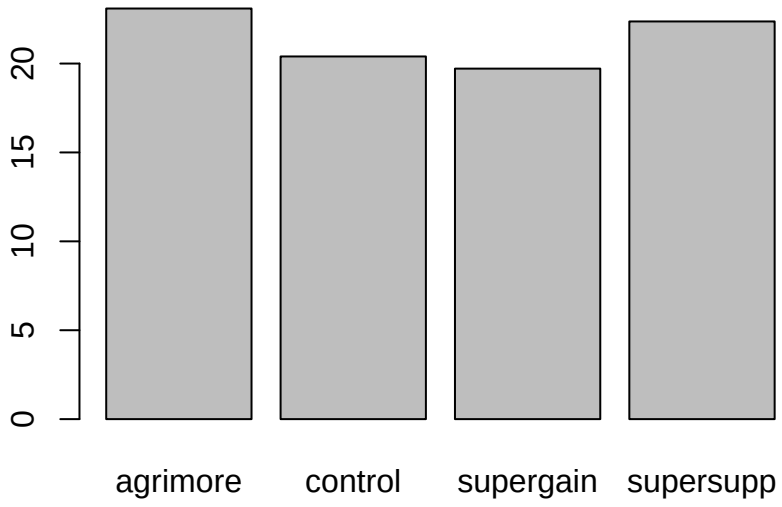
And now we get curious. Will the growth level differ by supplements? Or by diets? We can use one-way ANOVA to see if they are different.

```
model1 ← aov(gain~supplement)
summary(model1)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
supplement   3   91.9  30.627   3.823 0.0161 *
Residuals   44  352.5   8.011
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can also plot the data to see the relationship between growth (gain) and supplement.

```
barplot(tapply(gain, supplement, mean))
```



At a glance, you can see that the bars for agrimore and suppersup are slighly higher than those for control and supergain. But is the difference statistically significant?

```
TukeyHSD(model1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = gain ~ supplement)

$supplement
                        diff        lwr        upr     p adj
control-agrimore    -2.6967005 -5.7818024  0.3884014 0.1058141
supergain-agrimore  -3.3814586 -6.4665605 -0.2963567 0.0267429
supersupp-agrimore  -0.7273521 -3.8124540  2.3577498 0.9220050
supergain-control   -0.6847581 -3.7698601  2.4003438 0.9337824
supersupp-control    1.9693484 -1.1157536  5.0544503 0.3336830
supersupp-supergain  2.6541065 -0.4309954  5.7392084 0.1142704
```

You can check the p value (adjusted for multiple comparisons) to see if the difference is at the significant level. You can see that the difference between supergain and agrimore is statistically significant.

The last function we used `TukeyHSD()` is one of the post-hoc analysis you can use after an ANOVA model. An ANOVA model only tells you if any of the groups means is significantly different from the other group means, but it does not tell you which group means are different from one another. The Tukey 'honestly significant difference'

test will show you where the true difference lies.

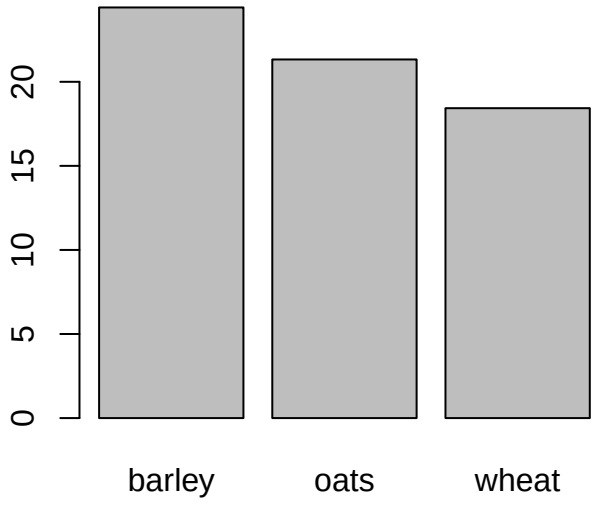
Now, let's take a look at diets. Will diets also make a difference in growth?

```
model2<-aov(gain~diet)
summary(model2)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
diet         2  287.2  143.59   41.11  7e-11 ***
Residuals   45  157.2    3.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F table shows that there is a significant difference by diet but let's see where the difference is.

```
barplot(tapply(gain, diet, mean))
```



You can see that barley leads to the highest gain and wheat the lowest. But which two groups means are significantly different from each other?

```
TukeyHSD(model2)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = gain ~ diet)

$diet
                  diff       lwr       upr     p adj
oats-barley  -3.092817 -4.694242 -1.491391 0.0000773
wheat-barley -5.990298 -7.591723 -4.388872 0.0000000
wheat-oats   -2.897481 -4.498906 -1.296055 0.0002006
```

As you can see, all three pairs are significantly different from each other.

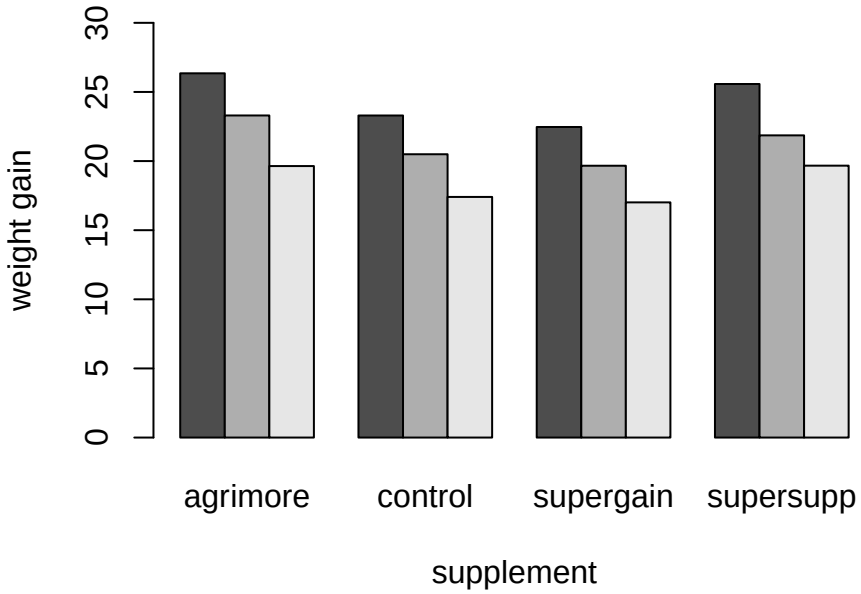
Now, we're done with analysis of variance that takes into consideration one factor at a time. Let's think about how we can take both factors into consideration. People often say that there is an interaction between factor A and factor B. What does that mean? Let's take a look.

2 Factorial ANOVA

```
model3 <- aov(gain~supplement*diet)
summary(model3)
```

```
                Df  Sum Sq Mean Sq F value   Pr(>F)
supplement       3   91.88   30.63   17.82 2.95e-07 ***
diet             2 287.17  143.59   83.52 3.00e-14 ***
supplement:diet  6    3.41    0.57    0.33    0.917
Residuals       36   61.89    1.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
barplot(tapply(gain, list(diet, supplement), mean), beside = T, ylab = "weight gain", xlab = "
    supplement", ylim = c(0, 30))
```



The plot shows the general pattern we observed in one-way ANOVAs. The supplements control and supergain resulted in lower weight gains than agrimore and supersupp did. Also, barley led to higher weight gain than oats did, which led to higher weight gains than wheat. The ANOVA table also shows that there is no particular interaction. This means that the effect of diet and the effect of supplement that were observed independently do not result in any different effects when the two factors act together.