STA258
Analysis of Variance

Al Nosedal.
University of Toronto.

Winter 2017

The Data Matrix

The following table shows last year's sales data for a small business. The sample is put into a matrix format in which each of the three rows corresponds to one of the three countries in which the company does business, and each of the four columns corresponds to one of its four salespersons. So a cell in the matrix corresponds to one of 12 salesperson/country combinations. The numbers in the cell represent the sales (in units of $1000) made by that particular salesman in that country last year. This data will be used throughout the chapter to develop the theory underlying **Analysis of Variance** or, for short, **ANOVA**.

The Data Matrix

	Country A	Country B	Country C	Average
Salesperson 1	6, 7, 8	10, 10	12, 13, 14	10
Salesperson 2	10, 15	10, 10	11, 16	12
Salesperson 3	10, 15	7, 13	11, 16	12
Salesperson 4	2, 3, 4	10, 10	8, 9, 10	7
Average	8	10	12	10

Altogether, there were 28 sales last year that totaled \$280 - so the average sale was \$10. The row (salesperson) averages are:
Row 1 (Salesperson 1) \$10.
Row 2 (Salesperson 2) \$12.
Row 3 (Salesperson 3) \$12.
Row 4 (Salesperson 4) \$ 7.

The column (country) averages are:
Column 1 (Country A) $8.
Column 2 (Country B) $10.
Column 3 (Country C) $12.

Now we will begin our study of how to make a statistically valid prediction of the next sales figure. In that regard, there are four possible situations that can occur.

1. Neither the country nor the salesperson of the next sale (observation) is known.
2. The country of the next sale is known, but the salesperson is not known.
3. The salesperson of the next sale is known, but the country is not known.
4. Both the country and the salesperson of the next sale are known.

Situation 1.

Without any additional information, the best prediction is the sample mean \$10. This prediction is best in the least squares sense - that is, if \$10 had been used to predict each of the 28 observations in the sample, then the total of the squared errors SS_{TOTAL} would be as small as possible. In our data set, SS_{TOTAL} equals 354. That figure can be verified by calculating $\sum(x_i - 10)^2$ for each observation x_i of the sample.

Situation 2. One-factor ANOVA Model.
If only the country of the next sale is known, then two different predictions are possible for the next sales figure:

- The sample mean \$10.
- The mean of the sales of the country in which the next sale will occur. (In this case, \$8 if the next sale will occur in Country A, \$10 in Country B, or \$12 in Country C.) This prediction **ignores** the information present in the sales figures from the other two countries.

Situation 3. One-factor ANOVA Model.
If only the salesperson of the next sale is known, then two different predictions are possible for the next sales figure:

- The sample mean \$10.
- The mean of the previous sales of the salesperson who will make the next sale. (In this case, \$10 if the next sale will be made by Salesperson 1, etc.) This prediction **ignores** the information present in the sales figures from the other three salespersons.

Situations 2 and 3 are called **one-factor ANOVA models**, because only one-factor is known about the next sale. As noted, if a prediction is either a row mean or a column mean, then it **ignores** the observations in the other rows or columns. To make that kind of prediction, it's necessary to statistically verify that the ignored observations are indeed different populations and therefore not relevant to the prediction.

Situation 4. Two-factor ANOVA Model.
We are not covering this kind of model in our course.

The Null Hypothesis for One-Factor ANOVA

We have discussed the prediction possibilities for one-factor ANOVA models. Now, we will learn how to test the statistical significance of a one-factor ANOVA model.
Let's suppose that we want to predict the next sales figure, and that we know the country in which this sale will occur but the identity of the salesperson is NOT known. Without any statistical testing, we can always by default use the sample mean \$10 to predict the next sale. The default prediction, the sample mean, doesn't use any information about the country (column) in which the sale will occur.

The Null Hypothesis for One-Factor ANOVA

However, if instead we use the mean of the observations in only one column (the column that corresponds to the particular country in which we know the next sale will occur), then we have to test the null hypothesis

$$H_0 : \mu_{COL1},\ \mu_{COL2},\ \mu_{COL3}, \quad \text{are equal}$$

and reject it in favor of the alternative hypothesis

$$H_a : \mu_{COL1},\ \mu_{COL2},\ \mu_{COL3}, \quad \text{are NOT all equal}$$

The Null Hypothesis for One-Factor ANOVA

If the null hypothesis is rejected, then we can be statistically confident that the column means are not all equal, and therefore that the individual column means (i. e., \$8, \$10, \$12) can be used to predict the amount of the next sale. If the next sale sale was going to occur in Country A, then the prediction would be \$8. If the next sale was going to occur in Country B, then the prediction would be \$10. If the next sale was going to occur in Country C, then the prediction would be \$12.

The One-Factor ANOVA F Test

To test the null hypothesis stated above, we have to calculate an F-statistic. If $F_{STAT} > F_{(c-1,n-c),\ \alpha}$, then reject H_0, and use the sample column means to predict future observations. Otherwise, do not reject H_0 and use the overall sample mean to predict future observations.

ANOVA Table

To see how this F_{STAT} is calculated, see the ANOVA Table below.

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Explained	c -1	SS_{EXP}	$\frac{SS_{EXP}}{c-1}$	$F = \frac{MSS_{EXP}}{MSS_{UNEXP}}$
Unexplained	n -c	SS_{UNEXP}	$\frac{SS_{UNEXP}}{n-c}$	
Total	n -1	SS_{TOTAL}		

Calculation of SS_{TOTAL}

If **no** model is used, then the predictions for each of the 28 observations (in dollar amounts) will be 10. If these predictions are used, the squared error of these 28 predictions is given in the table below.

	Country A	Country B	Country C
Salesperson 1	16, 9, 4	0, 0	4, 9, 16
Salesperson 2	0, 25	0, 0	1, 36
Salesperson 3	0, 25	9, 9	1, 36
Salesperson 4	64, 49, 36	0, 0	4, 1, 0

Prediction Errors Squared when NO Factor is used (Total) = 354.

Calculation of $SS_{UNEXPLAINED}$

If the column model is used, then the 28 observations would have the following 28 predictions, where $8 is the average for the first column, $10 is the average for the second column, and $12 is the average for the third column.

	Country A	Country B	Country C
Salesperson 1	8, 8, 8	10, 10	12, 12, 12
Salesperson 2	8, 8	10, 10	12, 12
Salesperson 3	8, 8	10, 10	12, 12
Salesperson 4	8, 8, 8	10, 10	12, 12, 12

Calculation of $SS_{UNEXPLAINED}$

Using the above 28 predictions, the errors squared are shown in the table below.

	Country A	Country B	Country C
Salesperson 1	4, 1, 0	0, 0	0, 1, 4
Salesperson 2	4, 49	0, 0	1, 16
Salesperson 3	4, 49	9, 9	1, 16
Salesperson 4	36, 25, 16	0, 0	16, 9, 4

Errors Squared when the Column Factor is used (Total) $= 274$.

Calculation of $SS_{EXPLAINED}$

The units explained by the row model are calculated by finding the square of each prediction change when moving from NO model to the column model. The following table presents the square of each prediction change:

	Country A	Country B	Country C
Salesperson 1	4, 4, 4	0, 0	4, 4, 4
Salesperson 2	4, 4	0, 0	4, 4
Salesperson 3	4, 4	0, 0	4, 4
Salesperson 4	4, 4, 4	0, 0	4, 4, 4

Table of the Square of the Prediction Change when Moving from NO Model to the Row Model (Total) = 80.

ANOVA Table

The ANOVA Table for the column factor can now be filled in as shown below:

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Explained	2	80	$\frac{80}{2} = 40$	$\frac{40}{10.96} = 3.65$
Unexplained	25	274	$\frac{274}{25} = 10.96$	
Total	27	354		

So for this one-factor ANOVA model, $F_{STAT} = 3.65$.

Conclusion

If the null hypothesis is true, then the F-statistic should be a value from the $F_{2,\ 25}$ distribution. Referring to the table that contains the upper 0.05 cut-off points of F distributions, we see that $F_{(2,25),0.05} = 3.39$. Since 3.65 is greater than 3.39, this tells us that the F-statistic is in the upper 0.05 of the $F_{2,\ 25}$ distribution. Therefore we can reject the null hypothesis at the 0.05 significance level, and we conclude that the country means are not all the same. Thus, the prediction for the next sale in a known country is the mean of all the previous sales in that country.

R Code

```r
# Step 1. Entering data;

sales1=c(6,7,8,10,15,10,15,2,3,4);

sales2=c (10,10,10,10,7,13,10,10);

sales3=c(12,13,14,11,16,11,16,8,9,10);

sales=c(sales1,sales2,sales3);

country=c(rep(1,10),rep(2,8),rep(3,10));
```

R Code

```
# Step 2. ANOVA;

oneway.test(sales~country,var.equal=TRUE);

##
##  One-way analysis of means
##
## data:  sales and country
## F = 3.6496, num df = 2, denom df = 25, p-value = 0.04068
```

This time for the Row Factor

We have just performed the F test to verify that the country (column) one-factor ANOVA model is statistically significant. There is another one-factor ANOVA model that also could be examined - the salesperson (row) factor model. Let's test

$$H_0 : \mu_{ROW1},\ \mu_{ROW2},\ \mu_{ROW3},\ \mu_{ROW4} \quad \text{are equal}$$

and reject it in favor of the alternative hypothesis

$$H_a : \mu_{ROW1},\ \mu_{ROW2},\ \mu_{ROW3},\ \mu_{ROW4} \quad \text{are NOT all equal}$$

ANOVA Table

The resulting ANOVA table for the salesperson (row) factor is shown below:

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Explained	3	120	$\frac{120}{3} = 40$	$\frac{40}{9.75} = 4.10$
Unexplained	24	234	$\frac{234}{24} = 9.75$	
Total	27	354		

So for this one-factor ANOVA model, $F_{STAT} = 4.10$.

Conclusion

Consulting the upper 0.05 cut-off table for the F distribution, we find that

$F_{(3,\ 24),0.05} = 3.01$

Since F-statistic $= 4.10 > 3.01$, the null hypothesis can once again be rejected at the 0.05 level, and we can use the salesperson factor to predict sales, concluding that it is statistically valid to predict either \$10, \$12, or \$7, respectively, for Salespersons 1, 2, 3, or 4.

R Code

```r
# Step 1. Entering data;

sales1=c(6, 7, 8, 10, 10, 12, 13, 14);

sales2=c (10, 15 , 10, 10 , 11, 16 );

sales3=c(10, 15 , 7, 13 , 11, 16 );

sales4= c(2, 3, 4 , 10, 10 , 8, 9, 10 );

sales=c(sales1,sales2,sales3,sales4);

person=c(rep(1,8),rep(2,6),rep(3,6),rep(4,8));
```

R Code

```
# Step 2. ANOVA;

oneway.test(sales~person,var.equal=TRUE);

##
##  One-way analysis of means
##
## data:  sales and person
## F = 4.1026, num df = 3, denom df = 24, p-value = 0.01748
```

Underlying Assumptions

Officially, to use the predictions from an ANOVA model, three assumptions about the populations from which the sample was taken must be satisfied:

1. Each population has a Normal distribution.
2. Each population has the **same** standard deviation σ.
3. The observations are mutually independent of one another.

Formulas

Sum of Squares for Treatments (a.k.a. between-treatments variation or Explained)

$$SST = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

Sum of Squares for Error (a.k.a. within-treatments variation or Unexplained)

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = (n_1 - 1)s_1^2 + ... + (n_k - 1)s_k^2.$$

Formulas

Mean Square for Treatments

$$MST = \frac{SST}{k-1}$$

Mean Square for Error

$$MSE = \frac{SSE}{n-k}$$

Formulas

Test Statistic

$$F = \frac{MST}{MSE}$$

Exercise

A statistics practitioner calculated the following statistics:

		Treatment	
Statistic	1	2	3
n	5	5	5
$\bar{x}$	10	15	20
s^2	50	50	50

Complete the ANOVA table.

Solution

$$\bar{\bar{x}} = \frac{5(10)+5(15)+5(20)}{5+5+5} = 15$$

$$SST = 5(10-15)^2 + 5(15-15)^2 + 5(20-15)^2 = 250$$

$$SSE = (5-1)(50) + (5-1)(50) + (5-1)(50) = 600$$

ANOVA Table

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Treatments	2	250	$\frac{250}{2} = 125$	$\frac{125}{50} = 2.50$
Error	12	600	$\frac{600}{12} = 50$	
Total	14	850		

Exercise

A statistics practitioner calculated the following statistics:

		Treatment	
Statistic	1	2	3
n	4	4	4
$\bar{x}$	20	22	25
s^2	10	10	10

Complete the ANOVA table.

Solution

$$\bar{\bar{x}} = \frac{4(20)+4(22)+4(25)}{4+4+4} = 22.33$$
$$SST = 4(20-22.33)^2 + 4(22-22.33)^2 + 5(25-22.33)^2 = 50.67$$
$$SSE = (4-1)(10) + (4-1)(10) + (4-1)(10) = 90$$

ANOVA Table

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Treatments	2	50.67	$\frac{50.67}{2} = 25.33$	$\frac{25.33}{10} = 2.53$
Error	9	90	$\frac{90}{9} = 10$	
Total	11	140.67		

Exercise

A consumer organization was concerned about the differences between the advertised sizes of containers and the actual amount of product. In a preliminary study, six packages of three different brands of margarine that are supposed to contain 500ml were measured. The differences from 500 ml are listed here. Do these data provide sufficient evidence to conclude that differences exist between the three brands? Use $\alpha = 0.05$.

Brand 1	Brand 2	Brand 3
1	2	1
3	2	2
3	4	4
0	3	2
1	0	3
0	4	4

Solution

Step 1. State Hypotheses.

μ_i = population mean for differences from 500 ml (brand i, where $i = 1, 2, 3$).

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_a : At least two means differ.

Solution

Step 2. Compute test statistic.

	Brand 1	Brand 2	Brand 3
Mean	1.33	2.50	2.67
Variance	1.87	2.30	1.47

Grand mean $= \bar{\bar{x}} = 2.17$.

$SST = 6(1.33 - 2.17)^2 + 6(2.50 - 2.17)^2 + 6(2.67 - 2.17)^2 = 6.387 \approx 6.39$

$SSE = (6-1)(1.87) + (6-1)(2.30) + (6-1)(1.47) = 28.20$

Solution

Grand mean $= \bar{\bar{x}} = 2.17.$
$SST = 6(1.33 - 2.17)^2 + 6(2.50 - 2.17)^2 + 6(2.67 - 2.17)^2 =$
$6.387 \approx 6.39$
$SSE = (6-1)(1.87) + (6-1)(2.30) + (6-1)(1.47) = 28.20$

ANOVA Table

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Treatments	2	6.39	$\frac{6.39}{2} = 3.195$	$\frac{3.195}{1.88} = 1.70$
Error	15	28.20	$\frac{28.20}{15} = 1.88$	
Total	17	34.59		

Solution

Step 3. Find Rejection Region.
We reject the null hypothesis only if

$$F > F_{\alpha,k-1,n-k}$$

If we let $\alpha = 0.05$, the rejection region for this exercise is

$$F > F_{0.05,\ 2,15} = 3.682$$

Solution

Step 4. Conclusion.
We found the value of the test statistic to be $F = 1.70$. Since $F = 1.70 < F_{0.05,\ 2,15} = 3.682$, we **can't** reject H_0. Thus, there is **not** evidence to infer that the average differences differ between the three brands.

R Code

```r
# Step 1. Entering data;

brand1=c(1,3,3,0,1,0);

brand2=c (2,2,4,3,0,4);

brand3=c(1,2,4,2,3,4);

differences=c(brand1,brand2,brand3);

brand=c(rep(1,6),rep(2,6),rep(3,6));
```

R Code

```
# Step 2. ANOVA;

oneway.test(differences~brand,var.equal=TRUE);

##
##  One-way analysis of means
##
## data:  differences and brand
## F = 1.6864, num df = 2, denom df = 15, p-value = 0.2185
```

APPENDIX

Balanced one-way analysis of variance:theory

For balanced analysis of variance, let $n = n_1 = n_2 = ... = n_a$ be the number of observations in each sample. In particular, we assume the data structure

Sample	Data		Distribution
1	$y_{11}, y_{12}, ..., y_{1n}$	iid	$N(\mu_1, \sigma^2)$
2	$y_{21}, y_{22}, ..., y_{2n}$	iid	$N(\mu_2, \sigma^2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	$y_{a1}, y_{a2}, ..., y_{an}$	iid	$N(\mu_a, \sigma^2)$

with all samples independent.

The data structure can be rewritten as the balanced one-way ANOVA model

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \text{ independent } N(0, \sigma^2)$$

$i = 1, 2, ..., a$, $j = 1, ..., n$

We focus on testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_a.$$

The test of H_0 is based on estimating σ^2. We construct two estimates of the variance. The first estimate is always valid. The second estimate is valid **only** when $\mu_1 = \mu_2 = ... = \mu_a$. If the μ_is are all equal, we have two estimates of σ^2, so they should be about the same.

The first estimate of the variance, the one that is always valid. The mean squared error is

$$MSE = \frac{S_1^2 + S_2^2 + ... + S_a^2}{a} = \frac{1}{a(n-1)} \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

The degrees of freedom associated with the MSE are the degrees of freedom for error (dfE), we we have

$$dfE = a(n-1).$$

The distribution of the MSE is related to the χ^2 family of distributions.

$$\frac{dfE \times MSE}{\sigma^2} \sim \chi^2(dfE).$$

A commonly used terminology in analysis of variance is the sum of squares for error (SSE). This is defined to be

$$SSE = dfE \times MSE = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

Note that

$$\frac{SSE}{\sigma^2} \sim \chi^2(dfE).$$

The second estimate of σ^2 is to be valid only when $\mu_1 = \mu_2 = ... = \mu_a$. Consider the distributions of the $\bar{y}_i$s. Each is the sample mean of n observations, so each has the distribution of a sample mean. In particular,

$$\bar{y}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$$
$$\bar{y}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n}\right)$$
$$\vdots$$
$$\bar{y}_a \sim N\left(\mu_a, \frac{\sigma^2}{n}\right)$$

The a different samples are independent of each other, so $\bar{y}_1, \bar{y}_2, ..., \bar{y}_a$ are all independent.

The variance of the $\bar{y}_i$s is σ^2/n and the sample variance of the $\bar{y}_i$s is

$$S_{\bar{y}}^2 = \frac{1}{a-1}\sum_{i=1}^{a}(\bar{y}_i - \bar{\bar{y}})^2$$

where $\bar{\bar{y}} = \frac{1}{a}\sum_{i=1}^{a}\bar{y}_i$.

We have $S_{\bar{y}}^2$ as an estimate of σ^2/n but we set out to find an estimate of σ^2. The obvious choice is

$$MSTrts = nS_{\bar{y}}^2 = \frac{n}{a-1}\sum_{i=1}^{a}(\bar{y}_i - \bar{\bar{y}})^2$$

where $MSTrts$ abbreviates the commonly used term **mean squared treatments**. $MSTrts$ has $a-1$ degrees of freedom.

The sum of squares for treatments is defined as

$$SSTrts = dfTrts \times MSTrts = n \sum_{i=1}^{a} (\bar{y}_i - \bar{\bar{y}})^2$$

The estimate $S_{\bar{y}}^2$ is the sample variance of a random sample of size a from a Normal population with variance σ^2/n, so

$$\frac{(a-1)S_{\bar{y}}^2}{\sigma^2/n} = \frac{(a-1)MSTrts}{\sigma^2} \sim \chi^2(a-1).$$

A decision regarding the validity of the claim $\mu_1 = \mu_2 = ... = \mu_a$ is based on comparing $MSTrts$ with MSE. The hypothesis $\mu_1 = \mu_2 = ... = \mu_a$ is rejected at the α level if

$$\frac{MSTrts}{MSE} \geq F(1-\alpha, a-1, dfE).$$

Here $F(1-\alpha, a-1, dfE)$ is the number below which $(1-\alpha)100\%$ of the possible F ratios when the μ_is are all equal.