

Chapter 11: Analysis of variance

Note made by: Timothy Hanson
Instructor: Peijie Hou

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Comparing more than two means

- In Chapter 7 we had two groups and tested $H_0 : \mu_1 = \mu_2$.
- In Chapter 11 we will have I groups and test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$.
- We are still interested in whether the population means are the same across groups, there's just more than two.
- The alternative hypothesis is H_A : one or more of $\mu_1, \mu_2, \dots, \mu_I$ are different.
- Let's look at an example where $I = 5$.

Example 11.1.1: Organic farming of corn

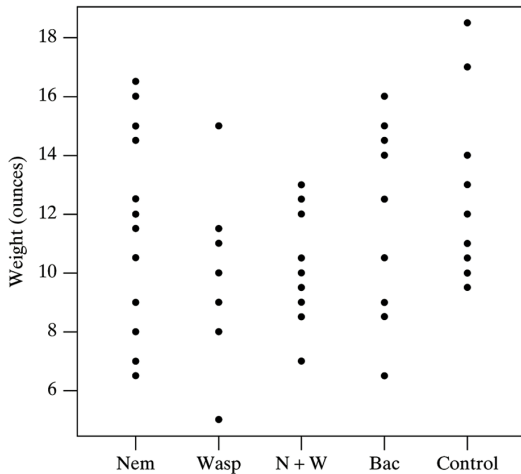
When growing sweet corn, can organic methods be used successfully to control harmful insects and limit their effect on the corn? In a study researchers compared the weights of ears of corn under five organic conditions: (a) using a beneficial soil nematode, (b) parasitic wasp, (c) both the nematode and the wasp, (d) a bacterium, and (e) a control with no treatment. In summary, the treatments are

- Treatment 1: Nematodes
- Treatment 2: Wasps
- Treatment 3: Nematodes and wasps
- Treatment 4: Bacteria
- Treatment 5: Control

Corn data

Table 11.1.1 Weights (ounces) of ears of sweet corn					
	Treatment				
	1	2	3	4	5
	16.5	11.0	8.5	16.0	13.0
	15.0	15.0	13.0	14.5	10.5
	11.5	9.0	12.0	15.0	11.0
	12.0	9.0	10.0	9.0	10.0
	12.5	11.5	12.5	10.5	14.0
	9.0	11.0	8.5	14.0	12.0
	16.0	9.0	9.5	12.5	11.0
	6.5	10.0	7.0	9.0	9.5
	8.0	9.0	10.5	9.0	18.5
	14.5	8.0	10.5	9.0	17.0
	7.0	8.0	13.0	6.5	10.0
	10.5	5.0	9.0	8.5	11.0
Mean	11.5	9.6	10.3	11.1	12.3
SD	3.5	2.4	2.0	3.1	2.9
<i>n</i>	12	12	12	12	12

Side-by-side dotplots of corn data

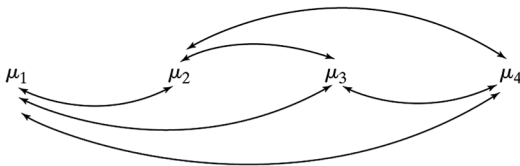


Comparing more than two means at the same time

- It's natural to ask: why not now just compare all possible pairs $\mu_1 - \mu_2, \mu_1 - \mu_3, \mu_2 - \mu_3$, etc., with a bunch of pairwise t-tests?
- This results in *multiple comparisons*: the same data set is used to make multiple inferences on $I > 2$ associated parameters.
- The overall Type I error is then larger (perhaps much larger) than α .
- ANOVA keeps α fixed, even with lots of comparisons all at once.

Comparing four population means requires six comparisons

$$\begin{array}{lll} H_0: \mu_1 = \mu_2 & H_0: \mu_1 = \mu_3 & H_0: \mu_1 = \mu_4 \\ H_0: \mu_2 = \mu_3 & H_0: \mu_2 = \mu_4 & H_0: \mu_3 = \mu_4 \end{array}$$



Comparing 3 means requires 3 comparisons, 4 means requires 6 comparisons, 5 means requires, 10 comparisons, etc. In general, there's $I(I - 1)/2$ pairwise comparisons to make.

Table 11.1.2 Overall risk of Type I error in using repeated t tests at $\alpha = 0.05$

I	Overall risk
2	0.05
3	0.12
4	0.20
6	0.37
8	0.51
10	0.63

This is an example of what the true Type I error might be if we didn't use ANOVA, but rather used $I(I - 1)/2$ t -tests.

Building the test

- We will test $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$ via *analysis of variance* (ANOVA).
- ANOVA compares how variable the sample means $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_I$ are to how variable observations are *around each mean*.
- Assumptions: Observations in each group are independently normally distributed with the same variance σ^2 .
- The data in different groups are also independent.

Sums of Squares

- $SS(\text{between}) = \sum_{i=1}^I n_i (\bar{y}_i - \bar{\bar{y}})^2$ measures the variability *explained by letting each group have its own mean.*
- $SS(\text{within}) = \sum_{i=1}^I (n_i - 1) s_i^2$ measures the variability *that is left over, also called “pure error.”*
- $SS(\text{total}) = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ measures *the total variability in the data ignoring groups.*
- $SS(\text{total}) = SS(\text{between}) + SS(\text{within}).$
- The variability in the data is equal to the variability explained by the model plus the slop that's left over.

- A mean square is the average of the squared deviations from a central value. It is a sum of squares divided by the number of informative values in the sum of squares, called the “degrees of freedom”, or df.
- The df for between is $df_b = I - 1$; the df for within is $df_w = n - I$ where $n = n_1 + n_2 + \cdots + n_I$.
- $MS(\text{between}) = SS(\text{between}) / (I - 1)$,
 $MS(\text{within}) = SS(\text{within}) / (n - I)$.
- We collect the sums of squares, mean squares, df, in a table called an ANOVA table.

ANOVA formulae

ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between groups	$I - 1$	$\sum_{i=1}^I n_i (\bar{y}_i - \bar{\bar{y}})^2$	SS/df
Within groups	$n_{\bullet} - I$	$\sum_{i=1}^I (n_i - 1) s_i^2$	SS/df
Total	$n_{\bullet} - 1$	$\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$	

Lamb weights

Table 11.2.1 Weight gains of lambs (lb)*

	Diet 1	Diet 2	Diet 3
	8	9	15
	16	16	10
	9	21	17
		11	6
		18	
n_i	3	5	4
Sum = $\sum_{j=1}^{n_i} y_{ij}$	33	75	48
Mean = \bar{y}_i	11.000	15.000	12.000
SD = s_i	4.359	4.950	4.967
*Extra digits are reported for accuracy of subsequent calculations.			

Table 11.2.3 ANOVA table for lamb weight gains

Source	df	SS	MS
Between diets	2	36	18.00
Within diets	9	210	23.33
Total	11	246	

Estimate of common $\sigma^2 = \text{MS}(\text{within})$.

- Want to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$.
- Test statistic is $F = MS(Between)/MS(Within)$, which has an $F(df_b, df_w)$ distribution if H_0 is true.
- R uses this F-distribution to get the P-value. Read your text if you are interested in more details.
- In R, need to define two lists. One list has the response measurement of interest, the other has which group each response is from. The grouping list needs to be a 'factor' – the sample code shows how to do this.
- In R, $SS(within)$ is called 'Residuals Sum Sq' and $SS(between)$ takes the same name as the group list name.
- Obtain P-value from R. Reject H_0 if P-value is less than α as usual.

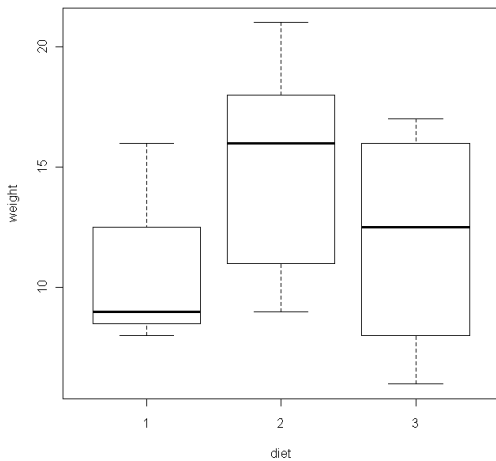
R code for lamb diet data

```
weight=c( 8,16, 9, 9,16,21,11,18,15,10,17, 6)
diet  =c( 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3)
diet=factor(diet)
boxplot(weight~diet)
fit=aov(weight~diet)
summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet	2	36	18.000	0.7714	0.4907
Residuals	9	210	23.333		

We accept $H_0 : \mu_1 = \mu_2 = \mu_3$ at the $\alpha = 0.05$ level. There is no significant difference in weight across diet.

Lamb data



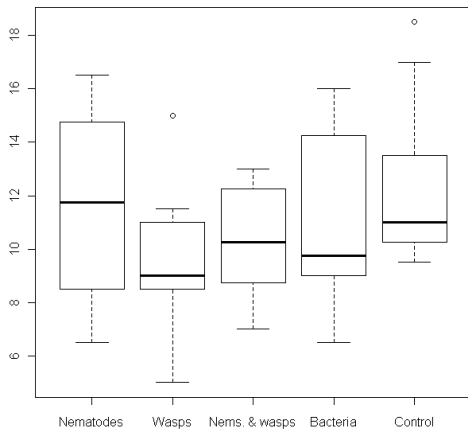
R code for corn growth data

```
weight=c(16.5,11.0, 8.5,16.0,13.0,15.0,15.0,13.0,14.5,10.5,
         11.5, 9.0,12.0,15.0,11.0,12.0, 9.0,10.0, 9.0,10.0,
         12.5,11.5,12.5,10.5,14.0, 9.0,11.0, 8.5,14.0,12.0,
         16.0, 9.0, 9.5,12.5,11.0, 6.5,10.0, 7.0, 9.0, 9.5,
         8.0, 9.0,10.5, 9.0,18.5,14.5, 8.0,10.5, 9.0,17.0,
         7.0, 8.0,13.0, 6.5,10.0,10.5, 5.0, 9.0, 8.5,11.0)
treat= c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,
         1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,
         1,2,3,4,5,1,2,3,4,5)
treat=factor(treat)
boxplot(weight~treat,names=c("Nematodes","Wasps","Nems. & wasps","Bacteria","Control"))
fit=aov(weight~treat)
summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	4	52.31	13.0771	1.6461	0.1758
Residuals	55	436.94	7.9443		

We accept $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ at the $\alpha = 0.05$ level.
There is no significant difference in weight across treatment.

Organic sweet corn boxplots



Example 1.1.4: MOA and schizophrenia

- Monoamine oxidase (MOA) enzyme thought to regulate behavior.
- Blood from $n = 42$ schizophrenia patients collected, *stratified* by diagnosis (I, II, III).
- Is there an association between MOA and diagnosis?

Example 1.1.4: MOA and schizophrenia

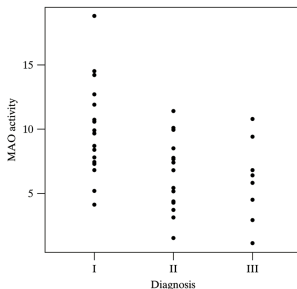


Figure 1.1.2 MAO activity in schizophrenic patients

- What happens to MOA as severity of diagnosis increases? Is the relationship perfect?
- Let's test $H_0 : \mu_1 = \mu_2 = \mu_3$ in R.

MAO activity (Fig. 1.1.2)

```
moa=c(6.8,4.1,7.3,14.2,18.8,9.9,7.4,11.9,5.2,7.8,7.8,8.7,12.7,14.5,10.7,8.4,9.7,  
      10.6,7.8,4.4,11.4,3.1,4.3,10.1,1.5,7.4,5.2,10.0,3.7,5.5,8.5,7.7,6.8,3.1,  
      6.4,10.8,1.1,2.9,4.5,5.8,9.4,6.8)  
group=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3)  
group=factor(group)  
boxplot(moa~group,names=c("Schizo I","Schizo II","Schizo III"))  
fit<-aov(moa~group)  
summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	136.1	68.06	6.346	0.00411 **
Residuals	39	418.3	10.72		

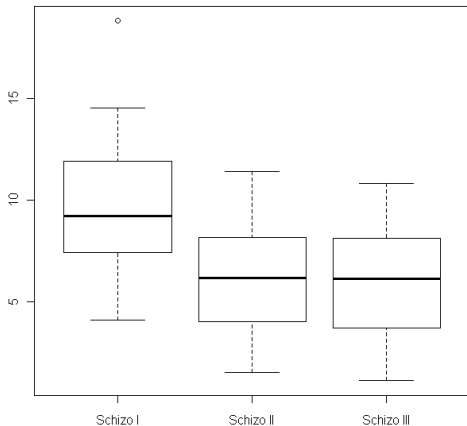
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	136.12	68.059	6.3461	0.004111 **
Residuals	39	418.25	10.724		

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Reject $H_0 : \mu_1 = \mu_2 = \mu_3$ at $\alpha = 0.05$ because $0.004 < 0.05$.
There is a significant difference in mean MOA activity across the three diagnoses.

MAO activity in schizophrenia



Side-by-side boxplots from R code.