

**Class 21: Analysis of Variance (Text: Sections 12.1)**

**Where We Are?**

<i>Quantitative Variable</i>	<i>Categorical Variables</i>
<b>1 or 2 Samples</b> Use Z test if $\sigma$ known or T test if $\sigma$ not known	<b>1 or 2 Samples</b> Use Z test
<b>More than 2 samples</b> Use ANOVA	<b>More than 2 samples</b> Use Chi-Square test

ANOVA stands for “Analysis of Variance”, but in spite of its name, it tests whether the *means* of several populations are equal.

**ANOVA: The Big Idea**

As with chi-square, we’ll start with examples where there are only two samples (which we could have done with a Z or T test) and then go on and use the method when there are three or more samples.

**In which of the following examples do the two samples appear to have come from populations with the same mean and which look like they came from populations with different means?**

**Example 1**

Sample 1	8	9	11	12
Sample 2	18	19	21	22

**Example 2**

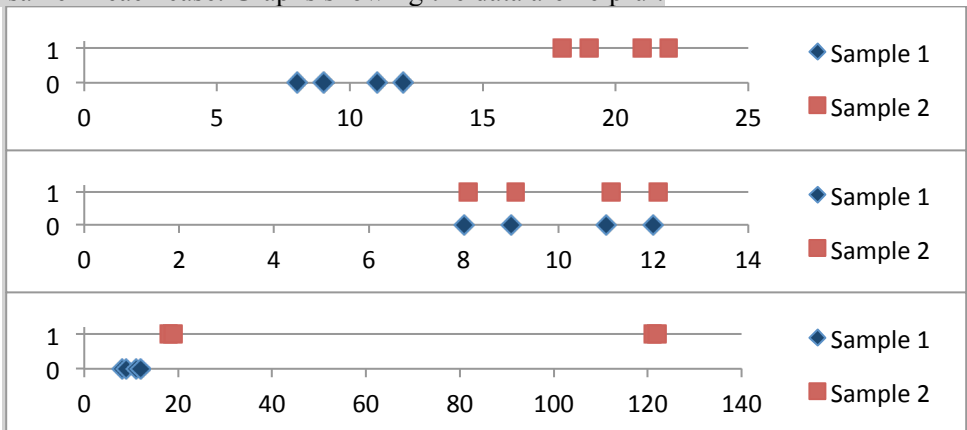
Sample 1	8	9	11	12
Sample 2	8.1	9.1	11.1	12.1

**Example 3**

Sample 1	8	9	11	12
Sample 2	18	19	121	122

Notice that Sample 1 is the same in each case. Graphs showing the data are helpful:

**Example 1**



**Example 2**

**Example 3**

**What can we see?**

In Example 1, it looks very unlikely that the samples came from the same population. The means of population 2 looks larger than the mean of population 1.

In Example 2, the samples could easily come from the same population. The means of population 2 looks similar to the mean of population 1.

In Example 3, it is hard to tell. The means of the samples are very different, but the standard deviation of Sample 2 is so big that the difference in means could just be the result of random variation.

**What can we conclude?**

- If the difference between the sample means is *small* compared with the variation within the samples, then the means of the populations could be the same.
- If the difference between the sample means is *large* compared with the variation within the samples, then the means of the populations are not likely to be the same.

The samples are often called *groups*.

**How do we test this?**

We use variation, which is of the form

$$\sum (\text{Values} - \text{Mean})^2,$$

and then we create the *F*-statistic, which is the ratio

$$F = \frac{\text{Average between group variation}}{\text{Average within group variation}}.$$

The within group variation is often called the *error*, although it does not mean there is a mistake.

The *F*-statistic has the **F-distribution** which has two degrees of freedom, one determined by the numerator and one by the denominator. The calculations to find *F* are messy, so we will mostly look at them on the computer.

**The F-test**

**Null hypothesis:** The means of all the populations are equal.

**Alternative hypothesis:** The means of all the populations are not all equal. Thus at least one population has a different mean, but ANOVA does not tell us which. (We have to look back at the data.)

This is exactly like the chi-squared test, when if we rejected the null hypothesis, we knew there was an interaction, but we didn't know what it was.

**Example 1:****Step 1:** Null is that the two populations have the same mean;**Alternate** is that the populations do not have the same mean.**Step 2:** Test statistic from Excel

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	4	40	10	3.333333		
Row 2	4	80	20	3.333333		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	200	1	200	60	0.000243	5.98737758
Within Groups	20	6	3.333333			
Total	220	7				

**Step 3:** The  $P$ -value is  $P = 0.000243$ , which is very small.**Step 4:** We reject the null hypothesis and conclude there is a difference in the population means.**Example 2: What sort of  $P$ -value do we expect?**

Large; lets check the output.

**Step 1:** Null is that the two populations have the same mean;**Alternate** is that the populations so not both have the same mean.**Step 2:** Test statistic from Excel

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	4	40	10	3.333333		
Row 2	4	40.4	10.1	3.333333		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.02	1	0.02	0.006	0.940776	5.98737758
Within Groups	20	6	3.333333			
Total	20.02	7				

**Step 3:** The  $P$ -value is  $P = 0.940776$ , which is very large.**Step 4:** We cannot reject the null hypothesis. We do not have evidence of a difference in the population means.**Example 3: What kind of  $P$ -value do we expect?** In the middle.**Excel says  $P = 0.09029$ ; what can we conclude?** We cannot reject the null—there is not sufficient evidence that the means are different.

**Notation**

We have  $k$  groups (that is  $k$  samples) of sizes  $n_1, n_2, \dots, n_k$ , respectively. (The book uses  $I$  for  $k$ .)

The total number of data points in all the samples combined is  $N$ , so  $N = \sum_{i=1}^k n_i$ .

The means of each of the samples are  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ , and the overall mean of all the samples together is  $\bar{x}$ .

The standard deviation of each of the samples are  $s_1, s_2, \dots, s_k$ . The pooled standard deviation is  $s_p$ . (Formula for  $s_p$ . in next class.)

**What does the other information in the table represent?****Example 1:**

Sample 1	8	9	11	12
Sample 2	18	19	21	22

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	4	40	10	3.333333		
Row 2	4	80	20	3.333333		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	200	1	200	60	0.000243	5.98737758
Within Groups	20	6	3.333333			
Total	220	7				

**SUMMARY gives the mean and variance of each sample**

Why are the means of the groups what they are?

Sample 1 centered on 10; Sample 2 centered on 20. That is  $\bar{x}_1 = 10$ ,  $\bar{x}_2 = 20$ ; Overall mean is  $\bar{x} = 15$

Variance of Sample 1 is  $((8 - 10)^2 + (9 - 10)^2 + (11 - 10)^2 + (12 - 10)^2)/3 = 10/3 = 3.333$

Why are the variances equal?

Both samples are equally spread out about their respective means.

What are the standard deviations of each of the groups?

$s_1 = s_2 = \sqrt{3.333} = 1.826$ .

**ANOVA gives the results of the test****What is the df?**

Degrees of freedom:

Between groups  $DF = k - 1 = 2 - 1 = 1$

Within groups  $DF = N - k = 8 - 2 = (4 - 1) + (4 - 1) = 6$

Total  $DF = N - 1 = 8 - 1 = 7$

Notice that  $7 = 6 + 1$ , that is

Total DF = Between group DF + Within group DF

**Sums of Squares, SS**

Sums of squares about the relevant mean.

**Between groups, SSG:** Sums of squares of deviation of group mean from the overall mean, weighted by the size of group:

$$SSG = \sum_{\text{All groups}} n_i(\bar{x}_i - \bar{x})^2 = 4(10 - 15)^2 + 4(20 - 15)^2 = 200$$

This measures *between* group variation, though not yet the average variation.

**Within groups, SSE:** Sums of squares of deviations of individual observations from the group (sample) means, 10 or 20, summed across groups. SSE is often called *error*, so if  $x_{ij}$  is a typical element in Group  $i$

$$SSE = \sum_{\text{Groups}} \left( \sum_{\text{Group } i} (x_{ij} - \bar{x}_i)^2 \right) = (8 - 10)^2 + \dots + (12 - 10)^2 + (18 - 20)^2 + \dots + (22 - 20)^2 = 20$$

This measures *within* group variation, though not yet the average.

**Total:** Sum of all the squares of the deviations of the individual observations from the **overall mean**, 15:

$$SST = \sum_{\text{All observations, all Groups}} (x_{ij} - \bar{x})^2 = (8 - 15)^2 + (9 - 15)^2 + \dots + (22 - 15)^2 = 220$$

Notice that  $220 = 200 + 20$ , that is

$$\text{Total SST} = \text{Between group SSG} + \text{Within group SSE}$$

**Mean Squares, MS**

These are the averages, we divide the sums of squares by the relevant degrees of freedom.

**Between Groups:** Mean between group variation

$$MSG = \frac{SSG}{DFG} = \frac{200}{1} = 200.$$

This is the average *between* group variation.

**Within Groups:** Mean within group variation, or error

$$MSE = \frac{SSE}{DFE} = \frac{20}{6} = 3.333.$$

This is the average *within* group variation.

**F-statistic**

We divide the measure of the separation of the sample means,  $MSG$ , by the variation in the samples,  $MSE$ :

$$F = \frac{MSG}{MSE} = \frac{200}{3.333} = \frac{200}{20/6} = 60.$$

**What is the distribution of the F-statistic?** What table do we look it up in?

$F$  has DF numerator = 1 and DF denominator = 6.

Table: Off the chart

Calculator:  $Fcdf(60, 1000, 1, 6) = 0.000243$  ...the 1000 is any very large positive number.

In general,  $F$  has DF numerator =  $k - 1$  and DF denominator =  $N - k$ . written  
 $F(k - 1, N - k)$

**Example 2: Fill in the eight missing values in the table. (Check with the original table.)**

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	4	40	10	3.333333		
Row 2	4	40.4	10.1	3.333333		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.02	1	0.02	0.006	0.940776	5.98737758
Within Groups	20	6	3.333333			
Total	20.02	7				

**Example 3: Fill in the missing values in the table.**

Sample 1	8	9	11	12		
Sample 2	18	19	121	122		
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Row 1	4	40	10	3.333333		
Row 2	4	280	70	3536.667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	7200	1	7200	4.067797	0.09029	5.98737758
Within Groups	10620	6	1770			
Total	17820	7				

### **What is F-crit? (Optional)**

Instead of using the  $P$ -value, some people use  $F$ -crit to determine if the null hypothesis should be rejected. If  $F > F$ -crit, then we reject the null. Thus  $F$ -crit is the  $F$ -value that corresponds to the significance level, here 5%;

Table: Find  $F = 5.99$  in DF num = 1, DF demon = 6,  $P$ -value = 0.05.

Calculator:  $Fcdf(5.98737758, 1000, 1, 6) = 0.05$ .