

Analysis of Variance (ANOVA)

Recall, when we wanted to compare two population means, we used the 2-sample t procedures. Now let's expand this to compare $k \geq 3$ population means. As with the t -test, we can graphically get an idea of what is going on by looking at side-by-side boxplots. (See Example 12.3, p. 748, along with Figure 12.3, p. 749.)

1 Basic ANOVA concepts

1.1 The Setting

Generally, we are considering a quantitative response variable as it relates to one or more explanatory variables, usually categorical. Questions which fit this setting:

- (i) Which academic department in the sciences gives out the lowest average grades? (Explanatory variable: *department*; Response variable: *student GPA's for individual courses*)
- (ii) Which kind of promotional campaign leads to greatest store income at Christmas time? (Explanatory variable: *promotion type*; Response variable: *daily store income*)
- (iii) How do the type of career and marital status of a person relate to the total cost in annual claims she/he is likely to make on her health insurance. (Explanatory variables: *career* and *marital status*; Response variable: *health insurance payouts*)

Each value of the explanatory variable (or value-pair, if there is more than one explanatory variable) represents a population or group. In the Physicians' Health Study of Example 3.3, p. 238, there are two factors (explanatory variables): aspirin (values are "taking it" or "not taking it") and beta carotene (values again are "taking it" or "not taking it"), and this divides the subjects into four groups corresponding to the four cells of Figure 3.1 (p. 239). Had the response variable for this study been quantitative—like systolic blood pressure level—rather than categorical, it would have been an appropriate scenario in which to apply (2-way) ANOVA.

1.2 Hypotheses of ANOVA

These are always the same.

H_0 : The (population) means of all groups under consideration are equal.

H_a : The (pop.) means are not all equal. (Note: This is different than saying "they are all unequal"!)

1.3 Basic Idea of ANOVA

Analysis of variance is a perfectly descriptive name of what is actually done to analyze sample data acquired to answer problems such as those described in Section 1.1. Take a look at Figures 12.2(a) and 12.2(b) (p. 746) in your text. Side-by-side boxplots like these in both figures reveal differences between samples taken from three populations. However, variations like those depicted in 12.2(a) are much less convincing that the population means for the three populations are different than if the variations are as in 12.2(b). The reason is because the ratio of variation between groups to variation within groups is much smaller for 12.2(a) than it is for 12.2(b).

1.4 Assumptions of ANOVA

Like so many of our inference procedures, ANOVA has some underlying assumptions which should be in place in order to make the results of calculations completely trustworthy. They include:

- (i) Subjects are chosen via a simple random sample.
- (ii) Within each group/population, the response variable is normally distributed.
- (iii) While the population means may be different from one group to the next, the population standard deviation is the same for all groups.

Fortunately, ANOVA is somewhat **robust** (i.e., results remain fairly trustworthy despite mild violations of these assumptions). Assumptions (ii) and (iii) are close enough to being true if, after gathering SRS samples from each group, you:

- (ii) look at normal quantile plots for each group and, in each case, see that the data points fall close to a line.
- (iii) compute the standard deviations for each group sample, and see that the ratio of the largest to the smallest group sample s.d. is no more than two.

2 One-Way ANOVA

When there is just *one* explanatory variable, we refer to the analysis of variance as *one-way ANOVA*.

2.1 Notation

Here is a key to symbols you may see as you read through this section.

k = the number of groups/populations/values of the explanatory variable/levels of treatment

n_i = the sample size taken from group i

x_{ij} = the j th response sampled from the i th group/population.

\bar{x}_i = the sample mean of responses from the i th group = $\frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

s_i = the sample standard deviation from the i th group = $\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

n = the (total) sample, irrespective of groups = $\sum_{i=1}^k n_i$.

\bar{x} = the mean of all responses, irrespective of groups = $\frac{1}{n} \sum_{ij} x_{ij}$

2.2 Splitting the Total Variability into Parts

Viewed as one sample (rather than k samples from the individual groups/populations), one might measure the total amount of variability among observations by summing the squares of the differences between each x_{ij} and \bar{x} :

$$\text{SST (stands for sum of squares total)} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2.$$

This variability has two sources:

1. Variability between group means (specifically, variation around the overall mean \bar{x})

$$\text{SSG} := \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2, \quad \text{and}$$

2. Variability within groups means (specifically, variation of observations about their group mean \bar{x}_i)

$$\text{SSE} := \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1) s_i^2.$$

It is the case that

$$\text{SST} = \text{SSG} + \text{SSE}.$$

2.3 The Calculations

If the variability between groups/treatments is large relative to the variability within groups/treatments, then the data suggest that the means of the populations from which the data were drawn are significantly different. That is, in fact, how the F statistic is computed: it is a measure of the variability between treatments divided by a measure of the variability within treatments. If F is large, the variability between treatments is large relative to the variation within treatments, and we reject the null hypothesis of equal means. If F is small, the variability between treatments is small relative to the variation within treatments, and we do not reject the null hypothesis of equal means. (In this case, the sample data is consistent with the hypothesis that population means are equal between groups.)

To compute this ratio (the F statistic) is difficult and time consuming. Therefore we are always going to let the computer do this for us. The computer generates what is called an ANOVA table:

Source	SS	df	MS	F
Model/Group	SSG	$k - 1$	$\text{MSG} = \frac{\text{SSG}}{k - 1}$	$\frac{\text{MSG}}{\text{MSE}}$
Residual/Error	SSE	$n - k$	$\text{MSE} = \frac{\text{SSE}}{n - k}$	
Total	SST	$n - 1$		

What are these things?

- The *source* (of variability) column tells us SS=Sum of Squares (sum of squared deviations):

SST measures variation of the data around the overall mean \bar{x}

SSG measures variation of the group means around the overall mean

SSE measures the variation of each observation around its group mean \bar{x}_i

- Degrees of freedom

$\underline{k - 1}$ for SSG, since it measures the variation of the k group means about the overall mean

$\underline{n - k}$ for SSE, since it measures the variation of the n observations about k group means

$\underline{n - 1}$ for SST, since it measures the variation of all n observations about the overall mean

- **MS = Mean Square = $\frac{SS}{df}$** :

This is like a standard deviation. Look at the formula we learned back in Chapter 1 for sample standard deviation (p. 51). Its numerator was a sum of squared deviations (just like our SS formulas), and it was divided by the appropriate number of degrees of freedom.

It is interesting to note that another formula for MSE is

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)},$$

which may remind you of the pooled sample estimate for the population variance for 2-sample procedures (when we believe the two populations have the same variance). In fact, the quantity MSE is also called s_p^2 .

- The F statistic = $\frac{MSG}{MSE}$

If the null hypothesis is true, the F statistic has an F distribution with $k - 1$ and $n - k$ degrees of freedom in the numerator/denominator respectively. If the alternate hypothesis is true, then F tends to be large. We reject H_0 in favor of H_a if the F statistic is sufficiently large.

As with other hypothesis tests, we determine whether the F statistic is large by finding a corresponding P -value. For this, we use Table E. Since the alternative hypothesis is always the same (no 1-sided vs. 2-sided distinction), the test is single-tailed (like the chi-squared test). Nevertheless, to read the correct P -value from the table requires knowledge of the number of degrees of freedom associated with both the numerator (MSG) and denominator (MSE) of the F -value.

Look at Table E. On the top are the numerator df, and down the left side are the denominator df. In the table are the F values, and the P -values (the probability of getting an F statistic larger than that if the null hypothesis is true) are down the left side.

Example: Determine

$$P(F_{3,6} > 9.78) = 0.01$$

$$P(F_{2,20} > 5) = \text{between } 0.01 \text{ and } 0.025$$

Example: A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee.

Program 1	Program 2	Program 3	Program 4
9	10	12	9
12	6	14	8
14	9	11	11
11	9	13	7
13	10	11	8

The data was entered into *Stata* which produced the following ANOVA table.

Source	Analysis of Variance			F	Prob > F
	SS	df	MS		
Between groups	54.95	3	18.3166667	7.04	XXXXXX
Within groups	41.6	16	2.6		
Total	96.55	19	5.08157895		

Bartlett's test for equal variances: $\text{chi2}(3) = 0.5685$ Prob>chi2 = 0.904

The X-ed out P -value is between 0.001 and 0.01.

Stata gives us a line at the bottom—the one about Bartlett’s test—which reinforces our belief that the variances are the same for all groups.

Example: In an experiment to investigate the performance of six different types of spark plugs intended for use on a two-stroke motorcycle, ten plugs of each brand were tested and the number of driving miles (at a constant speed) until plug failure was recorded. A partial ANOVA table appears below. Fill in the missing values:

Source	df	SS	MS	F
Brand	<u>5</u>	<u>55961.5</u>	<u>11192.3</u>	<u>2.3744</u>
Error	<u>54</u>	<u>254539.26</u>	4,713.69	
Total	<u>59</u>	310,500.76		

Note: One more thing you will often find on an ANOVA table is R^2 (the *coefficient of determination*). It indicates the ratio of the variability between group means in the sample to the overall sample variability, meaning that it has a similar interpretation to that for R^2 in linear regression.

2.4 Multiple Comparisons

As with other tests of significance, one-way ANOVA has the following steps:

1. State the hypotheses (see Section 1.2)
2. Compute a test statistic (here it is $F_{\text{df numer.}, \text{df denom.}}$), and use it to determine a probability of getting a sample as extreme or more so under the null hypothesis.
3. Apply a decision rule: At the α level of significance, reject H_0 if $P(F_{k-1, n-k} > F_{\text{computed}}) < \alpha$. Do not reject H_0 if $P > \alpha$.

If $P > \alpha$, then we have no reason to reject the null hypothesis. We state this as our conclusion along with the relevant information (F -value, df -numerator, df -denominator, P -value). Ideally, a person conducting the study will have some preconceived hypotheses (more specialized than the H_0 , H_a we stated for ANOVA, and ones which she held before ever collecting/looking at the data) about the group means that she wishes to investigate. When this is the case, she may go ahead and explore them (even if ANOVA did not indicate an overall difference in group means), often employing the method of *contrasts*. We will not learn this method as a class, but if you wish to know more, some information is given on pp. 762–769.

When we have no such preconceived leanings about the group means, it is, generally speaking, inappropriate to continue searching for evidence of a difference in means if our F -value from ANOVA was not significant. If, however, $P < \alpha$, then we know that at least two means are not equal, and the door is open to our trying to determine which ones. In this case, we follow up a significant F -statistic with *pairwise comparisons* of the means, to see which are significantly different from each other.

This involves doing a t -test between each pair of means. This we do using the pooled estimate for the (assumed) common standard deviation of all groups (see the **MS** bullet in Section 2.3):

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s_p \sqrt{1/n_i + 1/n_j}}$$

To determine if this t_{ij} is statistically significant, we could just go to Table D with $n - k$ degrees of freedom (the df associated with s_p). However, depending on the number k of groups, we might be doing many comparisons, and recall that statistical significance can occur simply by chance (that, in fact, is built into the interpretation of the P -value), and it becomes more and more likely to occur as the number of tests we conduct on the same dataset grows. If we are going to conduct many tests, and want the overall probability of rejecting any of the null hypotheses (equal means between group pairs) in the process to be no more than α , then we must adjust the significance level for each individual comparison to be much smaller than α . There are a number of different approaches which have been proposed for choosing the individual-test

significance level so as to get an overall family significance of α , and we need not concern ourselves with the details of such proposals. When software is available that can carry out the details and report the results to us, we are most likely agreeable to using which ever proposal(s) have been incorporated into the software.

The most common method of adjusting for the fact that you are doing multiple comparisons is a method developed by Tukey. *Stata* provides, among others, the Bonferroni approach for pairwise comparisons, which is an approach mentioned in our text, pp. 770-771.

Example: Recall our data from a company looking at training programs for manual tasks carried out by its employees. The results were statistically significant to conclude that not all of the training programs had the same mean result. As a next step, we use Bonferroni multiple comparisons, providing here the results as reported by *Stata*

		Comparison of post-test by program (Bonferroni)		
Row Mean-				
Col Mean		1	2	3
2		-3		
		0.057		
3		.4	3.4	
		1.000	0.025	
4		-3.2	-.2	-3.6
		0.038	1.000	0.017

Where the row labeled '2' meets the column labeled '1', we are told that the sample mean response for Program 2 was 3 lower than the mean response for Program 1 (Row Mean - Col Mean = -3), and that the adjusted Bonferroni probability is 0.057. Thus, this difference is not statistically significant at the 5% level to conclude the mean response from Program 1 is actually different than the mean response from Program 2. Which programs have statistically significant (at significance level 5%) mean responses?

Program 3 is different from program 2, with program 3 apparently better.
 Program 4 is different from program 1, with program 1 apparently better.
 Program 4 is different from program 3, with program 3 apparently better.

Apparently, programs 1 and 3 are the most successful, with no statistically-significant difference between them. At this stage, other factors, such as how much it will cost the company to implement the two programs, may be used to determine which program will be set in place.

Note: Sometimes instead of giving P -values, a software package will generate P confidence intervals for the differences between means. Just remember that if the CI includes 0, there is no statistically significant difference between the means.

3 Two-Way ANOVA

Two-way ANOVA allows to compare population means when the populations are classified according to two (categorical) factors.

Example. We might like to look at SAT scores of students who are male or female (first factor) and either have or have not had a preparatory course (second factor).

Example. A researcher wants to investigate the effects of the amounts of calcium and magnesium in a rat's diet on the rat's blood pressure. Diets including high, medium and low amounts of each mineral (but otherwise identical) will be fed to the rats. And after a specified time on the diet, the blood pressure will be

measured. Notice that the design includes nine different treatments because there are three levels to each of the two factors.

Comparisons of Two-way to One-Factor-at-a-Time

- usually have a smaller total sample size, since you're studying two things at once [*rat diet example, p. 800*]
- removes some of the random variability (some of the random variability is now explained by the second factor, so you can more easily find significant differences)
- we can look at **interactions** between factors (a significant interaction means the effect of one variable changes depending on the level of the other factor).

Examples of (potential) interaction.

- Radon (high/medium/low) and smoking.
High radon levels increase the rate of lung cancer somewhat. Smoking increases the risk of lung cancer. But if you are exposed to radon and smoke, then your lung cancer rates skyrocket. Therefore, the effect of radon on lung cancer rates is small for non-smokers but big for smokers. We can't talk about the effect of radon without talking about whether or not the person is a smoker.
- age of person (0-10, 11-20, 21+) and effect of pesticides (low/high)
- gender and effect of different legal drugs (different standard doses)

Two-way ANOVA table

Below is the outline of a two-way ANOVA table, with factors A and B, having I and J groups, respectively.

<u>Source</u>	<u>df</u>	<u>SS</u>	<u>MS</u>	<u>F</u>	<u>p-value</u>
A	$I - 1$	SSA	MSA	MSA/MSE	
B	$J - 1$	SSB	MSB	MSB/MSE	
$A \times B$	$(I - 1)(J - 1)$	SSAB	MSAB	MSAB/MSE	
Error	$n - IJ$	SSE	MSE		
Total	$n - 1$	SST			

The general layout of the ANOVA table should be familiar to us from the ANOVA tables we have seen for regression and one-way ANOVA. Notice that this time we are dividing the variation into four components:

1. the variation explained by factor A
2. the variation explained by factor B
3. the variation explained by the interaction of A and B
4. the variation explained by randomness

Since there are three different values of F , we must be doing three different hypothesis tests at once. We'll get to the hypotheses of these tests shortly.

The Two-way ANOVA model

The model for two-way ANOVA is that each of the IJ groups has a normal distribution with potentially different means (μ_{ij}), but with a common standard deviation (σ). That is,

$$x_{ijk} = \underbrace{\mu_{ij}}_{\text{group mean}} + \underbrace{\epsilon_{ijk}}_{\text{residual}}, \quad \text{where } \epsilon_{ijk} \sim N(0, \sigma)$$

As usual, we will use two-way ANOVA provided it is reasonable to assume normal group distributions and the ratio of the largest group standard deviation to the smallest group standard deviation is at most 2.

Main Effects

Example. We consider whether the classifying by diagnosis (anxiety, depression, DCFS/Court referred) and prior abuse (yes/no) is related to mean BC (Being Cautious) score. Below is a table where each cell contains the mean BC score for people who were in that group.

Diagnosis	Abused	Not abused	Mean
Anxiety	24.7	18.2	21.2
Depression	27.7	23.7	26.6
DCFS/Court Referred	29.8	16.4	20.8
Mean	27.1	19.4	

Here is the ANOVA table:

	df	SS	MS	F	p-value
Diagnosis	2	222.3	111.15	2.33	.11
Ever abused	1	819.06	819.06	17.2	.0001*
D * E	2	165.2	82.60	1.73	.186
Error	62	2958.0	47.71		
Total	67				

The table has three P -values, corresponding to three tests of significance:

- I. H_0 : The mean BC score is the same for each of the three diagnoses.
 H_a : The mean BC score is not the same for all three diagnoses.
The evidence here is not significant to reject the null hypothesis. ($F = 2.33, df_1 = 2, df_2 = 62, P = 0.11$)
- II. H_0 : There is no main effect due to ever being abused.
 H_a : There is a main effect due to being abused.
The evidence is significant to conclude a main effect exists. ($F = 17.2, df_1 = 1, df_2 = 62, P = 0.0001$)
- III. H_0 : There is no interaction effect between diagnosis and ever being abused.
 H_a : There is an interaction effect between the two variables.
The evidence here is not significant to reject the null hypothesis. ($F = 1.73, df_1 = 2, df_2 = 62, P = 0.186$)

When a main effect has been found for just one variable without variable interactions, we might combine data across diagnoses and perform one of the other tests we know that is applicable. (Two-sample t or One-way ANOVA are both options here, since the combining of information leaves us with just two groups.) But we might also perform a simpler task: Draw a plot of the main effects due to abuse.

Interaction Effects

Example. We consider whether the mean BSI (Belonging/Social Interest) is the same after classifying people on the basis of whether or not they've been abused and diagnosis.

Diagnosis	Abused	Not abused	Mean
Anxiety	27.0	26.8	29.7
Depression	27.3	31.7	32.4
DCFS/Court Referred	23.0	37.1	26.9
Mean	26.6	31.7	

ANOVA table:

	df	SS	MS	F	p-value
Diagnosis	2	118.0	59.0	1.89	.1602
Ever abused	1	483.6	483.6	15.5	.0002*
D X E	2	387.0	193.5	6.19	.0035*
Error	62	1938.12	31.26		
Total	67				

Since the interaction is significant, let's look at the individual mean BSI at each level of diagnosis on an interaction plot.

Knowing how many people were reflected in each category (information that is not provided here) would allow us to conduct 2-sample t tests at each level of diagnosis. Such tests reveal that there is a significant difference in mean BSI between those who have ever been abused and those not abused only for those who have a DCFS/Court Referred disorder. There is no statistically significant difference between these two groups for those with a Depressive or Anxiety disorder (though it's pretty close for those with an Anxiety disorder).

Example. Promotional fliers. [Exercise 13.15, p. 821 in Moore/McCabe]

Means:

		discount			
promos		10	20	30	40
1		4.423	4.225	4.689	4.920
3		4.284	4.097	4.524	4.756
5		4.058	3.890	4.251	4.393
7		3.780	3.760	4.094	4.269

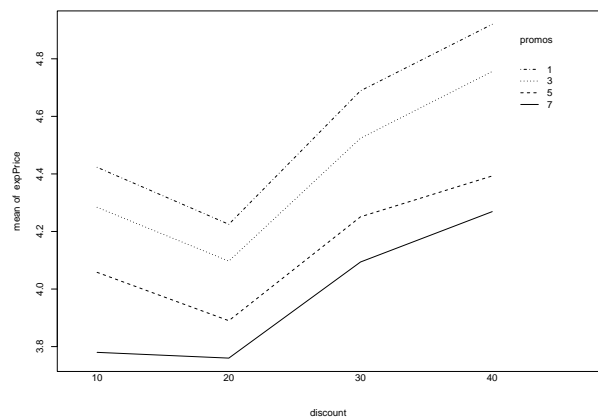
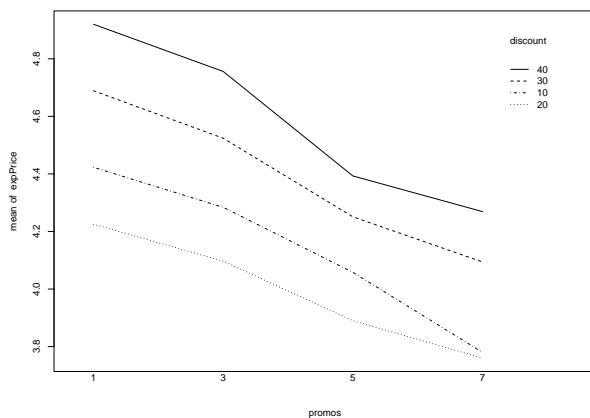
Standard Deviations:

		discount			
promos		10	20	30	40
1		0.18476	0.38561	0.23307	0.15202
3		0.20403	0.23462	0.27073	0.24291
5		0.17599	0.16289	0.26485	0.26854
7		0.21437	0.26179	0.24075	0.26992

Counts:

		discount			
promos		10	20	30	40
1		10	10	10	10
3		10	10	10	10
5		10	10	10	10
7		10	10	10	10

	Sum Sq	Df	F value	Pr(>F)
promos	8.36	3	47.73	<2e-16
discount	8.31	3	47.42	<2e-16
promos:discount	0.23	9	0.44	0.91
Residuals	8.41	144		



Question: Was it worth plotting the interaction effects, or would we have learned the same things plotting only the main effect?