ANOVA: Comparing More Than Two Means

10.1 ANOVA: The Completely Randomized Design

Elements of a Designed Experiment

Before we begin any calculations, we need to discuss some terminology. To make this easier let's consider an example of a designed experiment: Suppose we are trying to determine what combination of sun exposure (4hrs, 6hrs, or 8hrs), fertilizer (A, B, or C), and water (3 tpw, 5 tpw, or 7 tpw) will produce the largest Milkweed plant. We can place 27 identical plants in a greenhouse and control the amount of water, fertilizer, and sun given to each plant. Every plant will get exactly one of each of the three options for each item above. This will produce 27 unique combinations. We will weigh the plants at the end of the experiment to see which combination produced the heaviest plant.

Definitions:

The **response variable** or ***dependent variable*** is the variable of interest to be measured.

In our example, this would be the plant weight.

Factors are those variables whose effect on the response is of interest to the experimenter.

In our example, they would be the sun exposure, fertilizer, and water.

Quantitative Factors are measured on a numerical scale.

In our example, water and sun exposure would be quantitative factors.

Qualitative Factors are not naturally measured on a quantitative scale.

In our example, fertilizer is a qualitative factor.

Factor Levels are the values of the factor utilized in the experiment.

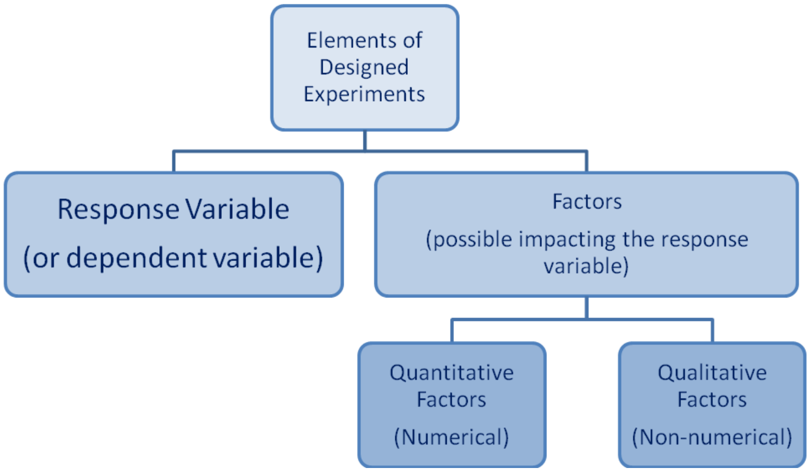
The factors plus their levels for our example are sun exposure (4hrs, 6hrs, or 8hrs), fertilizer (A, B, or C), and water (3 tpw, 5 tpw, or 7 tpw).

The **treatments** are the factor-level combinations utilized.

In our example, one possible treatment would be 4 hrs of sun, fertilizer A, and watering 3 times per week.

An **experimental unit** is the object on which the response and factors are observed or measured.

The plants are the experimental units in our example.



Elements of Designed Experiments

Response Variable (or dependent variable)

Factors (possible impacting the response variable)

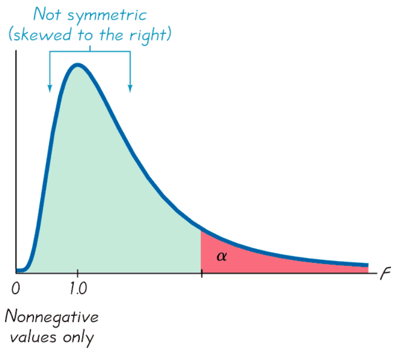
Quantitative Factors (Numerical)

Qualitative Factors (Non-numerical)

The next section will introduce the method of **analysis of variance** (ANOVA), which is used for tests of hypotheses that three or more population means are all equal.

❖ **For example:** $\boldsymbol{H_0: \mu_1 = \mu_2 = \mu_3 = . . . \mu_k}$ **vs.** $\boldsymbol{H_1}$**: At least one mean is different**

The ANOVA methods require use of the *F*-Distribution



The Completely Randomized Design (CRD)

A **Completely Randomized Design** ***(CRD)*** is a design for which the treatments are randomly assigned to the experimental units.

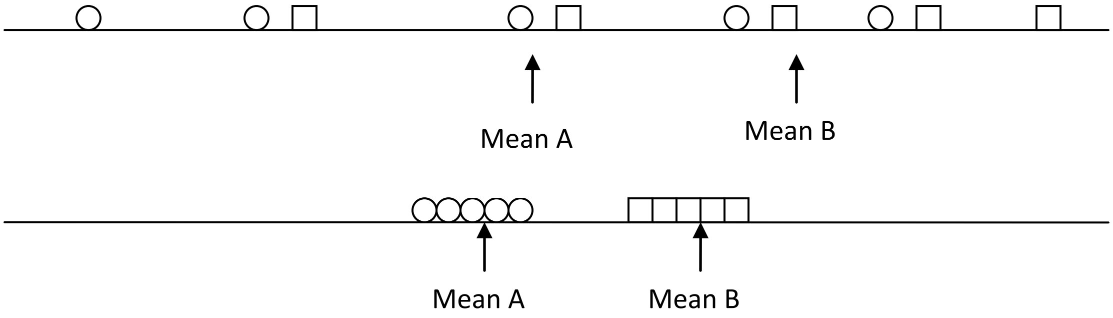
Example: If we wanted to compare three different pain medications: Aleve, Bayer Aspirin, and Children's Tylenol, we could design a study using 18 people suffering from minor joint pain. If we randomly assign the three pain medications to the 18 people so that each drug is used on 6 different patients, we have what is called a **Balanced Design** because an equal number of experimental units are assigned to each drug.

The objective of a CRD experiment is usually to compare the k treatment means.

The competing pair of hypotheses will be as follows:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_A : \text{At least two means differ}$$

How can we test the above Null hypothesis? The obvious answer would be to compare the sample means; however, before we compare the sample means we need to consider the amount of sampling variability among the experimental units. If the difference ***between*** the means is small relative to the sampling variability ***within*** the treatments, we will decide not to reject the null hypothesis that the means are equal. If the opposite is true we will tend to reject the null in favor of the alternative hypothesis. Consider the drawings below:



It is easy to see in the second picture that Treatment B has a larger mean than Treatment A, but in the first picture we are not sure if we have observed just a chance occurrence since the sampling variability is so large within treatments.

To conduct our statistical test to compare the means we will then compare the variation ***between*** treatment means to the variation ***within*** the treatment means. If the variation ***between*** is much larger than the variation ***within*** treatments, we will be able to support the alternative hypothesis.

The variation ***between*** is measured by the **Sum of Squares for Treatments** (SST):

$$\mathrm{SST}=\sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{n_i}\left(\overline{X}_i-\overline{X}\right)^2$$

The variation ***within*** is measured by the **Sum of Squares for Error** (SSE):

$$SSE=\sum_{j=1}^{n_1}\left(x_{1j}-\overline{X}_1\right)^2+\sum_{j=1}^{n_2}\left(x_{2j}-\overline{X}_2\right)^2+\ldots+\sum_{j=1}^{n_k}\left(x_{kj}-\overline{X}_k\right)^2$$

From these we will be able to get the **Mean Square for Treatments** (MST) and the **Mean Square for Error** (MSE):

$$MST=\frac{SST}{k-1} \qquad \& \qquad MSE=\frac{SSE}{n-k}$$

It is the two above quantities that we will compare, but as usual we will compare them in the form of a test statistic. The test statistic we will create in this case has an F distribution with (k – 1, n – k)* degrees of freedom:

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

(The F-tables can be found among our course documents)

*Recall, n = number of experimental units total and k = number of treatments.

Since the F statistic is a ratio of two chi-squared random variables divided by their respective degrees of freedom, we can interpret the ratio as follows: If F > 1, the treatment effect is greater than the variation within treatments. If $\mathrm{F}\leq 1$, the treatment effect is not significant. Do you see why this is? The F statistic is a ratio of the two sets of variation between/within. If the within variation is relatively small compared to the between variation, we will have a ratio greater than one. When F is greater than one, the treatment causes more variation in the response variable than what naturally occurs. Thus the treatment is effective.

You might recall that earlier, we conducted a t-test to compare two independent population means, and we pooled their sample variances as an estimate of their (assumed equal) population variances. That t-test and the above F-test we have described are equivalent and will produce the same results when there are just two means being compared.

Let's list what we have discussed so far and talk about the assumptions of this F test:

F-test for a CRD experiment

Hypotheses: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
$H_A : \text{At least two means differ}$

Test Statistic: $F = \dfrac{\dfrac{SST}{k-1}}{\dfrac{SSE}{n-k}}$

Rejection Region: $F > \mathrm{F}_\alpha$ (reject the null when F is larger than some critical value)

*Note the F-test for ANOVA is always right-tailed. Since MST is always the numerator of the fraction which forms our F-statistic, we are looking for a ratio that is large in order to state there is a significant treatment effect.

Conditions required for a Valid ANOVA *F*-Test: Completely Randomized Design

1. The samples are randomly selected in an independent manner from the *k* treatment populations.
2. All *k* sampled populations have distributions that are approximately normal.
3. The *k* population variances are equal.

Example 151: (Part B) In the following study, three kinds of fertilizer were compared. The researchers applied each fertilizer to five orchids. Each of the fifteen orchids was randomly assigned to one of the three fertilizers. At the end of the one year, the heights (in ft) of the plants were recorded. Test the claim at the 1% significance level that at least one of the three fertilizers has a different effect on orchid height.

A	B	C
1.2	1.3	1.5
1.4	1.2	1.5
1.3	1.1	1.6
1.3	1.0	1.7
1.5	0.9	1.6

Steps worked out

($n =$ total number of observations , k = number of treatments ,& T = total for treatment)

Claim: The fertilizers have an effect on plant height.

Hypotheses: $H_0 : \mu_A = \mu_B = \mu_C$
H_A : At least two means differ from each other significantly.

Correction Factor: $CF = \frac{\left(\sum y_i\right)^2}{n} = \frac{(\text{Sum of all observations})^2}{n}$ = 26.934

Sum of Squares Total: $SS(Total) = \sum y_i^2 - CF$ = 27.69 – 26.934 = 0.756

= (Square each observation then add them up) – CF

Sum of Squares for Treatments: $SST = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \cdots + \frac{T_k^2}{n_k} - CF$ = (Square each treatment total, divide by the number of observations in each treatment, and then add those results up) – CF = 0.576

Sum of Square for Error: $SSE = SS(Total) - SST$ = 0.18

Mean Square Treatment: $MST = \dfrac{SST}{k-1}$ = 0.288

Mean Square Error: $MSE = \dfrac{SSE}{n-k}$ = 0.015

Test Statistic: $F = \dfrac{MST}{MSE}$ = 19.2 (This is a very large F-value)

Critical Value: In order to properly form our conclusion we need either a p-value or a critical value. To find the critical value, we will first consider how our F-test stat was formed. It was the ratio of MST and MSE. The MST is on top so its degrees of freedom will be the numerator degrees of freedom and the degree of freedom for MSE will serve as the denominator degree of freedom (we need these quantities for the F-table). Then we need an alpha value. Let's use 1% in this case. Using these values, we will find the following critical value on the F-table: $f_{k-1,n-k,\alpha}$ = $f_{2,12,0.01} = 6.93$ Now if $F = \dfrac{MST}{MSE}$ > $f_{k-1,n-k,\alpha}$, we reject the null.

Conclusion: Reject the Null; there is sufficient evidence to support the claim that at least two fertilizers produce different average orchid heights.

It is common to organize all of these calculations into an easy to read table called the ANOVA table:

Source	Df	SS	MS	F
Treatments	K – 1	SST	MST = SST/k-1	MST/MSE
Error	N – k	SSE	MSE = SSE/n-k	
Total	N – 1	SS(total)		

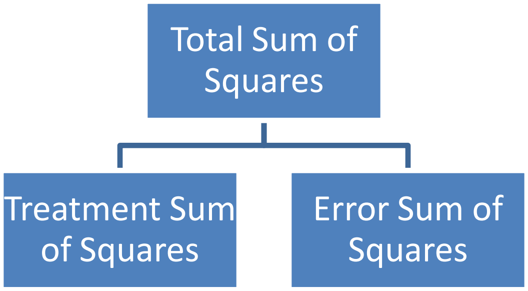
For our example we would have:

Source	Df	SS	MS	F
Treatments	2	0.576	0.288	19.2
Error	12	0.18	0.015	
Total	14	0.756		

Note: The variance within samples, SSE (also called variation due to error) is an estimate of the common population variance σ^2 based on the sample variances.

In our orchid example above, 0.756 is the quantity we call the **Sum of Squares Total.** There is a very useful relationship that we will need to understand in this chapter:

SS(total) = SST + SSE



The required calculations to find the various sum-of-squares values for an ANOVA test can be tedious, so in practice, statisticians turn to software to perform these calculations. Consider the following example:

Example 151 Tech: A group of undergraduate biology majors were randomly assigned to one of four GRE test prep strategies, and their post-prep GRE exam scores for the quantitative reasoning portion of the GRE were recorded. The first group participated in a Kaplan's GRE preparation course. The second group participated in a Test Masters GRE preparation course. The third group enrolled in a self-paced GRE prep course online, and the last group studied on their own without participating in a prep course. All of the students in the study had taken the GRE once before the experiment began. There was no significant difference between their previous GRE scores on the quantitative reasoning section. The GRE scores on the quantitative reasoning portion of the test for the four groups are included below along with part of the output from a statistical software package called SPSS. Fill in the missing parts of the ANOVA table, and answer the set of questions that follow.

Test Preparation Data:

Kaplan	Test Masters	Self-Paced	Control
167	156	149	141
148	150	154	162
162	145	151	164
143	156	145	165
150	167	158	145
147	152	145	161
153	147	155	160

SPSS Output

Dependent Variable: Quantitative Score

Source	Sum of Squares	df	Mean Square	F	Sig.
PrepProgram	125.857		41.952		.574
Error					
Total	1609.000				

Complete the ANOVA table above and answer the following questions:

a) What is the null hypothesis this particular ANOVA procedure is testing?
b) What is the p-value for this test?
c) What is the decision regarding the null hypothesis?
d) Based on the results of this experiment, do prep courses help Biology majors improve their quantitative GRE scores?
e) Based on the design of this experiment, is it possible to conclude that test prep is not helpful for this population of students when attempting to improve their score on the quantitative section of the GRE?

Look at the example below in Figure 1 to see what has to happen for us to be able to reject the null hypothesis (that all treatment means are equal). Notice how in case B, we change sample 1 so that it now has a much larger mean than the other samples. Look what that does to the F statistic. Also, notice the very small p-value which is a result of the large F-value. Remember that in a right-tailed test like this, a large test stat produces a small p-value which implies rejection of the null hypothesis.

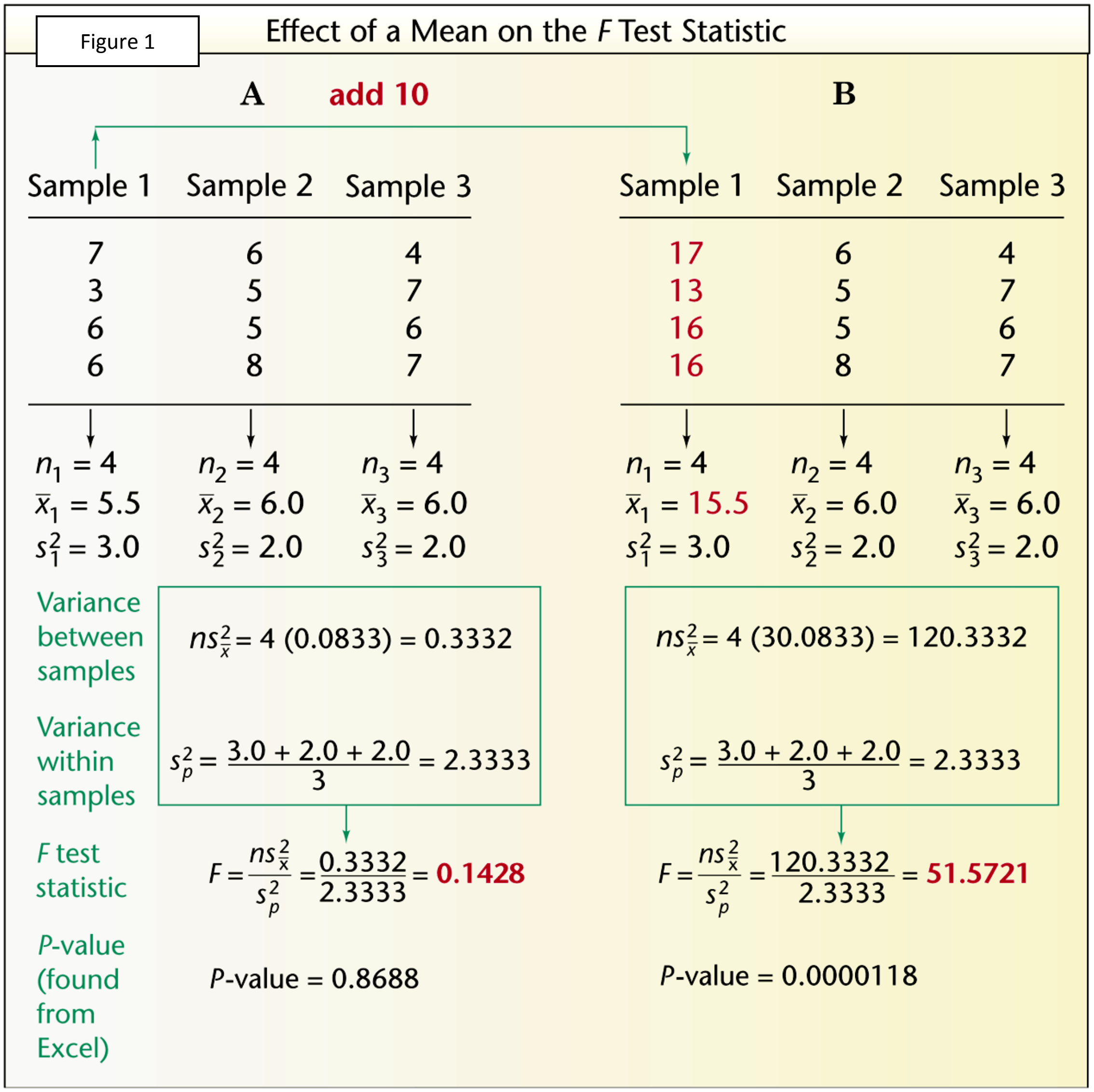


Figure 1

Effect of a Mean on the *F* Test Statistic

A add 10 **B**

Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
7	6	4	17	6	4
3	5	7	13	5	7
6	5	6	16	5	6
6	8	7	16	8	7
$n_1 = 4$	$n_2 = 4$	$n_3 = 4$	$n_1 = 4$	$n_2 = 4$	$n_3 = 4$
$\bar{x}_1 = 5.5$	$\bar{x}_2 = 6.0$	$\bar{x}_3 = 6.0$	$\bar{x}_1 = 15.5$	$\bar{x}_2 = 6.0$	$\bar{x}_3 = 6.0$
$s_1^2 = 3.0$	$s_2^2 = 2.0$	$s_3^2 = 2.0$	$s_1^2 = 3.0$	$s_2^2 = 2.0$	$s_3^2 = 2.0$

	A	B
Variance between samples	$ns_{\bar{x}}^2 = 4\ (0.0833) = 0.3332$	$ns_{\bar{x}}^2 = 4\ (30.0833) = 120.3332$
Variance within samples	$s_p^2 = \dfrac{3.0 + 2.0 + 2.0}{3} = 2.3333$	$s_p^2 = \dfrac{3.0 + 2.0 + 2.0}{3} = 2.3333$
F test statistic	$F = \dfrac{ns_{\bar{x}}^2}{s_p^2} = \dfrac{0.3332}{2.3333} = \mathbf{0.1428}$	$F = \dfrac{ns_{\bar{x}}^2}{s_p^2} = \dfrac{120.3332}{2.3333} = \mathbf{51.5721}$
P-value (found from Excel)	*P*-value = 0.8688	*P*-value = 0.0000118

Example 152: A cosmetic company wants to produce silver nitrate for use in its cosmetics. It is interested in knowing the most productive procedure for producing the silver nitrate from dissolved silver. It is believed that stirring of the mixture of silver and nitric acid during the dissolving process has an effect on the yield of silver nitrate crystals produced. To determine the optimal number of revolutions while stirring the company has set up an experiment involving 15 identical samples randomly assigned to one of three stirring scenarios. The yields (in tons) for the three stirring options are shown below. At the 2.5% level of significance test the claim that the number of revolutions while stirring has an effect on silver nitrate yield.

Revs	10rpm	20rpm	30rpm
Yields	3.9	3.2	3.5
	3.6	3.1	3.3
	3.7	3.3	3.2
	3.3	3.3	3.4
	3.8	3.4	3.4
Totals	18.3	16.3	16.8

Example 152 Tech: A study was conducted to determine the factor that reduces LDL cholesterol the most: medication, diet, or exercise. Twenty-seven patients at a hospital with comparable levels of LDL cholesterol are randomly assigned to each treatment group. After eight weeks, the drop in LDL cholesterol for each patient was measured. Fill in the missing parts of the ANOVA table, answer the set of questions that follow, and use a 5% significance level to test the claim that all three of the treatments produce the same drop in LDL cholesterol.

Treatments		
Medication	Diet	Exercise
40	32	20
16	61	18
26	63	27
25	58	33
21	44	36
35	40	24
30	39	21
29	52	15
28	51	19

Dependent Variable: LDL reduction

Source	Sum of Squares	df	Mean Square	F	Sig.
Treatment				22.832	.000
Error	1732.444				
Total	5028.667	26			

Complete the ANOVA table above and answer the following questions:

a) What is the null hypothesis this particular ANOVA procedure is testing?
b) What is the p-value for this test?
c) What is the decision regarding the null hypothesis?
d) Based on the results of this experiment, are all of these treatments equally effective at reducing LDL cholesterol?
e) Based on the result of this experiment, is it possible to determine if there is a significant difference between the average reduction in LDL cholesterol achieved through the use of medication and the reduction achieved by exercise?

Assumptions for CRD (compare these with the assumptions for the independent t-test):

1. The samples are randomly selected in an independent manner from the k treatment populations.
2. All k sampled populations are approximately normal.
3. The k population variances are equal.

Finally, there is one issue we should be concerned with. If we do reject the null hypothesis in an ANOVA CRD test, how do we know which means differ significantly? For example, in the first problem above which fertilizer was the best? Was it C? It seems that C and B must be significantly different, but what about C and A? Are they significantly different? What if C costs twice as much as A, if they are not significantly different we could save money by buying A. We will address this issue in the next section.

10.2 Multiple Comparisons of Means

Multiple Comparisons of the Means

If we had a completely randomized design with three treatments, A, B, and C and we determine that the treatment means are statistically different using the ANOVA F-test, how can we know the proper ranking of the three means? In other words, how can we know which means are statistically different from one another? How can we put them in order according to size?

We want to be able to form all of the possible pairwise comparisons. The number of pairwise comparisons possible is given by the combination formula:

Number of pairwise comparisons = $\binom{k}{2} = \frac{k!}{(k-2)!2!} = \frac{k \cdot (k-1)}{2}$

k = number of treatments

For our cosmetics example from the CRD section above, there were three treatments, so the number of pairwise comparisons would be $\binom{3}{2} = \frac{3!}{2!1!} = \frac{3*2*1}{2*1*1} = \frac{6}{2} = 3$.

Example 153: Suppose a CRD experiment has 6 treatments. How many pairwise comparisons are there?

To make these comparisons, we want to form confidence intervals to estimate the true difference between the treatment means, but we need to be careful about our confidence level for these comparisons. If we want an overall confidence level of 95%, we cannot have each interval have a 95% confidence level. This would lower the overall level to a value below 95%. Let's determine what the overall confidence level would be if each of our intervals were at a 95% confidence level and we had three comparisons to make:

P(All the intervals capture the true difference between the means) = $0.95^3 \approx 0.857$

This means the overall confidence level would be as low as 85.7%. This means we could be only 85.7% confident that all of our intervals capture the true differences between the pairs of means. To avoid this problem, we will use another approach to generate our comparisons.

Several methods exist to tackle this problem. In our class, in order to guarantee that the overall confidence level is 95%, we will use three different methods depending on the type of experiment we run and the kinds of comparisons we wish to make.

Guidelines for selecting a multiple comparison method in ANOVA		
Method	**Treatment Sample Sizes**	**Types of Comparisons**
Tukey	Equal	Pairwise
Bonferroni	Equal or Unequal	Pairwise
Scheffe`	Equal or Unequal	General Contrasts

Note: For equal sample sizes and pairwise comparisons, Tukey's method will yield simultaneous confidence intervals with the smallest width, and the Bonferroni intervals will have smaller widths than the Scheffe` intervals.

The approach needed to create the intervals using each of the above listed methods are given at the end of this section; however, in our class we are going to focus on being able to interpret the results of the multiple comparisons. Consider the following example:

Example 154: Suppose we want to conduct pairwise comparisons for the treatments involved in the following experiment. Which multiple comparison method would you use to produce the shortest possible confidence intervals for the differences between the treatment means?

A	B	C
12.5	13.2	12.3
11.3	13.8	11.5
10.0	14.1	14.8
13.1		15.1

Example 155: If a CRD experiment involving five different antibacterial solutions randomly sprayed onto 30 different plates filled with the common cold virus (each of the five sprays will be used on six different plates), what multiple comparison method would be used if you wanted to make pairwise comparisons?

Example 156: An experiment was conducted to determine if there is a difference between drying times for four different brands of outdoor paints. A multiple comparison procedure was used to compare the different average drying times and produced the following intervals:

$$\mu_A - \mu_B = (3,9)$$
$$\mu_A - \mu_C = (5,7)$$
$$\mu_A - \mu_D = (-15,4)$$
$$\mu_B - \mu_C = (-2,8)$$
$$\mu_B - \mu_D = (-12,-5)$$
$$\mu_C - \mu_D = (-10,-4)$$

Rank the means from smallest to highest. Which means are significantly different?

There is a nice way to summarize the information obtained from the multiple comparison procedure. The diagram shows which means are significantly different by the use of lines and position. The means that are connected by a line are not significantly different from one another.

The proper diagram for the example above would be:

$$\overline{C\quad B}\quad\overline{A\quad D}$$

From the diagram we can see that C & B are not significantly different and A & D are not significantly different, but both A & D are significantly larger than both C & B.

If we determined that the means for treatments (paints) C and B are the smallest (this means these paints are fastest at drying), we might want to form a confidence interval for the mean drying time for these two paints. The formula below is equivalent to our usual t-interval:

$$\overline{X_T} \pm t_{\alpha/2} S\sqrt{1/n_T}$$

Here $S = \sqrt{MSE}$, n_T = number of repetitions for the treatment, and t has degrees of freedom = n – k (the error degrees of freedom).

Example 157: (Part B) The results of a CRD experiment for corn crop yields are given below. The multiple comparison intervals to compare the treatments A, B, C, and D were provided by SPSS. Using the SPSS output below, rank the treatments from lowest to highest, express the results in a diagram using bars to join the means that are not significantly different, and form a confidence interval for the true mean yield (in tons) for the highest of the four treatments.

Treatments	Yields			
A	3.7	3.6	3.5	3.5
B	1.2	3.4	3.1	2.5
C	0.5	1.0	0.7	0.6
D	1.3	3.0	3.3	2.6

Source	DF	SS	MS	F	P-value
Treatments	3	17.212	5.737	12.875	0.0005
Error	12	5.348	0.446		
Total	15	22.56			

Multiple Comparisons

Dependent Variable: Yield
Tukey HSD

(I) Treatments	(J) Treatments	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	1.0250	.47203	.186	-.3764	2.4264
	3.00	2.8750*	.47203	.000	1.4736	4.2764
	4.00	1.0250	.47203	.186	-.3764	2.4264
2.00	1.00	-1.0250	.47203	.186	-2.4264	.3764
	3.00	1.8500*	.47203	.010	.4486	3.2514
	4.00	.0000	.47203	1.000	-1.4014	1.4014
3.00	1.00	-2.8750*	.47203	.000	-4.2764	-1.4736
	2.00	-1.8500*	.47203	.010	-3.2514	-.4486
	4.00	-1.8500*	.47203	.010	-3.2514	-.4486
4.00	1.00	-1.0250	.47203	.186	-2.4264	.3764
	2.00	.0000	.47203	1.000	-1.4014	1.4014
	3.00	1.8500*	.47203	.010	.4486	3.2514

Based on observed means.

*. The mean difference is significant at the .05 level.

Note: The comparisons are repeated above, so only look at the ones that are unique: 12, 13, 14, 23, 24, and 34.

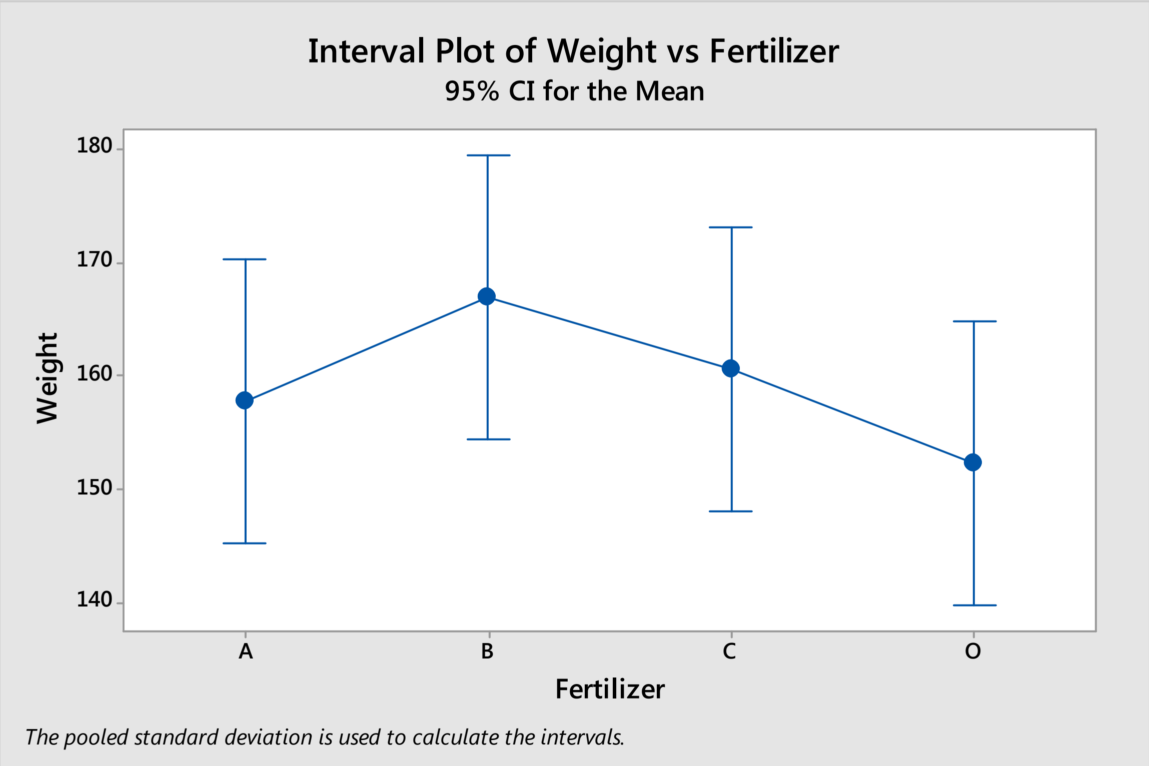
Confidence Interval for a Treatment Mean

$$\overline{X_T} \pm t_{\alpha/2} S \sqrt{1/n_T}$$

Here $S = \sqrt{MSE}$, n_T = number of repetitions for the treatment, and t has degrees of freedom = n – k (the error degrees of freedom).

Example 157.4 Orange trees at a citrus farm near Orlando were randomly assigned to one of three new fertilizers or the traditional fertilizer being used already. The new fertilizers are supposed to produce heavier oranges. Researchers tested the claim that at least one of the new fertilizers produce a heavier orange. The following result were produced using Minitab (a statistical software package). Interpret the output.

```
Null hypothesis         All means are equal
Alternative hypothesis  At least one mean is different
Significance level      α = 0.05

Equal variances were assumed for the analysis.

Factor      Levels  Values
Fertilizer       4  A, B, C, O

Analysis of Variance
Source      DF  Adj SS  Adj MS  F-Value  P-Value
Fertilizer   3    1203   400.9     0.95    0.427
Error       40   16922   423.1
Total       43   18125

Means
Fertilizer   N    Mean  StDev        95% CI
A           11  157.85  17.19  (145.32, 170.39)
B           11  166.91  21.09  (154.38, 179.44)
C           11  160.71  13.34  (148.18, 173.24)
O           11  152.41  27.82  (139.88, 164.94)

Pooled StDev = 20.5682
```



The pooled standard deviation is used to calculate the intervals.

Example 157.5 A study reported in 2010 in the Journal "Archive of Sexual Behavior" looked at sexual practices in healthy Euro-American, Asian, and Hispanic female populations. A portion of the research looked at the age of sexual debut (the age that women first engaged in intercourse)for women in the three study populations. An ANOVA CRD was conducted using SPSS to determine if there was a difference in the age of sexual debut among the three groups of women. The results are reported below. Based on the results below, does there appear to be a significant difference between the groups of women? Based on the results of the ANOVA procedure below, what (if any) multiple comparison procedure should be conducted on the data?

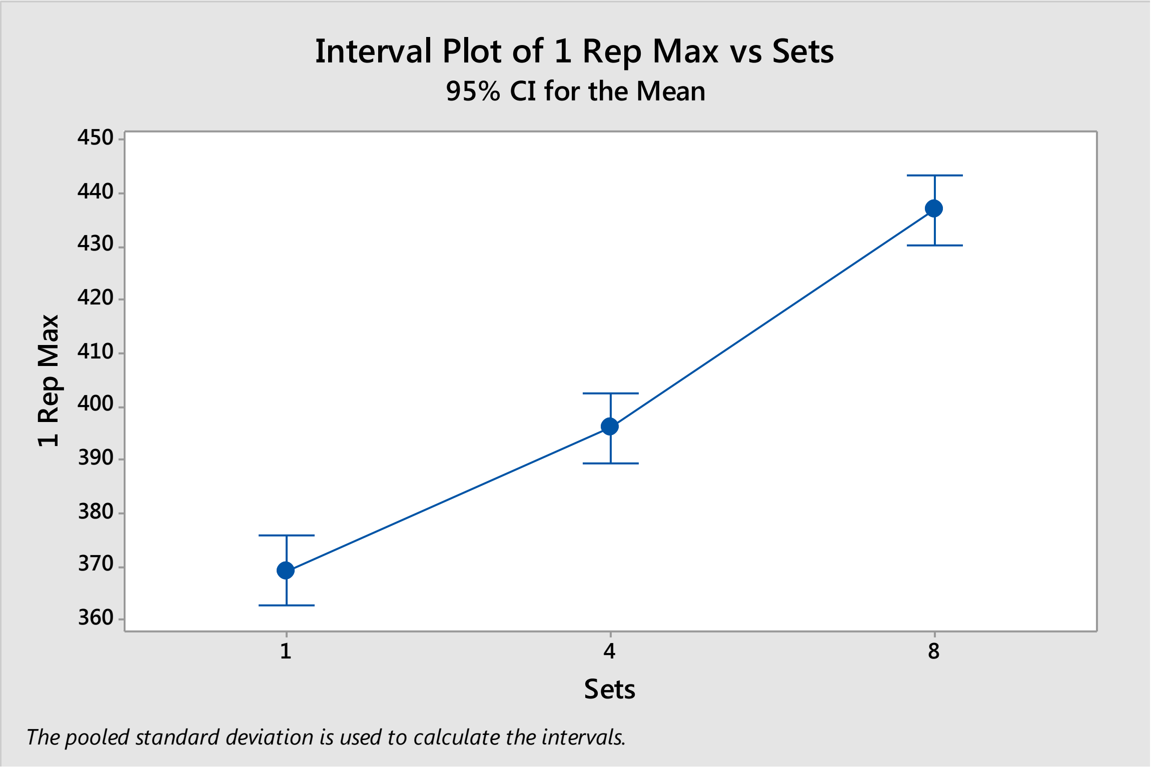
Tests of Between-Subjects Effects

Dependent Variable:Age@firstInter

Source	Sum of Squares	df	Mean Square	F	Sig.
Race	10.953	2	5.476	1.037	.364
Error	221.867	42	5.283		
Corrected Total	232.820	44			

Example 157.6 Exercise researchers compared the effect of different volumes of training on strength gains in the squat exercise. Thirty-three participants performed either one set, four sets, or eight sets of squats. The participants were randomly assigned to each of the groups, and each participant had experience with weight training. The study lasted for 12 weeks, and each participant performed their assigned squat workout twice per week. Participants did not have statistically different one-rep squat maximums before beginning the study. The following result were produced using Minitab (a statistical software package). Interpret the output.

```
Null hypothesis          All means are equal
Alternative hypothesis   At least one mean is different
Significance level       α = 0.05

Equal variances were assumed for the analysis.

Factor  Levels  Values
Sets         3  1, 4, 8

Analysis of Variance
Source  DF  Adj SS   Adj MS  F-Value  P-Value
Sets     2   25566  12782.8   114.05    0.000
Error   30    3363    112.1
Total   32   28928

Means
Sets   N    Mean  StDev       95% CI
1     11  369.22   8.77  (362.70, 375.74)
4     11  395.95  11.40  (389.43, 402.46)
8     11  436.90  11.37  (430.38, 443.42)

Pooled StDev = 10.5869
```



10.3 ANOVA: The Randomized Block Design

The Randomized Block Design

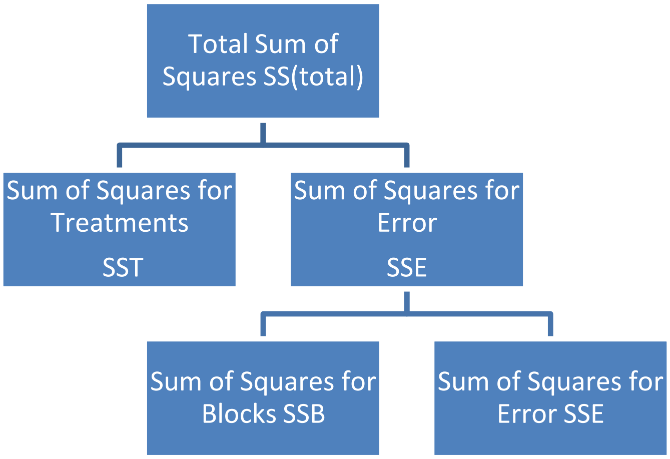
In the matched paired t-test examples earlier, we saw that when the sampling variability is large, we may be unable to reject the null hypothesis (i.e. - detect a difference between the means). Instead of using independent samples of experimental units as in the CRD, the Randomized Block Design uses experimental units in matched sets. We will assign one member from the matched pair sets to each treatment.

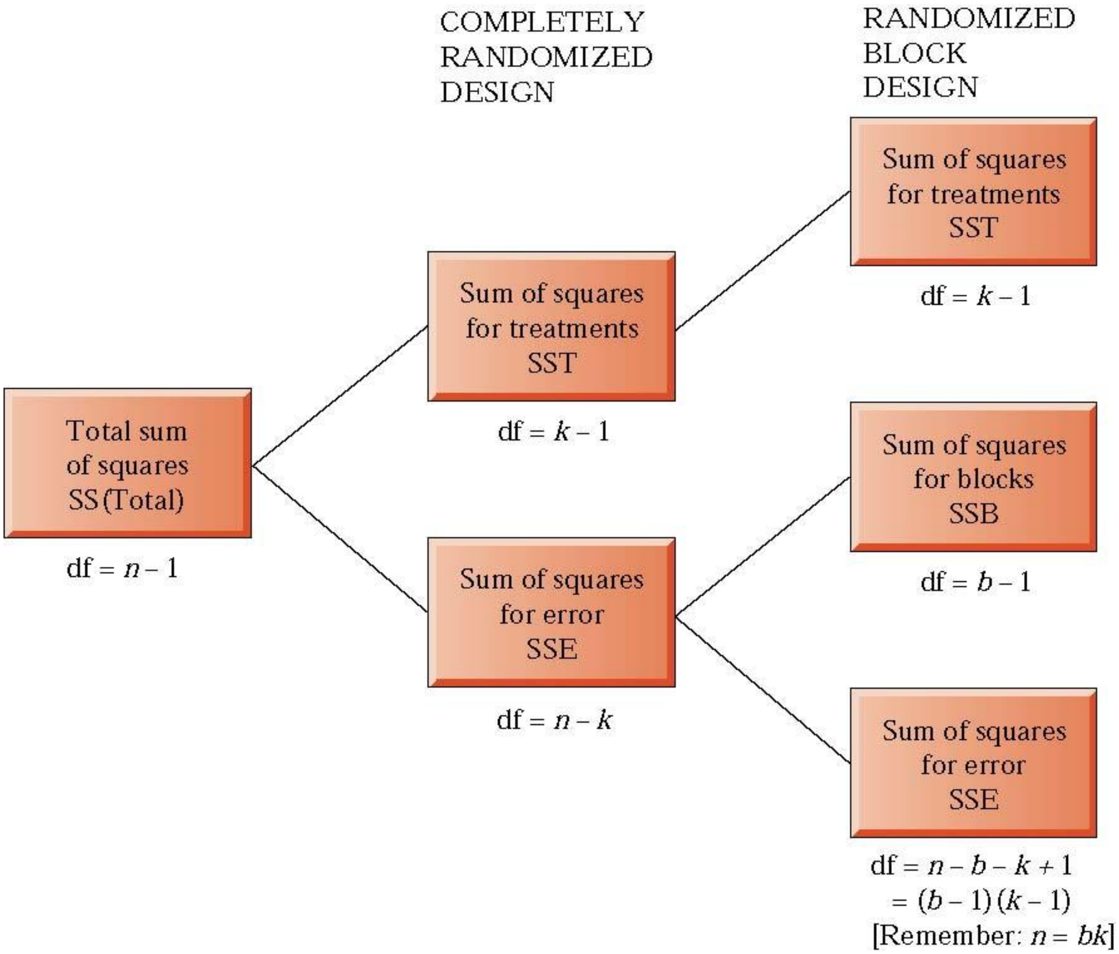
The RBD consists of:

1. Matched sets of experimental units, called **blocks**, are formed consisting of k experimental units. The b blocks should consist of experimental units that are as close as possible.
2. One experimental unit from each block is randomly assigned to each treatment, resulting in a total of n = bk responses.

Note: The matched pair t-test and the below F-test we will describe are equivalent and will produce the same results when there are just two means being compared.

Recall how we partitioned the variability of our sample data when we used CRD, for RBD the partitioning is extended by further partitioning the sum of squares for error as the following two diagrams illustrate:


Total Sum of Squares SS(total)
Sum of Squares for Treatments SST
Sum of Squares for Error SSE
Sum of Squares for Blocks SSB
Sum of Squares for Error SSE


COMPLETELY RANDOMIZED DESIGN
RANDOMIZED BLOCK DESIGN
Total sum of squares SS(Total)
df = $n - 1$
Sum of squares for treatments SST
df = $k - 1$
Sum of squares for error SSE
df = $n - k$
Sum of squares for treatments SST
df = $k - 1$
Sum of squares for blocks SSB
df = $b - 1$
Sum of squares for error SSE
df = $n - b - k + 1$
= $(b - 1)(k - 1)$
[Remember: $n = bk$]

Example 158: (Part B) Four fertilizers (treatments: A, B, C, & D) were compared and three watering schemes (blocks: 3 times per week, 5tpw, & 7tpw) using a RBD experiment. The results (crop yields in kilograms) are shown below. Assume there is no interaction between type of fertilizer and the different watering schemes and use a 5% significance level to test if the fertilizers all produce the same yields.

	A	B	C	D	**Totals**
3 tpw	7	8	9	7	**31**
5 tpw	14	15	15	12	**56**
7 tpw	13	10	11	9	**43**
Totals	**34**	**33**	**35**	**28**	**130**

As with CRD, we will want to compare the MS for treatments to the MS for error, but we will make another comparison as well. We will also compare the MS for blocks to the MS for error, and as before, we will place the answers to our calculations in an ANOVA table:

Source	Df	SS	MS	F
Treatment	K – 1	SST	MST	MST/MSE
Block	B – 1	SSB	MSB	MSB/MSE
Error	N – k – b + 1	SSE	MSE	
Total	N – 1	SS(total)		

The formulas to find these quantities are given below:

Correction Factor: $CF = \frac{\left(\sum y_i\right)^2}{n} = \frac{\left(\text{Sum of all observations}\right)^2}{n}$ = 1408.33

Sum of Squares Total: $SS(Total) = \sum y_i^2 - CF$

= (Square each observation then add them up) – CF = 95.67

Sum of Squares for Treatments: $SST = \frac{T_1^2}{b} + \frac{T_2^2}{b} + \cdots + \frac{T_k^2}{b} - CF$ = (Sum of squared Treatment totals divided by the number of observations in each treatment) – CF = 9.67

Sum of Squares for Blocks: $SSB = \frac{B_1^2}{k} + \frac{B_2^2}{k} + \cdots + \frac{B_k^2}{k} - CF$ = (Sum of squared block totals divided by the number of observations in each block) – CF = 78.17

Sum of Square for Error: $SSE = SS(Total) - SST - SSB$ = 7.83

Mean Square Treatment: $MST = \frac{SST}{k-1}$ = 3.22

Mean Square for Blocks: $MSB = \frac{SSB}{b-1}$ = 39.09

Mean Square Error: $MSE = \frac{SSE}{n-k-b+1}$ = 1.305

Test Statistics: $F = \frac{MST}{MSE}$ = 2.47 & $F = \frac{MSB}{MSE}$ = 29.95

Now we can fill in that ANOVA table:

Source	Df	SS	MS	F
Treatment	3	9.67	3.22	2.47
Block	2	78.17	39.09	29.95
Error	6	7.83	1.305	
Total	11	95.67		

Now we are ready to run our tests…

Our pair of Competing Hypotheses will be:

$$\begin{aligned} &H_0: \mu_1 = \mu_2 = \ldots = \mu_k \\ &H_A: \text{At least two treatment means differ} \end{aligned}$$

and

$$\begin{aligned} &H_0: \mu_1 = \mu_2 = \ldots = \mu_b \\ &H_A: \text{At least two block means differ} \end{aligned}$$

To test the above claims at the 5% significance level, we need to compare our test stats to our critical values:

$$\begin{aligned} &F_{treatments} = 2.47 < F_{3,6,0.05} = 4.76 \\ &F_{blocks} = 29.95 > F_{2,6,0.05} = 5.14 \end{aligned}$$

From the above comparisons, we can conclude that there is a significant difference between blocks, but not a significant difference between treatments at the 5% significance level.

Example 158 Tech: Cardiac stents are used to open a closed artery in order to reduce the risks of a heart attack in at risk patients. Medical researchers are interested in determining if the length of time spent in the hospital after a patient undergoes the implantation of a cardiac stent is affected by the patient’s view from the hospital bed. The researchers randomly assigned 90 cardiac stent patients to one of three different types of hospital rooms for their recovery: a room with a view (V) of nature (trees, grass, and open sky), a room with pictures (P) of nature on the walls but no window view, or a room without a view (N) and without images of nature on the walls. To block out the effect of different surgeons, the researchers used patients from three different surgeons. The response variable of interest for the study was the length of hospital stay. Complete the ANOVA table below and use the results to answer the following questions:

Tests of Between-Subjects Effects

Dependent Variable: Length Of Stay

Source	Sum of Squares	df	Mean Square	F	Sig.
Room	29.042				.000
Surgeon	1.400				.042
Error		85			
Corrected Total	48.496				

Multiple Comparisons

Length Of Stay

(I) Room	(J) Room	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
N	P	1.2000*	.11900	.000	.9161	1.4839
	V	1.2100*	.11900	.000	.9261	1.4939
P						
	V	.0100	.11900	.996	-.2739	.2939

Based on observed means.

*. The mean difference is significant at the 0.05 level.

a) Which variable represents the treatments in this experiment?
b) Which variable represents the blocks in this experiment?
c) What is the null hypothesis for a test of the treatment effect?
d) What is the p-value for the test of the treatment effect?
e) What is the decision regarding the null hypothesis for the test of the treatment effect?
f) Based on the results of this experiment, do people in these different room types all have the same average length of hospital stay?
g) At a 5% significance level, is the length of hospital stay affected by the choice of surgeon?
h) Use the results of the multiple comparison procedure, included with the SPSS output, to construct a diagram that ranks the means for our different room types from small to large.
i) Summarize your conclusions for this ANOVA RBD experiment. Be sure to state which means are significantly different from each other, if any.

Finally, a note on the assumptions for RBD:

Assumptions for RBD (compare these assumptions with those for the matched pair t-test):

1. The b blocks are randomly selected and all k treatments are applied in random order to each block.
2. The distributions of observations corresponding to all bk block-treatment combinations are approximately normal.
3. The bk block-treatment distributions have equal variances.

Example 159: Four methods of blending penicillin were compared in a randomized block design. The blocks are blends of the raw material. Construct the ANOVA table. Are there differences between the methods? the blends? Use a 10% significance level for both.

	Method				Totals
Blend	A	B	C	D	
1	89	88	97	94	368
2	84	77	92	79	332
3	81	87	87	85	340
4	87	92	89	84	352
5	79	81	80	88	328
Totals	420	425	445	430	1720

Example 159 (Tech): Four methods of blending penicillin were compared in a randomized block design. The blocks are blends of the raw material. Below the table of yields, a partial ANOVA table has been provided using the statistical software called Minitab. Complete the ANOVA table and test for differences between the methods and the blends. Use a 10% significance level for both tests.

	Method				Totals
Blend	A	B	C	D	
1	89	88	97	94	368
2	84	77	92	79	332
3	81	87	87	85	340
4	87	92	89	84	352
5	79	81	80	88	328
Totals	420	425	445	430	1720

```
Factor Information

  Factor    Levels     Values
Method          4    A, B, C, D
Blend           5    1, 2, 3, 4, 5

Analysis of Variance

Source             DF    Adj SS    Adj MS    F-Value    P-Value
   Method           ?        70         ?          ?     0.3387

   Blend            ?       264         ?          ?     0.0407
Error               ?                   ?
Total               ?       560
```

10.4 ANOVA: Factorial Experiments

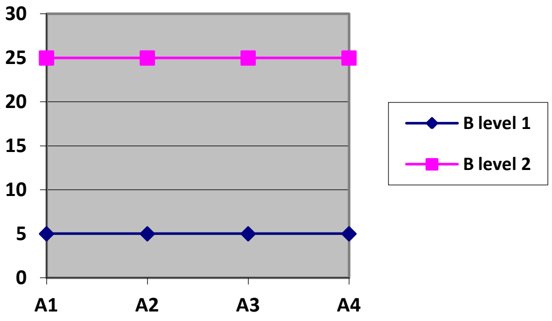
In the first section of this chapter, we only considered experiments involving a single factor that could possibly affect the response variable. In the previous section, we considered a randomized block design experiment, which involved two factors that could possibly influence the response variable. However, in that section, we assumed that the two factors did not interact with each other. In this section, we will consider the scenario where there are two factors which may interact with each other.

If an experiment involves two or more factors and those factors can potentially interact with each other, the experiment is referred to as a full factorial experiment (or a complete factorial experiment). When analyzing the results of a full factorial experiment, we need to consider all possible treatments that can be formed by combining the different levels of each of the factors in all possible combinations. This can produce many possible treatments, so to keep things manageable, we will only look at two-factor factorial experiments.

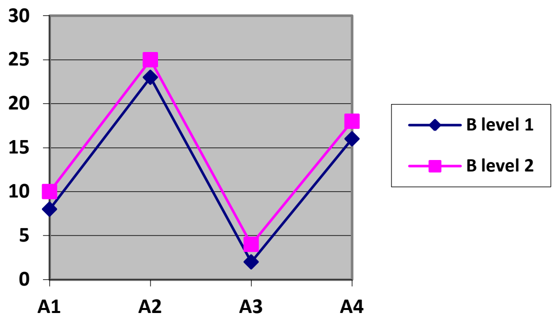
We will assume that all of the experimental units are assigned to the different possible treatments in a completely random and independent manner. This will allow us to initially analyze the data using the same approach we used for the single-factor completely randomized design experiments, but if we reject the null hypothesis, we will carry our analysis further to better determine what caused the rejection of the null hypothesis.

Let's consider a few diagrams to see what sort of results can occur during a two-factor factorial experiment:

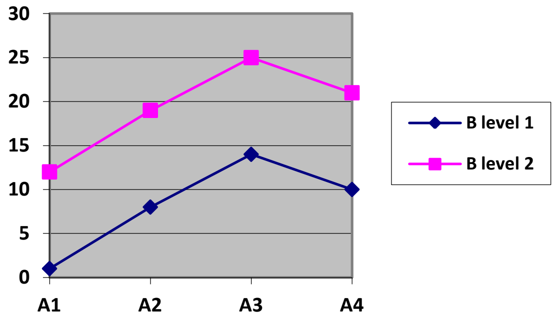
Case 1: In this scenario, you can see that there is a difference between the levels of factor B. We can then assume there is a main effect due to factor B, which means that B has an effect on the response variable. You can also see that varying the different levels of A produces no change in the response variable.



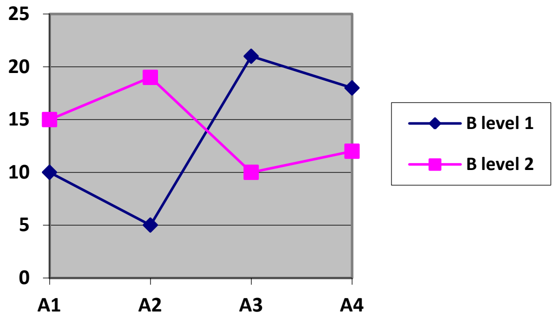
Case 2: In this scenario, you can see that there is a difference between the levels of factor A. We can then assume there is a main effect due to factor A, which means that A has an effect on the response variable. You can also see that varying the different levels of B produces no real change in the response variable.



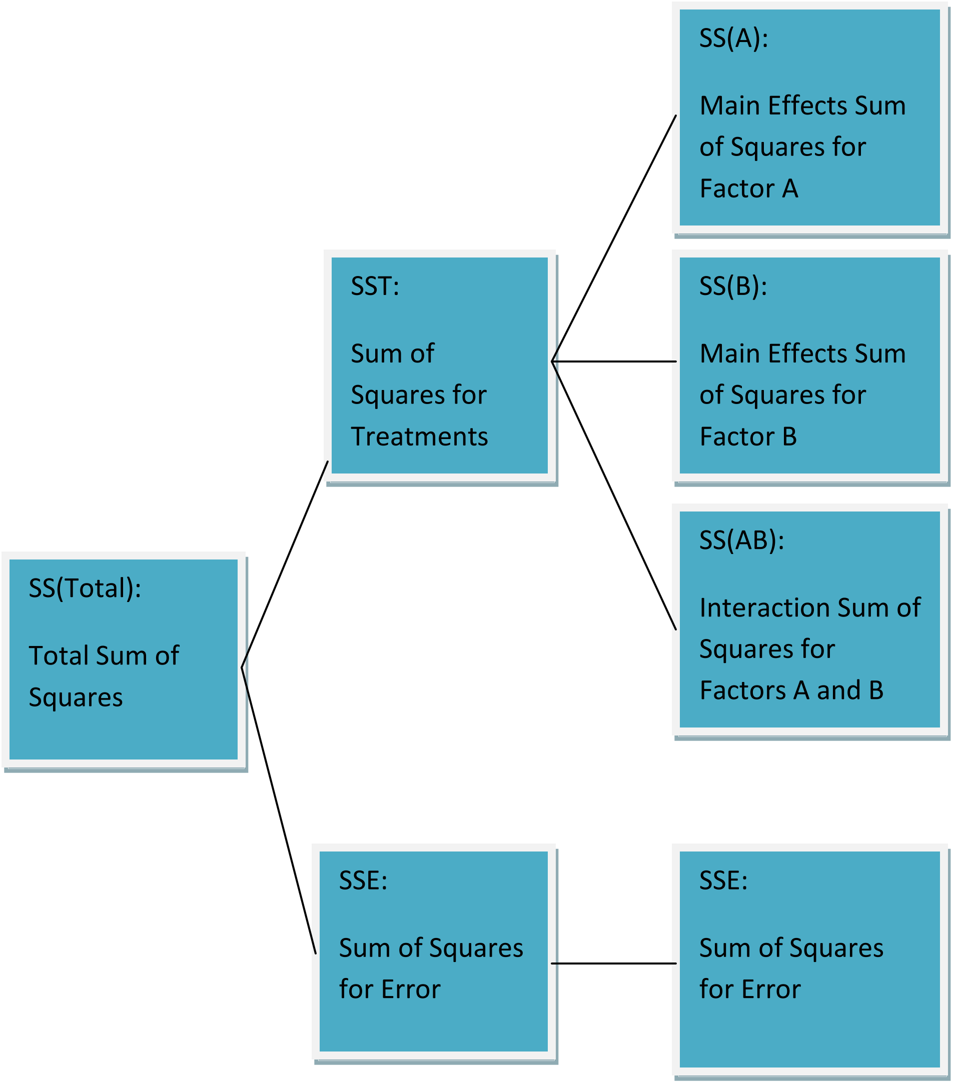
Case 3: In this scenario, you can see that varying the levels of factor A produces a change in the response variable, and you can see that varying the levels of factor B also produces a change in the response variable. Because the lines in the graph are parallel to each other, we do not have evidence of an interaction effect.



Case 4: In this scenario, you can see that factors A and B interact with each other. This is indicated by the fact that the two lines in the diagram intersect each other.

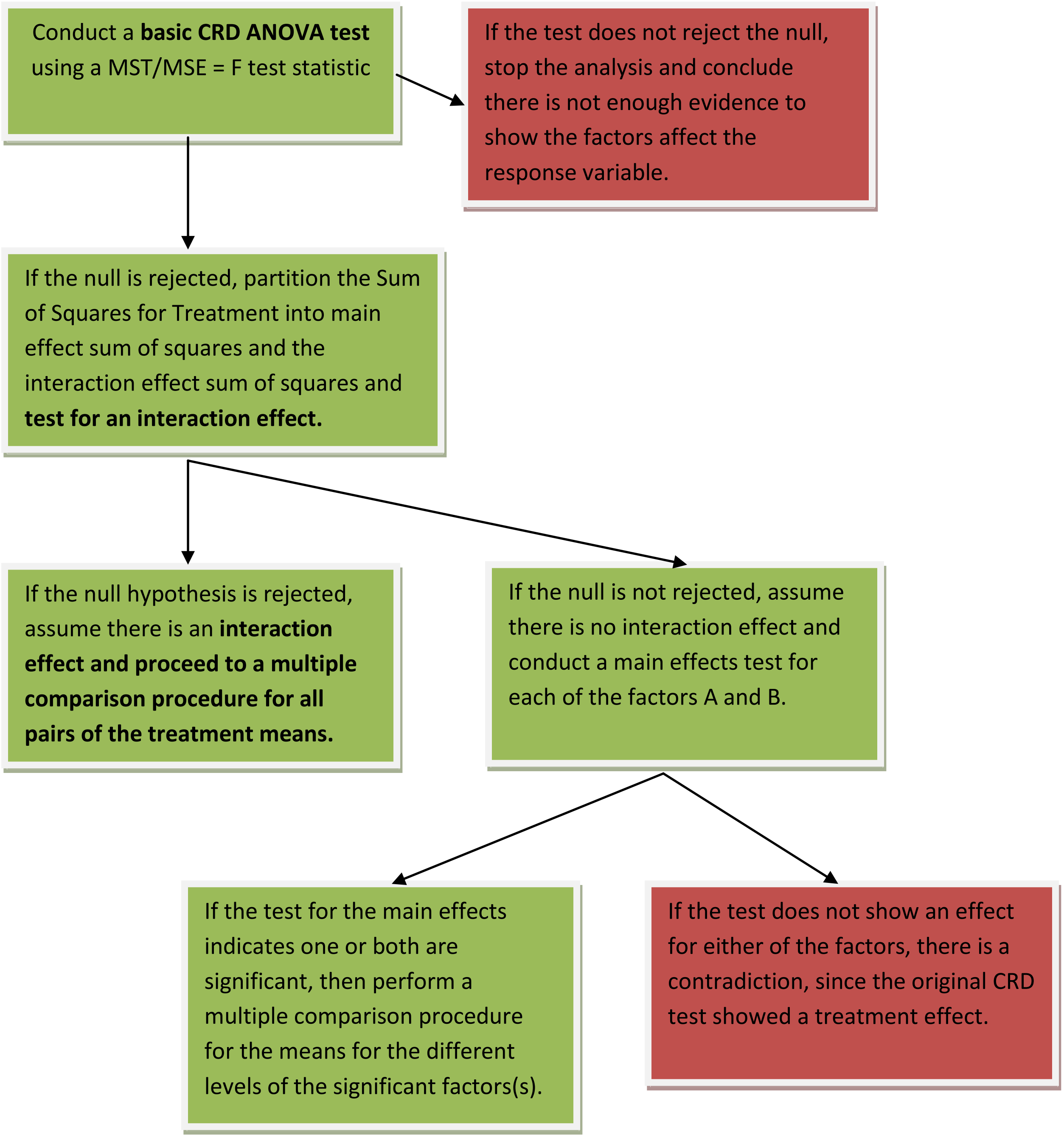


For the two-factor factorial experiment, the total sum of squares will be partitioned as follows:


SS(A):
Main Effects Sum of Squares for Factor A
SST:
Sum of Squares for Treatments
SS(B):
Main Effects Sum of Squares for Factor B
SS(AB):
Interaction Sum of Squares for Factors A and B
SS(Total):
Total Sum of Squares
SSE:
Sum of Squares for Error
SSE:
Sum of Squares for Error

Here is a broad overview of the procedure for conducting the analysis of a two-factor factorial experiment:

- Conduct a basic ANOVA CRD test using MST/MSE as your test statistic
- If the results of the CRD test are significant, partition the SST into SS(A), SS(B), and the interaction effect.
- Test for the interaction effect
- If the interaction effect appears to exist, move to a multiple comparison procedure to determine which means are significantly different from the others.
- If the interaction effect does not appear to exist, conduct the main effect tests.
- If one or both of these test are significant, move to a multiple comparison procedure to determine which means are significantly different from the others.



Conduct a **basic CRD ANOVA test** using a MST/MSE = F test statistic

If the test does not reject the null, stop the analysis and conclude there is not enough evidence to show the factors affect the response variable.

If the null is rejected, partition the Sum of Squares for Treatment into main effect sum of squares and the interaction effect sum of squares and **test for an interaction effect.**

If the null hypothesis is rejected, assume there is an **interaction effect and proceed to a multiple comparison procedure for all pairs of the treatment means.**

If the null is not rejected, assume there is no interaction effect and conduct a main effects test for each of the factors A and B.

If the test for the main effects indicates one or both are significant, then perform a multiple comparison procedure for the means for the different levels of the significant factors(s).

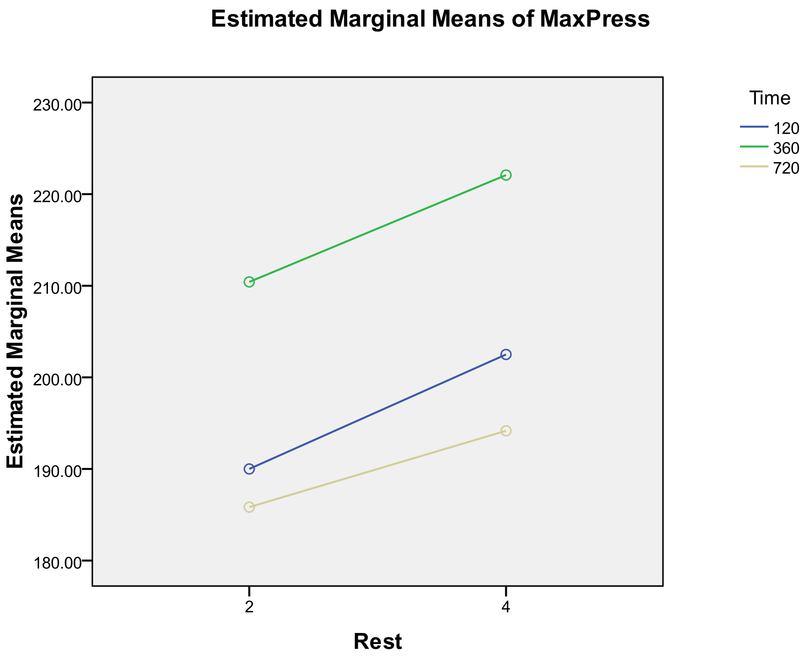
If the test does not show an effect for either of the factors, there is a contradiction, since the original CRD test showed a treatment effect.

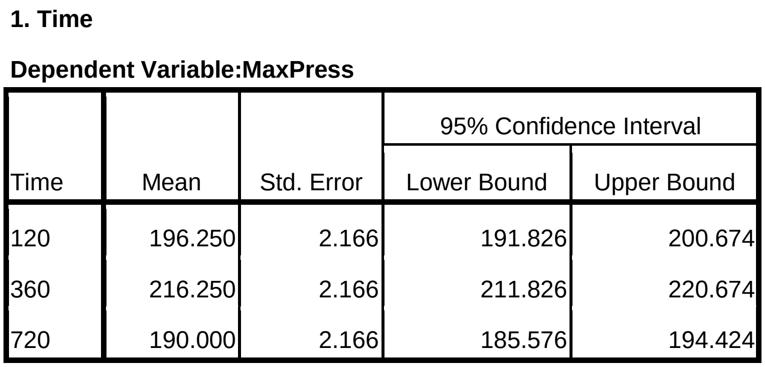
Layout of ANOVA Table for Factorial Design Experiments

Source	DF	SS	MS	F
A	k - 1	SSA	SSA/(k -1)	MSA/MSE
B	b - 1	SSB	SSB/(b-1)	MSB/MSE
A*B	(k-1)(b-1)	SSAB	SSAB/[(k-1)(b-1)]	MSAB/MSE
Error	n - kb	SSE	SSE/(n-kb)	
Total	n - 1	SStotal		

Example 159.1 (Tech): Exercise researchers are interested in the effects of two variables on maximum bench press in trained males. The two variables of interest are the amount of time (120 seconds, 360 seconds, or 720 seconds) the pectoral muscles spend under tension during four training sessions and the number of days (2 days or 4 days) of recovery between the training sessions. The participants did not have significantly different maximum bench presses prior to the start of the intervention, and the four training sessions were conducted with sets involving a load of 75% of the participant's baseline maximum bench press. Thirty-six participants were randomly assigned to six different groups, and the response variable measured was the amount of weight lifted during the maximum bench press at the completion of the training intervention. Use the SPSS outputs below to answer the following questions:

	Two Days Rest	Four Days Rest
120s	190, 190, 195, 200, 185, 180	200, 190, 210, 205, 195, 215
360s	220, 210, 205, 200, 215, 212.5	230, 220, 225, 217.5, 225, 215
720s	180, 180, 190, 195, 190, 180	185, 185, 190, 200, 205, 200




1. Time

Dependent Variable:MaxPress

Time	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
120	196.250	2.166	191.826	200.674
360	216.250	2.166	211.826	220.674
720	190.000	2.166	185.576	194.424

2. Rest

Dependent Variable:MaxPress

Rest	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
2	195.417	1.769	191.804	199.029
4	206.250	1.769	202.638	209.862

Tests of Between-Subjects Effects

Dependent Variable:MaxPress

Source	Sum of Squares	df	Mean Square	F	Sig.
Time	4512.500				.000
Rest	1056.250				.000
Time * Rest	29.167			.259	.774
Error					
Corrected Total	7287.500				

Multiple Comparisons

MaxPress

Tukey HSD

(I) Time	(J) Time	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
120	360	-20.0000*	3.06375	.000	-27.5530	-12.4470
	720	6.2500	3.06375	.120	-1.3030	13.8030
360	120	20.0000*	3.06375	.000	12.4470	27.5530
	720	26.2500*	3.06375	.000	18.6970	33.8030
720	120	-6.2500	3.06375	.120	-13.8030	1.3030
	360	-26.2500*	3.06375	.000	-33.8030	-18.6970

Based on observed means.

The error term is Mean Square(Error) = 56.319.

*. The mean difference is significant at the .05 level.

a) Identify the factors and levels for this experiment.

b) This two-factor factorial experiment can be referred to as a 3 X 2. Where does the 3 X 2 come from?

c) Give an example of a treatment for this experiment. How many different treatments are there?

d) How many replications were used for this experiment? Why is it necessary to have more than one?

e) What does the plot of the marginal means indicate?

f) Complete the missing parts of the ANOVA table above.

g) What is the p-value for the F test statistic related to the interaction effect? What should we conclude about the interaction between these factors?

h) Based on the results of the test for an interaction effect, is it appropriate to test for main effects?

i) At a 5% significance level, does the amount of rest between workouts affect the amount of weight lifted during the maximum bench press?

j) At a 5% significance level, does the amount of time the pectoral muscles spend under tension affect the amount of weight lifted during the maximum bench press?

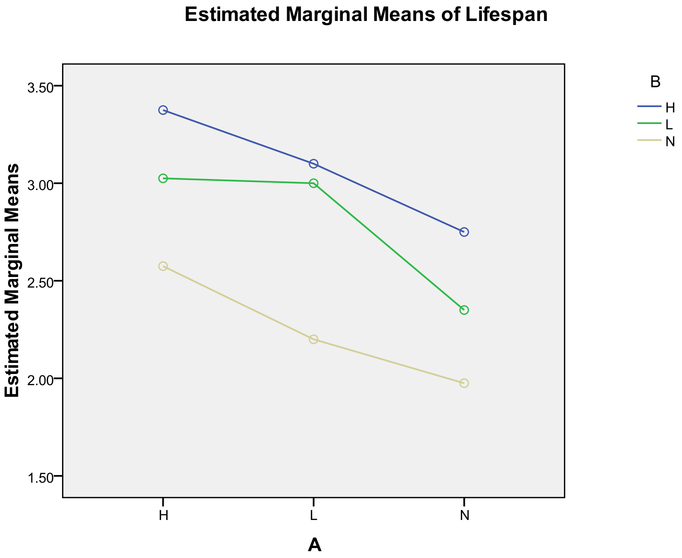
k) Use the results of the multiple comparison procedure, included with the SPSS output, to construct a diagram that ranks the means for the different times under tension.

l) Why is there no multiple comparison output for the rest factor? Which level of rest produces the greater maximum bench press?

m) Summarize your conclusions for this ANOVA two-factor factorial experiment.

Example 159.2 (Tech): Nutrition researchers are interested in the effects of two antioxidant supplements (A and B) on the longevity of rats. The response variable, lifespan, is measured in years. A group of 36 rats were divided evenly and randomly into nine groups. The supplements A and B were provided at three different dose levels (high, low, and none). Use the various SPSS outputs below to answer the questions that follow:

	No A	Low A	High A
No B	1.9, 2.1, 2.0, 1.9	2.0, 2.1, 2.3, 2.4	2.6, 2.5, 2.6, 2.6
Low B	2.2, 2.2, 2.4, 2.6	2.9, 3.0, 3.0, 3.1	2.9, 3.0, 3.1, 3.1
High B	2.9, 2.8, 2.6, 2.7	3.1, 3.2, 3.0, 3.1	3.4, 3.3, 3.5, 3.3



Tests of Between-Subjects Effects

Dependent Variable:Lifespan

Source	Sum of Squares	df	Mean Square	F	Sig.
A	2.474			85.635	.000
B	4.217				.000
A * B					.015
Error	.390				
Corrected Total	7.299				

a) Identify the factors and levels for this experiment.

b) This two-factor factorial experiment can also be referred to as a 3 X 3. Where does the 3 X 3 come from?

c) Give an example of a treatment for this experiment. How many different treatments are there?

d) How many replications were used for this experiment? Why is it necessary to have more than one?

e) What does the plot of the marginal means indicate?

f) Complete the missing parts of the ANOVA table above.

g) What is the p-value for the F test statistic related to the interaction effect? At the 5% significance level, what should we conclude about the interaction between these factors?

h) Based on the results of the test for an interaction effect, is it appropriate to test for main effects?

i) *At a 5% significance level, use the multiple comparison output below to determine if rats receiving a high dose of supplement B, live significantly longer with a high or low dose of supplement A.

j) *At a 5% significance level, use the multiple comparison output below to determine if rats receiving a high dose of supplement A, live significantly longer without B supplementation or with a low dose of supplement B.

*optional material

Pairwise Comparisons

Dependent Variable:Lifespan

B	(I) A	(J) A	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
H	H	L	.275*	.085	.003	.101	.449
		N	.625*	.085	.000	.451	.799
	L	H	-.275*	.085	.003	-.449	-.101
		N	.350*	.085	.000	.176	.524
	N	H	-.625*	.085	.000	-.799	-.451
		L	-.350*	.085	.000	-.524	-.176
L	H	L	.025	.085	.771	-.149	.199
		N	.675*	.085	.000	.501	.849
	L	H	-.025	.085	.771	-.199	.149
		N	.650*	.085	.000	.476	.824
	N	H	-.675*	.085	.000	-.849	-.501
		L	-.650*	.085	.000	-.824	-.476
N	H	L	.375*	.085	.000	.201	.549
		N	.600*	.085	.000	.426	.774
	L	H	-.375*	.085	.000	-.549	-.201
		N	.225*	.085	.013	.051	.399
	N	H	-.600*	.085	.000	-.774	-.426
		L	-.225*	.085	.013	-.399	-.051

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Pairwise Comparisons

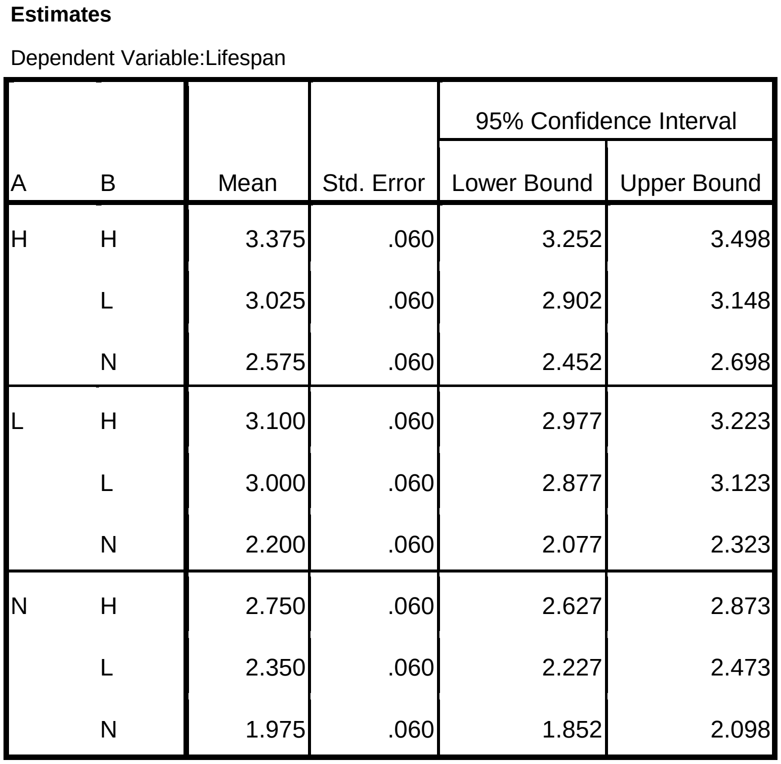
Dependent Variable:Lifespan

A	(I) B	(J) B	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
H	H	L	.350*	.085	.000	.176	.524
		N	.800*	.085	.000	.626	.974
	L	H	-.350*	.085	.000	-.524	-.176
		N	.450*	.085	.000	.276	.624
	N	H	-.800*	.085	.000	-.974	-.626
		L	-.450*	.085	.000	-.624	-.276
L	H	L	.100	.085	.250	-.074	.274
		N	.900*	.085	.000	.726	1.074
	L	H	-.100	.085	.250	-.274	.074
		N	.800*	.085	.000	.626	.974
	N	H	-.900*	.085	.000	-1.074	-.726
		L	-.800*	.085	.000	-.974	-.626
N	H	L	.400*	.085	.000	.226	.574
		N	.775*	.085	.000	.601	.949
	L	H	-.400*	.085	.000	-.574	-.226
		N	.375*	.085	.000	.201	.549
	N	H	-.775*	.085	.000	-.949	-.601
		L	-.375*	.085	.000	-.549	-.201

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).



Estimates

Dependent Variable:Lifespan

A	B	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
H	H	3.375	.060	3.252	3.498
	L	3.025	.060	2.902	3.148
	N	2.575	.060	2.452	2.698
L	H	3.100	.060	2.977	3.223
	L	3.000	.060	2.877	3.123
	N	2.200	.060	2.077	2.323
N	H	2.750	.060	2.627	2.873
	L	2.350	.060	2.227	2.473
	N	1.975	.060	1.852	2.098