CHAPTER 4

Analysis of Variance (ANOVA)

Eager to continue with her work on BBQ sauce viscosity, Maria arrives at her office early the next day thinking she would look into statistical analysis techniques for comparing more than two samples before the additional data arrives. Her plan is to start with the textbook from Dr Wang's class. It had occurred to her that she could certainly perform *t*-tests as she had done before, one for each pair of lines. With five lines there are quite a few pairs to compare: line 2 versus line 3, line 2 versus line 4, and so on. She lists them all on a sheet of paper and counts 10 pairs. She has already performed the *t*-test for line 4 versus line 6, so nine more *t*-tests would be required. And then, she would have to compare the results of all 10 tests. It looks to be a lot of work even when using Excel. Maybe there is a more efficient way to tackle this problem.

She is surprised to find an email from Lisa with the data from three additional lines. The text of the email reads:

I have included viscosity data from lines 2, 3, and 5 to compare with the data you have for lines 4 and 6. I can't wait to see what you find!

"That woman must never sleep! With all of the things she is responsible for at the plant, I can't believe that she was able to get this information to me so quickly," Maria exclaims.

The data files look similar to the files she had received earlier. The first few lines in the data file look like this:

line	Date and Time	viscosity
2	4/28/2013 8:09	5109
2	4/28/2013 8:18	4978
2	4/28/2013 8:26	4929
2	4/28/2013 8:34	4916
2	4/28/2013 8:43	4430
2	4/28/2013 8:51	4822
2	4/28/2013 8:59	5006
2	4/28/2013 9:08	3987
2	4/28/2013 9:16	4487
2	4/28/2013 9:24	4219
2	4/28/2013 9:33	4494
2	4/28/2013 9:41	4633

Statistics for Food Scientists
http://dx.doi.org/10.1016/B978-0-12-417179-4.00004-4

There is now a column indicating the line where the product was produced. Maria checks and sees that there is the same number of data points from the same date as the previous files.

After taking care of a few minor tasks, Maria opens her textbook and scans the table of contents. She finds a chapter that contains the *t*-test that she performed. Two chapters after this, she finds another chapter titled "Comparing *k* Means—One-way Analysis of Variance (ANOVA)." Guessing that this may be what she is looking for, she turns to that chapter and begins reading.

Maria quickly learns that analysis of variance (often abbreviated as ANOVA) could be used for her comparison of the production lines. Its name relates to the concept that the variability in a set of data can be broken into different components. In the simplest form, there are two components:

1. Variability between the factor level means
2. Variability of the individual values within each factor level

Factors designate the group of things being compared; in her case, these are the production lines. Levels are the names of the elements in the factors; in her case, these are the line numbers 2 through 6. An illustration in the book demonstrates these concepts:

	Factor		
sample	**A**	**B**	**C**
1	90	94	100
2	92	96	115
3	88	98	97
mean score	90	96	104

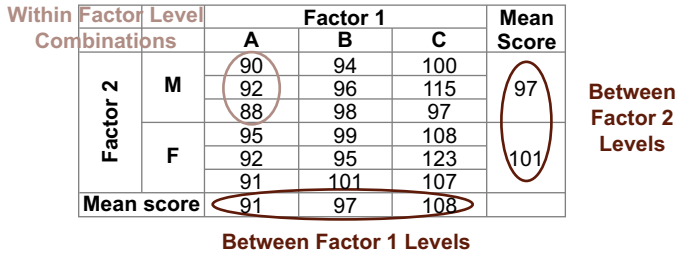
Within Factor Level

Between Factor Levels

The statistical test for differences between the factor level means uses a ratio of these two variability components, with the variability between factor levels in the upper part of the ratio and the pooled within factor level variability in the lower part of the ratio. A larger ratio suggests treatment differences.

For her application, the factor levels would be the different production lines. Maria learns that if there are only two factor levels, ANOVA is identical to the *t*-test that she has performed. ANOVA is just an extension of the analysis she already completed. The technique will compare the variability between the lines and the variability within the lines.

Looking ahead in the book, she sees that the next chapter is titled "Multi-way Analysis of Variance." The ANOVA technique is quite flexible and can accommodate more complex problems, in particular additional factors. The example in the book describes a problem investigating different drugs (factor 1) for both male and female patients (factor 2).



This gets Maria thinking about the data Lisa provided her. It is from one production day, but from both the day and evening production shifts. There are different operators for the lines on each shift, and they may run the machinery differently. Perhaps she should consider the shift as an additional treatment in her analyses.

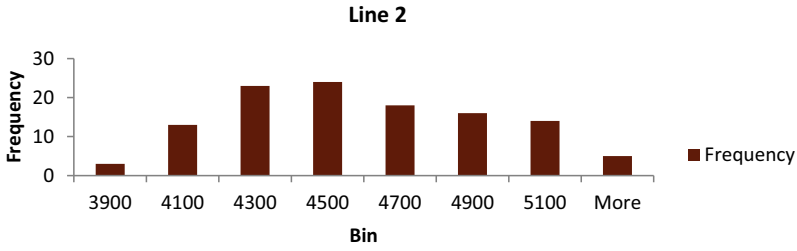
"Maybe I should not bite off more than I can chew," Maria thinks aloud. "Let me start with the simpler one-way ANOVA to compare the lines and see what I can learn from that analysis."

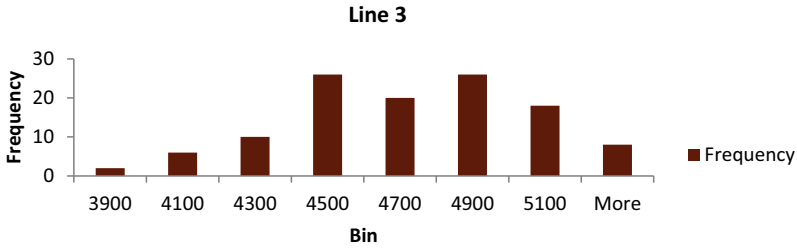
Maria sees that in the Analysis ToolPak add-in, there are three choices for ANOVA: Single Factor; Two-Factor with Replication; and Two-Factor without Replication.

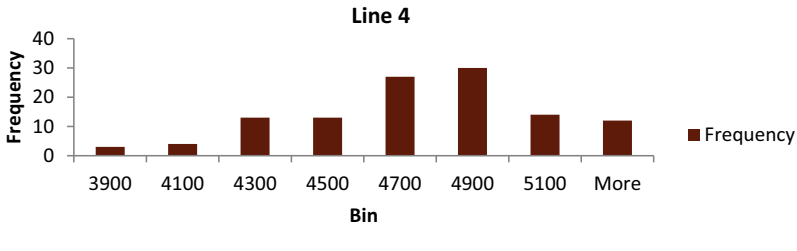
Maria arranges the data in Excel so that the viscosities from the five lines are in separate columns. The first few lines of the rearranged data sheet look like this:

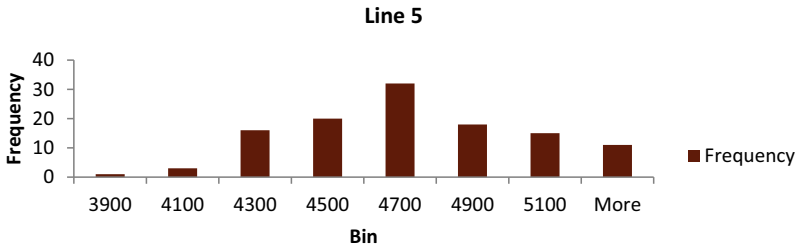
Line 2	Line 3	Line 4	Line 5	Line 6
5109	5172	5076	5289	4046
4978	5042	4618	5059	4724
4929	4796	4663	5125	4584
4916	4784	4747	4353	4541
4430	4840	4132	4807	4432
4822	5423	4726	4258	4489
5006	4918	4816	5113	4950
3987	4731	5385	4789	4603
4487	5070	4885	4801	4552
4219	4621	4793	4355	4643
4494	4423	5423	4260	4353
4633	4635	5001	5249	4500
4388	4473	4873	4784	4954
4425	4981	4490	4648	4849
4914	5066	4521	4565	4317
4607	4789	5374	4233	4719
4838	4586	4609	4688	4607
4179	5005	4689	4579	4876
4839	4948	4807	4028	4737
4269	4464	4617	4815	4911
4777	4666	5263	5421	4736
4221	4574	4108	4608	4372
5014	5048	4957	4776	4519

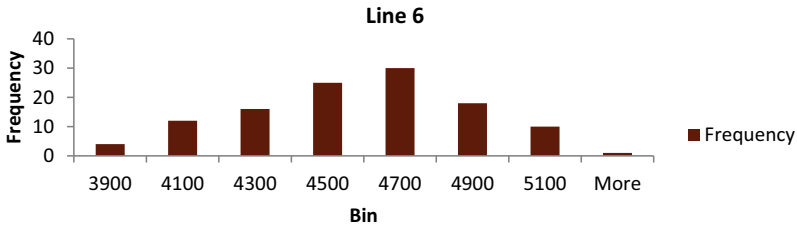
As she did when comparing only lines 4 and 6, Maria starts by making histograms of viscosity for each of the lines. Her graphs look like this:





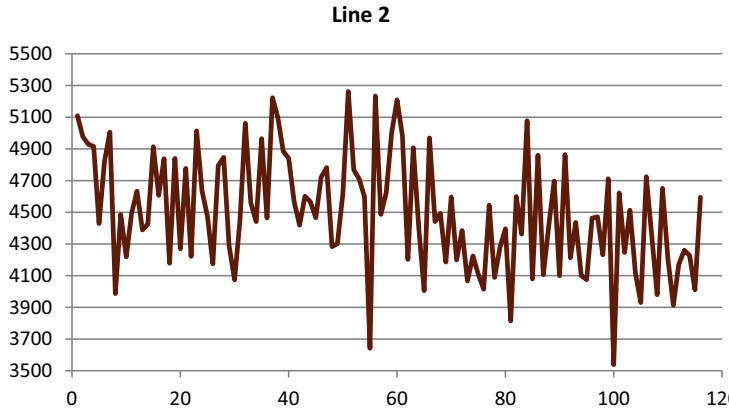


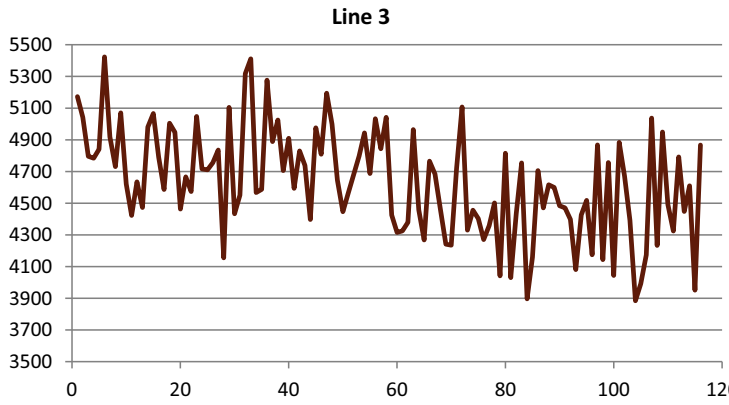


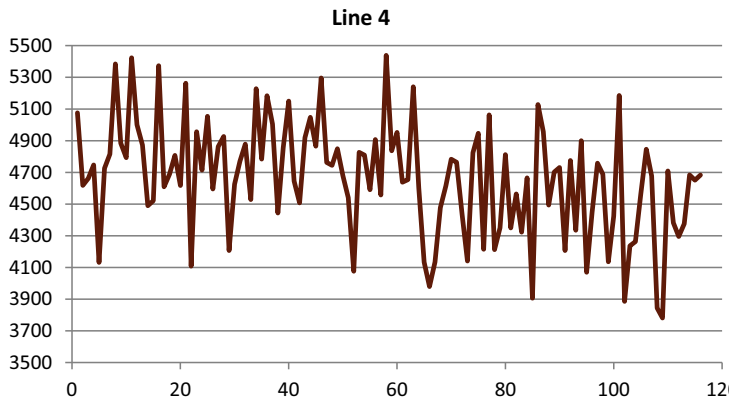


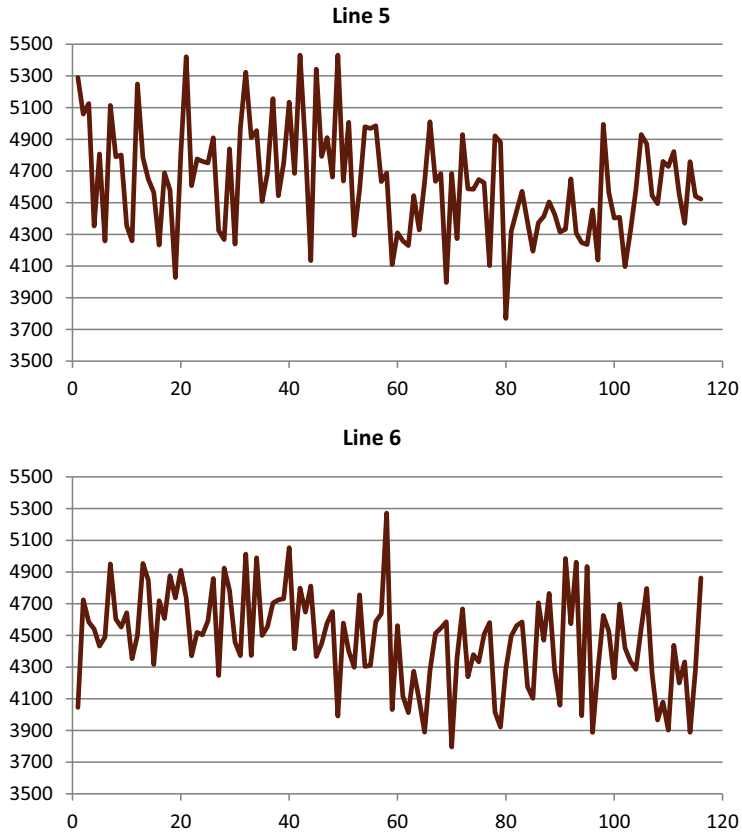
Maria notes that line 2 in addition to line 6 may have a lower average than the other lines. She also observes that the variability of line 6 still looks to be smaller than that of the other lines.

As she did when comparing only lines 4 and 6, Maria also makes plots of the viscosity data in time sequence for all five production lines. Her plots look like this:









She does not see an increasing or decreasing trend for the lines except for line 3, which does seem to display diminishing viscosities over time. She makes a note to mention that to Lisa.

In the menu of her Analysis ToolPak add-in, she selects ANOVA: Single Factor. She selects the full input range and checks the box indicating that the column labels are in the first row. The analysis output that appears in a new work sheet looks like this:

ANOVA: Single Factor

Summary

Groups	Count	Sum	Average	Variance
Line 2	116	521,847	4499	129,823
Line 3	116	537,496	4634	111,617
Line 4	116	540,892	4663	125,955
Line 5	116	536,978	4629	113,420
Line 6	116	519,734	4480	91,052

ANOVA

Source of Variation	SS	df	MS	F	p-value	F crit
Between groups	3,325,273	4	831,318	7.27	0.00001	2.39
Within groups	65,764,690	575	114,373			
Total	69,089,962	579				

Most of the analysis details are easily understood. There is a table with summary statistics for each of the lines with the sum, average, and variance. The ANOVA table below the summary table contains a *p*-value for the between groups source of variation that seems to be the statistical test for the differences between the lines, exactly what she is looking for. She recalls that smaller *p*-values indicate significant differences. As the *p*-value here rounds down to zero, it would indicate the average viscosities are different between the lines.

"OK, now I'm getting somewhere!" Maria looks at the means for each of the five lines and sees that for line 3 and line 5, the means are pretty similar at 4634 and 4629. And line 2 and line 6 are also similar with means of 4499 and 4480. Line 4 has a mean of 4663, which is pretty close to the means for line 3 and line 5.

"Do I decide on my own what differences are significant? Lisa would want something more definitive I'm sure," Maria thinks out loud as she builds her analysis summary. "Maybe there is a way to determine this more objectively."

Maria goes back to her textbook to see if this question can be addressed. As she reads on, she comes across the concept of multiple paired comparison (MPC) procedures, which seems to be a way to determine individual treatment differences after an ANOVA analysis is performed. A number is calculated to determine the minimum significant difference between treatments. Any two treatments with a mean difference smaller than this number would not be considered different; any two treatments with a mean difference larger than this number would be considered different. There are several ways of calculating the number, and she is not sure which one to choose. The textbook mentions Fisher's least significant difference (LSD), Duncan's, Student-Newman-Keuls, Tukey's Honest Significant Difference (HSD), and Scheffe's as possible choices in certain situations. For some, the minimum significant difference is smaller and the significance test is more liberal. For others, the minimum significant difference is larger and the test is more conservative. But, she does see one statement that has relevance: If you are interested in all pairwise

comparisons, you should use Tukey's HSD. She is interested in comparing all of the lines to each other, so she follows this advice.

Maria looks up MPC procedures in Excel and finds that none are available in the Analysis ToolPak add-in. "This is a problem; why not stop with the ANOVA analysis and try to determine the individual paired differences myself?" Maria grumbles in frustration. Fortunately the book describes the calculation for the minimum significant difference.

It uses a new tabled distribution that she is not familiar with: the studentized range. A value from this distribution is the q in this formula:

$$\frac{q_{k,\nu,\alpha/2} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}{\sqrt{2}}$$

In this formula s is the square root of within groups variance from the ANOVA; $q_{k,\nu,\alpha/2}$ is the alpha upper significance level of the studentized range for k means; ν is the number of degrees of freedom for the within groups variance (labeled d*f* in her ANOVA output); and n_i and n_j are the sample sizes for each of the treatments being compared.

Maria looks at the studentized range table in her textbook and sees that it stops at $\nu = 120$. The degrees of freedom for her analysis are 575, not even close to this! But, she also sees that there is a table entry for infinity, and that number is very close to the one for 120; so, for her degrees of freedom of 575, the entry for infinity should be OK to use. This value is 3.87 for an alpha $= 0.05$. Using this, she calculates the minimum significant difference to be 121.5.

Maria now examines the table of averages for each line. The two smallest averages are for line 2 and line 6 at 4499 and 4480, respectively. The difference between the two is smaller than 121.5, so the average viscosity of these two lines is not significantly different. Line 3, line 4, and line 5 have averages of 4634, 4663, and 4629, respectively. These three averages are all within 121.5 of each other and are then not significantly different. But, each of these three averages is greater than 121.5 from the averages of both line 2 and line 6. So, the lines form two groups: line 2 and line 6 have significantly smaller average viscosities than line 3, line 4, and line 5.

Maria adds these details to her analysis summary by creating a table ordering the lines by decreasing average viscosity. She feels confident now that she is prepared to meet with Lisa to deliver the results of her analyses. She is about to arrange a meeting for the following day when she remembers that she had thought about analyzing the data as a two-way

ANOVA with shift as the second treatment. The two-way ANOVA would add an additional dimension to the analysis that may be interesting.

Line	Average Viscosity	Group
4	4663	A
3	4634	A
5	4629	A
2	4499	B
6	4480	B
minimum significant difference	121.5	

Maria now starts working with the data to create a new column that indicates the shift that produced and collected the data based on the time stamp. All data from the first shift is from 8:00 AM to 4:00 PM. The second shift data is from 4:00 PM until midnight. Her data file now looks like this:

shift	Line 2	Line 3	Line 4	Line 5	Line 6
1	5109	5172	5076	5289	4046
1	4978	5042	4618	5059	4724
1	4929	4796	4663	5125	4584
1	4916	4784	4747	4353	4541
1	4430	4840	4132	4807	4432
1	4822	5423	4726	4258	4489
1	5006	4918	4816	5113	4950
1	3987	4731	5385	4789	4603
1	4487	5070	4885	4801	4552
1	4219	4621	4793	4355	4643
1	4494	4423	5423	4260	4353
...	...	...	...	...	...
2	5002	4426	4836	4110	4032
2	5209	4316	4953	4310	4561
2	4981	4325	4638	4260	4116
2	4203	4379	4654	4230	4013

In the menu of her Data Analysis add-in, she selects ANOVA: Two-Factor with Replication since she has repeated viscosity measurements for each line and shift combination. She chooses the full input range and indicates that there are 58 rows for each sample (Shift). The analysis output that appears in a new window looks like this:

ANOVA: Two-Factor with Replication

Summary	Line 2	Line 3	Line 4	Line 5	Line 6	Total
Shift 1						
Count	58	58	58	58	58	290
Sum	268,403	279,231	278,452	276,659	267,248	1,369,993
Average	4628	4814	4801	4770	4608	4724
Variance	109,425	70,285	95,764	119,372	61,583	97,848

Continued

ANOVA: Two-Factor with Replication—cont'd

Summary	Line 2	Line 3	Line 4	Line 5	Line 6	Total
Shift 2						
Count	58	58	58	58	58	290
Sum	253,444	258,265	262,439	260,319	252,486	1,286,953
Average	4370	4453	4525	4488	4353	4438
Variance	118,653	88,423	119,576	69,080	89,163	100,079
Total						
Count	116	116	116	116	116	
Sum	521,847	537,496	540,892	536,978	519,734	
Average	4499	4634	4663	4629	4480	
Variance	129,823	111,617	125,955	113,420	91,052	

ANOVA

Source of Variation	SS	*df*	MS	*F*	*p*-value	*F* crit
Sample	11,889,085	1	11,889,085	126.30	0.000000	3.86
Columns	3,325,273	4	831,318	8.83	0.000001	2.39
Interaction	220,237	4	55,059	0.58	0.673685	2.39
Within	53,655,367	570	94,132			
Total	69,089,962	579				

In her analysis, the "Sample" source of variation is from the shifts and the "Column" source of variation is from the lines. This appears to be the standard naming convention in Excel. The analysis validates her earlier determination about the line differences since the *p*-value listed for "Columns" is smaller than 0.05. To her surprise, the analysis also indicates that there is a significant difference between the two shifts since the *p*-value for "Sample" is also smaller than 0.05. Shift 1 has a higher average viscosity than shift 2. She looks at the means for the two shifts for each of the production lines and sees that the differences between shift 1 and shift 2 for each of the lines seem to be much larger than the differences between the lines!

There is also a source of variation listed as "Interaction." She looks back at her textbook and finds out that a significant interaction is when the effect of one factor is not the same for different levels of another factor. In her analysis, the interaction source of variation is not significant. So this difference between shifts is consistent across the lines. Why would the viscosities be so different between the shifts?

Maria is now even more puzzled about the line performances at the plant. She is happy that she performed the additional analysis to account for

shifts since that seems to be an important finding. She is certain she will need to defend this finding to Lisa when she shares the information.

Maria sets up a meeting with Lisa for the next morning. She carefully describes the steps in her analysis and how she has come to the conclusion that while lines do perform differently, there is an even bigger difference in viscosity for products produced between shift 1 and shift 2.

"I think that you have made an important discovery Maria," Lisa says after Maria presents her case. "I'm convinced that your analysis is thorough and correct from the details you have shown. But let's think this through some more. The data show clearly that the viscosities are consistently higher on shift 1 regardless of the production line. Although there are different operators on each line, the viscosity measurements for all lines are made in the quality lab by the same technician. The technician is different between the shifts, can we be sure that this is a true shift difference and not some sort of measurement problem?"

Maria realizes that Lisa is correct and that what she had assumed was a production difference could be a problem with the measurements. "I'll see what I can do to determine what is really going on here and get back to you with what I find," she promises Lisa.