Week 3 Lecture 8

Excel ANOVA Example

In our on-going investigation of whether or not males and females are paid equally for equal work, we have come up with contradicting results so far, average salaries are clearly different but average compa-ratios are not. We need to examine reasons that might impact these differences to see if we can explain what is going on. For possible factors influencing individual salaries, we need to be able to, paraphrasing what they say in TV cop shows, "rule it out as a suspect" in causing differences or keep it in as a cause of differences between the gender pay practices.

One key issue in our question that has not clearly been examined yet is the impact of grades on salaries. Clearly, grade differences have the potential to complicate the issue as the work done differs by grade. One question to ask here is, "are average salaries equal across grade levels?" This becomes our research question.

Example

For the research question of: are average salaries equal across the grades, we have the following hypothesis test.

Step 1: Ho: All salary means are equal.

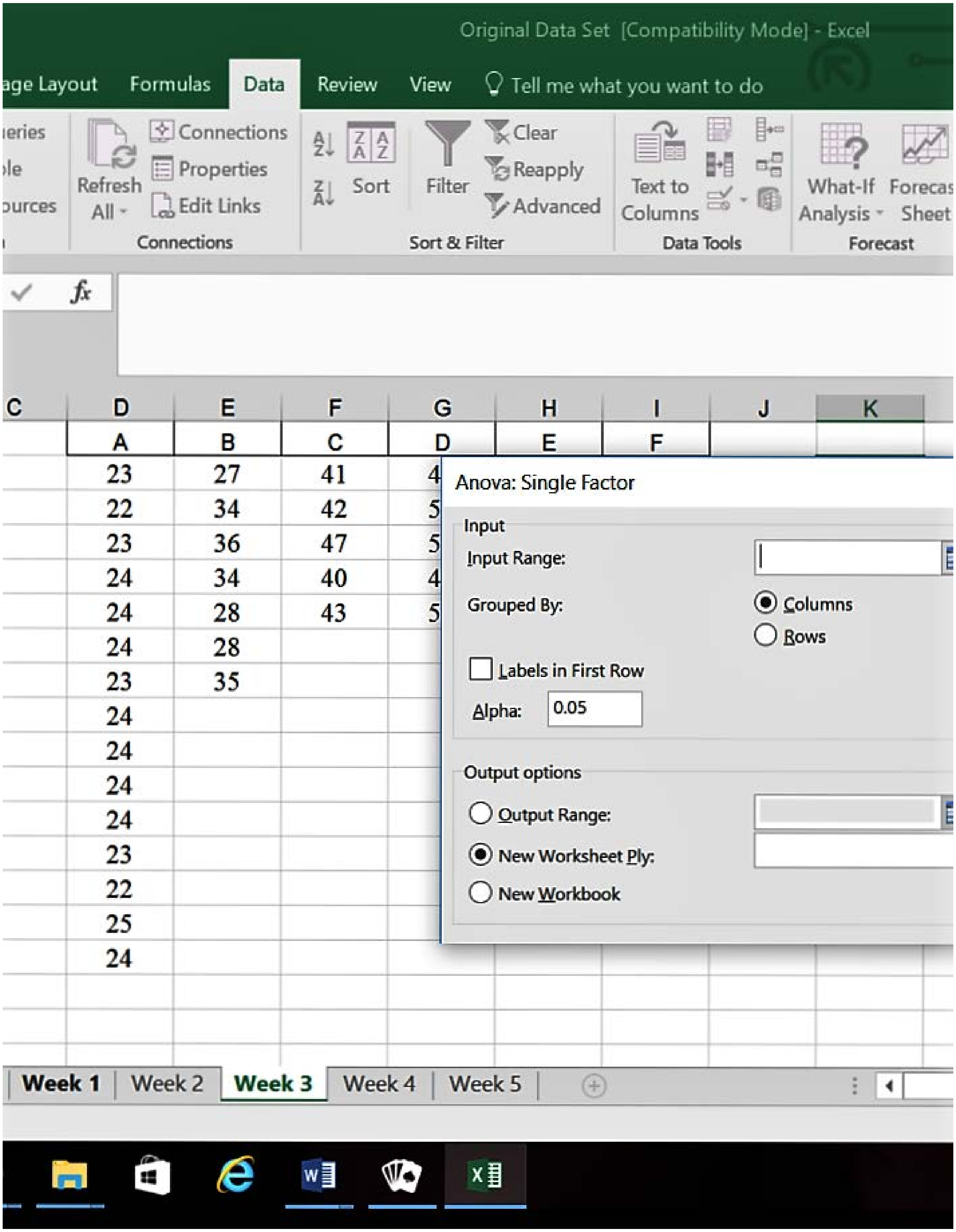
Ha: At least one mean differs.

Step 2: Reject the null if the p-value < alpha = .05.

Step 3: Statistical Test: Single-factor ANOVA. (Note: salary variance in some of the grades may violate the equal variance requirement. We will ignore this for the purposes of this example.)

Step 4: Perform Test.

The input box for Excel's Single factor ANOVA is



C	D	E	F	G	H	I	J	K
	A	B	C	D	E	F		
	23	27	41	4				
	22	34	42	5				
	23	36	47	5				
	24	34	40	4				
	24	28	43	5				
	24	28						
	23	35						
	24							
	24							
	24							
	24							
	23							
	22							
	25							
	24							

Anova: Single Factor

Input

Input Range: |

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Alpha: 0.05

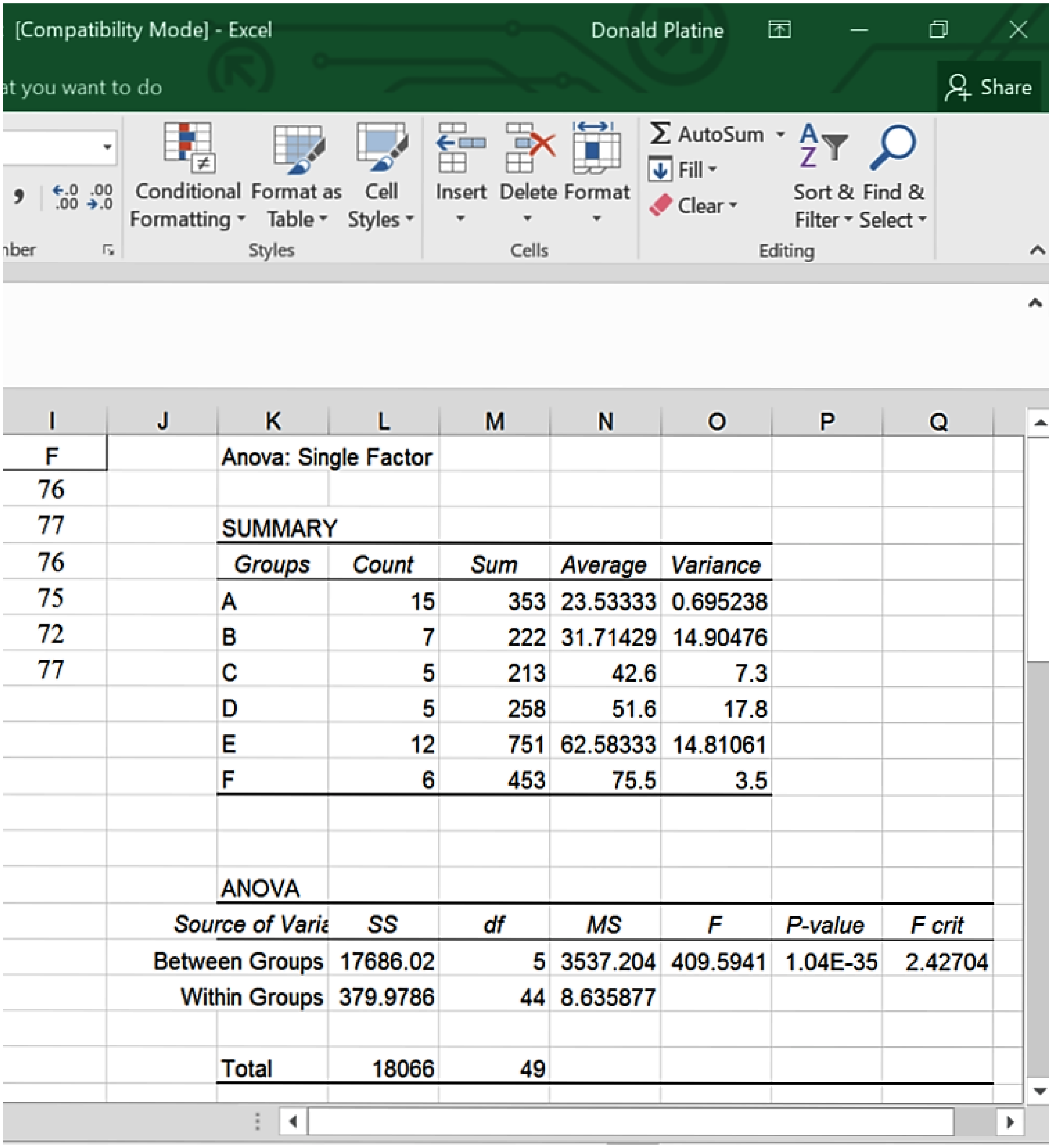
Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

The input range for this example would be D1:F16; we would click on Labels in the first row, and select any output range desired (This would be given in the assignment for consistency’s sake). Completing the input screen and clicking OK gives us an output table.



I	J	K	L	M	N	O	P	Q
F		Anova: Single Factor						
76								
77		SUMMARY						
76		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
75		A	15	353	23.53333	0.695238		
72		B	7	222	31.71429	14.90476		
77		C	5	213	42.6	7.3		
		D	5	258	51.6	17.8		
		E	12	751	62.58333	14.81061		
		F	6	453	75.5	3.5		
		ANOVA						
		Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
		Between Groups	17686.02	5	3537.204	409.5941	1.04E-35	2.42704
		Within Groups	379.9786	44	8.635877			
		Total	18066	49				

Reading the ANOVA output tables

The first thing we see is the test name in cell K-1: Anova: Single Factor. This is just a check to ensure we have the right test. Next we see a summary table. Under the Groups column we should see the data labels (in this case our grades). If not, and we see something such as a number, an input error has been made, the labels were not included but the Labels box was checked. If this happens, just redo the data set up and overwrite the output.

For each variable, we see the count, sum, average, and variance. If we had some question about having equal variance, we could perform an F-test on the variables with the extreme values. (Again, for purposed of this example, we are going to ignore the requirement for equal variances.)

The next table is the ANOVA output. While, technically for our hypothesis test, we only need to look at the p-value result, the other columns provide some useful information.

Note: this is somewhat technical, and is presented only as an explanation of the table. The source of variation column gives us our two variation measures; *Between groups* refers to the overall variation while *Within Groups* refers to the average variation for all the groups. The SS column (Sum of Squares) is an estimate of the variation (slightly different than our variance formula). This value is divided by the df (degrees of freedom) value for each group. This df is conceptually the same as that discussed with the t-test; and the total df is N-1, where N is the number of data points. Looking at this value (49 in this example) confirms we entered the right number of data points of 50.

MS stands for Mean Square and is the SS divided by the df. The F value is determined by dividing the MS for the between row by the MS of the Within groups row. The p-value and the critical F statistic complete the table.

Step 5: Conclusion and Interpretation: The F is much larger than the F critical, and the p-value is much less than 0.05 (Note: 1.04E-35 means move the decimal point 35 places to the left (0.0000000000000000000000000000000000104). If the E (for exponent) had been positive, we would have moved the decimal to the right, example 1.04E4 = 10400.)

So, according to our decision rule, since the p-value is < (less than) 0.05, we reject the null hypothesis and conclude that at least one mean differs. This suggests that grade level has an impact on salary, and that measuring pay in salary terms could be creating some issues in answering our questions.

Determining Differences

When we reject the null hypothesis, a logical follow-up question is often, which differences are meaningful? There are several approaches to answering this question; all involve a pair by pair comparison, and most require access to statistical tables not available within Excel.

One approach that we can use in our Excel worksheet involves developing *confidence intervals* around the difference in group means. (Note: Confidence intervals allow us to develop a range that contains the value we are looking for with a known level of confidence such as 95%. We will discuss this again in Week 5.)

All of the required information for these intervals is available from the ANVOA output. The basic approach is to

1. Find the difference between each pair of means
2. To this value, add and subtract a measure of the variation in the data (due to sample error, we know our sample means are not exactly equal to the population parameter, so we need to take this sample error into account, our real difference might be a bit larger or smaller than the samples show).

3. Examine the ranges to see if 0 is included (alternately, do the endpoints have different signs a + and -); if so the real population difference could be 0 and the means do not significantly differ.

The formula for the interval that we will build in Excel is:

(mean1 – mean2) +/- t*sqrt(MSW * (1/n1 + 1/n2)) (Lind, Marchel, & Wathen, 2008).

Here is an example of how we work out the formula, and what each term means. The value of the means for each variable is found in the Summary table, as is the count (n) for each variable. The MSW is the MS for within that is found in the ANOVA table, and we find t with the t.inv function from Excel.

So, let's walk thru constructing an interval for grades A and B, and then we can look at what it might look like in an Excel spreadsheet. From our example output above, we have:

Mean A = 23.5 (rounded)

Mean B = 31.7 (rounded)

n for A = 15

n for B = 7

MSW = 8.64 (rounded)

T has a df equal to that of MSW (44 in this case), and the probability is our 0.05 for a 95% interval. T.inv(0.05, 44) equals 2.015 (rounded).

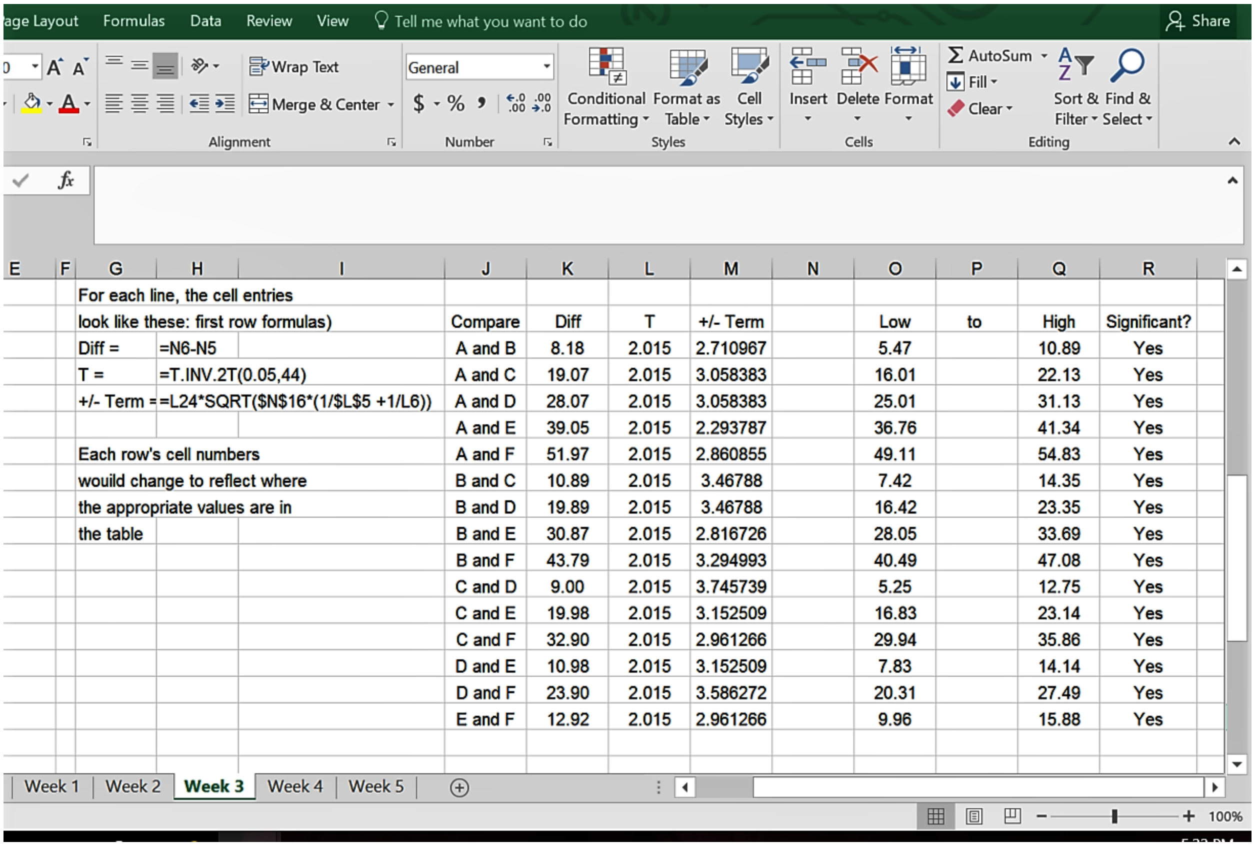
So, for grades A and B, our mean difference = 31.7 – 23.5 = 8.2

The +/- term is t * sqrt(MSW * (1/n1 +1/n2)). Plugging in our values gives us

2.015* sqrt(8.64 * (1/15 + 1/7) = 2.71.

So, our interval is 8.2 +/- 2.71 = 5.49 to 10.91 (rounded).

Since 0 is not in this range, we can say that the mean salaries for grades A and B differ significantly. Setting this up in Excel (using cell references as the examples on the left show) give us the following:



E	F	G	H	I	J	K	L	M	N	O	P	Q	R
		For each line, the cell entries											
		look like these: first row formulas)			Compare	Diff	T	+/- Term		Low	to	High	Significant?
		Diff =	=N6-N5		A and B	8.18	2.015	2.710967		5.47		10.89	Yes
		T =	=T.INV.2T(0.05,44)		A and C	19.07	2.015	3.058383		16.01		22.13	Yes
		+/- Term =	=L24*SQRT(N16*(1/L5 +1/L6))		A and D	28.07	2.015	3.058383		25.01		31.13	Yes
					A and E	39.05	2.015	2.293787		36.76		41.34	Yes
		Each row's cell numbers			A and F	51.97	2.015	2.860855		49.11		54.83	Yes
		wouild change to reflect where			B and C	10.89	2.015	3.46788		7.42		14.35	Yes
		the appropriate values are in			B and D	19.89	2.015	3.46788		16.42		23.35	Yes
		the table			B and E	30.87	2.015	2.816726		28.05		33.69	Yes
					B and F	43.79	2.015	3.294993		40.49		47.08	Yes
					C and D	9.00	2.015	3.745739		5.25		12.75	Yes
					C and E	19.98	2.015	3.152509		16.83		23.14	Yes
					C and F	32.90	2.015	2.961266		29.94		35.86	Yes
					D and E	10.98	2.015	3.152509		7.83		14.14	Yes
					D and F	23.90	2.015	3.586272		20.31		27.49	Yes
					E and F	12.92	2.015	2.961266		9.96		15.88	Yes

So, all of the grade average salary differences are significantly different from each other. Grade is definitely a factor in an employee's salary, and introduces a source of variation that is not an equal work measure.

We have not yet found an answer to our question, as we have not yet figured out how to get a measure of equal work to base our comparisons on.

More to follow next week.

References

Lind, D. A., Marchel, W. G., & Wathen, S. A. (2008). *Statistical Techniques in Business & Finance*. (13th Ed.) Boston: McGraw-Hill Irwin.