Using MS Excel in teaching Design of Experiment

MOHAMMAD SALEHI
Mathematics, Statistics, and Physics
Qatar University
Doha, PO Box 2713
QATAR
salehi@qu.edu.qa

Abstract: - Choosing right software to teach an applied statistical course is very crucial. When one teaches applied statistical courses to statistics students s/he needs to use an interactive software. Instructors should communicate with the software rather getting all the statistical results by clicking tabs. The mean reason is that the instructor can simultaneously teach the techniques and the data analysis. Perhaps the most straightforward option is to use the free software R but it is not easy for average students to use R. In this paper, I will share the experience of using MS Excel in teaching Design of Experiment. I find this experience as a success one.

Key-Words: - Interactive teaching approach**;** Statistical package; Problem-oriented; Technique-oriented.

1 Introduction

The first challenge of teaching an applied statistics course is to choose appropriate Statistical Packages. Instructors should solve "problem-oriented“ and "technique-oriented” exercises [1]. For “problem-oriented” exercises those software with clickable interfaces like SPSS and Minitab are very useful as student does not require extensive learning and training of them. The software provides the required output with few clicks and students can focus on solving the problem and interpret the output . However teaching a statistical course to statistics students, instructor ought to present “technique-oriented” exercises in order to teach the technique and review some important concepts. Even though using artificial data sets were discouraged for many reasons but for the first year courses you may prepare a simple artificial data set and use calculator to teach and clarify the technique [5]. However, it is very difficult if not impossible to use only calculator to teach techniques in a more advance courses like Design of Experiment course. An option is to use R [6] which is reasonably good software to be used for teaching data analysis and techniques simultaneously. Moreover, it is a free software and designed for teaching. Personally, I like the software very much for many its advantages particularly for its interactive nature. However, it requires extensive learning and training with some programming skills [3]. R is taught in a course called Statistical Packages (STAT371) in Statistics program, Qatar University. However, not all students have passed this course when they take Design of Experiment (STAT332). Even those students who passed the course they are reluctant to use R.

In Spring 2016, I was supposed to teach STAT381 (Categorical data analysis) and I posted its syllabus on Blackboard at the night of the first class. In the syllabus, I mentioned that we are going to use R. Right after posting it, students dropped the course except one! My conjecture was that their favorite software is not R. Like many universities, we should have one more extra contact hour for each applied statistics course in which a TA works on R so students feel comfortable to use it. But this opportunity is not available in our program. However, our students are usually familiar with Excel either from High School or Foundation program of Qatar University so that they are very comfortable to use Excel. This motivated me to find the possibility of using Excel in this course. In the first step, I did a literature survey on the subject. As I expected there are quiet number of literatures in which the advantage and disadvantage of using Excel in the first year level statistical courses were discussed. Goos and Leemans (2004) [2] discussed using Excel to teach optimal design of experiments in which he presented an interactive teaching approach to use Matrix operations in Excel for teaching optimal design. I did also Google-searched to find out whether some colleagues use Excel in teaching Design of Experiment but I could not find any. I decided to use Excel beside SPSS and Minitab in teaching. I will present an Example solved in Excel to demonstrate how plausible using Excel in Design of Experiment course.

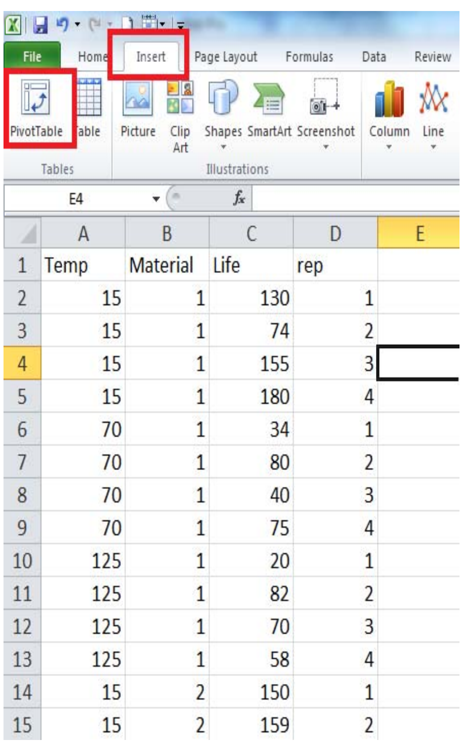
2 Used Features of MS Excel

An important advantage of MS Excel is that it can interactively be used in class, avoids the "black box" approaches offered by some statistical software. Using MS Excel helps students to understand the problem and its solution. After teaching the techniques and methods by an interactive software we may use a software like SPSS or Minitab to speed up the teaching and emphasize on the data analysis part of the course. Our text book is Montgomery (2013) [4] and I solved the most of its examples using MS Excels during the class time with details. Used Excel features were very few. The most frequent used features will be presented below.

2.1 PivotTable

PivotTable creates Multi-ways table of functions, including frequencies, summations, the means, the variances, etc. *PivotTable* is on the *Insert* tab in Excel. The data set files provided by the text book are in a stack format so that *PivotTable* is very useful to get sums of the factor levels as well as replications. Figure 1 shows *PivotTable* and *Insert* tab.

Figure 1. PivotTable in Insert tab is shown by red boxes.



	A	B	C	D	E
1	Temp	Material	Life	rep	
2	15	1	130	1	
3	15	1	74	2	
4	15	1	155	3	
5	15	1	180	4	
6	70	1	34	1	
7	70	1	80	2	
8	70	1	40	3	
9	70	1	75	4	
10	125	1	20	1	
11	125	1	82	2	
12	125	1	70	3	
13	125	1	58	4	
14	15	2	150	1	
15	15	2	159	2	

2.2 Simple functions

To compute formulas in the experimental design problems one needs to use the four basic arithmetic operations, the square/square root operations and *summation*, *SUM()*, for a row, column or a matrix which can easily be done in a nice interactive way in Excel. I also used *F.DIST.RT* for computing P-Values in the ANOVA tables.

3 Examples

As it is mentioned, Excel was used for almost all taught methods in the course. I will present an example to show that a practical example can be presented in the class with details of methods and computation of formulas in a reasonable time in the class.

We present Example 5.3. of the text book [4]. As an example of a factorial design involving two factors, an engineer is designing a battery for use in a device that will be subjected to some extreme variations in temperature. The only design parameter that he can select at this point is the plate material for the battery, and he has three possible choices. When the device is manufactured and is shipped to the field, the engineer has no control over the temperature extremes that the device will encounter, and he knows from experience that temperature will probably affect the effective battery life. However, temperature can be controlled in the product development laboratory for the purposes of a test. "The engineer decides to test all three plate materials at three temperature levels-15, 70, and 125°F- because these temperature levels are consistent with the product end-use environment. Because there are two factors at three levels, this design is sometimes called a 3^2 factorial design. Four batteries are tested at each combination of plate material and temperature, and all 36 tests are run in random order. The experiment and the resulting observed battery life data are given in Figure 1 format.
The general model is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

We are interested in testing hypotheses about the equality of row treatment effects, say

$$\begin{cases} H_0: \tau_1 = \tau_2 = \tau_3 = 0 \\ H_1: \ Atleat\ one\ \tau_i \neq 0 \end{cases} \qquad (1)$$

and the equality of column treatment effects, say

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: \ Atleat\ one\ \beta_i \neq 0 \end{cases} \qquad (2)$$

We are also interested in determining whether row and column treatments interact.

$$\begin{cases} H_0: (\tau\beta)_{ij} = 0 \qquad for\ all\ i\ and\ j \\ H_1: \ Atleat\ one\ (\tau\beta)_{ij} \neq 0 \end{cases} \qquad (3)$$

To test (1), (2) and (3) we should compute

$$y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{i..} = \frac{y_{i..}}{bn} \qquad i = 1, 2, \ldots, a$$

$$y_{.j.} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{.j.} = \frac{y_{.j.}}{an} \qquad j = 1, 2, \ldots, b$$

$$y_{ij.} = \sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{ij.} = \frac{y_{ij.}}{n} \qquad \begin{matrix} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{matrix}$$

$$y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad \bar{y}_{...} = \frac{y_{...}}{abn}$$

Which can be done quickly by using *PivotTable* in Excel and the result will be as follows

Table 1: Using PivotTable to compute summations over the rows, the columns factors and replication.

Sum of I Colur	15				15 Total	70				70 Total	125				125 Total	(Grand Tot
Row	1	2	3	4		1	2	3	4		1	2	3	4		(blank)
1	130	74	155	180	539	34	80	40	75	229	20	82	70	58	230	998
2	150	159	188	126	623	136	106	122	115	479	25	58	70	45	198	1300
3	138	168	110	160	576	174	150	120	139	583	96	82	104	60	342	1501
(blank)																
Grand T	418	401	453	466	1738	344	336	282	329	1291	141	222	244	163	770	3799

This is reproduction of Table 5.4 of the text book.
At this stage, to copy and paste Table 1 in two tables of observations and the rows totals, values in the gray cells, to other cells, Table 2 (a) and (b). It is better to paste them as *Values* to avoid unwanted changes.

Figure 1: *Values (V)* Paste option.

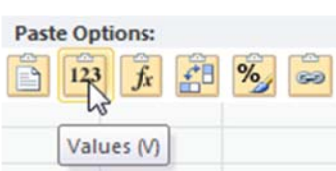


Table 2:

a) Copy and Paste of observations from Table1.

	15				70				125				
Row	1	2	3	4	1	2	3	4	1	2	3	4	
1	130	74	155	180	34	80	40	75	20	82	70	58	998
2	150	159	188	126	136	106	122	115	25	58	70	45	1300
3	138	168	110	160	174	150	120	139	96	82	104	60	1501
Totals	418	401	453	466	344	336	282	329	141	222	244	163	3799

b) Copy and Paste of the row totals from Table1.

15 Total	70 Total	125 Total	Grand Total
539	229	230	998
623	479	198	1300
576	583	342	1501

Then, using square opration, ^2, and drag methods preduce squares of Table 2 (a) and (b).

We can now calculate the sum of squares required in ANOVA table from Table 2.

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn}$$

$$SS_A = \frac{1}{bn}\sum_{i=1}^{a} y_{i..}^2 - \frac{y_{...}^2}{abn}$$

$$SS_{\text{Subtotals}} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij.}^2 - \frac{y_{...}^2}{abn}$$

$$SS_{AB} = SS_{\text{Subtotals}} - SS_A - SS_B$$

$$SS_E = SS_T - SS_{AB} - SS_A - SS_B$$

$$SS_E = SS_T - SS_{\text{Subtotals}}$$

$$SS_B = \frac{1}{an}\sum_{j=1}^{b} y_{.j.}^2 - \frac{y_{...}^2}{abn}$$

They are computed just using row, column *sum()*and the arithmetic operations. We are now able to create the table of ANOVA, Table 3. The first column is just the sum of squares, the second is degree of freedoms, the next column is produced by dividing the first column by the second column and the forth column is dividing the third column by *SSE*. The last column is the P-values for which function “=F.DIST.RT(F0,Df,DFE)” is used. When one uses those software with clickable interfaces the P-values are provided out of the dark so that students gradually forget its concept and they only know that H_0 is rejected if it is small but we can emphasize that it is probability of observing the current observed data under H_0 using Excel. We can also have a flashback on the F distribution structure.

Table 3: The ANOVA table.

	SS	DF.	MS	F_0	P-value
SSA=	10683.7	2	5341.8611	7.911	0.001976
SSB=	39118.7	2	19559.361	28.97	1.91E-07
SSAB=	9613.78	4	2403.4444	3.56	0.018611
SSE=	18230.8	27	675.21296		
SST=	77647	35	2218.4849		

Solving this example took less than 30 minutes in the class while I could review all computations and concepts. More importantly, students were following the steps very well.

4 Students feedback

Students were quite comfortable with Excel and they expressed their appreciation for using Excel. To test their level of learning through Excel I gave the following question in the final exam and restricted them to use Excel only. The question is an Analysis of Covariance problem from Chapter 15 of the test book.

Question 4

An engineer is studying the effect of cutting speed on the rate of metal removal in a machining operation. However, the rate of metal removal is also related to the hardness of the test specimen. Five observations are taken at each cutting speed. The amount of metal removed (*y*) and the hardness of the specimen (*x*) are shown in the following table. Analyze the data using and analysis of covariance. Use α=0.05.

		Cutting	Speed	(rpm)	
1000	1000	1200	1200	1400	1400
y	x	y	x	y	x
8	60	52	105	58	115
30	80	34	80	22	72
38	90	5	60	13	64
17	65	14	65	32	81
28	76	25	73	20	70

a) Write appropriate model indicating the response, the factor, and the covariate variables.

b) Compute all necessary S.., T.. and E..'s, and test the hypothesis that there is no treatment effect at level 0.05.

c) Test the hypothesis that the covariate variable (the hardness) has no effect at 0.05.

I had 23 students in this course. The mean mark for the final exam was 23.12 out of 35 while the mean score for Question 4 was 7.1 out of 10. This means that the mean percentage for Question 4 was 71% while for that of final exam was 66%. This shows that student did slightly better in Question 4. An example of a student's Excel sheet in the final exam is giving in Figure 3.

4 Conclusion

I used Excel effectively in teaching Design of Experiment for statistics students and I found it useful to teach "technique-oriented" exercises in the class. Many example were taught using Excel. It was mainly used as a calculator. Our students were comfortable to use it and express their preference of using Excel. However, I do not claim Excel is the best software to be used but we significantly spent less time to teach it compare to a statistical packages like R that requires extensive learning and training with some programming skills. Nevertheless, it can be a good solution when no TA available to assist instructor for teaching statistical package and students have problem to use a programmable software.

References:

[1] Chatfield C. Teaching a Course in Applied Statistics, *Journal of the Royal Statistical Society. Series C (Applied Statistics),* Vol. 31, No. 3, 1982 pp. 272-289

[2] Goos P. and Leemans H., *Journal of Statistics Education* Vol.12, No 3, 2004), www.amstat.org/publications/jse/v12n3/goos.html.

[3] Montgomery, D. C. Design and Analysis of Experiments, 8th edition, John Wiley & Sons, New York, 2013.

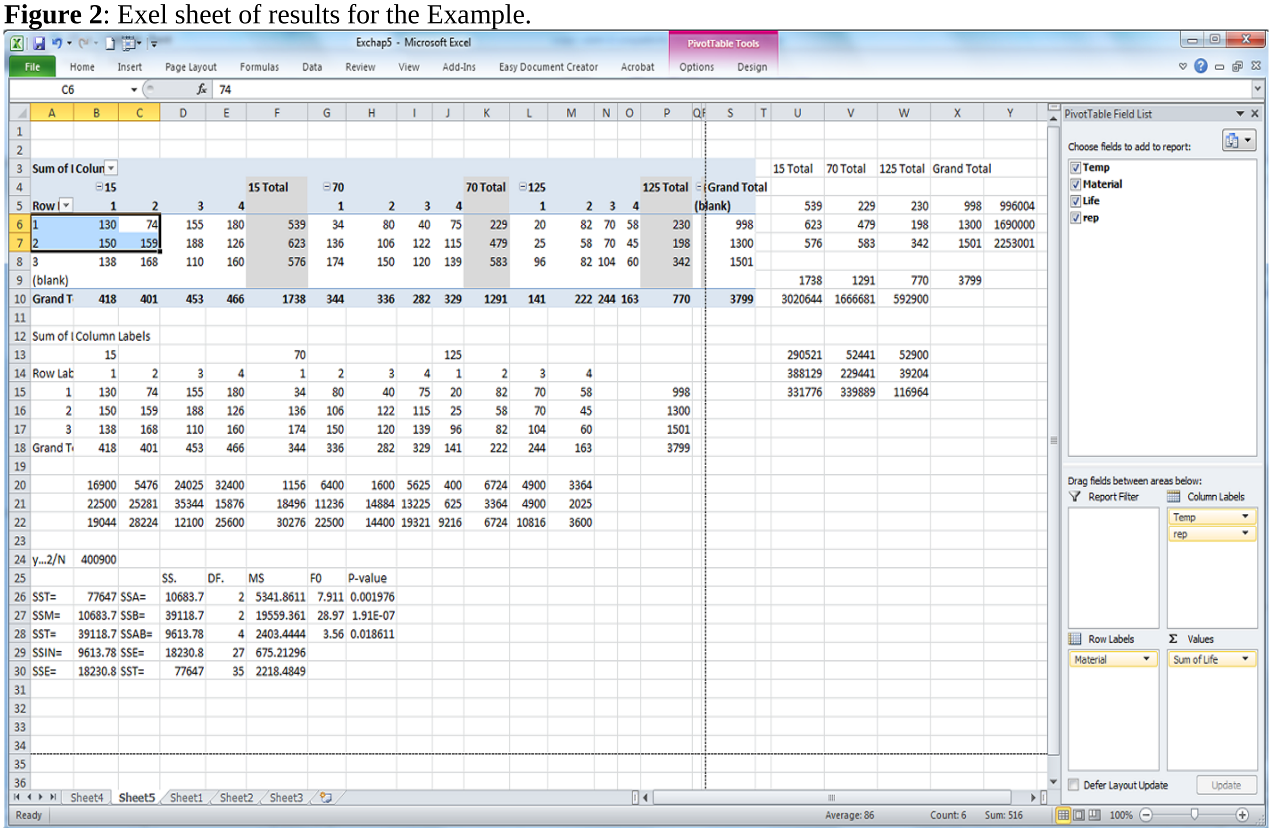
[4] Paura L., Arhipova I. Advantages and Disadvantages of Professional and Free Software for Teaching Statistics, *Information*

Technology and Management Science, Vol. 15, 2012, pp 9-14.

[5] Singer J. D. and Willett J. B, Improving the Teaching of Applied Statistics: Putting the Data Back into Data Analysis, *The American Statistician*, Vol. 44, No.3, 1990, 224-230.

[6] The R Project for Statistical Computing. [Online]. Available: http://www.R-project.org/

Figure 2: Exel sheet of results for the Example.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	S	U	V	W	X	Y
3	Sum of l Colun																	15 Total	70 Total	125 Total	Grand Total	
4		15				15 Total	70				70 Total	125				125 Total	Grand Total					
5	Row l	1	2	3	4		1	2	3	4		1	2	3	4		(blank)	539	229	230	998	996004
6	1	130	74	155	180	539	34	80	40	75	229	20	82	70	58	230	998	623	479	198	1300	1690000
7	2	150	159	188	126	623	136	106	122	115	479	25	58	70	45	198	1300	576	583	342	1501	2253001
8	3	138	168	110	160	576	174	150	120	139	583	96	82	104	60	342	1501					
9	(blank)																	1738	1291	770	3799	
10	Grand T	418	401	453	466	1738	344	336	282	329	1291	141	222	244	163	770	3799	3020644	1666681	592900		
12	Sum of l	Column Labels																				
13		15				70				125								290521	52441	52900		
14	Row Lab	1	2	3	4	1	2	3	4	1	2	3	4					388129	229441	39204		
15	1	130	74	155	180	34	80	40	75	20	82	70	58			998		331776	339889	116964		
16	2	150	159	188	126	136	106	122	115	25	58	70	45			1300						
17	3	138	168	110	160	174	150	120	139	96	82	104	60			1501						
18	Grand T	418	401	453	466	344	336	282	329	141	222	244	163			3799						
20		16900	5476	24025	32400	1156	6400	1600	5625	400	6724	4900	3364									
21		22500	25281	35344	15876	18496	11236	14884	13225	625	3364	4900	2025									
22		19044	28224	12100	25600	30276	22500	14400	19321	9216	6724	10816	3600									
24	y...2/N	400900																				
25				SS.	DF.	MS	F0	P-value														
26	SST=	77647	SSA=	10683.7	2	5341.8611	7.911	0.001976														
27	SSM=	10683.7	SSB=	39118.7	2	19559.361	28.97	1.91E-07														
28	SST=	39118.7	SSAB=	9613.78	4	2403.4444	3.56	0.018611														
29	SSIN=	9613.78	SSE=	18230.8	27	675.21296																
30	SSE=	18230.8	SST=	77647	35	2218.4849																

Figure3: Student's example in final exam.

finalExam Jawaher Saad AlMesaifri - Microsoft Excel

E19 =E12-B$22*(E15-E$18)

	A	B	C	D	E	F
1						
2			Cutting	Speed	(rpm)	
3	1000	1000	1200	1200	1400	1400
4	y	x	y	x	y	x
5	8	60	52	105	58	115
6	30	80	34	80	22	72
7	38	90	5	60	13	64
8	17	65	14	65	32	81
9	28	76	25	73	20	70
10						
11						
12	Syy=	3173.6		Y1.Bar	24.2	
13	Sxx=	3556.93		Y2.Bar	26	
14	Sxy=	3307.6		Y3.Bar	29	
15	Tyy=	58.8		X1.Bar	74.2	
16	Txx=	97.7333		X2.Bar	76.6	
17	Txy=	75.8		X3.Bar	80.4	
18	Eyy=	3114.8		X..Bar	77.0666667	
19	Exx=	3459.2		adjY1.	26.8782185	
20	Exy=	3231.8		adjY2.	26.4359891	
21				adjY3.	25.8857925	
22	Bat=	0.93426				
23						
24						
25						

	I	J	K	L	M	N
1	X	Y	CuttingSpeed	X^2	Y^2	XY
2	60	8	1000	3600	64	480
3	80	30	1000	6400	900	2400
4	90	38	1000	8100	1444	3420
5	65	17	1000	4225	289	1105
6	76	28	1000	5776	784	2128
7	105	52	1200	11025	2704	5460
8	80	34	1200	6400	1156	2720
9	60	5	1200	3600	25	300
10	65	14	1200	4225	196	910
11	73	25	1200	5329	625	1825
12	115	58	1400	13225	3364	6670
13	72	22	1400	5184	484	1584
14	64	13	1400	4096	169	832
15	81	32	1400	6561	1024	2592
16	70	20	1400	4900	400	1400
17	1156	396	18000	92646	13628	33826
18		10454.4				
19						
20	Row Labels	Sum of Y	Sum of X	Xi.^2	Yi.^2	Xi.Yi.
21	1000	121	371	137641	14641	44891
22	1200	130	383	146689	16900	49790
23	1400	145	402	161604	21025	58290
24	Grand Total	396	1156	445934	52566	152971