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. 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 Excells 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.

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

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad i = 1, 2, ..., a$$

$$j = 1, 2, ..., b$$

$$k = 1, 2, ..., n$$

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

$$H_0: \tau_1 = \tau_2 = \tau_3 = 0$$

$$H_1: \text{Atleast one } \tau_i \neq 0$$

(1)
and the equality of column treatment effects, say

\[
\begin{align*}
H_0: \beta_1 &= \beta_2 = \beta_3 = 0 \\
H_1: \text{Atleast one } \beta_i &\neq 0
\end{align*}
\] (2)

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

\[
\begin{align*}
H_0: (\tau\beta)_{ij} &= 0 \text{ for all } i \text{ and } j \\
H_1: \text{Atleast one } (\tau\beta)_{ij} &\neq 0
\end{align*}
\] (3)

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

\[
\begin{align*}
\bar{y}_c &= \frac{\sum_{j=1}^{b} \sum_{i=1}^{a} y_{ij}}{bn} & i &= 1, 2, \ldots, a \\
\bar{y}_r &= \frac{\sum_{j=1}^{b} \sum_{i=1}^{a} y_{ij}}{an} & j &= 1, 2, \ldots, b \\
\bar{y}_b &= \frac{\sum_{j=1}^{b} \sum_{i=1}^{a} y_{ij}}{n} & j &= 1, 2, \ldots, a \\
\bar{y}_a &= \frac{\sum_{j=1}^{b} \sum_{i=1}^{a} y_{ij}}{bn} & i &= 1, 2, \ldots, a \\
\end{align*}
\]

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.

<table>
<thead>
<tr>
<th>Row</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>70</td>
<td>130</td>
<td>220</td>
<td>350</td>
<td>230</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>1300</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>90</td>
<td>160</td>
<td>250</td>
<td>380</td>
<td>260</td>
<td>180</td>
<td>270</td>
<td>300</td>
<td>1350</td>
</tr>
<tr>
<td>3</td>
<td>140</td>
<td>110</td>
<td>180</td>
<td>270</td>
<td>400</td>
<td>290</td>
<td>210</td>
<td>320</td>
<td>350</td>
<td>1410</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
<td>130</td>
<td>200</td>
<td>300</td>
<td>430</td>
<td>320</td>
<td>240</td>
<td>360</td>
<td>400</td>
<td>1420</td>
</tr>
<tr>
<td>Total</td>
<td>480</td>
<td>340</td>
<td>470</td>
<td>620</td>
<td>800</td>
<td>520</td>
<td>350</td>
<td>470</td>
<td>510</td>
<td>2040</td>
</tr>
</tbody>
</table>

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 Table 1.

<table>
<thead>
<tr>
<th>Row</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>70</td>
<td>130</td>
<td>220</td>
<td>350</td>
<td>230</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>1300</td>
</tr>
<tr>
<td>2</td>
<td>120</td>
<td>90</td>
<td>160</td>
<td>250</td>
<td>380</td>
<td>260</td>
<td>180</td>
<td>270</td>
<td>300</td>
<td>1350</td>
</tr>
<tr>
<td>3</td>
<td>140</td>
<td>110</td>
<td>180</td>
<td>270</td>
<td>400</td>
<td>290</td>
<td>210</td>
<td>320</td>
<td>350</td>
<td>1410</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
<td>130</td>
<td>200</td>
<td>300</td>
<td>430</td>
<td>320</td>
<td>240</td>
<td>360</td>
<td>400</td>
<td>1420</td>
</tr>
<tr>
<td>Total</td>
<td>480</td>
<td>340</td>
<td>470</td>
<td>620</td>
<td>800</td>
<td>520</td>
<td>350</td>
<td>470</td>
<td>510</td>
<td>2040</td>
</tr>
</tbody>
</table>

Then, using square operation, $^2$, and drag methods produce squares of Table 2 (a) and (b).

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

\[ SS_T = \sum_{i=1}^{k} \sum_{j=1}^{b} \sum_{i=1}^{a} y_{ij}^2 - \frac{y^2}{abn} \]

\[ SS_A = \frac{1}{bn} \sum_{i=1}^{k} \sum_{j=1}^{b} y_{i.}^2 - \frac{y^2}{abn} \]

\[ SS_A = SS_{Subtotals} - SS_A - SS_B \]

\[ SS_E = SS_T - SS_A - SS_B \]

\[ SS_E = SS_T - SS_{Subtotals} \]

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.

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>$F_0$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSA</td>
<td>10683.7</td>
<td>2</td>
<td>5341.8611</td>
<td>7.911</td>
<td>0.001976</td>
</tr>
<tr>
<td>SSB</td>
<td>39118.7</td>
<td>2</td>
<td>19559.361</td>
<td>28.97</td>
<td>1.91E-07</td>
</tr>
<tr>
<td>SSAB</td>
<td>9613.78</td>
<td>4</td>
<td>2403.4444</td>
<td>3.56</td>
<td>0.018611</td>
</tr>
<tr>
<td>SSE</td>
<td>18230.8</td>
<td>27</td>
<td>675.21296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SST</td>
<td>77647</td>
<td>35</td>
<td>2218.4849</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\alpha=0.05$.

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

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:


Figure 2: Excel sheet of results for the Example.
Figure 3: Student’s example in final exam.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>Y</td>
<td>CuttingSpeed</td>
<td>X^2</td>
<td>Y^2</td>
<td>XY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
<td>8</td>
<td>1000</td>
<td>3600</td>
<td>64</td>
<td>2400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>1200</td>
<td>1200</td>
<td>1400</td>
<td>1400</td>
<td>80</td>
<td>30</td>
<td>1000</td>
<td>6400</td>
<td>900</td>
<td>2400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>110</td>
<td>120</td>
<td>130</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>65</td>
<td>70</td>
<td></td>
</tr>
</tbody>
</table>

The Excel spreadsheet shows the data related to cutting speed and its impact on the final exam results. The columns A to P contain various data points, with columns X and Y representing the cutting speed and its square values, respectively. The spreadsheet also includes calculations for the sum of squares and other statistical measures.