Learning ANOVA Concepts Using Simulation

Leslie Chandrakantha

Abstract: **Analysis of Variance (ANOVA) is an important topic in introductory statistics. Many students struggle to understand the ANOVA concepts. Statistical concepts are important in engineering education. In this paper, we describe how to use simulation with Excel Data Tables and standard functions to perform one-way ANOVA. We calculate different values of the *F*-statistic by resampling from the original sample and compute the *p*-value of the test. Using this approach, students will be able to get a better feel about the *p*-value concept. Our preliminary assessment shows that student learning is enhanced by incorporating this approach in the classroom.**

Index Terms: **ANOVA, Data Table, *F* statistic, Resampling.**

I. INTRODUCTION

STATISTICS can be applied in many situations in our everyday lives. Statistical knowledge is relevant in many areas such as engineering, business, natural science, and social science.Therefore, students need to take at least one course in statistics in many college degree programs. Understanding of statistical concepts such as sampling distributions, confidence intervals, hypothesis testing, and *p*-values is not easy for many students and many educators are conducting research on teaching and learning [2], [10], [13]. Computers have had a tremendous impact on teaching statistics at any level. Tishkovskaya et al. [14] have shown that the teaching and learning of statistics has benefited from the development of technological resources. The traditional way of teaching using books, lectures, and mathematical derivations does not give a good understanding of the concepts to many students in introductory level. Statistics instructors have been using computer simulation methods (CSM) in the classroom to teach difficult concepts. Computer simulation methods allow students to experiment with random samples for the purpose of clarifying abstract and difficult concepts of statistics. Cobb [4] noted that incorporating computer simulation to illustrate the key concepts and to allow students to discover important principles themselves enhances their knowledge. delMas et al. [5] evaluated the benefits of the computer simulation where students drew multiple samples from a variety of populations in order to observe the sampling distribution of the mean.

Leslie Chandrakantha is with the Department of Mathematics and Computer Science, John Jay College of Criminal Justice of the City University of New York, New York, NY 10019, USA. (phone: 212-237-8835,email:lchandra@jjay.cuny.edu).

Their evaluation of the simulation suggests that it provided an effective supplement to book and lecture based methods of instruction. Johnson [9] used Minitab macros to resample in hypothesis testing. In his work, he pointed out that minimal or no mathematical skills are required and that the resampling approach helps students understanding the statistical concepts. Black [1] was one of the first authors to recognize the usefulness of spreadsheets in a simulation context for teaching complex concepts of statistics. He used spreadsheet simulation to calculate the power of the tests as a planning tool. Hagtvedt et al. [7], [8] have developed a VBA based application tool that is integrated into Excel spreadsheets to simulate sampling distributions and confidence intervals. Their assessment showed that the students who used their applications did significantly better than the others. A comprehensive review of the literature of computer simulation methods used in all areas of statistics to help students understand difficult concepts has been done by Mills [11].

There are many on-line sources that allow students to simulate random variables and test statistics and to visualize their distributional properties. However, almost without exception, these on-line sources either build their simulation application using something other than a simple spreadsheet, or may not provide the source code. These web based applications are written in languages such as Java, JavaScript, or some other programming languages. Generally, introductory statistics students do not have necessary skills to produce codes that would replicate what can be found in these on-line sites. Furthermore, web-based applications require an internet connection to be used in the classroom or anywhere else. An Excel demonstration can be used with or without an internet connection. As long as the students have an access to a computer with Excel, they can perform these simulations and visualize the intended concepts.

In this paper, we describe how to use Excel standard formulas and the Data Table facility to perform simulation using resampling in teaching one-way analysis of variance (ANOVA). ANOVA is one of many important topics in the introductory statistics curriculum. ANOVA is commonly used in many areas to compare means of several populations. Rossman/Chance Applet Collection [12] includes a Java applet that simulates the ANOVA table for given sample sizes.

978-1-4799-5233-5/14/$31.00 ©2014 IEEE

Combination of Excel standard functions and Data Tables provides a convenient way to conduct simulation and resampling. Christie [3] used Excel Data Tables for estimating the population mean, correlation, and hypothesis testing.

This paper is organized as follows. The next section gives a brief overview of one-way ANOVA and an example. The next section introduces the simulation of ANOVA using Excel Data Tables. The later sections provide *p*-value calculations and the empirical distribution of the *F*-statistic, a comparison of two teaching methods, and concluding remarks.

II. OVERVIEW OF ANOVA

The one-way analysis of variance procedure, referred to as ANOVA, is used to test the following hypothesis: the means of three or more populations are the same against the alternative that not all population means are the same. ANOVA test requires that the populations from which samples are drawn have the same variance, σ^2.

The test performed by calculating two estimates of the variance, σ^2, of population distributions: the variance between samples and the variance within samples. The variance between samples is also known as mean square between samples (MSB) and the variance within samples is also known as mean square within samples (MSW). Both MSB and MSW estimate the variance of populations, σ^2. MSB is based on the values of the means of the samples taken from populations and MSW is based on the individual values in the samples. If the means of the populations under consideration are not equal, the variation among the means of respective samples is expected to be large, and, therefore, the value of MSB is expected to be large.

The value of the test statistic, *F,* for the ANOVA test is calculated as

$$F = \frac{Variance\ between\ samples}{Variance\ within\ samples} = \frac{MSB}{MSW}.$$

This test statistic has the *F* distribution with degrees of freedom *k-1* and *n-k* respectively, where *k* is the number of populations under consideration and *n* is the number of data values in all samples. The formulas for calculating MSB and MSW can be found in any introductory statistics text. The one-way ANOVA test is always right-tailed and the *p*-value is computed using the right tail of the *F* distribution curve.

Example: The following example is used to demonstrate the simulation of ANOVA test to the class. A university employment office wants to compare the amount of time taken by graduates with three different majors to find their first full time job after graduation. *Table I* lists the times (in days) for random samples of business, computer science, and engineering majors who graduated recently.

TABLE I
SAMPLE OF TIMES (IN DAYS)

Business	208	162	240	180	148	312	176	292
Computer Science	136	113	281	128	305	147	232	
Engineering	126	275	363	146	298	392		

Can we conclude that the mean time taken to find their first full time job for all graduates in these fields is the same?

Under the null hypothesis, the data comes from a single population and we are interested in how often the value of the test statistic as large as the one we observe will occur by chance if, in fact, this hypothesis is true.

III. SIMULATION USING EXCEL DATA TABLES

Data Tables are part of a group of commands that are called what-if analysis tools in Excel. When you use Data Tables, you are doing a what-if analysis. What-if analysis is the process of changing the values in cells to see how those changes will affect the outcome of formulas on the worksheet. We can use the Data Table function to compute the values of a test statistic for different random samples of data. A valuable introduction to Data Tables is given in Ecklund [6].

To generate the values of a statistic for different samples using the Data Table, first, calculate the value of the statistic based on a random sample. This will be our original output value. Now set up our Data Table by putting this value (formula for this output value) in the top cell of the right column. Leave the left column blank and select the table and access the Data Table dialog box using menu bar **Data > What If Analysis > Data Tables**. For the column input cell in the Data Table dialog box, type an empty cell reference that has no part in this Data Table set up. Excel generates a new sample and computes the value of the statistic for each substitution of this empty input cell and fills the table. Copying the formula down of the output cell does not work in this case. If we do this manually, we need to recalculate the statistic by repeated sampling by pressing **F9** key and recording these values in the column. The Data Table function recalculates the values whenever we change the spreadsheet. Some times this may slow down the speed of the work. This can be avoided by turning off the calculation by using the menu bar **File> Options > Formulas > Automatic except for data tables > OK**. You can recalculate the table again by pressing the **F9** key.

Spreadsheet Implementation of the Example

In this section, we show how to resample from given data, generate 1000 resamples, and calculate the *F*- test statistic values assuming data come from a single population. The *Figure1* shows the spreadsheet implementation of this example. We separate our data into three sets at random and calculate the value of the test statistic. This value is recorded in the Data Table in *Figure 1*.

Put the data in a single column with the three groups going from B3:B10 (Business majors), B11:B17 (Computer Science majors), and B18:B23 (Engineering majors). Under the null hypothesis, times for all three majors have a single population, and it will not matter how they are allocated between groups. Column D contains simple random numbers generated from the RAND function. Column E finds the rank order of these random numbers using RANK function to produce an ordering to resample from the original data in B3:B23. Now VLOOKUP function and the random ordering created in column E are used to generate three samples from the data in B3:B23. These resampled values are given in F3:F23. Sum of the data values for three groups are calculated in cells B25:B27 and F25:F27 for the original data and resampled data respectively. In cells B28:B29 and F28:F29, sum of the data values and sum of the squares of data values for entire sample are calculated for the original and resampled data. Cells B30:B31 and F30:F31 give the sum of squares between groups (SSB) and sum of the squares within groups (SSW). Finally in cells B32 and F32, the value of the *F* test statistic is calculated for the original data and resampled values.

The values of the resampled test statistic are calculated using the Data Table. The Data Table goes down as far as you want, but in this work, we create 1000 value. The value in cell F32 will now be the base value of our Data Table. *Figure 1* shows only a portion of the spreadsheet of this implementation. Our Data Table has the cell range H2:I1001 which calculates 1000 values of the statistic. The top right cell of the table contains the formula = F32. The left column of the table is left blank. Select the cell range H2:I1001, and use the commands **Data > What-if Analysis > Data Table**. In the Data Table dialog box, leave Row input cell blank and select any empty cell (say J1) for Column input cell and click **OK**. This will fill the right column with 1000 values of the statistic for repeated resamples. Pressing the **F9** key will recalculate a different set of values of the statistic. *Table II* shows the summary of the Excel formulas written in the spreadsheet implementation in *Figure 1*.

TABLE II
EXCEL FORMULAS

Cell	Formula	Purpose
D3	**=RAND( )**	Generate simple random numbers.
E3	**=RANK(D3,$D$3:$D$23)**	Create a rank ordering in values in column D.
F3	**=VLOOKUP(E3,$A$3:$B$23,2)**	Resample the values in column B.
B25	**=SUM(B3:B10)**	Compute the sum of the values for Business majors.
B26	**=SUM(B11:B17)**	Compute the sum of the values for Computer Science majors
B27	**=SUM(B18:B23)**	Compute the sum of the values for Engineering majors
B28	**=SUM(B3:B23)**	Compute the sum of the values in the sample.
B29	**= SUMPRODUCT(B3:B23,B3:B23)**	Compute the sum of the squares of the values in the sample
B30	**=POWER(B25,2)/8+POWER(B26, 2)/7+POWER(B27,2)/6 - POWER(B28,2)/21**	Compute the sum of squares between groups (SSB).
B31	**=B29-(POWER(B25,2)/8 + POWER(B26,2)/7 + POWER(B27,2)/6)**	Compute the sum of squares within groups (SSW).
B32	**=(B30/2)/(B31/18)**	Compute the value of the *F*- statistic.
I2	**=F32**	Top cell of the Data Table. Use this value to compute the Data Table
L3	**=(COUNTIF(I2:I1001,">=1.297796"))/1000**	Compute the *p*-value. This will count how many cells in range I2:I1001 are greater than or equal to 1.297796 and divide this count by 1000 to find the proportion

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Original				Resample		Data Table					
2	Index	Times		Random	Index	Times			3.20951				
3	1	208		0.9092	4	180			1.16		*P*-value	0.297	
4	2	162		0.18528	19	146			0.55474				
5	3	240		0.35892	14	147			0.17712				
6	4	180		0.95777	2	162			0.37713				
7	5	148		0.75416	10	113			0.37965				
8	6	312		0.90762	5	148			7.25522				
9	7	176		0.8735	7	176			1.59987				
10	8	292		0.12696	20	298			0.11441				
11	9	156		0.24643	16	126			0.78645				
12	10	113		0.876	6	312			1.90691				
13	11	281		0.03976	21	392			0.27625				
14	12	128		0.64031	11	281			2.52372				
15	13	305		0.93458	3	240			0.00723				
16	14	147		0.80715	9	156			0.99758				
17	15	232		0.20703	18	363			1.16592				
18	16	126		0.98571	1	208			0.12488				
19	17	275		0.61304	12	128			0.86743				
20	18	363		0.82509	8	292			3.4172				
21	19	146		0.23214	17	275			0.26973				
22	20	298		0.31722	15	232			2.08012				
23	21	392		0.50636	13	305			0.01661				
24									0.4262				
25	T1	1718				1370			0.70363				
26	T2	1362				1870			0.97547				
27	T3	1600				1440			2.65829				
28	$\sum x$	4680				4680			3.51441				
29	$\sum x^2$	1182958				1182958			0.41701				
30	SSB	17642.02				36798.214			4.02755				
31	SSW	122344.5				103188.36			1.85904				
32	F-Ratio	1.297796				3.2095087			0.88721				
33									1.53665				

Fig. 1: Portion of the spreadsheet showing simulated values of the statistic and the *p*-value

IV. *p*-VALUE AND EMPIRICAL DISTRIBUTION

In order to evaluate whether a value of the test statistic as large as the one from observed observations is likely, we simply compute the proportion of the 1000 resamples where the tests statistic values given in the Data Table exceed the observed value which is calculated in cell B32. This is known as the *p*-value of the test and it is the proportion of values of the test statistic, assuming data come from a single population, is as extreme or more extreme than the observed value of the test statistic. This is calculated in cell L3 in *Figure 1* using the formula = (COUNTIF(I2:I1001, ">=1.297796"))/1000 as 0.297. The actual *p*-value based on the *F* distribution with 2 and 18 degrees of freedom can be calculated using the Excel formula =F.DIST.RT(1.297796,2,18,) and it is 0.2975. This value agrees with our simulated value up to three decimal places. Interpreting this *p*-value, we can say that it is fairly likely one would observe a value of the statistic as large as, or larger than the one produced even when there is no difference in terms of the times to find their first full time job for graduates of three majors under consideration. This relatively larger *p*-value provides sufficient evidence for the claim that "the mean time taken to find their first full time job for all graduates in these fields is the same".

The Excel histogram in *Figure 2* shows the empirical distribution of the test statistic for 1000 resamples. The shaded area approximately represents the number of values exceeding the observed value of the statistic. The density curve of the *F* distribution with 2 and 18 degrees of freedom is shown in the smaller graph. The histogram of values of the statistic suggests that they are very closely distributed as an *F* distribution with 2 and 18 degrees of freedom. This will confirm the idea that we can use the simulation approach to perform the ANOVA test.

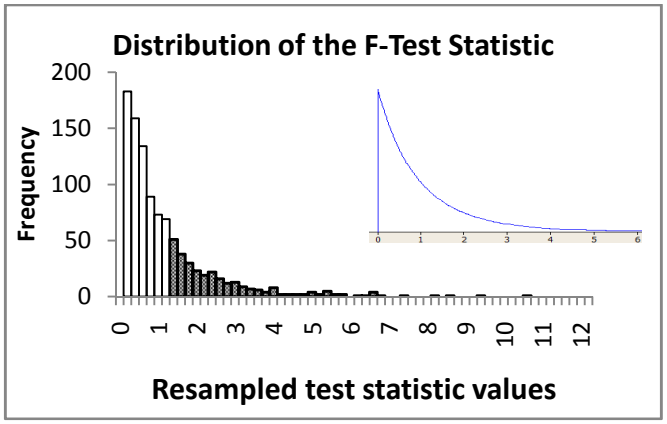

Fig. 2: Histogram of the values of the *F* - test statistic

V. COMPARISON WITH EXISTING TEACHING METHODS

We have taught two introductory statistics sections last semester, one using computer simulation methods (CSM) with Excel described in this paper and the other by the traditional method of not using simulation. Both classes have the same course content, same exams, same quizzes, and same assignments. At the end of the lesson, the following survey was conducted to know students' attitudes about the method of teaching.

Q1. Method helps me to understand the concepts.
Q2. Feeling like I am part of the discussion.
Q3. Feeling comfortable taking part in the lesson.
Q4. Visual representation of outcomes is useful in understanding.
Q5. Recommend this approach to other students.

Table III gives the summary of their responses:

TABLE III
SUMMARY OF STUDENT RESPONSES

	CSM Method Class			Traditional Method Class		
Question	Yes	No	No Opinion	Yes	No	No Opinion
Q1	56%	24%	20%	39%	35%	26%
Q2	68%	20%	12%	35%	43%	22%
Q3	64%	24%	12%	57%	17%	26%
Q4	60%	24%	16%	57%	13%	30%
Q5	64%	16%	20%	52%	35%	13%

Table IV shows the final exam scores statistics:

TABLE IV
EXAM SCORE STATISTICS

Class	n	Mean	Median	Std. Dev.
CSM used	25	76.20	77	15.12
Traditional method used	23	68.61	68	14.34

The majority of the students in the CSM class answered yes to all five questions while majority of traditional method class students answered yes to only questions 3, 4, and 5. Looking at the percentages, we observe that students are understanding concepts better and feeling more comfortable in computer simulation approach. The two sample t-test was performed using Excel to test the hypothesis that the CSM class performs better on average than the traditional method class. The *p*-value produced by Excel was 0.0408 which indicates that CSM class performs significantly better at 0.05 level of significance. We have to caution that these sample sizes are not large enough to make a firm judgment on the conclusion. We plan to use larger sample sizes in the future in order to give a comprehensive survey and to make a formal assessment.

V. CONCLUSION

Until recent years, many statistics instructors used the traditional way of teaching which concentrated on the derivation of the test statistics whose sampling distribution can be derived mathematically. These mathematical derivations require skills beyond most students in introductory statistics classes. In contrast, the computer simulation approach described in this paper requires no mathematical skills and leads students to better understanding of the concepts. With the advancement of technology, we can now use computer intensive methods as an alternative to traditional methods of teaching statistics. We have demonstrated how to teach ANOVA concepts using simulation. This approach of using Excel Data Tables and formulas is relatively easy for students to comprehend the concepts by implementing in a spreadsheet and visualizing the sampling distributions and *p*-values. Excel is easy to use software and many students have access to it. Our preliminary assessment suggests that the computer simulation approach will enhance students understanding of concepts.

REFERENCES

[1] T. R. Black, "Simulation on spreadsheets for complex concepts: Teaching statistical power as an example," *International Journal of Mathematical Education in Science and Technology*, vol. 30, no. 4, pp. 473-481, 1999.

[2] H. Callaert, 2002. "Understanding statistical misconceptions." *Proc. of the 6th Int. Conference on the Teaching of Statistics.* Cape Town, South Africa, July 2002..

[3] D. Christie, "Resampling with excel." *Teaching Statistics*, vol. 26. no. 1, pp. 9 –14, 2004.

[4] P. Cobb, "Where is the mind? Constructivist and sociocultural perspectives on mathematical development." *Educational Researcher*, vol. 23, pp. 13-20, 1994.

[5] R. C. delMas, J. Garfield, and B. L. Chance, " A model of classroom research in actions: Developing simulation activities to improve students'sttatistical reasoning." *Journal of Statistics Education*, vol. 7, no.3,1999,Available: http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm.

[6] P. Ecklund, "Introduction to excel 2007 data tables and data table exercises,"2009, Available:http://faculty.fuqua.duke.edu/~pecklund/ExcelReview/Excel%202007%20Data%20Table%20Notes.pdf.

[7] R. Hagtvedt, G. T. Jones, and K. Jones, K. "Pedagogical simulation of sampling distributions and the central limit Theorem." *Teaching Statistics*, vol. 29, no. 3, pp. 94 – 97, 2007.

[8] R. Hagtvedt, G. T. Jones, and K. Jones, "Teaching confidence intervals using simulation." *Teaching Statistics*, vol. 30, no. 2, pp. 53 – 56, 2008.

[9] R. W. Johnson, " An introduction to the bootstrap." *Teaching Statistics*, vol. 23, no. 2. pp. 49 – 54, 2001.

[10] C. Keeler and K. Steinhorst, "A new approach to learning probability in the first statistics course." *Journal of Statistics Education*. Vol. 9, no. 3, 2001,Available:. http://www.amstat.org/publications/jse/v9n3/keeler.html. .

[11] J. D. Mills, "Using computer simulation methods to teach statistics: Review of the Literature." *Journal of Statistical Education,* vol. 10, no. 1,2002,Available:. http://www.amstat.org/publications/jse/v10n1/mills.html.

[12] A. Rossman and B. Chance, "Rossman/Chance applet collection." 2011, available: http://www.rossmanchance.com/applets/.

[13] D.Tempelaar, "Modeling students learning of introductory statistics." *Proc. of the 6th Int. Conference on the Teaching of Statistics.* Cape Town, South Africa, July 2002.

[14] S. Tishkovskaya and G. Lancaster, "Statistical Education in the 21st century: a review of challenges, teaching innovations and srategies for reform," *Journal of Statistical Education*, vol. 20, no. 2, 2012, Available: http://www.amstat.org/publications/jse/v20n2/tishkovskaya.pdf.