# Analysis Of Variance With Summary Statistics In Microsoft® Excel®

David A. Larson, University of South Alabama, USA
Ko-Cheng Hsu, University of South Alabama, USA

**ABSTRACT**

*Students regularly are asked to solve Single Factor Analysis of Variance problems given only the sample summary statistics (number of observations per category, category means, and corresponding category standard deviations). Most undergraduate students today use Excel for data analysis of this type. However, Excel, like all other statistical software packages, requires an input data set in order to invoke its Anova: Single Factor procedure. The purpose of this paper is therefore to provide the student with an Excel macro that, given just the sample summary statistics as input, generates an equivalent underlying data set. This data set can then be used as the required input data set in Excel for Single Factor Analysis of Variance.*

**Keywords:** Analysis of Variance, Summary Statistics, Excel Macro

## INTRODUCTION

Most students have migrated to Excel for data analysis because of Excel's pervasive accessibility. This means, given just the summary statistics for Analysis of Variance, the Excel user is either limited to solving the problem by hand or to solving the problem using an Excel Add-in. Both of these options have shortcomings. Solving by hand means there is no Excel counterpart solution that puts the entire answer 'right there in front of the student'. Using an Excel Add-in is often not an option because, typically, Excel Add-ins are not available. The purpose of this paper, therefore, is to explain how to accomplish analysis of variance using an equivalent input data set and also to provide the Excel user with a straight-forward macro that accomplishes this technique. The technique works for Single Factor Analysis of Variance because, given a set of summary statistics (which are also the sufficient statistics in this instance), it is easy to verify the following two equations do generate an artificial data set with identical summary statistics [(2)]:

$$y_i = \bar{x} + \frac{s}{\sqrt{n}}$$

$$i = 1, 2, \ldots, n-1$$

$$y_n = n * \bar{x} - (n-1) * y_i$$

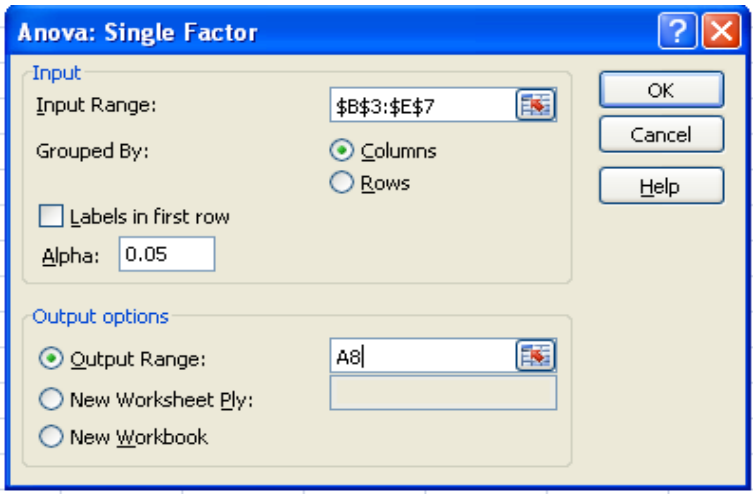
where, $\bar{x}$ = sample mean, s = sample standard deviation, and n = number of category observations

### Single Factor Analysis Of Variance In Excel When The Actual Sample Data Set Is Available

To begin with, assume that the required input data set is available with which to do an analysis of variance in Excel. An example input data set is shown below.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | Category | | | | | |
| 2 | Obs. | A | B | C | D | | |
| 3 | 1 | 7.6 | 8.5 | 6.8 | 7.4 | | |
| 4 | 2 | 8.3 | 8.7 | 6.7 | 6.5 | Actual | |
| 5 | 3 | 7.6 | 7.7 | 6.6 | 6.8 | Data Set | |
| 6 | 4 | | 8.3 | 6.4 | | | |
| 7 | 5 | | 8.7 | | | | |

In Excel, the following sequence is invoked in order to generate the analysis of variance solution for the above data: Data (tab) – Analysis (panel on right) – click on Data Analysis (in Analysis panel) – select Anova: Single Factor on the Data Analysis screen – OK. At this point in the sequence, the Anova: Single Factor screen (below) is filled in [note: the first requirement is the input data set (in above cells B3:E7, in this case)].

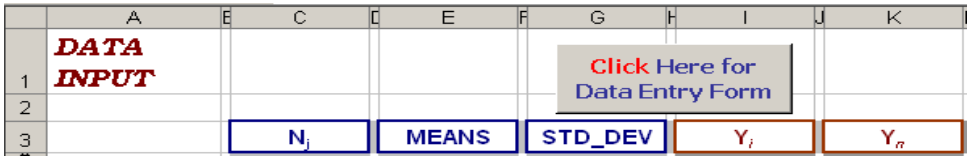



When OK is clicked on the above screen, the following Anova: Single Factor solution is generated by Excel (note: since Excel is a 'live' spreadsheet, the default output in this case has been 'cleaned up' appropriately).

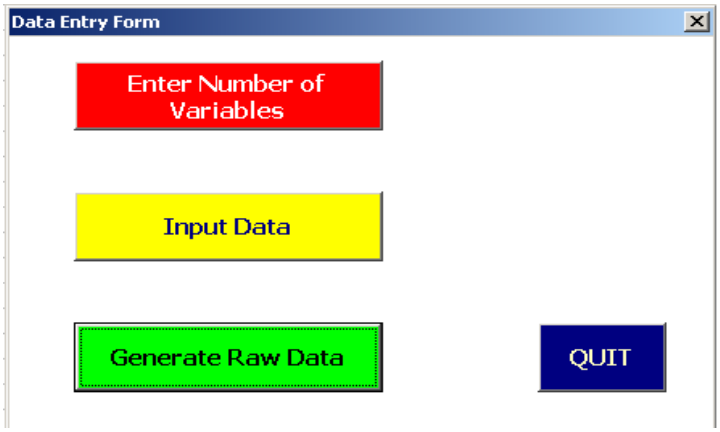
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | Category | | | | | |
| 2 | Obs. | A | B | C | D | | |
| 3 | 1 | 7.6 | 8.5 | 6.8 | 7.4 | | |
| 4 | 2 | 8.3 | 8.7 | 6.7 | 6.5 | Actual | |
| 5 | 3 | 7.6 | 7.7 | 6.6 | 6.8 | Data Set | |
| 6 | 4 | | 8.3 | 6.4 | | | |
| 7 | 5 | | 8.7 | | | | |
| 8 | Anova: Single Factor | | | | | Corresponding | |
| 9 | *Category (Cat)* | *Nj* | *Cat_Sum* | *Cat_Avg* | *Cat_Variance* | Cat_Std Dev | |
| 10 | A | 3 | 23.5 | 7.833 | 0.16333 | 0.40415 | Summary |
| 11 | B | 5 | 41.9 | 8.380 | 0.17200 | 0.41473 | Statistics |
| 12 | C | 4 | 26.5 | 6.625 | 0.02917 | 0.17078 | |
| 13 | D | 3 | 20.7 | 6.900 | 0.21000 | 0.45826 | |
| 14 | 4 | 15 | 112.6 | 7.50667 | 0.13838 | | |
| 15 | K | Nt | ∑allcats | Wgt Avg Avgs | Wgt Avg Var | | |
| 16 | ANOVA | | | | | | |
| 17 | *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value(ob)* | *F*.05,df=3,11* |
| 18 | Between Categories | 8.347 | 3 | 2.782 | 20.11 | 0.00009 | 3.5874 |
| 19 | Within Categories | 1.522 | 11 | 0.13838 | | P-value(*) | 0.05 |
| 20 | Total | 9.869 | 14 | 0.705 | | | |

**Single Factor Analysis Of Variance In Excel When The Actual Sample Data Set Is Not Available (Only Summary Statistics Are Available)**

Next, the Excel generated summary statistics above (Nj in cells B10:B13, $\bar{X}$'s in cells D10:D13 and s's in cells F10:F13) are used as input into the macro in order to generate the required equivalent input data set.
The Excel macro is displayed below. After accessing the macro, the first screen, as shown immediately below, instructs you to Click Here for Data Entry Form [(2)].




The first macro screen above indicates that the inputs are the number of observations per category (Nj), the category means ( $\bar{X}$)'s, and the corresponding standard deviations (s's) for each category. The Yi and Yn are the corresponding generated observations which are used to build the equivalent data set. Click Here for Data Entry Form moves the user to the second screen (shown below).




On the second screen (shown above), you are asked to first enter the number of variables (≡ to categories or groups), which is 4 for the above example (A, B, C, and D). Then you click on Input Data and you begin with category A and, on separate screens, do the following: Enter Sample Size (Nj), Enter Sample Mean, and Enter Sample Standard Deviation. You do the same entry sequence for categories B, C, and D. When you have done this, you click on Generate Raw Data and the following information is displayed. [Note: the screens involved are clearly labeled and permit you to correct mistakes, if any]. After you select Generate Raw Data, the following results are displayed.

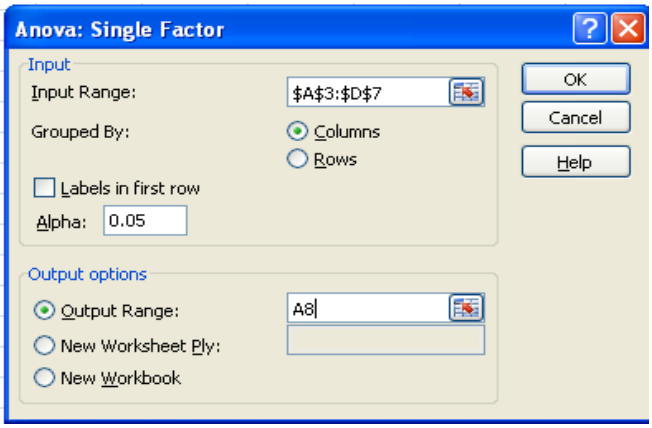
| | A | C | E | G | I | K | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | DATA INPUT | Number of Variable: | 4 | Click Here for Data Entry Form | | | RAW DATA | | | |
| 2 | | | | | | | VARIABLE | | | |
| 3 | | $N_j$ | MEANS | STD_DEV | $Y_i$ | $Y_n$ | 1 | 2 | 3 | 4 |
| 5 | 1 | 3 | 7.8333 | 0.4042 | 8.0666 | 7.3666 | 8.0666 | 8.5655 | 6.7104 | 7.1646 |
| 6 | 2 | 5 | 8.3800 | 0.4147 | 8.5655 | 7.6381 | 8.0666 | 8.5655 | 6.7104 | 7.1646 |
| 7 | 3 | 4 | 6.6250 | 0.1708 | 6.7104 | 6.3688 | 7.3666 | 8.5655 | 6.7104 | 6.3708 |
| 8 | 4 | 3 | 6.9000 | 0.4583 | 7.1646 | 6.3708 | | 8.5655 | 6.3688 | |
| 9 | | | | | | | | 7.6381 | | |

The above displayed values in C5:C8, E5:E8, and G5:G8 allow you to verify that you have entered correctly, for all four categories respectively, the number of observations (Nj), Means and Standard Deviations. The

values in I5:I8 and K5:K8 display the values used to construct the equivalent data set. The constructed equivalent data set is displayed in columns M through P on the above screen. In addition, the equivalent data set shown above is also placed (by itself) on a separate Excel sheet labeled DATA. This is done in order to facilitate entering the equivalent data set into Excel (especially in the cases of large equivalent data sets). In this case, the equivalent data set on the DATA sheet appears as follows;

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **A** | **B** | **C** | **D** |
| 2 | | | | |
| 3 | 8.066636111 | 8.565473 | 6.71039 | 7.164576534 |
| 4 | 8.066636111 | 8.565473 | 6.71039 | 7.164576534 |
| 5 | 7.366627777 | 8.565473 | 6.71039 | 6.370846931 |
| 6 | | 8.565473 | 6.36883 | |
| 7 | | 7.638108 | | |

The above data set is then easily used as the required input data set in the standard way for the Anova: Single Factor procedure in Excel. The Anova: Single Factor screen in this case is simply filled in as follows (Note: the above equivalent data set is obviously an artificial data set. In our opinion this is an important property of the surrogate data set because it reinforces the idea that the summary statistics are the sufficient statistics in this case.);




After OK is clicked on the above screen, the following Excel solution is generated;

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | **A** | **B** | **C** | **D** | | | |
| 2 | | | | | | | |
| 3 | 8.066636111 | 8.565473 | 6.71039 | 7.164576534 | | | |
| 4 | 8.066636111 | 8.565473 | 6.71039 | 7.164576534 | | | |
| 5 | 7.366627777 | 8.565473 | 6.71039 | 6.370846931 | | | |
| 6 | | 8.565473 | 6.36883 | | | | |
| 7 | | 7.638108 | | | | | |
| 8 | Anova: Single Factor | | | | | | |
| 9 | *Category (Catj)* | *Nj* | *Catj_Sum* | *Catj_Avg* | *Catj_Var* | *Catj_Std* | |
| 10 | A | 3 | 23.50 | 7.83 | 0.1633 | 0.4042 | |
| 11 | B | 5 | 41.90 | 8.38 | 0.1720 | 0.4147 | |
| 12 | C | 4 | 26.50 | 6.63 | 0.0292 | 0.1708 | |
| 13 | D | 3 | 20.70 | 6.90 | 0.2100 | 0.4583 | |
| 14 | 4 | 15 | 112.6 | 7.5067 | 0.1384 | | |
| 15 | K | Nt | Σallcats | Wgt Avg Avgs | Wgt Avg Var | | |
| 16 | ANOVA | | | | | | |
| 17 | *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value(ob)* | *F*.05,df=3,11* |
| 18 | Between Categories | 8.3471 | 3 | 2.7824 | 20.1067 | 0.0001 | 3.5874 |
| 19 | Within Categories | 1.5222 | 11 | 0.1384 | | P-value(*) | 0.05 |
| 20 | Total | 9.8693 | 14 | 0.7049 | | | |

## CONCLUSION

As the reader can verify, the Excel solution above is identical to the prior Excel solution which was obtained using the actual underlying data set. This means the student, given just summary statistics, can follow the standard Excel steps in generating the Excel answer. He or she, therefore, has, among other things, 'everything right there in front of them' for confirmation purposes when working through this type of problem. In addition, and very importantly, once the Excel solution is obtained using the equivalent data set methodology, the Tukey pair wise comparisons (if warranted), for example, are easily added to the Excel output obtained using the equivalent data set.

### Practice Problem

Golf balls were tested for hitting distance on a ball-driving machine. Four brands of golf balls were tested using random samples of seven golf balls of each brand. Summary statistics of the results are given below. In Excel, perform the correct Single Factor Analysis of Variance (use $\alpha = .05$) in this case to determine if mean distances are equal for all four brands of golf balls. What is the decision on the null hypothesis in this case? Explain.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Golf | # of | Sample | Sample |
| 2 | Ball | Balls Hit | Mean ( $\bar{X}$ ) Distance | Standard |
| 3 | Brands | Nj | in Yards | Deviation (s) |
| 4 | 1 | 7 | 280 | 11.4018 |
| 5 | 2 | 7 | 292 | 10.9545 |
| 6 | 3 | 7 | 276 | 11.0454 |
| 7 | 4 | 7 | 300 | 11.3137 |

Hint: Fob = 6.9746; P-value(ob) = .0018. [The interested reader can contact the authors if they wish to learn how to add the Tukey pair wise comparisons efficiently to the Excel solution for this problem.)

## AUTHOR INFORMATION

**David A. Larson** is a Lecturer in Statistics in the Mitchell College of Business at the University of South Alabama. He received his Ph.D. in Economics from the University of Maryland in College Park, Maryland.

**Ko-Cheng Hsu** is an Associate Professor of Accounting in the Mitchell College of Business at the University of South Alabama. He received his Ph.D. in Accounting from Memphis State University.

## REFERENCES

1. Microsoft® Office Excel® 2007, Copyright© Microsoft Corporation, 2006.
2. Larson, David A. "Analysis of Variance With Just Summary Statistics As Input", *The American Statistician*, 46, 151-152 (1992).

## NOTES

The Excel macro is available from the authors upon request. Additionally, a complete set of instructions for implementing the macro are will also be included. To obtain the items contact: dlarson@usouthal.edu [Practitioners who use Excel and obtain their data only in summary form may, of course, also obtain the macro and the corresponding implementation instructions.]

**NOTES**