Two-way ANOVA for Unbalanced Data: The Spreadsheet Way

K. Thomas Klasson

United States
Department of
Agriculture

Agricultural
Research Service

Southern Regional
Research Center

Technical Report
June 2019
Revised February 2020

K. Thomas Klasson is a Supervisory Chemical Engineer at USDA-ARS, Southern Regional Research Center, 1100 Robert E. Lee Boulevard, New Orleans, LA 70124; email: thomas.klasson@ars.usda.gov

Two-way ANOVA for Unbalanced Data: The Spreadsheet Way

K. Thomas Klasson

U.S. Department of Agriculture
Agricultural Research Service
Southern Regional Research Center
New Orleans, Louisiana, USA

Technical Report
June 2019, Revised February 2020

Abstract

The potential benefits of using spreadsheets in education is well documented and more use of spreadsheets examples have been encouraged. Understanding two-way Analysis of Variance (ANOVA) with unbalanced data is challenging and is often dismissed and handed over to dedicated statistical software program without knowing how the data are handled by those programs. This paper allows students, instructors, and researchers to use Excel spreadsheets to explore two-way ANOVA scenarios with unbalanced data. Supplementary material includes a complete spreadsheet with the examples used in the text and three different approaches on how to handle unbalanced data.

1. Introduction

The use of technology in education has long existed and implemented for different reasons [1]. Sometimes as a more effective method of transferring information. Sometimes for uniform training purposes and sometimes as a problem-based learning tool. When it comes to spreadsheets, educational research studies have shown that spreadsheet applications aid in promoting problem-solving skills in students and expanding their capabilities [1, 2, 3]. There is probably no better argument for the benefit of using spreadsheets in education than the excellent review article written for the first issue of *Spreadsheets in Education* by Baker and Sugden [4], who summarized the findings of many others from K-12 mathematics education to several secondary education areas of science such as Number Theory, Combinatorics, Numerical Analysis, Statistics, Physical Sciences, Computer Sciences, etc. The article concludes with "There is no longer a need to question the potential for spreadsheets to enhance the quality and experience of learning that is offered to students" and promotes improving access to computers and encouraging the development of more types of spreadsheets covering more topics. It is in that spirit that the enclosed spreadsheet was developed as a further expansion of spreadsheet use when teaching, understanding, and using two-way Analysis of Variance (ANOVA) statistics. The examples used are simply used to frame the discussion and spreadsheet development and does not constitute limitation of the area of implementation.

Students and researchers are often introduced to ANOVA statistics by first studying one-way ANOVA examples. Textbooks then move on to factorial ANOVA statistics, for example two-way ANOVA, but often this is limited to balanced data. Balanced data occur when the number of data values (replications) for each of the categories (or groups) is the same and all the desired categories contain data values. When this is the case, the calculations are relatively simple and equations can be written to construct an ANOVA table. It is rare that classes, even secondary education classes, ever address the case of unbalanced data; however, unbalanced data are almost always encountered in some science disciplines (often with human, animal, plant, or environmental subjects). To illustrate an example of unbalanced data, we can take data from Smith and Cribbie [5] and the results from a hypothetical survey of gambling behaviour of men and women with different personal connections to sports (former athletes, current athletes, and non-athletes). This is a classic 2x3 two-factor ANOVA (gender being one factor, with two levels, and athlete status being the second factor, with three levels). We will designate the first factor with the letter A and the second with the letter B. The data taken from Smith and Cribbie [5] are shown in Table 1. Noted is the unbalanced data with different number of scores for the different categories. We can imagine that a data value represents the summary score from a survey taken by one of the individuals belonging to a specific category (e.g., a female former athlete) and we are interested in how gender and athletic status impact gambling behaviour. When discussing data like this, it is convenient to say that the data are contained in cells (there are six data cells in Table 1) and the cells are arranged in rows and columns.

Table 1. Hypothetical gambling behaviour of women and men with different athlete status. The values within parenthesis are the mean values for the cells.

		Athletic status (Factor B)		
		Current athlete (B=1)	Former athlete (B=2)	Non-athlete (B=3)
Gender (Factor A)	Male (A=1)	3.0, 2.8, 3.0 (2.93)	5.1, 4.7, 4.9, 5.2, 4.9, 5.0 (4.97)	2.1, 2.0, 1.9, 1.8 (1.95)
	Female (A=2)	2.3, 2.1, 2.4 (2.27)	3.9, 3.8, 4.1 (3.93)	1.2, 1.1, 1.3, 1.1, 1.0 (1.14)

In most cases when ANOVA information is sought, the unbalanced data are entered into a statistical program such as SAS and the results are reported without much knowledge to the process and procedures that are used. There are, in general, three accepted types of ANOVA treatment of unbalanced data when all categories are represented (i.e., no empty cells). These types have been summarized by, for example, Herr [6] and Shaw and Mitchell-Olds [7]. Discussions about the preferred treatment have been a topic of many articles; even as recently as 2014 [5], which is remarkable as the methods originate from the 1930's [8]. This manuscript is not intended to make any unambiguous recommendations on which ANOVA treatment to use; it is simply an example on how spreadsheets can be used to explore the different ANOVA types of sums of squares. It also highlights how statistical software packages calculate the sums of squares. It should be noted that not all statistical software support the three types or the software defaults to one of the types [5, 9]. The spreadsheet developed here does not exclude or default to any of the three types. The typical ANOVA table for a two-way design is shown in Table 2. The three types of ANOVA treatments differ in the values of the entities *SSA* and *SSB* and the underlying hypotheses [5].

Table 2. Typical spreadsheet structure of a two-way ANOVA table with shaded cells where values are located. Below, the number of Factor A levels is N_A, the number of Factor B levels is N_B, and the total number of data values is N.

Source of Variability	Sums of Squares	Degrees of Freedom (*df*)	Mean Square Error (*MSE*)	*F* Ratio	Critical *F* value	p value
Factor A	SSA	$N_A–1$	SSA/df_A	MSE_A/MSE_E		
Factor B	SSB	$N_B–1$	SSB/df_B	MSE_B/MSE_E		
Interaction AxB	$SSAB$	$(N_A–1)(N_B–1)$	$SSAB/df_{AB}$	MSE_{AB}/MSE_E		
Error	SSE	$N–N_AN_B$	SSE/df_E			
Total	SST	$N–1$				

2. Examples used

The first example that we will use was introduced above. The second hypothetical example has been used by others as well. Here we consider how the final botanical plant height depends on whether or not weeds were removed around plants of two different initial sizes. This is a classical 2x2 two-factor ANOVA (weed status being one factor, with two levels, and initial size being the second factor, with two levels.) Again, we will designate the first factor with the letter A and the second factor with the letter B. The data were taken from Hector, von Felten, and Schmid [10] and Shaw and Mitchell-Olds [7] and are shown in Table 3. We can imagine that a data value represents the final height of a single plant belonging to a specific category (e.g., weeds removed around a short plant in early-growth).

Table 3. Hypothetical final plant heights of short and tall plants that had or had not weeds removed around them at early growth.

		Early growth height (Factor B)	
		Shorter than 4 inches (B=1)	Taller than 4 inches (B=2)
Weed status (Factor A)	Weeds not removed (A=1)	50, 57 (53.5)	91, 94, 102, 110 (99.25)
	Weeds removed (A=2)	57, 71, 85 (71)	105, 120 (112.5)

3. Theory

The theory of the two-way ANOVA for balanced data can easily be found in most statistics textbooks or on the web. The handling of unbalanced data goes back to the 1930's and the work of Frank Yates [6, 11], who first published on agricultural experimental data that were unbalanced [8]. Since then, numerous articles have discussed the work, the use of various statistical computer software, and how to visualize the data [5, 7, 9, 10]. Much of the published work focuses on the type of sums of squares used in the ANOVA table and what hypotheses they are addressing. As the three types of sums of squares are different, their values are often different and, thus, leading to different interpretation of the statistical significance of the different factors. In this manuscript, we use the terms Type I, Type II, and Type III [5, 10] to designate the types. Other names that have been used over the years have been summarized by Smith and Cribbie [5].

3.1. Type I sums of squares

To briefly explain Type I, first consider the assumption that the gambling behaviour in Example 1 is completely independent on gender and athletic status and that the variability in the data within and between cells is simply a result of random error. In this scenario, the data would be best represented by an overall mean (μ) across all the data values. The goodness of fit of this model can be estimated from the square of the residual between the actual value and the model prediction (in this case, a single mean) summed over all the data. This would be calculated as

$$R^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - \mu\right)^2 \text{ or, in short hand, } R^2(\mu) \quad (1)$$

where i is the row counter, j is the column counter, and k is the item counter within the cells. Most will recognize that this is the first step in calculation of variance in a data set. It should be noted that R^2, as calculated by Equation 1, is equal to the Total Sum of Squares (SST) in the ANOVA table (Table 2). Now, let us assume that gender, but not athletic status, has an impact on the scores in Table 1. This would suggest that data scores for males would vary around a mean value and the scores for females would vary around a different mean value, and both of these gender means would vary around a common value. We are not going to refer to this common value as a mean but as a common value, which can or cannot be a mean. The means of the female and male group minus the common value (μ') is given as α_i and can be seen as impact values. They are simply the values by which μ' should be adjusted to give the predicted female and male mean scores. The goodness of fit of this model can be estimated from the square of the residual between the actual value and the model prediction (in this case, $\mu' + \alpha_i$), summed over all the data.

$$R^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - (\mu' + \alpha_i)\right)^2 \text{ or, in short hand, } R^2(\mu',\alpha) \quad (2)$$

The improvement in fit due to us considering gender, would be the difference between the results of Equation 1 and 2. This is known as the sum of squares for Factor A (gender) and would be part of the ANOVA table.

$$SSA = R^2(\mu) - R^2(\mu',\alpha)$$

Next, let us assume that both gender and athletic status have an impact on the scores and that they are independent of each other. This would suggest that, just as the data vary around row means (gender means), the data also vary around column means (athletic status means). That would lead to us to state that the values in a cell can be estimated by adjusting a common value (μ') for gender impact (α_i) and for athletic status impact (β_j). The goodness of this model to fit the data can be estimated from the square of the residual between the actual value and the model prediction (in this case, $\mu' + \alpha_i + \beta_j$), summed over all the data.

$$R^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - (\mu' + \alpha_i + \beta_j)\right)^2 \text{ or, in short hand, } R^2(\mu',\alpha,\beta) \quad (3)$$

The improvement in fit due to us considering athletic status impact, AFTER considering gender impacts, would be the difference between the results of Equation 2 and 3. This is known as the sum of squares for Factor B (athletic status), after Factor A (gender status) has been considered, and would be part of the ANOVA table.

$$SSB = R^2(\mu',\alpha) - R^2(\mu',\alpha,\beta)$$

The actual values of μ' and α_i in Equations 2 and 3 are not necessarily the same. Lastly, let us assume that both gender and athletic status have an impact on the scores and that they may be partially dependent of each other; i.e., there may be some interaction between the two factors that may cause the scores to be higher (or lower) than they were if the factors were completely independent from each other. That would lead to us to state that the values in a cell can be estimated by adjusting a common value (μ') for gender impact (α_i), for athletic status impact (β_j), and for interaction impact, $(\alpha\beta)_{ij}$. The goodness of fit of this model can be estimated from the square of the residual between the actual value and the model prediction [in this case, $\mu' + \alpha_i + \beta j + (\alpha\beta)_{ij}$], summed over all the data.

$$R^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - (\mu' + \alpha_i + \beta_j + (\alpha\beta)_{ij})\right)^2 \text{ or, in short hand, } R^2(\mu',\alpha,\beta,\alpha\beta) \quad (4)$$

The improvement in fit due to us considering interactions, AFTER considering gender and athletic status, would be the difference between the results of Equation 3 and 4. This is known as the sum of squares for interaction AxB (gender-athletic status) and would be part of the ANOVA table.

$SSAB = R^2(\mu',\alpha,\beta) - R^2(\mu',\alpha,\beta,\alpha\beta)$

It is notable that the square residuals, R^2, calculated by Equation 4 is equal to the sum of squares errors (SSE) in the ANOVA table. Together with the other calculations above, we can now construct a complete ANOVA table.

In the above description of calculation of the Type I sums of squares for Factor A and B, we considered Factor A first and Factor B second, but we could just as well have considered Factor B first and then Factor A. In that case, the resulting equations would have been

SSB (B first) = $R^2(\mu) - R^2(\mu',\beta)$ SSA (B first) = $R^2(\mu',\beta) - R^2(\mu',\alpha,\beta)$

It is important to realize that SSA (A first) ≠ SSA (B first) and SSB (A first) ≠ SSB (B first). Thus, there are two versions of Type I sums of squares; one that considers Factor A first and one that considers Factor B first. To summarize the Type I sums of squares.

Type IA: $SSA = R^2(\mu) - R^2(\mu',\alpha)$ $SSB = R^2(\mu',\alpha) - R^2(\mu',\alpha,\beta)$

Type IB: $SSA = R^2(\mu',\beta) - R^2(\mu',\alpha,\beta)$ $SSB = R^2(\mu) - R^2(\mu',\beta)$

Type IA & IB: $SSAB = R^2(\mu',\alpha,\beta) - R^2(\mu',\alpha,\beta,\alpha\beta)$ $SSE = R^2(\mu',\alpha,\beta,\alpha\beta)$ $SST = R^2(\mu)$

3.2. Types II and III sums of squares

The above procedure correctly calculates the Type I sums of squares for the ANOVA table by the method used in most statistical software. Calculation of SSA and SSB by Types II and III follow slightly different logic and has been outlined by Smith and Cribbie [5]; Shaw and Mitchell-Olds [7]; Speed, Hocking, and Hackney [12]. One may think of Type II as a method where improvement to a model, by adding a main factor (i.e., A and B), is evaluated after all the other main factors have been considered. Type III could be thought of as a method where improvement to a model, by adding a main factor (i.e., A and B), is evaluated after all the other factors (mains and interactions) have been considered. The values for SST, SSE, and $SSAB$ are common among all three types and are calculated by the same equations each time. The sums of squares for SSA and SSB differ between the types, and the calculations for Types II and III are given below.

Type II: $SSA = R^2(\mu',\beta) - R^2(\mu',\alpha,\beta)$ $SSB = R^2(\mu',\alpha) - R^2(\mu',\alpha,\beta)$

Type III: $SSA = R^2(\mu',\beta,\alpha\beta) - R^2(\mu',\alpha,\beta,\alpha\beta)$ $SSB = R^2(\mu',\alpha,\alpha\beta) - R^2(\mu',\alpha,\beta,\alpha\beta)$

In the case of Types II and III there is no such thing as considering the order of A or B. Two more types of (sums of) squares of residuals were introduced above, $R^2(\mu',\alpha,\alpha\beta)$ and $R^2(\mu',\beta,\alpha\beta)$, that were not shown in Equations 1-4. These additional entities are calculated by the following two equations:

$$R^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - (\mu' + \alpha_i + (\alpha\beta)_{ij})\right)^2 \text{ or, in short hand, } R^2(\mu',\alpha,\alpha\beta) \quad (5)$$

$$R^2 = \sum_i \sum_j \sum_k \left(y_{ijk} - (\mu' + \beta_j + (\alpha\beta)_{ij})\right)^2 \text{ or, in short hand, } R^2(\mu',\beta,\alpha\beta) \quad (6)$$

When using Equations 2-6, we also need to consider that ANOVA normally add the following restrictions [13], forcing sums of the impact values to be zero when looking at rows and columns.

$$\sum_i \alpha_i = 0 \qquad \sum_j \beta_j = 0 \qquad \sum_i (\alpha\beta)_{ij} = 0 \qquad \sum_j (\alpha\beta)_{ij} = 0 \quad (7, 8, 9, 10)$$

3.3. Calculations of μ', α_i, β_j, and $(\alpha\beta)_{ij}$

The calculation of μ', α_i, β_j, and $(\alpha\beta)_{ij}$ is done by least square regression; i.e., for each of Equations 2-6, μ', α_i, β_j, and $(\alpha\beta)_{ij}$ are found by seeking the minimum R^2 value for the equation. The use of least square regression models to calculate the different types of sums of squares is explained by Overall and Spiegel [13] but note that these authors called the different types of sums of squares "methods" rather than "types" and their numbering system differs from most others (Model 1 = Type III and Model 3=Type I) [5].

The least square regression technique to find μ', α_i, β_j, and $(\alpha\beta)_{ij}$ is aided by using dummy contrast variables. To summarize this method, consider trying to determine μ', α_i, β_j, and $(\alpha\beta)_{ij}$ in the model used in Equation 4 [$\mu' + \alpha_i + \beta_j + (\alpha\beta)_{ij}$]. For the scenario described in Example 1 (gambling scores), the predicted value (less random error) for any data value can be written in the most general sense as

Predicted value =

$$\mu' + \alpha_1 + \alpha_2 + \beta_1 + \beta_2 + \beta_3 + (\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{13} + (\alpha\beta)_{21} + (\alpha\beta)_{22} + (\alpha\beta)_{23} \quad (11)$$

With the acknowledgement that only some of the terms will be used when predicting a specific data value. When the restrictions in Equations 7-10 are imposed, Equation 11 simplifies to

Predicted value =

$$\mu' + \alpha_1\,(a_1) + \beta_1\,(b_1) + \beta_2\,(b_2) + (\alpha\beta)_{11}\,(a_1b_1) + (\alpha\beta)_{12}\,(a_1b_2) \quad (12)$$

Where a_1, b_1, b_2, a_1b_1, and a_1b_2 are dummy contrast variables that are assigned values of 1, 0, or –1, depending on which data value is predicted. For example, if we want to predict the gambling score of a male current athlete (first row, left column in Table 1), we use the following values for the dummy contrast variables:

$a_1 = 1$ $\quad b_1 = 1$ $\quad b_2 = 0$ $\quad a_1b_1 = a_1 \cdot b_1 = 1$ $\quad a_1b_2 = a_1 \cdot b_2 = 0$

Similarly, if we wanted to predict the gambling score for a female former athlete (bottom row, center column in Table 1), we use the following values for the dummy contrast variables:

$a_1 = -1$ $\quad b_1 = 0$ $\quad b_2 = 1$ $\quad a_1b_1 = a_1 \cdot b_1 = 0$ $\quad a_1b_2 = a_1 \cdot b_2 = -1$

By assigning values to the dummy contrast variables for each of the data values, we can use the multiple variable least square regression technique to determine α_1, β_1, β_2, $(\alpha\beta)_{11}$, and $(\alpha\beta)_{12}$ as the regression coefficients of Equation 12. These results can be used with Equations 7-10 to determine the other impact values. How to assign the values of all the dummy contrast variables is described later.

3.4. The unique case of no interactions between the main factors

In the above discussion, we included the possibility of interaction between the main factors which is the normal two-way ANOVA assumption. However, in some cases, it may be desirable not to consider the interactions and instead include that variance in the *SSE*. If this is the case, the above derivation of some sum of squares will not apply. Specifically, the "Interaction AxB" row in Table 2 is not applicable and the Degrees of Freedom for the Error in Table 2 is calculated as

df_{AB} (in Table 2) = $N - N_A N_B + (N_A - 1)(N_B - 1)$

and the sum of squares for the different types are

Type IA:	$SSA = R^2(\mu) - R^2(\mu',\alpha)$	$SSB = R^2(\mu',\alpha) - R^2(\mu',\alpha,\beta)$	
Type IB:	$SSA = R^2(\mu',\beta) - R^2(\mu',\alpha,\beta)$	$SSB = R^2(\mu) - R^2(\mu',\beta)$	
Type II:	$SSA = R^2(\mu',\beta) - R^2(\mu',\alpha,\beta)$	$SSB = R^2(\mu',\alpha) - R^2(\mu',\alpha,\beta)$	
Type III:	$SSA = R^2(\mu',\beta) - R^2(\mu',\alpha,\beta)$	$SSB = R^2(\mu',\alpha) - R^2(\mu',\alpha,\beta)$	
All Types:	$SSAB$ = not applicable	$SSE = R^2(\mu',\alpha,\beta)$	$SST = R^2(\mu)$

As is noted, SSA and SSB for Types II and III are exactly the same when interactions between main factors are not considered.

4. Development of a spreadsheet

The layout of the spreadsheet is shown in Figure 1 with sections shaded for the different regions. Below is the general description of each region. It is recommended that the spreadsheet which accompanies this report is downloaded from the report's web site or requested from the author and that Excel Help is used to look at the detailed description of each Excel function. It should be noted that the example spreadsheet is set up for handling up to a 5x5 classic ANOVA design with up to 100 data values. The spreadsheet can easily be expanded downward to allow for more data values (up to 999). Expanding the design beyond 5x5 becomes more challenging.

4.1. Data (A10:C111)

The data entry section has three columns. One for the data values and two for the A-and B-levels. From the examples above, we are showing the first 10 lines (Figure 2). There is one data value per row in the spreadsheet. Several other grey sections in the spreadsheet (Figure 1) are "extensions" of that data row.

4.1. Basic information about the data (D10:J13)

This section has basic information about the data which is needed for some calculations. The number of data values, the number of levels of A and B, and the location of each of these. The spreadsheet functions used are COUNT, MAX, and CONCATENATE.

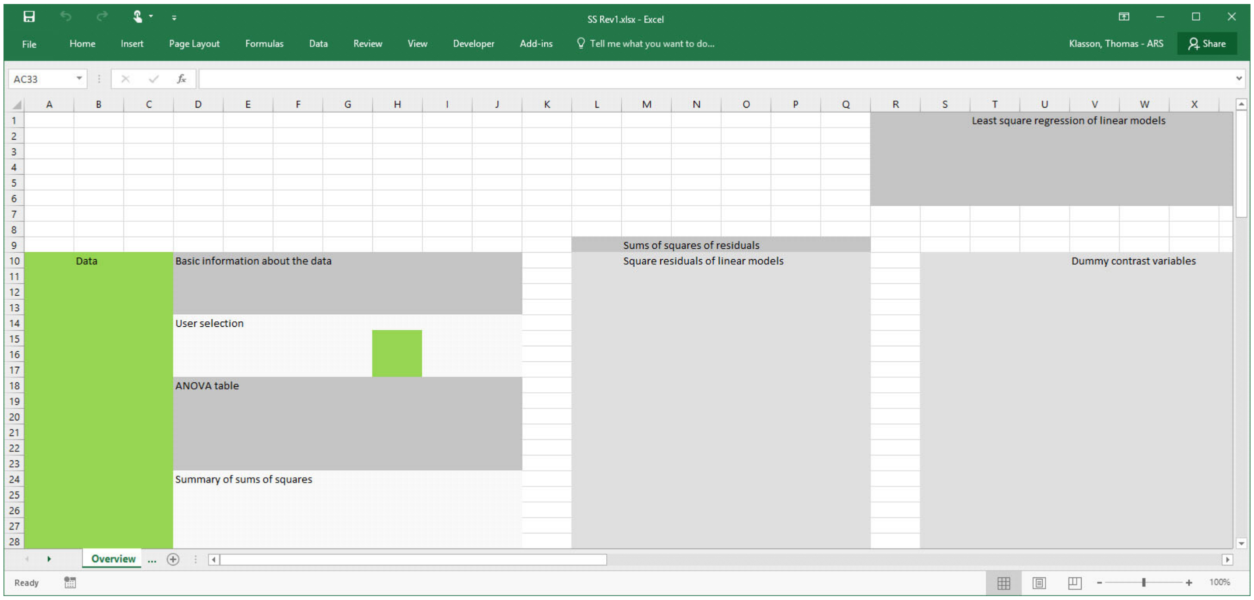

Figure 1. Layout of spreadsheet with different regions highlighted. Green sections represent user input sections. The gray sections carry out calculations and show results.

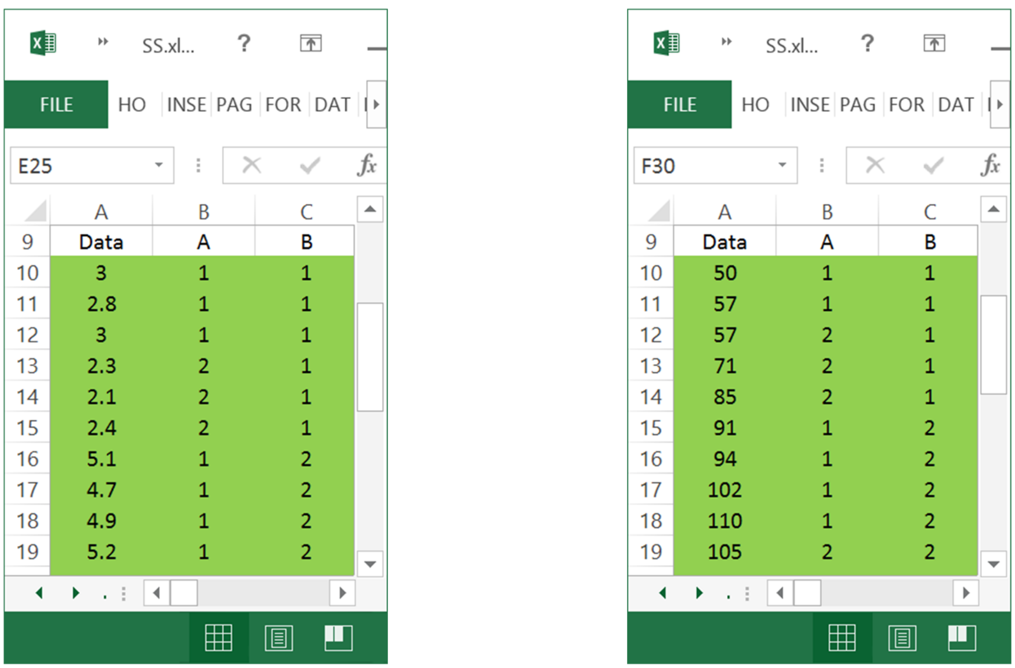

Figure 2. Data entry section in the spreadsheet stretches from A10 to C111. The first 10 lines of the data and A- and B-levels of Examples 1 and 2 (Tables 1 and 3).

4.2. Dummy contrast variables (S10:BJ111)

The number of dummy contrast variables is related to the number of levels of A and B. The number of dummy contrast variables for A is equal to N_A–1 and the number of dummy contrast variables for B is equal to N_B–1. Thus, in the example of gambling scores we will have one dummy contrast variable (a_1) associated with Level 1 of A and two dummy contrast variables (b_1 and b_2)

associated with Levels 1 and 2 of B. Every data value will each have a set of dummy contrast variable values depending on where its cell is located within the rows and columns. The values for the dummy contrast variables will be the same for all data values within a cell. The general procedure of how to assign dummy contrast variable values for a_i follows these simple principles:

a_1 = 1 for data where A=1

a_1 = –1 for data where A=the last (highest) level of A

a_1 = 0 for all other data, regardless of A

a_2 = 1 for data where A=2

a_2 = –1 for data where A=the last (highest) level of A

a_2 = 0 for all other data, regardless of A

and so on until all the a_i contrast variable have values.

The procedure of how to assign dummy contrast variable values for b_j follows the same principle. In the Excel spreadsheet, nested IF functions were used to assign the dummy contrast variable values based on above principles. There is also a set of dummy contrast variables that are associated with the interactions of A and B. There are (N_A–1)*(N_B–1) interaction dummy contrast variables which we will call *ab* and they are simply multiplications of *a*- and *b*-values for each data value. In Tables 4 and 5, values of the dummy contrast variables are shown for the two examples. Note that the three columns to the left in Tables 4 and 5 are the data entry columns, coloured green in Figure 2. In the spreadsheet, that can handle more levels of A and B, the dummy contrast variables that are not relevant are assigned values of zero.

Table 4. Dummy contrast variable values for cells in Example 1 (Table 1, Gambling Scores) as determined by the A- and B- levels of the cell.

	A	B	a_1	b_1	b_2	a_1b_1	a_1b_2
Data value	1	1	1	1	0	1	0
Data value	1	2	1	0	1	0	1
Data value	1	3	1	–1	–1	–1	–1
Data value	2	1	–1	1	0	–1	0
Data value	2	2	–1	0	1	0	–1
Data value	2	3	–1	–1	–1	1	1

Table 5. Dummy contrast variable values for cells in Example 2 (Table 3, Plant Height) as determined by the A- and B- levels of the cell.

	A	B	a_1	b_1	a_1b_1
Data value	1	1	1	1	1
Data value	1	2	1	–1	–1
Data value	2	1	–1	1	–1
Data value	2	2	–1	–1	1

4.3. Least squares regression of linear models (R1:AQ6)

There are six linear models (data value predictors) that need to be solved separately. Each model is a linear equation, based on a portion of Equation 12, and the model consists of dummy contrast variables (a_i, b_j, a_ib_j) and regression coefficients [α_i, β_j, $(\alpha\beta)_{ij}$]. Excel has a convenient function, LINEST, that takes two main range arguments, which determines coefficients of linear equations. The first argument is the range of dependent values (in Excel, these are referred to as y-values), the second argument is the range of independent values (in Excel, these are referred to as x-values). In our case, the range of dependent values is the range of data values and our range of independent values is the range of all dummy contrast variable values for the model in question. One of the restrictions of the LINEST function is that the y- and x-values must be in ranges that are uninterrupted and of specific sizes. In the spreadsheet created, the functions CONCATENATE and INDIRECT are used to select the ranges (shown in R1:R6) to be used for the various models. The result from LINEST is a small array of values, but only the value in the upper left corner of the array is normally displayed in the spreadsheet. In order to access other parts of the array, we use the Excel INDEX function with the COLUMN function as a counter. If you are not familiar with the LINEST and INDEX functions, please consult the Excel Help menu. It also explains the other arguments needed for LINEST.

4.4. Square residuals of linear models (L10:Q111)

For each of the six linear models, the square residual is calculated for each of the data values as the difference between the data value and its predicted value (according to the model used). That difference is then squared. The Excel function SUMPRODUCT is partially used for the predicted value calculation with range arguments of the dummy contrast variable values and the determined regression coefficients. An IF function is also used to make sure residuals are only calculated when needed.

4.5. Sums of squares of residuals (L9:Q9)

The sums of the squares of the residuals are simply that. Each of the square residuals are summed. The function used in Excel is SUM.

4.6. Summary of sums of squares (D24:J29)

Each of the sums of squares for the main effect and the interaction that make up the ANOVA table is the difference between two of the sums of squares of residuals (Equations 2-6.) The SST Sum of Squares (Equation 1) can be directly calculated with the DEVSQ Excel function. All the types (IA, IB, II, and III) of the sums of squares are summarized in this section of the spreadsheet.

4.7. User selection (D14:J17)

This section is intended to give the user flexibility over the ANOVA table. The significance level can be set and the sums of squares type [1 (IA), 2 (II), or 3 (III)] and can be selected for the ANOVA table and to evaluate importance. If the number 7 is entered for the Type, Type IB sums of squares will be used. The interaction between the main factors can also be changed.

4.8. ANOVA table (D18:J23)

The classic ANOVA table is simply a compilation of values taken from the "Basic information about data," the "User selection," and the "Summary of sums of squares" sections. The critical *F*-value and the p-value use the Excel functions FINV and FDIST. The choice of sums of squares is selected by the CHOOSE function.

5. Results of examples

The results of entering the data from Examples 1 and 2 into the Excel spreadsheet are displayed in Figure 3 and 4. Only information in the green sections were entered and the remaining parts of the spreadsheet were automatically calculated. An interesting aspect to point out is noted in the "Summary of sums of squares" section where it can be seen that the Type II *SSA* and *SSB* values are the same as some of the Type IA and IB values. Another thing to note is that *SSA* + *SSB* + *SSAB* + *SSE* only adds up to *SST* in the case of Type IA and IB types of squares, which is typical for balanced data and Type I unbalanced data [10]. The spreadsheet can be used for both balanced and unbalanced data.

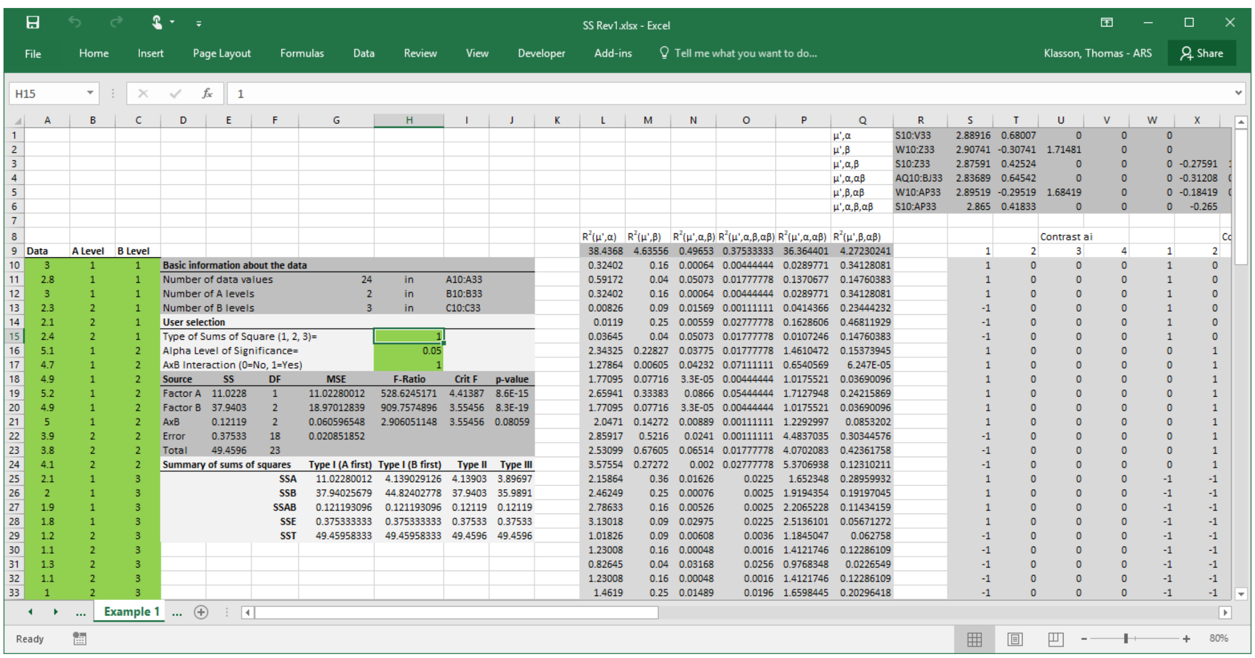

Figure 3. Results of ANOVA analysis for data in Example 1, using Type IA sums of squares.

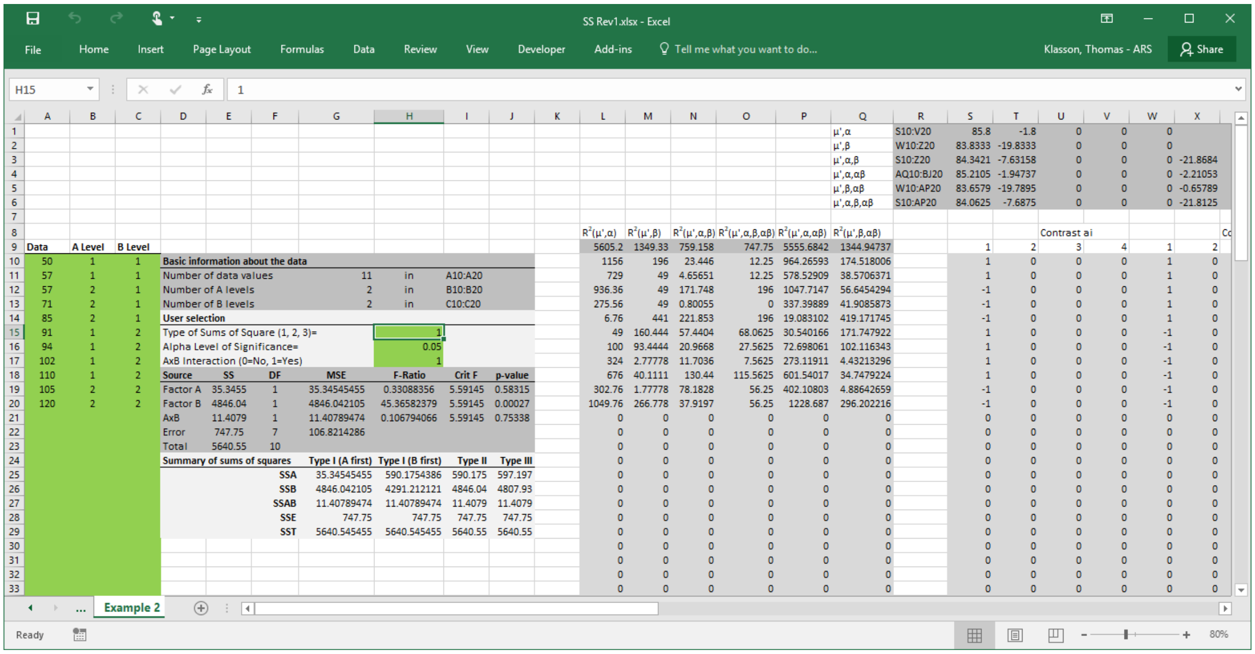

Figure 4. Results of ANOVA analysis for data in Example 2, using Type IA sums of squares.

6. Examples of the implementation of the spreadsheet as an education and research tool

ANOVA evaluation implies the use of *F* statistics and some significance level or acceptable error rate (i.e., alpha, cell H16 in the spreadsheet), often 0.05 (i.e., 5%). To explore the implication of selecting a particular type of sums of squares and using 0.05 as the acceptable error rate, consider the Example 1 data (Table 1) using the spreadsheet to calculate the p-value for the Factor A, Factor B, and the interaction between Factors A and B. The resulting p-values are listed in Table 6, left, and they strongly indicate that gambling behaviour is dependent on both gender (Factor A) and athlete status (Factor B) with p-values of <<0.05. We also find that the gambling is not dependent on the interaction between Factor A and B with a p-value of 0.081 (which is >0.05). In this example, it made no difference which type of sums of squares were chosen to calculate the p-values and evaluate the overall impact of the different factors.

Table 6. *F* test results (p-values), from the developed spreadsheet (range D18:J23), using different types of sums of squares to evaluate the impact of different factors for Examples 1 and 2.

	Example 1, p-values					Example 2, p-values			
Factor	IA	IB	II	III		IA	IB	II	III
A	9E-15	4E-11	4E-11	6E-11		0.58	0.051	0.051	0.050
B	8E-19	2E-19	8E-19	1E-18		0.00027	0.00039	0.00027	0.00028
A×B	0.081	0.081	0.081	0.081		0.75	0.75	0.75	0.75

Now consider the data in Example 2 (Table 2). Using the spreadsheet we can calculated the *F* test statistics and the p-values for this data set. The results (Table 6, right) show that, regardless of the type of sums of squares used, the final plant height is clearly dependent how tall the plants were during early growth (Factor B) as the p-values were <<0.05. We also see that, regardless of the type of sums of squares used, there was no interaction between the two Factors A and B as the p-values were >>0.05. However, when we try to determine the impact of weed removal (Factor A) on final plant height, we find the results conflicting. On one hand, the p-value using Type IA sums of squares indicate that there is no impact of weed removal on final plant height (p-value >>0.05 for Factor A). On the other hand, using Types IB, II, and III sums of squares, weed removal appear to have an impact on the final plant height because the p-values are either equal to, or slightly above, the acceptable error rate value of 0.05. The significance of this would be that an experimenter using the Type IA sums of squares would conclude with absolute certainty that weed removal had no impact on final plant height. But another experimenter using the same data and the Type III sums of squares would probably be very hesitant to state something definite about the impact of weed removal, and would likely recommend additional experiments.

So what sums of squares should be used? Unfortunately, there is not a "correct" answer to this question even though much has been written about the topic [5, 7, 9, 10, 14]. Most statistical data packages defaults to Type I and III sums of squares [5, 9], but Type I has fallen out of favour [7] as it can be dependent on the order of which the main factors are considered and also on the degree of data unbalance. This would suggest that Type III should be recommended, which is a recommendation shared by several others according to Shaw and Mitchell-Olds [7]. However, recent writings within the last two decades [5, 9, 14] have recommended that Type II should be the default method if no (or minor) interaction is noted because it is often more powerful than Type III, but in most cases these recommendations come with the caveat that Type II is dependent on the degree of unbalance and care must be taken when using Type II to investigate main effects. All recommendations warn against using two-way ANOVA treatment when interactions are very significant. For the sake of simplicity, it is safe to recommend Type III sums of squares as the basis of analysis, once it has been determined that interaction is not highly significant. However, it is always recommended that the details of any statistical method used is stated when results are presented. Using this recommendation, we can formulated the conclusions of ANOVA evaluation of the data in Examples 1 and 2.

In the case of Example 1: Using two-way ANOVA Type III sums of squares and a significant level of 0.05, it is concluded that both gender and athletic status had an impact on gambling behaviour. There was no statistical evidence that this conclusion was impacted by interactions between gender and athletic status.

In the case of Example 2: Using two-way ANOVA Type III sums of squares and a significant level of 0.05, it is concluded that the final plant height was clearly dependent how tall the plants were during early growth. There was no clear evidence that removal of weeds during the early growth impacted the final plant height; however, more experiments are recommended to confirm this initial finding. There was no statistical evidence that these conclusions were impacted by interactions between the weed removal scheme and early growth plant height.

The spreadsheet technique for calculating sums of squares for Types I, II, and III ANOVA treatment of unbalanced data was presented to a panel of chemists and engineers at the Annual Meeting of American Chemical Society, March 18-22, 2018, in New Orleans, Louisiana, USA.

7. Supplementary material

The Excel spreadsheet shown in Figures 3 and 4 can be requested from the author.

8. References

[1] Gasiorowski, J. H. (1998), The relationship between student characteristics and math achievement when using computer spreadsheets. A PhD Dissertation, West Virginia University, Morgantown, West Virginia.

[2] Abramovich, S. and Nabors, W. (1997), Spreadsheets as generators of new meanings in middle school algebra. *Comp Schools*, 13(1–2): 13–25.

[3] Molyneux-Hodgson, S., Rojano, T., Sutherland, R. and Ursini, S. (1999), Mathematical modelling: the interaction of culture and practice. *Educational Studies in Mathematics*, 39(1–3): 167–83.

[4] Baker, J. and Sugden, S. J. (2007), Spreadsheets in education–The first 25 years. *Spreadsheets in Education*, 1(1): 18–43.

[5] Smith, C. E. and Cribbie, R. (2014), Factorial ANOVA with unbalanced data: A fresh look at the types of sums of squares. *Journal of Data Science*, 12(3): 385–404.

[6] Herr, D. G. (1986), On the history of ANOVA in unbalanced, factorial designs: The first 30 years. *American Statistician*, 40(4): 265–270.

[7] Shaw, R. G. and Mitchell-Olds, T. (1993), ANOVA for unbalanced data: An overview. *Ecology*, 74(6): 1638–1645.

[8] Yates, F. (1934), The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29(185): 51–66.

[9] Langsrud, Ø. (2003), ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, 13(2): 163–167.

[10] Hector, A., von Felten, S. and Schmid, B. (2010), Analysis of variance with unbalanced data: An update for ecology & evolution. *Journal of Animal Ecology*, 79(2): 308–316.

[11] Nelder, J. A. and Lane, P. W. (1995), The computer analysis of factorial experiments: In memoriam-Frank Yates. *American Statistician*, 49(4): 382–385.

[12] Speed, F. M., Hocking, R. R. and Hackney, O. P. (1978), Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 73(361): 105–112.

[13] Overall, J. E. and Spiegel, D. K. (1969), Concerning least squares analysis of experimental data. *Psychological Bulletin*, 72(5): 311–322.

[14] Lewsey, J.D., Gardiner, W.P. and Gettinby, G. (2001), A study of Type II and Type III power for testing hypotheses from unbalanced factorial designs. *Communications in Statistics – Simulation and Computation*, 30(3): 597-609.