C.A.L.F
STEWARDSHIP

CATTLE. ANTIMICROBIAL. LAND. FINANCIAL.

Microsoft Excel for the Veterinary Practitioner: Basic Techniques of Animal Health Data Analysis

W. Isaac Jumper, DVM

MISSISSIPPI STATE UNIVERSITY™
COLLEGE OF VETERINARY MEDICINE

Microsoft Excel for the Veterinary Practitioner: Basic Techniques of Animal Health Data Analysis

W. Isaac Jumper, DVM

Introduction:

Software tools that aid in the collection, organization, and analysis of animal health data are increasingly accessible to the veterinary practitioner. Spreadsheet software, such as Microsoft Excel can store large amounts of data, and can perform a variety of tasks to evaluate that data. In order to make evidence-based treatment, management, and production decisions, an accurate method of data collection and organization must be established. Spreadsheet software allows the user to collect, transform, describe, and display data in ways that are useful for analysis and communication of results. The objective of this article is to describe basic techniques for collection, organization, and analysis of common types of data used in livestock production systems.

Getting started in Excel:

Before using spreadsheet software, the user must be familiar with some fundamental spreadsheet aspects, and the terminology used to describe them. In this article, Microsoft Excel will be used for examples, although other forms of spreadsheet software are available (e.g. Google sheets). When opening a spreadsheet in Microsoft Excel, the user can choose between a blank workbook or one of several different pre-made workbook templates designed for a specific purpose (e.g. calendar, budget, schedule, etc.). Figure 1 displays the menu screen presented to the user when opening a new workbook.

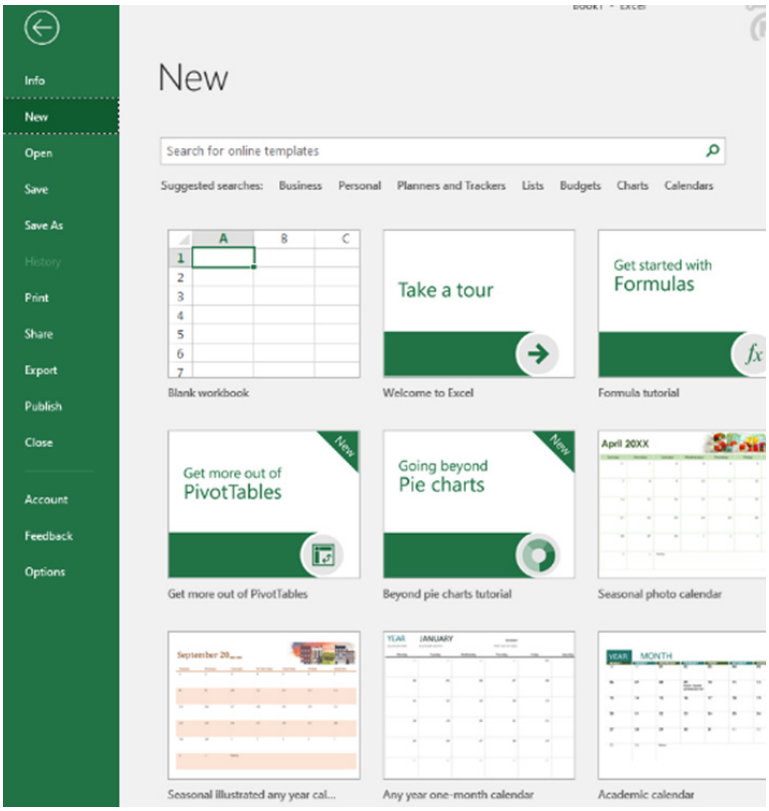

Figure 1: Microsoft Excel menu screen displaying options for opening a new workbook. Note that blank workbooks and workbooks with pre-made templates are available.

Each workbook is made up of at least one spreadsheet, but other spreadsheets may be added in each workbook. Once the user selects "Blank workbook" from the menu in Figure 1, the blank workbook is opened to "sheet1" and they are ready to begin working. At the bottom of each spreadsheet, tabs are present that allow the user to add additional spreadsheets to the workbook, or switch between existing spreadsheets. Figure 2 illustrates a workbook with more than one sheet. Additional spreadsheets are added to the workbook by clicking the circled plus icon adjacent to the last sheet. Each sheet can be named separately by right-clicking on the name of the tab (e.g. "Sheet4"in Figure 2) and selecting the "Rename" option, or by double-clicking on the tab itself. Different sheets within the same workbook may be used to store different sets of data, provide reference lists for functions and formulas, or contain tools such as pivot tables.

Spreadsheet organization:

Spreadsheets are an assortment of cells, arranged in rows and columns. Each cell may contain data of various types, or formulas that perform calculations based on data in other cells of the same spreadsheet or other spreadsheets. Cells are named by the column and row to which they belong. Columns are arranged in alphabetical order beginning with "A", while rows are arranged in numerical order beginning with "1". Spreadsheets are not infinite in the number of rows and columns that they contain, but they are very large. Each cell is identified by its column and row (e.g. A1, C5, D9, etc.). When an individual cell is selected, it will be surrounded by a green box, and the column and row to which it belongs will be highlighted in green as well. As you scroll through the spreadsheet, the column letter and row number will always be visible.

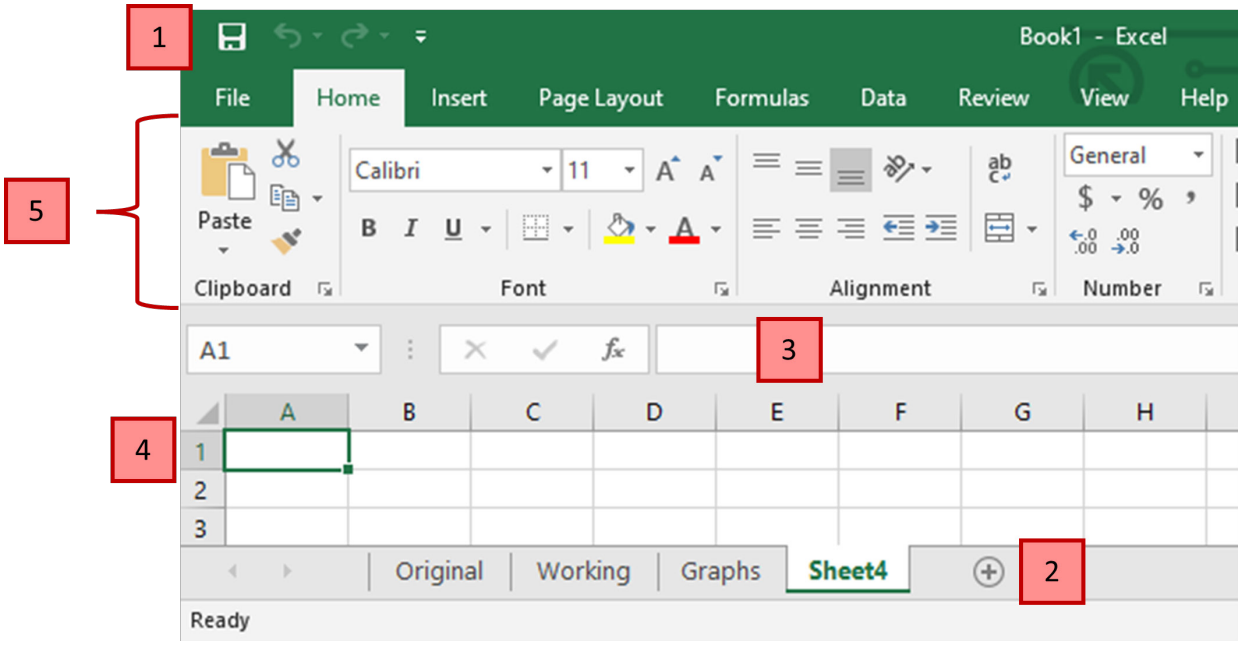

Figure 2: A blank excel worksheet. 1.) Save button; saves the current workbook using default .xlsx file. Directly adjacent to the save button are the "undo" (arrow pointed to left) and "redo" (arrow pointed to right) buttons. 2.) The workbook contains 3 sheets with the example names "Original", "Working", and "Graphs". "Sheet4" is the selected sheet and may be renamed by right-clicking on "Sheet4" and selecting the "Rename" option, or by double-clicking on "Sheet4". Sheets may be added to the workbook by clicking on the circled plus icon, and may be deleted from a workbook by right-clicking on the sheet name tab and selecting the "Delete" option. 3.) Formula bar which displays the content of a selected cell. 4.) Cell "A1" selected with corresponding column and row highlighted in green. 5.) Ribbon made up of individual tabs containing useful tools.

Before working with data in a spreadsheet, creating a copy of the data can help protect against mistakes or errors that cannot be corrected. In the event of such an error, having a copy of the data can save the user from re-entering, or re-collecting, the data. As an example, Figure 2 displays "Original" and "Working" spreadsheets. These spreadsheets represent the original, unaltered data, and data undergoing analysis, respectively.

By default, Excel workbooks save as a .xlsx file type. In some instances, it may be useful to save an individual worksheet as a comma delimited file (i.e. .csv file), such as if importing the data into a statistical software program. Note that each .csv file can save only one worksheet in a workbook at a time, whereas the default .xlsx file type will save the entire workbook.

Before working with data in a spreadsheet, creating a copy of the data can help protect against mistakes or errors that cannot be corrected. In the event of such an error, having a copy of the data can save the user from re-entering, or re-collecting, the data. As an example, Figure 2 displays "Original" and "Working" spreadsheets.

These spreadsheets represent the original, unaltered data, and data undergoing analysis, respectively.

By default, Excel workbooks save as a .xlsx file type. In some instances, it may be useful to save an individual worksheet as a comma delimited file (i.e. .csv file), such as if importing the data into a statistical software program. Note that each .csv file can save only one worksheet in a workbook at a time, whereas the default .xlsx file type will save the entire workbook.

Types of data:

Careful planning prior to data collection can save time and headache during analysis, as well as reduce the risk of inaccurate conclusions. Data is objective or factual information that is used to draw conclusions or make calculations. There are many types of data, and the method of analysis differs depending on the type of data collected.

All data may be described as either qualitative or quantitative. Qualitative data is categorical, and describes properties of an animal (e.g. breed, sex, etc.). Nominal and ordinal data are two types of qualitative data. Nominal data is categories without any apparent order. For example, recording individual calf sex as bull, heifer, or steer would be nominal data. This data fits into three convenient categories that may be labelled as “b”, “h”, and “s” in a spreadsheet. If numbers are used to represent categories of sex (i.e. 1=bull, 2=heifer, 3=steer) or any other nominal categories when collecting data in an Excel worksheet, it must be remembered that these data cannot be manipulated in an algebraic manner. For instance, you would not sum two rows of data that include a “1” and a “2” to get a total of “3”, because a bull plus a heifer does not equal a steer. The order of these nominal data categories also is not important, as 1 being assigned to “bulls” is arbitrary. Ordinal data is also categorical, but there is an inherent order to the data. One of the most common types of ordinal data collected by veterinary practitioners is body condition score. Each body condition score is essentially a category, and there is an order of increasing BCS as the number that represents each category increases. The interval between categories is not always known, not the same in all cases, and not equal between all categories (e.g. a cow with a BCS of 3 is not necessarily in half the condition of a cow with a BCS of 6). For ordinal data, it does not matter what numbers or letters are used to represent the categories, the order of the data is what is of most importance.

Quantitative data captures specific amounts, rather than just classes (e.g. weight, temperature, etc.). Quantitative data can be further divided into discrete and continuous data. Discrete data is data of which a fraction or proportion is not possible. Common examples of discrete data are numbers of calves born, or number of teats on a sow. A cow cannot have given birth to 2.5 calves, nor can a sow have 8.75 teats. Discrete data is most commonly associated with counts of data, or whole numbers. Continuous data is measurement data, and can have any value from negative infinity to positive infinity. Weight is an example of continuous ratio data, where there is a true zero or a true absence of the variable measured. Temperature, as it pertains to animal health data (i.e. the Fahrenheit and Celsius scales) is continuous interval data where zero is an arbitrary point on the scale, and does not represent a true absence of the variable (i.e. negative weights do not exist, but negative temperatures do).

When entering data into a spreadsheet, accuracy of data entry is imperative. Methods used to identify potential errors in the data will be discussed later in this article; however, there is no replacement for accurate, consistent data entry. Inconsistencies in data entry create problems with analysis. For instance, if cows that are pregnant are sometimes labeled with a “P” and sometimes labeled “B” (i.e. “Pregnant” vs. “Bred”), the user will struggle to acquire accurate summaries of pregnancy data. Data is often coded to make entry and analysis easier. One example of data coding may be treatment, where numbers are used to represent a different antimicrobial (i.e. 1 = LA-200, 2 = Excede, etc.). It should go without saying that if a coding system is used, it is very important to use the coding system consistently, and to maintain an accurate key for what each code represents.

Setting up the spreadsheet:

In most situations, each row in a spreadsheet is used to represent and individual animal, or an individual observation. Column headings are often contained in the first row, with each column describing some piece of data being collected on each individual. Typically, the first column is reserved for individual identification. Unique identification of individuals aids efforts by producers and veterinarians to measure individual

	A	B	C	D	E	F	G
1	CowID	Age	BCS	PregStat	Gest(Mon)	Dry	Date_checked
2	1026		6	P	5	0	8/19/2020
3	1027		5	P	5.5	0	8/19/2020
4	Red tag	8	5	O		0	7/30/2020
5	A112	8	6	P	5.5	0	7/30/2020
6	A117	8	6	P	5.5	0	7/30/2020
7	A128	8	5	P	6	0	7/30/2020
8	A145	8	5	P	5	0	7/30/2020
9	A162	8	6	P	5	0	7/30/2020
10	A169	8	4	P	5	0	7/30/2020
11	A170	8	6	P	5.5	0	7/30/2020

Figure 3: An example of pregnancy test data from a cow-calf operation. The first row contains column headings. Filters have been enabled for each column to facilitate sorting the data.

performance and health. Figure 3 provides an example of how a typical spreadsheet used to organize pregnancy examination data on a cow-calf operation might be set up. The first row in Figure 3 contains column headings that provide organization for the data in each subsequent row. An animal may appear more than once in the dataset (e.g. each row may represent a treatment, and some animals may be treated more than once). As can be seen in column A in Figure 3, the first two values for “CowID” in rows 2 and 3 are right-justified in the cell. When values within a cell are right justified, Microsoft Excel is reading the data within the cell as an integer, and can perform mathematical functions with the data. In other instances, such as rows 4-11 in Figure 3, the data within the cells is left-justified. Data that is left-justified within cells is being read as text by Microsoft Excel, and cannot be included in any mathematical functions. Data in column C, which contains “BCS” data, is being read as integer data, because it is right-justified, while Column D, which contains “PregStat” data is being read as text because the data is left-justified.

The drop-down arrows within each cell in the top row indicate that the “Sort & Filter” function is being used within the spreadsheet. This can be accomplished by selecting the “Home” tab at the top of the screen, and selecting the “Filter” option from the “Sort & Filter” button. Once filters have been enabled, clicking on an individual drop-down arrow provides the user with options to sort the data based on values in each column. The drop-down filter menu for each column will produce a list of all unique values found in the column. For example, Figure 4 has been sorted by “BCS” to show only those cows with a BCS of 5. Notice that the row numbers are blue, and not all numbers are visible. When data is filtered, only the data that matches the filter will be shown, and all other rows of data will be hidden from view. The double line between row number 4 and row number 7 indicates there is data hidden by the filter. Rows 5 and 6, which can be seen in Figure

3 to contain cows with a BCS of 6, are hidden in Figure 4. Filters can also be used to sort data in ascending or descending order (if the data is numerical), or alphabetically (if the data is text). When data is filtered, the number of records selected out of the total number filtered can be seen displayed at the bottom of the spreadsheet below the additional sheet tabs (Figure 4). In Figure 4, filtering "BCS" for all values equal to 5 results in 181 of 420 total cattle displayed. The "Sort & Filter" function can be very useful for identifying errors in data, and providing summaries of all rows that fall into specific categories of interest.

In some cases, spreadsheets may be very large (i.e. contain data that extends for numerous rows and columns). When dealing with large spreadsheets, it is often useful to "freeze" the top row (i.e. the row that contains column headings) so that no matter how far you scroll down the sheet, the column headings are always visible. This helps keep the user from needing to scroll to the top of the spreadsheet each time they need to know what values in a column represent. To "freeze" the top row, click on the "View" tab at the top of the screen. Select any cell in the first row, then click on the "Freeze panes" button under the "View" tab. Selecting the "Freeze Top Row" option from the drop-down menu will ensure the user can view the column headings while scrolling down the spreadsheet.

	A	B	C	D	E	F	G	H
1	CowID	Age	BCS	PregStat	Gest(Mon)	Dry	Date_checked	Days_bred
3	1027		5	P	5.5	0	8/19/2020	165
4	Red tag	8	5	O		0	7/30/2020	0
7	A128	8	5	P	6	0	7/30/2020	=E7*30
8	A145	8	5	P	5	0	7/30/2020	150
14	A260	8	5	P	5	0	7/30/2020	150
15	A284	8	5	P	5.5	0	7/30/2020	165
18	A324	8	5	P	5.5	0	7/30/2020	165
19	A351	8	5	P	5.5	0	7/30/2020	165
20	A353	8	5	P	5.5	0	7/22/2020	165
21	A363	8	5	P	6	0	7/30/2020	180

Original Working

Edit 181 of 420 records found

Figure 4: Example of data being filtered by "BCS"; only cattle in a BCS of 5 are displayed. "Days_bred" is calculated using a formula where estimated months gestation is multiplied by 30 to produce an estimation of days gestation. When data is filtered, the number of records filtered will be displayed at the bottom of the spreadsheet. The example above has 181 of 420 records displayed.

Table 1. Microsoft Excel formulas commonly used in analysis of livestock data

Formula name	Description	Application	Example
=AVERAGE	Returns the average (arithmetic mean) of a range of values selected by the user.	Use when calculating the mean of continuous data; used as a measure of central tendency	=AVERAGE(A2:A741)
=MEDIAN	Returns the median, or the number in the middle of a set of numbers arranged in numerical order	Useful in describing the central tendency of data, especially data that is does not follow the normal distribution	=MEDIAN (B2:B45)
=MODE.SNGL	Returns the most frequently occurring, or repetitive value within an array of data	Useful when determining when value occurs the most often within a dataset	=MODE.SNGL(D2:D97)
=MIN	Returns the smallest number in an array of data	Useful when describing the range of a dataset; may also be useful when evaluating a dataset data entry errors	=MIN(B2:B67)
=MAX	Returns the largest number in an array of data	Useful when describing the range of a dataset; may also be useful when evaluating a dataset data entry errors	=MAX(B2:B67)
=PERCENTILE.INC	Returns the value for which a specified percentile (e.g. 0.25, 0.50, etc.) of observations falls below (i.e. if 0.25 is selected, then 25% of the values in the dataset fall below the value returned by the formula.	Useful when describing the distribution of the data within a dataset	=PERCENTILE.INC(D2:D45,0.25) In this example, D2:D45 is the data array, and 0.25 is the k-value, or the percentile of interest.
=IF	If – then statement; logical test used to assign a value to the cell if the condition is met	The formula reads data in a defined cell and if the data meets a specific criterion defined by the user, then it returns the specified value.	=IF(D1="Bull",1,0) Read literally as if cell D1 equals "Bull", then return 1, otherwise return 0; note that text values must be in quotations while integers do not
=COUNT	Counts the number of cells in an array that contain numbers	Useful in finding the number of observations within a specific array; can indirectly help the user identify when numerical data is missing	=COUNT(D1:D100)
=COUNTA	Counts the number of cells in an array that are not empty	Useful when trying to determine if data are missing	=COUNTA(D1:D100)
=COUNTBLANK	Counts the number of cells in an array that do not contain data	Useful when trying to determine if data are missing	=COUNTBLANK(D1:D100)
=COUNTIF	Counts the number of cells in an array that meet a specific criterion specified by the user	Useful when counting observations of nominal or ordinal data within a given array	=COUNTIF(D1:D100,"B") Read as count the number of observations in the data array of D1 to D100 that exactly match the value "B"; note that text values must be in quotations while integers do not
=STDEV.S	Estimates standard deviation based on a sample	Useful for estimating the standard deviation of a dataset collected from a sample of a larger population	=STDEV.S(D2:D200) Assumes the data do not represent the entire population
=FREQUENCY	Calculated the number of values from a specified array that fit within user-defined bins (i.e. categories).	Useful when determining how many individual data observations fit into logical categories (i.e. how many calves in the dataset are 401-500 lbs., 501-600 lbs., 601-700 lbs. etc.)	=FREQUENCY(D2:D200,J1:J5) Where D2:D200 contains the array of data, and J1:J5 are the bins (categories). We

Using formulas:

Formulas are powerful tools for data analysis in spreadsheets. Formulas most often begin with an equal sign (=), but may also begin with plus (+) or minus (-) signs. There are a variety of formulas included in Microsoft Excel that are useful for summarizing, describing, and analyzing data. Formulas included in Excel can be found by clicking on the "Formulas" tab at the top of the spreadsheet, and exploring the "Function Library" group.

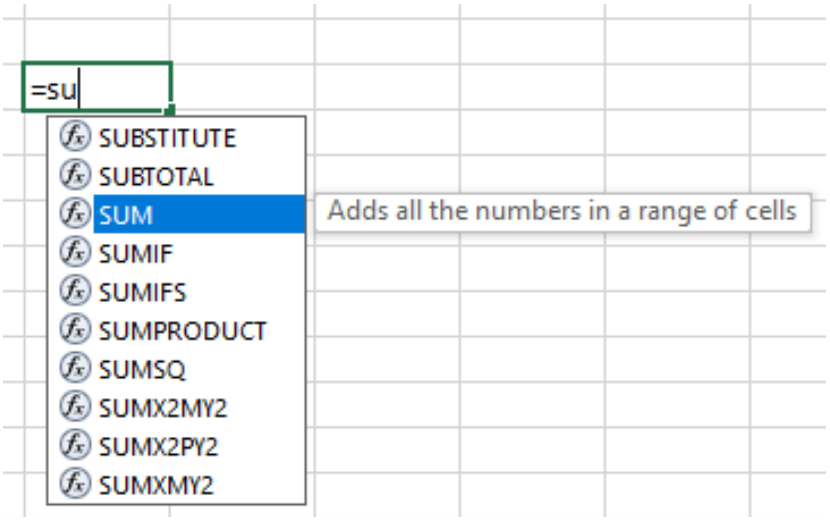

Figure 5: To enter a formula, select an empty cell, begin the formula with =, +, or – and start typing the formula. The "SUM" formula is selected above, and a brief description of the formula can be seen.

When working with data derived from livestock operations, practitioners often have a need to summarize or describe numerical data. Microsoft Excel contains many statistical formulas that are useful when summarizing data using descriptive statistics. Descriptive statistics that may be useful to the practitioner are mean, median, mode, range (minimum and maximum), and percentiles. Table 1 summarizes these and other formulas useful to the veterinary practitioner.

To begin a formula, an empty cell should be selected away from the rows and columns of data. Formulas can be based on data in separate sheets of the same workbook, but often formulas are used in the same spreadsheet that contains the data being analyzed. Once an empty cell is selected, begin typing the formula. Microsoft Excel will attempt to predict which formula the user needs by bringing up a list of formulas based on the letters that are typed following either the equal, plus, or minus sign. Figure 5 displays an example of the drop-down options provided by Microsoft Excel when the user begins to search for the "sum" function. The user does not need to type the entirety of the formula desired, but only a sufficient portion to make the desired formula appear in the drop-down box below the formula. A brief description of the function of each formula can be seen in the box that appears when a formula is selected. Users may select a formula from the drop-down list by clicking or using the arrow keys on the keyboard to scroll up and down. Once the desired formula is highlighted in blue (e.g. the "SUM" formula in Figure 5). The user may

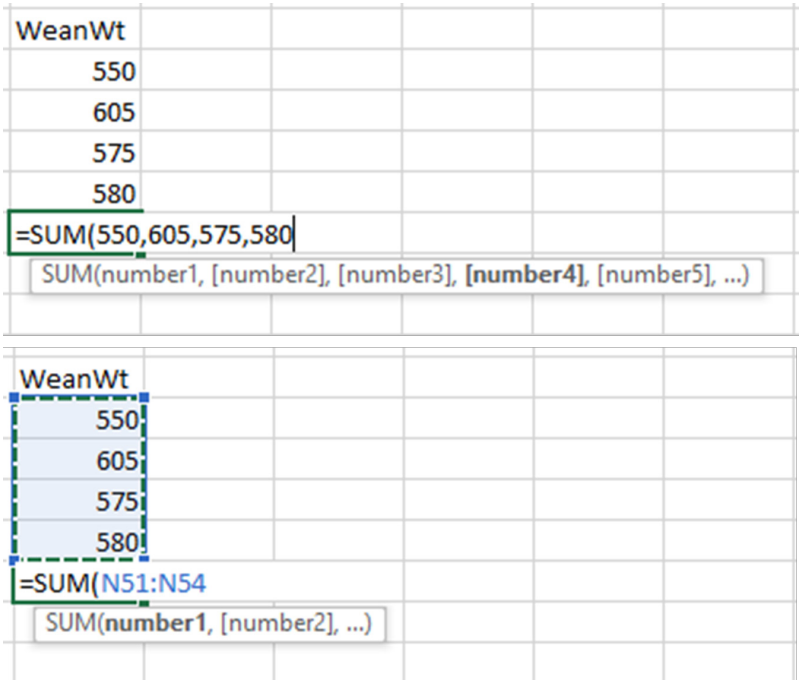

Figure 6: Formulas can be completed by a.) manually entering data, or b.) by selecting a range of cells.

either double-click on the selected formula or use the "Tab" key on the keyboard to populate the selected cell with the desired formula.

Once a formula has been selected and begun, the user must supply the formula with the needed information. For example, the "SUM" formula shown in Figure 5 would need the values that the user desires to be summed. There are two approaches to entering information such as this into formulas in Microsoft Excel. Figure 6a illustrates the first option, or entering individual values into the formula. This method is inefficient, especially in large datasets. However, when datasets contain many values, this method can be cumbersome, both when entering data and if the data must be edited (e.g. in Figure 6, if the 575 was actually determined to be 757, the user must open the formula and manually change the entry, as well as changing the value in the dataset).

Figure 6b illustrates the second method of entering data into a formula. Rather than entering each number individually, the user can select the range of cells that contains data to be included in the formula. In Figure 6, the range of cells is N51:N54. When formulas are completed using this method, they automatically update if the user changes any values in the range of cells selected for the formula. This reduces the risk that a value would be changed in the data, and not in the formula.

	mean	678		mean	503
	median	500.8354		median	502.0587
	mode	550		mode	550
	min	0		min	331
	max	57000		max	683
0.25	25th percent	457.31	0.25	25th percent	462.77
0.5	50th percent	500.84	0.5	50th percent	502.06
0.75	75th percent	539.64	0.75	75th percent	547.60

Figure 7: a.) Descriptive statistic results for weaning weight data containing errors, and b.) the resulting descriptive statistics when these errors have been corrected.

Evaluating datasets for errors:

Before extensive analysis is performed on a dataset, it is always beneficial to evaluate datasets for errors. Errors may be data that has been entered incorrectly, or not entered at all. Errors in data entry may be obvious (e.g. in a range of rectal temperatures taken on a set of feeder calves, a temperature of 1015.0 may be an instance of decimal movement, or addition of an extra integer; this temperature is not biologically plausible), or they may be subtle. Using the Sort & Filter function as well as descriptive statistics described previously can help the user find errors in the data. When errors are found, assumptions should not be made as to what the data should be, even if it seems obvious that a rectal temperature of 1015.0 F should actually be 101.5 F. When errors in data are suspected, the best practice to correct them is looking at the original data, or remeasuring the value if possible. Making assumptions and correcting values in a dataset can inadvertently introduce bias into the dataset. Even the best data curation methods may not detect every error, so critical thought and diligent effort should be devoted to the data collection and entry process to reduce the risk of errors or problems in the data. As an example of using descriptive statistics to evaluate a dataset for errors, consider the data in Figure 7. The descriptive statistics contained in Figure 7a describe weaning weights recorded on 313 calves. Often data is recorded chute-side during processing, either electronically or written by hand, and entered into Microsoft Excel for evaluation and analysis later. A quick look at the descriptive statistics in Figure 7a reveals obvious errors in the data. First, the minimum value in the dataset is currently 0, although it is not possible for a calf to weigh 0 lbs. Secondly, the maximum value in the dataset is 57,000 which is also not a biologically plausible weaning weight. These values, especially 57,000, are outliers, or extreme values that fall well out of the range of other values in the dataset. Outliers are not always errors; however, it is a good idea to critically evaluate values that seem extreme in a dataset.

Outliers may have a large impact on descriptive statistics. The mean, median, and mode are measures of

central tendency, or measures of where most of the data within a dataset tends to be centered. When data follows the Normal (i.e. Gaussian) distribution, the values of the mean and median will be very similar, if not identical. Outliers in a dataset cause the distribution to be skewed towards the outliers. The more extreme the outlier, the larger the impact on the mean. The median is more resistant to the effects of outliers than the mean, because one extreme value only moves the median one number.

Zeros within datasets may pose a problem for analysis if they are used in place on absent data (e.g. it is possible for a cow to have given birth to zero calves, and it is possible that a calf gained 0 lbs., but it is not possible for an animal to weigh 0 lbs.). Erroneous zeros are most often encountered as errors when no value for a particular data point is collected (i.e. one calf doesn't get weighed), and a 0 is entered accidentally or by default. When values are missing, it is best to leave these cells blank within Microsoft Excel, rather than entering zeros. These zeros are inaccurate data that can lead to inaccurate conclusions. When the nature of the data being collected is such that zeros are possible, the user should be cautious not to remove or correct all zeros.

Figure 7a above shows the results of descriptive statistics with outliers (i.e. 0 and 57,000) included. Notice the mean and median are separated by about 178 lbs. After evaluation of original data, or remeasuring the weights of these calves, the zero is removed and the 57,000 is corrected to 570. Making these corrections in the dataset produces the descriptive statistics shown in Figure 7b. In Figure 7b, the values of the mean and median are very similar. The minimum and maximum values are also plausible. If the errors in this dataset had not been found, incorrect conclusions about average weaning weight may have resulted.

Conclusion:

Microsoft Excel is a powerful data collection, organization, and analysis tool available to the veterinary practitioner. Learning to capture data accurately, organize the data in a useful manner, and evaluate a dataset for errors are the first steps in using data to explore problems with animal health and production, and create evidence-based solutions.

USDA

This work is supported by the Agriculture and Food Research Initiative Competitive Grants Program grant no. 2018-69003-28706 from the USDA National Institute of Food and Agriculture. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S Department of Agriculture.