

# Using Excel for Statistical Analysis

---

Microsoft Excel is spreadsheet software that is used to store information in columns and rows, which can then be organized and/or processed. Excel is a powerful program with an intuitive user interface, and can be a great option for entering, organizing, and cleaning data.

In addition to its spreadsheet functions, Excel provides a number of standard statistical and graphing procedures. However, these should be approached with caution, as statisticians have found numerous errors in Excel's statistical routines and distributions. Moreover, in recent years, professional statistical packages such as SPSS (a.k.a. PASW) and Stata have developed easy-to-use, point-and-click interfaces, complete with drop-down menus and dialogue boxes, making them easier to use for those not familiar with the command-line interface. For these reasons, we do not recommend using Excel for statistical analysis, beyond very basic descriptive statistics and getting a feel for your data. If you choose to enter and clean your data initially in Excel, we recommend transferring it to another program, such as Stata or SPSS, before conducting analyses. SSDS provides resources and individual consulting to assist with transferring data and with learning these statistical software packages.

This document begins with a brief review of the literature on the accuracy of Excel's statistical routines, and then offers suggestions on several procedures that can be run in Excel with confidence.

## Table of Contents

---

Caveats and Considerations .....	1
Using Formulas in Excel.....	2
Sorting.....	5
Filtering .....	8
SSDS Software Services at Stanford .....	10

## Caveats and Considerations

---

Professional statisticians have been critical of statistical procedures in Excel for many years, at least since the 1997 distribution of the program. Recent assessments have found that many of the errors in Excel's algorithms persist in the 2007 release. Yalta (2008) assessed Excel's computation of several statistical distributions, and found substantive errors in almost all. He finds that Excel will report more significant figures in its answer than it has accurately calculated. He compared Excel to two free, open-source programs, Gnumeric 1.7.11 and OpenOffice.org Calc 2.3.0, and found both of these to be more accurate than Excel.

McCullough and Heiser (2008) further find, "Excel 2007, like its predecessors, fails a standard set of intermediate-level accuracy tests in three areas: statistical distributions, random number generation, and estimation" (4570). Discussing Excel's procedure for exponential

smoothing, the authors find it is “grievously flawed; we wonder how such obvious errors could have been made.” They find however, that Excel’s procedures for univariate, ANOVA, and linear regression analysis are acceptable, but strenuously caution against using the Solver optimization tool. They additionally recommend against using the LOGEST and GROWTH functions, as well as the Normal Probability Plot, which is used to check the residuals for normality. Finally, they cite others’ work showing inaccurate t-test results in the presence of missing values, inaccurate p-values from a t-test, and incorrect labeling of t-test and z-test tables.

Beyond these considerable problems with the accuracy of statistics Excel reports, other critics decry misleading visuals in many Excel graphical features. The Department of Statistics and Actuarial Science at University of Iowa provide a summary:

<http://www.stat.uiowa.edu/~jcryer/JSMTalk2001.pdf>

Here is some further reading about statistical analysis in Excel:

McCullough, B.A. and David A. Heiser. 2008. “On the accuracy of statistical procedures in Microsoft Excel 2007.” *Computational Statistics and Data Analysis* 52: 4570–4578

Yalta, A. Talha. 2008. “The accuracy of statistical distributions in Microsoft Excel 2007.” *Computational Statistics and Data Analysis* 52: 4579–4586

<http://www.practicalstats.com/xlsstats/excelstats.html>

<http://people.umass.edu/evagold/excel.html>

<http://www.stat.uiowa.edu/~jcryer/JSMTalk2001.pdf>

## Using Formulas in Excel

---

Excel can be used with confidence to obtain basic descriptive statistics, such as mean, median, mode, maximum, and minimum. All of these functions can be accessed through Excel’s formula function.

To enter a formula, choose an empty cell. In this cell, type the equal sign “=”. Whatever you type after the “=” is considered the formula. For example, you can type

= A1 + A2

and then press the Enter button. The cell will now display the sum of cells A1 and A2. You can achieve the same result by typing “=”, then using your mouse and clicking on cell A1, typing “+”, and then clicking on A2 and hitting Enter.

**Note:** If either cell A1 or A2 contains non-numeric values, then the formula cell will display: “#VALUE!”

You can format the cell by right-clicking on the cell with your mouse and selecting **Format Cells...** If the error message persists, this is generally an indication of an error in your formula. Excel also provides a SUM function, which allows you to calculate a sum for a range of cells. For example, to use the SUM function on the first ten rows of column A, type in an empty cell:

=SUM(A1:A10)

You can use the SUM function on a row the same way:

=SUM(A1:M1)

You can also use the SUM function on a contiguous block of cells, for example, rows 1-5 of columns A-M:

=SUM(A1:M5)

Notice that as you type the range of cells into the formula cell, Excel outlines the range in color.

Instead of typing the range, you can select it by clicking and dragging the mouse. To do this, type:

=SUM(  
in the formula cell. Then click and drag to select the desired range. Excel will show the selected range in the formula cell:

=SUM(A1:M5

End by typing the closing parenthesis “”).

The formula interface can be used in exactly the same way on the following functions:

AVERAGE: the arithmetic mean of the selected data  
MEDIAN: the value at the 50<sup>th</sup> percentile of the selected data  
MODE: the most commonly occurring value in the selected data  
MIN: the smallest value in the selected data  
MAX: the largest value in the selected data

Other formulas for a wide range of statistical and probability functions can be found in the *Help* menu in Excel. Note: Caveats and Considerations about statistical and probability functions in Excel.

## Sorting

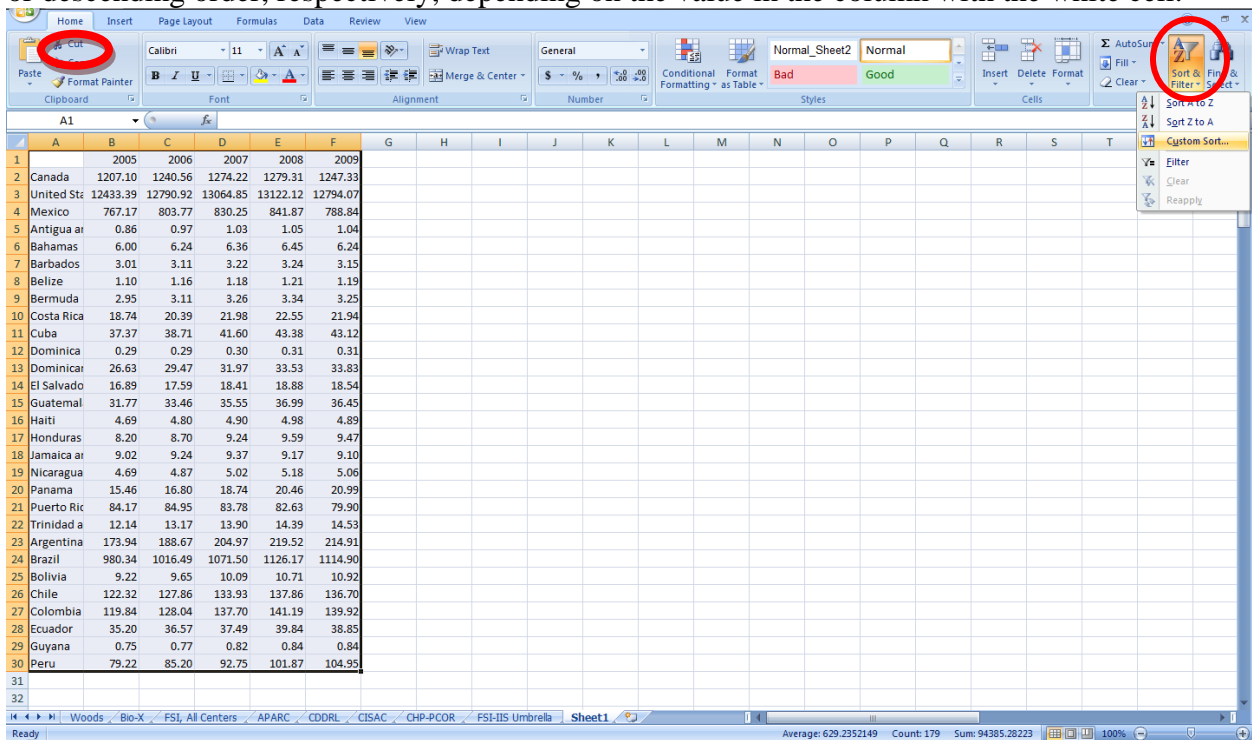
---

The SORT function will arrange your data in increasing, decreasing, alphabetical, or reverse-alphabetical order. Note: Be careful when sorting. If you sort only one row or column, you will effectively “scramble” these data relative to the rest of the spreadsheet. If the relationship

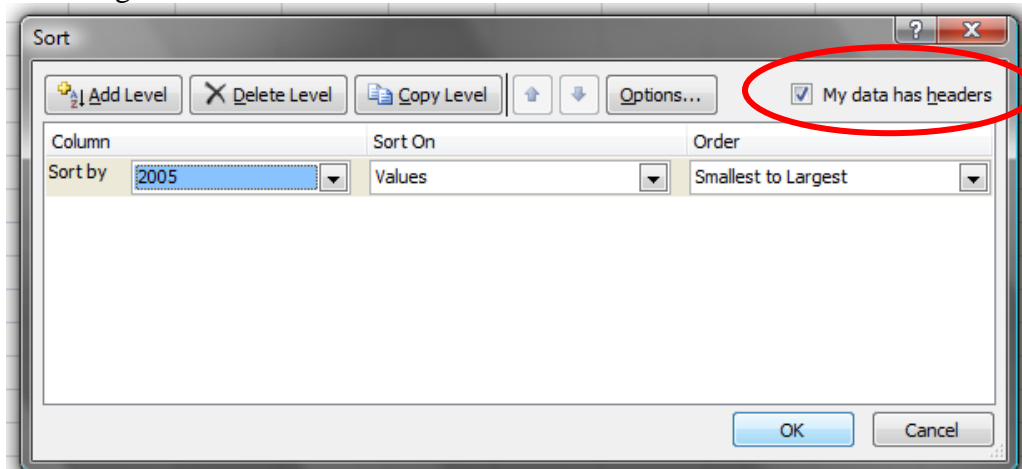
between data in different rows or columns must be preserved, always select the entire spreadsheet before sorting. Also, you can always undo a bad sort by typing “ctrl-Z” before you save.

The examples below are of GDP data for several countries in the western hemisphere. To sort this data, highlight the desired selection. With the “Home” tab selected on the top right, select the “Sort and Filter” menu from the top left.

Notice that whichever cell you last clicked in is white (in the example below it is cell A1). If you select “Sort A to Z” or “Sort A to A” from this menu, Excel will sort your data in ascending or descending order, respectively, depending on the value in the column with the white cell.

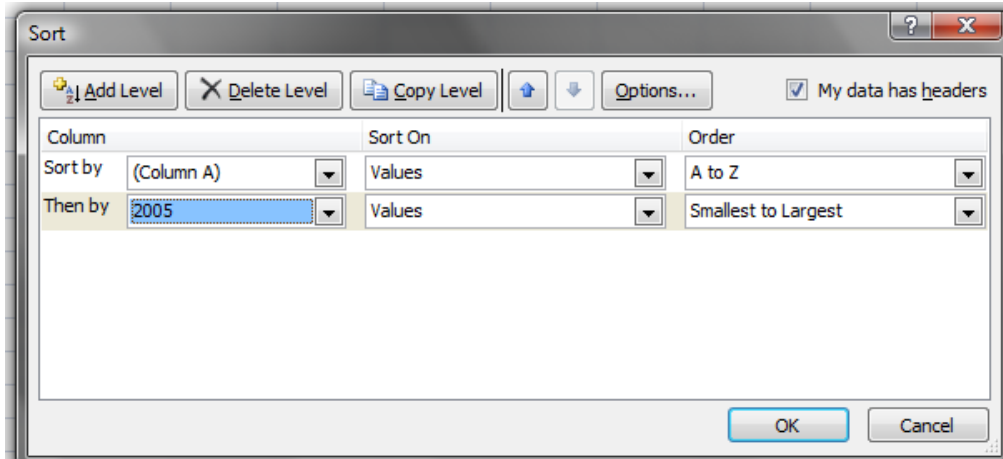


You can also choose “Custom Sort” from the “Sort and Filter” Menu, which will open the following box:



In this example, the data headers are “2005”, “2006”, “2007”, etc. Select the “My data has headers” checkbox in the top right corner so that headers will not be treated as values and mixed in with the sorting.

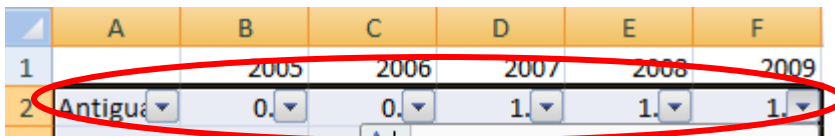
Next, in the “Sort by” drop-down menu, choose the column you would like to sort by. Leave the “Sort On” menu set to “Values”, and choose an order from the “Order” drop-down menu. Then click OK. If your data has some duplicate values, and you want to further sort within those, then you can use the “Add level” selection:



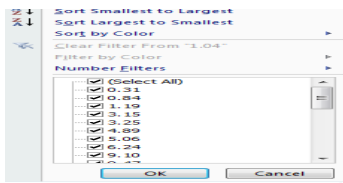
This selection would cause Excel to first sort according to country name (Column A) and then for any duplicates, sort those according to 2005 value.

## Filtering

The FILTER function allows you to select a subset of your data to display. From the same “Sort and Filter” menu used above, choose “Filter”. There will now be a small box with an arrow in the box on the first cell of each column:



If you click on one of these boxes, a dialogue box will open:



	A	B	C	D	E	F
1		2005	2006	2007	2008	2009
2	Antigua	0.	0.	1.	1.	1.
3	Argentina	173.94				
4	Bahamas	6.00				
5	Barbados	3.01				
6	Belize	1.10				
7	Bermuda	2.95				
8	Bolivia	9.22				
9	Brazil	980.34	10			
10	Canada	1207.10	11			
11	Chile	122.32				
12	Colombia	119.84				
13	Costa Rica	18.74				
14	Cuba	37.37				
15	Dominica	0.29				
16	Dominican	26.63				
17	Ecuador	35.20				
18	El Salvado	16.89				
19	Guatemala	31.77				
20	Guyana	0.75				
21	Haiti	4.69	4.80	4.90	4.98	4.89
22	Honduras	8.20	8.70	9.24	9.59	9.47
23	Jamaica ar	9.02	9.24	9.37	9.17	9.10
24	Mexico	767.17	803.77	830.25	841.87	788.84
25	Nicaragua	4.69	4.87	5.02	5.18	5.06
26	Panama	15.46	16.80	18.74	20.46	20.99
27	Peru	79.22	85.20	92.75	101.87	104.95
28	Puerto Ric	84.17	84.95	83.78	82.63	79.90
29	Trinidad a	12.14	13.17	13.90	14.39	14.53
30	United Sta	12433.39	12790.92	13064.85	13122.12	12794.07

Initially, all values are selected. You can deselect a value by clicking on the checkbox next to it. If you click on the “(Select All)” check box, you can select or deselect all.

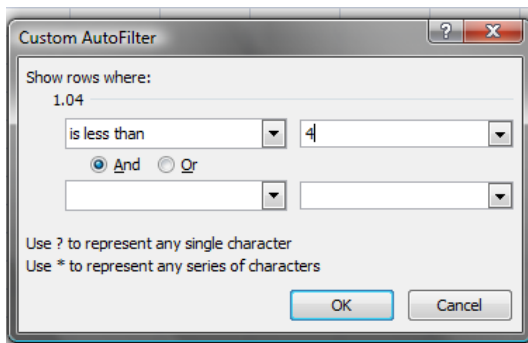
	A	B	C	D	E	F
1		2005	2006	2007	2008	2009
2	Antigua	0.	0.	1.	1.	1.
3	Argentina	173.94				
4	Bahamas	6.00				
5	Barbados	3.01				
6	Belize	1.10				
7	Bermuda	2.95				
8	Bolivia	9.22				
9	Brazil	980.34	10			
10	Canada	1207.10	11			
11	Chile	122.32				
12	Colombia	119.84				
13	Costa Rica	18.74				
14	Cuba	37.37				
15	Dominica	0.29				
16	Dominican	26.63				
17	Ecuador	35.20				
18	El Salvado	16.89				
19	Guatemala	31.77				
20	Guyana	0.75				
21	Haiti	4.69	4.80	4.90	4.98	4.89
22	Honduras	8.20	8.70	9.24	9.59	9.47
23	Jamaica ar	9.02	9.24	9.37	9.17	9.10
24	Mexico	767.17	803.77	830.25	841.87	788.84
25	Nicaragua	4.69	4.87	5.02	5.18	5.06
26	Panama	15.46	16.80	18.74	20.46	20.99
27	Peru	79.22	85.20	92.75	101.87	104.95
28	Puerto Ric	84.17	84.95	83.78	82.63	79.90
29	Trinidad a	12.14	13.17	13.90	14.39	14.53
30	United Sta	12433.39	12790.92	13064.85	13122.12	12794.07

Here, we have manually selected the first five values: 0.31, 0.84, 1.19, 3.15, and 3.25.

	A	B	C	D	E	F
1		2005	2006	2007	2008	2009
2	Antigua	0.	0.	1.	1.	1.
5	Barbados	3.01	3.11	3.22	3.24	3.15
6	Belize	1.10	1.16	1.18	1.21	1.19
7	Bermuda	2.95	3.11	3.26	3.34	3.25
15	Dominica	0.29	0.29	0.30	0.31	0.31
20	Guyana	0.75	0.77	0.82	0.84	0.84
31						
32						

Now only those rows with the selected values for 2009 are visible while other rows are hidden but not deleted. To restore all values, click on the Filter button on the 2009 column, and again Select All.

	A	B	C	D	E	F	G	H	I
1		2005	2006	2007	2008	2009			
2	Antigua	0.1	0.1	1.1	1.1	1.1			
3	Argentina	173.94							
4	Bahamas	6.00							
5	Barbados	3.01							
6	Belize	1.10							
7	Bermuda	2.95							
8	Bolivia	9.22							
9	Brazil	980.34	10.00						
10	Canada	1207.10	12.00						
11	Chile	122.32							
12	Colombia	119.84							
13	Costa Rica	18.74							
14	Cuba	37.37							
15	Dominica	0.29							
16	Dominican	26.63							
17	Ecuador	35.20							
18	El Salvador	16.89							
19	Guatemala	31.77							
20	Guyana	0.75							
21	Haiti	4.69	4.80	4.90	4.98	4.89			
22	Honduras	8.20	8.70	9.24	9.59	9.47			
23	Jamaica	9.02	9.24	9.37	9.17	9.10			
24	Mexico	767.17	803.77	830.25	841.87	788.84			
25	Nicaragua	4.69	4.87	5.02	5.18	5.06			
26	Panama	15.46	16.80	18.74	20.46	20.99			
27	Peru	79.22	85.20	92.75	101.87	104.95			
28	Puerto Rico	84.17	84.95	83.78	82.63	79.90			
29	Trinidad	12.14	13.17	13.90	14.39	14.53			
30	United States	12433.39	12790.92	13064.85	13122.12	12794.07			



You can also click on the specific column filter button to achieve the same effect. Click on the 2009 Filter button, and choose “Number Filters”. A second menu will open off to the side. From this, choose “Less Than”.

In the above example, all the values selected were less than 4. You can also choose the same values by selecting rows where the value is less than 4.

	A	B	C	D	E	F
1		2005	2006	2007	2008	2009
2	Antigua	0.1	0.1	1.1	1.1	1.1
3	Argentina	173.94				
4	Bahamas	6.00				
5	Barbados	3.01				
6	Belize	1.10				
7	Bermuda	2.95				
8	Bolivia	9.22				
9	Brazil	980.34	10.00			
10	Canada	1207.10	12.00			
11	Chile	122.32				
12	Colombia	119.84				
13	Costa Rica	18.74				
14	Cuba	37.37				
15	Dominica	0.29				
16	Dominican	26.63				
17	Ecuador	35.20				
18	El Salvador	16.89				
19	Guatemala	31.77				
20	Guyana	0.75				
21	Haiti	4.69	4.80	4.90	4.98	4.89
22	Honduras	8.20	8.70	9.24	9.59	9.47
23	Jamaica	9.02	9.24	9.37	9.17	9.10
24	Mexico	767.17	803.77	830.25	841.87	788.84
25	Nicaragua	4.69	4.87	5.02	5.18	5.06
26	Panama	15.46	16.80	18.74	20.46	20.99
27	Peru	79.22	85.20	92.75	101.87	104.95
28	Puerto Rico	84.17	84.95	83.78	82.63	79.90
29	Trinidad	12.14	13.17	13.90	14.39	14.53
30	United States	12433.39	12790.92	13064.85	13122.12	12794.07

If your columns have text values, there is also a corresponding “Text Filters” menu for columns.

# Conditional Statements: Using IF, AND, OR

## IF Formulas

The formula interface can be used for conditional statements, using the IF function. These can be very useful in cleaning data, for example checking for matching values in a range of cells. This comes in handy if you have cut-and-pasted selections from two different spreadsheets, and you want to verify that an ID column from each selection matches.

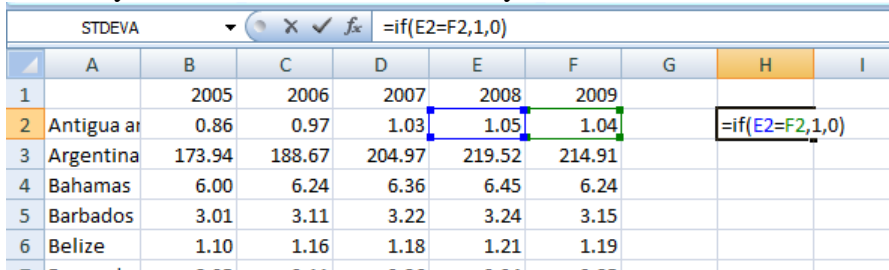
The general syntax for the IF function is:

=IF(condition, value if true, value if false)

If you want to check that values in column E match values in column F you can type in an empty cell:

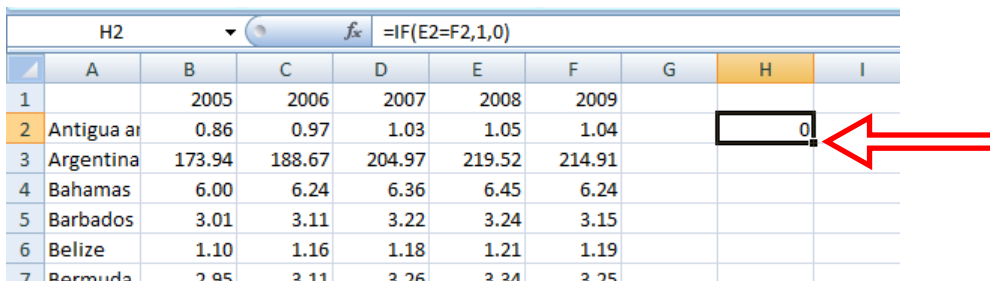
=IF(E2=F2, 1,0)

Note: if you didn't have a header row, you would use E1 and F1.



	A	B	C	D	E	F	G	H	I
1		2005	2006	2007	2008	2009			
2	Antigua ar	0.86	0.97	1.03	1.05	1.04		=if(E2=F2,1,0)	
3	Argentina	173.94	188.67	204.97	219.52	214.91			
4	Bahamas	6.00	6.24	6.36	6.45	6.24			
5	Barbados	3.01	3.11	3.22	3.24	3.15			
6	Belize	1.10	1.16	1.18	1.21	1.19			

Then hit Enter.



	A	B	C	D	E	F	G	H	I
1		2005	2006	2007	2008	2009			
2	Antigua ar	0.86	0.97	1.03	1.05	1.04		0	
3	Argentina	173.94	188.67	204.97	219.52	214.91			
4	Bahamas	6.00	6.24	6.36	6.45	6.24			
5	Barbados	3.01	3.11	3.22	3.24	3.15			
6	Belize	1.10	1.16	1.18	1.21	1.19			
7	Bermuda	2.95	3.11	3.26	3.34	3.25			

Notice that there is a 0 in the formula cell because, in this case, E2 and F2 are not equal. Now click again on the cell in which you just typed this formula. Notice that Excel highlights this cell by outlining it in black, with a small black square on the bottom right corner. Click and hold the square, and drag it down as many rows as you wish. This will carry the formula down through these rows; each new cell will display a 1 or 0, indicating whether the corresponding cells from columns E and F match.



	A	B	C	D	E	F	G	H
1		2005	2006	2007	2008	2009		
2	Antigua ar	0.86	0.97	1.03	1.05	1.04		0
3	Argentina	173.94	188.67	204.97	219.52	214.91		0
4	Bahamas	6.00	6.24	6.36	6.45	6.24		0
5	Barbados	3.01	3.11	3.22	3.24	3.15		0
6	Belize	1.10	1.16	1.18	1.21	1.19		0
7	Bermuda	2.95	3.11	3.26	3.34	3.25		0
8	Bolivia	9.22	9.65	10.09	10.71	10.92		0
9	Brazil	980.34	1016.49	1071.50	1126.17	1114.90		0
10	Canada	1207.10	1240.56	1274.22	1279.31	1247.33		0
11	Chile	122.32	127.86	133.93	137.86	136.70		0
12	Colombia	119.84	128.04	137.70	141.19	139.92		0
13	Costa Rica	18.74	20.39	21.98	22.55	21.94		0
14	Cuba	37.37	38.71	41.60	43.38	43.12		0
15	Dominica	0.29	0.29	0.30	0.31	0.31		0
16	Dominicar	26.63	29.47	31.97	33.53	33.83		0
17	Ecuador	35.20	36.57	37.49	39.84	38.85		0
18	El Salvado	16.89	17.59	18.41	18.88	18.54		0
19	Guatemal	31.77	33.46	35.55	36.99	36.45		0
20	Guyana	0.75	0.77	0.82	0.84	0.84		0
21	Haiti	4.69	4.80	4.90	4.98	4.89		0
22	Honduras	8.20	8.70	9.24	9.59	9.47		0
23	Jamaica ar	9.02	9.24	9.37	9.17	9.10		0
24	Mexico	767.17	803.77	830.25	841.87	788.84		0
25	Nicaragua	4.69	4.87	5.02	5.18	5.06		0
26	Panama	15.46	16.80	18.74	20.46	20.99		0
27	Peru	79.22	85.20	92.75	101.87	104.95		0
28	Puerto Ric	84.17	84.95	83.78	82.63	79.90		0
29	Trinidad a	12.14	13.17	13.90	14.39	14.53		0
30	United Sts	12433.39	12790.92	13064.85	13122.12	12794.07		0
31								

In this case, no cells from columns E and F match, so all formula cells are "0".

Similarly, you can use the IF statement to look for duplicates. First sort your data by column of interest by following the instructions in the previous section on sorting. For example, to check for duplicates in column A after sorting, type in an empty cell in the top row:

=IF(A1=A2,1,0)

Then select this cell, click on the small square in the bottom right corner, and drag it down to match the length of column A. Any 1's in your new column will indicate that the corresponding cell in column A matches the cell below it. If the new column contains only 0's, then there are no duplicates in column A.

## AND and OR Formulas

Linking the AND and OR functions with IF allows you to evaluate sophisticated conditionals. AND checks whether two logical statements are both true, while OR checks whether either is true. Building on the prior example, suppose you want to check whether the value from column E matches the value from column F, and at the same time, whether E1 equals 2.

Recall that the formula below tells whether the values in A1 and G1 match:

=IF(E1=F1, 1,0)

To check whether the value is 2, type in an empty cell:

=AND(IF(E1=F1,1,0), IF(E1=2,1,0))

This will display TRUE if *both* statements are true (i.e. if E1 equals F1, *and* E1=2), and FALSE otherwise.

To check whether either statement is true, use OR:

=OR(IF(E1=F1,1,0), IF(E1=2,1,0))

This will display TRUE if *either* statement is true, and FALSE otherwise.

Note: it is important to have the 1's and 0's in the right order in your IF statements. Excel equates 1 with TRUE and 0 with FALSE. In an AND or OR statement, it does not directly check whether the statements are true, only whether the IF statement returned a "1" or a "0". When evaluating an AND statement, it will check whether both IF statements returned "1"; when evaluating an OR statement, it will check whether either IF statement returned "1". You will not obtain correct results if you type:

=OR(IF(E1=F1,0,1), IF(E1=2,0,1))

## SSDS Software Services at Stanford

---

The software consultants at Social Science Data and Software (SSDS) provide technical support for SPSS users at Stanford. Users can view documents, access information about our drop-in hours, and submit questions from our web page at:

<http://ssds.stanford.edu>

*Note: This document is based on Excel 2007 for Windows*

---

Copyright © 2010, by The Board of Trustees of the Leland Stanford Junior University. Permission granted to copy for non-commercial purposes, provided we receive acknowledgment and a copy of the document in which our material appears. No right is granted to quote from or use any material in this document for purposes of promoting any product or service.

---

*Social Science Data and Software  
Document revised: 9/21/2010*