

GEOG 360: Excel Tutorial

Winter 2016: Basic & advanced functions

1. Formula entry & built-in functions

We have collected a dataset of bird occurrences in a meadow and have counted the number of birds passing through the meadow over 2 hours (see worksheet 'Shannon index'). Now we want to calculate a common biodiversity index called the Shannon index, which quantifies the biodiversity of a sample through the following formula:

$$H' = -\sum p_i \ln(p_i)$$

where p_i are the proportion of individuals (or relative abundance) belonging to bird group i .

- A. Calculate the total number of birds in the sample by using built-in functions, e.g. **=sum(cell1:cell2)**.
 - You can either type the formula into a cell directly (being sure to use '=' before the formula) if you know the syntax. Alternatively, you can go to the 'Formulas' tab where you will find 'Insert Function', which gives you a window where the different built-in functions are grouped into categories, listed, and their function is explained.
- B. Calculate the relative abundances (i.e., p_i) of each bird group. This is simply the abundance divided by the total. The sum of all relative abundances should always equal 1.
- C. Calculate the natural logarithm (ln) of each relative abundance using the built-in formula for ln().
 - If you do not know the syntax for a formula go to 'Help' (the lightbulb at the top of the page) or 'Formulas'>'Insert Function'>search for a function and search for the function you are looking for. It will list a couple of functions that match your search criteria and you can inspect which fits your needs from there.
- D. Multiply $\ln(p_i)$ by p_i (multiplication is denoted by '*').
- E. Now, calculate the sum of the value from step D and multiply the sum by -1. This is the Shannon index (compare your final calculation to the formula above).

2. Fixed versus relative cell references, multiple worksheets & graphs

We have downloaded data from FAOSTAT (<http://faostat.fao.org/>) about the total agricultural land area and the subset of that agricultural area managed organically for 8 geographical regions (see spreadsheets 'agriculture1' and 'agriculture2'). We want to calculate the percentage of agricultural land under organic production and plot this data in a graph.

- A. In the spreadsheet 'agriculture3', in the table under section 2.1, calculate the percentage of all agricultural land that is managed organically for Central America in 2004 (i.e. cell D7) through the formula:

$$\text{percentage land managed organically} = \left(\frac{\text{organic land}}{\text{agricultural land}} \right) * 100$$

You can either insert **=(agriculture2!D6/agriculture1!D6)*100** directly or you can first enter '=', then navigate with the mouse to click on the cells (within other worksheets, if needed) you want to use within your calculation. Note the syntax for referencing cell: the string of characters preceding the '!' refer to the worksheet name, and after '!' is the column letter and row number of the cell.

- B. Now that you have the formula entered in one cell, you can simply move your cursor over that cell to the bottom right corner. When your cursor changes to a dark black '+', drag the corner of cell D7 down till cell D14 and the same equation will be copied to the other cells and other regions. The same can be done horizontally across years. Note that Excel moves the position of the reference cells in spreadsheets 'agriculture1' and 'agriculture2' to match the position of cells to which we have copied the formula. This is an example of using relative cell references.
- C. Now we want to calculate the percentage of total (worldwide) agricultural land occurring in different regions. In the spreadsheet 'agriculture3' in the table under section 2.2 calculate the percentage of total organic land that is in Central America in 2004 (i.e. cell D20) through the formula:

$$\text{percentage of total organic land} = \left(\frac{\text{organic land}}{\text{total organic land}} \right) * 100$$

Again, you can insert **=(agriculture2!D7/agriculture1!D13)*100** directly or navigate with the mouse to the cells & worksheets you want to use for your calculation.

- D. Now, if we drag the corner of cell D20 down to cell D26 we get error messages since Excel now uses the wrong cells as reference. To fix the reference cell for the total organic land from a relative reference to an absolute reference, we need to add \$ signs in front of the row and/or column we want to remain fixed, i.e. **=(agriculture2!D7/agriculture2!D\$13)*100**. If we now drag down the corner of cell D20 till D26 you notice that Excel will continue using cell D13 in worksheet 'agriculture2' as reference for the total organic land. While if we drag the corner of cell D20 across to the right to calculate the percentage for different years, Excel will keep the row fixed (as we entered a \$ for the row) but will still change the column. If we wanted to fix the column as well, and use the total organic land in 2004 as reference for all years and regions we would need to enter **=(agriculture2!D7/agriculture2!\$D\$14)*100**.
- E. Now we want to graph the data. Select the data in Table 2.1 (cells A6:I14), go to the 'Insert' tab, and click on 'Scatter with straight lines'. The graph looks weird. To see what's going on, RIGHT click the graph and then on 'Select Data'. A popout appears. It shows each line that is plotted on the graph as a 'series', and the cells each series references for its labels, x values, and y values. Excel has automatically put the regions

on the x-axis and plotted a series for each year. We want to flip these – lets make each series correspond to a region, and have ‘years’ on the x-axis. On the popout, click ‘Switch Row/Column’ (top middle of the popout). *(Note: In some versions of Excel, it may automatically interpret the row of years as data, in which it will be read in as ‘series1’. But this is not data; the years are actually column headers. Check to see if this is the case for you. If so, remove the first series (click on it in the window on the left, and then press ‘Remove’). For each series now define the Name, the Y-values (i.e. the cells where the percentage data is) and the X-values (i.e. the cells where the years are). Click ok.)*

- F. Add a label to the y-axis (Design>Chart Layout/Add chart element>Axis Titles) and try different formatting options from the Home, Chart Layout, and Format menus (e.g. reduce the number of decimals shown on the y-axis, change the thickness and colour of the lines of the regions, add a title to the graph, change the scale of the y-axis). Or right click on axis, and see options

3. Interpolation/Extrapolation & conditional formatting

We have a time series of fictional SO₂ emission data from the year 1980 to the year 2005 (see spreadsheet ‘SO₂’, Table 3.1). We want to extrapolate from this data to project SO₂ emissions up to the year 2035. We will use Excel to generate a linear trend from the existing data, projecting SO₂ out to 2035.

- A. In cell AB5 insert the built-in function TREND. Use ‘Help’/’Search for a function’ to learn what this function does and what values you need to enter to use this function. Within the function, you first need to specify the cell range that contains the ‘known y-values’ (i.e. the emission values from 1980 - 2005) then the cell range that contains your ‘known x-values’ (i.e. the years 1980 - 2005), then the ‘new x-value’ for which you want to calculate a new y (i.e. 2006) and finally you need to specify whether you want to force the intercept of the linear trendline to be zero. Since emissions in year zero was *not* zero, enter TRUE.
- B. Now drag this formula down to fill in all the values for 2006, you will notice that the cell references changes using wrong cells as inputs to the function. We therefore need to fix certain cell references. For the ‘known y-values’ we want to only fix the column, as we want the rows to change with different regions. For the ‘known x-values’ we want to fix both row and column values since we always want to use line 4, and columns B through AA as reference data. For the ‘new x-value’ we want, instead, to fix only the row since we want the column to change for different years. Our final formula for cell AB5 that we can then drag-copy both horizontally and vertically to the rest of the data is therefore =TREND(\$B5:\$AA\$5,\$B\$4:\$AA\$4,AB\$4,TRUE). Already that formula?

We now want to use the same time series from 1980 to 2005 to examine how much SO₂ each country can emit if we want to reach a certain target in the year 2035. Lets assume the target for 2035 is 30,000 t SO₂ for each country. We can use the same TREND function to

interpolate to this target value. We simply have to use the trend between our current emissions (i.e. 2005) and our target emissions for 2035 (i.e. 30,000) to interpolate the emission values for the years 2006-2034. Note we do not want to use any historical data other than 2005 (why?).

- C. First create a column for 2035 right next to the column for 2005 (***Note: you can already find this in column B29 – it just needs to be moved***). The idea is to treat the 2035 targets as if they are data (using the targets within our referenced ‘known y-values’ and ‘known x-values’). Now after 2035 begin the new ‘unknown’ columns for 2006 - 2034.
 - D. Use the TREND function as described above to interpolate values for the years 2006 to 2034 in Table 3.2.
 - E. To improve visualization we now want to format cells so that we can see which countries should decrease their SO₂ emissions after 2005 versus which countries can still increase their emissions, if we adopt the equal-emissions-by-country targets. Select the column of emission data for the year 2005 and go to ‘Home’ tab, ‘Conditional formatting > Highlight Cell Rules > Greater Than...’. Enter 30,000 and choose a colour scheme (e.g. light red fill) that suits your preference. Then go to ‘Conditional Formatting’ again create a rule to highlight cells that have a value smaller than 30,000 in a different colour (e.g. light green fill).
4. In-class, we will walk through three more advanced functions or tools using a larger dataset from the ‘CountryAreaHarvested’ and ‘CountryYields’ spreadsheets to populate the respective fields for wheat and maize in Table 4 and Table 5 of the MajorCropProducers sheet. These data were downloaded from FAOSTAT in January 2016. They show national harvested areas (in hectares) and yields (in hectograms, hg, per hectare) for wheat and maize for all producing countries globally in the year 2014. Note that there are 10,000 hectograms (hg) in one metric tonne (t).

We will use the following functions or tools to populate Table 4 and Table 5.

The ‘Sort’  and ‘Filter’  features.

VLOOKUP(lookup_value, table_array, column_index_num, FALSE)

IF(condition, result if true, result if false)

5. Functions to try out on your own:

Other useful functions with which you should become familiar, if you are not already. These may help you in later assignments, and possibly in your other coursework, recordkeeping, or job:

- A. AVERAGE(range)
- B. MEDIAN(range)
- C. COUNT(range)
- D. COUNTIF(range, reference)
- E. INDEX (array, row_number, column_number)
- F. Pivot Tables