Using Excel to Analyze Experimental Data, Part II

Tips and Tricks that Simplify Handling Large Amounts of Data

May 22, 2013

NASA Advanced Supercomputing Division

Preface

- The webinar assumes basic knowledge of Excel
- The advice here is somewhat unconventional and is influenced by:
  - The author’s computer science background
  - The author’s experiences with large workbooks for analyzing
    - Benchmarking data
    - PBS usage data
- There is a lot of material here
  - You may need to invest some time after the talk to experiment with the techniques described
  - To help with that there is an accompanying workbook:
    - Excel_Webinar_Examples.xlsx
- Some of the techniques here are Excel-specific
  - Some will not work with OpenOffice, Numbers (Apple), or Google Tools

Question? Use the Webex chat facility to ask the Host

Recap of first webinar

1. Useful to think of Excel as a <u>functional</u> programming language
2. Large tables of similar records a useful paradigm
   - Columns with computations can be constructed with copy & paste
3. Pivot tables are useful for assigning data records into buckets and then performing a data reduction operation on each bucket
4. A useful technique to improve reliability and facilitate maintenance is to use multiple sheets to separate data from formulas
   - Can "copy" data into formula sheet with **=INDIRECT(ADDRESS(…))**
   - For identical analysis of multiple datasets:
     - For each dataset, one sheet for each of:
       - Experimental data, organized as a table of records with a header
       - Parameter sheet for analysis parameters (optional)
       - Formula sheet that pulls data values and per-experiment parameters
     - Sheets are named using a pattern
       - Allows data & parameter sheet names to be calculated on formula sheet
       - (see sheets with names beginning with "Exp" in examples workbook)

Question? Use the Webex chat facility to ask the Host

Q & A from Part I

1. When separating data from formulas, what's the best way to modify multiple formula sheets?
   - From Part I: delete all but one; make changes; make copies
   - Alternative (if not making changes to graphs or pivot tables):
     - Select all formula sheet tabs
     - Make editing changes; changes will occur on each selected sheet
       - Assumes all formula sheets are identical
2. ?

Question? Use the Webex chat facility to ask the Host

Looking up Values in a Table

- Suppose you have a table of data values, e.g.,

day_id	job_id	uid	gid	queue	MAUs	SBUs
20110701	1068006	11127	40836	normal	8	11.6
20110701	1068007	11127	40836	normal	8	11.6
20110701	1068102	11127	40836	normal	8	0.3
…	…	…	…	…	…	…

and you want to lookup items there for another table:

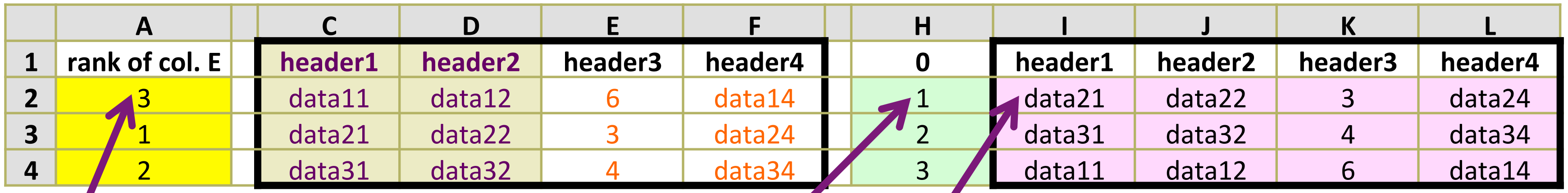
host	PBS	gid	nodes	jobname	mem(k)	wait hrs	run hrs	job_id	queue
columbia22	1121693.pbs1.0	s0836	128	nbody	970,915,904	0	45.93	1121693	vlong
columbia24	1121703.pbs1.0	s0836	128	nbody	926,170,816	0	10.35	1121703	vlong
columbia22	1091589.pbs1.0	s0836	128	nbody	916,032,832	139.65	100.1	1091589	vlong
…	…	…	…	…	…	…	…	…	…

- Use the **VLOOKUP(*x, table, c, approx*)** function
  - lookup **x** in 1st column of table
  - return item in column **c**

Question? Use the Webex chat facility to ask the Host

Trick #5: Sorting as a Functional Operation

- Not talking about **Data:Sort** (that is *imperative*); instead:
  - Have unsorted data, say on experimental data sheet
  - Want formulas to produced sorted copy of that data:



	A		C	D	E	F		H	I	J	K	L
1	rank of col. E		header1	header2	header3	header4		0	header1	header2	header3	header4
2	3		data11	data12	6	data14		1	data21	data22	3	data24
3	1		data21	data22	3	data24		2	data31	data32	4	data34
4	2		data31	data32	4	data34		3	data11	data12	6	data14

Table to sort using Column E

Table sorted using column E (K)

H2 is **=H1+1**

A2 is **=RANK.EQ(E2,E$2:E$4,1)**

I2 is **=VLOOKUP($H2,$A$2:$F$4,COLUMN()-COLUMN($F2), FALSE)**

  - Where:
    - **RANK.EQ(*x*,*range*,*order*)** is the rank of ***x*** in ***range*** (***order***=1 for ascending)
  - Other cells in same color region can be pasted from **A2**, **H2**, and **I2**
  - (See “Sort” sheet of examples workbook)

Reductions across Multiple Sheets

- Excel sheets are (obviously) 2-dimensional
- You can get 3-D "arrays" by using multiple sheets
- (Some) functions allow 3-D ranges to be supplied
  - **SUM('Sheet1:Sheet3'!A1:D4)**
  - Similarly: **PRODUCT**, **MAX**, **MIN**, **AVERAGE**
- Suppose you have a collection of experiments on sheets named:

  Exp1 Exp2 … Exp*N*

  - Tip: put in empty "sentinel" sheets on either side:

    AllExp< Exp1 Exp2 … Exp*N* >

  - Can *average* location D3 across all experiments with:

    =AVERAGE('AllExp<:>'!D3)

    (the empty cells don't cause a problem)
  - Adding new experiment sheets between the sentinels will cause update
  - (see "Sum over Sheets" in example workbook)

Question? Use the Webex chat facility to ask the Host

Array Formulas

- Suppose you want to compute a weighed average:

	A	B	C	D
1		weight	case I	case II
2	A	1.0	2	2
3	B	2.0	0	
4	C	1.0	4	4
5	weighted average		1.50	3.00

- Could compute **C5** with:

  =SUMPRODUCT(B2:B4,C2:C4) / SUM(B2:B4)

- But if we want to ignore blank values (e.g. **D3**):

  {=SUMPRODUCT(B2:B4,D2:D4) / SUM(IF(ISNUMBER(D2:D4),B2:B4,0))}

  - An “Array Formula”
  - Type *control-shift-return* instead of *return* when entering
  - (see “Array Formulas” in examples workbook)
  - Note: **IF(condition, then-value, else-value)** is a conditional expression; e.g. **IF(A1=A2, 1, 0)**

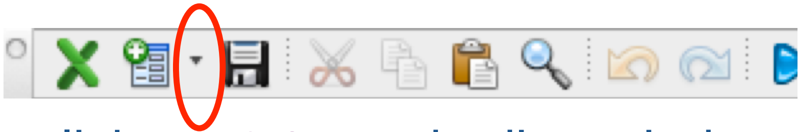
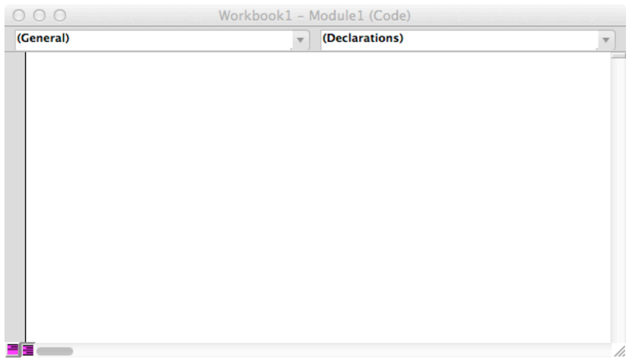
Question? Use the Webex chat facility to ask the Host

Approaching the Dark Side: Visual Basic

- Visual Basic for Applications (VBA)
  - Imperative-style programming in a BASIC-like language
  - Can simplify very complicated formulas
  - Can provide functionality not available otherwise
    - E.g. “name of worksheet #n”
- Recommendations:
  - In general, stick to functional programming style
    - I.e., no side effects in the VBA code!
  - Exception might be to provide macro-like support for administering workbook
    - E.g., updating program sheets from a template when changes have been made
- Note: workbook will need to be .xlsm file (it has “macros”)

Question? Use the Webex chat facility to ask the Host

Adding User-Defined Functions in VBA

- Enable **Developer** ribbon items
  - (Mac) **Preferences:Ribbon** check **Developer** box in list
- Click **Developer** ribbon item
- Click **Editor** item
- Click 
  - Then click **Module** and editor window pops up: 

- Type VB code in window, e.g.

```
Public Function WorksheetName(num As Long) As String
   If Worksheets.Count < num Then
      WorksheetName = ""
   Else
      WorksheetName = Worksheets(num).Name
   End If
End Function
```

  - Use in formula as: **=WorksheetName(7)** which returns name of 7th sheet

Question? Use the Webex chat facility to ask the Host

Issue: Graphs & Pivot Tables on Formula Sheets

- Issue with copying a formula sheet with a graph or a pivot table
  - Graph/table will copy, but will refer to data on original sheet not the new sheet ☹
- Workaround:
  - Copy the template worksheet to a new workbook
  - Copy the new worksheet back to original book
    - Repeat as needed to get all the copies you need
  - Rename the new sheets as appropriate
  - With a pivot table: must also "**Change Source**" to refer to new sheet name

Question? Use the Webex chat facility to ask the Host