

Spreadsheets

Capabilities
Data management

Spreadsheets

- Spreadsheets are programs designed to work with data arranged in rows and columns
- Modern spreadsheets are extremely flexible...
 - Data can be organized in many different ways
- ...and capable
 - Used for data management, graphing, data analysis
 - Extensions can add even more capabilities
- Biologists love them

Spreadsheets are Swiss Army knives

- Multi-tool
- Not designed to be the best tool for any single purpose
- Can do many different tasks well enough to get by
- Excel is Microsoft's spreadsheet program, but all work in a similar way



Spreadsheet capabilities we will use

- Managing data
- Graphing
- Data analysis
- Numerical analysis
- Programming

Worksheets

	A	B	C	D	E	F	G	H
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								

Very flexible

Cells in rows and columns

Each cell can receive data

Excel sets the data to a variable type automatically

Can use cell formulas, which allow you to do calculations on one or more cells

Rows are identified with numbers, columns are identified with letters

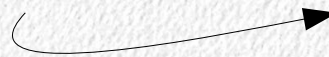
Cell A1 is selected

Cell formulas

- A cell formula starts with an equal sign
- To add a number in cell A2 to one in cell A3, you would enter:

=a2 + a3

- Symbols used for basic arithmetic operations



Operation	Symbol
Addition	+
Subtraction	-
Multiplication	*
Division	/
Exponent	^

Cell functions

- Excel comes with many built-in functions
- Functions have a name, with arguments for the function enclosed in parentheses

=average(a2, a3, a4) ← gives the average of the numbers in cells a2, a3, and a4

=average(a2:a4) works for contiguous cells

- To average the masses of leaves...

	A	B	C	D	E	F	G	H
1	Mass of leaves on trees					Mass of leaves on the ground		
2								
3	4	2.4				0.4	3.1	
4	3.1	1.7				1.5	1.6	
5	3	6.1				1	3.3	
6								
7								
8	Average	3.383333				Average	=AVERAGE(F3:G5)	
9								

Some Excel function types

- Everything on your calculator, plus many, many more

<http://office.microsoft.com/en-us/excel-help/excel-functions-by-category-HP005204211.aspx>

Data organization

- Excel is extremely flexible about data organization
 - Any data type can be entered in any cell
 - Cells have variable types, but columns and rows do not – can mix data types in a column and row
- This doesn't mean that every way you could choose to organize your data is a good idea
- Some organizations look nice on the screen, but make it difficult to work with the data later

A really bad way

	A	B	C	D	
1	Leaves from the tree				
2	Leaves from the ground				
3					
4		3.1		2.4	
5			0.4		
6		4			
7					
8			3.1		
9			1.5		
10		1.6		6.1	
11					
12	3		1		
13					
14		3.3	1.7		
15					

12 sycamore leaves, 6 picked from trees, 6 from the ground

Nothing to stop you from doing something like this, but...

Maximally inconvenient for data analysis, graphing

We could average the leaves on the tree with the command:

`=average(d4,b6,c8,d10,a12,c14)`

Leaves from the ground with:

`=average(b4,c5,c9,b10,c12,b14)`

A more sensible way

	A	B	C	D	E	F	G	H
1	Mass of leaves on trees					Mass of leaves on the ground		
2								
3	4	2.4				0.4	3.1	
4	3.1	1.7				1.5	1.6	
5	3	6.1				1	3.3	
6								

Advantages? Disadvantages?

A better way to organize the data: unstacked data

- Each column is a measurement for a different group
- Column headings are assigned to indicate the group (and should probably also indicate the measurement type)
- Does better with larger data sets than our first method

	A	B	C
1	Leaves on the ground mass	Leaves in the trees mass	
2	0.4	4	
3	1.5	3.1	
4	1	3	
5	3.1	2.4	
6	1.6	1.7	
7	3.3	6.1	
8	3.3	6.1	
9	0.8	2.5	
10	3.3	1.2	
11	0.6	2.7	
12	1	2.6	
13	2.9	5.9	
14	5	1.6	
15	1.1	2	
16	0.4	2.3	
17	0.8	1	
18	1.6	2	
19	3.5	1.8	
20	2.7	4	
21	3.5	2.1	
22	3.7	7.9	
23	0.9	6.5	
24	1.3	2.3	
25	1	1.1	
26	1.2	8	
27	2	3	
28	0.5	2.3	
29	2.8	4.9	
30	4.1	7.6	
31	1.1	7	
32	2	5	
33	1.7	9.8	
34	2.5	6.5	
35	3.4	8	
36	0.6	10.2	
37	0.7	6.5	

Better?

Problems:

Rows don't mean anything

Column names have to be complex to indicate both the group and the quantity measured

What if we measured both mass and length of leaves? How would we indicate that two measurements are from the same leaf?

Stacked data

- The most flexible data arrangement
- Rows are individual observations
- Columns are variables recorded for the observations
- Add rows as more data are collected
- Add columns if additional variables are added

	A	B	C	D	E	F	G	
1	mass	petiole.diam	max.vein.len	num.veins	leaf.thick	max.width	type	
2	4	2.7	18.7	21	0.3	154	T	
3	3.1	2	19.7	23	0.17	235	T	
4	3	2.2	11.4	23	0.25	131	T	
5	2.4	1.7	17.5	25	0.11	199	T	
6	0.4	1.2	9.4	27	0.03	105	G	
7	1.7	1.9	12	28	0.05	155	T	
8	1.5	1.4	12.1	30	0.15	155	G	
9	6.1	2.5	21.9	30	0.05	256	T	
10	6.1	3.1	19.4	30	0.31	195	T	
11	2.5	1.7	15	31	0.1	211	T	
12	1.2	1.5	12.2	31	0.21	161	T	
13	2.7	1.8	16.1	31	0.13	184	T	
14	2.6	1.8	15.2	31	0.1	214	T	
15	5.9	2.15	25.4	32	0.15	278	T	
16	1.6	1.7	10.8	32	0.14	153	T	
17	1	1.9	13.2	33	0.1	156	G	
18	2	1.3	10.2	33	0.05	164	T	
19	2.3	1.9	14.2	33	0.05	179	T	
20	3.1	2.4	18.4	34	0.22	219	G	
21	1.6	1.8	10.7	34	0.13	186	G	
22	1	1.5	10.9	34	0.07	116	T	
23	2	2.2	13	34	0.25	174	T	
24	1.8	1.8	15	34	0.05	165	T	
25	3.3	2.5	15.8	35	0.2	206	G	
26	2.2	2.2	22	35	0.2	211	G	


Stacked leaf data

Each row is a leaf

Group (tree or ground)
indicated by a column

Each measured variable
is a column, with a
heading that indicates
only what the
measurement is

Big advantage of stacked data: Pivot Tables

- Pivot Tables are a feature of Excel that works with stacked data
- Excel reads column headings, and presents them as variable names
- You can then build a data summary by dragging variable names into a template with:
 - Column fields
 - Row fields (used for grouping the data)
 - Values (i.e. the data to be summarized)
 - Filters (used for subsetting the data)
- The subject of our first exercise