Array formulas

Using array functions
Matrix operations

Arrays and matrices

- Arrays = computer science term, referring to data values organized in rows and/or columns
  - Useful for iterating = applying on operation to each element of the array
- In math, a matrix is a set of values organized in rows and columns
  - Useful for analysis of data held in rows and columns
- Excel offers array formulas for working with ranges of cells → repetitive operations on cells, or matrix math functions

Array formulas

- In Excel, an array formula is any that repeats an operation on a range of cells
- Array formulas can return a single cell as a result, or may return a range of cells
- Best seen by example...

Single cell returned: calculating your grade in a class

- You have several different assignments worth varying numbers of points
- You want to know your average percentage
- You could:
  - Calculate your percentage for each assignment
  - Calculate the average of the percents
- Or, you could:
  - Use an array formula to calculate the average of the percentages in one formula

Array formula, non-array formula approaches

C8 | {=AVERAGE(B2:B5/C2:C5)} ← *Curly braces show that this is an array formula*

	A	B	C	D	E	F
1	Assignment	My score	Points possible		My percentages	
2	Homework 1	8	10		80%	
3	Homework 2	25	26		96%	
4	Homework 3	19	29		66%	
5	Homework 4	132	140		94%	
6						
7			Average by array formula		Average of percentages	
8			84%		84%	
9						
10						

Use CTRL+SHIFT+ENTER to make it an array formula →

Non-array formula way

Calculate each percentage

Average the percentages

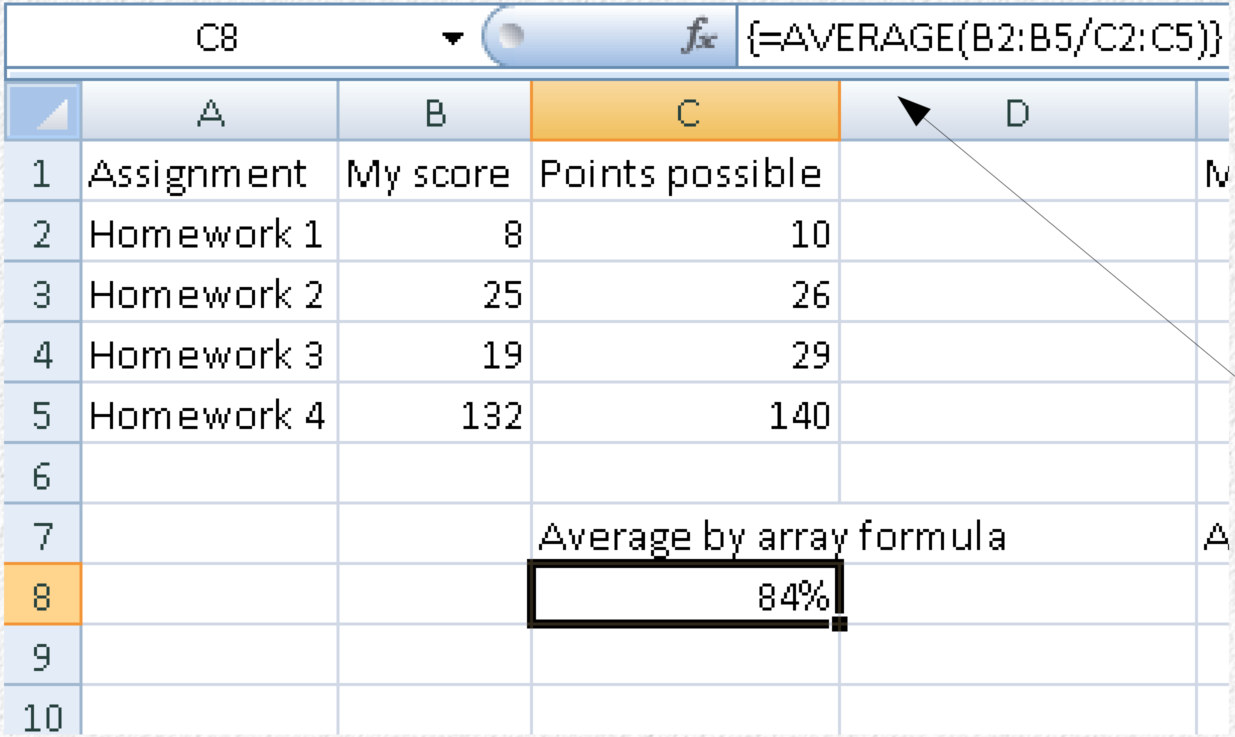
Array formula

Each percentage calculated as:

Range with scores (B2:B5) divided by range with points possible (C2:C5) → new array of proportions created (not displayed)

The array of proportions are then averaged to get the final score

Unpacking the formula...



C8 {=AVERAGE(B2:B5/C2:C5)}

	A	B	C	D
1	Assignment	My score	Points possible	
2	Homework 1	8	10	
3	Homework 2	25	26	
4	Homework 3	19	29	
5	Homework 4	132	140	
6				
7			Average by array formula	
8			84%	
9				
10				

1. Calculate all the ratios, row by row:

Matching cells in the ranges used –

B2/C2, B3/C3, B4/C4, B5/C5

The result is an array of these proportions

2. Average all the ratios

average() calculates the mean of the array of ratios:

average(B2/C2, B3/C3, B4/C4, B5/C5)

Mean of grouped data

- Consider – we have data on numbers of osprey seen at the San Elijo Lagoon during weekly surveys over two years
- We don't have the raw counts, but instead have a frequency table
- How do we calculate the mean and standard deviation of number of osprey?

C	D
Number of osprey	Frequency
0	40
1	20
2	30
3	6
4	4

Mean without an array formula

$$\frac{\sum x_i}{n}$$

=E8/E9

C	D	E
Number of osprey	Frequency	Products
0	40	0
1	20	20
2	30	60
3	6	18
4	4	16
	Sum of products	114
	Total counts	100
	Mean	1.14

← Osprey seen on the 40 counts with 0 osprey

=c2*d2

Total osprey seen across all counts

=sum(e2:e6)

Total counts

=sum(d2:d6)

Mean with an array formula

$$\frac{\sum x_i}{n}$$

{=SUM(C2:C6*D2:D6)/SUM(D2:D6)}

C	D
Number of osprey	Frequency
0	40
1	20
2	30
3	6
4	4
Mean	1.14

Array formulas and matrix algebra

- We can do matrix algebra in Excel using array functions
- An example: response to selection in scarlet gilia

Flower characteristics in *Ipomopsis aggregata*

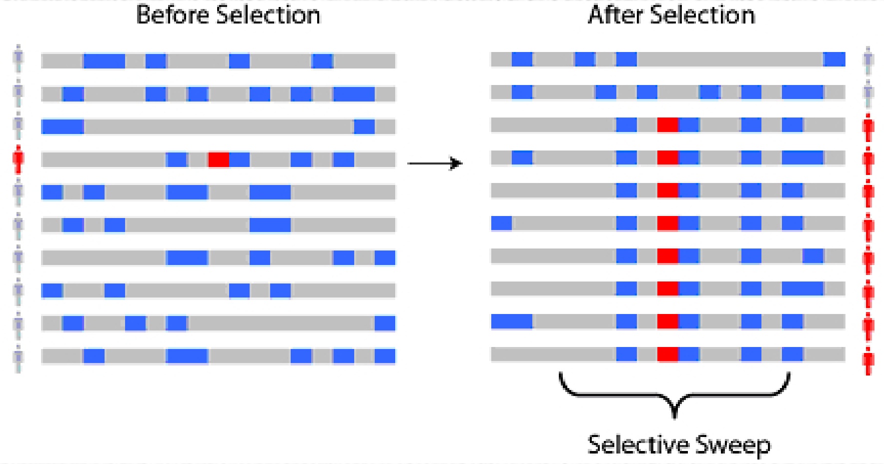
- Work by Diane Campbell (1996)
- Wanted to understand:
  - What selective pressures pollinators put on these flowers
  - How the flowers evolve in response
- Three variables:
  - Corolla length
  - Corolla width
  - Proportion pistillate (proportion of time flower can receive pollen)

Response to direct and indirect selection pressure

- Organisms respond to selection as an entire organism
- Selection on a trait is expected to cause that trait to change = direct selection
- Selection on a trait is expected to cause other traits to change that are correlated with it = indirect selection
- How can traits be correlated?

Indirect selection on correlated traits

- Indirect selection can be caused by genetic linkage
  - Selective sweeps
- Can be due to constraints
  - Resource allocation (somatic growth vs. reproduction)
  - Physical constraints (egg output constrained by body size, size-related variation in multiple traits)



Predicting responses to selection when traits are correlated

- Response due to **direct** selection will be based on the amount of heritable phenotypic variation in the trait, and the strength of selection on it
- Response due to **indirect** selection will be based on the amount of covariation between the two traits, and the strength of direct selection on the second, correlated trait
- Total response will be the sum of the direct + indirect selection
- Simplified example with two traits: Trait #1 and Trait #2

Response to direct selection

$$R_{1,d} = \sigma_1^2 \beta_1$$

Amount of change in trait #1 due to direct selection

Variance in trait

Selection gradient #1 (strength of direct selection on trait 1)

Response to indirect selection

$$R_{1,i} = cov_{1,2}\, \beta_2$$

Amount of change in trait #1 due to indirect selection

Covariance between trait 1 and 2

Selection gradient #2 (strength of direct selection on trait 2)

Trait #1

Total response for trait #1

$$R_1 = \sigma_1^2 \beta_1 + cov_{1,2}\, \beta_2$$

Direct

Indirect

Response to direct selection

Trait #2

$$R_{2,d} = \sigma_2^2 \beta_2$$

Amount of change in trait #2 due to direct selection

Variance in trait #2

Selection gradient #2 (strength of direct selection on trait 2)

Total response for trait #2

$$R_2 = \sigma_2^2 \beta_2 + cov_{1,2} \beta_1$$

Direct

Indirect

Response to indirect selection

$$R_{2,i} = cov_{1,2} \beta_1$$

Amount of change in trait #2 due to indirect selection

Covariance between trait 1 and 2

Selection gradient #1 (strength of direct selection on trait 1)

Response for both Trait #1 and Trait #2

Responses for each trait

$$R_1 = \sigma_1^2 \beta_1 + cov_{1,2} \beta_2$$

$$R_2 = \sigma_2^2 \beta_2 + cov_{1,2} \beta_1$$

Rearranged to align the selection gradients

$$\sigma_1^2 \beta_1 + cov_{1,2} \beta_2 = R_1$$

$$cov_{1,2} \beta_1 + \sigma_2^2 \beta_2 = R_2$$

Expressed as a matrix multiplication

$$\begin{bmatrix} \sigma_1^2 & cov_{1,2} \\ cov_{1,2} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$$

Matrix algebra

- Data that can be held in a matrix (rows and columns) can be manipulated using matrix algebra
- Certain types of calculations become much more efficient using a matrix approach
- We’ll focus on matrix multiplication

Matrices

- Made up of rows and columns
- A 2x2 matrix has two rows, two columns
- Element 2,1 is c
- Symbolized by bold, capital letters

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Multiplying matrices: across and down

For element 1,1 of output multiply row 1 by column 1, and add products

$$\begin{bmatrix} a & b \\ . & . \end{bmatrix} \times \begin{bmatrix} e & . \\ g & . \end{bmatrix} = \begin{bmatrix} ae+bg & . \\ . & . \end{bmatrix}$$

For element 1,2 multiply row 1 by column 2, and add products

$$\begin{bmatrix} a & b \\ . & . \end{bmatrix} \times \begin{bmatrix} . & f \\ . & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ . & . \end{bmatrix}$$

Multiplying matrices: across and down

For element 2,1 multiply row 2 by column 1, and add products

$$\begin{bmatrix} . & . \\ c & d \end{bmatrix} \times \begin{bmatrix} e & . \\ g & . \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & . \end{bmatrix}$$

For element 2,2 multiply row 2 by column 2, and add products

$$\begin{bmatrix} . & . \\ c & d \end{bmatrix} \times \begin{bmatrix} . & f \\ . & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

Is matrix multiplication commutative?

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

$$\begin{bmatrix} e & f \\ g & h \end{bmatrix} \times \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} ea+fc & eb+fd \\ ga+hc & gc+hd \end{bmatrix}$$

Some matrix multiplication facts:

- To multiply matrices, the number of columns in the left matrix has to be equal to the number of rows on the right
- The matrix produced will have the number of rows of the left matrix, and the number of columns of the right matrix

Our example response to selection as a matrix multiplication

$$\begin{bmatrix} \sigma_1^2 & cov_{1,2} \\ cov_{1,2} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 \beta_1 + cov_{1,2}\beta_2 \\ cov_{1,2}\beta_1 + \sigma_2^2 \beta_2 \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$$

*The **G matrix***

*The **β matrix***

Same total responses we started with

*The **R matrix***

Campbell’s G matrix, selection gradients

Covariances in off-diagonal elements

Indirect selection acts through the covariances

	Corolla length	Corolla width	Proportion pistillate
Corolla length	1.092	0.021	-0.039
Corolla width	0.021	0.025	0.004
Proportion pistillate	-0.039	0.004	0.002

Variances on the main diagonal – heritable genetic variation

Direct selection on a trait applies to the variances

Direct selection on:	β
Corolla length	0.05
Corolla width	1.22
Proportion pistillate	0.96

Responses to selection

Positive selection on all three traits found in the field, increases in all three predicted

G

	CL	CW	PP
CL	1.092	0.021	-0.039
CW	0.021	0.025	0.004
PP	-0.039	0.004	0.002

x

β

β
0.05
1.22
0.96

=

R

R
0.043
0.035
0.005

Negative covariance between CL and PP dampens response for CL

Positive covariance between CW with everything, PP with CW, results in bigger responses due to indirect than expected with only direct

In Excel

- Matrix multiplication is the mmult() function
  - An array function
  - Takes two arguments, the left matrix and the right matrix
  - Output is a range of cells with the number of rows of the left matrix, number of columns of the right

		G					
	CL	CW	PP		β		R
CL	1.092	0.021	-0.039		0.05		=mmult(C6:E8,G6:G8)
CW	0.021	0.025	0.004	x	1.22	=	
PP	-0.039	0.004	0.002		0.96		

CTRL+SHIFT+ENTER

		G					
	CL	CW	PP		β		R
CL	1.092	0.021	-0.039		0.05		0.04278
CW	0.021	0.025	0.004	x	1.22	=	0.03539
PP	-0.039	0.004	0.002		0.96		0.00485

Can use Excel to ask how direct and indirect selection affects the traits

	G						
	CL	CW	PP		β		R
CL	1.092	0.021	-0.039		0.05		0.0546
CW	0.021	0.025	0.004	x	0	=	0.00105
PP	-0.039	0.004	0.002		0		-0.00195

	G						
	CL	CW	PP		β		R
CL	1.092	0.021	-0.039		0		0.02562
CW	0.021	0.025	0.004	x	1.22	=	0.0305
PP	-0.039	0.004	0.002		0		0.00488

	G						
	CL	CW	PP		β		R
CL	1.092	0.021	-0.039		0		-0.03744
CW	0.021	0.025	0.004	x	0	=	0.00384
PP	-0.039	0.004	0.002		0.96		0.00192

Selecting on just one trait still causes responses in other traits because of covariances

The negative correlation between CL and PP causes direct positive selection on one to produce a decrease in the other

Conclusion: reasons to use array formulas

- Some functions can only be used as array formulas (e.g. frequency())
- Using array formulas can simplify the spreadsheet
- Using array formulas can save you work

Reasons to be cautious with array formulas

- They are easy to make mistakes with
  - Recommend you do an example calculation without array formula, then do it with array formula to check accuracy
- They need to be handled differently in the spreadsheet
  - If an array formula returns multiple cells, you can't edit part of the array