Intro Stats with Google Sheets

By: Alexander Clark
Columbia University

This version: August 27, 2023
Latest version

Preface

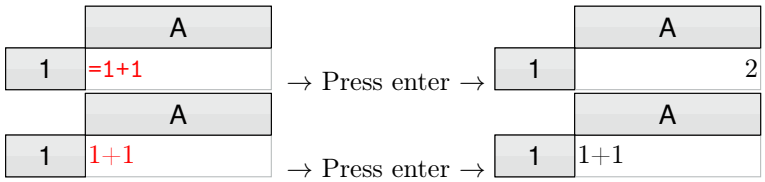
This is not a celebration of Google Sheets, though Sheets is excellent for its low scare-factor and its free availability. Life is full of tradeoffs and this is an embrace of limits. If you choose to continue with these notes, you are someone who needs to manipulate data or make statistical calculations, but not so often that you should suffer the investment required to use languages like Python or R or to learn/pay for software like Excel. Google Sheets will get you from point A to point B and with only an internet browser.

Contents

1 Basics

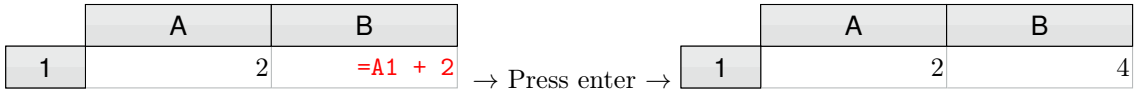
1.1 Formulas: Not `1+1` but `=1+1`

Arithmetic can be done as you would expect, as long as you enter `=` and then the formula. Without the equals sign, you will only have text.



1.2 Cell References: `A1`, `A$1`, and `$A1`

You can reference the value of another cell using its column and row coordinate. For example, `A1` refers to the value in column A, row 1.



1.3 Cell Ranges: `A1:A2`, `A:A` and `A1:B2`

You can select a range of cells with the mouse or by specifying a range of cell coordinates. When using coordinates, there are three types of ranges you might use.

1. Single column subset, `A1:A2`. This will select the cells in column A from row 1 to row 2.
2. An entire column, `A:A`. This will select every cell in column A.

3. Rectangular selection, `A1:B2`, This will select `A1:A2` and `B1:B2`. Note, you can use `A1:C1` to select all cells in row 1 for columns A-C.

4. An entire row, `1:1`. This will select every row in row 1.

Selecting a range of cells will be useful later when we use formulas. Note, if you use the reference `=A1:B2` by itself, as shown below, you will get back an error because the 2×2 range cannot be inserted into a single cell.[1]

	A	B	C
1	2		=A1:B2
2			

1.4 Copying and Pasting Cell References

To leverage the power of Sheets, you should be copying and pasting your references and formulas as much as possible. There are three types of references that behave differently when copying and pasting. A plain `=A1` reference

1. `=A1`, relative referencing
2. `=A$1`, mixed referencing where the column is relative
3. `=$A1`, mixed referencing where the row is relative
4. `=$A$1`, absolute referencing

1.4.1 Relative Column and Row `=A1`

	A	B	C	D
1	1	2	=A1	
2	10			

↓ Paste `C1` contents to `C2` and `D1`.

	A	B	C	D
1	1	2	1	2
2	10		10	

1.4.2 Relative Column, Absolute Row `=A$1`

The column is relative, because if you paste a reference to cell `A1` from `C1` to `D1`, then the cell reference is essentially updated to `B1` so that the movement in columns is reflected in the cell reference. On the other hand, if you paste `A1` from `C1` to `C2`, the row does not update because of absolute reference. The value pasted to `C2` will be that found in `A1` because there is no column movement and the row is absolute.

	A	B	C	D
1	1	2	=A$1	
2	10			

↓ Paste `C1` contents to `C2` and `D1`.

	A	B	C	D
1	1	2	1	2
2	10		1	

[1] Use the formula `=ARRAYFORMULA(A1:B2)` to copy the contents of `A1:B2` to `C1:D2`.

1.4.3 Absolute Column, Relative Row `=$A1`

The row is relative, because if you paste a reference to cell `A1` from `C1` to `C2`, then the cell reference is essentially updated to `A2` so that the movement in rows is reflected in the cell reference.

	A	B	C	D
1	1	2	=A$1	
2	10			

↓ Paste `C1` contents to `C2` and `D1`.

	A	B	C	D
1	1	2	1	1
2	10		10	

1.4.4 Absolute Column and Row `=$A$1`

	A	B	C	D
1	1	2	=A$1	
2	10			

↓ Paste `C1` contents to `C2` and `D1`.

	A	B	C	D
1	1	2	1	1
2	10		1	

2 Functions

Google Sheets offers hundreds of functions across several categories. We are especially interested in the statistical, math, and logical categories.[2]

in progress

2.1 Mathematical and Statistical

Below is a selection of commonly used function. Their purpose should be apparent from the name. To me, the trickiest thing is remembering to use `AVERAGE` instead of trying `MEAN`, which doesn't exist.

Descriptive Statistics

Function	Sample Usage	Notes
AVERAGE	`AVERAGE(A1:A10)`	
CORREL	`CORREL(A1:A10, B1:B10)`	Correlation, same as `PEARSON`.
COVAR	`COVAR(A1:A10, B1:B10)`	
MEDIAN	`MEDIAN(A1:A10)`	
PERCENTILE	`PERCENTILE(A1:A10, 0.5)`	
VAR	`VAR(A1:A10)`	This is SD^{+2}.
VARP	`VARP(A1:A10)`	
SUM	`SUM(A1:A10)`	
STDEV	`STDEV(A1:A10)`	This is SD^{+}, not SD, per Freedman et al. 2007.
STDEVP	`STDEVP(A1:A10)`	This is SD, per Freedman et al. 2007.

[2] Find the full function list at https://support.google.com/docs/table/25273?hl=en.

Next, we have some common probability distributions and the accompanying functions. You might also see the functions named differently, `NORM.DIST` instead of `NORMDIST` for example. Sometimes they are the same and sometimes they are slightly different. For example, `CHIDIST` calculates the right-tail probability and `CHISQ.DIST` calculates the left tail probability. The `cumulative` parameter should be set to `True` or `False`.

Probability Distributions

Function	Sample Usage	Notes
NORMDIST	`NORMDIST(x, mean, std dev, cumulative)`	
BINOMDIST	`BINOMDIST(num successes, trials, probability success, cumulative)`	
CHIDIST	`CHIDIST(x, degrees of freedom)`	right-tailed
CHISQ.DIST	`CHISQ.DIST(x, degrees of freedom, cumulative)`	left-tailed
TDIST	`TDIST(x, degrees of freedom, tails)`	
NORMINV	`NORMINV(probability, mean, std dev)`	
TINV	`TINV(probability, degrees of freedom)`	Two-tailed inverse

2.2 Logical

2.3 Lookup Functions

References

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). WW Norton & Company. https://clio.columbia.edu/catalog/6285515