

Database Functions

Using the Microsoft Excel 2007 AVERAGE, STDEV, DAVERAGE and DSTDEV Functions

Microsoft Excel 2007 has a whole host of built in functions: routines that are pre programmed to do things that we might use over and over again. Alternatively, we might work with things that are complicated that would be tedious to keep reprogramming. Additionally, to minimise the possibility of programming badly and not realizing it; we can rely on pre programmed functions. We will look at the following five functions:

- AVERAGE (**average** or **arithmetic mean**) function
- STDEV (**standard deviation**) function
- DAVERAGE (**database average**) function
- DSTDEV function: the **database sample standard deviation function**
- DSTDEVP function: : the **database population standard deviation function**

The AVERAGE and STDEV functions are probably familiar to you; but they are included here to provide a convenient starting point for the more powerful and flexible DATABASE functions, DAVERAGE, DSTDEV and DSTDEVP.

The purpose of the latter three functions is to allow us to build or load a table into Excel 2007 and analyse the averages (arithmetic means) and standard deviations of one, two, three or more variables, spread over one, two, three or more columns. Of course, the AVERAGE and STDEV can analyse data spread over several columns; but that's not the point as we will see now.

Children's Teeth

Let's take the example of children's teeth! Let's imagine that we wish to analyse the dental records of children from the age of 5 to 16 years; and look at the damaged, missing and filled teeth of these children. For the purposes of this demonstration, we will not be distinguishing between boys and girls, although we could build that in too. The table that follows contains the frequency of each of the categories of problem we are looking at; and they do this by the age of the child, increasing in years:

	B	C	D	E
4	Children's Teeth			
5	age	decayed	missing	filled
6	5	30	10	10
7	6	20	10	10
8	7	40	0	20
9	8	40	10	20
10	9	10	10	10
11	10	30	20	10
12	11	40	20	20
13	12	10	10	20
14	13	40	10	10
15	14	10	10	10
16	15	30	20	20
17	16	20	20	20

Table 1

For example, table 1 shows that in our survey of 'n' children, we found that there were 30 decayed teeth in children aged 5 years, there were 10 missing teeth in children aged 5 years and there were 10 filled teeth in the mouths of children aged 5 years; and so on for each age group.

Basic Statistics

We can analyse table 1 in a few ways without the need to know much about mathematics, statistics and Excel 2007. Just look at the next table and see the work we can do, see table 2:



	F	G	H	I
27	Basic Statistics			
28				
29	age	decayed	missing	filled
30	5	30	10	10
31	6	20	10	10
32	7	40	0	20
33	8	40	10	20
34	9	10	10	10
35	10	30	20	10
36	11	40	20	20
37	12	10	10	20
38	13	40	10	10
39	14	10	10	10
40	15	30	20	20
41	16	20	20	20
42				
43	Total	320	150	180
44	Average	26.6667	12.5000	15.0000
45	Standard Deviation	12.3091	6.2158	5.2223
46	Minimum	10	0	10
47	Maximum	40	20	20
48	Range	30	20	10

Table 2

The DAVERAGE Function

For all ages of all children, we know the average number of decayed teeth now, the average number of missing teeth and the average number of filled teeth. That's interesting but suppose we are testing a new kind of toothpaste, a new brushing technique or the fluoridation of water. How will the total arithmetic mean that we have just calculated help us? Well, of course it might, but let's assume that the new toothpaste was introduced, say, 2 years ago; and there has been a control group of children who are now in the age range 7 – 10 years who have been using that toothpaste, and no other, ever since. The new brushing technique might have been introduced just two years ago; and the fluoridation programme might have started only a year ago.

To test the efficacy of the new products and methods, we need to be **selective** about the statistics we compile and analyse. The database functions in Excel 2007 help us to be selective. Keeping the work as simple as possible, and taking the toothpaste case as our demonstration case, what we would like now is just to concentrate on the decayed teeth of the control group. The DAVERAGE function lets us do this as follows:

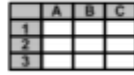
=DAVERAGE(database,field,criteria)

So not only can we see the total number of decayed, missing and filled teeth for all children in the survey, but we also know the averages of all of these problems, as well as the standard deviations, the minima, the maxima and the ranges. There are more statistics we could derive from these data; but let's concentrate on the matter in hand.

To complete the early part of this discussion, table 3 shows the detail of the functions for the totals, averages and standard deviations: we won't refer to these functions again in any detail!

	J	K	L	M
52	age	decayed	missing	filled
53		7		
54		8	>10	
55		9	>20	
56		10	>20	
57				
58	Age	decayed	missing	filled
59		5	30	10
60		6	20	10
61		7	40	0
62		8	40	10
63		9	10	10
64		10	30	20
65		11	40	20
66		12	10	10
67		13	40	10
68		14	10	10
69		15	30	20
70		16	20	20
71				
72	DAVERAGE: 7 – 10	30	30.0000	
73	DAVERAGE: 7 – 10 constrained	36.667	36.6667	
74				
75	STDEV all data	10.37	10.3701	
76				
77	DSTDEV decayed: 7 - 10	14.142	14.1421	
78	DSTDEVP decayed: 7 – 10	12.247	12.2474	
79				

Table 3



where **database** is the database, or range, where our data is stored in the spreadsheet
field is the variable we want to test: in this case it is “decayed”
criteria is one or more constraints that helps us home in on our quest!

Table 4 that follows shows us all we need to know and then we’ll explain what we’ve done.

	J	K	L	M	N	O	P
52	age	decayed	missing	filled			
53		7					
54		8	>10				
55		9	>20				
56		10	>20				
57							
58	Age	decayed	missing	filled			
59		5	30	10	10		
60		6	20	10	10		
61		7	40	0	20		
62		8	40	10	20		
63		9	10	10	10		
64		10	30	20	10		
65		11	40	20	20		
66		12	10	10	20		
67		13	40	10	10		
68		14	10	10	10		
69		15	30	20	20		
70		16	20	20	20		
71							
72	DAVERAGE: 7 – 10	30	=DAVERAGE(J58:M70,"decayed",J52:J56)				
73	DAVERAGE: 7 – 10 constrained	36.667	=DAVERAGE(J58:M70,"decayed",J52:K56)				
74							
75	STDEV all data	10.37	=STDEV(K59:M70)				
76							
77	DSTDEV decayed: 7 - 10	14.142	=DSTDEV(J58:M70,"decayed",J52:J56)				
78	DSTDEVP decayed: 7 – 10	12.247	=DSTDEVP(J58:M70,"decayed",J52:J56)				
79							

Table 4

Notice, firstly, the introduction of a new section at the top of the table (rows 7 – 11): this is the criteria section: it doesn’t have to go at the top; but if you put it at the bottom and then wanted to add more data, you might have problems if you accidentally overwrote this section.

Look at row 72, the DAVERAGE: 7 - 10 row. Here we find the DAVERAGE is 30.000 and the function we have programmed is

=DAVERAGE(J58:M70,"decayed",J52:J56)

Note: we could use a range name here instead of the range J58:M70

This means that we have asked Excel 2007 to look at the whole table of data: J58:M70 (both titles *and* values), to concentrate on the decayed variable (notice that the name of the variable we need to look at is enclosed in "" even in the function) and then to look at the criteria, in the range J52:J56 in this case. Why does this give us a DAVERAGE of 30? Well, what it has found is that when we look at the criteria range of J52:J56, this tells us to look at the decayed data for children aged 7, 8, 9 and 10 ONLY and calculate the average number of decayed teeth they have:

$$= (40 + 40 + 10 + 30) \div 4 = 120 \div 4 = 30.000$$

Look at row 73 now. Now we have asked Excel 2007 to look at children aged 7, 8, 9 and 10 BUT only to calculate the average number of decayed teeth under the following constraints:

=DAVERAGE(J58:M70,"decayed",J52:K56)

	Column K
Age 7	all decayed teeth
Age 8	only if there are more than 10 decayed teeth in this age group
Age 9	only if there are more than 20 decayed teeth in this age group
Age 10	only if there are more than 20 decayed teeth in this age group

	A	B	C
1			
2			
3			

So, the average number of decayed teeth under these constraints is:

$$= (40 + 40 + 30) \div 4 = 110 \div 4 = 36.667$$

since there are <20 decayed teeth in the 9 year age group, this group is ignored by DAVERAGE.

Play around with this function: what results do think you will get if you change the function to

`=DAVERAGE(J58:M70,"decayed",J52:L56)?`

or

`=DAVERAGE(J58:M70,"decayed",J52:M56)?`

and how do you explain your results?

Suppose now that you changed the function back to

`=DAVERAGE(J58:M70,"decayed",J52:K56)`

and then in cell K56 you entered the new criterion >30: how does this change the result you had when you calculated the DAVERAGE without this criterion?

Powerful and flexible, you should agree. You can set up all sorts of hypotheses and test them using DAVERAGE.

The DSTDEV Function

As a starting point to this part of the discussion, we have entered the STDEV function in row 75 of the table above where we have a standard deviation result for ALL damaged teeth of 10.370. However, look at row 77 and determine what is happening there: what have we done, and what does the result mean?

The format of the DSTDEV function is

`= DSTDEV(database,field,criteria)`

It's remarkably similar to the DAVERAGE function, isn't it?

Here's row 77 again:

DSTDEV decayed: 7 - 10	14.142	=DSTDEV(J58:M70,"decayed",J52:J56)
------------------------	--------	------------------------------------

So, we can use the same data, the same field and the same criteria for this function as we did with DAVERAGE. Here, we have calculated the standard deviation of the decayed teeth subject to the data in the range J58:M70. Check through this and make sure it's clear!

	A	B	C
1			
2			
3			

Work through this example yourself and change the criteria from J52:J56 to J52:K56 and see what happens: why does it happen and what does it mean?

Work through this example again and change the criteria from J52:K56 to J52:L56 and see what happens: why does it happen and what does it mean?

The DSTDEVP Function

Finally, we have included the DSTDEVP function for the sake completeness! This function is related to the DSTDEV function and the difference between the two is:

DSTDEV assumes our data are based on **SAMPLE** data; and
DSTDEVP assumes our data are based on **POPULATION** data.

This may be important if you are an advanced statistician or you need to distinguish between sample and population data and situations. Excel 2007's Help screen and a statistics text will give you more information on this point.

As a matter of interest, the STDEV function is SAMPLE based; and it has its own STDEVP brother!

Conclusion

This page has demonstrated in brief the average (AVERAGE) and standard deviation (STDEV) functions of Microsoft Excel spreadsheets. Moreover, it has introduced the idea of the database versions of these functions, DAVERAGE and DSTDEV: these versions allow us to interrogate any basic data we have by setting constraints that allow us to tease out a variety of relationships and test hypotheses. We demonstrated that the database average and standard deviation functions are both powerful and flexible; and add a dimension to basic statistical analysis that may not be available otherwise.

Duncan Williamson
August 2000 updated August 2009

The Excel 2007 file database_functions.xlsx accompanies this file and it can be downloaded from www.excelmaster.co.uk free of charge.