What is a Datafile?

Microdata arranged into a spreadsheet-like format with each row representing a **case** and each column a **variable**.

Cases are what (or who) you have data about. aka *Participant*, *Subject, Observation, Response*

ex. Each **person** who answered a survey
Each **county** in Virginia
Each **sample** of water from a lake
Each **attempt** at a stimulus
Each **year** in the past century

Variables represent characteristics collected about each case.

Recognizing a **datafile**:

- ✓ variable names in the first row
- ✓ data for each case on all other rows
- ✓ no other text or explanation

What type of data should I download?
If you can download the data set up for the statistical software you will use, that is the best choice.

What if I have data from other software programs?
See if you can save it in a format accepted by your software. There is also software that can convert many formats.

What is a database?
Databases allow you to store multiple datafiles in one place and getting data based on relationships between the data.

How do Text Formats and others differ?
Text formats include delimited (e.g., comma, tab, piped, etc), or fixed format. Proprietary formats can only be opened by a particular software but can hold additional information about the data and values, such as labels.

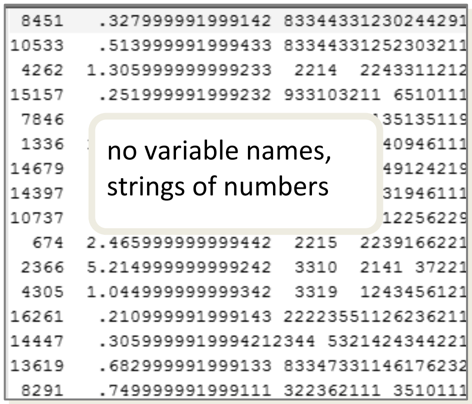
Common Datafile Extensions

	Extension	**File Type**
Common Statistical Software	.sav	SPSS
	.dta	Stata
	.sas7bdat	SAS
	.rData	R
	.rda	R (Windows)
Other Statistical Software	.jmp	JMP
	.mat	Matlab
	.mtw	MiniTab
	.rat	RATS
	.sta	Statistica
	.sys	SYSTAT
Generic Databases	.xls	Excel (Older)
	.xlsx	Excel (Newer)
	.ods	ODF Spreadsheet
	.mdb	Access
	.dbf	Databases
	.db	Databases
Text Formats	.tab	Tab Delimited
	.tsv	Tab Delimited
	.csv	Comma Delimited
	.txt	Fixed or Delimited
	.dat	Usu. has setup file

Same Data, Different Formats

Fixed Format

When hard drive space was expensive, this would give the smallest file size

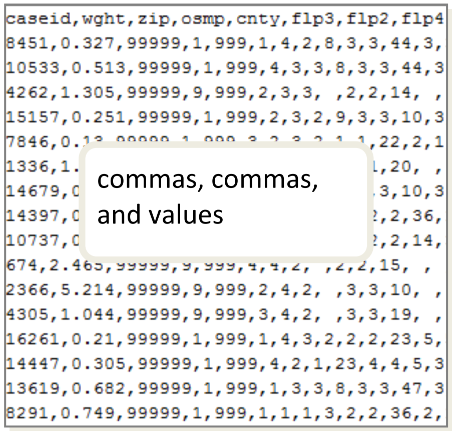


```
 8451   .327999991999142 833443312302442 91
10533   .513999991999433 83344331252303211
 4262  1.305999999999233  2214  2243311212
15157   .251999991999232 933103211 6510111
 7846                              135135119
 1336  1                           40946111
14679                              49124219
14397                              31946111
10737                              12256229
  674  2.465999999999442  2215  2239166221
 2366  5.214999999999242  3310  2141 37221
 4305  1.044999999999342  3319  1243456121
16261   .210999991999143 22223551126236211
14447   .30599999199942123 44 5321424344221
13619   .682999991999133 83347331146176232
 8291   .749999991999111 322362111 3510111
```

no variable names, strings of numbers

Comma Delimited

Use this unless commas are in your data

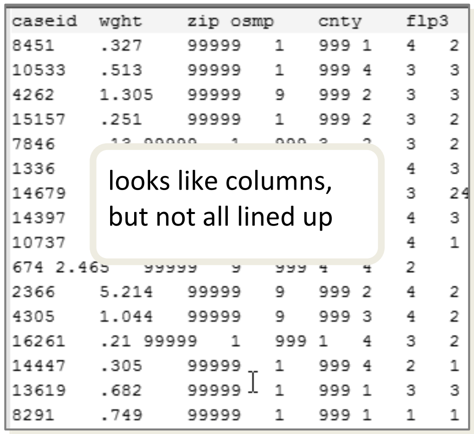


```
caseid,wght,zip,osmp,cnty,flp3,flp2,flp4
8451,0.327,99999,1,999,1,4,2,8,3,3,44,3,
10533,0.513,99999,1,999,4,3,3,8,3,3,44,3
4262,1.305,99999,9,999,2,3,3, ,2,2,14, ,
15157,0.251,99999,1,999,2,3,2,9,3,3,10,3
7846,0.13,99999,1,999,3,2,3,2,1,1,22,2,1
1336,1.                           1,20, ,
14679,0                          ,3,10,3
14397,0                          2,2,36,
10737,0                          2,2,14,
674,2.465,99999,9,999,4,4,2, ,2,2,15, ,
2366,5.214,99999,9,999,2,4,2, ,3,3,10, ,
4305,1.044,99999,9,999,3,4,2, ,3,3,19, ,
16261,0.21,99999,1,999,1,4,3,2,2,2,23,5,
14447,0.305,99999,1,999,4,2,1,23,4,4,5,3
13619,0.682,99999,1,999,1,3,3,8,3,3,47,3
8291,0.749,99999,1,999,1,1,1,3,2,2,36,2,
```

commas, commas, and values

Tab Delimited

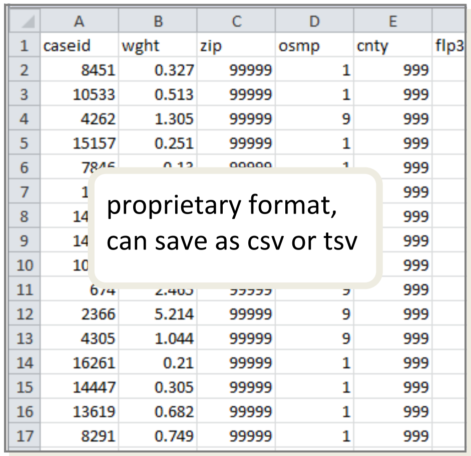
Easier to examine, if your data has no tabs



```
caseid  wght    zip osmp    cnty    flp3
8451    .327    99999   1   999 1   4   2
10533   .513    99999   1   999 4   3   3
4262    1.305   99999   9   999 2   3   3
15157   .251    99999   1   999 2   3   2
7846    .13 99999   1   999 3   2   3   2
1336                                4   3
14679                               3   24
14397                               4   3
10737                               4   1
674 2.465   99999   9   999 4   4   2
2366    5.214   99999   9   999 2   4   2
4305    1.044   99999   9   999 3   4   2
16261   .21 99999   1   999 1   4   3   2
14447   .305    99999   1   999 4   2   1
13619   .682    99999   1   999 1   3   3
8291    .749    99999   1   999 1   1   1
```

looks like columns, but not all lined up

Spreadsheet



	A	B	C	D	E	
1	caseid	wght	zip	osmp	cnty	flp3
2	8451	0.327	99999	1	999	
3	10533	0.513	99999	1	999	
4	4262	1.305	99999	9	999	
5	15157	0.251	99999	1	999	
6	7846	0.13	99999	1	999	
7	1				999	
8	14				999	
9	14				999	
10	10				999	
11	674	2.465	99999	9	999	
12	2366	5.214	99999	9	999	
13	4305	1.044	99999	9	999	
14	16261	0.21	99999	1	999	
15	14447	0.305	99999	1	999	
16	13619	0.682	99999	1	999	
17	8291	0.749	99999	1	999	

proprietary format, can save as csv or tsv