

Entering Raw Data Into R

The following data are from Tanya Winston's Psyc 497 project, completed in Spring 1999. Her sample consisted of 33 CCU students, not all of whom answered the questionnaire correctly, which explains why the subject identifier (numbers in the first column of the data table below) are not quite consecutive. Subject 5 is missing, for example. Students were asked to respond to the following statements by circling the best reply that characterized their own driving behavior. Only a subset of the questions that were asked are represented in this data set. Question numbers are in parentheses.

```
# (2) I have received a speeding ticket.  never / 1 to 3 (recoded as some) /
#      4 or more (recoded as many)
# (3) When I drive I tailgate.  never / sometimes / always
# (5) When driving I am very cautious.  never / sometimes / always
# (10) When driving I wear a seat belt.  never / sometimes / always
# (11) I talk on my cellphone while driving.  never / sometimes / always
# (12) I drive over the speed limit.  never / sometimes / always
```

subj	sex	age	ticket	tailgate	cautious	seatbelt	cellphone	speedlimit
1	F	33	never	never	always	always	sometimes	always
2	F	37	never	sometimes	sometimes	always	sometimes	sometimes
3	F	50	some	never	sometimes	always	sometimes	sometimes
4	F	20	never	sometimes	always	always	never	sometimes
6	F	27	some	never	always	always	sometimes	sometimes
7	F	22	never	sometimes	always	always	sometimes	always
8	F	21	some	never	sometimes	sometimes	never	sometimes
9	F	49	some	sometimes	always	sometimes	never	sometimes
10	F	21	some	sometimes	sometimes	always	sometimes	sometimes
11	M	43	some	sometimes	sometimes	sometimes	sometimes	sometimes
12	F	21	never	sometimes	sometimes	always	never	sometimes
13	F	22	some	never	sometimes	always	never	always
14	F	26	many	sometimes	sometimes	always	sometimes	sometimes
15	F	22	some	sometimes	sometimes	sometimes	never	sometimes
17	F	19	never	sometimes	sometimes	always	sometimes	always
18	M	37	some	never	sometimes	always	never	sometimes
20	M	21	some	sometimes	sometimes	sometimes	never	always
21	M	21	some	never	always	always	sometimes	always
22	M	21	some	never	always	always	never	sometimes
23	F	19	never	sometimes	always	always	sometimes	sometimes
24	M	26	many	sometimes	sometimes	sometimes	never	sometimes
25	F	20	some	never	sometimes	sometimes	never	sometimes
27	F	25	some	sometimes	sometimes	always	sometimes	always
28	F	49	some	sometimes	sometimes	always	never	sometimes
29	F	20	many	never	sometimes	always	sometimes	sometimes
30	M	20	never	never	always	always	never	never
31	M	22	some	sometimes	always	always	never	sometimes
32	F	21	never	sometimes	sometimes	always	never	sometimes
33	F	21	some	sometimes	always	always	never	always

Most stat programs (SPSS, SAS, etc.) have a built-in data editor that allows data entry in a spreadsheet-like format. R is no exception. However, it is generally more convenient to enter data sets other than

very simple ones using more capable software, such as a spreadsheet program. That procedure will be illustrated here.

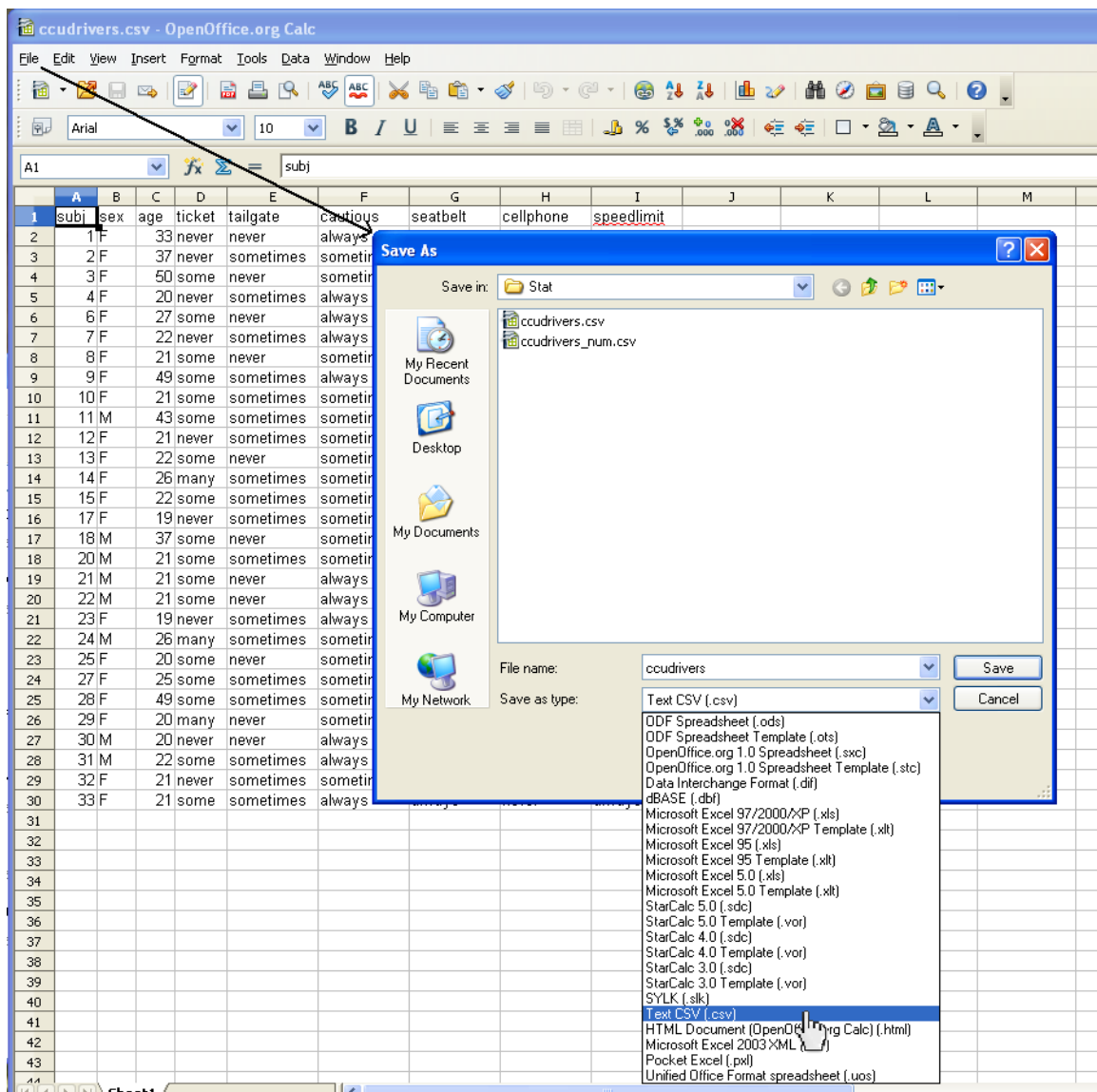
If you are "not good with Excel" (as students have told me in the past), don't worry. You won't be using any of the complex features of the program. In fact, you shouldn't use any of its complex features. Basically, you will be typing data values into boxes and then saving the file in a specific format. No specialized Excel knowledge is required. I typed these data into an OpenOffice spreadsheet, but the result will look the same, or very similar, in Excel.

The screenshot shows an OpenOffice Calc spreadsheet titled "ccudrivers.csv". The spreadsheet contains a table with 10 columns (A-J) and 33 rows. The data is as follows:

	A	B	C	D	E	F	G	H	I	J
1	subj	sex	age	ticket	tailgate	cautious	seatbelt	cellphone	speedlimit	
2	1	F	33	never	never	always	always	sometimes	always	
3	2	F	37	never	sometimes	sometimes	always	sometimes	sometimes	
4	3	F	50	some	never	sometimes	always	sometimes	sometimes	
5	4	F	20	never	sometimes	always	always	never	sometimes	
6	6	F	27	some	never	always	always	sometimes	sometimes	
7	7	F	22	never	sometimes	always	always	sometimes	always	
8	8	F	21	some	never	sometimes	sometimes	never	sometimes	
9	9	F	49	some	sometimes	always	sometimes	never	sometimes	
10	10	F	21	some	sometimes	sometimes	always	sometimes	sometimes	
11	11	M	43	some	sometimes	sometimes	sometimes	sometimes	sometimes	
12	12	F	21	never	sometimes	sometimes	always	never	sometimes	
13	13	F	22	some	never	sometimes	always	never	always	
14	14	F	26	many	sometimes	sometimes	always	sometimes	sometimes	
15	15	F	22	some	sometimes	sometimes	sometimes	never	sometimes	
16	17	F	19	never	sometimes	sometimes	always	sometimes	always	
17	18	M	37	some	never	sometimes	always	never	sometimes	
18	20	M	21	some	sometimes	sometimes	sometimes	never	always	
19	21	M	21	some	never	always	always	sometimes	always	
20	22	M	21	some	never	always	always	never	sometimes	
21	23	F	19	never	sometimes	always	always	sometimes	sometimes	
22	24	M	26	many	sometimes	sometimes	sometimes	never	sometimes	
23	25	F	20	some	never	sometimes	sometimes	never	sometimes	
24	27	F	25	some	sometimes	sometimes	always	sometimes	always	
25	28	F	49	some	sometimes	sometimes	always	never	sometimes	
26	29	F	20	many	never	sometimes	always	sometimes	sometimes	
27	30	M	20	never	never	always	always	never	never	
28	31	M	22	some	sometimes	always	always	never	sometimes	
29	32	F	21	never	sometimes	sometimes	always	never	sometimes	
30	33	F	21	some	sometimes	always	always	never	always	
31										
32										
33										

Here are the steps you should follow:

- Open the spreadsheet software to a new spreadsheet page.
- Enter a row of variable names across the first row (Row 1). These variable names should be descriptive but short and should NOT contain spaces or other strange characters. The first column should be a "subjects identifier," which is called "subj" in the example above. You can call it "subj," "subject," "case," or even "rowname" if you prefer. For that matter, you can even leave the name of this column blank.
- The spreadsheet will be set up in such a way that every variable will have its own column in the spreadsheet, and every subject (or case) will have its own row.
- Be very careful when entering data values. Once again, the most common reason for statistical errors is incorrect data entry. Remember R is case (and spelling) sensitive. Thus, each of the following is a different data value: "some", "Some", "SOME", "some " (with a space on the end, a common typing error), "som", "smoe", and so on.
- When you are finished entering data, pull down the File menu, select Save As..., and save the spreadsheet in CSV (comma separated variables) format.



CSV is sort of a spreadsheet universal language. CSV files can be read by any spreadsheet software worthy of the name and can even be examined in a text editor or word processor. (CSV files are plain text files.) While CSV format won't save fancy formatting in spreadsheets--and Excel will warn you of this over and over--you won't have any fancy formatting anyway. You should not have underlined or boldfaced anything, put boxes around anything, colored any cells, or for that matter even skipped any cells. Thus, CSV format will work just fine, and most statistical software will read data files in this format as well.

Here is another way the same data can be entered.

The screenshot shows the OpenOffice.org Calc interface with a spreadsheet titled 'ccudrivers_num.csv'. The spreadsheet has 11 columns labeled A through J and 34 rows. The data is as follows:

	A	B	C	D	E	F	G	H	I	J
1	subj	sex	age	ticket	tailgate	cautious	seatbelt	cellphone	speedlimit	
2	1	2	33	1	1	3	3	2	3	
3	2	2	37	1	2	2	3	2	2	
4	3	2	50	2	1	2	3	2	2	
5	4	2	20	1	2	3	3	1	2	
6	5	2	27	2	1	3	3	2	2	
7	6	2	22	1	2	3	3	2	3	
8	7	2	21	2	1	2	2	1	2	
9	8	2	49	2	2	3	2	1	2	
10	9	2	21	2	2	2	3	2	2	
11	10	1	43	2	2	2	2	2	2	
12	11	2	21	1	2	2	3	1	2	
13	12	2	22	2	1	2	3	1	3	
14	13	2	26	3	2	2	3	2	2	
15	14	2	22	2	2	2	2	1	2	
16	15	2	19	1	2	2	3	2	3	
17	16	1	37	2	1	2	3	1	2	
18	17	1	21	2	2	2	2	1	3	
19	18	1	21	2	1	3	3	2	3	
20	19	1	21	2	1	3	3	1	2	
21	20	2	19	1	2	3	3	2	2	
22	21	1	26	3	2	2	2	1	2	
23	22	2	20	2	1	2	2	1	2	
24	23	2	25	2	2	2	3	2	3	
25	24	2	49	2	2	2	3	1	2	
26	25	2	20	3	1	2	3	2	2	
27	26	1	20	1	1	3	3	1	1	
28	27	1	22	2	2	3	3	1	2	
29	28	2	21	1	2	2	3	1	2	
30	29	2	21	2	2	3	3	1	3	
31										
32										
33										
34										

The difference here is in the way the data values have been coded. The variable "age," a numerical variable, has not changed. Numbers are numbers. But the categorical variables, such as "sex," have now been coded as numbers. (This does not make them numerical variables! The numbers are just an alternative way of labeling the data values.) The advantage of doing this is that spreadsheets are better with numbers than they are with character values (words), typing mistakes are less likely, and so on. The disadvantage is that the data are harder to make sense of. You have to make a data code sheet or codebook to keep track of what the numbers stand for. Even so, I frequently find myself asking questions like, "What does 3 mean again?" I end up referring to the code sheet a lot, which is inconvenient.

It's not always a matter of preference, by the way. While R will handle both of these data files (saved in CSV format) without a problem, some statistics software requires that all variables be coded as numbers. Check your software manual if you are not using R. If you enter the data in the second format (as numeric codes), these can easily be converted to character coding (words) within the R program itself, so it's not like you're stuck with a bunch of cryptic numbers forever if you choose this method of coding.

If you do choose to use character coding (words), remember the three cardinal rules of naming things in R: 1) spelling counts, 2) capitalization counts, and 3) don't put spaces in anything. Thus, the name of the last column in the example above is not "speed limit," but "speedlimit," although "speed.limit" would have worked just as well.

Once the file has been created and saved in CSV format, drop a copy of it into your working directory (remember the `getwd()` command if you're not sure which folder that is), then read it into R as follows:

```
> drivers = read.csv("ccudrivers.csv", row.names=1)
```

This assumes that you saved the file as "ccudrivers.csv." If you saved it with some other filename, then use that filename, of course. It also assumes that you put the subject identifiers in column 1 of the spreadsheet. These values will be read into R as row names. If you want to use some other column as row names, change the value of the `row.names=` option accordingly. (Generally, this is not a good idea, but R is very flexible.)

If all goes well, R will return the command prompt. You will not get a message that says "mission accomplished," or anything like that. The data set will be stored in your workspace with the name "drivers." You can confirm that as follows:

```
> ls()
[1] "drivers"
```

The format the data set is in is called a dataframe. All that means is that it is a data table with variables in the columns and cases in the rows. We will talk in depth about dataframes and how to deal with them at another time.