The best way to store dates

Steve Simon

2022-10-30

If you are given data that includes dates, expect confusion. Dates can be represented in many different ways.

### Two digit years

The year value of the date may have only two digits. This leads you to guess whether 10/28/29 is referring to the year 2029, a hopeful date not too far in our future, of the year 1929, when the stock market crash started off the Great Depression. Maybe 12/9/41 is the year 2041, when you see your first grandchild. Or maybe it is 1941, the day that Franklin Delano Roosevelt said would “live in infamy.”

When you feed a two digit year into your statistical software, it will make a guess as to which century it belongs to. Some systems will assume that any two digit date is 19xx. Others will treat large values, say greater than 30, as 19xx and smaller values as 20xx.

If your software guesses wrong, you may end up with a pediatric study with patients aged 101 or -99. It’s easy to mess this up because every package is different and the guesses that your software makes may change over time.

If you get data with two digit years, be sure to redisplay the data with four digit years after you import it. Then do a careful review.

### Day and month order

In the United States, we ordinarily list dates with the month first, then the day, then the year. In most of the rest of the world, the day is listed first, followed by the month and the year.

For any day 13 and above, it will be obvious what the order is, but there’s plenty of opportunities for confusion. Next year’s Star Wars day will be 05/04/2023 here in the U.S., but they will celebrate Guerre Des Etoilles on 04/05/2023 in France.

So know what continent your data is coming from before you start analyzing it.

Month names, of course, will also vary. In Mexico, Santa Claus visits in Dic, not Dec.

### Internal storage

When you import dates, your software will either store your data as a string or as a number. You should try to get dates stored as numbers because it makes calculation of time intervals easier.

But each package chooses a different numbering system. Each system defines an index date, the date corresponding to the number zero. The index date is January 1, 1960 in SAS and Stata and January 1, 1970 in R. If you are foolish enough to manipulate dates in Microsoft Excel, the index date might be December 31, 1899 or January 1, 1904. There are some bugs in how Excel handles the leap year in 1900.

SPSS is the oddball of the bunch. It uses an index date of October 14, 1582, the start of the Gregorian calendar. It also stores not the number of days since the index date, but the number of seconds. This means that today’s date is stored as a number just shy of 14 billion. SPSS makes this choice to avoid rounding errors. It may be harder for you and me, but for the computer arithmetic operations of numbers in the 14 billion range is no harder than single digit numbers.

### ISO 8601 to the rescue

There is a standard format for displaying dates, ISO 8601. ISO, the International Standards Organization, tries to bring order to a variety of computer systems by enforcing a common standard. For dates, ISO requires you to specify a four digit year, followed by a two digit month, and ending with a two digit day. The year, month, and day are separated by dashes. So the Declaration of Independence was signed on 1776-07-04.

It’s important to include the leading zeros for single digit months and days. So 1776-7-4 is not consistent with the standard. You can’t use “July” or “Jul” either. It has to be a two digit month.

The ISO standard is nice in that if your software package stores the date as a string, the dates still sort properly.

If you have control over how dates are entered, insist on the ISO 8601 standard. Scott Munro, author of the XKCD comic series, satirizes all the variations of date storage in a cartoon,



### Conclusions

Dates come in all different shapes and sizes. This can lead to all sorts of confusion. Two digit years might be placed in the wrong century. Sometimes the months are listed before the days, sometimes afterwards. You have to do a lot of quality checks when you are using dates that someone else entered. If you are entering dates yourself, use the ISO 8601 standard.