Planning a Survival Analysis Study
… and data management issues

Steve Simon

Abstract

Planning and data management issues for survival data. Planning a study with a survival outcome requires you to specify both the number of patients and the duration of follow-up time. You'll compute power for hypothetical studies and compare power across different research designs. Then you'll review the data management needs of a survival study, with a special emphasis on the problems associated with date variables.

Selecting an Appropriate Sample Size

Formula for number of deaths

You can find a brief explanation of power in section 9.7 of Hosmer, Lemeshow, and May.

The total number of deaths to achieve a specified power, β, is

$$m = \frac{(z_{\alpha/2}+z_{\beta})^2}{\theta^2\pi(1-\pi)}$$

where π is the proportion of deaths in the first group and θ is the log of the hazard ratio.

Selecting an Appropriate Sample Size

Example of calculation

If you wish to use a two-sided test at an alpha level of 0.05, then

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

If you want to have at least 80% power, then

$$z_{\beta} = z_{0.20} = 0.84.$$

If you consider a doubling of the hazard rate as the minimum clinically important difference, then

$$\theta = ln(2) = 0.693.$$

You expect to see half as many deaths in the treatment group compared to the control group, so

$$\pi = 0.33$$

Selecting an Appropriate Sample Size

Example of calculation

Then the number of deaths total, across both groups is

$$m = \frac{(z_{\alpha/2}+z_{\beta})^2}{\theta^2\pi(1-\pi)} = \frac{(1.96+0.84)^2}{0.693^2 0.33(1-0.33)} = 73.5.$$

Round this up to 75, with 25 deaths in the treatment group and 50 deaths in the control group.

Selecting an Appropriate Sample Size

You're not done yet!

So, how many patients do you need to follow and for how long in order to get 75 deaths total?

You need to account for deaths that you never see

- Because they occur after your study ends, or
- Because of early dropouts

You need to start making assumptions. In this example assume that

- You follow the average patient for three years, and
- You will have a 20% early dropout rate,
- Deaths follow an exponential distribution, and
- The baseline hazard rate is 0.4

You can (and should) modify these assumptions to check sensitivity.

Selecting an Appropriate Sample Size

You're not done yet!

The survival function for the exponential distribution is

$$S(t) = e^{-\lambda t}$$

This produces two adjustment factors.

$e^{-1.2} = 0.3$ in the first group and $e^{-2.4} = 0.09$ in the second group.

Note that the probability of survival/death is not halved/doubled when you double the hazard rate.

Divide the number of deaths by 1 - the probability of survival and by 1 - the probability of early dropout to get the total number studied in each group.

$n_1 =$ 25 / ((1-0.2)(1-0.3)) = 44.7.

$n_2 =$ 50 / ((1-0.2)(1-0.09)) = 68.7.

To keep things simple, you might wish to use the larger of the two sample sizes in both groups.

Selecting an Appropriate Sample Size

What does a hazard ratio of 2.0 really mean?

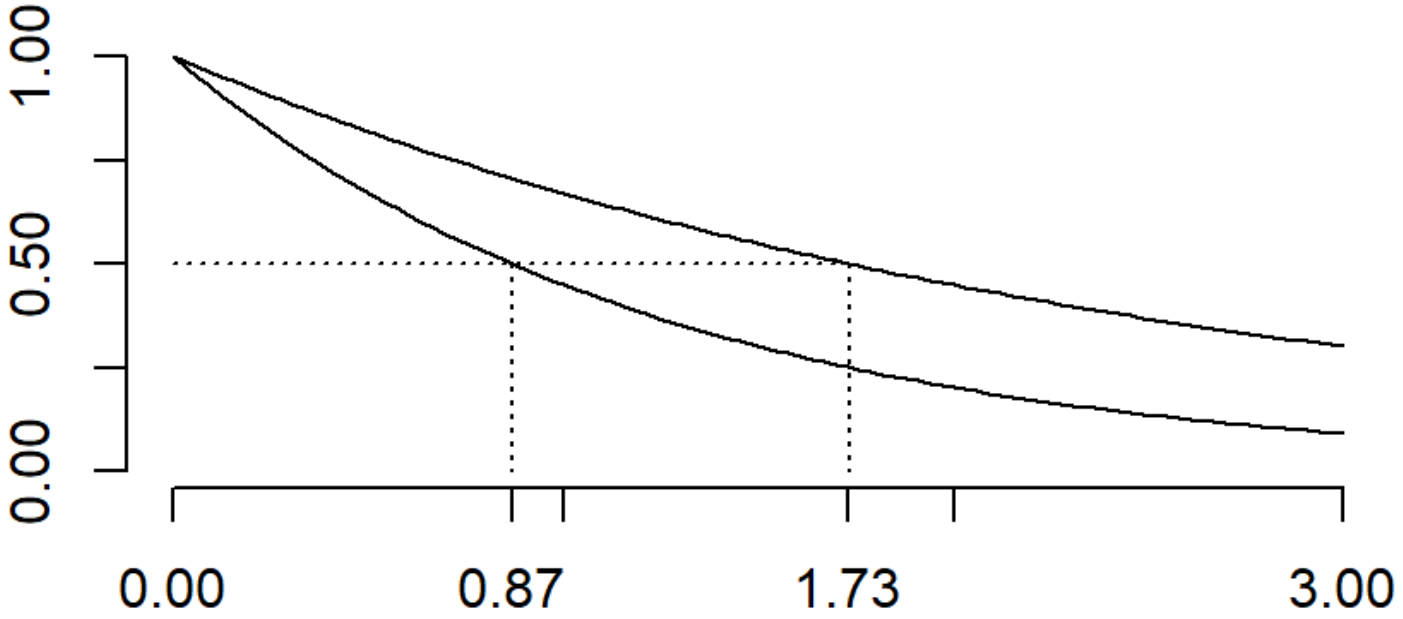
The hazard ratio is difficult to interpret for three reasons:

1. It is a unitless quantity.
2. It is a relative measure.
3. It involves rates rather than probabilities.

You should see how a hazard ratio of 2.0 changes the median survival or the probability that you survive beyond a target time point.

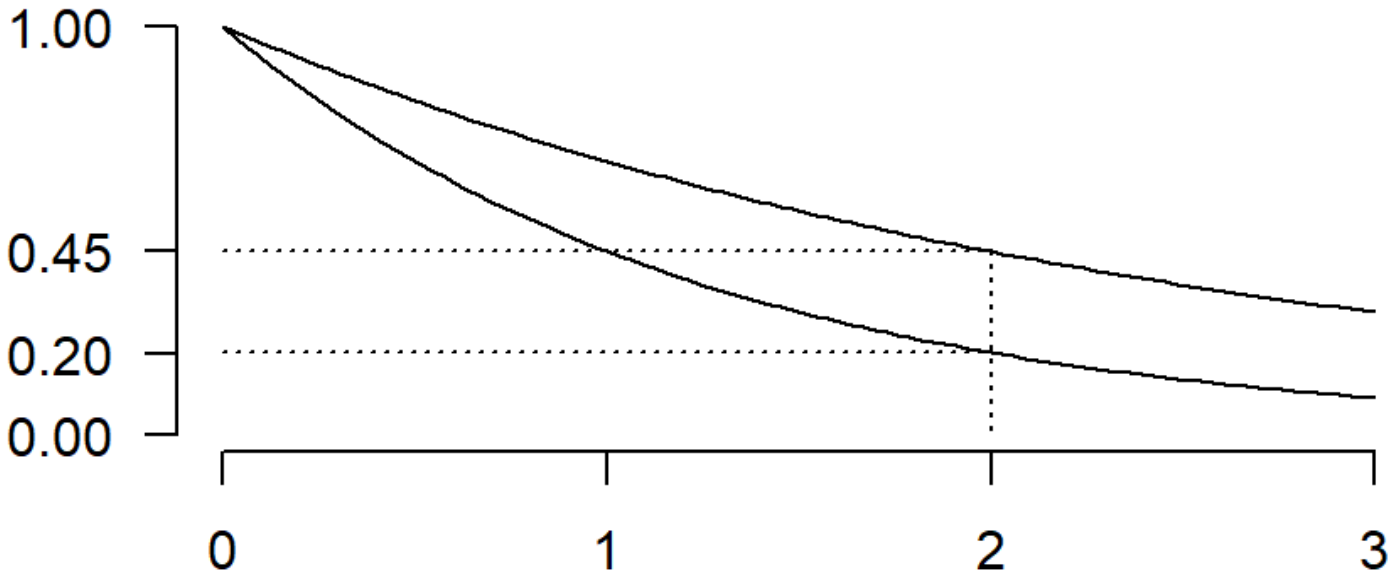
Selecting an Appropriate Sample Size

Ten extra months for the median survival time.



Selecting an Appropriate Sample Size

25% greater chance of surviving to two years.



Selecting an Appropriate Sample Size

Simulation of power

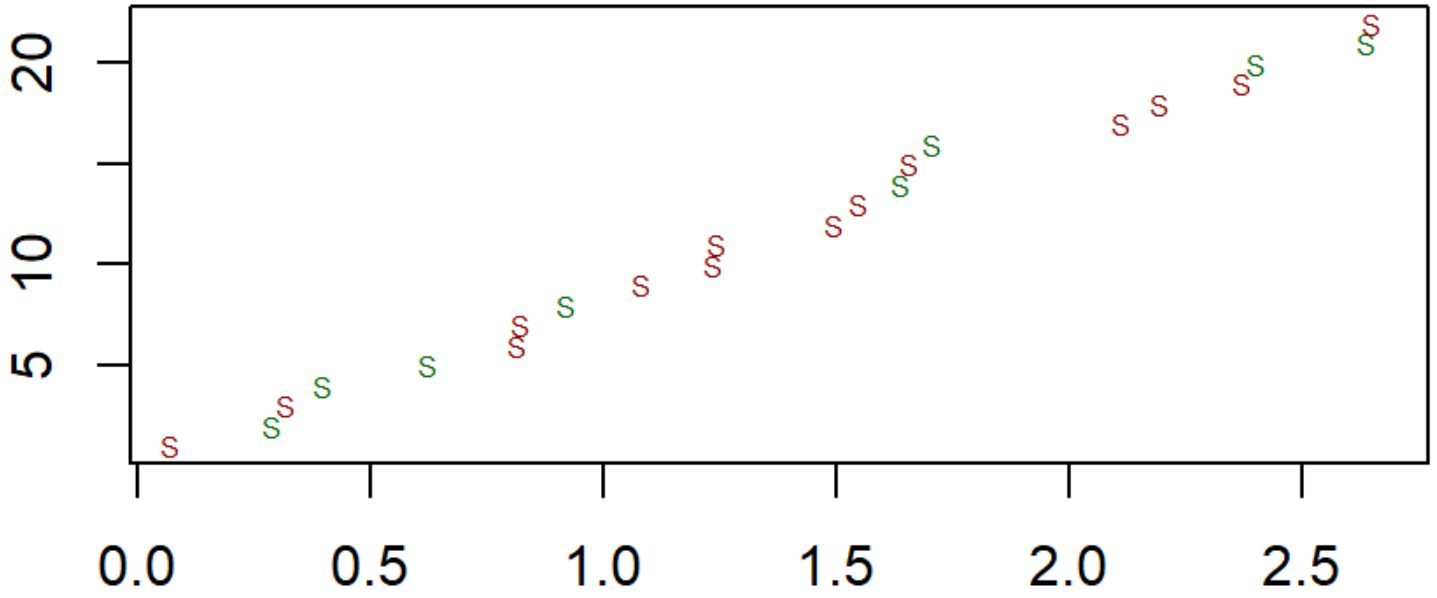
Many researchers consider a formula for sample size to be inadequate. The reasons include

1. The simplistic adjustments for censoring and dropouts.
2. The tendency for the sample size and power formulas to become unwieldy for even minor complications.
3. The inability to examine the sensitivity of your assumptions.

Simulation of power is fairly easy and very flexible.

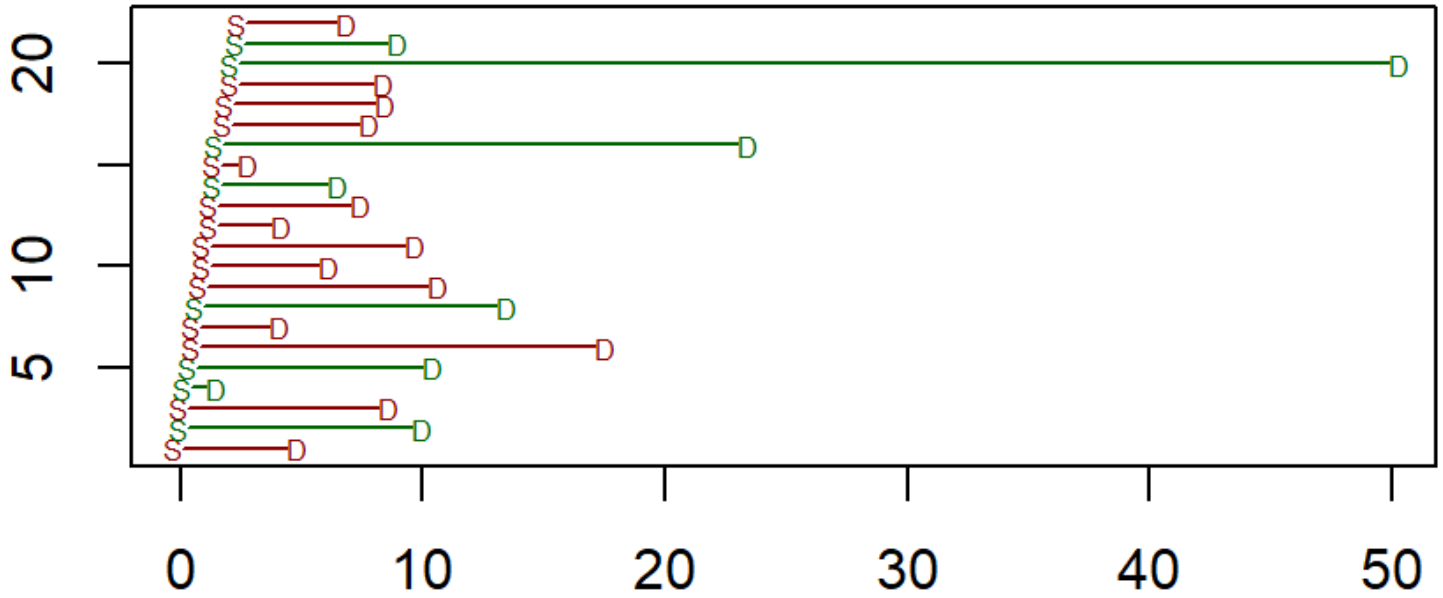
Selecting an Appropriate Sample Size

Generate random starting times



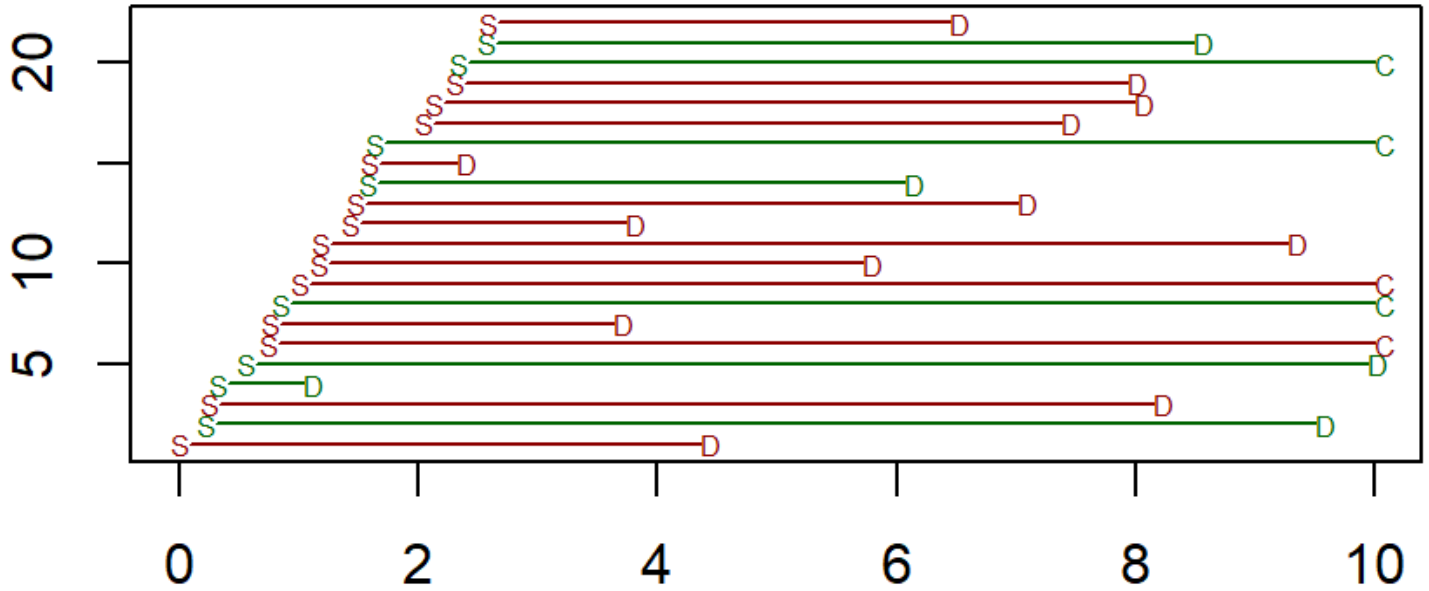
Selecting an Appropriate Sample Size

Generate random deaths



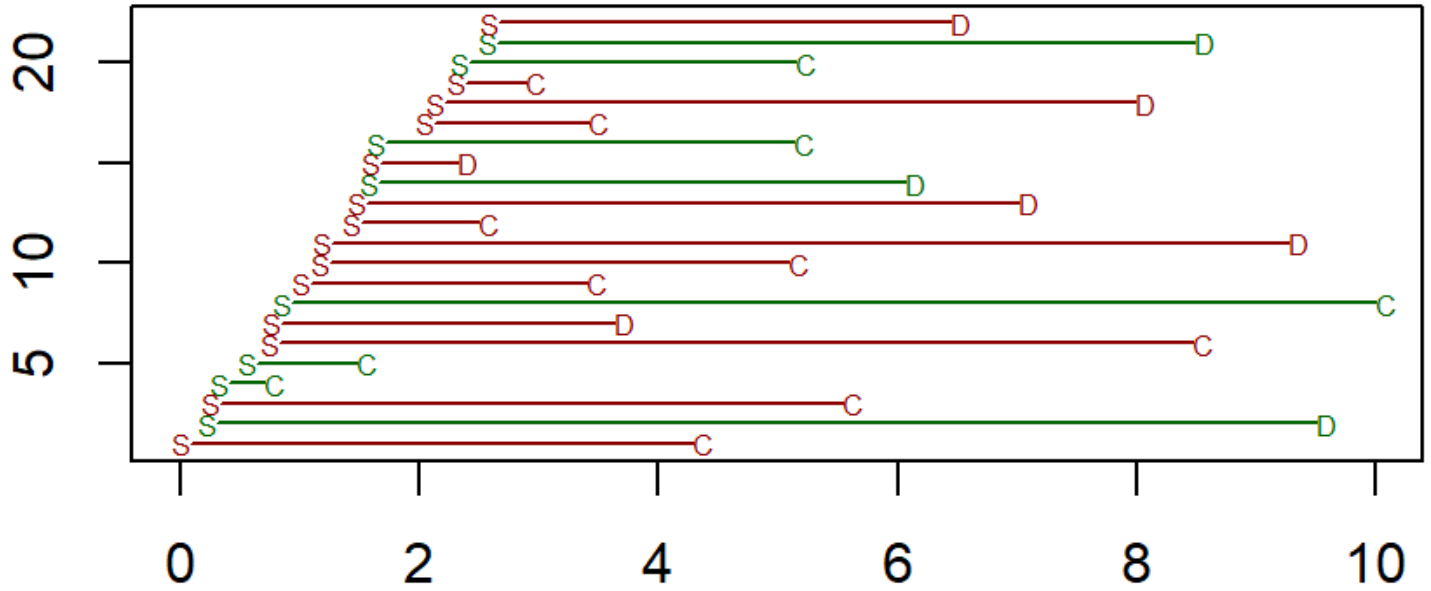
Selecting an Appropriate Sample Size

Truncate the study at a maximum study duration time.



Selecting an Appropriate Sample Size

Simulate a random process for early dropouts.



Selecting an Appropriate Sample Size

Repeat 1,000 times.

Calculate the test statistic. Repeat 1,000 times. The proportion of test statistics that reject the null hypothesis is your power.

Consider sensitivity analyses. How much does your power/sample size change as you change some of the underlying assumptions.

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS ***THE*** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. $27\frac{1}{2}$-13 2013.158904109
MMXIII-II-XXVII MMXIII $\frac{\text{LVII}}{\text{CCCLXV}}$ 1330300800
$((3+3)\times(111+1)-1)\times 3/3-1/3^3$ ~~2013~~ 

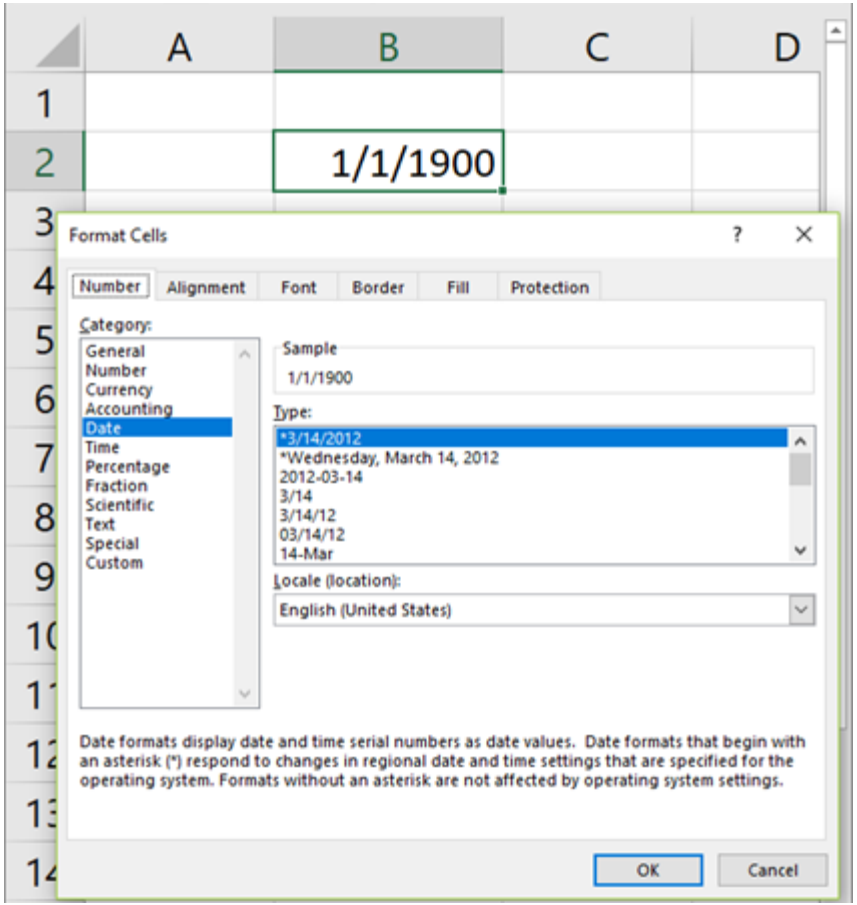
10/11011/1101 02/27/20/13 $\overset{2}{0}\overset{3}{1}\overset{1}{2}\overset{4}{3}\overset{}{7}_{5}\ _{67}\ _{8}$

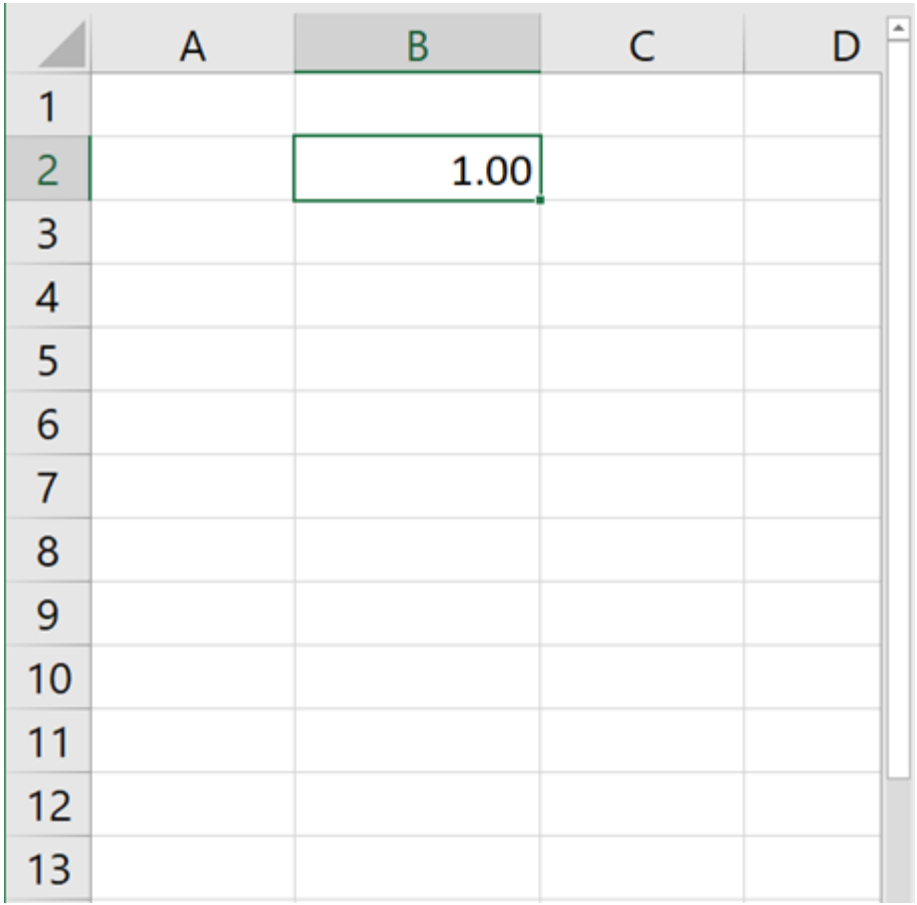
Data Management

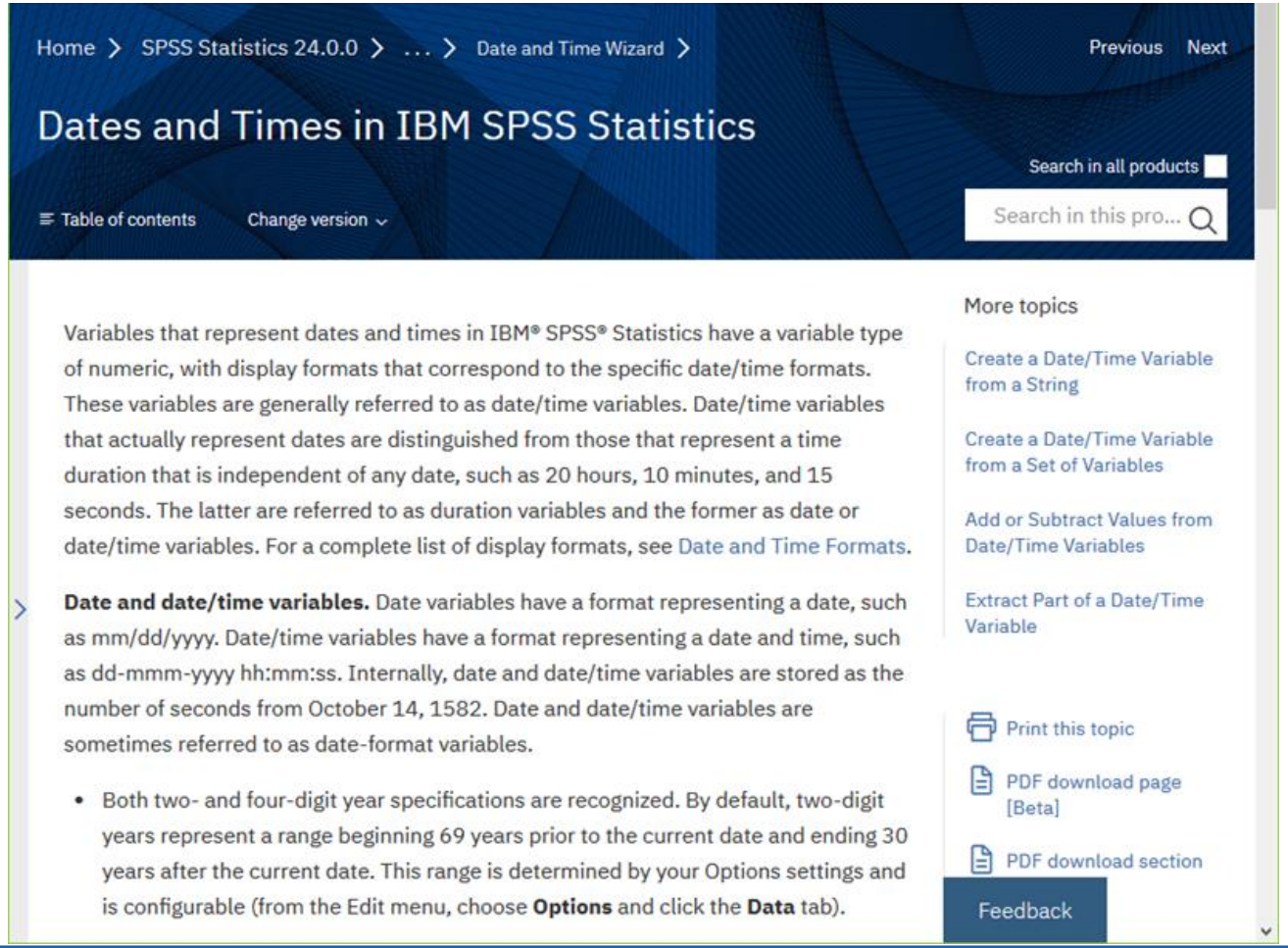
Internal storage of dates.

Every software program stores dates a little bit differently. R stores dates as internally as the number of days since January 1, 1970.

```
baseline <- as.Date("1/1/1970", format="%m/%d/%Y")
print(baseline)
## [1] "1970-01-01"
print(as.numeric(baseline))
## [1] 0
```







Data Management

Importing dates

Every software package is different, but in general when you import dates from another program, you will get

1. a date stored in the correct internal format.
2. a date stored in incorrectly.
3. a string that looks like a date but cannot be used in calculating time intervals.
4. a number related to the internal storage format of the system you are importing from.

Data Management

Importing dates

If you are entering the data yourself,

1. always specify year with four digits, and
2. use a consistent pre-defined format.

ISO 8601 format (yyyy-mm-dd) is best for two reasons

1. It sorts well, even if stored as a string.
2. It is different from both the informal American (mm-dd-yyyy) and the European (dd-mm-yyyy) standards.

Data Management

Importing dates

If your data imports incorrectly,

1. check for the wrong century,
2. check for origin problems (e.g., 1960-01-01 versus 1970-01-01),
3. try a more rigid or structured format for the import, or
4. fix things by hand.

Always document your choices.

Data Management

Manipulating dates

The computer software wants time intervals, not dates.

The time interval is usually just the subtraction of two dates.

- Use built-in functions if you can.
- Know your units of conversion
  - 86,400 seconds in a day,
  - 30.4375 days in a month,
  - 365.25 days in a year.

Data Management

The three dates you need

If you are given a data set where the censoring time is given for you, great! If you have to figure out the censoring time, you need at least three dates.

1. the date of origin,
2. the date of the event (if it occurred),
3. the date of last contact with the patient.

Data Management

The date of origin

The date of origin seems like it should be easy, but it is not. It could be

1. the date of birth,
2. the date of diagnosis,
3. the date of randomization,
4. the date when therapy was first initiated.

Data Management

The date of origin

For particular events, the date of origin might change.

- Rehospitalization, use date of first discharge.
- Failure of a mechanical device, use date of implant.
- Divorce, use date of marriage.
- Loan default, use date of loan contract.
- Infectious disease, use date of first exposure.

Data Management

The date of event

The date of the event is dependent on how you define the event.

- All-cause mortality versus mortality related to the health condition.
- Composite endpoints (death or relapse) require comparing the earlier of two dates.
- If the event did NOT happen, leave this field blank/missing.

Data Management

The date of last contact

Every patient will have a date of last contact. It will be the last time that you have been able to contact the patient and confirm that the event has not yet occurred.

If the event has occurred, you have several reasonable choices:

1. put the event date in this field also,
2. leave the field blank/missing,
3. put the last date of contact,

   (a date after the event date may represent a data error!)

Research Design Issues

Prospective analyses

Compared to most other studies, a prospective survival analysis

1. takes a lot more time,
2. requires more intensive follow-up, and
3. suffers more from dropouts.

You should plan for aggressive follow-up and/or rely on outside sources to ascertain why someone dropped out of the study.

You should collect and manage data longitudinally.

Research Design Issues

Interim analysis

Interim analyses look at early stopping for

1. vastly superior efficacy,
2. vastly worse side effects, or
3. futility.

Question: How do you handle patients with very little follow-up time at an interim analysis?

Answer: surrogate outcomes

Research Design Issues

Retrospective analyses

Retrospective studies are almost always

- faster,
- cheaper, but
- require a longitudinal data source.

You can often supplement with data from an external sources (e.g., National Death Index).

Research Design Issues

Survival analysis is an alternative to “intention to treat”

Intention to treat analysis (ITT) provides special handling of patients who

1. discontinue the assigned therapy, or
2. switch to the competing therapy.

Survival analysis with time-varying covariates provides an interesting alternative to ITT.

Conclusion

What have you learned today?

1. The initial sample size calculations are in terms of the number of events.
2. You have to extrapolate with parametric assumptions to get the number of patients and the required follow-up time.
3. Data management requires special attention to dates.
4. Survival models raises difficult, but interesting research design issues.