Lab 5, Sample Size and Power: Solutions

Dave Harrington

May 2018

Problem 1: Thinking about sample size

When designing a study, one needs to decide how long the study will run (for budgeting and analysis reasons) and estimate how many patients need to be enrolled in order to observe a certain power of tests for a given Type I error and magnitude of the difference of interest.

a) What is a potential issue if too few patients are recruited for a study?

b) Are there any issues if too many patients are recruited?

Problem 1 Solution.

a) If too few individuals are in the study, there may not be enough information in the data to declare a significant difference between treatment groups; i.e., the power may be too low. Underpowered studies may have little clinical value.

b) If there are too many patients in the study, more patients than necessary will be exposed to a new intervention which may not be an improvement and in some cases, may be harmful. More patients than necessary will experience the inconvenience of participating in a clinical trial. It is also the case that the time and resources of the study team will not be used efficiently.

   The definition of too many patients is, of course, subjective. In some instances, it may be perfectly acceptable to have power 0.80 of detecting a treatment effect or an important association. In other cases, where a new intervention is particularly expensive, it might be desirable to have relatively high power (0.90 or larger) to detect a treatment effect.

Problem 2: Calculating power

Consider a phase III cancer clinical trial in which the time to tumor progression is the primary endpoint. Suppose that the failure times in the standard and new treatment arms are exponentially distributed, with a rate of progression in the standard treatment arm of 50% per year and, under the alternative, a median time to progression of 4 years in the new treatment arm. Suppose there are 10 progression events observed during the study and the two-sided significance level is 0.05.

a) What is the power of the study to detect a difference in hazards between the two arms?

b) If the desired power is 0.80, how many events do investigators need to observe?

Problem 2 Solution.

a) As shown in the lecture slides,

$$d = \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log(\theta))^2}$$
$$\Rightarrow z_{1-\beta} = \sqrt{\frac{d\log(\theta)^2}{4}} - z_{1-\alpha/2}$$

For this problem,

$$\left.\begin{array}{l}\tau_c = 1\\ \tau_e = 4\end{array}\right\} \Rightarrow \theta = \frac{\tau_c}{\tau_e} = \frac{1}{4} = 0.25$$

$$z_{1-\beta} = \sqrt{\frac{10\log(0.25)^2}{4}} - 1.96 = 0.232$$
$$\Longrightarrow \text{Power} = P(Z \leq z_{1-\beta}) = 0.592$$

b) If the desired power is 0.80, investigators need to observe d events, with

$$\begin{aligned} d &= \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log(\theta))^2} \\ &= \frac{4(1.96 + 0.842)^2}{(\log(0.25))^2} \\ &\approx 17 \end{aligned}$$

Problem 3: Reproducing the power calculations for the SPRINT trial

Here is the quote at the beginning of the sample size slides from the SPRINT paper describing the sample size calculation for the trial.

> We planned a 2-year recruitment period, with a maximum follow-up of 6 years, and anticipated a loss to follow-up of 2% per year. With an enrollment target of 9,250 participants, we estimated that the trial would have 88.7% power to detect a 20% effect with respect to the primary outcome, assuming an event rate of 2.2% in the standard-treatment group.

a) Use `nSurv()` in the package `gsDesign` to reproduce the design of the SPRINT trial. You will have to spend a bit of time reading the help page for `nSurv` and scanning the published manuscript for the trial. The link to the manuscript is in the Unit 6 lecture slides.

b) How many events were anticipated to provide the 88.7% power?

c) How much power would the study have had if there had been only a 15% reduction expected in the rate of the primary endpoint?

d) How can the power of the study from part c) be increased to 90% if the study size does not change (i.e., the enrollment rate and enrollment period does not change)?

Problem 3 Solution.

a) Like many design summaries in published papers, the terms used are not always clearly defined.

- The primary outcome in the SPRINT trial was the time to a composite outcome consisting of myocardial infarction, other acute coronary syndromes, stroke, heart failure, or death from cardiovascular causes.
- The event rate of 2.2% in the control group is probably a constant hazard rate in the control group of 0.022 per year. This is `lambdaC`.
- A 20% "effect with respect to the primary outcome" means a 20% reduction, which implies a hazard ratio of 0.80 (intervention hazard / control hazard) .
- The accrual period R is two years.
- The accrual rate per year is 4,625 patients.
- The "anticipated loss to follow-up" is the anticipated attrition rate, `eta` and `etaE` in both arms.
- The value of α is found in the methods section of the paper, and is $\alpha = 0.05$. `nSurv()` is set up by default to calculate power for one-sided tests, so the solution uses a one-sided $\alpha = 0.025$.

There are several ways to confirm the design as stated in the paper using `nSurv()`. The function returns the value of whatever key parameter has been omitted.

The first solution below sets power as 88.7% and returns the study duration. That likely reflects how the study team would have done the calculation, although the team probably looked at many alternative designs before settling on this one. This solution leaves the

minimum follow-up `minfup` unspecified and returns a value of approximately 4 years (3.89). The minimum follow-up is the length of time between the enrollment of the last patient and the analysis time.

Another way to confirm the design is to solve for the power by specifying the study duration (i.e., "maximum follow-up", `T`) and the minimum follow-up duration ($6 - 4 = 2$ years) and setting `beta = NULL`. This solution results in power of 89.36%.

The two methods are approximations and will not show identical answers, but they should be close.

```
library(gsDesign)
```

```
## Loading required package: xtable
```

```
## Loading required package: ggplot2
```

```
#solving for follow-up duration
nSurv(lambdaC = 0.022, hr = 0.80, hr0 = 1.0,
      eta = 0.02, etaE = 0.02,
      R = 2, gamma = 4625,
      sided = 1, alpha = 0.025, beta = 1 - 0.887)
```

```
## Fixed design, two-arm trial with time-to-event
## outcome (Lachin and Foulkes, 1986).
## Solving for:  Follow-up duration
## Hazard ratio                  H1/H0=0.8/1
## Study duration:                   T=5.8796
## Accrual duration:                   2
## Min. end-of-study follow-up: minfup=3.8796
## Expected events (total, H1):        810.8162
## Expected sample size (total):       9250
## Accrual rates:
##     Stratum 1
## 0-2      4625
## Control event rates (H1):
##       Stratum 1
## 0-Inf     0.022
## Censoring rates:
##       Stratum 1
## 0-Inf      0.02
## Power:                 100*(1-beta)=88.7%
## Type I error (1-sided):   100*alpha=2.5%
## Equal randomization:          ratio=1
```

```
#solving for power
nSurv(lambdaC = 0.022, hr = 0.80, hr0 = 1.0,
      eta = 0.02, etaE = 0.02,
      R = 2, gamma = 4625,
      sided = 1, alpha = 0.025, beta = NULL,
      T = 6, minfup = 4)
```

```
## Fixed design, two-arm trial with time-to-event
## outcome (Lachin and Foulkes, 1986).
## Solving for:  Power
## Hazard ratio                   H1/H0=0.8/1
## Study duration:                    T=6
## Accrual duration:                    2
## Min. end-of-study follow-up: minfup=4
## Expected events (total, H1):        828.913
## Expected sample size (total):       9250
## Accrual rates:
##     Stratum 1
## 0-2      4625
## Control event rates (H1):
##       Stratum 1
## 0-Inf     0.022
## Censoring rates:
##       Stratum 1
## 0-Inf      0.02
## Power:                  100*(1-beta)=89.3584%
## Type I error (1-sided):   100*alpha=2.5%
## Equal randomization:          ratio=1
```

b) The two methods show between 810-830 events.

c) Power drops to aproximately 66% if the hazard ratio is 0.85 and no other parameters are changed. Power calculations are very sensitive to assumptions about the hazard ratio, or effect size.

```
library(gsDesign)

#solving for power, with HR 0.85
nSurv(lambdaC = 0.022, hr = 0.85, hr0 = 1.0,
      eta = 0.02, etaE = 0.02,
      R = 2, gamma = 4625,
      sided = 1, alpha = 0.025, beta = NULL,
      T = 6.0, minfup = 4.0)
```

```
## Fixed design, two-arm trial with time-to-event
## outcome (Lachin and Foulkes, 1986).
## Solving for:  Power
## Hazard ratio                   H1/H0=0.85/1
## Study duration:                    T=6
## Accrual duration:                    2
## Min. end-of-study follow-up: minfup=4
## Expected events (total, H1):        851.0172
## Expected sample size (total):       9250
## Accrual rates:
##     Stratum 1
## 0-2      4625
```

```
## Control event rates (H1):
##       Stratum 1
## 0-Inf     0.022
## Censoring rates:
##       Stratum 1
## 0-Inf      0.02
## Power:                  100*(1-beta)=65.8965%
## Type I error (1-sided):   100*alpha=2.5%
## Equal randomization:          ratio=1
```

d) Increasing the follow-up time increases the number of events that will be observed in the same study population. Setting `minfup = NULL` yields a minimum follow-up of 9.4 years and extends the study length to 11.4 years.

```
library(gsDesign)

#solving for minfup, with HR 0.85
nSurv(lambdaC = 0.022, hr = 0.85, hr0 = 1.0,
      eta = 0.02, etaE = 0.02,
      R = 2, gamma = 4625,
      sided = 1, alpha = 0.025, beta = 1 - 0.90,
      T = 6.0, minfup = NULL)
```

```
## Fixed design, two-arm trial with time-to-event
## outcome (Lachin and Foulkes, 1986).
## Solving for:  Follow-up duration
## Hazard ratio                   H1/H0=0.85/1
## Study duration:                    T=11.3763
## Accrual duration:                    2
## Min. end-of-study follow-up: minfup=9.3763
## Expected events (total, H1):        1594.084
## Expected sample size (total):       9250
## Accrual rates:
##     Stratum 1
## 0-2      4625
## Control event rates (H1):
##       Stratum 1
## 0-Inf     0.022
## Censoring rates:
##       Stratum 1
## 0-Inf      0.02
## Power:                  100*(1-beta)=89.9999%
## Type I error (1-sided):   100*alpha=2.5%
## Equal randomization:          ratio=1
```

Problem 4: Power in observational studies

Some investigators in the environmental health department want to conduct a study to assess the effects of exposure to toluene on time to pregnancy. They will conduct a cohort study involving women who work in a chemical factory in China.

It is estimated that 20% of the women will have workplace exposure to toluene. Furthermore, it is known that among unexposed women, 80% will become pregnant within a year. Suppose that among exposed women, 70% will become pregnant within a year.

The investigators will be able to enroll 200 women per year into the study and plan an additional year of follow-up at the end of accrual. Assuming 2 years of accrual, are there enough recruited women to be able to detect a reduction in the 1-year pregnancy rate with 85% power?

Problem 4 Solution.

Begin by calculating the required number of events:

- $S_c(1) = 1 - 0.80 = 0.2$
- $S_e(1) = 1 - 0.70 = 0.3$

$$\theta = \frac{\log(S_e)}{\log(S_c)} = \frac{-1.20}{-1.61} = 0.748$$

$$\begin{aligned} d &= \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{(\log(\theta))^2} \\ &= \frac{4(1.96 + 1.036)^2}{(\log(0.748))^2} \\ &\approx 427 \end{aligned}$$

If 200 women are enrolled per year during the 2-year accrual period, then a total of only 400 women will be enrolled in the study. Even if all women were followed until pregnancy (no censoring), there would not be enough recruited women to be able to detect a reduction in the 1-year pregnancy rate with 85% power.

How does the fact that an estimated 20% of the women will have workplace exposure to toluene come into play? It does not, in the question posed. The power depends on d, the number of required events, which is unaffected by the proportion exposed. If the question had asked for the sample size N to obtain d, then the proportion exposed would be part of the calculations.