Oph dataset (Time-Fixed Example)

Teeranan Pokaprakarn; Hiroto Udagawa; Danyi Xiong; Nusrat Rabbee (rabbee@berkeley.edu)

Department of Statistics, University of California, Berkeley, CA

In this section we will look at the Ophthalmology dataset (Oph) from the powerSurveEpi package [1]. This data is from a clinical trial to test the efficacy of different vitamin supplements in preventing visual loss. It has a time-fixed exposure with no transitions with N = 356 and a duration from 1 to 6 years. Thus it is a very simple case to test and validate our results from simulations using Method 1 in our package.

To set our parameters for our function, we retrieve the following results:

```
#============== Estimate Paramters =================

censor_time = median(Oph$times[Oph$status==0])
cox = coxph(Surv(time = times,event = status)~factor(group), data=Oph)
beta = summary(cox)$coef[,"coef"]

exp = nrow(Oph[Oph$group=='E',])/nrow(Oph)
fit = survfit((Surv(time = times,event = status)~group) ,dat=Oph)
```

Table 1: Estimates from Oph time-fixed dataset

Effective exposure proportion	Beta	Median control survival time	Median time to censor
0.4858757	-0.3734941	6	5

- The effective exposure proportion is simply the proportion of our subjects who are in the exposed group (estimated directly from the Oph data).
- The estimate for the beta value is calculated by fitting our sample data to a coxph regression model and getting its estimate (estimated from coxph() from survival package).
- The median control survival time is calculated by fitting a survival curve to the data and extracting the estimated median survival time (estimated from survfit() from survival package).
- The median time to censor was generated by taking the simple median times for all of the censored patients (estimated directly from the Oph data).

We now run our simulation of data generation using Method 1 (see above) with the same number of subjects as Oph data and check the results. We compare and diagnose our simulation results in the following steps using two ways.

Step 1: Compare the simulation results with the real data:

Most importantly, we check the number of control events (d_c), exposed events (d_exp), and power (pow) in our simulation output with the real data.

```
#Run power simulation on 450 subjects to get an N_effective value close to 354

power1 = getpower.method1(500, 450, duration = 6, med.TTE.Control = 6,
                          rho = 1, med.TimeToCensor = 5, b=beta, exp.prop=exp,
```

```
                    type='fixed', prop.fullexp = full_exp,
                    scenario = 'fixed(500)',maxrelexptime = 1/6,
                    min.futime = 1, min.postexp.futime = 1)
```

Table 2: Simulation results vs Data

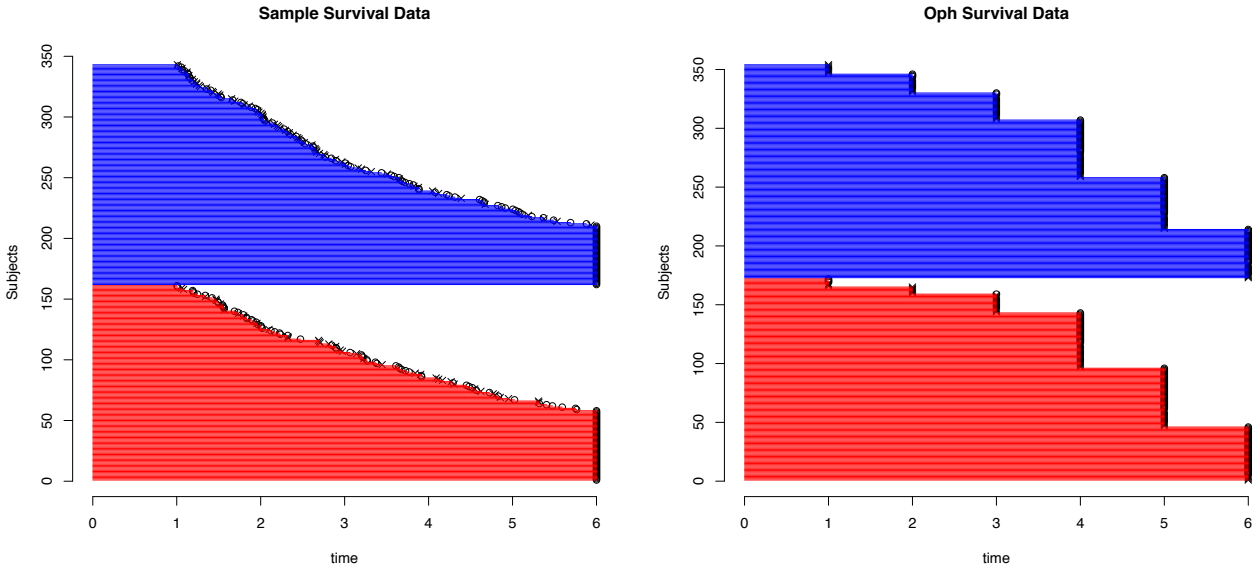
Parameter	Simulation Results	Oph Data Event Count
N	450.00	NA
N Effective	354.54	354
Total Events	100.62	154
Control Events	58.21	89
Exposed Events	42.40	65
Power	0.43	0.638*

*results based on analytical formula, $d = \frac{(z_{1-\frac{\alpha}{2}} + z_1 - \beta)^2}{\pi(1-\pi)\theta_R^2}$, where d is the number of required deaths,π is the proportion in group 1, and θ is the hazards ratio

Note that in our simulation, we input an N value (number of subjects to be scanned), but only get an output of N_eff (number of effective) subjects. This is because we pass our subjects through a filter to ensure a minimum follow-up time is met. This simulates similar practices when subjects are sampled for real studies. In our case we match the N_eff count with the real Oph data subjects for comparison.

We can now see that our simulation has too few events, which results results in the low power compared to the real data.

Step 2: We compare the incidence plots of our real data with the simulated data with the objective of comparing distribution of min(event,censored) over time in exposed vs control groups.



From the comparison of these plots, we make several observations.

a. The shapes of our plots are different. The Oph data has a more convex shape while our simulated data is more concave. The underlying distribution of event and censoring times is exponential under our simulation, which potentially differs from the real data.

b. Our data assumes continuous time while the real data is discrete by each year. This explains why the Oph data plot is collected in intervals.

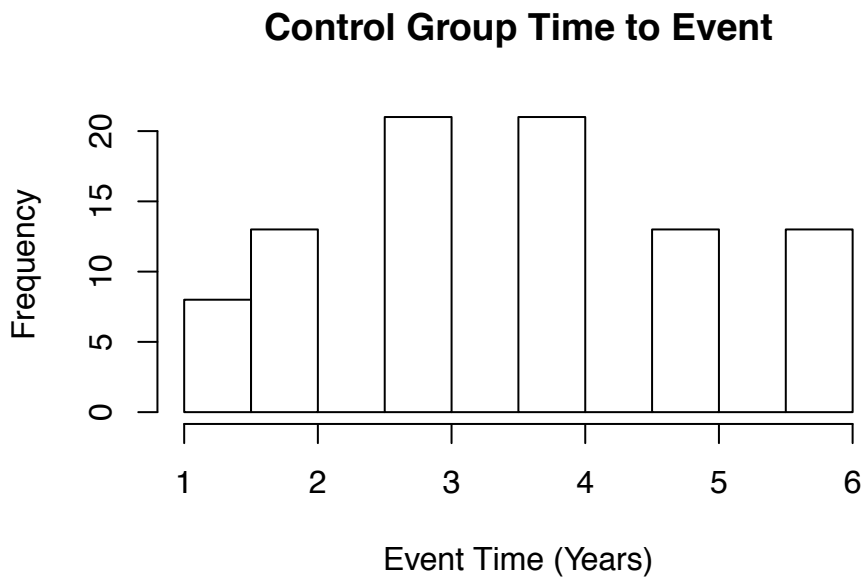
c. In our simulation, we experience relatively more censored events at time 6 when compared to the real data. In fact, our simulation is designed so that all remaining subjects at the end are censored. As seen in the table below, this is not the case with the real data where there are several events occurring at the very end.

Table 3: Oph Data Group vs Event Status (at time=6)

	Control	Exposed
0 (no event)	29	41
1 (event)	13	5

The differences are due to intrinsic assumptions of our simulation engine. We attempt to increase the number of events by making a better estimate of the median time to event for control group.

We infer that the median control event time from the survival object of the Oph data set was high; the length of the study was six years while our median control event time was also six years. We adjust our estimate of the median control event time by plotting the histogram of event times in the control group.



Based on the histogram, our previous estimated median control group time to event of 6 seems to be an overestimate. So we alter this parameter to 3.5 for our next simulation step.

We realize that this estimate based on the histogram may be an underestimate. Simply looking at this histogram does not take into account the censored subjects who may have longer event times. However, our approach is to continue to tweak these parameters until we get simulation results that reflect the parameters of the real data set (Oph).

Our new simulation plot as well as simulation results are shown below.

```
power2 = getpower.method1(1500, 480, duration = 6, med.TTE.Control = 3.5,
                          rho = 1, med.TimeToCensor = 5, b=beta, exp.prop=exp,
                          type='fixed', prop.fullexp = full_exp,
```

```
                        scenario = 'fixed(500)',maxrelexptime = 1/6, min.futime = 1,
                        min.postexp.futime = 1, simu.plot=T)
```

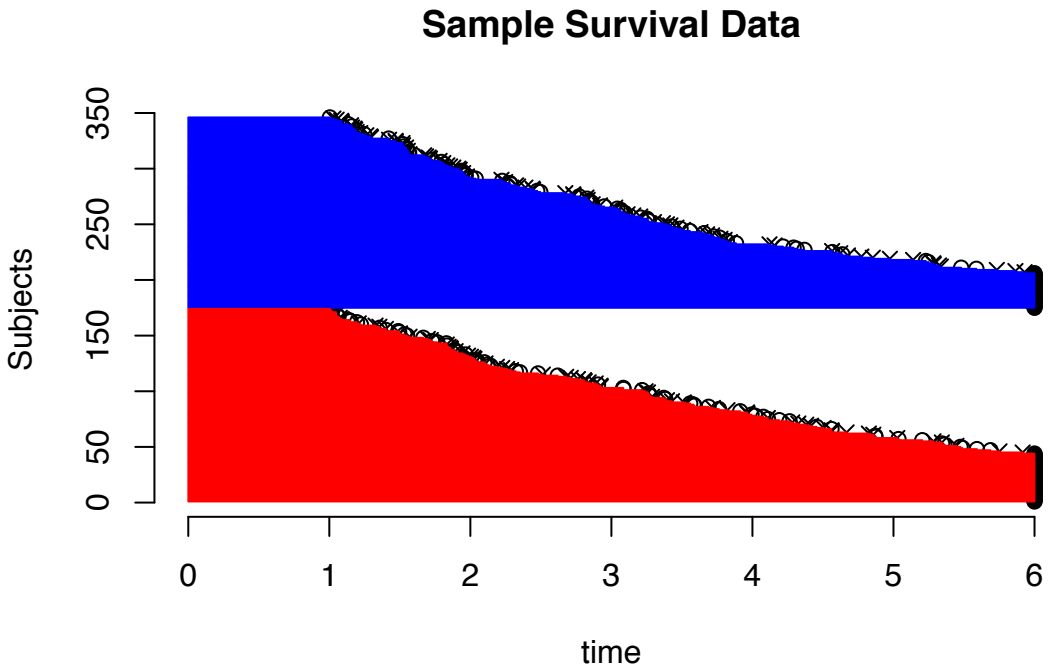


Table 4: Simulation results vs Data

Parameter	Simulation Results	Oph Data Event Count
N	480.00	NA
N Effective	353.15	354.00
Total Events	149.71	154.00
Control Events	84.31	89.00
Exposed Events	65.40	65.00
Power	0.64	0.64

As we can see, the results of our second simulation very closely mimics our sample data set, Oph. We can now use these parameters to run other simulations of this data and form more power estimations.

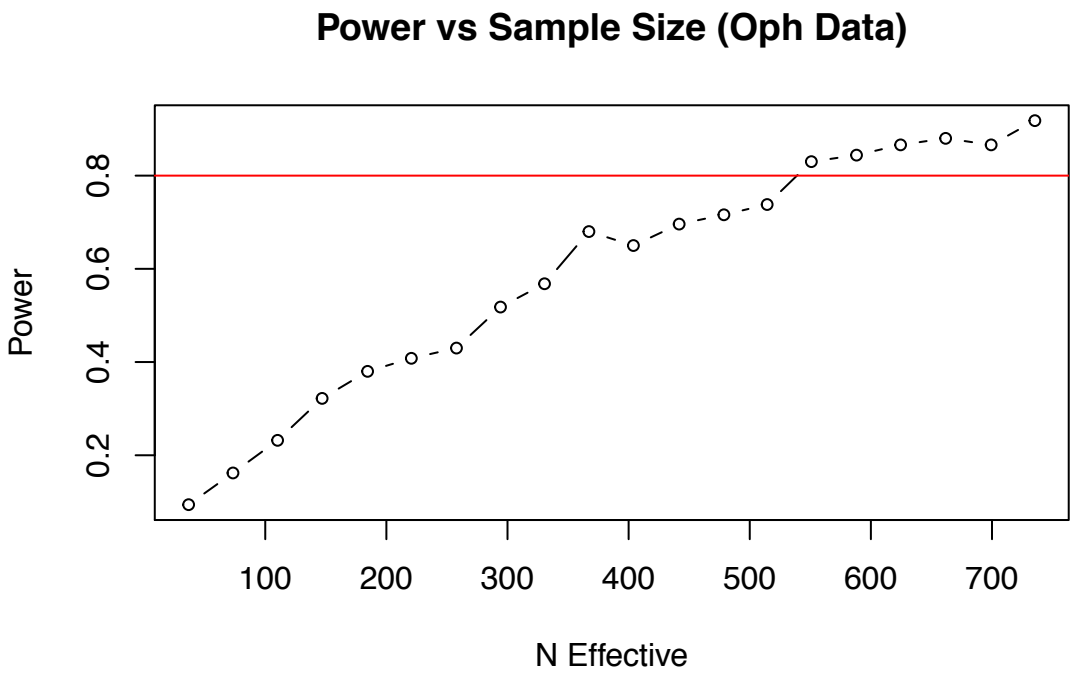
Again note that we simulate on 480 subjects in order to get an N_eff of 353. This N value is slightly higher than our original simulation, where we used N=450, since we have reduced the median control group time to event, and thus more subjects fail to pass the filter of minimum follow-up time.

This is an important characteristic of our simulation framework: the user will need to vary total # of initial subjects (N) in order to get an effective sample size (N_eff) that match the desired number of events of the data set they want to emulate (e.g., Oph).

Finally, one may be interested in the number of subjects needed to achieve a power of 0.80. This can be easily measured using several simulations of our data.

```
for(i in seq(50,1000,50)){
    power1 = getpower.method1(500, i, duration = 6, med.TTE.Control = 3.5,
                              rho = 1, med.TimeToCensor = 5, b=beta,
                              exp.prop=exp, type='fixed',
                              prop.fullexp = full_exp, scenario = 'fixed(500)'
```

```
                                ,maxrelexptime = 1/6, min.futime = 1,
                                min.postexp.futime = 1, simu.plot=F)
}
```



As we can see based on this plot, we would need a effective sample size of around 525 in order to reach the desired power.

References

[1] Qiu, W. (2015) Power and Sample Size Calculation for Survival Analysis of Epidemiological Studies (0.0.9) [software]