Exercise 4.2: For R software

1. For this exercise we will refer to the CASS data that was analysed in Reilly AJE 1996. We are interested in estimating the association between left ventricular end diastolic blood pressure (LVEDP) and operative mortality.

Let's suppose that the binary outcome (operative mortality) and a number of demographic and clinical covariates (age, sex, weight, surgery, angina, chf) are available for the full cohort (N=8088 patients), but that we have to retrieve data on LVEDP by clinical charts. We expect a High LVEDP in 23% (LVEDP>17 mmHg) of the subjects and an effect size (OR) nearly 1.65.

From the CASS cohort we know that:

- the proportion of females is 16%
- the hazard ratio on operative mortality for females compared to males is approximately 2.19

i. Due to a budget limitation we need to sample a subgroup of subjects for measuring LVEDP (n=400 subjects). Use the **PowerIIPhase()** function to find the power for different designs: a simple random sample (SRS), a random case-control (CC) sample and a CC sample stratified on sex. Compare the results. Is there a gain in power by stratifying for sex?

   Be aware that computation time might be long. For this reason, we suggest to limit number of simulations to 150 (**B**=150). Please note that in order to simulate CASS data we need to specify the following parameters (**tau** = 1, **lambda** = 0.02; **cens** = 0, see definition below at **Hints**).

ii. Estimate the power of each design mentioned in (i) varying the subgroup sample size from 300 to 400patients. Show the power curves using the function **PlotPower()**.

iii. Suppose that we have an auxiliary variable of LVEDP in the full cohort (note: auxiliary variable is a variable associated with the marker of interest) with sensitivity and specificity of 0.80 and 0.80 respectively (where sensitivity is defined as Prob(aux=high|LVEDP=high) and specificity is Prob(aux=low|LVEDP=low)). Compute the power for a case-control design stratified for an auxiliary variable and compare it with the one obtained in (i).

iv. (optional) The full cohort data is in **chap8_CASS.dta**. Perform the following steps:
   (a) Create a categorical variable for LVEDP considering high level with LVEDP(lve) > 17 mmHg.
   (b) Randomly select a case-control subsample of 400 subjects.
   (c) Run weighted logistic regression analysis using survey package to obtain the crude OR for dichotomized LVEDP from the case-control subsample in step (b).
   (d) (advanced) Reproduce step (b) and (c) 100 times and count the number of subsamples that achieved a significant effect of LVEDP (type I error of 0.05). Compare the power with the results in part (i).

   NOTE: Your sample size may be less than 400 in the final analysis because some cohort members are missing LVEDP!

Hints

Use the design2phase package implemented in R software. The package can be installed from the GitHub repository as follows:

```
> devtools::install_github("Fgraziano/design2phase")
```

Packages required are survival (version ≥3.2-7), survey, ggplot2, multipleNCC, reshape2, caret and Epi.

Please see the definition of the parameters of the `PowerIIphase function`:

`pBM` expected prevalence of the biomarker of interest in the cohort (phase I).

`betaBM` expected beta coefficient (ln(HR)) of the new biomarker of interest on the time-to-event endpoint.

`pstrata` prevalence of the stratum variable to be used in the design. Default is NULL for not stratified sampling.

`betastrata` expected beta coefficient (ln(HR)) of the stratum to be used in the design on the time-to-event endpoint. Default is NULL for not stratified sampling.

`acc.aux` vector of the expected sensibility and specificity (accuracy) of the auxiliary with respect to the biomarker (default NULL for no auxiliary variable used in the design). See the details for the definition.

`design2p` a list of extra sampling designs performed during the sampling process, with SRS and Case-Control (CC) fitted by default. Optional choice are "PPS", "NCC" and "CM". "CM" is possible only if auxiliary variable is available. If stratum or auxiliary variable are available, CC and PPS stratified by these variables are automatically performed. See the details for the definition.

`N` a number providing the sample size of the full cohort (phase I).

`n` a vector providing the sample size of the subsample (phase II).

`cens` rate of censoring from exponential distribution. Parameter useful for simulating censoring in the cohort (phase I). Default is a rate=0.1.

`tau` maximum follow-up time in the cohort. Parameter useful for simulating time-to-event in the cohort (phase I). Defa ult is 2.

`lambda` scale parameter of Weibull baseline hazard of event. Parameter used for simulating time-to-event in the cohort (phase I). Default is 0.1.

`k` shape parameter of Weibull baseline hazard of event. Parameter used for simulating time-to-event in the cohort (phase I). Default is 0.9.

`B` number of datasets generated by the simulation. Default is 1000.

`seed` of the random number generation. Default is NULL.

For example, to answer (i) above:

```
ex1<-PowerIIphase(pBM =0.23, betaBM = 0.50,pstrata = 0.16,
                  betastrata =  0.78, cens=0,tau=1,
N=8088,lambda=0.02, n=400, seed=1234, B=150)
```