YZ_BMC_PARTII - Coding In R - INPUT PART

Yueran Zhang(yz4188@cumc.columbia.edu)

2022-10-05

Assume we are going to study effects of certain treatment strategies on a specific population. The first step would be to create the population of interest (cohort). Using R, create a .csv file representing the cohort. In this file, each row should represent a person and each column should represent one of the person's attributes.

Each person should have the following attributes:

1. ID (from 1 to population size)
2. Age in months (an integer)
3. Sex ("male" or "female")
4. Injection drug use status or idu ("current", former", "none")
5. Seropositivity status: 0 or 1 where 1 represents being seropositive
6. Infection status: 0 or 1 where 1 represent having active infection

- seropositivity means the person has been exposed to the virus at some time in his/her life so he/she would have virus antibodies in his/her blood. Being seropositive does not necessarily mean having active infection for all diseases (such as Hepatitis C). So a person might be seropositive but uninfected while the reverse is not true - an infected person cannot be seronegative.

Create an input file (Rscript) that includes general statistics of cohort: 1. Population size = 10k

2. Age is a normal distribution with mean of 18 (years) and standard deviation of 5 (years). Convert years to months before drawing. If a drawn value is less than minimum age in months or greater than maximum age in months, replace it with minimum/maximum age. Minimum age = 10 (years), maximum age = 40 (years)
3. Probability of being seropositive among male and female population is 0.3 and 0.2 respectively.
4. Probability of having active infection given being seropositive is 0.8 (conditional probability).
5. Probability of being "current", "former" and "none" idu is 0.2, 0.3 and 0.5 respectively.

Create a separate Rscript that reads this input file and creates the corresponding cohort .csv file. This Rscript should not have any direct user input in it. In the other words, you should be able to change the values inside the input file and get a new cohort output without changing anything inside the code file.

INPUT PART - CREAT THE POPULATION COHORT/DATA SIMLUALTION

```
df <- data.frame(matrix(ncol = 6, nrow = 0))
colnames(df) <- c('IDs', 'age', 'sex', 'idu', 'seropos','inf')
ages <- floor(rnorm(10000, 18*12, 5*12))
for (id in 1:10000){
  id
  age <- max(120,min(480,ages[id]))
  sex <- sample(c('male', 'female'), 1)
  if (sex == 'male'){
    seropos <- sample(c(0, 1),1,TRUE,c(0.7,0.3))
  } else {
    seropos <- sample(c(0, 1),1,TRUE,c(0.8,0.2))
  }
  if (seropos == 1){
    inf <- sample(c(0, 1),1,TRUE,c(0.2,0.8))
  } else {
    inf <- 0
  }
  idu <- sample(c('current', 'former', 'none'),1,TRUE,c(0.2,0.3,0.5))
  df[id,] <- c(id,ages[id],sex,idu,seropos,inf)
}
write.csv(df,"Cohort.csv", row.names = FALSE)
```

Double Check in case:)

```
head(df,10)
```

```
##    IDs age    sex     idu seropos inf
## 1    1 247 female    none       0   0
## 2    2 268   male current       0   0
## 3    3 323 female  former       0   0
## 4    4 267 female  former       0   0
## 5    5 158   male  former       0   0
## 6    6 239   male    none       0   0
## 7    7 152   male current       1   1
## 8    8 175 female  former       0   0
## 9    9 224 female  former       1   1
## 10  10 231 female    none       0   0
```