Computing in Epidemiology and Biostatistics

Programming with functions

Wan-Yu Lin

Ex 6: Please rearrange the data in Seizure2 as the format in Seizure1. (Please answer this question with “for” loop)

R built-in functions: var, lm, glm, etc.

User-defined functions

Demo 1: Please calculate the mean and variance of y in Seizure1.

```
setwd('D:/Comp/1')
Seizure1 <- read.csv('seizure.csv')
y.mean <- mean(Seizure1$y)
```

$$\mathrm{var}(Y) = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}$$

```
sum((Seizure1$y - y.mean)^2)/(length(Seizure1$y) - 1)      #    sum : summation

var(Seizure1$y)     # R built-in command to calculate variance
```

Missing data in R

```
x <- NA
is.na(x)        # Is it a missing value?


x <- c(3600, 5000, 12000, NA, 1000, 2000, 600, 7500, 1800, 9000)    # Money in red envelopes
is.na(x)
mean(x)
mean(x, na.rm=T)             # NAs can be removed
var(x, na.rm=T)
sd(x, na.rm=T)               # standard deviation
sqrt(var(x, na.rm=T))        # square root
var(x, na.rm=T)^0.5
```

Logical expressions in R

Only two possible outcomes: TRUE (1) or FALSE (0)

A logical expression is formed using the comparison operators
<, >, <=, >=,
== (equal to)
!= (not equal to)
& (and)
| (or)   # holding down the shift key and press \
! (not)

Note that A|B is TRUE if A or B or both are TRUE.
If you want exclusive disjunction, that is either A or B is TRUE but not both, then use xor(A,B).

Ex: Mom told John: “if you get A+ in this semester, I will buy you a toy or a comic book.”

Ex 7: Please use R code to answer the following questions related to “Money in red envelopes”:

(1) How many students received more than 5,000 dollars?
(2) No. 3 and no. 8 both received more than 5,000 dollars?
(3) No. 3 or no. 8 received more than 5,000 dollars?
(4) Only one of no. 3 or no. 8 received more than 5,000 dollars?
(5) In (1)~(4), please revise the money cutoff to be 6,000 and 10,000, respectively. And answer (1)~(4).
(6) In (1)~(4), please revise the money cutoff to be 1000, 2000, 3000,…,20000, respectively. And answer (1)~(4).

Sampling in R

sample(1:N, size=n)   # Sampling n numbers from integers 1 to N, default: sampling without replacement

sample(1:39, size=5)   # Sampling 5 numbers from integers 1 to 39, default: sampling without replacement

sample(1:200, size=10)   # Sampling 10 students from 200 students, every student has at most one chance to be called

```
sample(1:200, size=201)

sample(1:200, size=201, replace=T)      # sampled balls will be put into the box again, each ball has more than one chance to be sampled
```

Replace: Should sampling be with replacement?

Ex: Instructor would like to sample 12 students to answer questions, from 50 students. Students sampled may be sampled again (very lucky). Please provide R code.

Set operations in R

```
(x <- c(sort(sample(1:20, 9)), NA))
(y <- c(sort(sample(3:23, 7)), NA))
union(x, y)
intersect(x, y)
setdiff(x, y)
setdiff(y, x)
setequal(x, y)

## True for all possible x & y :
setequal( union(x, y), c(setdiff(x, y), intersect(x, y), setdiff(y, x)) )

is.element(x, y)              # x %in% y
is.element(y, x)              # y %in% x
```

Ex 8: There were 50 students in a class. All odd-numbered students were boys, and all even-numbered students were girls.

There were 7 students passing the midterm exam: 11,16,23,31,36,47,50.

There were 9 students passing the final exam: 3,9,16,20,27,31,36,49,50.

Please use the R commands regarding set operations to answer the following four questions:

(1) Please list the boys who passed both the midterm and final exams

(2) Please list the girls who passed both the midterm and final exams

(3) Please list the boys who passed the midterm exam but failed the final exam

(4) Please list the girls who failed the midterm exam but passed the final exam

User-defined functions

Demo 2: Please write a function to demonstrate the central limit theorem. Let the function input be the sample size.

[Central limit theorem, C.L.T.] Let the population mean be μ, standard deviation be σ. The distribution of the sample mean $\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$, where *n* is the sample size.

```
par(mfrow = c(3,2))
CLT <- function(n){
   N <- 10000
   score <- rnorm(N,75,5)
   hist(score, xlim=c(50,100), breaks=seq(50,100,1))
   mean.score <- c()
   for(i in 1:1000){
      mean.score[i] <- mean(score[sample(1:N, size=n)])
   }
   hist(mean.score, xlim=c(50,100), breaks=seq(50,100,1))
   return( c(mean(mean.score), var(mean.score)) )
}

CLT(n=5)
CLT(n=10)
CLT(n=20)
```

According to C.L.T., $\frac{\sigma^2}{n}$ should be

25/5
25/10
25/20

Ex: What if the population comes from a chi-square distribution with degrees of freedom 1? Please revise the above R function.

Ex 9: **Please write a function on your own** to estimate the regression coefficients of a simple linear regression, where the response variable is “y” from “seizure.csv” (Week 1 course material) and the predictor variable is “ltime” from “seizure.csv”. (Hint: the regression coefficients should include an intercept term and a slope term)