Computing in Epidemiology and Biostatistics

Modern statistical computing in R: Monte-Carlo simulations (Interval estimates)

Wan-Yu Lin

Simulation is a process of generating random numbers and performing the experiments you want.

```
set.seed(7)
rnorm(10)
```

If you do not specify any seed number, R will randomly assign a seed number for you.

Use chi-square distribution as an example	
dchisq(x, df)	probability density function, p.d.f.
pchisq(q, df)	cumulative distribution function, c.d.f.
qchisq(p, df)	quantile function
rchisq(n, df)	generate random numbers

These functions provide information about the chi-square distribution with **df** degrees of freedom. **dchisq** gives the density, **pchisq** gives the distribution function, **qchisq** gives the quantile function, and **rchisq** generates random variables.

Ex：confidence interval (C.I.)

If people on average sleep 7 hours every day, with a standard deviation of 2 hours. If we randomly sample 100 subjects, please check the coverage of 95% C.I. and 99% C.I. for mean μ. Number of replications = 1000, and seed numbers from 1 to 1000, respectively.

(Scenario 1) The population standard deviation (σ) is known

The $(1-\alpha)\times 100\%$ C.I. for $\bar{X}$ is $\left[\bar{X} \mp z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right]$.

```
mu <- 7; sigma <- 2; n <- 100; no.rep <- 1000
l95 <- rep(NA,no.rep)    # lower bound
r95 <- rep(NA,no.rep)    # upper bound
l99 <- rep(NA,no.rep)
r99 <- rep(NA,no.rep)
for(i in 1:no.rep){
  #print(i)     # to see the progress of this simulation
  set.seed(i)
```

```
  x <- rnorm(n,mu,sigma)
# above: data generation process
# below: data analysis process
  l95[i] <- mean(x)-qnorm(0.975)*sqrt(sigma^2/n)
  r95[i] <- mean(x)+qnorm(0.975)*sqrt(sigma^2/n)
  l99[i] <- mean(x)-qnorm(0.995)*sqrt(sigma^2/n)
  r99[i] <- mean(x)+qnorm(0.995)*sqrt(sigma^2/n)
}
mean((l95<=mu) & (mu<=r95))     # check the coverage probability
mean((l99<=mu) & (mu<=r99))
```

(Scenario 2) The population standard deviation (σ) is unknown

The $(1-\alpha)\times 100\%$ C.I. for $\bar{X}$ is $\left[\bar{X} \mp t_{1-\alpha/2;n-1}\sqrt{\frac{s^2}{n}}\right] = \left[\bar{X} \mp t_{1-\alpha/2;n-1}\frac{s}{\sqrt{n}}\right]$.

```
lt95 <- rep(NA,no.rep)
rt95 <- rep(NA,no.rep)
lt99 <- rep(NA,no.rep)
rt99 <- rep(NA,no.rep)
for(i in 1:no.rep){
  #print(i)
  set.seed(i)
  x <- rnorm(n,mu,sigma)

  lt95[i] <- mean(x)-qt(0.975,n-1)*sqrt(var(x)/n)
  rt95[i] <- mean(x)+qt(0.975,n-1)*sqrt(var(x)/n)
  lt99[i] <- mean(x)-qt(0.995,n-1)*sqrt(var(x)/n)
  rt99[i] <- mean(x)+qt(0.995,n-1)*sqrt(var(x)/n)
}
mean((lt95<=mu) & (mu<=rt95))     # check the coverage probability
mean((lt99<=mu) & (mu<=rt99))

mean(r95-l95)     # the mean length of the C.I. in Scenario 1 (Z test)
mean(rt95-lt95)   # the mean length of the C.I. in Scenario 2 (t test)
mean(r99-l99)
mean(rt99-lt99)
```

Ex：Continuing the above example, if the population standard deviation (σ) is unknown, but a student computes the C.I. from a *Z* test, what would happen? Please check the coverage probability of 95% C.I. and 99% C.I. for μ while considering a sample size of 5, 10, 15, 20, …, 95, and 100, respectively. Number of replications = 1000. Please plot a figure with the x-axis as "sample size" and y-axis as "coverage probability".

Homework (8 points, please pay attention to all the words in this orange box)

Ex 21: Please generate random numbers from a Binomial distribution with *n*=20 (number of trials) and *p*=0.15 (probability of success), and compare the coverage and length of 95% asymptotic confidence intervals and 95% exact confidence intervals. Number of replications = 1000, and seed numbers from 1 to 1000, respectively.

Hint: rbinom(1,size=20,prob=0.15)

1 class, **20** students enrolled in this class, the probability of passing the final exam is **0.15**.

Requirements:

1. Please use the Newton-Raphson method to calculate exact C.I., but you don't need to present your plots when deciding initial values. (You may plot by yourself. To make this homework simpler, we will not score on them.)

2. Please calculate 95% asymptotic confidence interval based on $\hat{p} \pm z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (asymptotic: based on the central limit theorem). If the lower bound of a C.I. is less than 0, please regard it as 0, because the probability of success will always range from 0 to 1.

Please note: Using the "binom" package to answer this homework will be scored as 0, although you may use that package to check your own answers.