Homework Assignment 4

DS4043, Spring 2022

Due on April 19, 2022 at 11:59 pm

1. Consider random variables $X_1, \ldots, X_n$ are i.i.d. $\mathrm{N}\left(\mu = 30, \sigma^2 = 100\right)$, given $n = 50$ and $\alpha = 0.05$.

   a) Obtain the **Monte Carlo estimate of the confidence level** for the 95% confidence interval includes the true value of μ. Let the number of replicate as $m = 1000$. (Hint: you need to construct a 95% confidence interval of μ; the statistic is the sample mean.)

   Solution:

```
m = 1000; n = 50; mu <- 30
set.seed(15)
y <- numeric(m)
for(i in 1:1000){
  x <- rnorm(50,30,10)
  zq <- qnorm(0.975)  # quantile
  c.i <- c(mean(x)-zq*10/sqrt(n), mean(x)+zq*10/sqrt(n) ) # Confidence interval
  y[i] <- (mu> c.i[1]) & (mu< c.i[2])  # record whether \mu is in the interval
}
mean(y)
```

```
## [1] 0.959
```

   The Monte Carlo estimate of the confidence level is 0.959 close to 95%.

   b) For the hypotheses, $H_0 : \mu = 30$ vs $H_1 : \mu \neq 30$, use Monte Carlo method to compute an empirical probability of type-I error, and compare it with the true value. Let the number of replicate as $m = 10000$.

   Solution:

```
m = 10000; n = 50; mu <- 30
set.seed(13)
y <- numeric(m)
for(i in 1:10000){
  x <- rnorm(50,30,10)
  zq <- qnorm(0.975)
  c.i <- c(mean(x)-zq*10/sqrt(n), mean(x)+zq*10/sqrt(n) )
  y[i] <- (mu< c.i[1]) | (mu> c.i[2])
}
mean(y)
```

```
## [1] 0.0476
```

   The empirical probability of type-I error is 4.76% close to 5%.

2. Consider the random variables $X_1, \ldots, X_n$ are i.i.d. with a mixture normal density, i.e.

$$(1-p)N\left(\mu=0,\sigma^2=1\right)+pN\left(\mu=1,\sigma^2=9\right)$$

We have $\alpha = 0.05, p = 0.4$ and $n = 50$. Let β_1 denote the skewness of random variable X and its sample estimate is denoted by b_1. The hypotheses are $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Use the Monte Carlo method to estimate **empirical power** of the hypotheses. For finite samples one should use

$$\text{Var}\left(b_1\right) = \frac{6(n-2)}{(n+1)(n+3)}.$$

Let the number of replicate as $m = 10000$. To generate number from mixture density. Suppose $X_1 \sim N(0,1)$ and $X_2 \sim N(3,1)$ are independent. We can define a 50% normal mixture X, denoted $F_X(x) = 0.5F_{X_1}(x) + 0.5F_{X_2}(x)$. Unlike the convolution, the distribution of the mixture X is distinctly non-normal; it is bimodal. To simulate the mixture:

1. Generate an integer $k \in \{1, 2\}$, where $P(1) = P(2) = 0.5$.
2. If $k = 1$ deliver random x from N$(0, 1)$; if $k = 2$ deliver random x from N$(3, 1)$.

Solution:

```
set.seed(19)
## function to calculate b1 skewness
sk <- function(x) {
xbar <- mean(x)
m3 <- mean((x - xbar)^3)
m2 <- mean((x - xbar)^2)
return( m3 / m2^1.5 )
}

n=50;
m <- 10000;  #num. repl.
sktests <- numeric(m)  #test decisions
cv <- qnorm(.975, 0, sqrt(6*(n-2) / ((n+1)*(n+3))))

for (j in 1:m) {
  # generate data x
  for(i in 1:n){
  k <- sample(1:2,1,prob = c(0.6,0.4) )
  if(k ==1) x[i] = rnorm(1,0,1)
  if(k ==2) x[i] = rnorm(1,1,3)
  }
# test decision is 1 (reject) or 0
sktests[j] <- as.integer(abs(sk(x)) >= cv)
}
p.reject <- mean(sktests) # proportion rejected
print(p.reject)
```

```
## [1] 0.5053
```

Then the empirical power of the hypotheses is 0.5053.

3. Compute a jackknife estimate of the bias and the standard error of the correlation statistic in the *law* data example. Compare the result with the bootstrap method.

Solution:

```
library(bootstrap)
n <- nrow(law)
theta.jack <- numeric(n)
theta.b <- cor(law$LSAT,law$GPA)
for(i in 1:n){
  theta.jack[i] <- cor(law$LSAT[-i],law$GPA[-i])
}
bias <- (n - 1) * (mean(theta.jack) - theta.b)
se <- sqrt((n-1)*mean((theta.jack-mean(theta.jack))^2))
c(bias,se)
```

```
## [1] -0.006474  0.142519
```

The biases of jackknife and bootstrap are -0.006474 and -0.005910101. The standard errors of jackknife and bootstrap are 0.142519 and 0.1378404. (Note: students can use the bootstrap method and show the results or directly use the results in our lecture slides. The results of boostrap are not unique.)

4. Refer to the air-conditioning data set *aircondit* provided in the *boot* package. The 12 observations are the times in hours between failures of airconditioning equipment:

$$3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487.$$

Assume that the times between failures follow an exponential model Exp(λ). Obtain the MLE of the hazard rate λ and use bootstrap to estimate the bias and standard error of the estimate. Let the number of replicates as $m = 200$.

Solution: The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

Then the natural logarithm of $L(\theta)$ is

$$l(\theta) = \ln(L(\theta)) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i,$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Let $\frac{\partial l(\theta)}{\partial \theta} = 0$, we can get $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$.

```
library(boot)
set.seed(20)
x <- aircondit
B <- 200 #number of replicates
n <- nrow(aircondit) #sample size
B.lambda <- numeric(B) #storage for replicates
h.lambda <- n/sum(x)  # the estimator of the sample
#bootstrap method
for (b in 1:B) {
#randomly select the indices
i <- sample(1:n, size = n, replace = TRUE)
x <- aircondit[i,]
B.lambda[b] <- n/sum(x)
}
sd(B.lambda)
```

```
## [1] 0.003984
```

```
bias <- mean(B.lambda- h.lambda);
bias
```

```
## [1] 0.0007585
```

So the bias is 0.0007585 and standard error is 0.003984.