Homework Assignment 4

DS4043, Spring 2022

Due on April 20, 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)

ucl <- replicate(m, expr = {
  x <- rnorm(n, mean=mu, sd=10)
  mean(x) + 10 * qnorm(0.025)/sqrt(n)
})
lcl <- replicate(m, expr = {
  x <- rnorm(n, mean=mu, sd=10)
  mean(x) - 10 * qnorm(0.025)/sqrt(n)
})
(sum(lcl>30) - sum(ucl>30)) / m
```

```
## [1] 0.951
```

   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)
alpha <- 0.05
sigma <- 10
p <- numeric(m)
for (i in 1:m) {
  x <- rnorm(n, mu, sigma)
  ttest <- t.test(x, alternative = "greater", mu = mu)
  p[i] <- ttest$p.value
}
p_hat <- mean(p < alpha)
se_hat <- sqrt(p_hat * (1-p_hat)/m)
c(p_hat, se_hat)
```

```
## [1] 0.051900 0.002218
```

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

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

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 $\mathrm{N}(0,1)$; if $k=2$ deliver random x from $\mathrm{N}(3,1)$.

Solution:

```
set.seed(19)

n <- 50
sk <- function(x) {
  m3 <- mean((x-mean(x))^3)
  m2 <- mean((x-mean(x))^2)
  return (m3 /m2^1.5)
}

skewness.c1 <- function(n, cv){
  m <- 10000
  sktests <- numeric(m)
  # mixture of two distribution
  for (i in 1:m) {
    x1 <- rnorm(n, 0, 1)
    x2 <- rnorm(n, 1, 3)
    u <- runif(n)
    p <- as.integer(u>0.4)
    x <- p*x1 + (1-p)*x2
    sktests[i] <- as.integer(abs(sk(x)) >= cv)
  }
  p.reject <- mean(sktests)
  return(p.reject)
}
cv<-qnorm(0.975,0,sqrt(6*(n-2)/((n+1)*(n+3))))
print(skewness.c1(n, cv))
```

```
## [1] 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:

For jackknife estimate method:

```r
library(bootstrap)
n <- nrow(law)
theta_hat<-cor(law$LSAT,law$GPA)
theta_jack <- numeric(n)
# print(law$LSAT)
for (i in 1:n){
  theta_jack[i] <- cor(law$LSAT[-i], law$GPA[-i])
}
bias_jack <- (n - 1) * (mean(theta_jack) - theta_hat)
se_jack <- sqrt((n-1) * mean((theta_jack - mean(theta_jack))^2))
c(bias_jack, se_jack)
```

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

For bootstrap method:

```r
n <- 10000
r <- nrow(law)
theta_hat<-cor(law$LSAT,law$GPA)
theta_boot <- numeric(n)
for (i in 1:n){
  s <- sample(1:r, size=r, replace = TRUE)
  theta_boot[i] <- cor(law$LSAT[s], law$GPA[s])
}
bias_boot <- mean(theta_boot - theta_hat)
se_boot <- sd(theta_boot)
c(bias_boot, se_boot)
```

```
## [1] -0.004509  0.132941
```

<!-- solution end -->

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 $\text{Exp}(\lambda)$. 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:

Step1: $L(x_i, \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x} = \lambda^n * e^{-\lambda \sum_{i=1}^{n} x_i}$

Step2: $I = InL(x_i, \lambda) = n * \ln \lambda - \lambda \sum_{i=1}^{n} x_i$

Step3: $0 = \frac{\partial I}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$

MLE: Thus, $\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

```r
library(boot)
set.seed(20)
# bootstrap
m <- 200
data <- c(3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487)
n <- length(data)
lambda <- 1 / mean(data)
theta <- numeric(m)
for (i in 1:m){
  s <- sample(1:n, size=n, replace = TRUE)
  theta[i] <- 1 / mean(data[s])
}
bias <- mean(theta - lambda)
se <- sd(theta)
c(bias, se)
```

```
## [1] 0.0007585 0.0039845
```

<!-- solution end -->