DATA130004: Homework 5

董晴园 14300680173

2017.11.12

1. Prove that the k-level trimmed mean estimator has expectation zero when n random samples are independently generated from standard normal distribution.

   Proof: Generate i.i.d. n r.v. X_i from N(0,1). The k-level trimmed mean estimator is:

   $$\overline{X}_{[-k]} = \frac{1}{n-2k}\sum_{j=k+1}^{n-k} X_{(j)},$$

   where $X_{(j)}$ is the corresponding ordered sample.
   Since $X_{(1)}, ..., X_{(n)}$ are independent,

   $$\begin{aligned}\therefore \quad E\big(\overline{X}_{[-k]}\big) &= \frac{1}{n-2k}E\left\{\sum_{j=k+1}^{n-k} X_{(j)}\right\} \\ &= \frac{1}{n-2k}\sum_{j=k+1}^{n-k} E\left\{X_{(j)}\right\} \\ &= 0.\end{aligned}$$

2. Exercise 6.1: Estimate the MSE of the level k trimmed means for random samples of size 20 generated from a standard Cauchy distribution. (The target parameter θ is the center or median; the expected value does not exist.) Summarize the estimates of MSE in a table for $k = 1, 2, ..., 9$.

   Steps:
   For each k, first generate 20 random variables from a standard Cauchy distribution, and sort them in order: $X_{(1)}, ..., X_{(20)}$. Then compute

   $$\overline{X}^{(j)}_{[-k]} = \frac{1}{20-2k}\sum_{i=k+1}^{20-k} X_{(i)}, j = 1, ..., m.$$

   Replicate this process for m times. Finally, compute

   $$\widehat{MSE}\big(X_{[-k]}\big) = \frac{1}{m}\sum_{j=1}^{m}\left\{\overline{X}^{(j)}_{[-k]} - \overline{X}^{(j)}_{[-9]}\right\}^2.$$

   Using the following codes behind in R, we can summariize the estimates of MSE.

```
n <- 20
m <- 1000
k <- 1
mse <- numeric(9)
for (k in 1:9){
```

```
    tmean <- numeric(m)
    med <- numeric(m)
    for (i in 1:m) {
      x <- sort(rcauchy(n))
      tmean[i] <- sum(x[(k+1):(n-k)]) / (n-2*k)
      med[i] <- median(x)
    }
    mse[k] <- mean((tmean-med)^2)
    k <- k+1
  }
  mse
```

Here's the result:

k	1	2	3	4	5	6	7	8	9
$\widehat{MSE}(X_{[-k]})$	1.2642	0.1662	0.0714	0.0470	0.0292	0.0145	0.0089	0.0042	0.0

3. Exercise 6.4: Suppose that $X_1, ..., X_n$ are a random sample from a lognormal distribution with unknown parameters. Construct a 95% confidence interval for the parameter μ. Use a Monte Carlo method to obtain an empirical estimate of the confidence level.

   Since standard deviation is unknown, we use standard error to replace. In this case, 95% confidence interval for μ is:

$$\left(\hat{\mu} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

   Steps to estimate the confidence level with MC method:

   - Generate random variables from lognormal distribution with parameters (μ, σ^2)
   - Compute C.I. for μ with this sample and $y = I_{\{\mu \in C.I.\}}$ 如果不知道 μ 真实值怎么办?
   - Replicate this process for m times
   - Compute the empirical confidence level $\overline{y}$

   Now assume $\mu = 0$, we can use the following codes in R to get an empirical estimate of the confidence level:

```
set.seed(520)
n <- 20
alpha <- .05
m <- 1000
y <- numeric(m)
for (i in 1:m){
  x <- log(rlnorm(n))
  U.CI <- mean(x) + qnorm(1 - alpha/2) * sqrt(var(x)/n)
  L.CI <- mean(x) - qnorm(1 - alpha/2) * sqrt(var(x)/n)
  y[i] <- ifelse(0 <= U.CI & 0 >= L.CI, 1, 0)
}
mean(y)
```

```
[1] 0.941
```

4. In Example 6.4, to construct a $(1-\alpha)\times 100\%$ confidence interval for the variance parameter σ^2, we assume that the lower bound is 0 and the upper bound corresponds to a quantity involving the α-quantile of a χ^2 distribution, we now consider using $\alpha/2$ and $(1-\alpha/2)$-quantiles of the same χ^2 distribution to construct another confidence interval. It certainly will excludes 0.

   (a) Give the explicit form of the new confidence interval and justify its validity by showing the theoretical confidence level is $1-\alpha$.

   The new confidence interval:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

   Since $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$,

$$P\left\{\frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2}}(n-1)\right\} = \frac{\alpha}{2}$$

$$P\left\{\frac{(n-1)S^2}{\sigma^2} \geq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\} = \frac{\alpha}{2}$$

$$\therefore \quad P\left\{\chi^2_{\frac{\alpha}{2}}(n-1) \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\} = 1-\alpha$$

$$\Rightarrow \quad P\left\{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right\} = 1-\alpha.$$

   (b) Repeat the experiments in Example 6.5 with the same parameter set-up. Compare the two types of confidence interval, such as empirical coverage probability and average confidence interval width.

   In Example 6.5, we have $\mu = 0, \sigma = 2, n = 20, m = 1000$ replicates, and $\alpha = 0.05$. Let's compare the two types of C.I.:

```
#OLD confidence interval
set.seed(123)
n <- 20
alpha <- .05
UCL <- replicate(1000, expr = {
  x <- rnorm(n, mean = 0, sd = 2)
  (n-1) * var(x) / qchisq(alpha, df = n-1)
} )
#compute empirical coverage probability
> mean(UCL > 4)
[1] 0.95
#compute average confidence interval width
> mean(UCL)
[1] 7.527213

#NEW confidence interval
set.seed(456)
n <- 20
alpha <- .05
m <- 1000
width.CI <- y <- numeric(m)
for (i in 1:m){
  x <- rnorm(n, mean = 0, sd = 2)
  U.CI <- (n-1) * var(x) / qchisq(alpha/2, df = n-1)
```

```
    L.CI <- (n-1) * var(x) / qchisq(1-alpha/2, df = n-1)
    y[i] <- ifelse(4 <= U.CI & 4 >= L.CI, 1, 0)
    width.CI[i] <- U.CI - L.CI
  }
  #compute empirical coverage probability
  > mean(y)
  [1] 0.946
  #compute average confidence interval width
  > mean(width.CI)
  [1] 6.330749
```

(c) Repeat the experiments in Example 6.6 with the same parameter set-up. Compare the two types of confidence interval, such as empirical coverage probability and average confidence width.

In example 6.6, we repeat the simulation, replacing the N(0,4) samples with $\chi^2(2)$ samples.

```
#old confidence interval
set.seed(444)
n <- 20
alpha <- .05
UCL <- replicate(1000, expr = {
  x <- rchisq(n, df = 2)
  (n-1) * var(x) / qchisq(alpha, df = n-1)
} )
#compute empirical coverage probability
> mean(UCL > 4)
[1] 0.794
#compute average confidence interval width
> mean(UCL)
[1] 7.651295

#new confidence interval
set.seed(666)
n <- 20
alpha <- .05
m <- 1000
width.CI <- y <- numeric(m)
for (i in 1:m){
  x <- rchisq(n, df = 2)
  U.CI <- (n-1) * var(x) / qchisq(alpha/2, df = n-1)
  L.CI <- (n-1) * var(x) / qchisq(1-alpha/2, df = n-1)
  y[i] <- ifelse(4 <= U.CI & 4 >= L.CI, 1, 0)
  width.CI[i] <- U.CI - L.CI
}
#compute empirical coverage probability
> mean(y)
[1] 0.744
#compute average confidence interval width
> mean(width.CI)
[1] 6.170019
```

(d) Which confidence interval would you recommend in practice? Explain why.

As we can see, the OLD confidence interval has a higher empirical coverage probability with a longer average confidence interval width in both experiments. In practice, I will recommend the OLD confidence interval, because it has a more accurate empirical confidence level. The NEW confidence interval, though, narrows the width a little bit, the empirical confidence level also declines, which is nott the best choice in practise.