Assignment 5

Student Name : 정석규

Student ID : 201724570

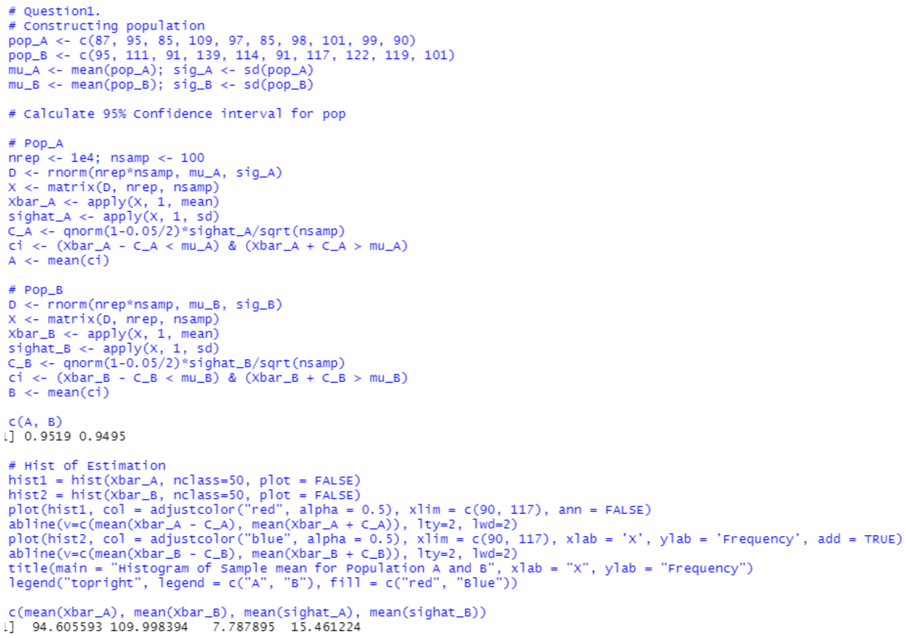
Q1.

1) **Source code :**

```
# Question1.
# Constructing population
pop_A <- c(87, 95, 85, 109, 97, 85, 98, 101, 99, 90)
pop_B <- c(95, 111, 91, 139, 114, 91, 117, 122, 119, 101)
mu_A <- mean(pop_A); sig_A <- sd(pop_A)
mu_B <- mean(pop_B); sig_B <- sd(pop_B)

# Calculate 95% Confidence interval for pop

# Pop_A
nrep <- 1e4; nsamp <- 100
D <- rnorm(nrep*nsamp, mu_A, sig_A)
X <- matrix(D, nrep, nsamp)
Xbar_A <- apply(X, 1, mean)
sighat_A <- apply(X, 1, sd)
C_A <- qnorm(1-0.05/2)*sighat_A/sqrt(nsamp)
ci <- (Xbar_A - C_A < mu_A) & (Xbar_A + C_A > mu_A)
A <- mean(ci)

# Pop_B
D <- rnorm(nrep*nsamp, mu_B, sig_B)
X <- matrix(D, nrep, nsamp)
Xbar_B <- apply(X, 1, mean)
```

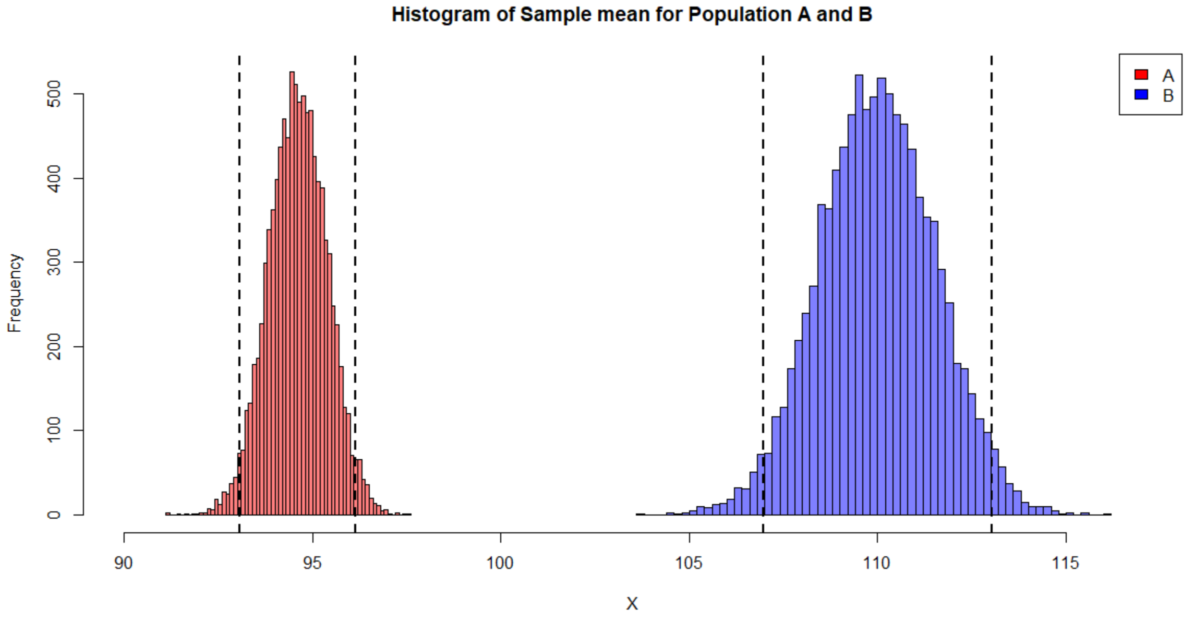
```
sighat_B <- apply(X, 1, sd)

C_B <- qnorm(1-0.05/2)*sighat_B/sqrt(nsamp)

ci <- (Xbar_B - C_B < mu_B) & (Xbar_B + C_B > mu_B)

B <- mean(ci)


c(A, B)


# Hist of Estimation

hist1 = hist(Xbar_A, nclass=50, plot = FALSE)

hist2 = hist(Xbar_B, nclass=50, plot = FALSE)

plot(hist1, col = adjustcolor("red", alpha = 0.5), xlim = c(90, 117), ann = FALSE)

abline(v=c(mean(Xbar_A - C_A), mean(Xbar_A + C_A)), lty=2, lwd=2)

plot(hist2, col = adjustcolor("blue", alpha = 0.5), xlim = c(90, 117), xlab = 'X', ylab = 'Frequency', add = TRUE)

abline(v=c(mean(Xbar_B - C_B), mean(Xbar_B + C_B)), lty=2, lwd=2)

title(main = "Histogram of Sample mean for Population A and B", xlab = "X", ylab = "Frequency")

legend("topright", legend = c("A", "B"), fill = c("red", "Blue"))


c(mean(Xbar_A), mean(Xbar_B), mean(sighat_A), mean(sighat_B))
```

2) R screenshot :



```
# Question1.
# Constructing population
pop_A <- c(87, 95, 85, 109, 97, 85, 98, 101, 99, 90)
pop_B <- c(95, 111, 91, 139, 114, 91, 117, 122, 119, 101)
mu_A <- mean(pop_A); sig_A <- sd(pop_A)
mu_B <- mean(pop_B); sig_B <- sd(pop_B)

# Calculate 95% Confidence interval for pop

# Pop_A
nrep <- 1e4; nsamp <- 100
D <- rnorm(nrep*nsamp, mu_A, sig_A)
X <- matrix(D, nrep, nsamp)
Xbar_A <- apply(X, 1, mean)
sighat_A <- apply(X, 1, sd)
C_A <- qnorm(1-0.05/2)*sighat_A/sqrt(nsamp)
ci <- (Xbar_A - C_A < mu_A) & (Xbar_A + C_A > mu_A)
A <- mean(ci)

# Pop_B
D <- rnorm(nrep*nsamp, mu_B, sig_B)
X <- matrix(D, nrep, nsamp)
Xbar_B <- apply(X, 1, mean)
sighat_B <- apply(X, 1, sd)
C_B <- qnorm(1-0.05/2)*sighat_B/sqrt(nsamp)
ci <- (Xbar_B - C_B < mu_B) & (Xbar_B + C_B > mu_B)
B <- mean(ci)

c(A, B)
1] 0.9519 0.9495

# Hist of Estimation
hist1 = hist(Xbar_A, nclass=50, plot = FALSE)
hist2 = hist(Xbar_B, nclass=50, plot = FALSE)
plot(hist1, col = adjustcolor("red", alpha = 0.5), xlim = c(90, 117), ann = FALSE)
abline(v=c(mean(Xbar_A - C_A), mean(Xbar_A + C_A)), lty=2, lwd=2)
plot(hist2, col = adjustcolor("blue", alpha = 0.5), xlim = c(90, 117), xlab = 'X', ylab = 'Frequency', add = TRUE)
abline(v=c(mean(Xbar_B - C_B), mean(Xbar_B + C_B)), lty=2, lwd=2)
title(main = "Histogram of Sample mean for Population A and B", xlab = "X", ylab = "Frequency")
legend("topright", legend = c("A", "B"), fill = c("red", "Blue"))

c(mean(Xbar_A), mean(Xbar_B), mean(sighat_A), mean(sighat_B))
1]  94.605593 109.998394   7.787895  15.461224
```

3) Answer :



Population 통계량에 대한 95% Confidence interval을 계산하기 위해 nrep과 nsamp를 1e4, 100으로 설정하고 Confidence interval을 계산했을 때, 추정된 Sample mean A와 Sample mean B의 분포는 위의 그래프와 같다, 이때 95%신뢰구간은 서로 overlap되지 않는다.

Xbar의 A, B의 sample mean의 불편추정량은 94.61, 109.99로, B의 sample mean의 평균이 더 높고, sighat의 평균은 각각 7.79, 15.46으로 신뢰구간의 넓이도 2배정도 차이남을 확인할 수 있다.

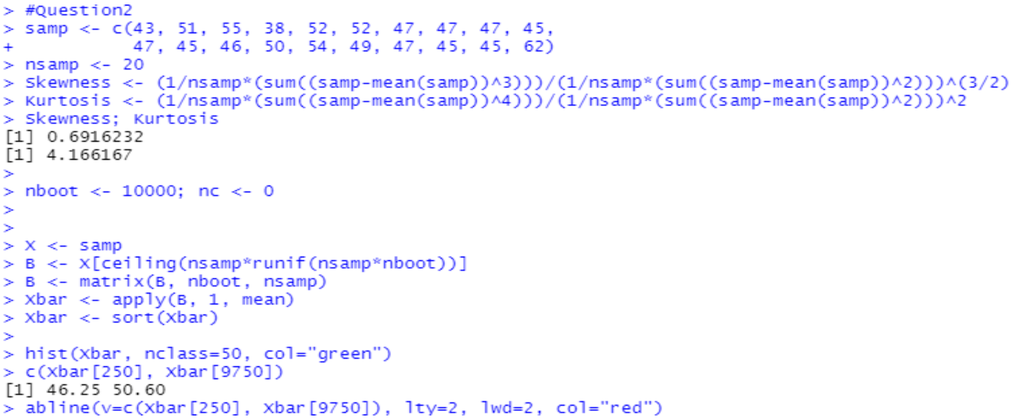
Q2.

1) Source code :

```
# Question2.

# Conclusion of Normality

samp <- c(43, 51, 55, 38, 52, 52, 47, 47, 47, 45,

          47, 45, 46, 50, 54, 49, 47, 45, 45, 62)

nsamp <- 20

Skewness <- (1/nsamp*(sum((samp-mean(samp))^3)))/(1/nsamp*(sum((samp-mean(samp))^2)))^(3/2)

Kurtosis <- (1/nsamp*(sum((samp-mean(samp))^4)))/(1/nsamp*(sum((samp-mean(samp))^2)))^2

Skewness; Kurtosis


# 95% Confidence interval for population mean
```

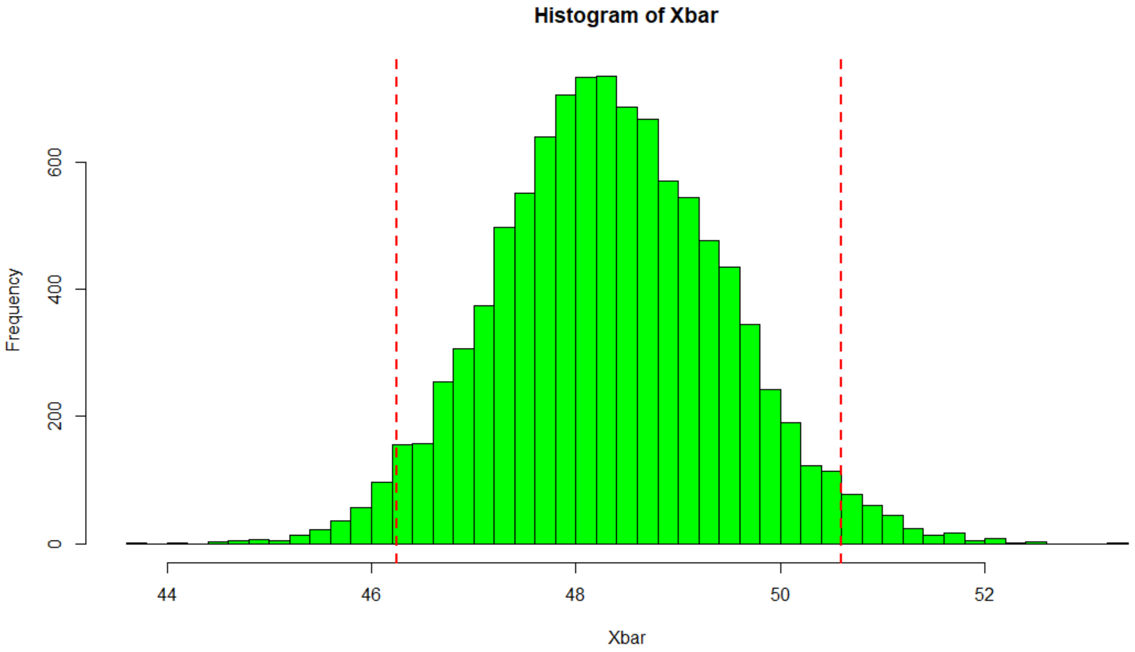
```
nboot <- 10000; nc <- 0

X <- samp

B <- X[ceiling(nsamp*runif(nsamp*nboot))]

B <- matrix(B, nboot, nsamp)

Xbar <- apply(B, 1, mean)

Xbar <- sort(Xbar)


hist(Xbar, nclass=50, col="green")

c(Xbar[250], Xbar[9750])

abline(v=c(Xbar[250], Xbar[9750]), lty=2, lwd=2, col="red")
```

2) R screenshot :



```
> #Question2
> samp <- c(43, 51, 55, 38, 52, 52, 47, 47, 47, 45,
+           47, 45, 46, 50, 54, 49, 47, 45, 45, 62)
> nsamp <- 20
> Skewness <- (1/nsamp*(sum((samp-mean(samp))^3)))/(1/nsamp*(sum((samp-mean(samp))^2)))^(3/2)
> Kurtosis <- (1/nsamp*(sum((samp-mean(samp))^4)))/(1/nsamp*(sum((samp-mean(samp))^2)))^2
> Skewness; Kurtosis
[1] 0.6916232
[1] 4.166167
>
> nboot <- 10000; nc <- 0
>
>
> X <- samp
> B <- X[ceiling(nsamp*runif(nsamp*nboot))]
> B <- matrix(B, nboot, nsamp)
> Xbar <- apply(B, 1, mean)
> Xbar <- sort(Xbar)
>
> hist(Xbar, nclass=50, col="green")
> c(Xbar[250], Xbar[9750])
[1] 46.25 50.60
> abline(v=c(Xbar[250], Xbar[9750]), lty=2, lwd=2, col="red")
```

3) **Answer :** Skewness의 값은 0.6916232, Kurtosis의 값은 4.166167으로 계산되어 Skewness = 0, Kurtosis = 3의 조건을 만족하지 않기 때문에 모집단은 Normal distribution을 따르지 않음을 추론할 수 있다. 모집단이 Normal distribution을 따르지 않으므로 Bootstrap을 이용하여 Population mean을 추론한다.



10,000 bootstrap replications를 진행했을 때, 생성된 10,000개의 sample mean의 Histogram은 위와 같은데,

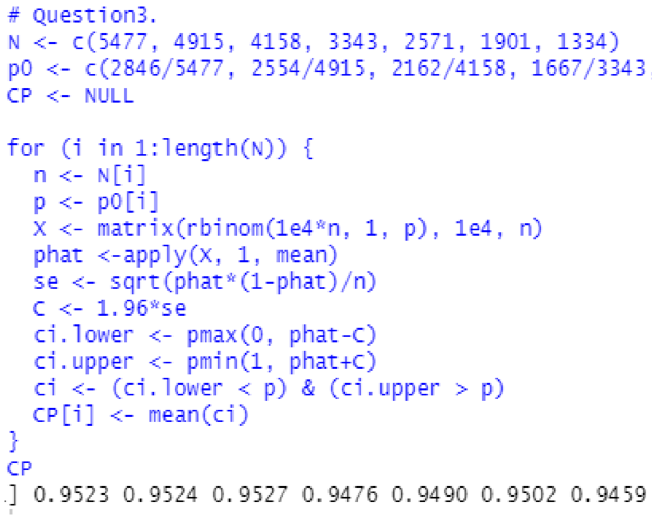
이때 Xbar[250]과 Xbar[9750]의 값은 각각 46.25 50.65로, 빨간 점선의 경계선안의 구간이 95% 신뢰구간을 형성하고 있음을 알 수 있다.

Q3.

1) **Source code :**

```
 X <- matrix(rbinom(1e4*n, 1, p0), 1e4, n)

 phat <-apply(X, 1, mean)

 se <- sqrt(phat*(1-phat)/n)

 C <- 1.96*se

 ci.lower <- pmax(0, phat-C)

 ci.upper <- pmin(1, phat+C)

 ci <- (ci.lower < p0) & (ci.upper > p0)

mean(ci)
```

2) **R screenshot :**



```
# Question3.
N <- c(5477, 4915, 4158, 3343, 2571, 1901, 1334)
p0 <- c(2846/5477, 2554/4915, 2162/4158, 1667/3343, 1341/2571, 987/1901, 666/1334)
CP <- NULL

for (i in 1:length(N)) {
  n <- N[i]
  p <- p0[i]
  X <- matrix(rbinom(1e4*n, 1, p), 1e4, n)
  phat <-apply(X, 1, mean)
  se <- sqrt(phat*(1-phat)/n)
  C <- 1.96*se
  ci.lower <- pmax(0, phat-C)
  ci.upper <- pmin(1, phat+C)
  ci <- (ci.lower < p) & (ci.upper > p)
  CP[i] <- mean(ci)
}
CP
] 0.9523 0.9524 0.9527 0.9476 0.9490 0.9502 0.9459
```

3) **Answer :** Order of birth의 확률이 모비율인지 표본비율인지 잘 모르겠음. 모비율의 bootstrap 적용방식을 모르겠음.