BST 232: Homework 1

Phuc Vu

09-13-2024

Question 1

(a)

$$\begin{aligned}
E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy \\
&= \int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f(x,y)dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f(x,y)dxdy \\
&= \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dx \\
&= E(X) + E(Y)\square
\end{aligned}$$

(b)

$$\begin{aligned}
E(cX) &= \int_{-\infty}^{\infty}(cx)f(x)dx \\
&= c\int_{-\infty}^{\infty}(x)f(x)dx \\
&= cE(X)
\end{aligned}$$

(c)

$$\begin{aligned}
Var(X) &= E((X-E(X))^2) \\
&= E(X^2 - 2XE(X) + E(X)^2) \\
&= E(X^2) - 2E(XE(X)) + E(E(X)^2) \\
&\textit{Note that: } E(XE(X)) = E(X)E(X) = E(X)^2 \textit{ (Result of 1b and } E(X) = c\textit{)} \\
&= E(X^2) - 2E(X)^2 + E(X)^2 \\
&= E(X^2) - E(X)^2\square
\end{aligned}$$

(d)

$$\begin{aligned}
Var(X+c) &= E((X+c)^2) - E(X+c)^2 \text{ (by 1c)} \\
&= E(X^2 + 2Xc + c^2) - (E(X)+c)^2 \\
&= E(X^2) + 2E(cX) + c^2 - E(X)^2 - 2cE(X) - c^2 \text{ (by 1a)} \\
&= E(X^2) - E(X)^2 \text{ (Since E(cX)=cE(X))} \\
&= Var(X) \text{ (by 1c)} \square
\end{aligned}$$

(e)

$$\begin{aligned}
Var(cX) &= E((cX)^2) - E(cX)^2 \text{ (by 1c)} \\
&= E(c^2X^2) - (cE(X))^2 \text{ (by 1b)} \\
&= c^2E(X^2) - c^2E(X)^2 \text{ (by 1b)} \\
&= c^2(E(X^2) - E(X)^2) \\
&= c^2Var(X) \text{ (by 1c)} \square
\end{aligned}$$

(f)

$$\begin{aligned}
E(E(X|Y)) &= \int_{-\infty}^{\infty} E(X|Y=y) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy \text{ (Since } f_{X|Y}(x|y) f_Y(y) = f_{X,Y}(x,y)) \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \text{ (Since } \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = f_X(x)) \\
&= E(X) \square
\end{aligned}$$

Question 2

(a)

$\hat{\mu}$ should be about the same as μ since both 100 and 200 are relatively large sample sizes and $\hat{\mu}$ is an unbiased estimator of μ

(b)

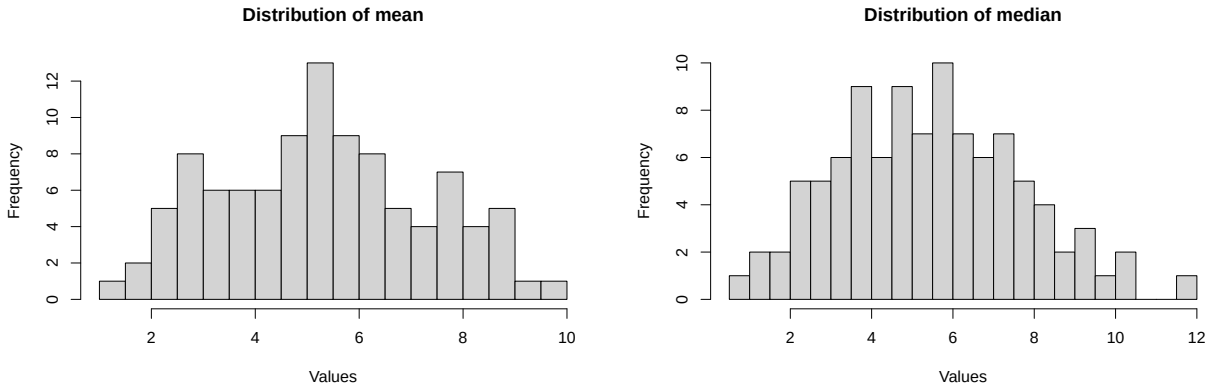
Since the $Sd(\hat{\mu}) = \sqrt{\frac{\sigma^2}{n}}$ and both of our data come from a same population (similar σ^2), $Sd(\hat{\mu})$ from the the colleague's larger data set (larger n) should be smaller

(c)

We have the p-value$= 2Pr(Z > |\frac{\hat{\mu}-\mu_0}{Sd(\hat{\mu})}|)$ (since both samples are large, the use of Z-value instead of T-value hardly make a difference). Since the standard normal distribution is symmetric and centered around 0, with similar $\hat{\mu}$, the colleague will have smaller p-value due to smaller $Sd(\hat{\mu})$

Question 3

(a)



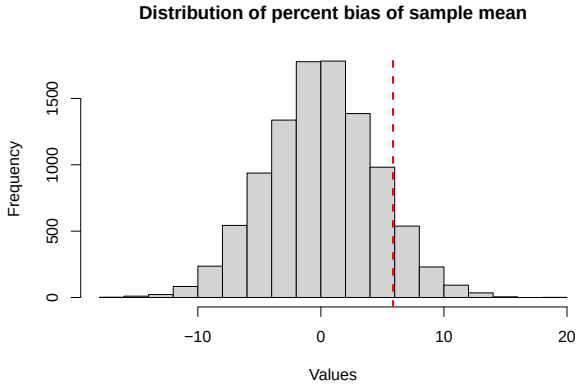
We normally call this distribution the sampling distribution

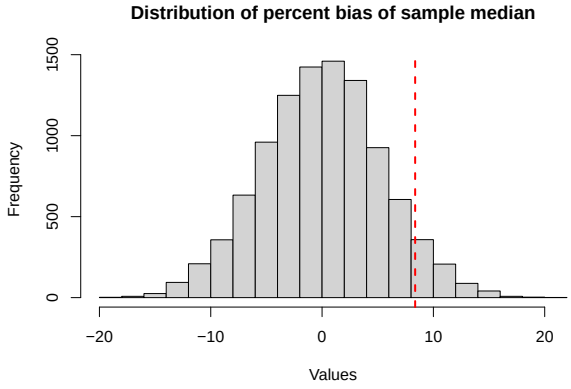
(b)

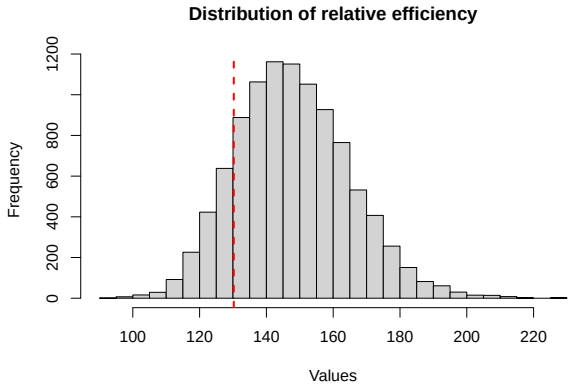
```
##              % Bias - muH % Bias - muT Rel. eff Rel. unc
## Simulation 1     5.859312     8.362184 130.2312 114.1189
```

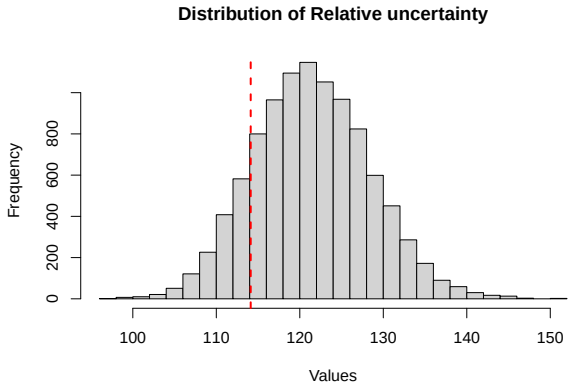
$\hat{\mu}$ has a lower bias (5.859 versus 8.362) and lower variance (relative efficiency: 130.231%) compared to $\tilde{\mu}$

(c)









Have I only conducted a single simulation of size R=100, I would have make a biased conclusion

(d)

The spread in the histograms can be understood as the variation of the estimator. The wilder the spread the higher variation.

The Monte Carlo error for the simulations are:

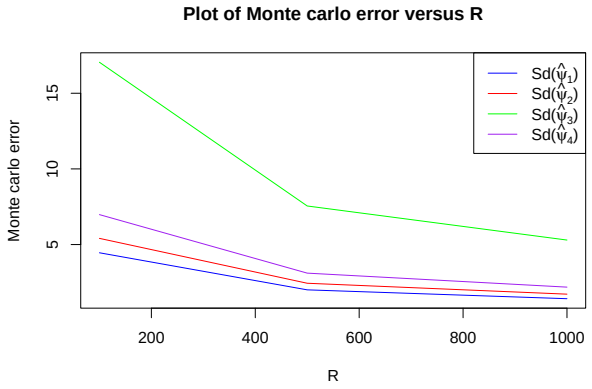
$Sd(\hat{\psi}_1) = 4.457$

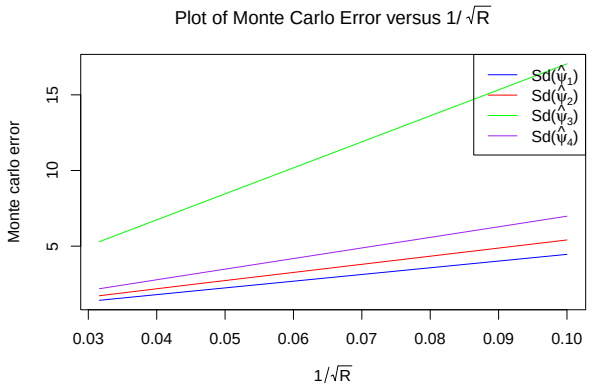
$Sd(\hat{\psi}_2) = 5.409$

$Sd(\hat{\psi}_3) = 17.049$

$Sd(\hat{\psi}_4) = 6.979$

(e)





Collectively, The lower the R value, the lower the Monte carlo errors.

Code

Important: The blank R chunk below automatically compiles all of the code chunks you wrote above into one long appendix. Do not remove it or write anything in it (unless you want to delete the code appendix for some reason). Again, ensure that you've labeled each chunk so that the teaching team knows which code pertains to which question.

```
knitr::opts_chunk$set(
  tidy = T, results = 'hold', echo = FALSE, warning = FALSE, message = FALSE,
  out.width = "50%"
)

# This code cell sets the template format.
# You can change `out.width` to change the plot size.
# Note you may need to run install.packages("formatR") for this cell to work.

# You can also load your libraries here, like so:
library(ggplot2)

source("~/Library/CloudStorage/OneDrive-HarvardUniversity/PhD/Year_1/MethodI/homework/runSimulation.R")
```

```r
### Question 3 (a) ###
set.seed(123)
q3_a<-runSimulation(mu=5,sigma=10,n=20,M=1,returnRaw = TRUE)

hist(q3_a$rawOutpu[,1,1],
     main = "Distribution of mean",  # Title of the histogram
     xlab = "Values",                      # Label for the x-axis
     ylab = "Frequency",breaks=20)                 # Label for the y-axis
hist(q3_a$rawOutpu[,2,1],
     main = "Distribution of median",  # Title of the histogram
     xlab = "Values",                      # Label for the x-axis
     ylab = "Frequency",breaks=20)                 # Label for the y-axis
### Question 3 (b) ###
q3_a$simResults
### Question 3 (c) ###
set.seed(123)
q3_c<-runSimulation(mu=5,sigma=10,n=20,M=10000,returnRaw = FALSE)
hist(q3_c[,1],
     main = "Distribution of percent bias of sample mean",  # Title of the histogram
     xlab = "Values",                      # Label for the x-axis
     ylab = "Frequency",breaks=20)  # Label for the y-axis

abline(v = q3_a$simResults[,1], col = "red", lwd = 2, lty = 2)
hist(q3_c[,2],
     main = "Distribution of percent bias of sample median",  # Title of the histogram
     xlab = "Values",                      # Label for the x-axis
     ylab = "Frequency",breaks=20)  # Label for the y-axis

abline(v = q3_a$simResults[,2], col = "red", lwd = 2, lty = 2)
hist(q3_c[,3],
     main = "Distribution of relative efficiency",  # Title of the histogram
     xlab = "Values",                      # Label for the x-axis
     ylab = "Frequency",breaks=20)  # Label for the y-axis

abline(v = q3_a$simResults[,3], col = "red", lwd = 2, lty = 2)
hist(q3_c[,4],
     main = "Distribution of Relative uncertainty",  # Title of the histogram
     xlab = "Values",                      # Label for the x-axis
     ylab = "Frequency",breaks=20)  # Label for the y-axis

abline(v = q3_a$simResults[,4], col = "red", lwd = 2, lty = 2)
### Question 3 (d) ###
sd_phi_1_3c<-sd(q3_c[,1])
sd_phi_2_3c<-sd(q3_c[,2])
sd_phi_3_3c<-sd(q3_c[,3])
sd_phi_4_3c<-sd(q3_c[,4])
### Question 3 (e) ###
set.seed(123)
q3_c_500<-runSimulation(mu=5,R=500,sigma=10,n=20,M=10000,returnRaw = FALSE)
q3_c_1000<-runSimulation(mu=5,R=1000,sigma=10,n=20,M=10000,returnRaw = FALSE)
sd_phi_1_3c_500<-sd(q3_c_500[,1])
sd_phi_2_3c_500<-sd(q3_c_500[,2])
sd_phi_3_3c_500<-sd(q3_c_500[,3])
```

```r
sd_phi_4_3c_500<-sd(q3_c_500[,4])
sd_phi_1_3c_1000<-sd(q3_c_1000[,1])
sd_phi_2_3c_1000<-sd(q3_c_1000[,2])
sd_phi_3_3c_1000<-sd(q3_c_1000[,3])
sd_phi_4_3c_1000<-sd(q3_c_1000[,4])
sd_phi_1<-c(sd_phi_1_3c,sd_phi_1_3c_500,sd_phi_1_3c_1000)
sd_phi_2<-c(sd_phi_2_3c,sd_phi_2_3c_500,sd_phi_2_3c_1000)
sd_phi_3<-c(sd_phi_3_3c,sd_phi_3_3c_500,sd_phi_3_3c_1000)
sd_phi_4<-c(sd_phi_4_3c,sd_phi_4_3c_500,sd_phi_4_3c_1000)
x<-c(100,500,1000)
plot(x, sd_phi_1, type = "l", col = "blue", ylim = range(c(sd_phi_1, sd_phi_2, sd_phi_3, sd_phi_4)),
     xlab = "R", ylab = "Monte carlo error", main = "Plot of Monte carlo error versus R")

# Add the other lines
lines(x, sd_phi_2, col = "red")
lines(x, sd_phi_3, col = "green")
lines(x, sd_phi_4, col = "purple")
legend("topright", legend = expression(
  paste("Sd(", hat(psi)[1], ")"),
  paste("Sd(", hat(psi)[2], ")"),
  paste("Sd(", hat(psi)[3], ")"),
  paste("Sd(", hat(psi)[4], ")")
), col = c("blue", "red", "green", "purple"), lty = 1)
x<-1/sqrt(c(100,500,1000))
plot(x, sd_phi_1, type = "l", col = "blue", ylim = range(c(sd_phi_1, sd_phi_2, sd_phi_3, sd_phi_4)),
     xlab = expression(1/sqrt(R)), ylab = "Monte carlo error", main = expression("Plot of Monte Carlo E

# Add the other lines
lines(x, sd_phi_2, col = "red")
lines(x, sd_phi_3, col = "green")
lines(x, sd_phi_4, col = "purple")
legend("topright", legend = expression(
  paste("Sd(", hat(psi)[1], ")"),
  paste("Sd(", hat(psi)[2], ")"),
  paste("Sd(", hat(psi)[3], ")"),
  paste("Sd(", hat(psi)[4], ")")
), col = c("blue", "red", "green", "purple"), lty = 1)
```