CS 6313 - Statistics for Data Science
Mini Project 2

Team Number - 29
Team Members - Niharika Rajaram, Tejas Prakash Bobhate

Contribution of each member:
Niharika Rajaram:

- Solved all the problems.
- Studied the theory of basic statistics to understand and implement both questions.
- Problem 1 requires understanding and implementing computing of Mean Square Error using Monte Carlo Simulation for different combinations of n and θ and applying required plotting methods for visual interpretations.
- Wrote R code for Question 1 and report for Question 2.

Tejas Prakash Bobhate:

- Solved all the problems.
- Studied the theory of basic statistics to understand and implement both questions
- Wrote R code for Question 2 and report for Question 1.
- Problem 2 requires an analytical solution for the Maximum Likelihood Estimator of a given probability density function.

Both of us solved all the questions and mentioned the best solution we had in the report.

Question 1

1.a)

To compute the mean squared error, we require uniform random samples (runif) with values of **n** and **theta** varied and supplied in the query. The values will be simulated from the uniform random distribution function and estimate the method of moment (mean of the random samples) and maximum likelihood estimation (maximum value of the random sample) then we compute the MSE1 i.e., the difference between the random sample e, and the method of moment whole squared and MSE2 i.e., the difference between the e of random sample and the MLE.

MSE is given by **MSE = or E[**$(\hat{\theta} - \theta)^2$**]**

1.b)

```
> n_s <- function(n, theta){
+   data = runif(n,0,theta)
+   mom = 2*mean(data)
+   mle = max(data)
+   return (c(mle, mom))
+ }
>
> comp <- function(n, theta){
+   estimate <- replicate(1000, n_s(n, theta))
+   mse <- (estimate - theta) ^2
+   mse.mle = mean(mse[c(TRUE, FALSE)])
+   mse.mom = mean(mse[c(FALSE, TRUE)])
+   return(c(mse.mle, mse.mom))
+ }
>
> res11 = comp(1,1)
> res11
[1] 0.3318478 0.3262722
> res550 = comp(1,10)
> res550
[1] 33.8854 33.1389
```

The MSE for n = 1 and Θ = 1 -> $\widehat{\Theta 1}$ = 0.3318 and $\widehat{\Theta 2}$ = 0.3262

The MSE for n = 1 and Θ = 10 -> $\widehat{\Theta 1}$ = 33.8854 and $\widehat{\Theta 2}$ = 33.1389

It can be inferred that for the given combinations of n = 1, θ = 1 and n=1 and θ = 10, the MSE values are approximately closer both for MSE and MLE.

1.c)

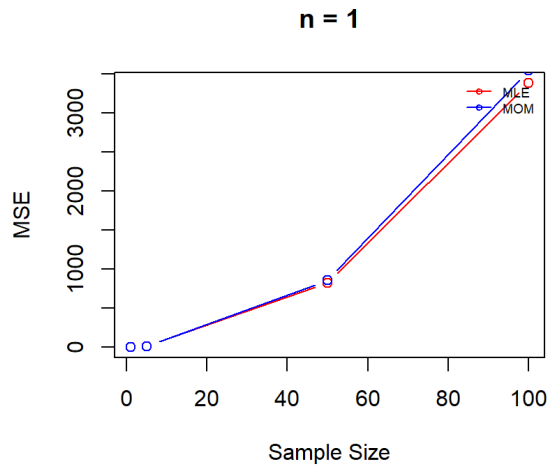
Given n = 1, 2, 3, 5, 10, 30 and θ = 1, 5, 50, 100,

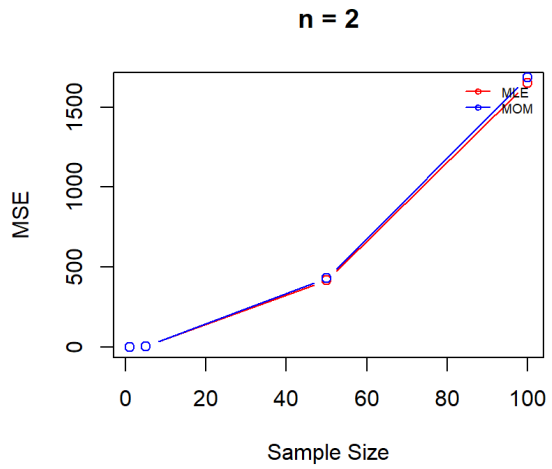
We use all the combinations of n and θ and perform Monte Carlo Simulation

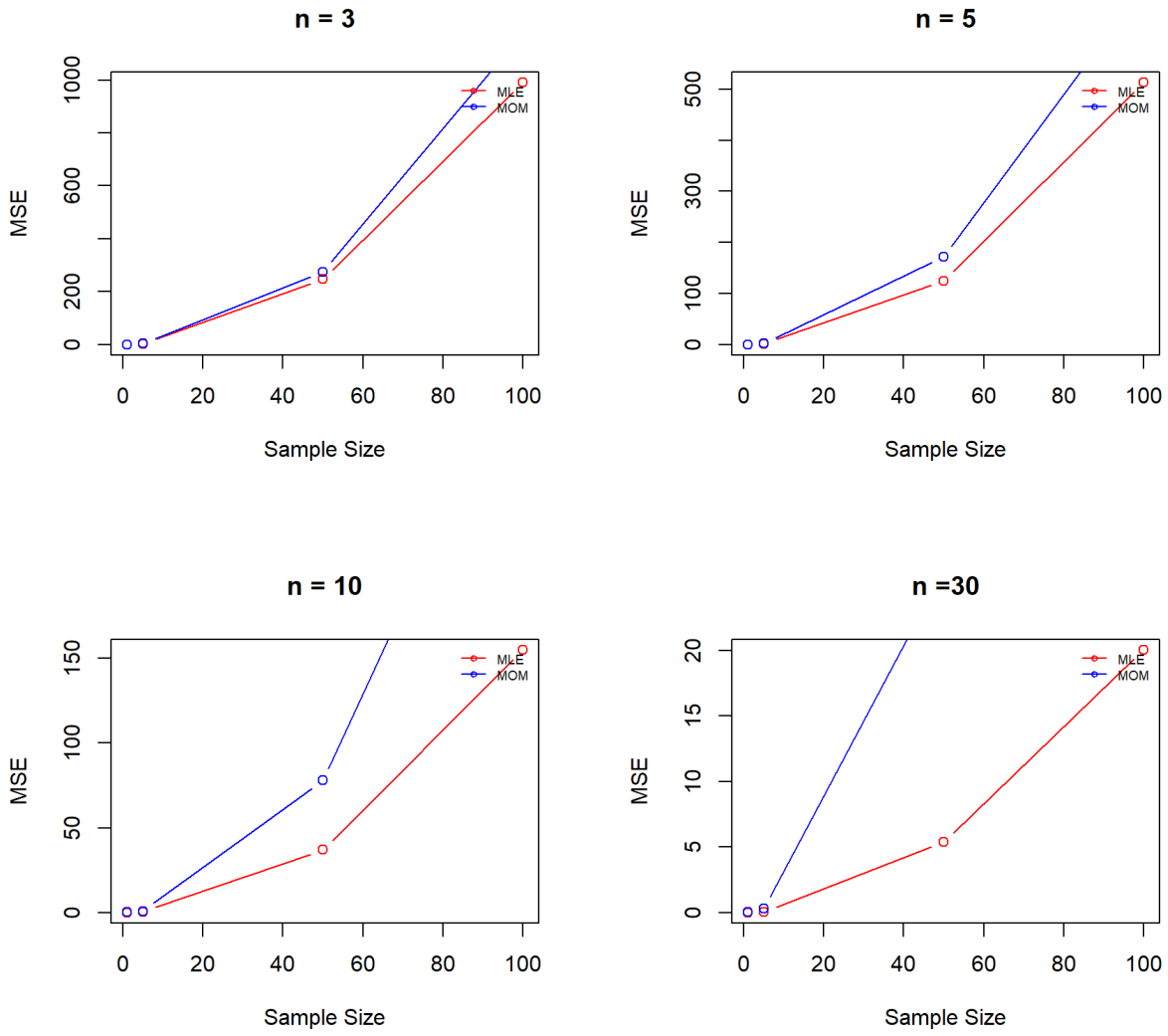
```
> res11 = comp(1,1)
> res11
[1] 0.3415798 0.3326653
> res15 = comp(1,5)
> res15
[1] 8.589885 8.119630
> res150 = comp(1,50)
> res150
[1] 822.416 850.619
> res1100 = comp(1, 100)
> res1100
[1] 3381.520 3538.599
> res21 = comp(2,1)
> res21
[1] 0.1639926 0.1612289
> res25 = comp(2,5)
> res25
[1] 4.355834 4.135820
> res250 = comp(2,50)
> res250
[1] 416.3793 429.6051
> res2100 = comp(2,100)
> res2100
[1] 1651.796 1686.873
> res31 = comp(3,1)
> res31
[1] 0.1072617 0.1152408
> res35 = comp(3,5)
> res35
[1] 2.487607 2.817341
> res350 = comp(3,50)
> res350
[1] 247.2300 273.8454
> res3100 = comp(3,100)
> res3100
[1]  990.8832 1178.0212
```

```
> res51 = comp(5,1)
> res51
[1] 0.04398459 0.06494585
> res55 = comp(5,5)
> res55
[1] 1.127400 1.688477
> res550 = comp(5,50)
> res550
[1] 124.5682 171.5998
> res5100 = comp(5,100)
> res5100
[1] 513.3072 702.9413
> res101 = comp(10,1)
> res101
[1] 0.01317505 0.03162926
> res105 = comp(10,5)
> res105
[1] 0.3758894 0.7863966
> res1050 = comp(10,50)
> res1050
[1] 36.91713 77.89337
> res10100 = comp(10,100)
> res10100
[1] 155.0445 333.8042
> res301 = comp(30,1)
> res301
[1] 0.001991779 0.011421823
> res305 = comp(30,5)
> res305
[1] 0.04611375 0.29063177
> res3050 = comp(30,50)
> res3050
[1]  5.382843 26.099487
> res30100 = comp(30,100)
> res30100
[1]  20.03455 112.03136
```

```r
thetas <- c(1,5,50,100)
MLE_t_1 = c(res11[1], res15[1], res150[1], res1100[1])
MLE_t_2 = c(res21[1], res25[1], res250[1], res2100[1])
MLE_t_3 = c(res31[1], res35[1], res350[1], res3100[1])
MLE_t_5 = c(res51[1], res55[1], res550[1], res5100[1])
MLE_t_10 = c(res101[1], res105[1], res1050[1], res10100[1])
MLE_t_30 = c(res301[1], res305[1], res3050[1], res30100[1])

MOM_t_1 = c(res11[2], res15[2], res150[2], res1100[2])
MOM_t_2 = c(res21[2], res25[2], res250[2], res2100[2])
MOM_t_3 = c(res31[2], res35[2], res350[2], res3100[2])
MOM_t_5 = c(res51[2], res55[2], res550[2], res5100[2])
MOM_t_10 = c(res101[2], res105[2], res1050[2], res10100[2])
MOM_t_30 = c(res301[2], res305[2], res3050[2], res30100[2])

plot(thetas, MLE_t_1, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 1")
lines(thetas, MOM_t_1, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_2, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 2")
lines(thetas, MOM_t_2, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_3, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 3")
lines(thetas, MOM_t_3, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_5, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 5")
lines(thetas, MOM_t_5, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_10, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 10")
lines(thetas, MOM_t_10, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_30, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n =30")
lines(thetas, MOM_t_30, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')
```



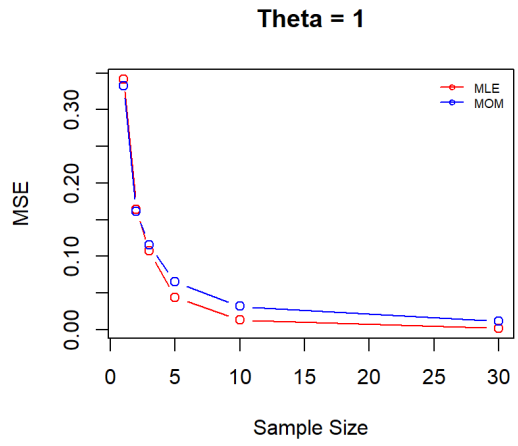


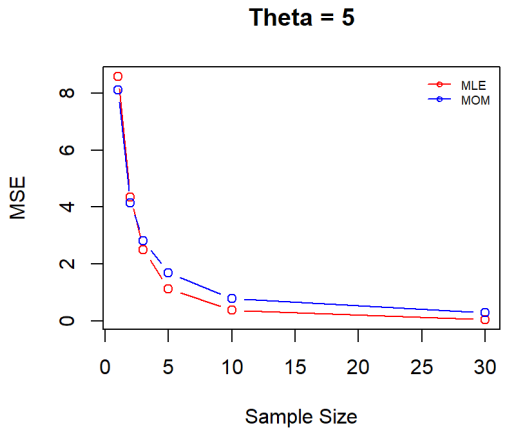


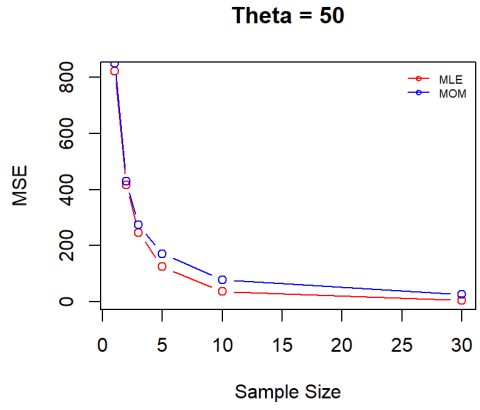
From the above graphs, for n = 1, 2, 3, 5, 10, 30, the curves are not the same. Based on this, we can state that the estimator depends on the value of n.

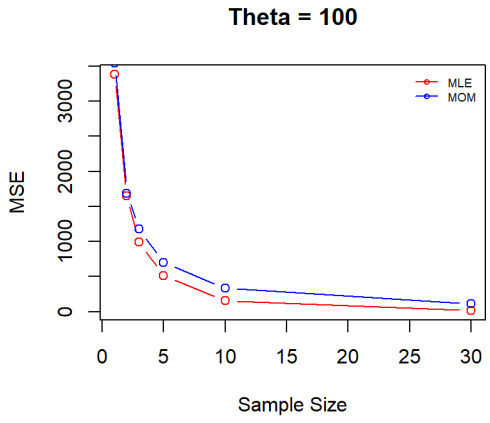
It can be seen that for the same value of n, both estimators are giving good results (very low MSE) for small θ values. But as the θ values are increased, the MSE outputs of Method of Moment Estimator is very high, while the Maximum Likelikehood Estimator is still performing good with high thetha values.

```r
par = (mfrow = c(2,2))

n = c(1,2,3,5,10,30)

MLE_1 <- c(res11[1], res21[1], res31[1], res51[1], res101[1], res301[1])
MLE_5 <- c(res15[1], res25[1], res35[1], res55[1], res105[1], res305[1])
MLE_50 <- c(res150[1], res250[1], res350[1], res550[1], res1050[1], res3050[1])
MLE_100 <- c(res1100[1], res2100[1], res3100[1], res5100[1], res10100[1], res30100[1])

MOM_1 <- c(res11[2], res21[2], res31[2], res51[2], res101[2], res301[2])
MOM_5 <- c(res15[2], res25[2], res35[2], res55[2], res105[2], res305[2])
MOM_50 <- c(res150[2], res250[2], res350[2], res550[2], res1050[2], res3050[2])
MOM_100 <- c(res1100[2], res2100[2], res3100[2], res5100[2], res10100[2], res30100[2])

plot(n, MLE_1, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 1")
lines(n, MOM_1, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')


plot(n, MLE_5, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 5")
lines(n, MOM_5, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')


plot(n, MLE_50, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 50")
lines(n, MOM_50, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')


plot(n, MLE_100, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 100")
lines(n, MOM_100, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1, pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')
```









From the above graphs, for Θ = 1, 5, 50, 100, the curves almost are the same. Based on this, we can state that the estimator does not depend on the value of Θ.

When we plot the MSE of MME and MLE across different n values, keeping θ constant, it is evident that increasing the sample size (n) leads to more stable values of estimated parameters with very low MSE values. Also, the shape for both MOM and MLE is extremely similar, interpreting that there is not much variation in any of the estimator if we the change the thetha.

1.d)

- For n = 1, MLE and MME are almost identical.
- For n <= 3, the predictions of MME are good. Therefore, for small n, we can use MME.
- For n > 3, MLE prediction is better than MME. Thus, for large n, MLE is preferred as MSE (Mean Square Error) for MLE is less when we compare it to MSE (Mean Square Error) for MME.

Based on above graphical data and visualization, we can conclude that –
i. When n is small, and θ is also low – Both estimators (MOM and MLE) – gives small MSE.
ii. When n is large, and θ is also low – MLE is performing quite better than MOM estimator, resulting in very low MSE.
iii. When θ is large – For both of the estimators, there is high degree of variance and both are producing high values of MSE (MLE is still slightly lower than MOM) Therefore, in totality it can be concluded that MLE performs better than MOM, but it also depend on the variation of n and θ.

Question 2

2.a)

Given

f(x) = $\left\{ \frac{\theta}{x^{\theta+1}},\ x \geq 1\ ;\ 0\,x\ <\ 1 \right\}$

Likelihood L(θ) = $\prod_{i=1}^{n} \left(\frac{\theta}{x_i^{\theta+1}} \right)$

Taking log on both sides,

$\log(L(\theta)) = \log\left(\prod_{i=1}^{n} \left(\frac{\theta}{x_i^{\theta+1}} \right) \right)$

$= \log\left(\theta^n \ * \ \prod_{i=1}^{n} \left(\frac{1}{x_i^{\theta+1}} \right) \right)$

$= n\log\theta + \sum_{i=1}^{n} \log\left(x_i^{-(\theta+1)} \right)$

$= n\log\theta - (\theta\ +\ 1) \sum_{i=1}^{n} \log x_i$

$= n\log\theta - \theta \sum_{i=1}^{n} \log x_i - \sum_{i=1}^{n} \log x_i$

Taking partial derivative with respect to θ and equating it to 0,

$\frac{\partial(log(L(\theta)))}{\partial\theta} \ = \ \frac{n}{\theta} - \sum_{i=1}^{n} logx_i \ = \ 0$

$\Rightarrow \ \frac{n}{\theta} \ = \sum_{i=1}^{n} logx_i$

$$\widehat{\theta}_{MLE} \ = \ \frac{n}{\sum_{i=1}^{n} logx_i}$$

2.b)

```
> n <- 5
> x <- c(21.72, 14.65, 50.42, 28.78, 11.23)
> mle <- n/ sum(log(x))
> mle
[1] 0.3233874
```

The MLE is 0.3233874.

2.c)

```
log_likelihood <- function(par, dat){
  result <- -1 * sum(log(par/dat^(par+1)))
  return (result)
}

MLE <- optim(par = 1, fn = log_likelihood, method="BFGS", hessian = TRUE, dat = x)
MLE
```

```
> MLE
$par
[1] 0.323387

$value
[1] 26.10585

$counts
function gradient
      22        9

$convergence
[1] 0

$message
NULL

$hessian
         [,1]
[1,] 47.81158
```

$par is 0.323387 which is the same as what we obtained in question 2b. So, the answers match.

2.d)

```
> se <- sqrt(diag(solve(MLE$hessian)))
> se
[1] 0.1446217
>
> CI_lower <- MLE$par - qnorm(1-(0.05/2)) * se
> CI_upper <- MLE$par + qnorm(1-(0.05/2)) * se
> CI_lower
[1] 0.03993372
> CI_upper
[1] 0.6068403
```

As we know that the confidence interval gives the interval out of 100% population, in which the true estimate value lies within that specified range, so here we can expect that out of the 100 trials for the same population having different samples, the true estimate value lies within that interval 95% of the times.

Section 2 : R Codes

Question 1b:

```
n_s <- function(n, theta){
  data = runif(n,0,theta)
  mom = 2*mean(data)
  mle = max(data)
  return (c(mle, mom))
}

comp <- function(n, theta){
  estimate <- replicate(1000, n_s(n, theta))
  mse <- (estimate - theta) ^2
  mse.mle = mean(mse[c(TRUE, FALSE)])
  mse.mom = mean(mse[c(FALSE, TRUE)])
  return(c(mse.mle, mse.mom))
}

res11 = comp(1,1)
res11
res550 = comp(1, 10)
res550
```

Question 1c:

```
res11 = comp(1,1)
res11
res15 = comp(1,5)
res15
res150 = comp(1,50)
res150
res1100 = comp(1, 100)
res1100
res21 = comp(2,1)
res21
res25 = comp(2,5)
res25
res250 = comp(2,50)
res250
res2100 = comp(2,100)
res2100
res31 = comp(3,1)
res31
```

```
res35 = comp(3,5)
res35
res350 = comp(3,50)
res350
res3100 = comp(3,100)
res3100
res51 = comp(5,1)
res51
res55 = comp(5,5)
res55
res550 = comp(5,50)
res550
res5100 = comp(5,100)
res5100
res101 = comp(10,1)
res101
res105 = comp(10,5)
res105
res1050 = comp(10,50)
res1050
res10100 = comp(10,100)
res10100
res301 = comp(30,1)
res301
res305 = comp(30,5)
res305
res3050 = comp(30,50)
res3050
res30100 = comp(30,100)
res30100

par = (mfrow = c(2,2))

n = c(1,2,3,5,10,30)

MLE_1 <- c(res11[1], res21[1], res31[1], res51[1], res101[1], res301[1])
MLE_5 <- c(res15[1], res25[1], res35[1], res55[1], res105[1], res305[1])
MLE_50 <- c(res150[1], res250[1], res350[1], res550[1], res1050[1], res3050[1])
MLE_100 <- c(res1100[1], res2100[1], res3100[1], res5100[1], res10100[1], res30100[1])

MOM_1 <- c(res11[2], res21[2], res31[2], res51[2], res101[2], res301[2])
MOM_5 <- c(res15[2], res25[2], res35[2], res55[2], res105[2], res305[2])
MOM_50 <- c(res150[2], res250[2], res350[2], res550[2], res1050[2], res3050[2])
MOM_100 <- c(res1100[2], res2100[2], res3100[2], res5100[2], res10100[2], res30100[2])
```

```
plot(n, MLE_1, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 1")
lines(n, MOM_1, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')


plot(n, MLE_5, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 5")
lines(n, MOM_5, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')


plot(n, MLE_50, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 50")
lines(n, MOM_50, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')


plot(n, MLE_100, xlab="Sample Size", ylab="MSE", type="b", col="red", main="Theta = 100")
lines(n, MOM_100, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

thetas <- c(1,5,50,100)
MLE_t_1 = c(res11[1], res15[1], res150[1], res1100[1])
MLE_t_2 = c(res21[1], res25[1], res250[1], res2100[1])
MLE_t_3 = c(res31[1], res35[1], res350[1], res3100[1])
MLE_t_5 = c(res51[1], res55[1], res550[1], res5100[1])
MLE_t_10 = c(res101[1], res105[1], res1050[1], res10100[1])
MLE_t_30 = c(res301[1], res305[1], res3050[1], res30100[1])

MOM_t_1 = c(res11[2], res15[2], res150[2], res1100[2])
MOM_t_2 = c(res21[2], res25[2], res250[2], res2100[2])
MOM_t_3 = c(res31[2], res35[2], res350[2], res3100[2])
MOM_t_5 = c(res51[2], res55[2], res550[2], res5100[2])
MOM_t_10 = c(res101[2], res105[2], res1050[2], res10100[2])
MOM_t_30 = c(res301[2], res305[2], res3050[2], res30100[2])

plot(thetas, MLE_t_1, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 1")
lines(thetas, MOM_t_1, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')
```

```
plot(thetas, MLE_t_2, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 2")
lines(thetas, MOM_t_2, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_3, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 3")
lines(thetas, MOM_t_3, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_5, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 5")
lines(thetas, MOM_t_5, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_10, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n = 10")
lines(thetas, MOM_t_10, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')

plot(thetas, MLE_t_30, xlab="Sample Size", ylab="MSE", type="b", col="red", main="n =30")
lines(thetas, MOM_t_30, type="b", col="blue")
legend("topright", legend=c("MLE", "MOM"), col=c('red', 'blue'), text.col = c('black', 'black'), lty=1,
pch=1, inset=0.01, ncol=1, cex=0.6, bty='n')
```

Question 2b:

```
n <- 5
x <- c(21.72, 14.65, 50.42, 28.78, 11.23)
mle <- n / sum(log(x))
mle
```

Question 2c:

```
log_likelihood <- function(par, dat){
  result <- -1 * sum(log(par/dat^(par+1)))
  return (result)
}

MLE <- optim(par = 1, fn = log_likelihood, method="BFGS", hessian = TRUE, dat = x)
MLE
```

Question 2d:

```
se <- sqrt(diag(solve(MLE$hessian)))
se
```

```
CI_lower <- MLE$par - qnorm(1-(0.05/2)) * se
CI_upper <- MLE$par + qnorm(1-(0.05/2)) * se
CI_lower
CI_upper
```