Solution’’

1a

Assume the service times follow an exponential distribution with a
rate of 3 people helped per hour.

3 people/hour

1 hour = 60min so each people will take 20min,

Average = 20min

Let X = amount of time (in minutes) sally will wait in line The time is known to have an exponential distribution with the average amount of time equal to four minutes.
X is a continuous random variable since time is measured.
m = 1/20 = 0.05
here standard deviation, σ, is the same as the mean. $\mu = \sigma$
distribution notation is $X \sim Exp(m)$. Therefore, $X \sim Exp(0.05)$
probability density function is $f(x) = me^{-mx}$. The number $e = 2.718$

Here Sally has to wait for 1/3 i.e. 0.33

1/lambda time

$$\int_a^b u\,dv = u.v\Big|_a^b - \int_a^b v.du$$

$$X \sim ExP(\lambda)\ f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x \geq 0 \\ 0, & \text{for } x = 0 \end{cases}$$

$$= \lambda e^{-\lambda x} \quad \text{where } x \geq 0$$

$$E(x) = \int_{-\infty}^{\infty} x . f(x)\,dx = \int_0^{\infty} x . \lambda e^{-\lambda x}\,dx$$

let $\boxed{u = x\lambda}$, $dv = e^{-\lambda x}\,dx$

$du = \lambda dx$, $\boxed{v = e^{-\lambda x}/\lambda}$

$$= \frac{-x\cancel{\lambda}e^{-\lambda x}}{\cancel{\lambda}}\Big|_0^{\infty} - \int_0^{\infty} \frac{-e^{-\lambda x}}{\cancel{\lambda}}\cancel{\lambda}\,dx$$

$$= \left[\frac{-x}{e^{\lambda x}}\Big|_0^{\infty}\right] + \int_0^{\infty} + e^{-\lambda x}\,dx = \boxed{\frac{-x}{e^{\lambda x}}\Big|_0^{\infty} - \frac{1}{\lambda e^{\lambda x}}\Big|_0^{\infty}} \text{ — } 1$$

So here for exponential distribution
by putting values in equ^n 1

$$= \left[\frac{-\infty}{e^{\lambda\infty}} - \frac{0}{e^{\lambda 0}}\right] - \left[\frac{1}{e^{\lambda\infty}} - \frac{1}{e^{\lambda 0}}\right]$$

$$= (0 - 0) - (0 - 1/\lambda)$$

$$= 1/\lambda \quad (\text{i.e. expected value of } x)$$

1B

$$Var(X) = E[X^2] - E[X]^2$$

$$\frac{d}{dx}\left(E(x^2) - E(x)^2\right) = \frac{-K(x^2) + K(x)\,E(x) + E(x^2) - E(x)^2}{x}$$

The variation around this estimate = 1/(3^2) = 0.111

1c

```r
set.seed(200)
power3 <- rexp(n=100000, rate =3 )
mean(power3)
var(power3)
```

0.334332405775988

0.11133097062678

Here the mean is 0.33 and Variance is 0.11 which we already got from the calculation in 1b

1d

Suppose Sally has been at the DMV for 10 minutes and has not been helped

From previous answers **1a** we know that Sally is already waiting 20mins in an average so mostly sally has to wait another 10mins.

2a

```r
for (i in 1:10000){
    x <- rnorm(100, mean=125, sd=8)
    mean(x)
    sd(x)
}
```

```r
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  110.0   119.2   125.9   125.6   130.9   144.7 
```

```r
mean(x)
```

125.621549865281

```r
sd(x)
```

7.43327164599275

```r
set.seed(500)
for (i in 1:10000){
    expo <- rexp(100,rate = 1.5)
    expo_m <- mean(expo)
    expo_sd <- sd(expo)
    }
```

```r
expo_m
```

0.707741241314281

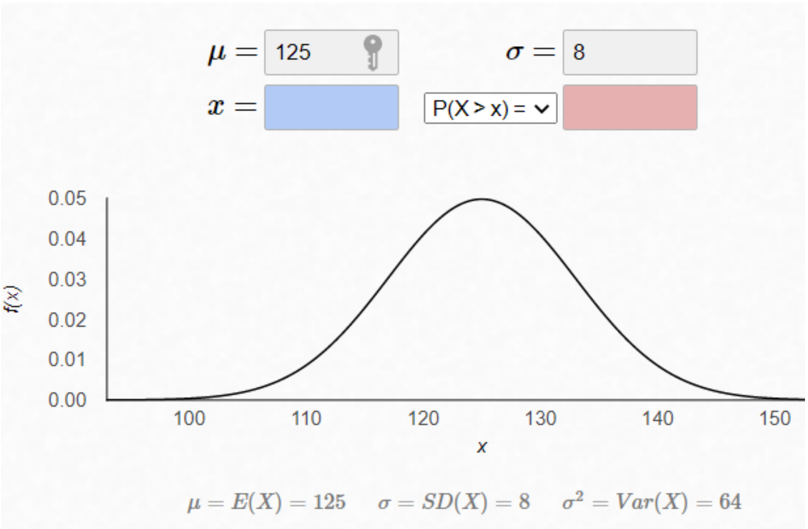
```r
summary(expo)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.02151 0.20286 0.43574 0.70774 1.11050 3.92578 
```

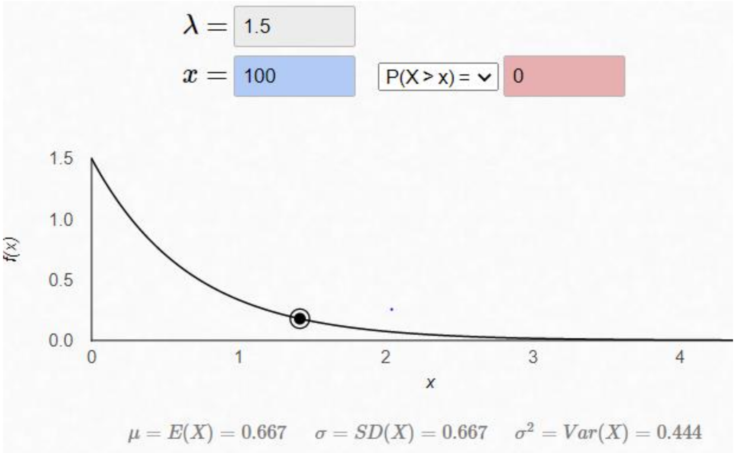
```r
expo_sd
```

0.716393634637044

2b

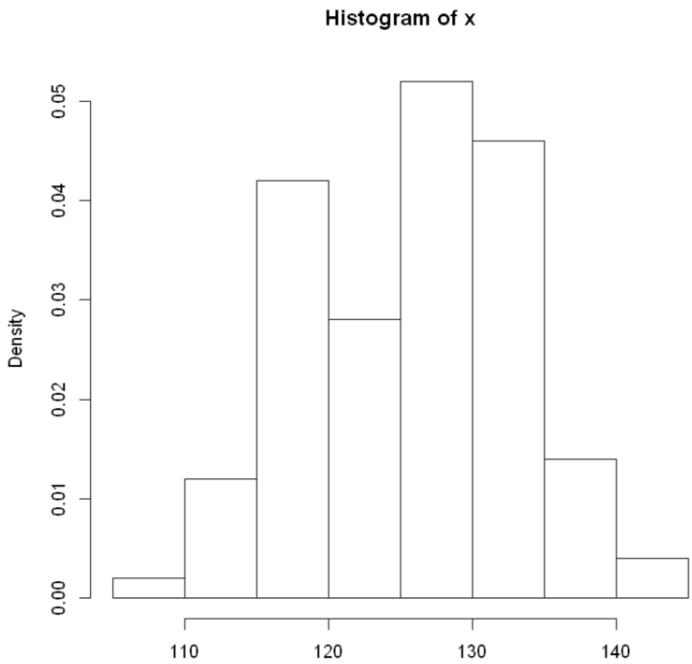


The standard deviation we got using function and manual calculator is having very minimal difference of 0.6 only



The standard deviation we got using function and manual calculator is having very minimal difference of 0.05 only

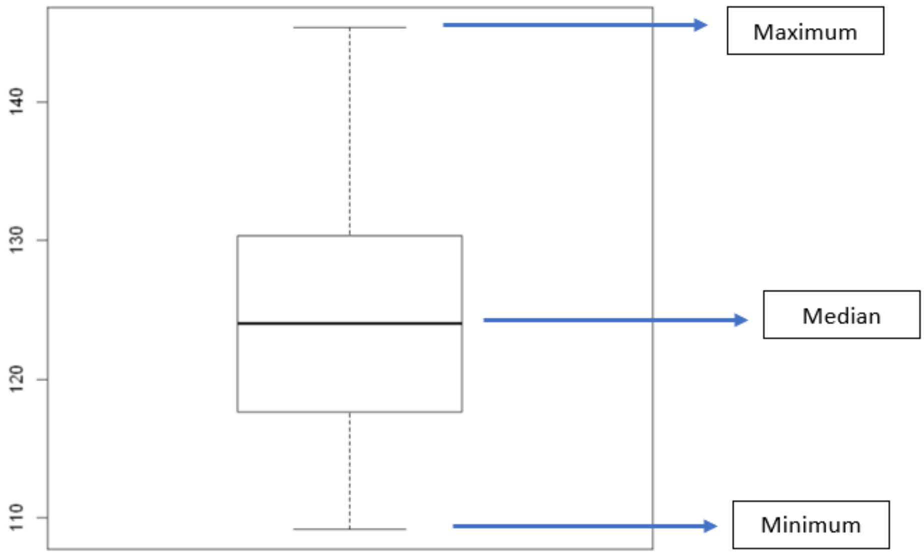
2c



Here the linear change can be seen based on the sample number. The normal distribution which starts from nearly 0.00 goes up to 0.05 and then again linearly going back nearly to 0.00.

The bell curve is clearly visible with mean of 125.62

```
boxplot(x)
```

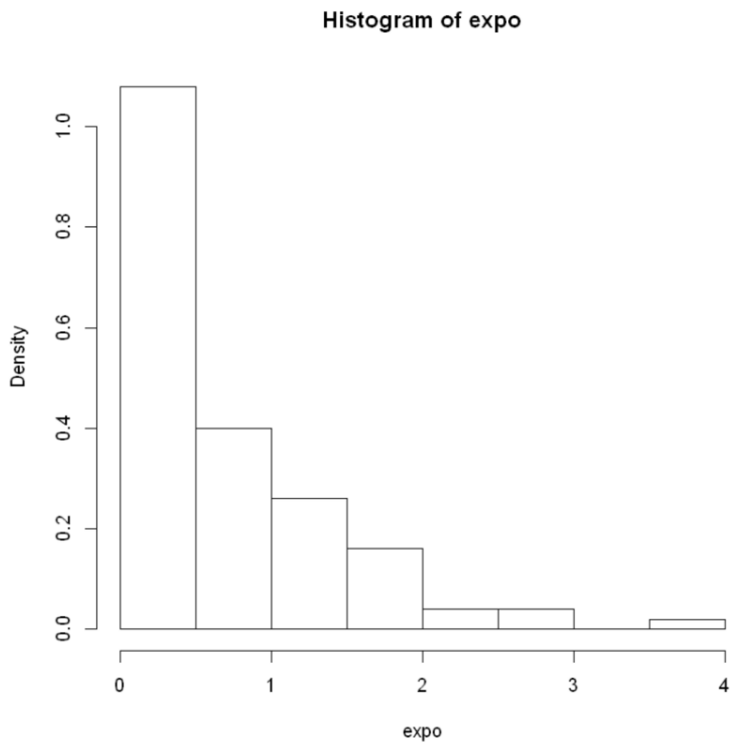


The deep black line shows the median i.e. 125.62

The box shows 25th and 75th percentile which is called Interquartile range that varies from 115 to 130 in this graph.

Then the line with bar show Minimum and maximum value i.e. Q1- 1.5IQR(min) and Q3+1.5IQR (max), here its 110 – 150

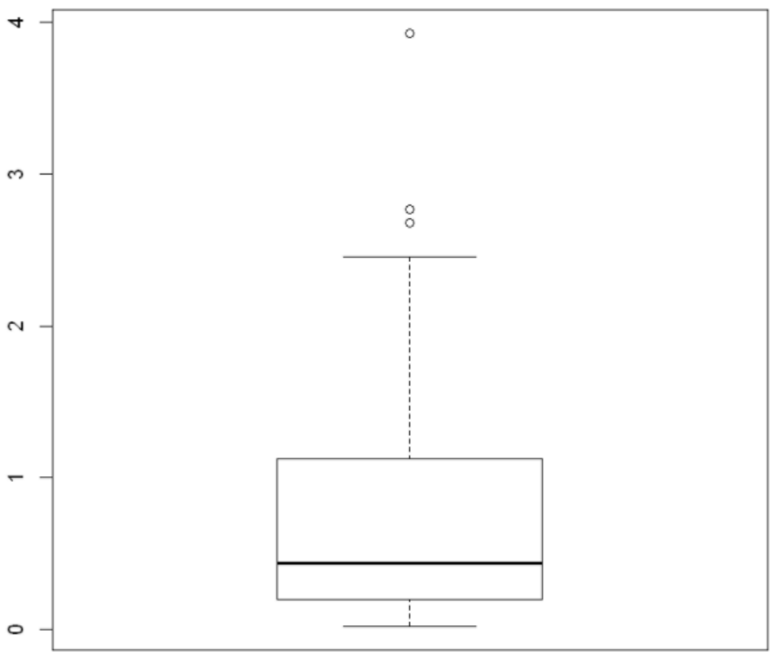
Here we don’t see any outliar.



Here the data are right-skewed, that means the mean is typically GREATER THAN the median.

The graph is linear decreasing from 1.0 to 0.0 from 0 to 4 range

```
boxplot(expo)
```



The deep black line shows the median i.e. 0.43

The box shows 25th and 75th percentile which is called Interquartile range that varies from 0.1 to 1.1 in this graph.

Then the line with bar show Minimum and maximum value i.e. Q1- 1.5IQR(min) and Q3+1.5IQR (max), here its 0.1 – 2.8

3 ouliars from 2.8,2.9 and 3.8

3a.

For a population that is normally distributed with mean 40 and s
median, and variance. Use 1,000 simulation iterations that each

```
for (i in 1:1000){
    sample = rnorm(10, mean=40, sd=10)
    sample_m <- mean(sample)
    sample_median <- median(sample)
    sample_var <- var(sample)
    }
```

```
sample_var
```

149.985688189827

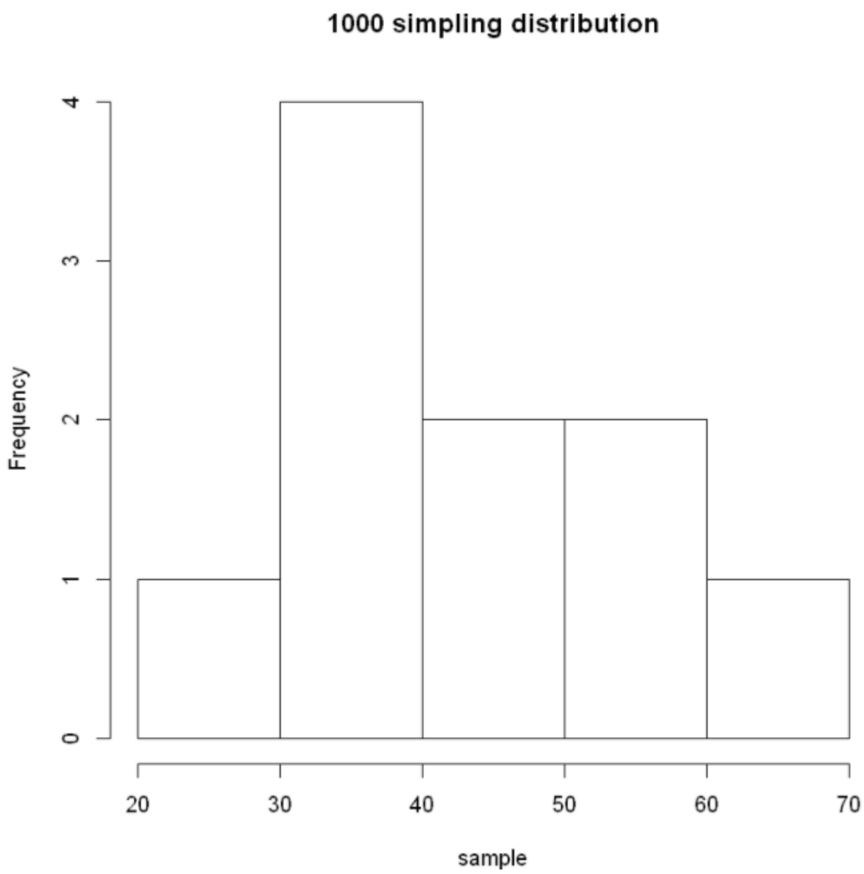
```
sample_m
```

44.1769840809494

```
summary(sample)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 29.29   34.82   40.02   44.18   54.38   66.48 
```

```
# Plot the graph.
hist(sample, main = "1000 simpling distribution")
```



For sample 20- 30 and 60-70 the frequency distribution is same.

In case of 30-40 and 40-60 it varies from 4 to 2.1

The normal distributed increase and decrease towards right -skewed, indicates the mean is typically GREATER THAN the median

3b. Based on theory, what is the distribution of the sample mean and sample median in this case (e.g., uniform, exponential, gamma, normal, etc.)?

It is somehow normally distributed.

mean: $\mu = m$

standard deviation: $\sigma = \frac{s}{\sqrt{n}}$

So function will be $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

approximately 68% are in the interval [μ-σ,μ+σ]

approximately 95% are in the interval [μ-2σ,μ+2σ]

almost all are in the interval [μ-3σ,μ+3σ]

Sample variance is 149.985688189827

Sample mean is 44.1769840809494

3c

Population is normally distributed, the sampling distribution of $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, n - 1 degrees of freedom i.e. 10-1 = 9 degrees of freedom
the sample mean is used in place of μ. This is the sum of n chi-square random variables but they are dependent due to the use of the sample mean in place of μ.

Proof,
$Z_i, i = 1, 2, \ldots, k$ are independent identically distributed N(0,1)random variables $\sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$.
By using Cochran's theorem
if we knew the population mean, and estimated the variance about it (rather than about the sample mean): $s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$
then $s_0^2/\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} Z_i^2, (Z_i = (X_i - \mu)/\sigma)$

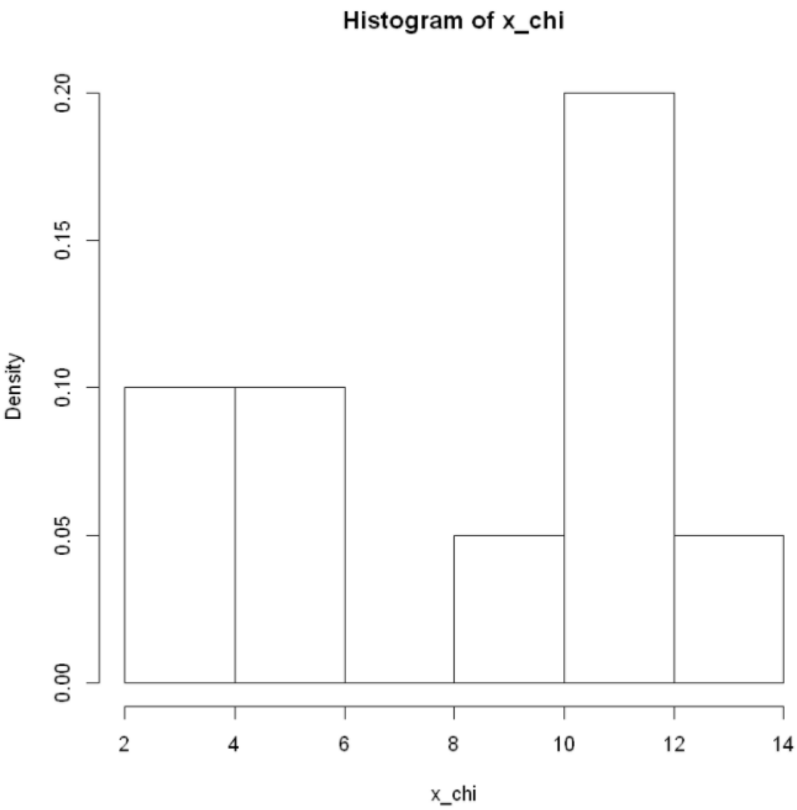
which will be 1/n times a χn2 random variable.

Here sample mean is used, instead of the population mean $(Z_i^* = (X_i - \bar{X})/\sigma)$ makes the sum of squares of deviations smaller $\sum_{i=1}^{n}(Z_i^*)^2 \sim \chi_{n-1}^2$

By using the theorem $ns_0^2/\sigma^2 \sim \chi_n^2$

We have $(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$

```
x_chi <- rchisq(sample, df = 9)
hist(x_chi, probability = TRUE)
```



We can see a normal change between 8 to 14 which shows the normal distribution of max density <= 0.20 where as for sample of 2 -6, there is a unformal density of 0.10

Here we got x_chi values as,

11.382377179703 8.12486295224 11.6835412752138 12.2418396841047 3.4731996921454 4.93478055201994 3.71376801538658 10.1218602602334 10.118685006315 5.90692569958763

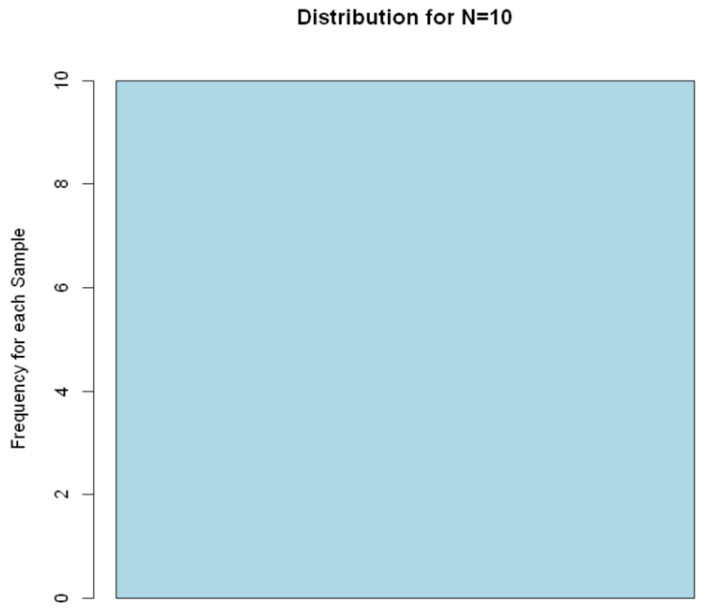
```
chisq.test(x_chi)
```

```
	Chi-squared test for given probabilities

data:  x_chi
X-squared = 12.773, df = 9, p-value = 0.1731
```

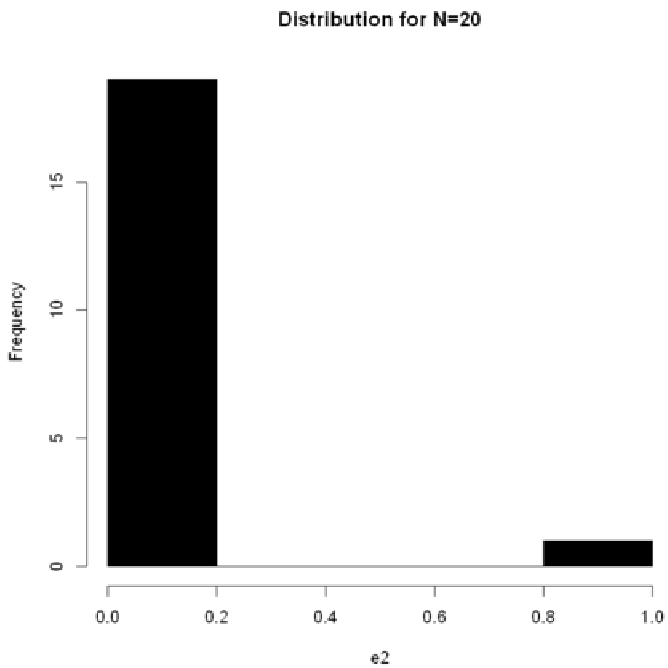
4 a,b and d

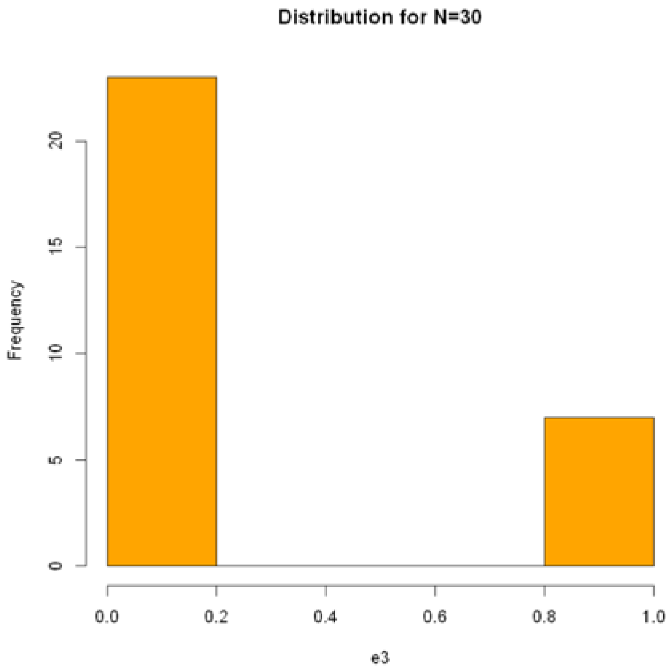
```
n <- 10 # individual sample size
N <- 500 #No of simulations/sample size
Mat <- 5 # number of columns in the matrices
mean_v <- matrix(NA, N, Mat)
std <- matrix(NA, N, Mat)
iter <- 0
for (n in seq(10, 50, 10)) {
  i <- i + 1
  for (i in 1:N) {
    expt <- rbinom(n, 1, 0.15)
    mean_v[i] <- mean(expt)
    std[i] <- sd(expt)
    print(mean_v[i])

  }
}
```

```
[1] 0.1
[1] 0.1
[1] 0
```



For sample N = 10 the distribution is uniform and having mean as 0

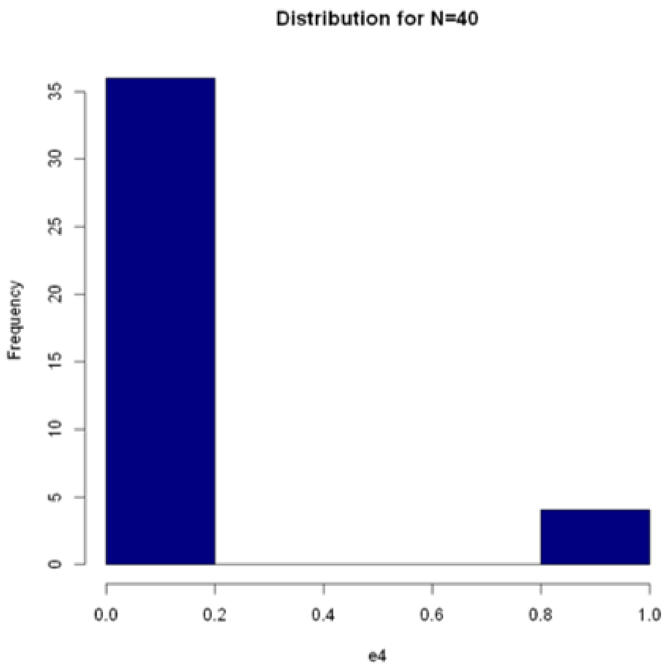


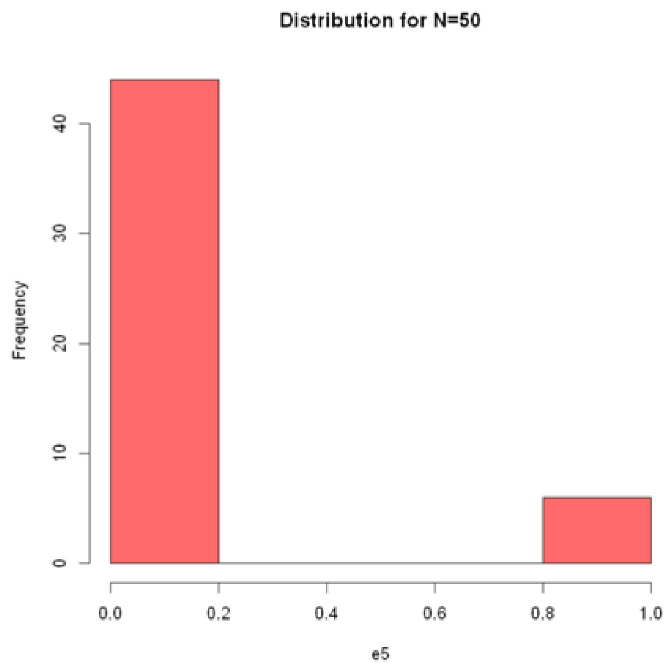


For N= 20 the frequency increases from 0.0 to 0.2 with highest of >15

For N= 30 the frequency increases from 0.0 to 0.2 with highest of >20

The normal distributed increase and decrease towards right -skewed, indicates the mean is typically GREATER THAN the median





For N= 40 the frequency increases from 0.0 to 0.2 with highest of >35

For N= 50 the frequency increases from 0.0 to 0.2 with highest of >40

The normal distributed increase and decrease towards right -skewed, indicates the mean is typically GREATER THAN the median

4c

mean and standard deviation associated with each of the five sets of ¯ x values.

```
mean_v <- matrix(NA, N, Mat)
std <- matrix(NA, N, Mat)
iter <- 0
for (n in seq(10, 50, 10)) {
  i <- i + 1
  for (i in 1:N) {
    expt <- rbinom(n, 1, 0.15)
    mean_v[i] <- mean(expt)
    mean_all5_set <- mean(mean_v[i])
    std[i] <- sd(expt)
    std_allset <- sd(std[i])
    #print(mean_v[i])

  }
}

```

```
mean_all5_set
```

0.12

```
std[i]
```

0.350509832753866

But in general if we calculate 1st 5 elements of each sample size then mean will be 0.2 with standard deviation of 1

4e

From all sample sizes of N from 10 to 50, I figured The normal distribution increase and decrease towards right -skewed, indicates the mean is typically GREATER THAN the median but clearly may be with increasing size of N, it might reach to a normal distribution.

Exercise 5:

```
set.seed(200)
for (i in 1:500){
    rc1 <- rcauchy(n = 10)
    rc1_m <- mean(rc1)
    rc1_sd <- sd(rc1)
    }
print(rc1)
print(rc1_m)
print(rc1_sd)
```

```
1]    1.07368008   0.93974256   0.35538983   0.06431938  -1.58082644
6]   -0.90274178 -17.18311407   1.18200674  -3.21658988   0.39978741
| -1.886835
| 5.54819
```

```
set.seed(200)
for (i in 1:500){
    rc5 <- rcauchy(n = 50)
    rc5_m <- mean(rc5)
    rc5_sd <- sd(rc5)
    }
print(rc5)
print(rc5_m)
print(rc5_sd)
```

```
 [1]  -1.21276139   3.08993905   0.92801139  -0.64831822  -0.52886908
 [6]   0.75069209   0.09048257  -0.43833479  -0.82618157   2.35557389
[11]  -0.53944986   0.25232688  -0.24287857   7.51824915   4.11764297
[16]   0.23595263 -15.46702307 -15.49562218   1.89657906  -0.36122027
[21]   0.57025279  -0.68638909   0.17211479  -1.14282179   4.57021006
[26]   0.55820792   0.20048265  -0.66960840  -0.78823291   1.22082233
[31]  -0.44987245   2.92342389  -0.08422889 -31.46436426  11.71096830
[36]  -2.64202151   3.33946033   0.12235671  -1.56479500  -1.18759006
[41]  -0.56608259   0.74877076  14.13946914 -14.31351619   0.68565910
[46]   4.38180222  -0.75063181   4.29658773   1.24774743  -0.31828002
[1] -0.4053062
[1] 6.697945
```

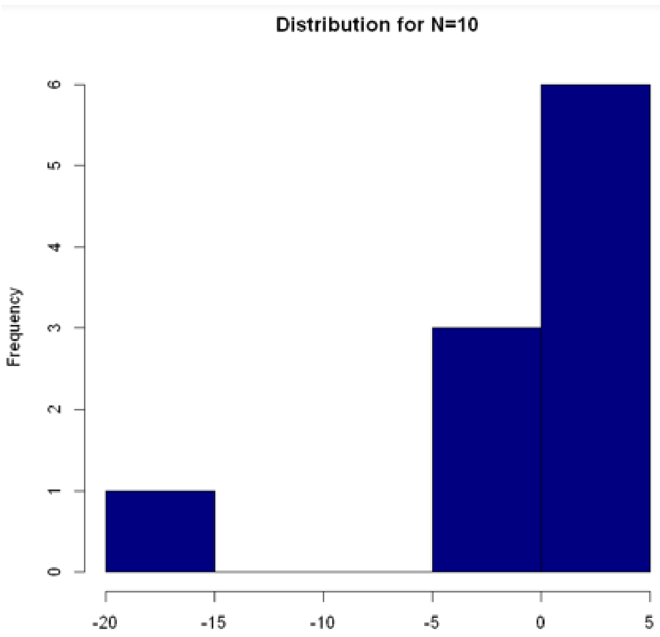
```r
set.seed(200)
for (i in 1:500){
    rc10 <- rcauchy(n = 100)
    rc10_m <- mean(rc10)
    rc10_sd <- sd(rc10)
    }
print(rc10)
print(rc10_m)
print(rc10_sd)
```

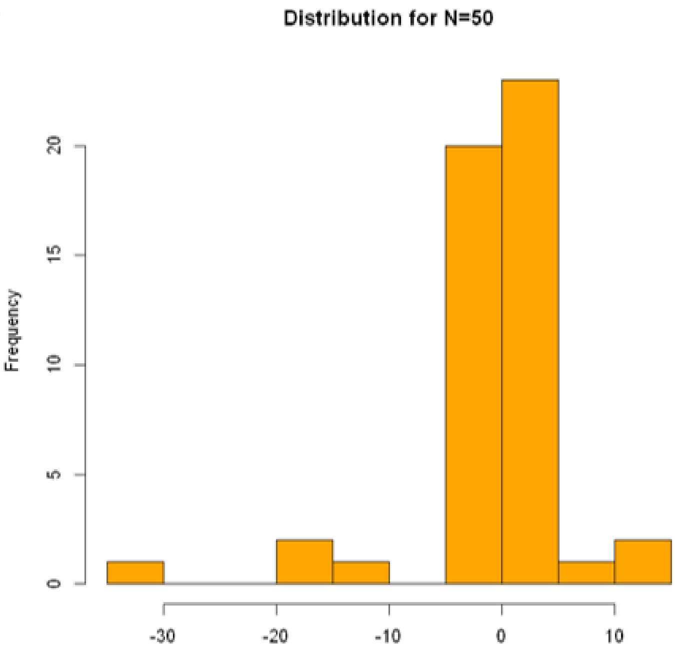
```
  [1]  -1.89937439   2.31864379   0.52128027  -0.02137322   3.89808989
  [6]   9.91005528  -0.78130051   0.45675663   0.72170565  10.88019204
 [11]   1.04991270   2.08210322   0.97042011   1.70817071   0.26399501
 [16]   0.06321905   1.13345678   6.32171846   0.90121497  -1.28016412
 [21]   0.31847897  -0.39878681  -3.40285155   0.28093668  -0.56245450
 [26]  -3.89276475   0.99691735  -7.19552143  -1.15624579  -1.96702526
 [31]  -0.33761041  -7.04775112   2.51544984  -1.31189954 260.55293177
 [36]  -0.32365465   0.03985993   2.82823444   0.36787006   6.71578037
 [41] 975.68239346  -1.70809475   0.10694933  -0.07523241  -0.12391360
 [46]   4.56379544  -0.25761176  -1.00057777  -0.37440275   0.42917523
 [51]  -1.27813836   1.17164639   0.24127961   0.04708423  -5.15376366
 [56]   0.96994915   0.37757110   1.74457266  -0.29465275  -0.39548883
 [61]   5.57125023  -0.01630676  -0.42689998   0.07802639   0.08945811
 [66]   0.46716088   0.74981193   0.21642051   0.18722396  -1.22251537
 [71]   0.41307225  -1.94935357  -0.92736298  -1.02878198   7.78509663
 [76] -35.97824774   0.27406469  -8.30443359   1.16112220  10.43208657
 [81] -23.30910172  -0.70748415   3.26396513  -3.76580078  -3.11713344
 [86]  -0.40951868   8.25859643   0.87387017   0.50379202  14.71960659
 [91]   1.90430095   0.83048646   0.75818004  -1.30310694  -6.74646269
 [96]   4.37978430   0.69103728   0.03750337  -1.18954262  -0.79435916
[1] 12.33361
[1] 100.8897
```

```r
set.seed(200)
for (i in 1:500){
    rc1k <- rcauchy(n = 1000)
    rc1k_m <- mean(rc1k)
    rc1k_sd <- sd(rc1k)
    }
print(rc1k)
print(rc1k_m)
print(rc1k_sd)
```

```
 [916]  6.378009e-02  1.057557e+00  7.848093e-01 -3.297926e+00  5.601752e-01
 [921]  3.422823e-01  2.645454e+00  2.510226e-01 -4.396887e+00 -3.310570e+00
 [926]  2.603945e-01  1.006670e+00 -9.186434e-02  1.407857e+01 -2.124655e-01
 [931]  5.252200e-01  1.201167e+00 -3.221863e-01 -7.236871e+01 -3.125235e+00
 [936]  1.483881e+00 -1.676219e-01  4.242165e+00 -2.834640e-01  3.479470e+00
 [941] -2.359142e+00 -3.921450e-01  2.524774e-01  1.037623e+00 -3.910781e-01
 [946] -4.617440e+01 -2.291347e+00 -1.737641e+00 -5.810136e-02  8.180828e+01
 [951]  1.851043e+00  1.411100e-01 -1.644872e-01 -1.511183e+00 -3.597709e+00
 [956]  2.861348e-01 -1.747105e+00  5.330847e-01  2.348741e-01  6.866454e-01
 [961] -4.598557e+00 -7.296900e+00  1.289415e+00  1.420140e+01 -2.473268e+00
 [966]  4.260242e-01  9.463109e-01 -2.598701e-01 -8.642223e+00 -4.859762e-01
 [971]  6.062226e+00 -1.278357e+00 -8.521563e-01 -1.835174e+00 -1.332874e+00
 [976]  3.563623e+03 -3.827428e-02 -1.298293e-01 -1.614157e-01  1.713963e-01
 [981]  3.378415e+00 -7.189037e-01 -3.986589e-01 -6.562718e-01  2.555141e+01
 [986]  8.102042e-02  1.118645e+00  1.146583e-01  1.354371e+00 -2.964862e-01
 [991]  1.035688e+00 -6.882426e+00  6.828994e-01  8.154515e-01  5.843174e+00
 [996]  3.200744e+00 -1.140176e+02 -3.357866e+01  3.715556e+00 -1.061284e+00

[1] -5.050868
[1] 311.6798
```

The cauchy() Calculates density, cumulative probability, quantile, and generate random sample for the Cauchy distribution (continuous)





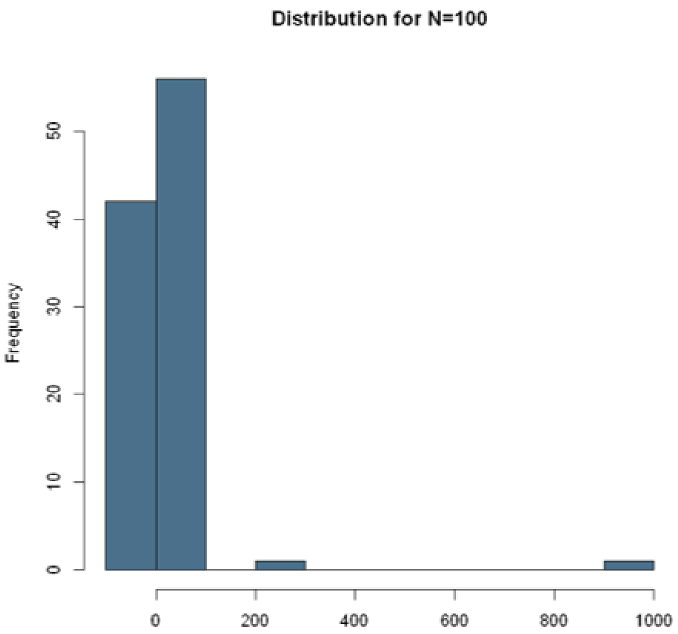
In question 4 when we calculated using rbinom() for these same N of 10 and 50

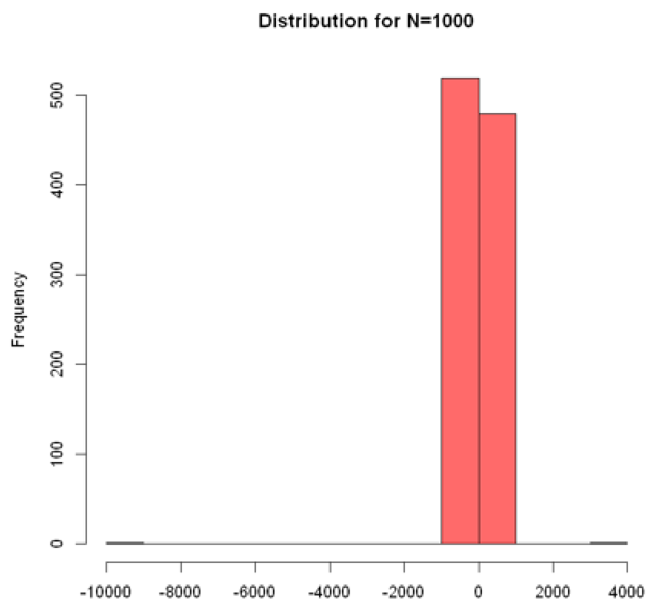
For sample N = 10 the distribution is uniform and having mean as 0

For N= 50 the frequency increases from 0.0 to 0.2 with highest of >40 where The normal distributed increase and decrease towards right -skewed, indicates the mean is typically GREATER THAN the median

In contrast,

By using Cauchy() we can see for N = 10 the data are left-skewed, then the mean is typically LESS THAN the median and in N= 50 also from 0-5 the frequency is > 20 only





For sample 100 to 1000 there is huge variation that frequency for N = 100 is >50 where as for 1000, its >500

Which is obvious as the sample size varies but in general I found Cauchy() approach is good to analyse smaller samples for example we calculated for N = 10 and 50, I believe the graphs are more intuitive compare to normal histograms using rbinom()

Cauchy distribution has no finite moments, i.e., mean, variance etc, but it can be normalized where we saw in these sample numbers where both type of skew is noticed.