Worksheet4

STAT414

2024-10-05

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

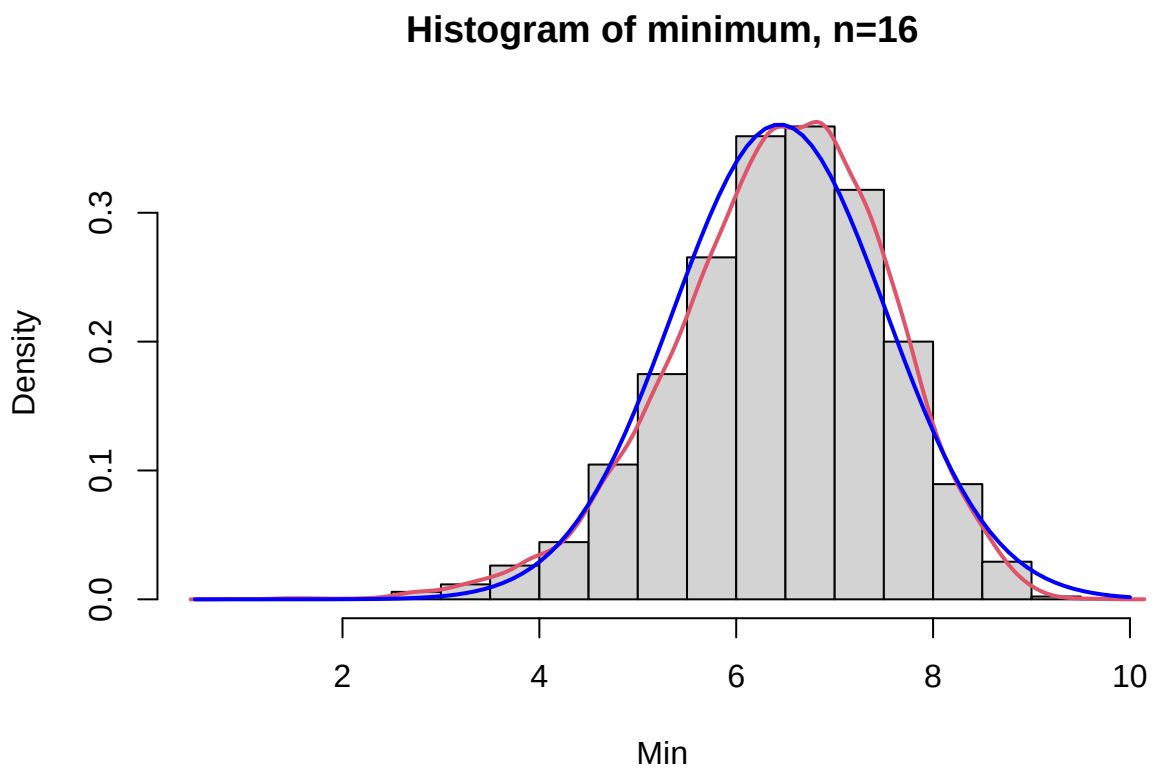
```
# #1 - In this first problem, we are drawing samples from an infinite population which is assumed
# to adequately modeled by a Normal distribution with mu = 10 and sig = 2. Objective is to
# study the sampling distribution of the widely used statistics mean(X^), median(X^), minimum
# (Min) and maximum (Max) of the observed sample.

mu = 10
sig = 2

# (a) Generate 10,000 random samples, each of size 16 from a normal population with mu=10,
# and sigma=2, and compute min, median, maximum, and mean from each sample. Store
# these results in a dataframe (10,000 rows and four columns appropriately labeled).
sample_count <- 10000
sample_size <- 16

#Create dataframe from matrix with 4 columns and 10000 entries(rows)
samplesdf <- as.data.frame(matrix(ncol=4, nrow=sample_count))
colnames(samplesdf) <- c("min", "median", "max", "mean") #name columns

#Generate 10000 entries into the dataframe, of size 16 random normal samples
#and calculate for each sample, min median max mean
set.seed(1)
for(i in 1:sample_count){
  sample <- rnorm(sample_size, mean=mu, sd=sig)
  samplesdf[i,] = c(min(sample), median(sample), max(sample), mean(sample))
}
#samplesdf
glimpse(samplesdf)
```

```
## Rows: 10,000
## Columns: 4
## $ min    <dbl> 5.570600, 6.021297, 7.245881, 7.741274, 6.390083, 6.952866, 7.4~
## $ median <dbl> 10.513151, 10.492507, 9.886882, 10.369122, 10.075447, 10.928505~
## $ max    <dbl> 13.19056, 12.71736, 12.20005, 14.80324, 14.34522, 13.17367, 13.~
## $ mean   <dbl> 10.183464, 10.282741, 10.065502, 10.660431, 9.869327, 10.743111~
```

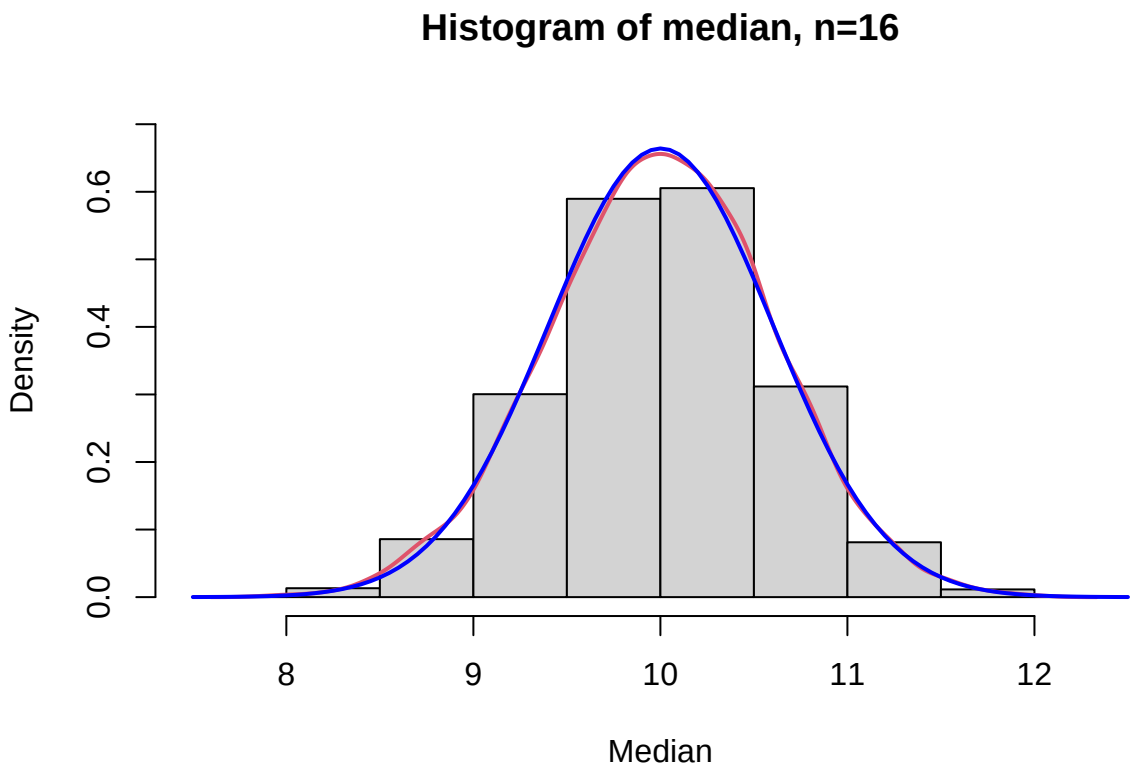
```r
# (b) Summarize the sampling distributions and plot histograms.
summary(samplesdf) #summarize the sampling distributions for each column (statistics)
```

```
##       min              median           max              mean       
##  Min.   :0.9158   Min.   : 7.853   Min.   : 9.714   Min.   : 7.964  
##  1st Qu.:5.7691   1st Qu.: 9.605   1st Qu.:12.764   1st Qu.: 9.664  
##  Median :6.5084   Median :10.010   Median :13.439   Median :10.002  
##  Mean   :6.4405   Mean   :10.002   Mean   :13.538   Mean   : 9.999  
##  3rd Qu.:7.2074   3rd Qu.:10.411   3rd Qu.:14.242   3rd Qu.:10.338  
##  Max.   :9.6866   Max.   :12.076   Max.   :18.627   Max.   :11.774
```

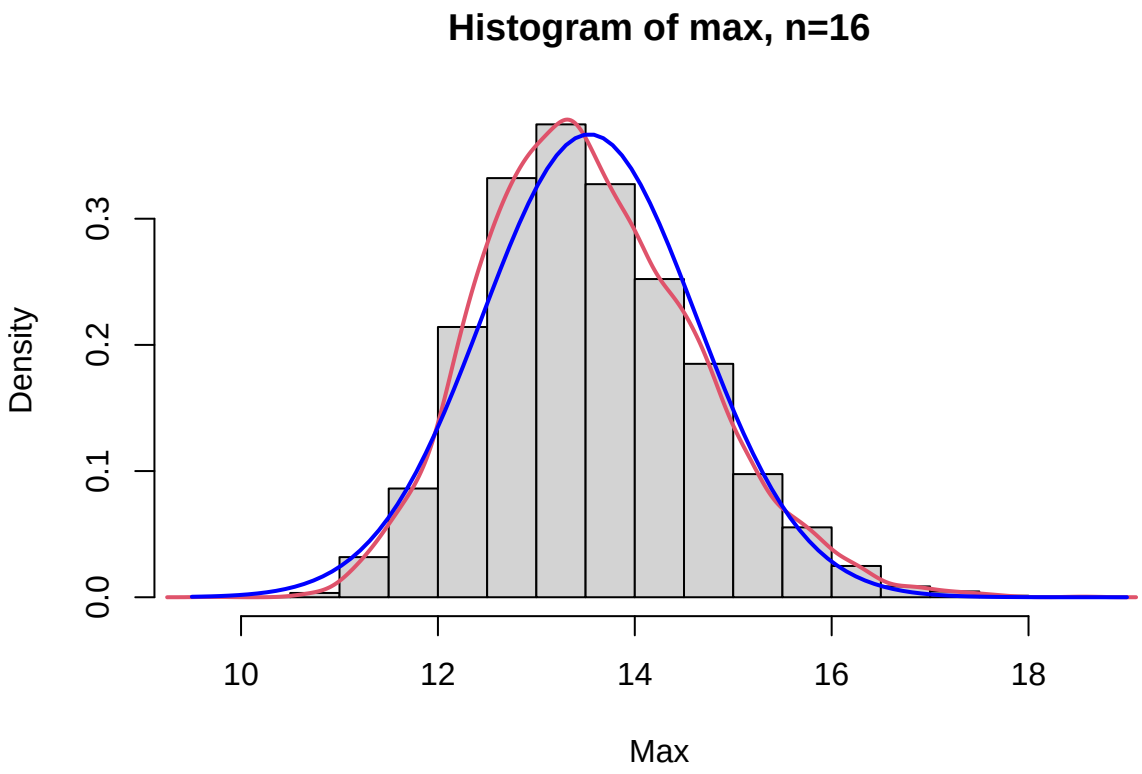
```r
#Histogram of min, curve and superimposed normal curve
hist(samplesdf$min, probability=TRUE, main = "Histogram of minimum, n=16", xlab="Min")
lines(density(samplesdf$min), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf$min), sd=sd(samplesdf$min)), lwd=2, col="blue", add=TRUE)
```



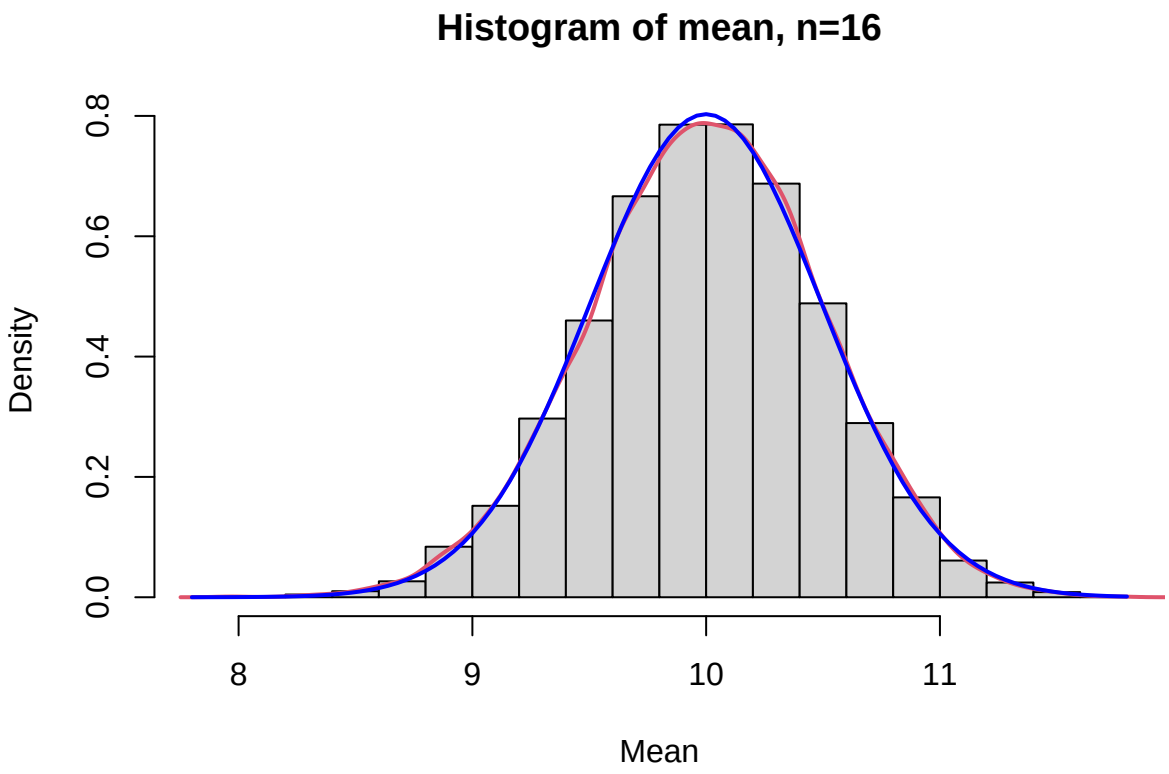
```r
#Histogram of median, curve and superimposed normal curve
hist(samplesdf$median, probability=TRUE, ylim = c(0,.7), main = "Histogram of median, n=16",xlab="Media
lines(density(samplesdf$median), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf$median), sd=sd(samplesdf$median)), lwd=2, col="blue", add=TRUE)
```




```r
#Histogram of max, curve and superimposed normal curve
hist(samplesdf$max, probability=TRUE, main = "Histogram of max, n=16", xlab="Max")
lines(density(samplesdf$max), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf$max), sd=sd(samplesdf$max)), lwd=2, col="blue", add=TRUE)
```




```
#Histogram of mean, curve and superimposed normal curve
hist(samplesdf$mean, probability=TRUE, main = "Histogram of mean, n=16", xlab="Mean")
lines(density(samplesdf$mean), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf$mean), sd=sd(samplesdf$mean)), lwd=2, col="blue", add=TRUE)
```

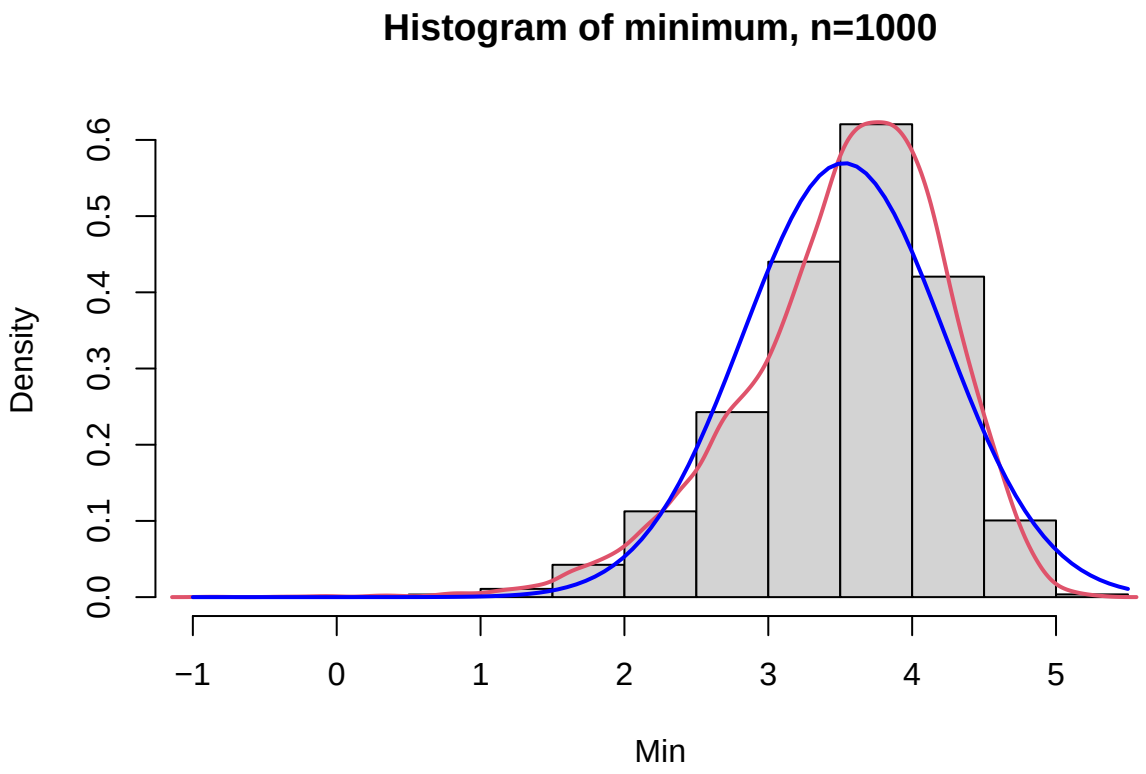



```
# (c) Compare the summary statistics and comment on the shapes of the histogram
cat("The summary statistics were already calculated in (b), refer to 1-b.
    The shapes of the mean and median histograms are very close to normal, as seen
    from the red lines which are the actual density lines of the histograms, and
    the blue lines which  are the superimposed normal lines from the dataframe.
    The histograms of max and min, are not perfectly curved and are a bit skewed,
    but still approximately normal.")
```

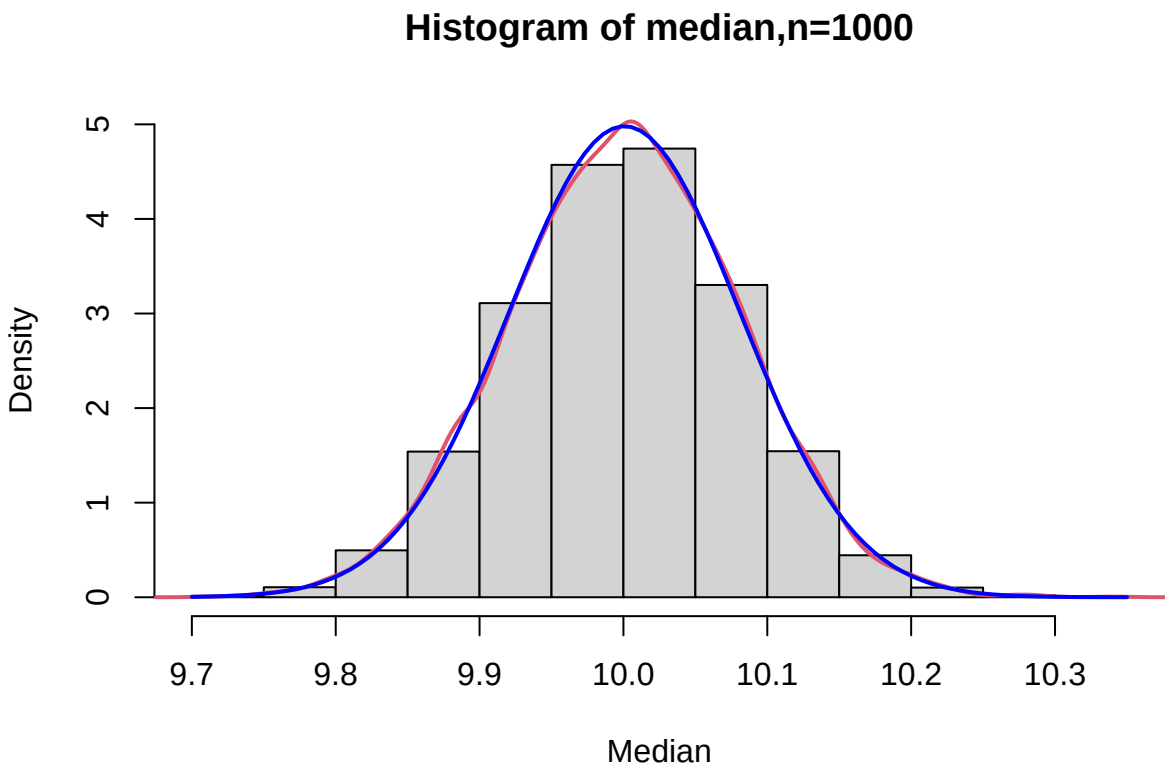
```
## The summary statistics were already calculated in (b), refer to 1-b.
##     The shapes of the mean and median histograms are very close to normal, as seen
##     from the red lines which are the actual density lines of the histograms, and
##     the blue lines which  are the superimposed normal lines from the dataframe.
##     The histograms of max and min, are not perfectly curved and are a bit skewed,
##     but still approximately normal.
```

```
# (d) Repeat (a)-(c) for a larger sample size.
sample_size <- 1000

samplesdf2 <- as.data.frame(matrix(ncol=4, nrow=sample_count))
colnames(samplesdf2) <- c("min", "median", "max", "mean") #name columns

for(i in 1:sample_count){
  sample <- rnorm(sample_size, mean=mu, sd=sig)
  samplesdf2[i,] = c(min(sample), median(sample), max(sample), mean(sample))
}
```

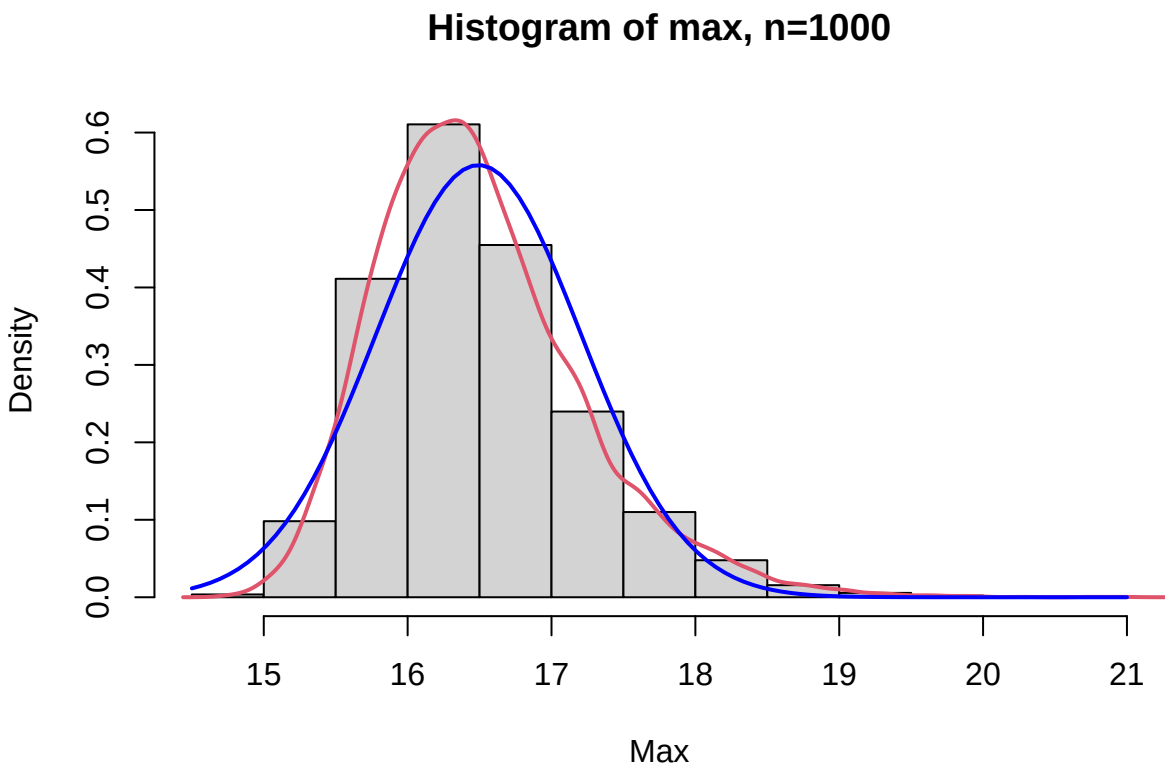
```r
#Histogram of min, curve and superimposed normal curve
hist(samplesdf2$min, probability=TRUE, main = "Histogram of minimum, n=1000", xlab="Min")
lines(density(samplesdf2$min), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf2$min), sd=sd(samplesdf2$min)), lwd=2, col="blue", add=TRUE)
```



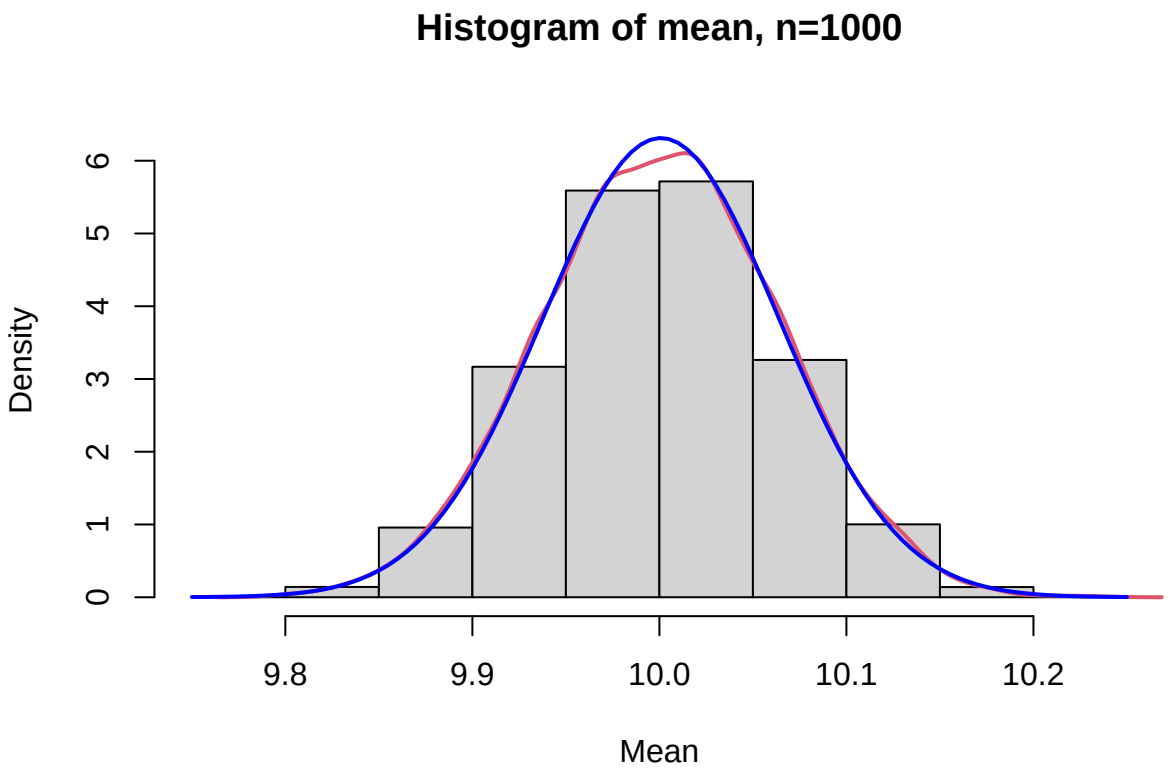

```r
#Histogram of median, curve and superimposed normal curve
hist(samplesdf2$median, probability=TRUE, ylim = c(0,5),
     main = "Histogram of median,n=1000",xlab="Median")
lines(density(samplesdf2$median), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf2$median), sd=sd(samplesdf2$median)), lwd=2, col="blue", add=TRUE)
```




```r
#Histogram of max, curve and superimposed normal curve
hist(samplesdf2$max, probability=TRUE, main = "Histogram of max, n=1000", xlab="Max")
lines(density(samplesdf2$max), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf2$max), sd=sd(samplesdf2$max)), lwd=2, col="blue", add=TRUE)
```




```r
#Histogram of mean, curve and superimposed normal curve
hist(samplesdf2$mean, probability=TRUE, ylim=c(0,6.5), main = "Histogram of mean, n=1000", xlab="Mean")
lines(density(samplesdf2$mean), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(samplesdf2$mean), sd=sd(samplesdf2$mean)), lwd=2, col="blue", add=TRUE)
```




```
summary(samplesdf2)
```

```
##       min                median           max              mean
##  Min.   :-0.8579   Min.   : 9.709   Min.   :14.73   Min.   : 9.791
##  1st Qu.: 3.1234   1st Qu.: 9.947   1st Qu.:15.99   1st Qu.: 9.957
##  Median : 3.6208   Median :10.002   Median :16.40   Median :10.001
##  Mean   : 3.5258   Mean   :10.001   Mean   :16.49   Mean   :10.001
##  3rd Qu.: 4.0209   3rd Qu.:10.055   3rd Qu.:16.89   3rd Qu.:10.044
##  Max.   : 5.2727   Max.   :10.348   Max.   :20.97   Max.   :10.242
```

```
("The shapes of the min and max histograms of n=1000 are less accurate to the
superimposed normal lines compared to the n=16 histograms. The mean and median histograms
  are still very accurate to the normal line. Intuitively, this is because if there are
  more samples in an entry, the likelihood of an extremely small or extremely large
  value increases, which causes the min and max distributions to become skewed. Min
  and max are highly influenced by extreme outliers in the data.")
```

```
## [1] "The shapes of the min and max histograms of n=1000 are less accurate to the \nsuperimposed norma
```

```
# #2 - Repeat steps of problem # for a skewed parent distribution such as a lognormal or a gamma
# distribution. Explain how would choose the parameters of these distributions so that they
# can be compared to the results from normal distribution.
library(dplyr)
cat("Choose gamma distribution for the repeated samples. As we recall from
```

```
    Worksheet3, the gamma distribution has the parameters: Shape and Scale, which
    can be calculated easily with the formulas mean=k*theta, and variance=k*theta^2,
    we know that sigma is 2 and mean is 10, therefore variance = 4. So 10 = k * theta,
    theta = 10/k. 4 = k * (10/k)^2 = k * (100/k^2) = 100/k, thus k = 100/4 = 25.
    And theta = 10/25 = .4 and k=shape and theta=scale")
```

```
## Choose gamma distribution for the repeated samples. As we recall from
##     Worksheet3, the gamma distribution has the parameters: Shape and Scale, which
##     can be calculated easily with the formulas mean=k*theta, and variance=k*theta^2,
##     we know that sigma is 2 and mean is 10, therefore variance = 4. So 10 = k * theta,
##     theta = 10/k. 4 = k * (10/k)^2 = k * (100/k^2) = 100/k, thus k = 100/4 = 25.
##     And theta = 10/25 = .4 and k=shape and theta=scale
```

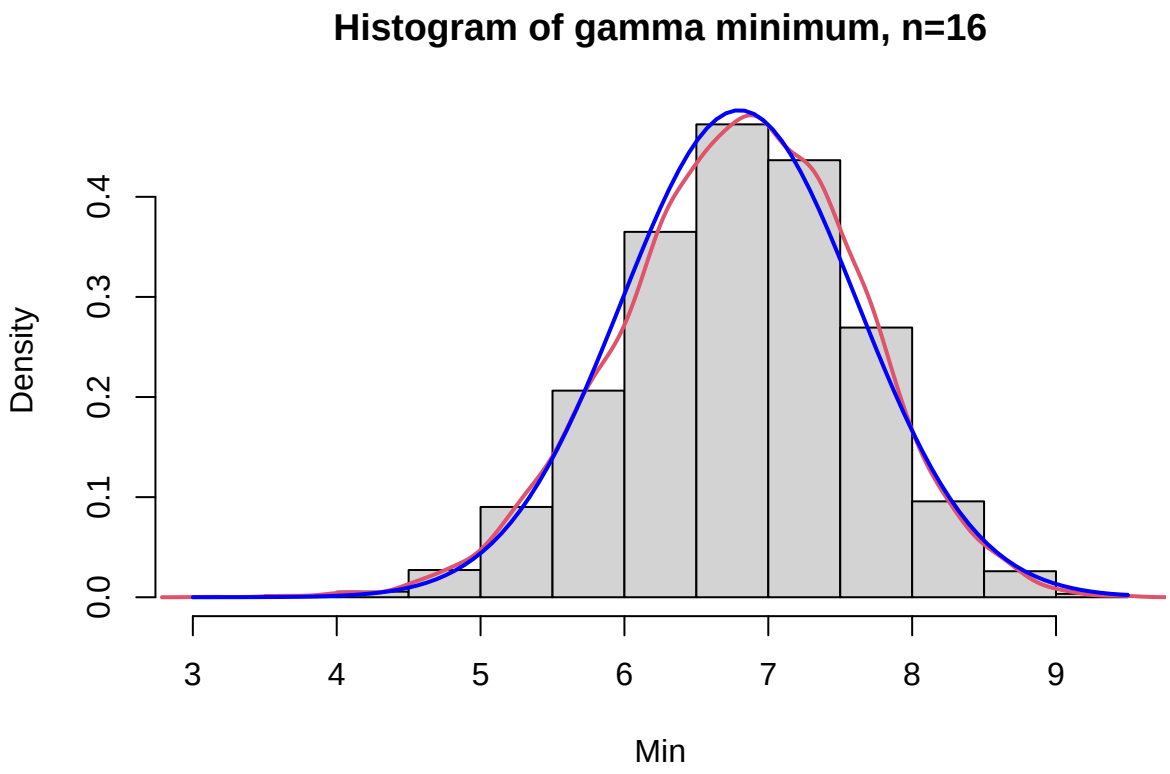
```
shape = 25
scale = 0.4
sample_count <- 10000
sample_size <- 16

gammadf <- as.data.frame(matrix(ncol=4, nrow=sample_count))
colnames(gammadf) <- c("min", "median", "max", "mean") #name columns

#Generate 10000 entries into the dataframe, of size 16 random normal samples
#and calculate for each sample, min median max mean
set.seed(1)
for(i in 1:sample_count){
  sample <- rgamma(sample_size, shape=shape, scale=scale)
  gammadf[i,] = c(min(sample), median(sample), max(sample), mean(sample))
}
glimpse(gammadf)
```

```
## Rows: 10,000
## Columns: 4
## $ min    <dbl> 6.988199, 6.257037, 6.948411, 7.181540, 7.435445, 6.755709, 7.0~
## $ median <dbl> 10.174689, 9.772867, 9.556629, 9.816832, 11.066700, 10.165113, ~
## $ max    <dbl> 12.60970, 12.09894, 11.38068, 14.57358, 13.19357, 13.61138, 12.~
## $ mean   <dbl> 10.002221, 9.913739, 9.664162, 9.907969, 10.813008, 10.155270, ~
```

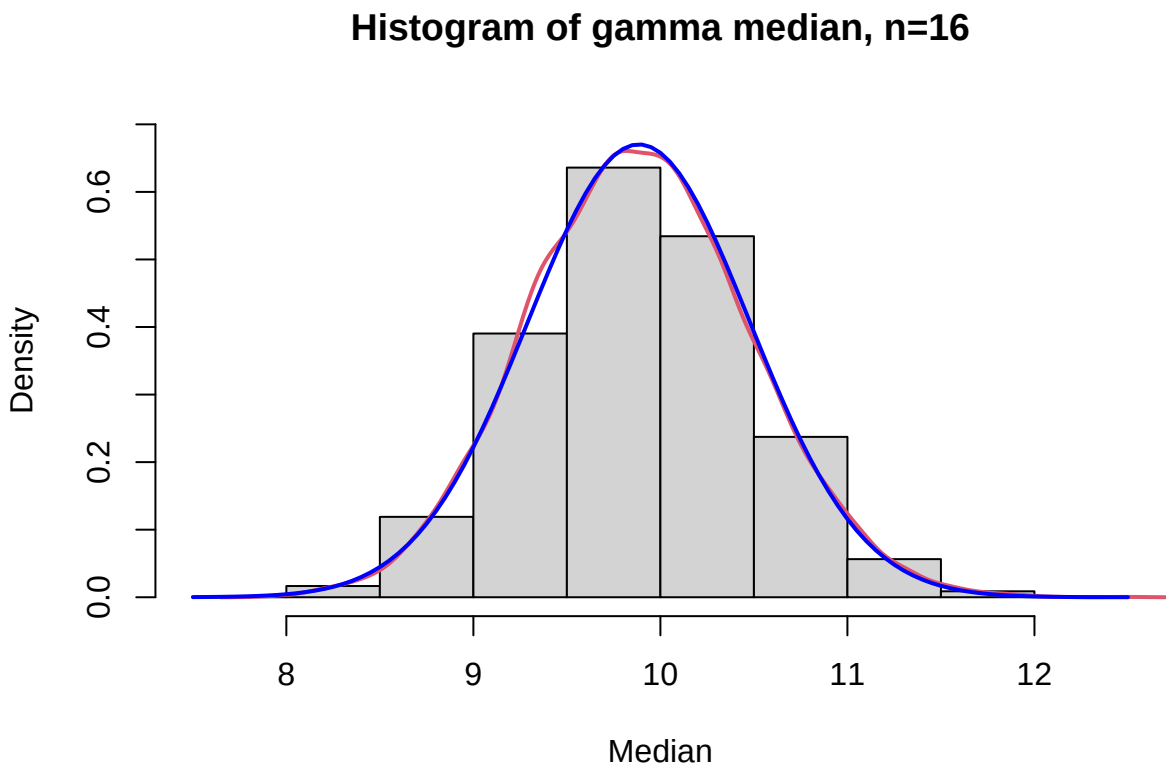
```
# (b) Summarize the sampling distributions and plot histograms.
summary(gammadf) #summarize the sampling distributions for each column (statistics)
```

```
##       min             median           max             mean      
##  Min.   :3.136   Min.   : 7.904   Min.   : 9.95   Min.   : 8.416  
##  1st Qu.:6.263   1st Qu.: 9.471   1st Qu.:12.86   1st Qu.: 9.658  
##  Median :6.827   Median : 9.877   Median :13.68   Median : 9.990  
##  Mean   :6.798   Mean   : 9.885   Mean   :13.84   Mean   :10.002  
##  3rd Qu.:7.372   3rd Qu.:10.281   3rd Qu.:14.67   3rd Qu.:10.336  
##  Max.   :9.474   Max.   :12.494   Max.   :20.74   Max.   :11.850
```

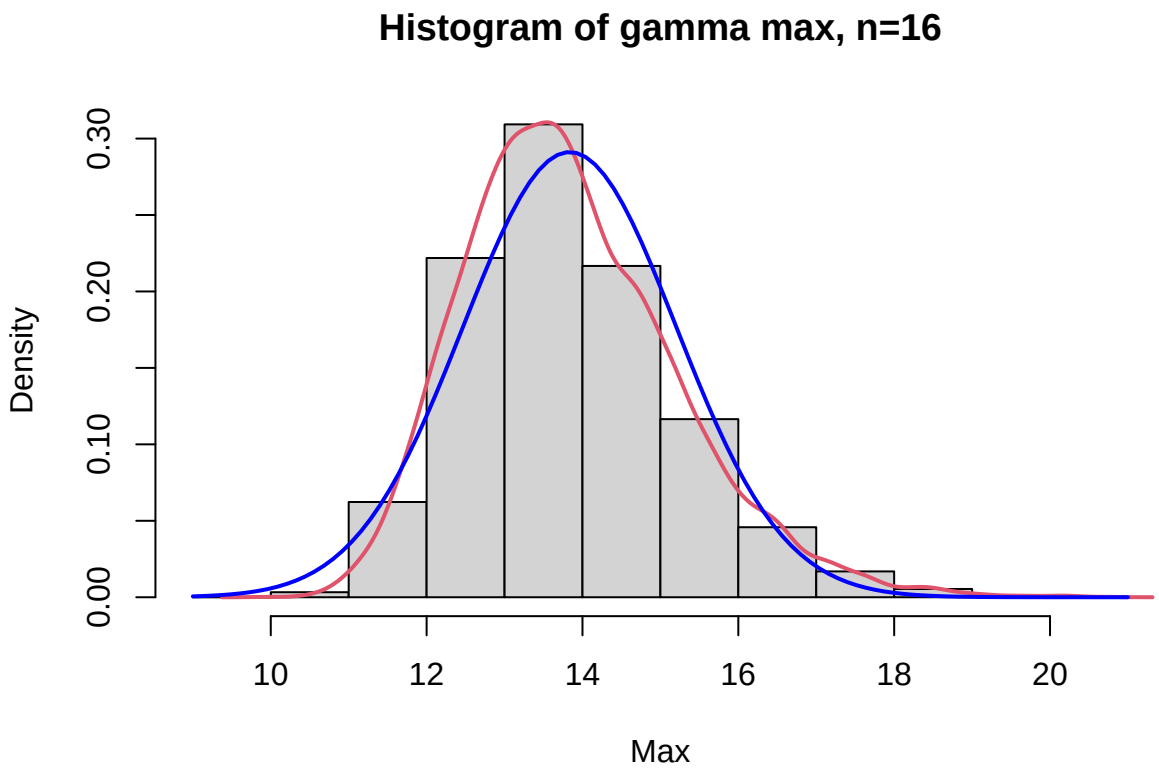
```
#Histogram of min, curve and superimposed normal curve
hist(gammadf$min, probability=TRUE, main = "Histogram of gamma minimum, n=16", xlab="Min")
lines(density(gammadf$min), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf$min), sd=sd(gammadf$min)), lwd=2, col="blue", add=TRUE)
```



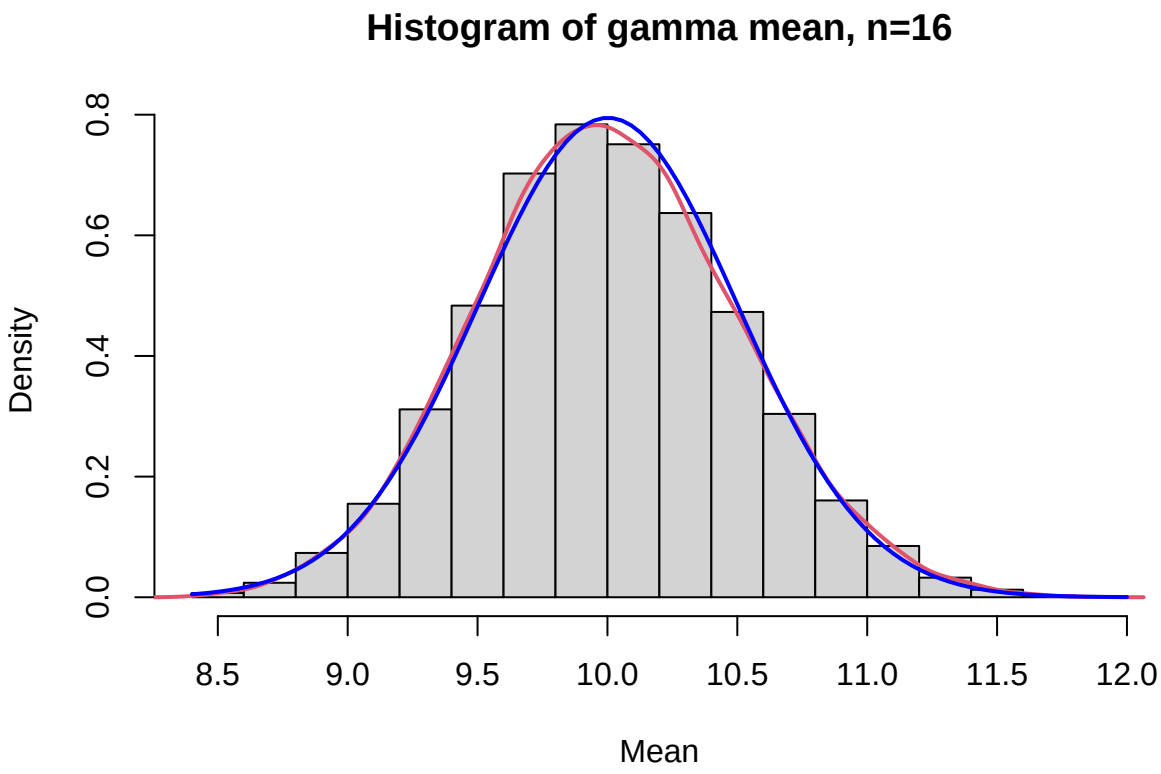

```
#Histogram of median, curve and superimposed normal curve
hist(gammadf$median, probability=TRUE, ylim = c(0,.7), main = "Histogram of gamma median, n=16",xlab="M
lines(density(gammadf$median), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf$median), sd=sd(gammadf$median)), lwd=2, col="blue", add=TRUE)
```




```
#Histogram of max, curve and superimposed normal curve
hist(gammadf$max, probability=TRUE, main = "Histogram of gamma max, n=16", xlab="Max")
lines(density(gammadf$max), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf$max), sd=sd(gammadf$max)), lwd=2, col="blue", add=TRUE)
```




```
#Histogram of mean, curve and superimposed normal curve
hist(gammadf$mean, probability=TRUE, main = "Histogram of gamma mean, n=16", xlab="Mean")
lines(density(gammadf$mean), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf$mean), sd=sd(gammadf$mean)), lwd=2, col="blue", add=TRUE)
```




```
# (c) Compare the summary statistics and comment on the shapes of the histogram
cat("The summary statistics were already calculated in (b), refer to 2-b.
    The shapes of the mean and median histograms are very close to normal, as seen
    from the red lines which are the actual density lines of the histograms, and
    the blue lines which  are the superimposed normal lines from the dataframe.
    The histograms of max and min, are not perfectly curved and are a bit skewed,
    but still approximately normal.")
```

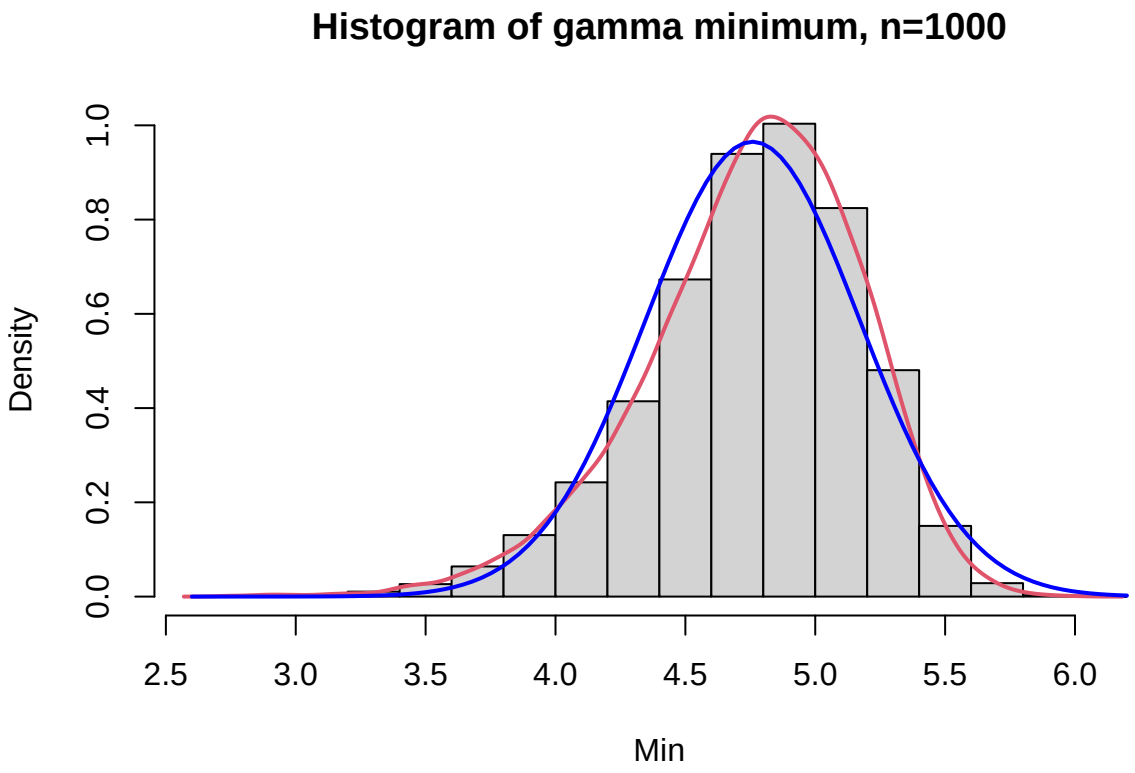
```
## The summary statistics were already calculated in (b), refer to 2-b.
##     The shapes of the mean and median histograms are very close to normal, as seen
##     from the red lines which are the actual density lines of the histograms, and
##     the blue lines which  are the superimposed normal lines from the dataframe.
##     The histograms of max and min, are not perfectly curved and are a bit skewed,
##     but still approximately normal.
```

```
sample_size <- 1000
gammadf2 <- as.data.frame(matrix(ncol=4, nrow=sample_count))
colnames(gammadf2) <- c("min", "median", "max", "mean") #name columns

#Generate 10000 entries into the dataframe, of size 16 random normal samples
#and calculate for each sample, min median max mean
set.seed(1)
for(i in 1:sample_count){
  sample <- rgamma(sample_size, shape=shape, scale=scale)
  gammadf2[i,] = c(min(sample), median(sample), max(sample), mean(sample))
```

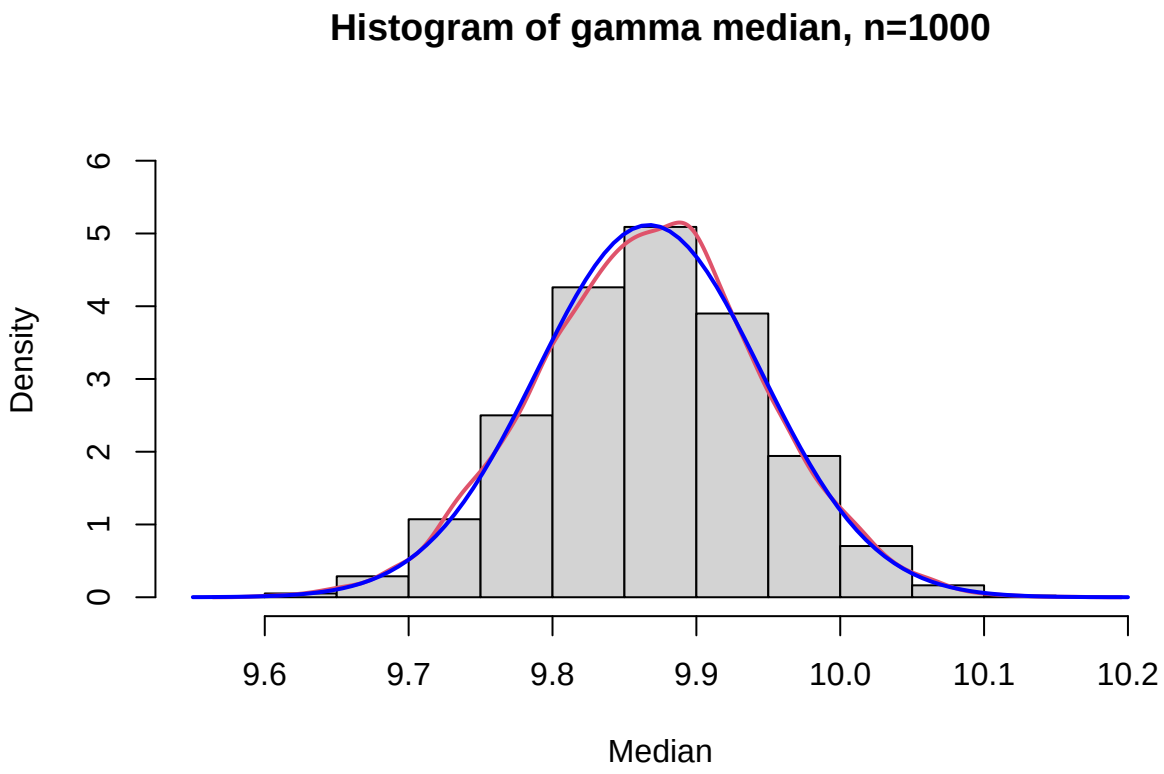
```
}
glimpse(gammadf2)
```

```
## Rows: 10,000
## Columns: 4
## $ min    <dbl> 5.160367, 4.044844, 4.477534, 4.897245, 4.503880, 4.888401, 5.0~
## $ median <dbl> 9.750309, 9.793803, 10.036507, 9.787397, 9.920642, 9.927141, 9.~
## $ max    <dbl> 18.79578, 18.24720, 16.86959, 16.92719, 18.28947, 17.06654, 17.~
## $ mean   <dbl> 9.939007, 9.987813, 10.109738, 9.976633, 10.083951, 10.041409, ~
```

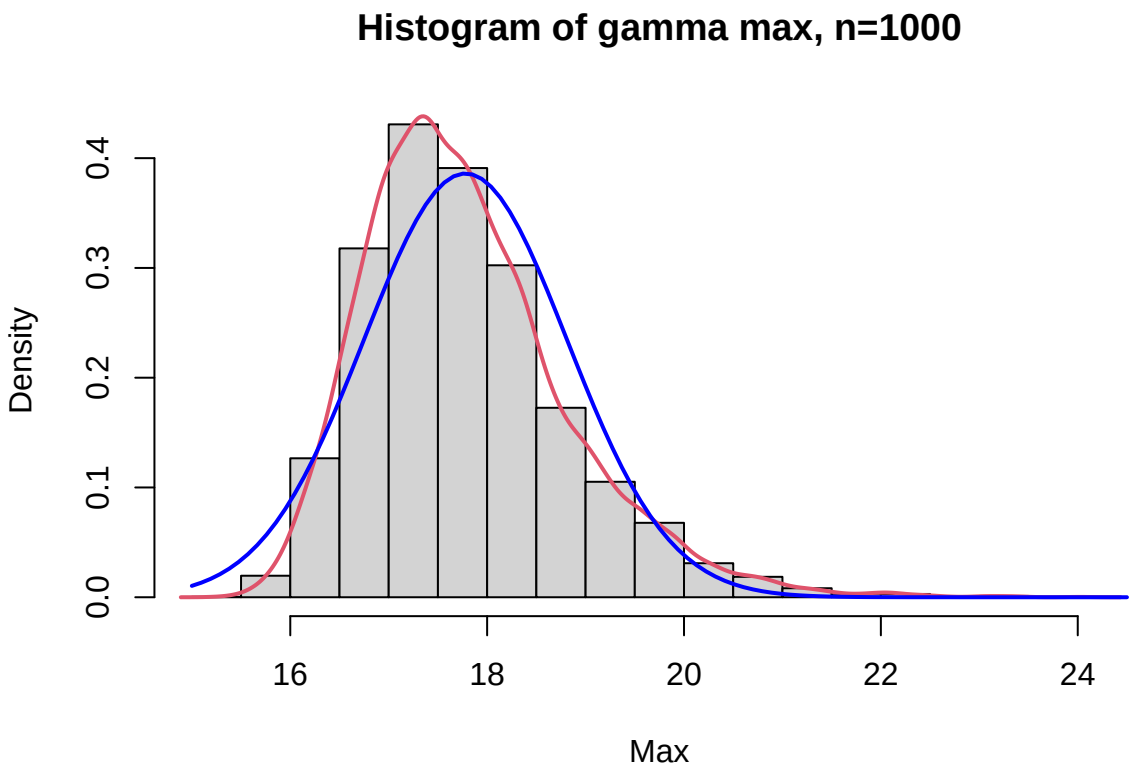
```
#Histogram of min, curve and superimposed normal curve
hist(gammadf2$min, probability=TRUE, main = "Histogram of gamma minimum, n=1000", xlab="Min")
lines(density(gammadf2$min), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf2$min), sd=sd(gammadf2$min)), lwd=2, col="blue", add=TRUE)
```



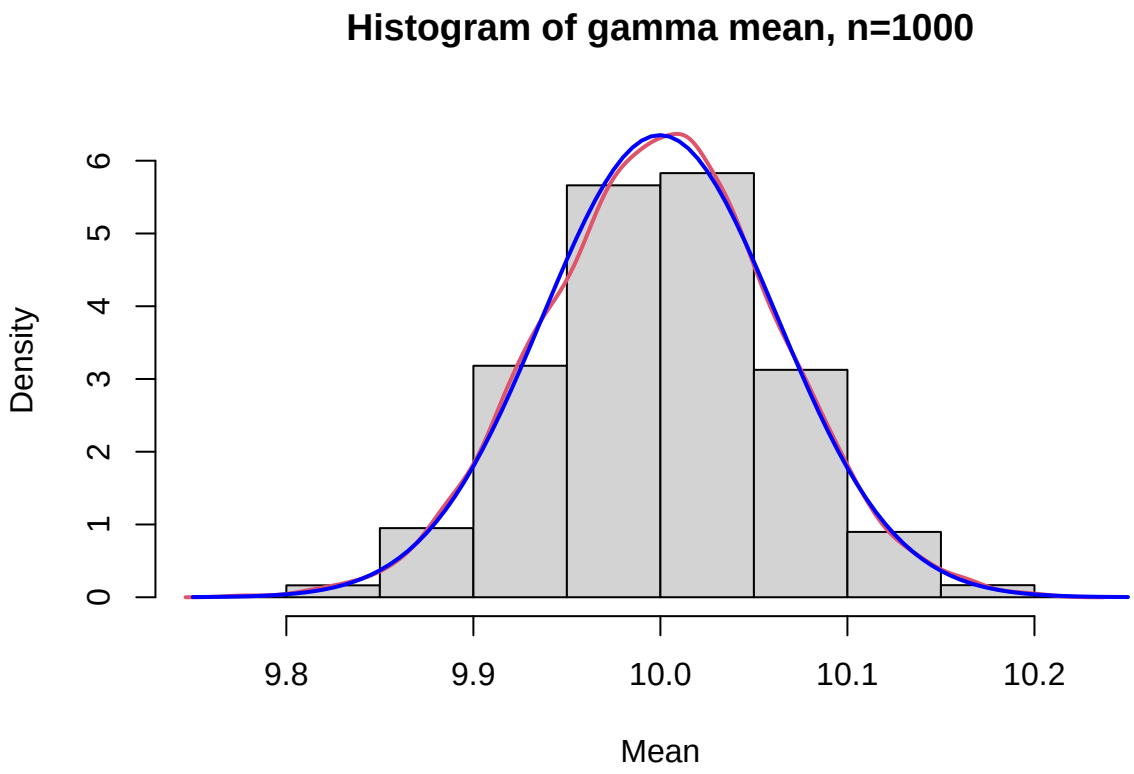

```
#Histogram of median, curve and superimposed normal curve
hist(gammadf2$median, probability=TRUE, ylim = c(0,6.5), main = "Histogram of gamma median, n=1000",xlab
lines(density(gammadf2$median), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf2$median), sd=sd(gammadf2$median)), lwd=2, col="blue", add=TRUE)
```




```
#Histogram of max, curve and superimposed normal curve
hist(gammadf2$max, probability=TRUE, main = "Histogram of gamma max, n=1000", xlab="Max")
lines(density(gammadf2$max), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf2$max), sd=sd(gammadf2$max)), lwd=2, col="blue", add=TRUE)
```




```
#Histogram of mean, curve and superimposed normal curve
hist(gammadf2$mean, probability=TRUE, main = "Histogram of gamma mean, n=1000", xlab="Mean",
     ylim=c(0,6.5))
lines(density(gammadf2$mean), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(gammadf2$mean), sd=sd(gammadf2$mean)), lwd=2, col="blue", add=TRUE)
```




```
summary(gammadf2)
```

```
##       min              median           max             mean
##  Min.   :2.741   Min.   : 9.597   Min.   :15.30   Min.   : 9.773
##  1st Qu.:4.513   1st Qu.: 9.815   1st Qu.:17.04   1st Qu.: 9.958
##  Median :4.797   Median : 9.868   Median :17.64   Median :10.000
##  Mean   :4.759   Mean   : 9.867   Mean   :17.78   Mean   :10.000
##  3rd Qu.:5.050   3rd Qu.: 9.919   3rd Qu.:18.34   3rd Qu.:10.041
##  Max.   :6.012   Max.   :10.162   Max.   :24.02   Max.   :10.210
```

```
# (c) Compare the summary statistics and comment on the shapes of the histogram
cat("The shapes of the min and max histograms of n=1000 are less accurate to the
superimposed normal lines compared to the n=16 histograms. The mean and median histograms
are still very accurate to the normal line. Intuitively, again, this is because of
the susceptibility of min and max to extreme outliers.")
```

```
## The shapes of the min and max histograms of n=1000 are less accurate to the
## superimposed normal lines compared to the n=16 histograms. The mean and median histograms
## are still very accurate to the normal line. Intuitively, again, this is because of
## the susceptibility of min and max to extreme outliers.
```