Worksheet5

STAT414

2024-10-16

```
#1 - Study Example 5.1 (204-205) from Chapter 5 of M&N. Suppose x1, x2, · · · , xm denote
# data obtained from a normally distributed population. Write down the formulas for the
# method of moments estimators (mme) of the parameters mu(mean), sig^2(variance), and sig
# (standard deviation).
```

Method of Moments estimates population parameters by calculating the samples, therefore

$$\text{MME}(\mu) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\text{MME}(\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$\text{MME}(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

```
#install.packages("EnvStats")
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
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
```

```
# (a) In EnvStats, find the tccb data set attach(EPA.94b.tccb.df) and subset the
# data corresponding to the Reference area. Plot the histogram of the logtransformed
# TcCB, and overlay a normal curve. Obtain a q-q plot. Comment on the
# normality assumption for the log-transformed data. (Please be aware that log in
# R is the natural log.)

tccb <- EPA.94b.tccb.df
head(tccb)
```

```
##   TcCB.orig TcCB Censored      Area
## 1      0.22 0.22    FALSE Reference
## 2      0.23 0.23    FALSE Reference
## 3      0.26 0.26    FALSE Reference
## 4      0.27 0.27    FALSE Reference
## 5      0.28 0.28    FALSE Reference
## 6      0.28 0.28    FALSE Reference
```

```
#Extract all entries in the tccb dataset where its Area attribute has the value
#of "Reference" The empty comma denotes that we extract every single attribute for each
#row, if specified a number, it would only spit out the value of the column
#index for rows with Area=="Reference"
data <- tccb[tccb$Area == "Reference",]
head(data)
```

```
##   TcCB.orig TcCB Censored      Area
## 1      0.22 0.22    FALSE Reference
## 2      0.23 0.23    FALSE Reference
## 3      0.26 0.26    FALSE Reference
## 4      0.27 0.27    FALSE Reference
## 5      0.28 0.28    FALSE Reference
## 6      0.28 0.28    FALSE Reference
```

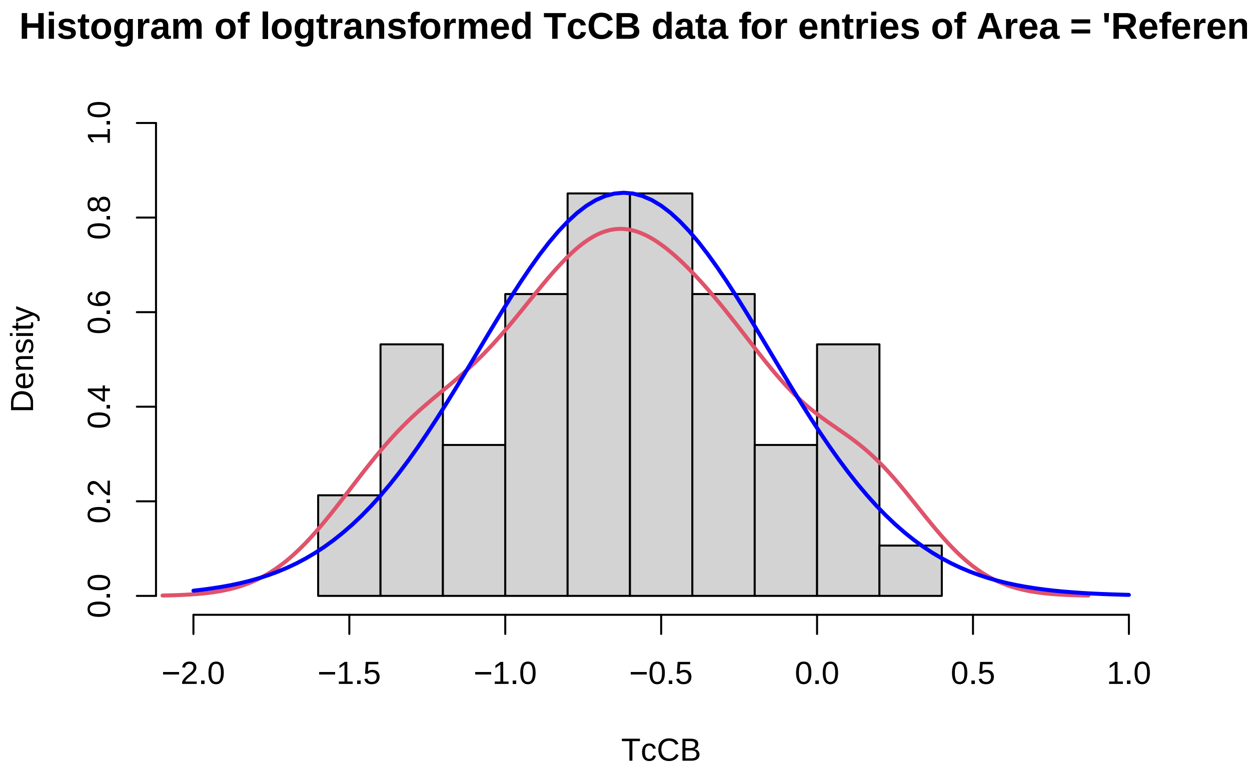
```
#Logtransform the TcCB data
logdata <- log(data$TcCB)
logdata
```

```
##  [1] -1.5141277 -1.4696760 -1.3470736 -1.3093333 -1.2729657 -1.2729657
##  [7] -1.2378744 -1.1086626 -1.0788097 -1.0498221 -0.9675840 -0.9416085
## [13] -0.9416085 -0.8675006 -0.8675006 -0.8439701 -0.7985077 -0.7765288
## [19] -0.7339692 -0.6931472 -0.6931472 -0.6733446 -0.6539265 -0.6161861
## [25] -0.5798185 -0.5798185 -0.5621189 -0.5621189 -0.5108256 -0.4780358
## [31] -0.4620355 -0.4004776 -0.3710637 -0.3285041 -0.3011051 -0.2744368
## [37] -0.2357223 -0.2107210 -0.1984509 -0.1743534 -0.1165338  0.1043600
## [43]  0.1222176  0.1310283  0.1310283  0.1823216  0.2851789
```

```
max(logdata)
```

```
## [1] 0.2851789
```

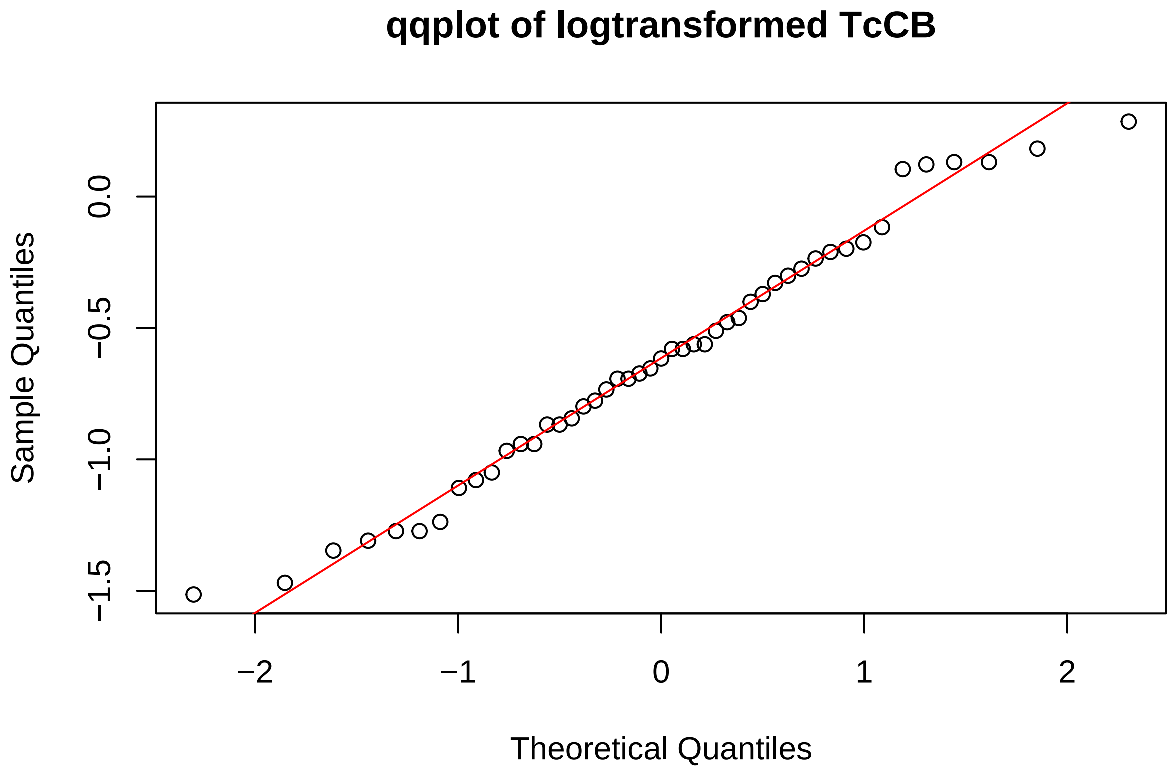
```
#Generate histogram with superimposed normal curve
hist(logdata, probability=TRUE,
     main = "Histogram of logtransformed TcCB data for entries of Area = 'Reference'",
     xlab="TcCB", xlim=c(-2,1), ylim=c(0,1))
lines(density(logdata), col = 2, lwd = 2)
curve(dnorm(x, mean=mean(logdata), sd=sd(logdata)), lwd=2, col="blue", add=TRUE)
```



```
cat("The curve of logtransformed TcCB is approximately similar to the superimposed normal
    curve, with noticeable but small differences in the curve, therefore the normality
    assumption of the dataset is mostly satisfied. ")
```

```
## The curve of logtransformed TcCB is approximately similar to the superimposed normal
##     curve, with noticeable but small differences in the curve, therefore the normality
##     assumption of the dataset is mostly satisfied.
```

```
#qqplot of TcCB and a reference line that shows the fit
qqnorm(logdata, main="qqplot of logtransformed TcCB")
qqline(logdata, col="red")
```



```r
cat("The majority of datapoints of TcCB fall along the 45-degree reference line,
    therefore we can say the assumption of normality is supported, and combined
    with the results of the histogram, normal distribution would be appropriate
    to model this transformed dataset.")
```

```
## The majority of datapoints of TcCB fall along the 45-degree reference line,
##     therefore we can say the assumption of normality is supported, and combined
##     with the results of the histogram, normal distribution would be appropriate
##     to model this transformed dataset.
```

```r
# (b) Obtain MME, MLE and MVUE for mu and sig^2 of the population of log-measurements
# from the Reference area.

#MME for mean and variance
mom1 <- mean(logdata)
mom2 <- 1/length(logdata) * sum((logdata - mean(logdata))^2)
cat("Moment 1(mean):", mom1," Moment 2(sample variance but divided by n, biased):", mom2)
```

```
## Moment 1(mean): -0.6195712  Moment 2(sample variance but divided by n, biased): 0.2143208
```

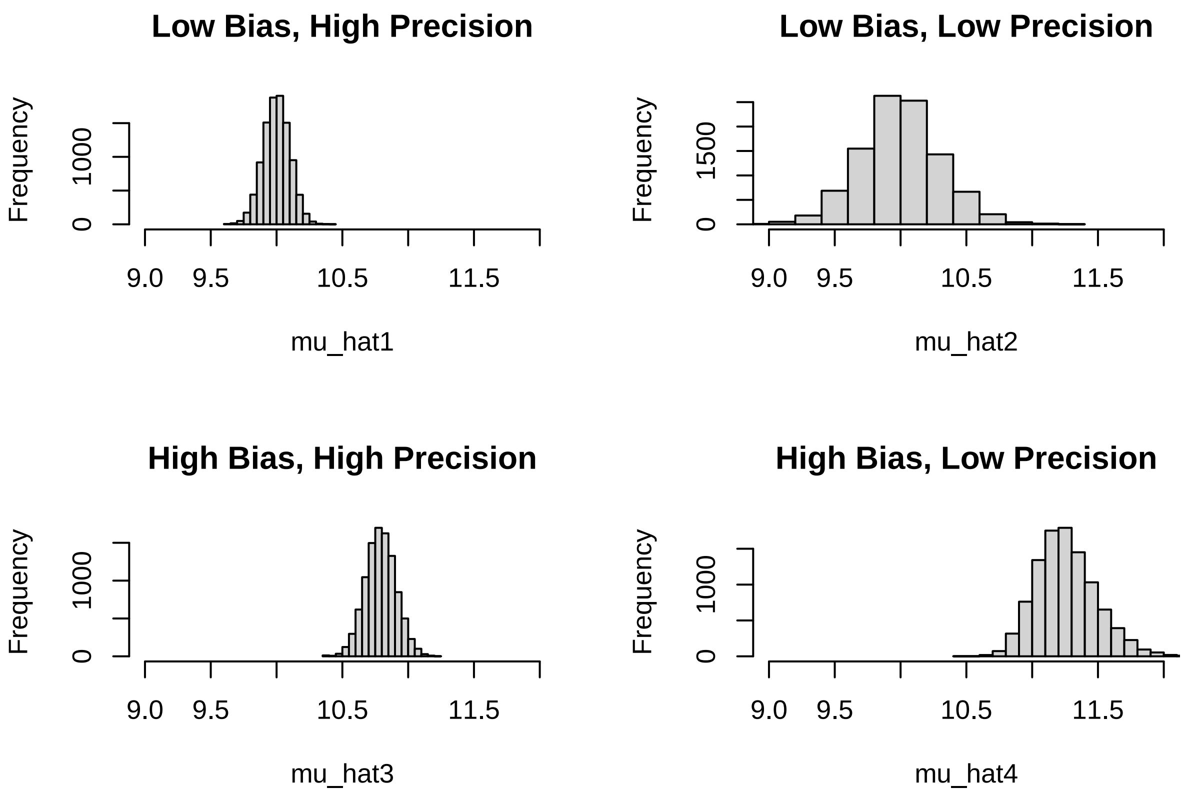
```r
#MLE for mean and variance, from the lecture material
mu_mle <- mean(logdata)
var_mle <- var(logdata)
cat("MLE of mu(mean):", mu_mle," MLE of variance(sample variance)", var_mle)
```

```
## MLE of mu(mean): -0.6195712  MLE of variance(sample variance) 0.21898
```

```
#MVUE for mean and variance
mu_mvue <- mean(logdata)
var_mvue <- var(logdata)
cat("MVUE of mu(mean):", mu_mvue," MVUE of variance(sample variance):", var_mvue)
```

```
## MVUE of mu(mean): -0.6195712  MVUE of variance(sample variance): 0.21898
```

```
#2 - Study the R code given on pages 22-23 of the lecture material Point Estimation.pdf.
# This code conducts a simulation study (also called Monte Carlo study) for comparing
# four different estimators proposed for a population parameter.

  set.seed(1)
  N <- 10000  # Number of samples to simulate
  # Generating samples of size n=100 from N(10,1)
  # and computing each of the four estimates of mu
  mu_hat1 <- c(1:N)
  mu_hat2 <- c(1:N)
  mu_hat3 <- c(1:N)
  mu_hat4 <- c(1:N)

  for (i in 1:N) {
    sample1 <- rnorm(n=100, mean=10, sd=1)
    mu_hat1[i] <- mean(sample1)

    sample2 <- rnorm(n=100, mean=10, sd=1)
    temp <- sort(sample2)
    mu_hat2[i] <- (temp[1] + temp[100]) / 2

    sample3 <- rnorm(n=100, mean=10, sd=1)
    temp <- sort(sample3)
    mu_hat3[i] <- mean(temp[51:100])

    sample4 <- rnorm(n=100, mean=10, sd=1)
    temp <- sort(sample4)
    mu_hat4[i] <- (temp[51] + temp[100]) / 2
  }

  par(mfrow=c(2,2))
  # Plot histograms for each estimator
  hist(mu_hat1, xlim=c(9,12), main="Low Bias, High Precision")
  hist(mu_hat2, xlim=c(9,12), main="Low Bias, Low Precision")
  hist(mu_hat3, xlim=c(9,12), main="High Bias, High Precision")
  hist(mu_hat4, xlim=c(9,12), main="High Bias, Low Precision")
```


Low Bias, High Precision

Low Bias, Low Precision

High Bias, High Precision

High Bias, Low Precision

```
#MSE = Bias^2 + Variance
# Simulated MSE of mu_hat1
mse_mu_hat1 <- var(mu_hat1) + (mean(mu_hat1) - 10)^2
mse_mu_hat1
```

```
## [1] 0.01004533
```

```
# Simulated MSE of mu_hat2
mse_mu_hat2 <- var(mu_hat2) + (mean(mu_hat2) - 10)^2
mse_mu_hat2
```

```
## [1] 0.09276702
```

```
# Simulated MSE of mu_hat3
mse_mu_hat3 <- var(mu_hat3) + (mean(mu_hat3) - 10)^2
mse_mu_hat3
```

```
## [1] 0.6392457
```

```
# Simulated MSE of mu_hat4
mse_mu_hat4 <- var(mu_hat4) + (mean(mu_hat4) - 10)^2
mse_mu_hat4
```

```
## [1] 1.639882
```

```
# (a) What is the population and what is the target population parameter? What are
# the nuisance parameters, if any?
cat("The population is N(10,1) where the population mean is 10 and population variance
    is 1. The target population parameter is mu, the mean. A nuisance parameter is
    any parameter that is not the target parameter(s), but still must be considered
    in order to analyze the targets. The only other parameter of the population is
    the variance, therefore the nuisance parameter is sig^2, the variance.")
```

```
## The population is N(10,1) where the population mean is 10 and population variance
##     is 1. The target population parameter is mu, the mean. A nuisance parameter is
##     any parameter that is not the target parameter(s), but still must be considered
##     in order to analyze the targets. The only other parameter of the population is
##     the variance, therefore the nuisance parameter is sig^2, the variance.
```

```
# (b) What are the different estimators proposed? Comment on the intuitive
#appropriateness of each estimator.
cat("The different estimators proposed are:
    1. Standard sample mean of rnorm(n=100, mean=10, sd=1)
      * The MVUE of the mean, which was shown in #1, is simply the sample mean,
        therefore this estimator is the most appropriate estimator for the population
        mean. The histgoram supports its low bias and high precision nature, because
        it considers all values in the data.
    2. Mean of the minimum and maximum values of rnorm(n=100, mean=10, sd=1), the midrange
      * From worksheet04 we can recall that min and max are very susceptible to
        extreme outliers, which skew their distributions even with very large
        sample sizes. Therefore because of their susceptibility to
        extreme outliers, this midrange estimator is not an appropriate estimator
        for the population mean. From the histogram, we see that the bias is low,
        because both sides of the mean are considered equally in min and max, but because
        it is susceptible to outliers, it has low precision.
    3. Mean of the top 50% values of sorted rnorm(n=100, mean=10, sd=1)
      * Intuitively, this is also a biased estimator because it ignores the bottom 50%
        lower values of the sample, which skews the mean greatly, making it not an appropriate
        estimator for the population mean, even though its precision is  high as seen from the
        histogram.
    4. Mean of the 51st value and the 100th value of sorted rnorm(n=100, mean=10, sd=1),
    which is the midrange of the top 50% of the random sample; (max-min)/2 .
      * This estimator calculates the midrange of the top 50% highest values of the data,
        which from the previous estimators and the histogram can conclude will be
        heavily biased towards the top 50% of the data and have low precision around
        the midpoint. It is the least appropriate estimator of the 4 for the
        population mean.
    ")
```

```
## The different estimators proposed are:
##     1. Standard sample mean of rnorm(n=100, mean=10, sd=1)
##       * The MVUE of the mean, which was shown in #1, is simply the sample mean,
##         therefore this estimator is the most appropriate estimator for the population
##         mean. The histgoram supports its low bias and high precision nature, because
##         it considers all values in the data.
##     2. Mean of the minimum and maximum values of rnorm(n=100, mean=10, sd=1), the midrange
##       * From worksheet04 we can recall that min and max are very susceptible to
##         extreme outliers, which skew their distributions even with very large
##         sample sizes. Therefore because of their susceptibility to
##         extreme outliers, this midrange estimator is not an appropriate estimator
##         for the population mean. From the histogram, we see that the bias is low,
##         because both sides of the mean are considered equally in min and max, but because
##         it is susceptible to outliers, it has low precision.
##     3. Mean of the top 50% values of sorted rnorm(n=100, mean=10, sd=1)
##       * Intuitively, this is also a biased estimator because it ignores the bottom 50%
##         lower values of the sample, which skews the mean greatly, making it not an appropriate
##         estimator for the population mean, even though its precision is  high as seen from the
##         histogram.
##     4. Mean of the 51st value and the 100th value of sorted rnorm(n=100, mean=10, sd=1),
##     which is the midrange of the top 50% of the random sample; (max-min)/2 .
##       * This estimator calculates the midrange of the top 50% highest values of the data,
##         which from the previous estimators and the histogram can conclude will be
##         heavily biased towards the top 50% of the data and have low precision around
##         the midpoint. It is the least appropriate estimator of the 4 for the
##         population mean.
##
```

```
# (c) The simulation assumes a specific value for the population parameters. Can we
# extend the conclusion of this simulation study for other values of the parameters?
cat("The simulations assume that the population mean is 10, which was given initially
    in N(10,1), meaning the population variance is also 1. If we compare the MSE to the
    variance of the target parameter, the lower the MSE the better the estimator.
    The conclusions of these simulations relate only to the specific parameter values
    of mu=10 and sig^2=1, in order to extend the study to other values of different
    parameters, the study would have to be redone using the new given parameters for
    a normal distribution, then would the study itself be stastically and empircally supported
    from testing with various population parameters. After it is shown, or not shown,
    that the estimators are robust among different parameters, only then can the
    estimators be extended for other values of the parameters, significantly.
    ")
```

```
## The simulations assume that the population mean is 10, which was given initially
##     in N(10,1), meaning the population variance is also 1. If we compare the MSE to the
##     variance of the target parameter, the lower the MSE the better the estimator.
##     The conclusions of these simulations relate only to the specific parameter values
##     of mu=10 and sig^2=1, in order to extend the study to other values of different
##     parameters, the study would have to be redone using the new given parameters for
##     a normal distribution, then would the study itself be stastically and empircally supported
##     from testing with various population parameters. After it is shown, or not shown,
##     that the estimators are robust among different parameters, only then can the
##     estimators be extended for other values of the parameters, significantly.
##
```

```
# (d) Modify the code to compare the MME to MVUE for estimating the variance sig^2
#of a normal distribution.
var_hat1 <- c(1:N)
var_hat2 <- c(1:N)

for (i in 1:N) {
    sample1 <- rnorm(n=1000, mean=10, sd=1)
    #MME for var = sample variance, but n instead of n-1
    var_hat1[i] <- 1/length(sample1) * sum((sample1 - mean(sample1))^2)

    sample2 <- rnorm(n=1000, mean=10, sd=1)
    #MVUE for var = sample variance
    var_hat2[i] <- 1/(length(sample2)-1) * sum((sample2 - mean(sample2))^2)
}

hist(var_hat1, main="MME")
hist(var_hat2, main="MVUE")

#MSE = Bias^2 + Variance
# Simulated MSE of var_hat2
#The actual variance parameter to calculate bias is 1, 10 is for the population mean
mse_var_hat1 <- var(var_hat1) + (mean(var_hat1) - 1)^2
mse_var_hat1
```

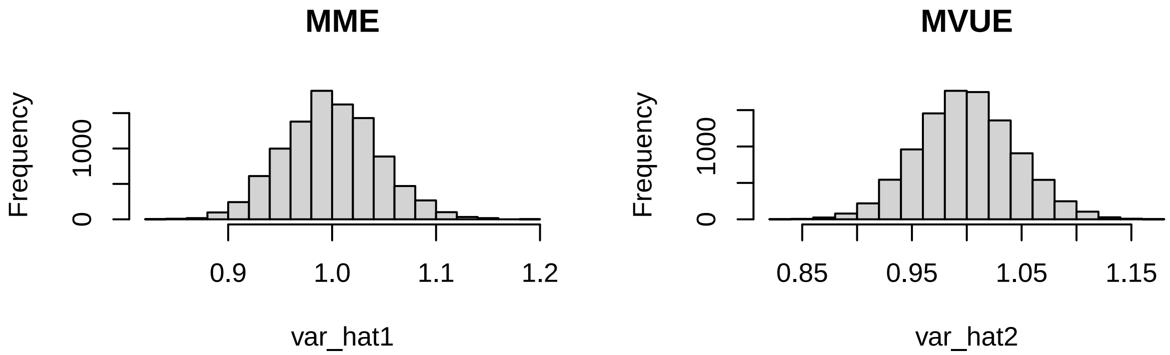
```
## [1] 0.002057116
```

```
# Simulated MSE of var_hat2
mse_var_hat2 <- var(var_hat2) + (mean(var_hat2) - 1)^2
mse_var_hat2
```

```
## [1] 0.001984034
```

```
cat("The MSE for MVUE is higher than the MME for the simulation done on variance,
    this discrepancy could be explained by the small sample size for generating
    n=100 random samples from N(10,1). If the code is rerun with n=1000, for
    rnorm(n=1000, mean=10, sd=1), the MSE of the MVUE for variance is lower than
    the MSE of MME, which is what the code above shows.
    ")
```

```
## The MSE for MVUE is higher than the MME for the simulation done on variance,
##     this discrepancy could be explained by the small sample size for generating
##     n=100 random samples from N(10,1). If the code is rerun with n=1000, for
##     rnorm(n=1000, mean=10, sd=1), the MSE of the MVUE for variance is lower than
##     the MSE of MME, which is what the code above shows.
##
```


MME
Frequency
0
1000
0.9
1.0
1.1
1.2
var_hat1
MVUE
Frequency
0
1000
0.85
0.95
1.05
1.15
var_hat2