# Norm_Asymp_WithMLE.R

This simulation generates data from a N$(0.5, 1)$ distribution and observes the asymptotic behavior of ECIC using the model set $\mathcal{M} = \{N(0,1), N(\mu,1), N(\mu,\sigma)\}$

```
#########################################################################
#                        Source Functions/Data Generation
#########################################################################

library(Rcpp)
```

```
## Warning: package 'Rcpp' was built under R version 4.0.5
```

```
library(RcppArmadillo)
```

```
## Warning: package 'RcppArmadillo' was built under R version 4.0.5
```

```
# load C++ script that is used for simulation steps in ECIC
sourceCpp("FnsInCPlusPlus\\Simulation_Rcpp_Fns.cpp")
# load function in written in R
source("FnsInR\\Simulation_R_Fns.R")

# set true normal model parameters
trueMu <- 0.5
trueSig <- 1
trueParams <- c("trueMu"=trueMu,"trueSig"=trueSig)
trueModel <- "N(0.5,1)"
MNames <- c("N(0,1)","N(mu,1)","N(mu,sigma)")
# model that should be selected
closestMod <- "N(mu,1)"
# cardinality of the model set
MLen <- length(MNames)
M <- list()
for(i in 1:MLen)
{
  M[[i]] <- MNames[i]
}
# set different sample sizes
ns <-  c(3,10,50,100,200)
# cardinality of the sample sizes
nsLen <- length(ns)
# set the number of draws for each sample size
noDraws <- 200
# sample size for estimating the probability of choosing the observed best
# model under the assumption an
# alternative model is true
```

```r
N1 <- 300
# sample size for simulating the DGOF distribution under the assumption that
# an alternative model is true
N2 <- 700
# pre-specified type-1 error rate
alpha <- 0.15
datList <- list()
set.seed(225)
# generate data from the true normal distribution
for(i in 1:nsLen)
{
  tempN <- ns[i]
  datList[[i]] <- generateData(tempN,noDraws,trueParams,"Normal")
}

# compute MLE estimates from the generated data
MLEList <- list()
for(i in 1:nsLen)
{
  MLEList[[i]] <- list()
}

for(i in 1:nsLen)
{
  tempN <- ns[i]
  tempDatList <- datList[[i]]
  tempMeanMLEs <- apply(X=tempDatList,MARGIN=2,FUN=function(x) mean(x))
  tempSigMLEs <- apply(X=tempDatList,MARGIN=2,FUN=function(x) sd(x))
  # convert from unbiased estimate to MLE
  tempSigMLEs <- tempSigMLEs*sqrt((tempN-1))/sqrt(tempN)
  # MLE's for N(0,1) (just 0,1 b/c no parameter estimation)
  MLEList[[i]][[1]] <- matrix(rep(c(0,1),noDraws),nrow=2,ncol=noDraws)
  rownames(MLEList[[i]][[1]]) <- c("fixedMean","fixedsig")
  # MLE's for N(mu,1)
  MLEList[[i]][[2]] <- rbind(tempMeanMLEs,1)
  rownames(MLEList[[i]][[2]]) <- c("MLEMean","fixedSig")
  # MLE's for N(mu,sigma)
  MLEList[[i]][[3]] <- rbind(tempMeanMLEs,tempSigMLEs)
  rownames(MLEList[[i]][[3]]) <- c("MLEMean","MLESig")
  colnames(MLEList[[i]][[1]]) <- colnames(MLEList[[i]][[2]]) <-
    colnames(MLEList[[i]][[3]]) <- paste("Draw",1:noDraws,sep="")
  names(MLEList[[i]]) <- paste("MLEs for Model",MNames)
}
names(MLEList) <- paste("Draws of n=",ns,sep="")

#########################################################################
#             Begin ECIC
#########################################################################

# ECIC step #1
# compute the IC under each model in the model set for each sample size
ICComps <- list()
for(i in 1:nsLen)
```

```
{
  ICComps[[i]] <- ICComputations(datMat=datList[[i]],M=M,MLen=MLen,
                                 noDraws=noDraws,
                                 ICType="BICNorm",MNames=MNames)
}
names(ICComps) <- paste("True Model",trueModel,",Draws of n=",ns,sep="")

# ECIC step #2
# determine the observed best models for each draw of each sample size
MbList <- list()
for(i in 1:nsLen)
{
  MbList[[i]] <- MbComputations(ICComps[[i]],MNames)
}
names(MbList) <- paste("Mbs for ","Draws of n=",ns,sep="")

# ECIC step #3
# compute the observed DGOFs for each draw of each sample size
obsDGOFs <- list()
obsDGOFs <- obsDGOFsComputations(ICComps,nsLen)

# ECIC step #4
set.seed(19)
ptm <- proc.time() #Start timing
simDat1List <- normDatSimRcpp(ns,MLEList,N1)
```

```
## Iteration 1 Complete
## Iteration 2 Complete
## Iteration 3 Complete
## Iteration 4 Complete
## Iteration 5 Complete
```

```
simDat2List <- normDatSimRcpp(ns,MLEList,N2)
```

```
## Iteration 1 Complete
## Iteration 2 Complete
## Iteration 3 Complete
## Iteration 4 Complete
## Iteration 5 Complete
```

```
proc.time() - ptm
```

```
##    user  system elapsed
##    6.35    0.31    4.66
```

```
names(simDat1List) <- paste("Draws of n=",ns,sep="")
names(simDat2List) <- paste("Draws of n=",ns,sep="")
# provide names for both simulated sets
for(i in 1:nsLen)
{
  for(j in 1:MLen) # indexes the models in the model set
```

```
  {
    names(simDat1List[[i]][[j]]) <- paste("Obs",1:noDraws,";",N1,
                                          "Simulated Draws")
    names(simDat2List[[i]][[j]]) <- paste("Obs",1:noDraws,";",N2,
                                          "Simulated Draws")
  }
  names(simDat1List[[i]]) <- paste("Generated from",MNames)
  names(simDat2List[[i]]) <- paste("Generated from",MNames)
}

# simulate distributions to estimate probabilities
ptm=proc.time()
ICsSimDat1 <- ICCompsRcpp(simDat1List,MNames)
```

```
## Its. for Sample Size Index 1 completed
## Its. for Sample Size Index 2 completed
## Its. for Sample Size Index 3 completed
## Its. for Sample Size Index 4 completed
## Its. for Sample Size Index 5 completed
```

```
proc.time() - ptm
```

```
##    user  system elapsed
##    5.32    0.14    5.61
```

```
# label the elements in ICsSimDat1
for(i in 1:nsLen)
{
  for(j in 1:MLen)
  {
    for(k in 1:noDraws)
    {
      colnames(ICsSimDat1[[i]][[j]][[k]]) <- paste("BIC Under", MNames)
    }
    names(ICsSimDat1[[i]][[j]]) <- paste("Normal Fit for Obs",1:noDraws)
  }
  names(ICsSimDat1[[i]]) <- paste("Generated from",MNames)
}
names(ICsSimDat1) <- paste("Draws of n=",ns,sep="")

# determine the model with the minimum IC for each set of draws
minICList <- list()
for(i in 1:nsLen)
{
  minICList[[i]] <- list()
  for(j in 1:MLen)
    minICList[[i]][[j]] <- list()
}
for(i in 1:nsLen) # i indexes sample size
{
  for(j in 1:MLen) # j indexes assumed true parameter
  {
```

```
    for(k in 1:noDraws)
    {
      minICList[[i]][[j]][[k]] <- apply(ICsSimDat1[[i]][[j]][[k]],
                                        MARGIN=1,FUN=function(x)
                                          MNames[which.min(x)])
    }
    names(minICList[[i]][[j]]) <- paste("Normal Fit for Obs",1:noDraws)
  }
  names(minICList[[i]]) <- paste("Generated from",MNames)
  print(i)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
```

```
names(minICList) <- paste("Draws of n=",ns,sep="")

# ECIC step #4b
# create a list of matrices that hold P_i(g(F)=M_b)
piHatList <- list()
for(i in 1:nsLen)
{
  piHatList[[i]] <- list()
  for(k in 1:noDraws)
  {
    piHatList[[i]][[k]] <- list()
  }
  names(piHatList[[i]]) <- paste("Normal Fit for Obs",1:noDraws)
}
names(piHatList) <- names(piHatList) <- paste("n=",ns,sep="")

# compute the probabilities
for(i in 1:nsLen) # indexes the sample size
{
  piHatList[[i]] <- piHatMatComputationsMLEs(minICList[[i]],MLen,N1,
                                             MNames,noDraws)
  print(i)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
```

```
names(piHatList) <- paste("n=",ns,sep="")

rm(ICsSimDat1)
rm(minICList)
gc()
```

```
##             used   (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells    635748   34.0    1351012   72.2   1351012   72.2
## Vcells 219058784 1671.3  335451302 2559.3 224880784 1715.8
```

```
# ECIC step #4c
DGOFList <- simDGOFsRcpp(simDat2List,MNames)
```

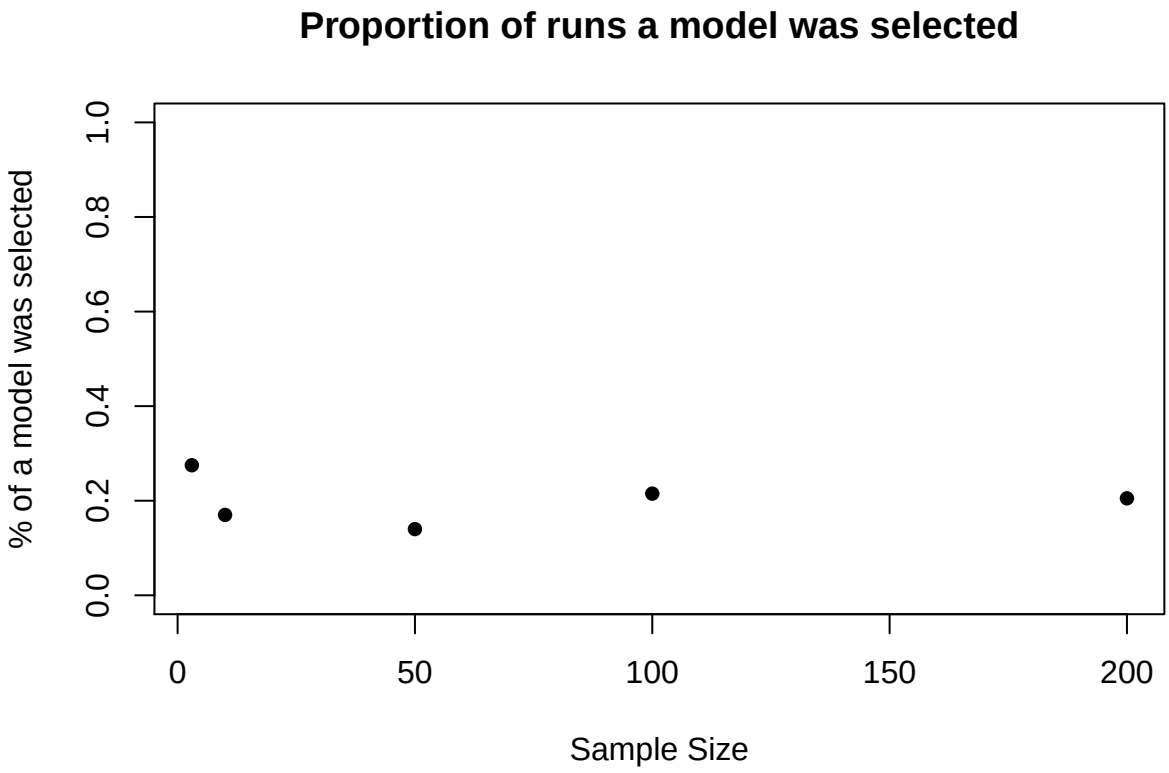
```
## Its. for Sample Size Index 1 completed
## Its. for Sample Size Index 2 completed
## Its. for Sample Size Index 3 completed
## Its. for Sample Size Index 4 completed
## Its. for Sample Size Index 5 completed
```

```
# label the elements in DGOFList
for(i in 1:nsLen)
{
  for(j in 1:MLen)
  {
    for(k in 1:noDraws)
    {
      colnames(DGOFList[[i]][[j]][[k]]) <- paste("DGOF Under", MNames,
                                                 " Observed Best")
    }
    names(DGOFList[[i]][[j]]) <- paste("Normal Fit for Obs",1:noDraws)
  }
  names(DGOFList[[i]]) <- paste("Generated from",MNames)
}
names(DGOFList) <- paste("Draws of n=",ns,sep="")

# ECIC steps 4d, 5, and 6
resultList <- ECICDecisionsMLEs(MbList,obsDGOFs,piHatList,DGOFList,alpha,
                                MNames,nsLen,MLen,noDraws)
```

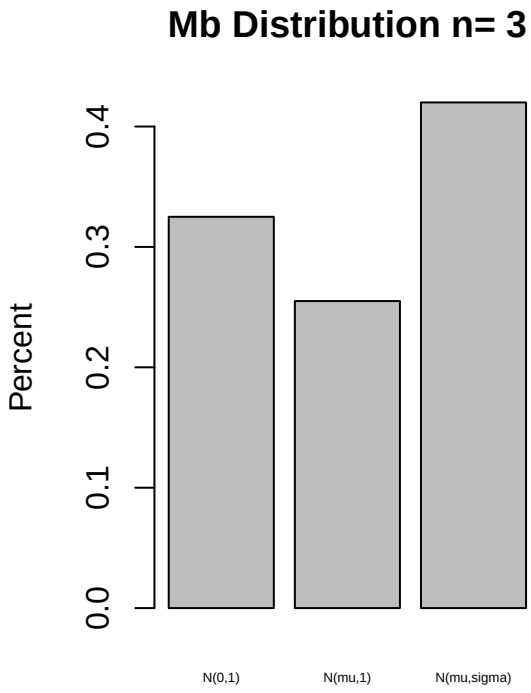
```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
```

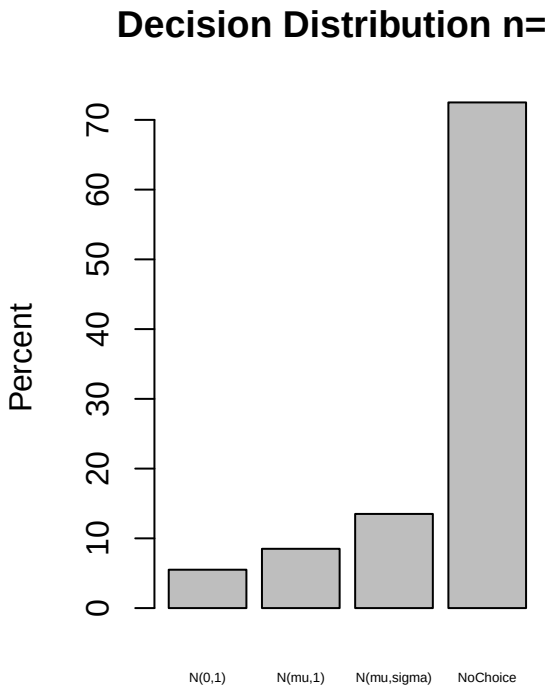
```
#########################################################################
#                       Plot Results
#########################################################################

# list to store the decision thresholds
thresholds <- resultList[[1]]
# decision list
aOrRList <- resultList[[2]]

# assess observed best model with ECIC choice
assessList <- list()
for(i in 1:nsLen)
{
```

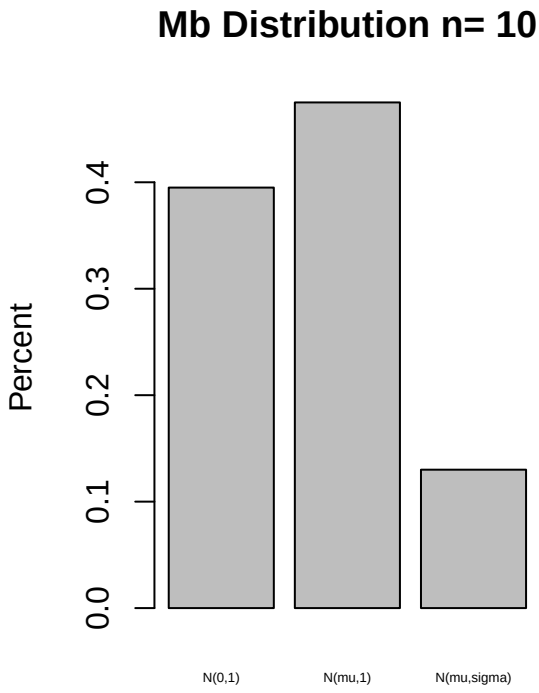
```
  assessList[[i]] <- rbind(MbList[[i]],aOrRList[[i]])
}

DecisionRates <- sapply(X=assessList,FUN=function(x) sum(as.numeric(x[2,]))/
                          noDraws)
plot(x=ns,y=DecisionRates,main="Proportion of runs a model was selected",
     xlab="Sample Size",ylab="% of a model was selected",pch=16,ylim = c(0,1))
```

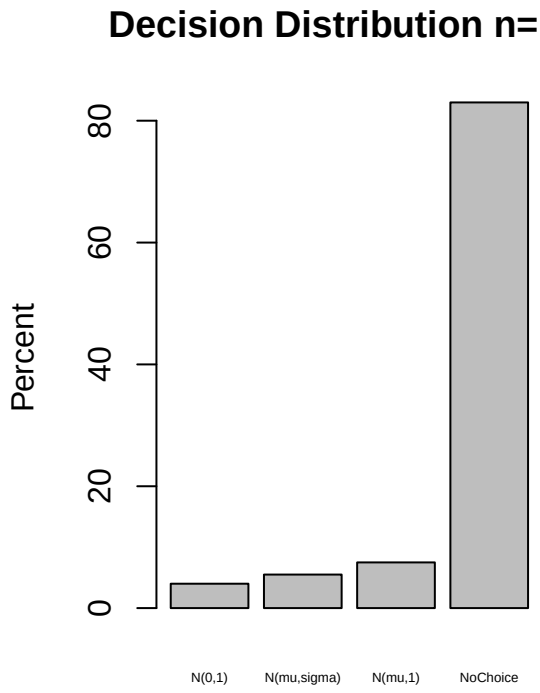


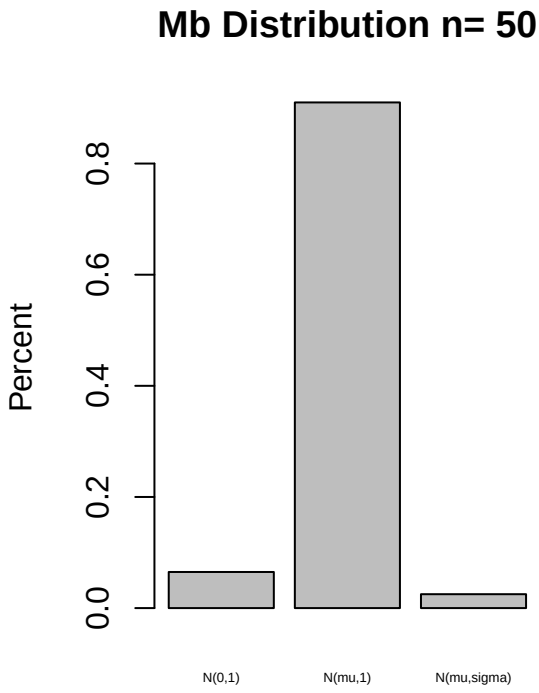

```
# take a look at type 1 error rate
# subset assess list by only when a decision was made i.e. second row = 1
decAssessList <- lapply(X=assessList,FUN=function(x) x[,x[2,]==1])
par(mfrow = c(1, 2))
for(i in 1:nsLen)
{
  # get the frequencies of models selected as best
  tempTable <- table(MbList[[i]])
  barplot(tempTable/noDraws,main=paste("Mb Distribution n=",ns[i]),
          cex.names=0.45,ylab="Percent")
  chosenModelRate <- table(decAssessList[[i]][1,])/noDraws*100
  unChosenModelRate <- (noDraws-ncol(decAssessList[[i]]))/noDraws*100
  decDist <- c(chosenModelRate,"NoChoice"=unChosenModelRate)
  decDist <- sort(decDist)
  barplot(decDist,main=paste("Decision Distribution n=",ns[i]),cex.names=0.45,
          ylab="Percent")
}
```

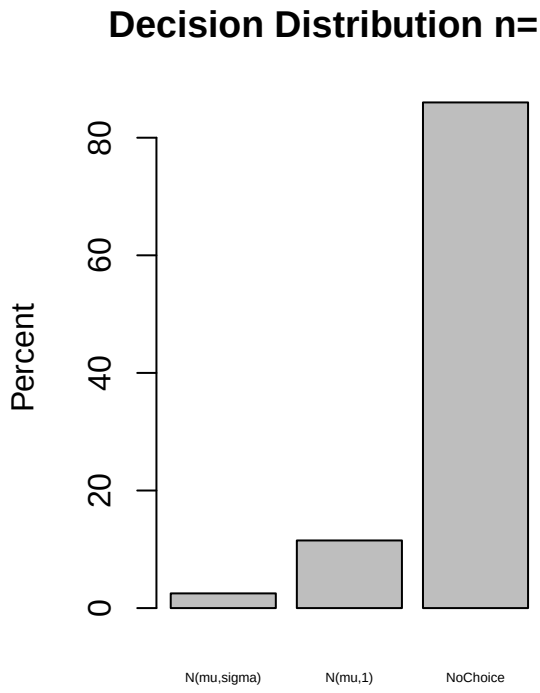

Mb Distribution n= 3
Percent
0.0
0.1
0.2
0.3
0.4
N(0,1)
N(mu,1)
N(mu,sigma)

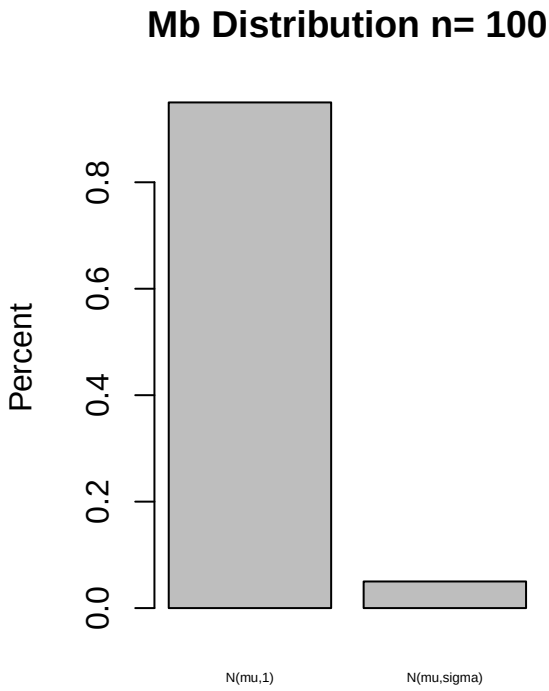

Decision Distribution n= 3
Percent
0
10
20
30
40
50
60
70
N(0,1)
N(mu,1)
N(mu,sigma)
NoChoice

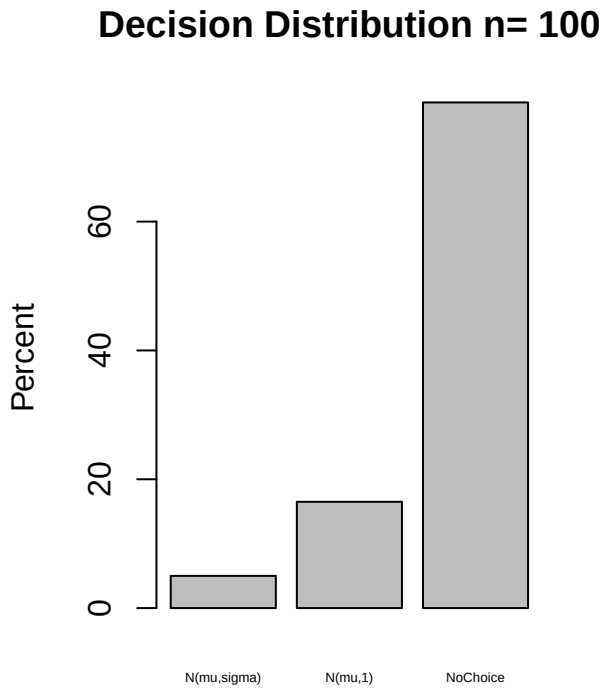

Mb Distribution n= 10
Percent
0.0
0.1
0.2
0.3
0.4
N(0,1)
N(mu,1)
N(mu,sigma)

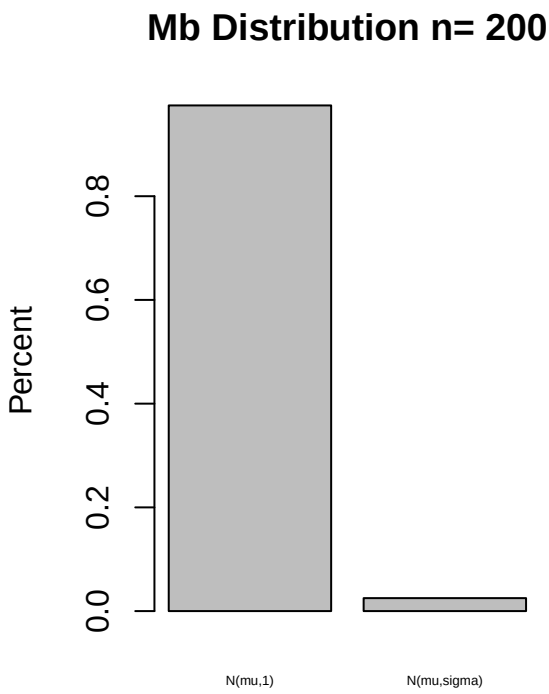

Decision Distribution n= 10
Percent
0
20
40
60
80
N(0,1)
N(mu,sigma)
N(mu,1)
NoChoice

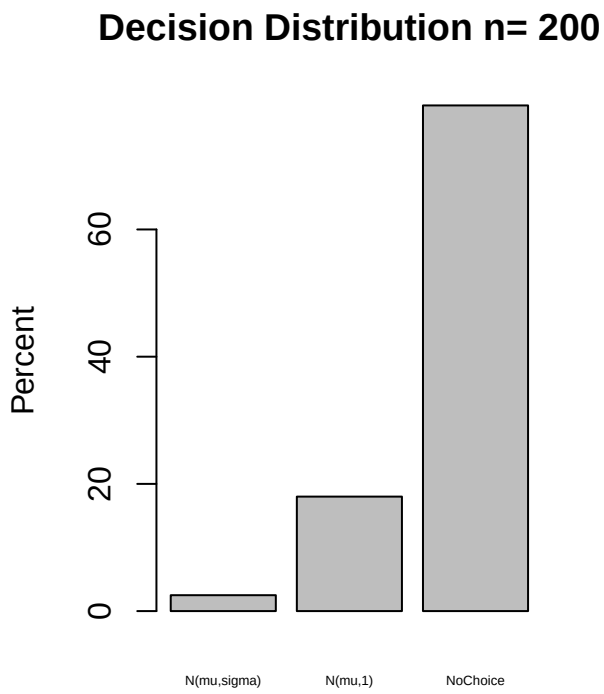

Mb Distribution n= 50
Percent
0.0
0.2
0.4
0.6
0.8
N(0,1)
N(mu,1)
N(mu,sigma)


Decision Distribution n= 50
Percent
0
20
40
60
80
N(mu,sigma)
N(mu,1)
NoChoice

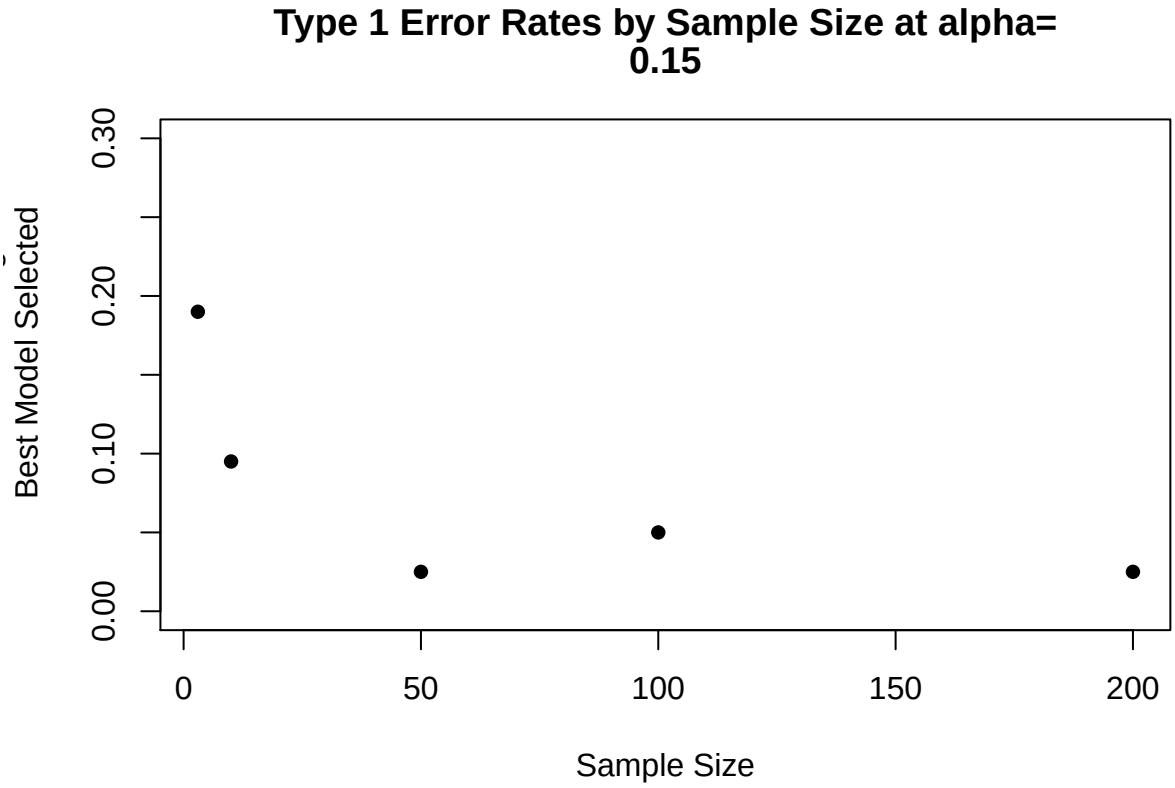

Mb Distribution n= 100
Percent
0.0
0.2
0.4
0.6
0.8
N(mu,1)
N(mu,sigma)


Decision Distribution n= 100
Percent
0
20
40
60
N(mu,sigma)
N(mu,1)
NoChoice

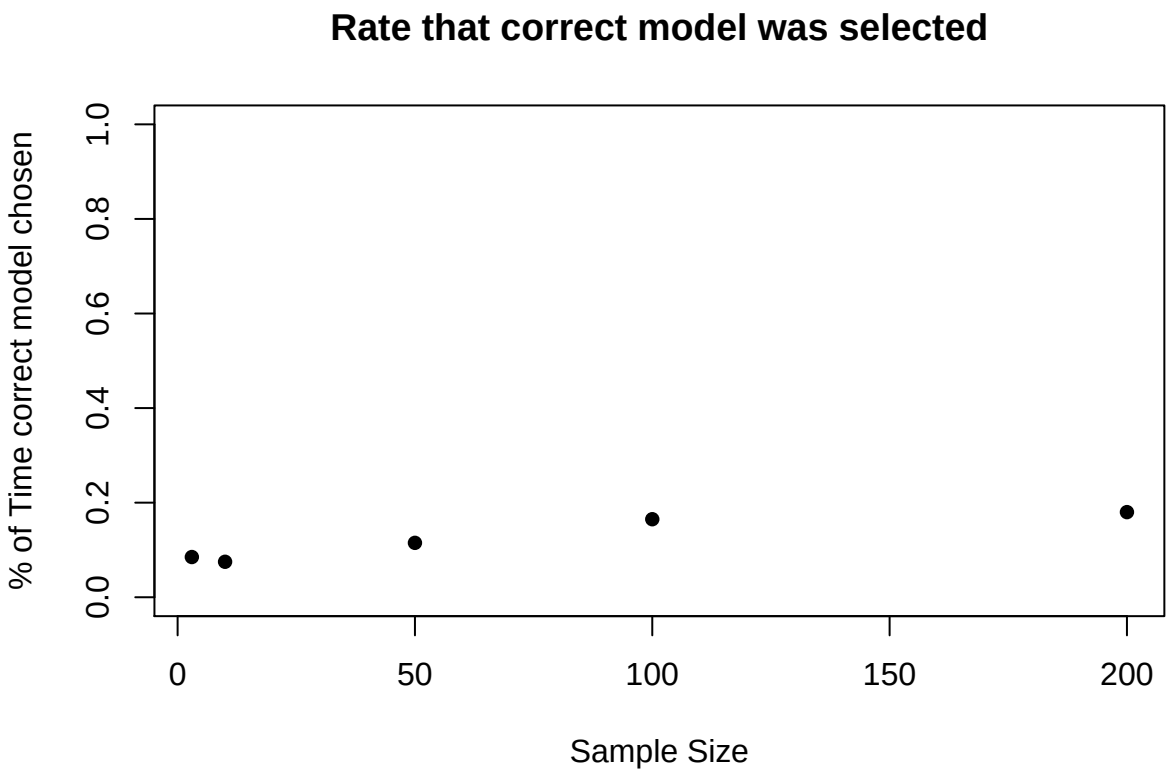



```
par(mfrow = c(1, 1))
# compute type 1 error rate
t1ErrorRates <- sapply(X=decAssessList,FUN=function(x) sum(x[1,]!=closestMod)/
                         noDraws)
plot(x=ns,y=t1ErrorRates,main=c("Type 1 Error Rates by Sample Size at alpha=",
                                alpha),xlab="Sample Size",ylab="% of Time Wrong
     Best Model Selected",pch=16,ylim = c(0,2*alpha))
```




```
CorrectRates <- sapply(X=decAssessList,FUN=function(x) sum(x[1,]==closestMod)/
                         noDraws)
plot(x=ns,y=CorrectRates,main="Rate that correct model was selected",
     xlab="Sample Size",ylab="% of Time correct model chosen",pch=16,
     ylim = c(0,1))
```


Rate that correct model was selected
% of Time correct model chosen
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Sample Size