MSDS_600_Week4_Assn.R

SeanOMalley1

Thu Jul 21 12:36:49 2016

```
# MSDS Assn Week 4
# Sean O'Malley

BN1 <- read.csv("/Users/SeanOMalley1/Documents/REGIS\ MSDS/MSDS_600/R\ Week\ 4\ Assn/
BN1.csv", header= FALSE)
BN2 <- read.csv("/Users/SeanOMalley1/Documents/REGIS\ MSDS/MSDS_600/R\ Week\ 4\ Assn/
BN2.csv", header=FALSE)
BNOM <- read.csv("/Users/SeanOMalley1/Documents/REGIS\ MSDS/MSDS_600/R\ Week\ 4\ Assn
/Binomial.csv", header=FALSE)
IN <- read.csv("/Users/SeanOMalley1/Documents/REGIS\ MSDS/MSDS_600/R\ Week\ 4\ Assn/l
n.csv", header=FALSE)
N1 <- read.csv("/Users/SeanOMalley1/Documents/REGIS\ MSDS/MSDS_600/R\ Week\ 4\ Assn/N
1.csv", header=FALSE)
N2 <- read.csv("/Users/SeanOMalley1/Documents/REGIS\ MSDS/MSDS_600/R\ Week\ 4\ Assn/N
2.csv", header=FALSE)

#BN1
dim(BN1) #shows the dimensions of the data
```

```
## [1] 100000      1
```

```
names(BN1) #shows the names of each variable/attribute
```

```
## [1] "V1"
```

```
summary(BN1) #summary of each variable/attribute
```

```
##        V1
##  Min.   : 1.781
##  1st Qu.: 8.643
##  Median : 9.993
##  Mean   : 9.994
##  3rd Qu.:11.343
##  Max.   :18.612
```

```
str(BN1)
```

```
## 'data.frame':    100000 obs. of  1 variable:
##  $ V1: num  10.11 11.95 8.65 11.47 6.19 ...
```

```
BN1mean = mean(BN1[,1])
BN1mean
```

```
## [1] 9.994328
```

```
var(BN1) #variance between numbers in data set
```

```
##          V1
## V1 4.001427
```

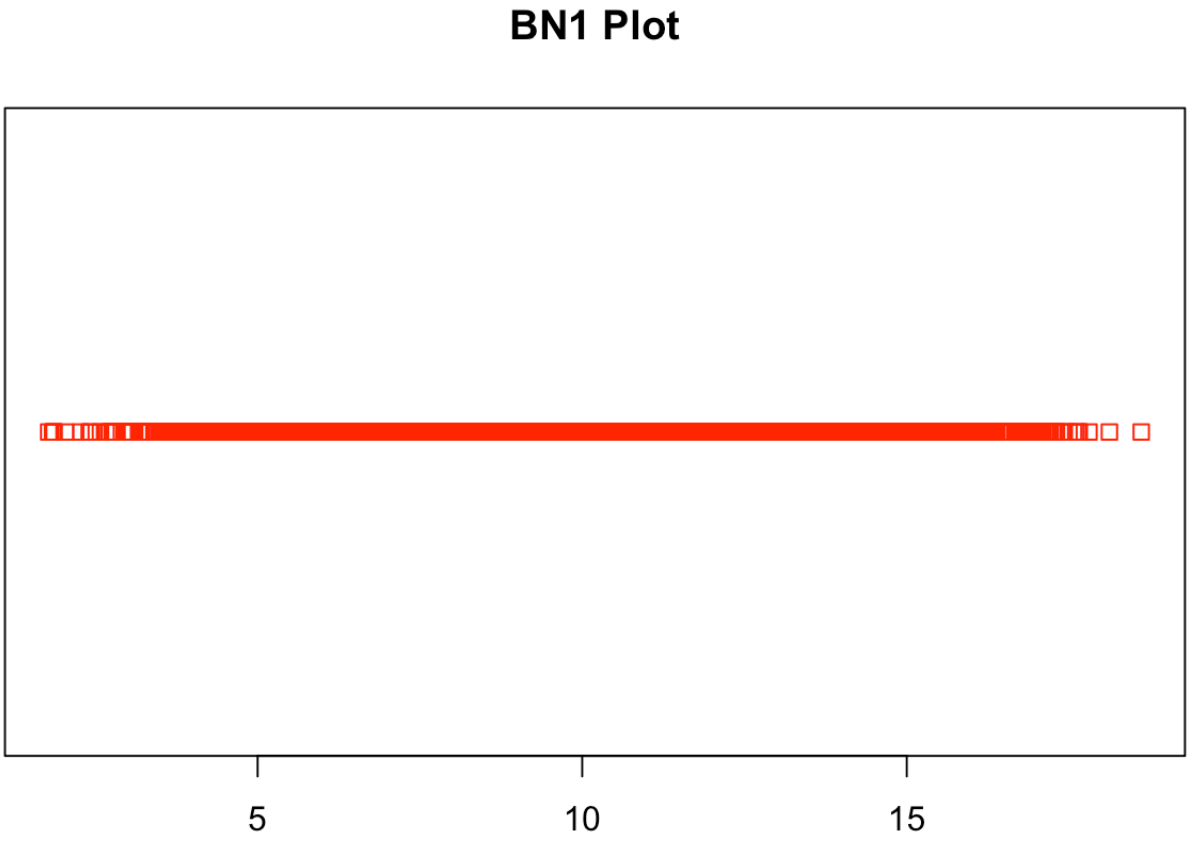
```
BN1sd1 <- sqrt(sum((BN1-BN1mean)^2)/(100000-1)) #standard deviation using traditional
math equation
BN1sd1
```

```
## [1] 2.000357
```

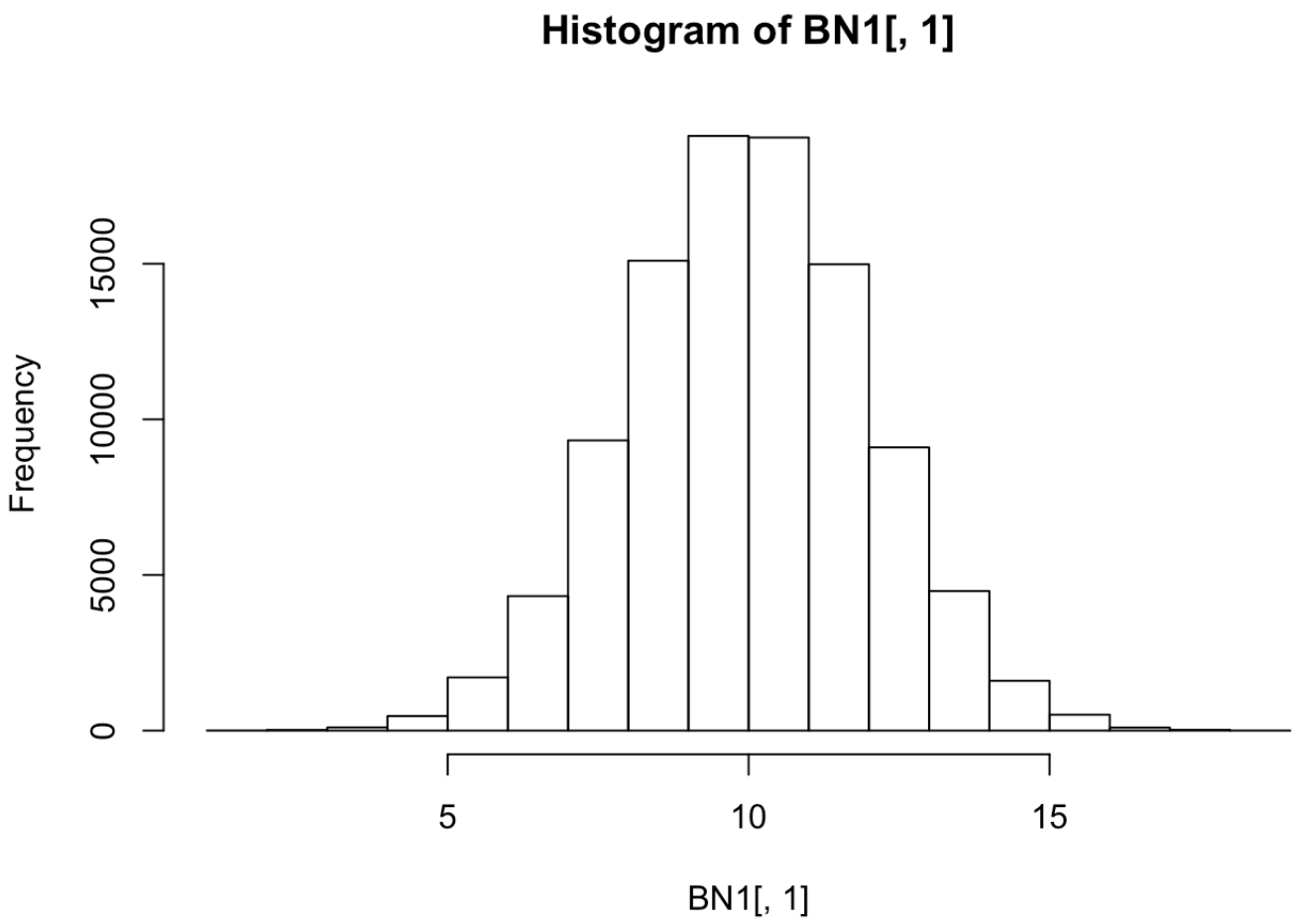
```
summary(BN1)
```

```
##        V1
##  Min.   : 1.781
##  1st Qu.: 8.643
##  Median : 9.993
##  Mean   : 9.994
##  3rd Qu.:11.343
##  Max.   :18.612
```

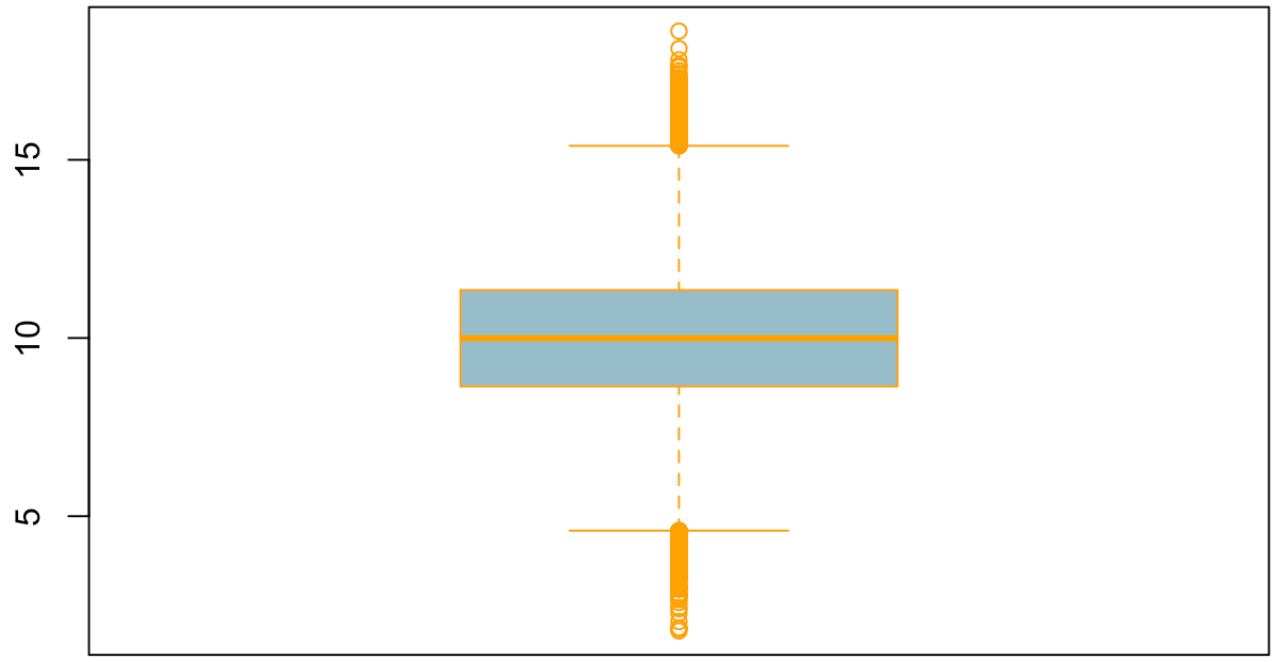
```
plot(BN1, main="BN1 Plot", col="red")
```




```
hist(BN1[,1])
```




```
boxplot(BN1,data=BN1 ,main="BN1 Boxplot",col="lightblue3",border="orange",fill="gray8
8")
```




```
#BN2
dim(BN2) #shows the dimensions of the data
```

```
## [1] 100000      1
```

```
names(BN2) #shows the names of each variable/attribute
```

```
## [1] "V1"
```

```
summary(BN2) #summary of each variable/attribute
```

```
##        V1
##  Min.   : 6.638
##  1st Qu.:10.321
##  Median :10.998
##  Mean   :10.997
##  3rd Qu.:11.667
##  Max.   :15.161
```

```
str(BN2)
```

```
## 'data.frame':    100000 obs. of  1 variable:
##  $ V1: num  11.57 9.55 10.88 11.29 10.84 ...
```

```
BN2mean = mean(BN2[,1])
BN2mean
```

```
## [1] 10.99686
```

```
var(BN2)
```

```
##           V1
## V1 0.9989431
```

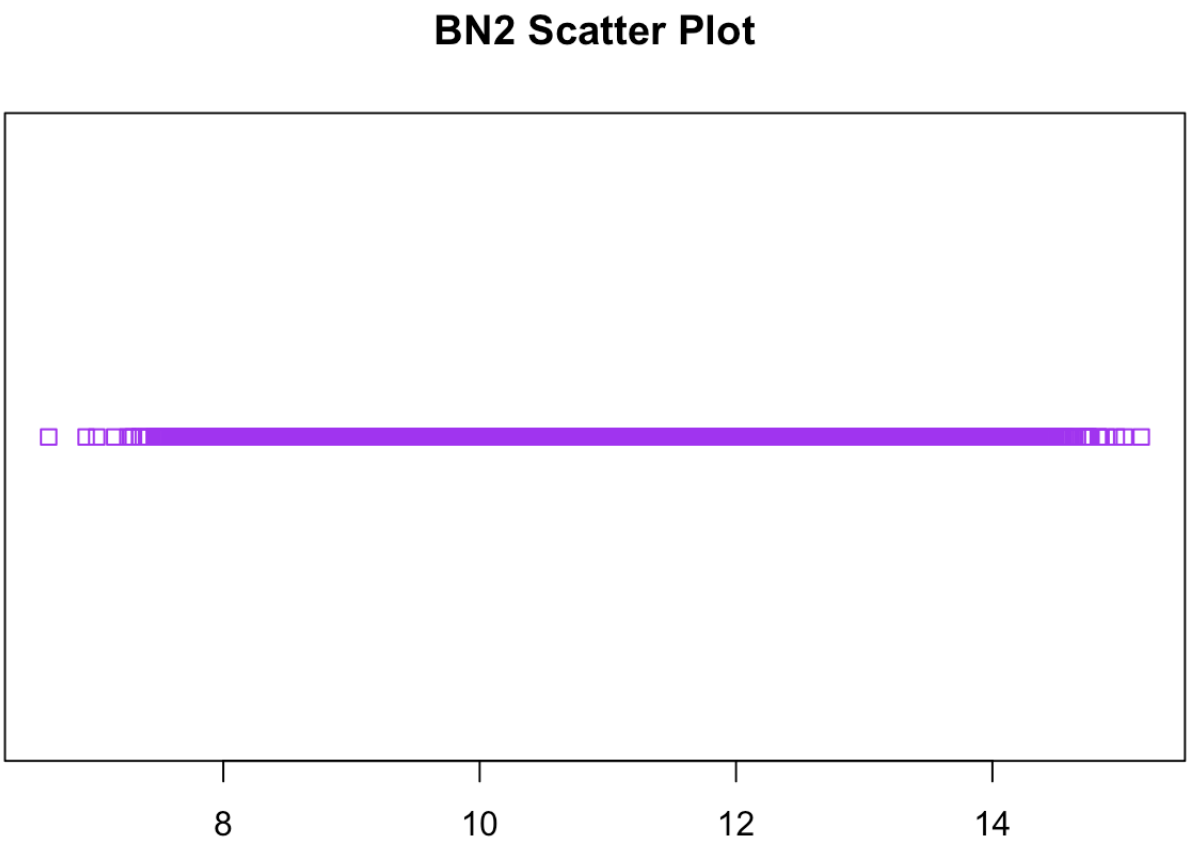
```
BN2sd1 <- sqrt(sum((BN2-BN2mean)^2)/(100000-1)) #standard deviation using traditional
math equation
BN2sd1
```

```
## [1] 0.9994714
```

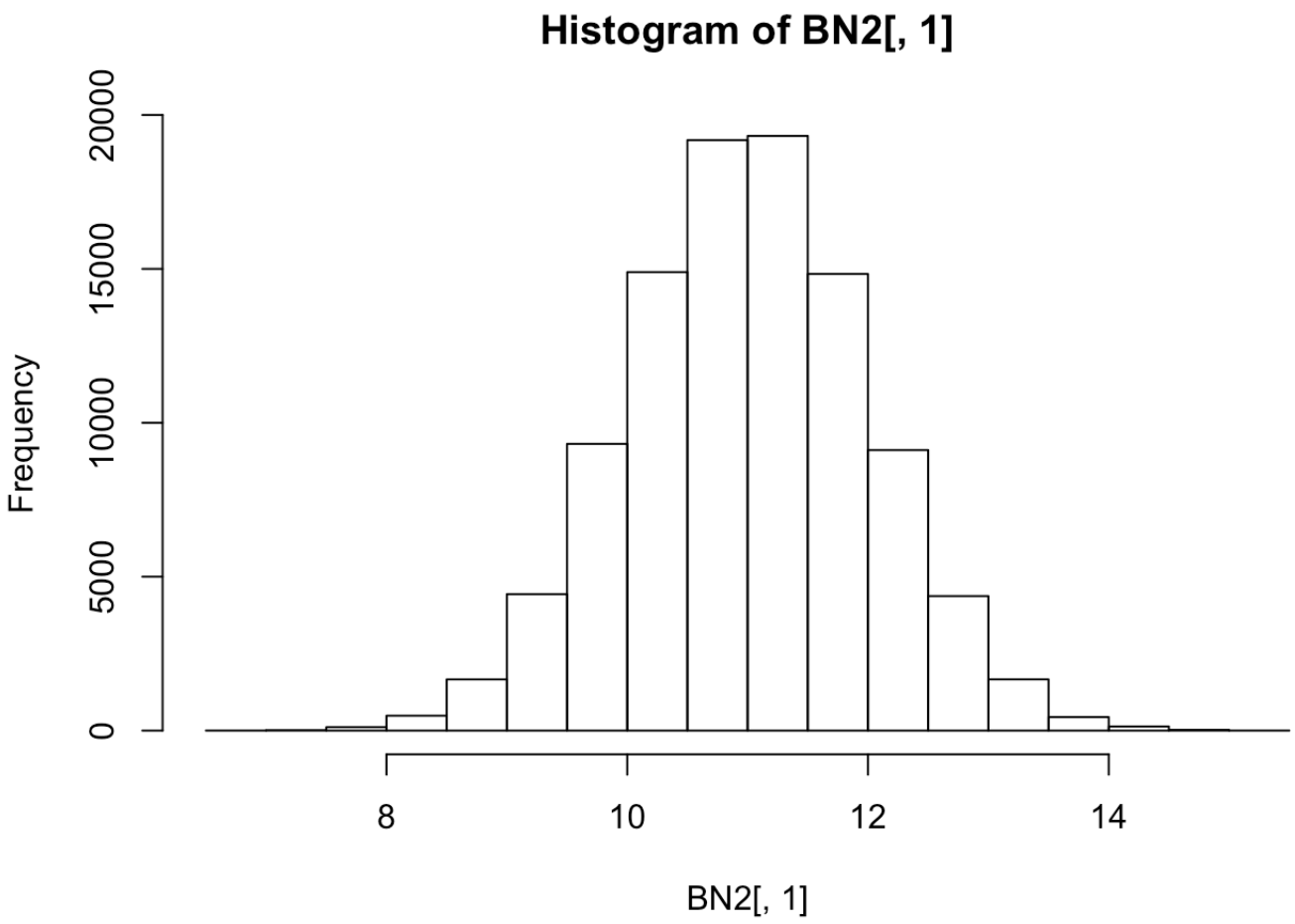
```
summary(BN2)
```

```
##        V1
##  Min.   : 6.638
##  1st Qu.:10.321
##  Median :10.998
##  Mean   :10.997
##  3rd Qu.:11.667
##  Max.   :15.161
```

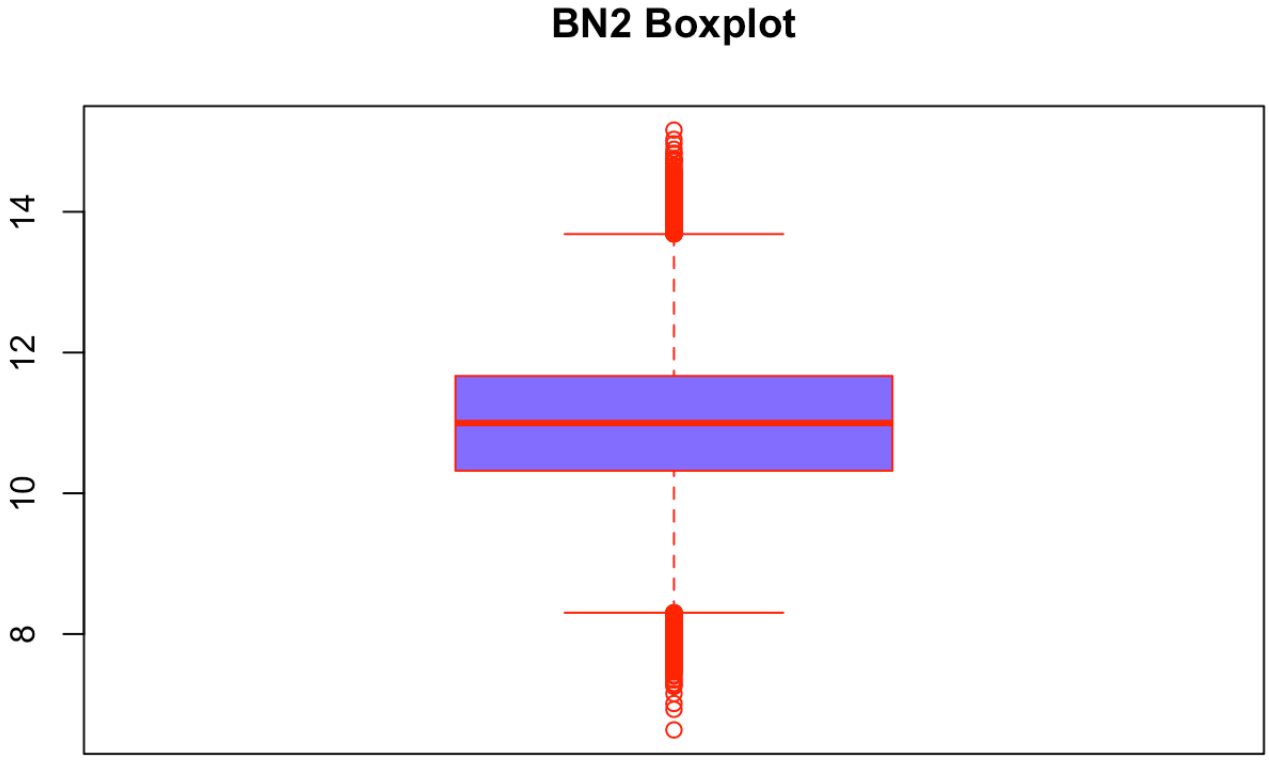
```
plot(BN2, main="BN2 Scatter Plot", col="purple")
```




```
hist(BN2[,1])
```




```
boxplot(BN2, data=BN2, main="BN2 Boxplot", col="slateblue1", border="red", fill="whit
e")
```




```
#BNOM
dim(BNOM) #shows the dimensions of the data
```

```
## [1] 1000    1
```

```
names(BNOM) #shows the names of each variable/attribute
```

```
## [1] "V1"
```

```
summary(BNOM) #summary of each variable/attribute
```

```
##        V1
##  Min.   :57.00
##  1st Qu.:67.00
##  Median :70.00
##  Mean   :70.17
##  3rd Qu.:73.00
##  Max.   :84.00
```

```
str(BNOM)
```

```
## 'data.frame':    1000 obs. of  1 variable:
##  $ V1: int  71 80 73 72 76 72 66 75 69 65 ...
```

```
BNOMmean = mean(BNOM[,1])
BNOMmean
```

```
## [1] 70.168
```

```
var(BNOM)
```

```
##          V1
## V1 21.98977
```

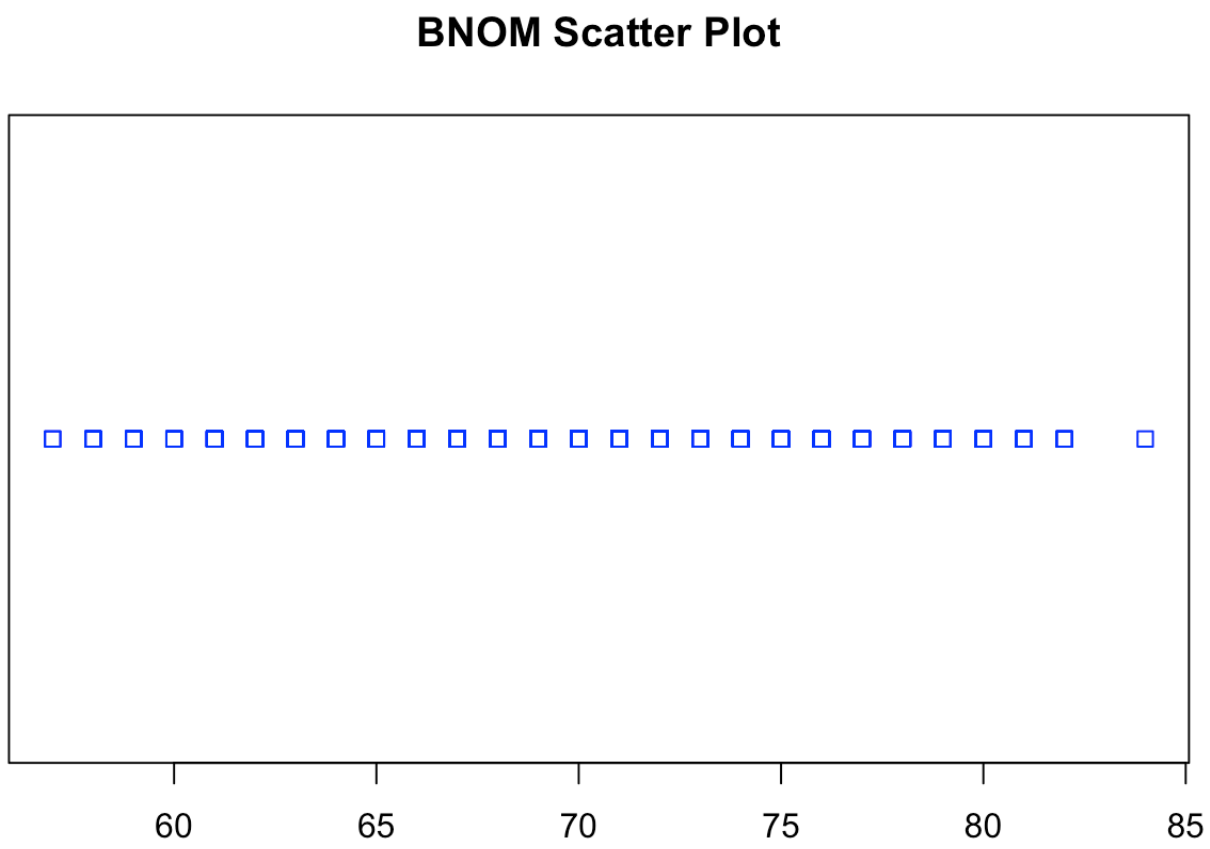
```
BNOMsd1 <- sqrt(sum((BNOM-BNOMmean)^2)/(1000-1))
BNOMsd1
```

```
## [1] 4.689325
```

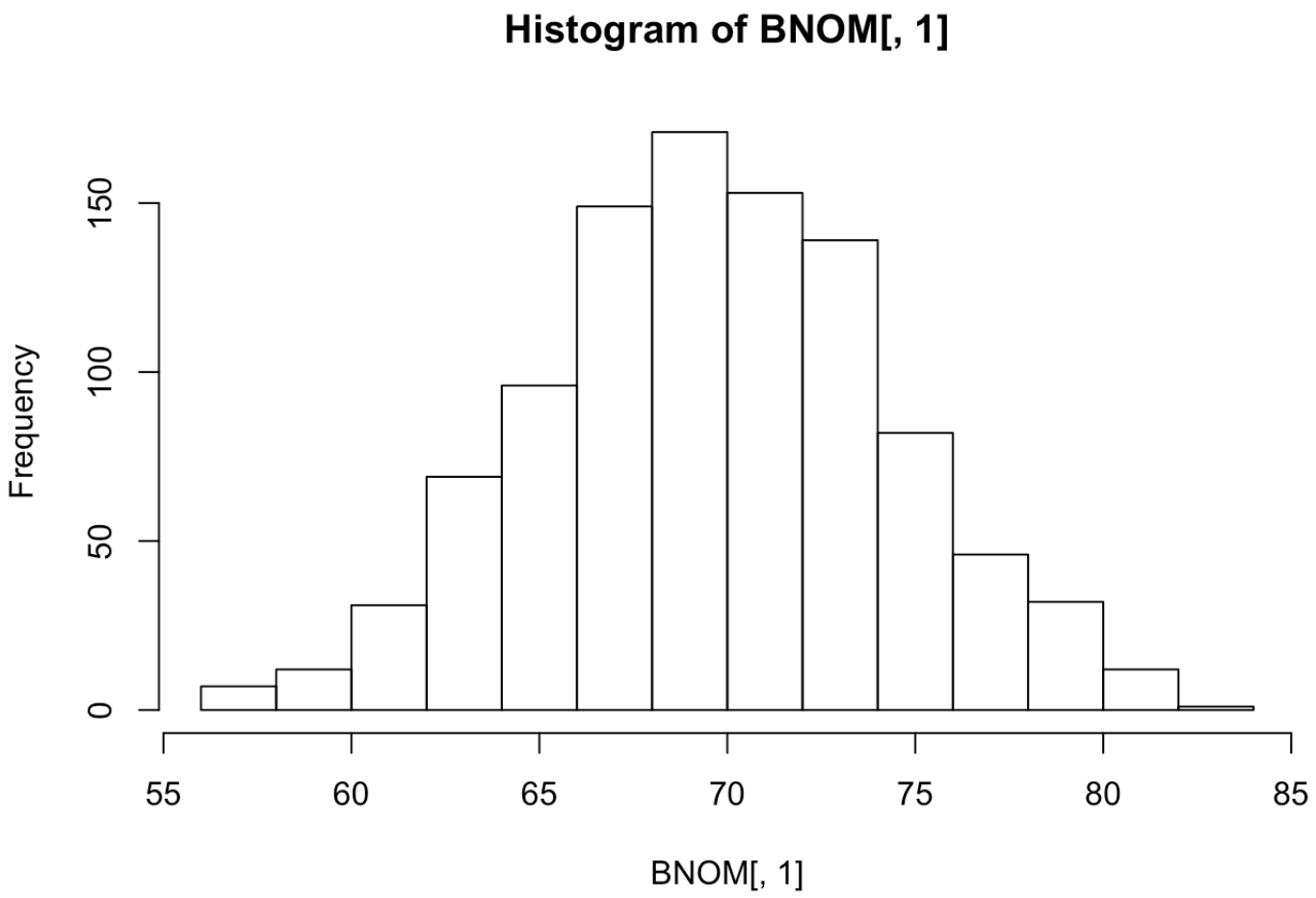
```
summary(BNOM)
```

```
##        V1
##  Min.   :57.00
##  1st Qu.:67.00
##  Median :70.00
##  Mean   :70.17
##  3rd Qu.:73.00
##  Max.   :84.00
```

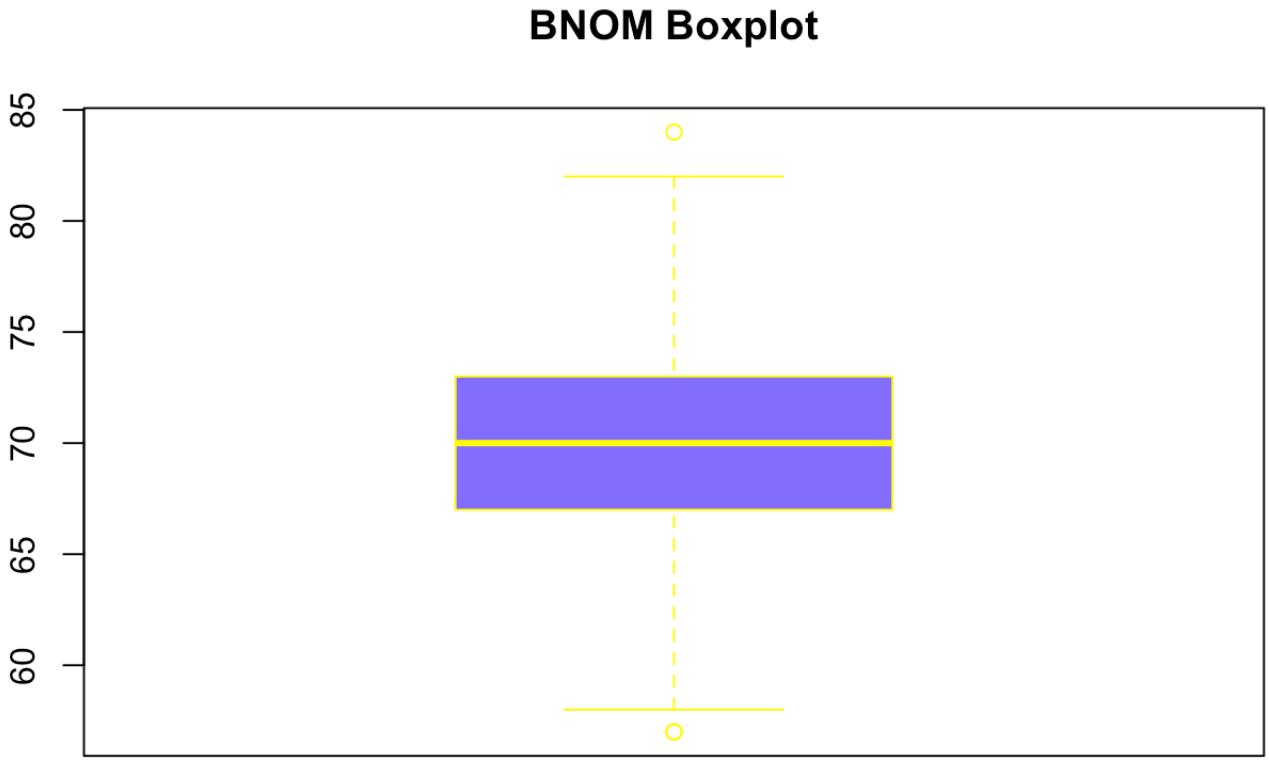
```
plot(BNOM, main="BNOM Scatter Plot", col="blue")
```




```
hist(BNOM[,1])
```




```
boxplot(BNOM, data=BNOM, main="BNOM Boxplot", col="slateblue1", border="yellow", fill
="white")
```




```
#IN
dim(IN) #shows the dimensions of the data
```

```
## [1] 1048576       1
```

```
names(IN) #shows the names of each variable/attribute
```

```
## [1] "V1"
```

```
summary(IN) #summary of each variable/attribute
```

```
##        V1
##  Min.   : 3.00
##  1st Qu.:16.00
##  Median :19.00
##  Mean   :18.99
##  3rd Qu.:22.00
##  Max.   :43.00
```

```
str(IN)
```

```
## 'data.frame':    1048576 obs. of  1 variable:
##  $ V1: int  20 25 13 12 20 18 19 14 24 17 ...
```

```
INmean = mean(IN[,1])
INmean
```

```
## [1] 18.99296
```

```
var(IN)
```

```
##          V1
## V1 19.03238
```

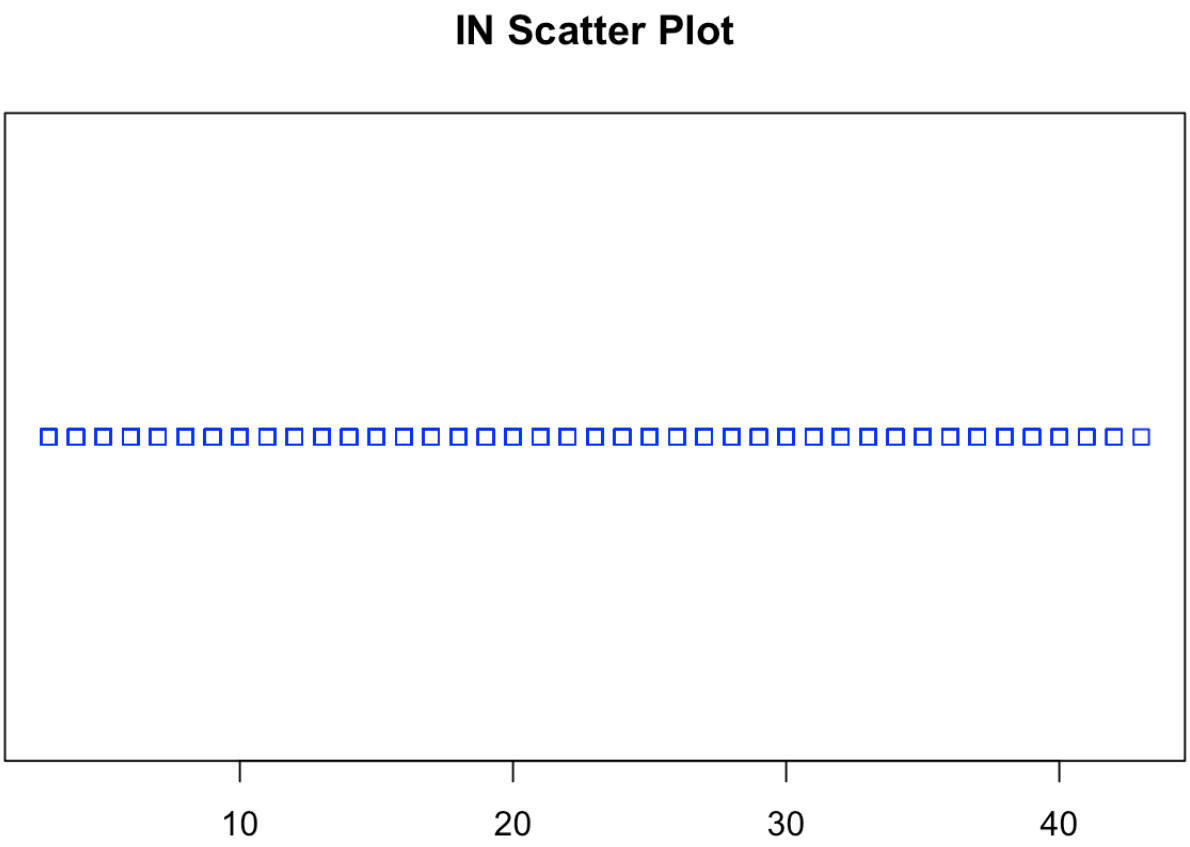
```
INsd1 <- sqrt(sum((IN-INmean)^2)/(1048576-1))
INsd1
```

```
## [1] 4.362612
```

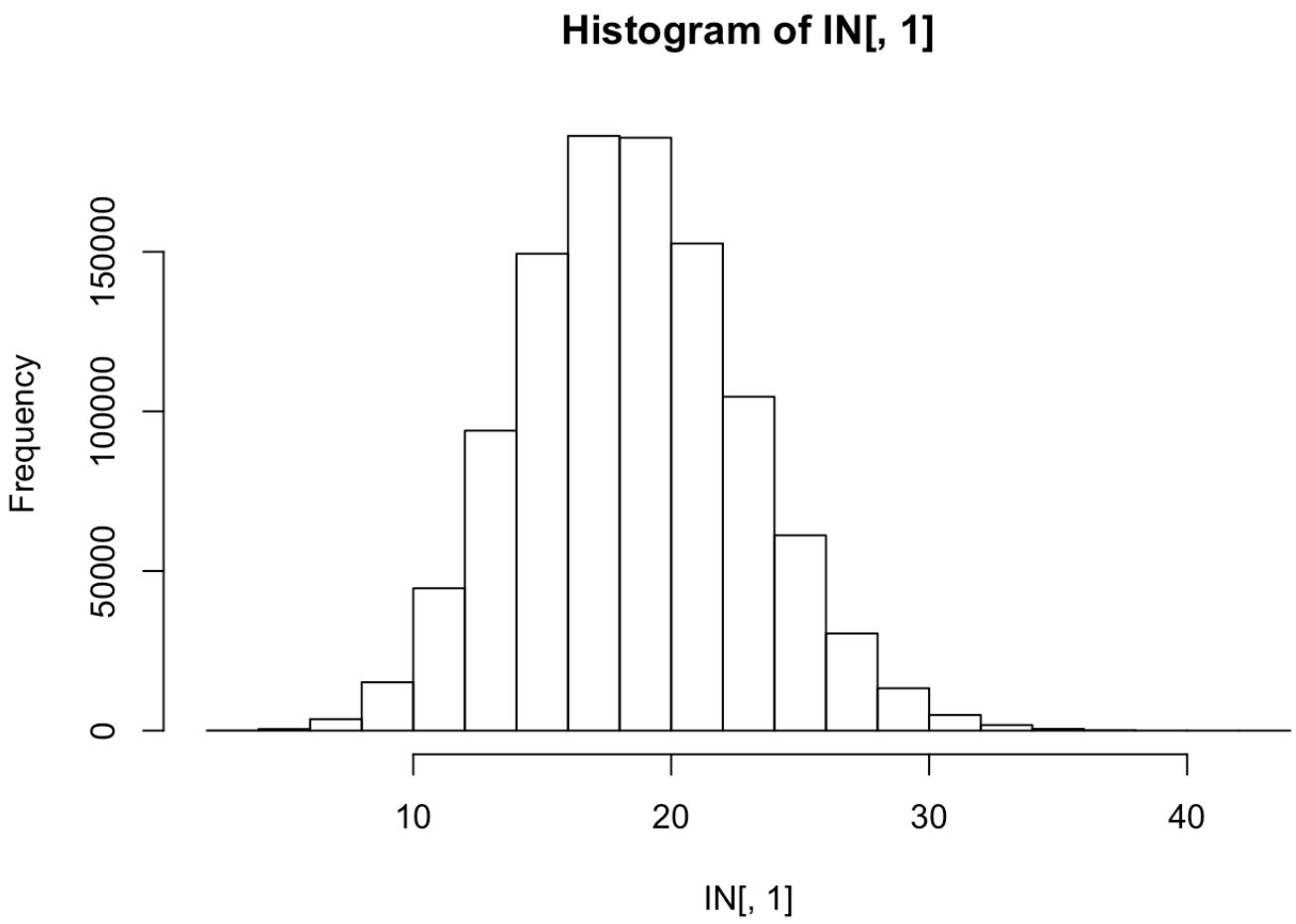
```
summary(IN)
```

```
##        V1
##  Min.   : 3.00
##  1st Qu.:16.00
##  Median :19.00
##  Mean   :18.99
##  3rd Qu.:22.00
##  Max.   :43.00
```

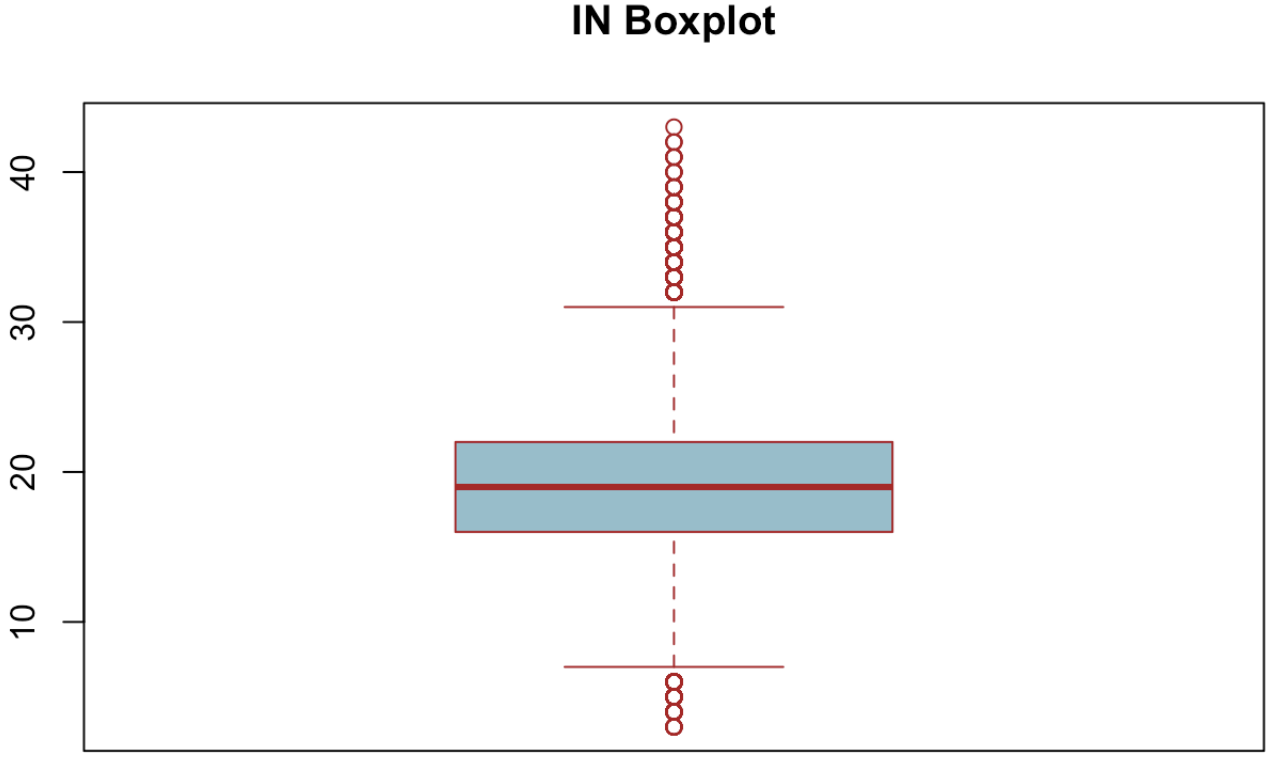
```
plot(IN, main="IN Scatter Plot", col="blue")
```




```
hist(IN[,1])
```




```
boxplot(IN, data=IN , main="IN Boxplot", col="lightblue3", border="brown", fill="whit
e")
```




```
#N1
dim(N1) #shows the dimensions of the data
```

```
## [1] 100   1
```

```
names(N1) #shows the names of each variable/attribute
```

```
## [1] "V1"
```

```
summary(N1) #summary of each variable/attribute
```

```
##        V1
##  Min.   : 5.064
##  1st Qu.: 8.774
##  Median :10.093
##  Mean   :10.134
##  3rd Qu.:11.458
##  Max.   :15.767
```

```
str(N1)
```

```
## 'data.frame':    100 obs. of  1 variable:
##  $ V1: num  13.57 10.99 9.82 6.87 8.3 ...
```

```
N1mean <- mean(N1[,1])
N1mean
```

```
## [1] 10.13424
```

```
var(N1)
```

```
##          V1
## V1 3.982841
```

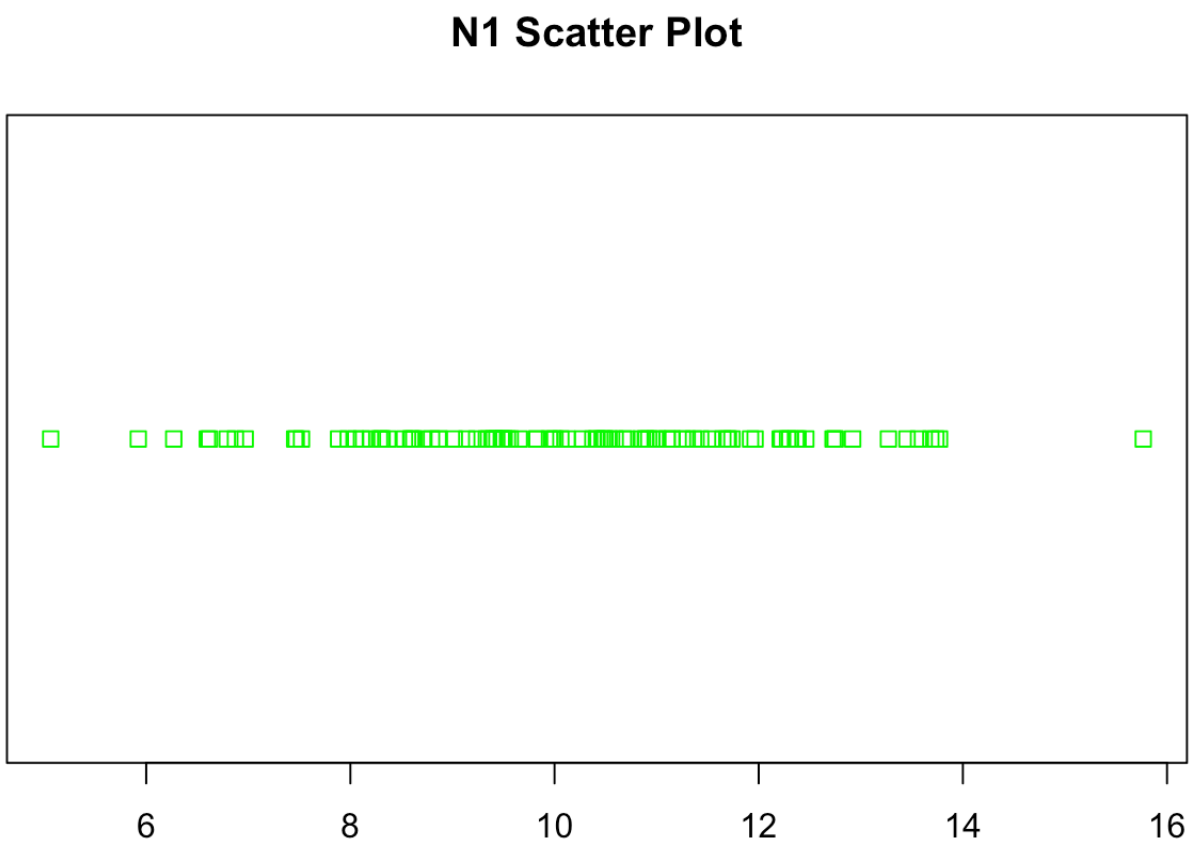
```
N1sd1 <- sqrt(sum((N1-N1mean)^2)/(100-1))
N1sd1
```

```
## [1] 1.995706
```

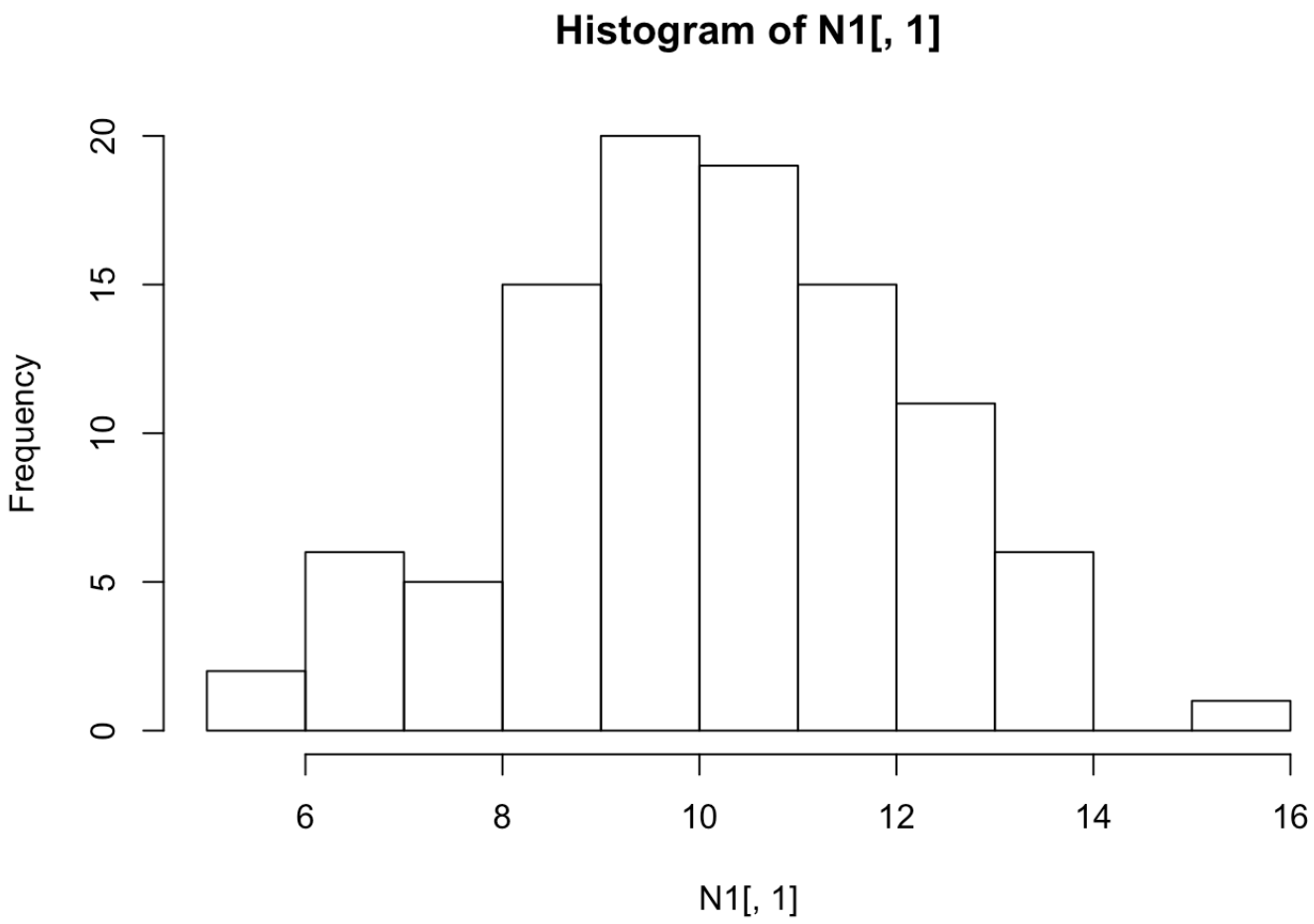
```
summary(N1)
```

```
##        V1
##  Min.   : 5.064
##  1st Qu.: 8.774
##  Median :10.093
##  Mean   :10.134
##  3rd Qu.:11.458
##  Max.   :15.767
```

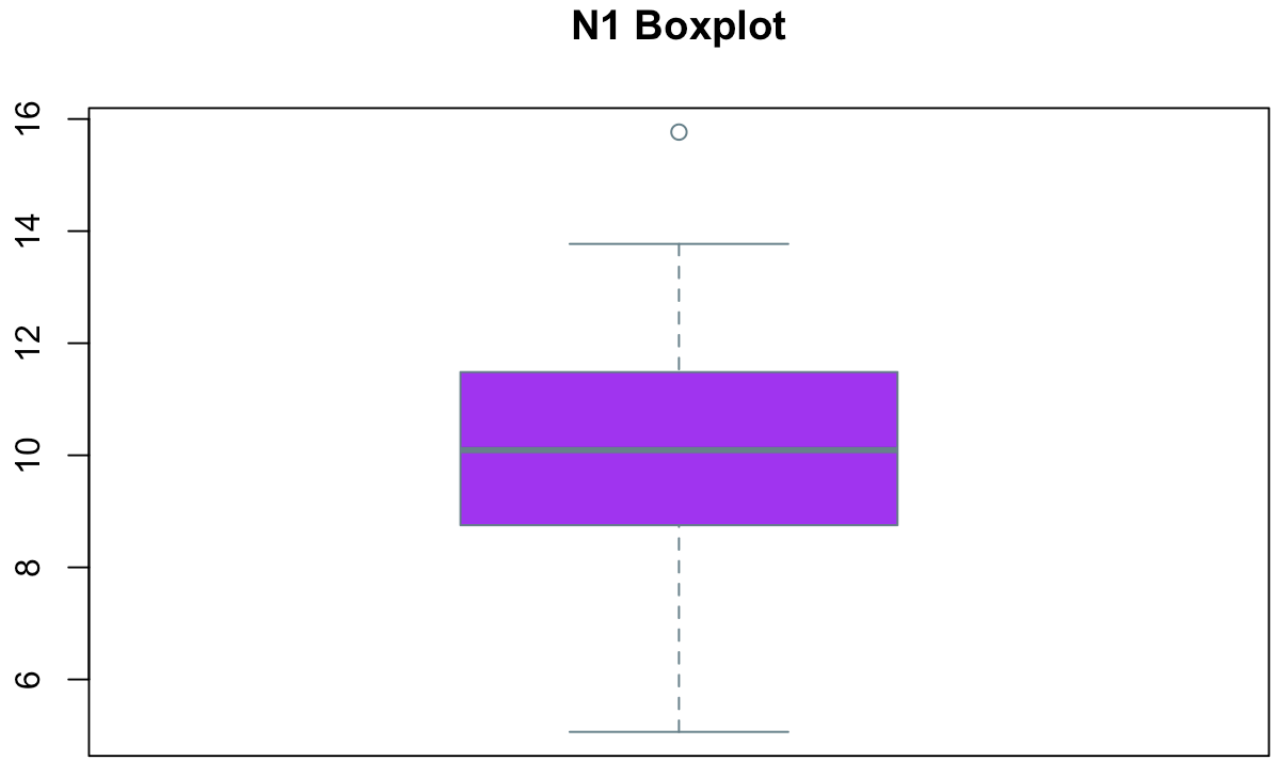
```
plot(N1, main="N1 Scatter Plot", col="green")
```




```
hist(N1[,1])
```




```
boxplot(N1, data=N1 , main="N1 Boxplot", col="purple", border="lightblue4",fill="whit
e")
```




```
#N2
dim(N2) #shows the dimensions of the data
```

```
## [1] 100   1
```

```
names(N2) #shows the names of each variable/attribute
```

```
## [1] "V1"
```

```
summary(N2) #summary of each variable/attribute
```

```
##        V1
##  Min.   : 0.6503
##  1st Qu.: 9.0030
##  Median :11.9056
##  Mean   :11.7082
##  3rd Qu.:14.1246
##  Max.   :22.6275
```

```
str(N2)
```

```
## 'data.frame':    100 obs. of  1 variable:
##  $ V1: num  14.2 13.2 12.6 14.2 13.3 ...
```

```
N2mean <- mean(N2[,1])
N2mean
```

```
## [1] 11.70822
```

```
var(N2)
```

```
##          V1
## V1 18.41505
```

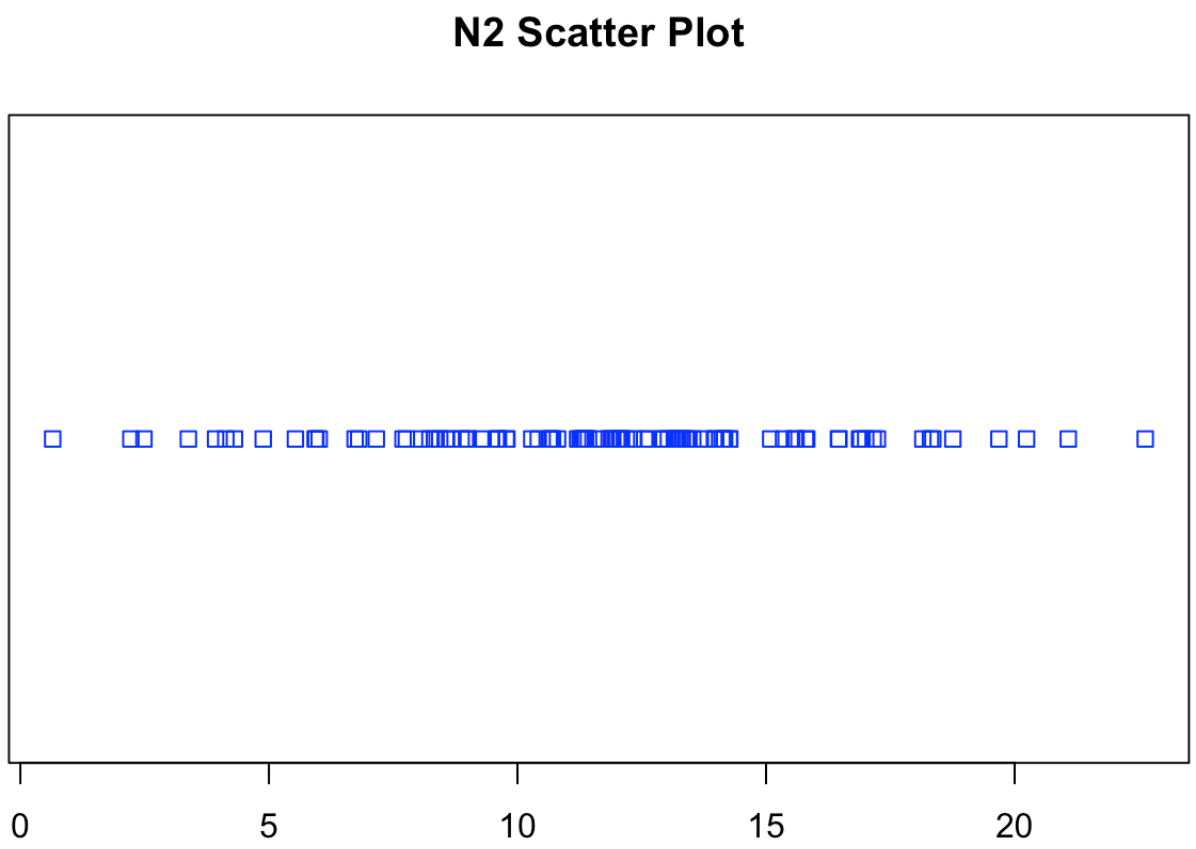
```
N2sd1 <- sqrt(sum((N2-N2mean)^2)/(100-1))
N2sd1
```

```
## [1] 4.291276
```

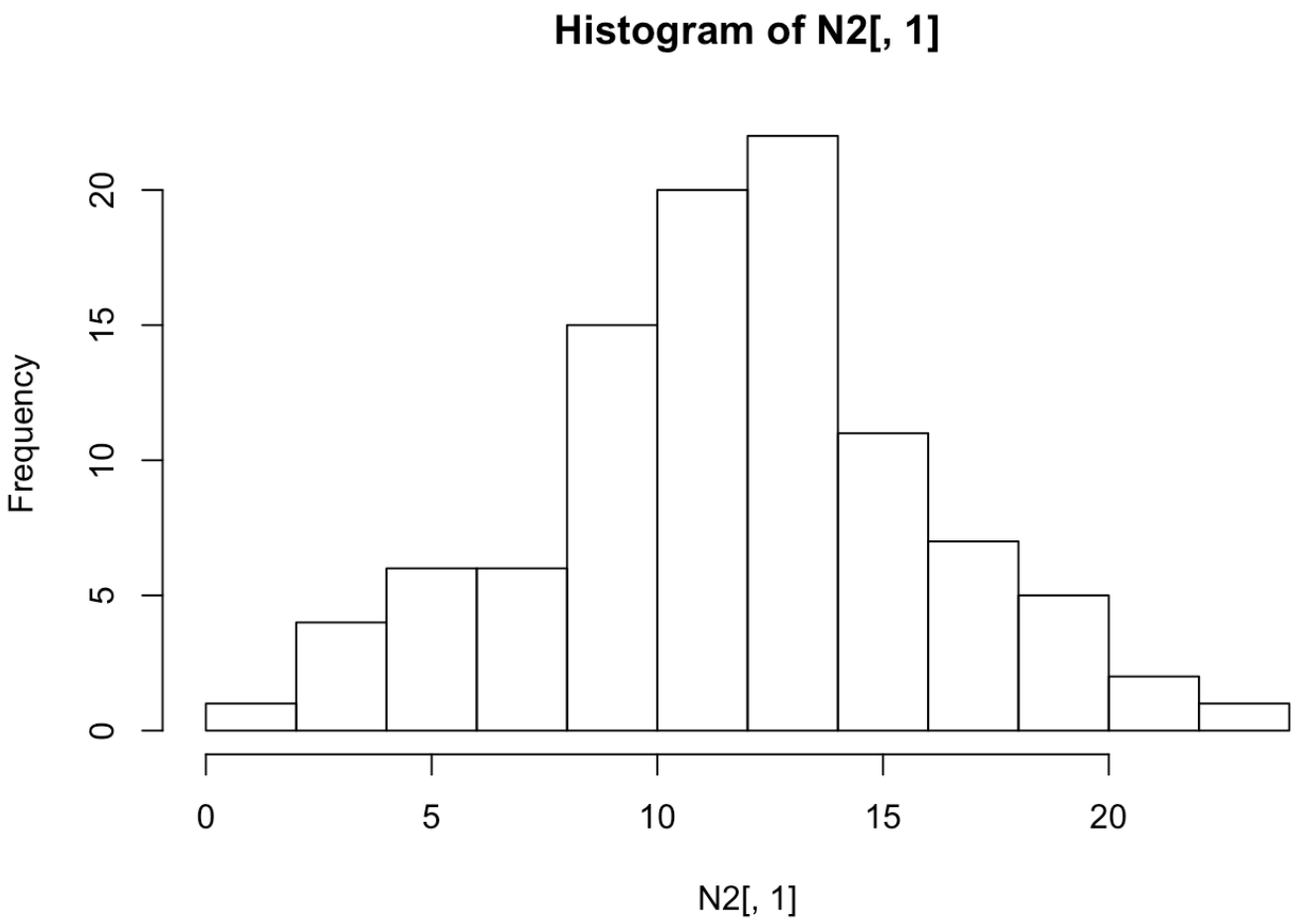
```
summary(N2)
```

```
##        V1
##  Min.   : 0.6503
##  1st Qu.: 9.0030
##  Median :11.9056
##  Mean   :11.7082
##  3rd Qu.:14.1246
##  Max.   :22.6275
```

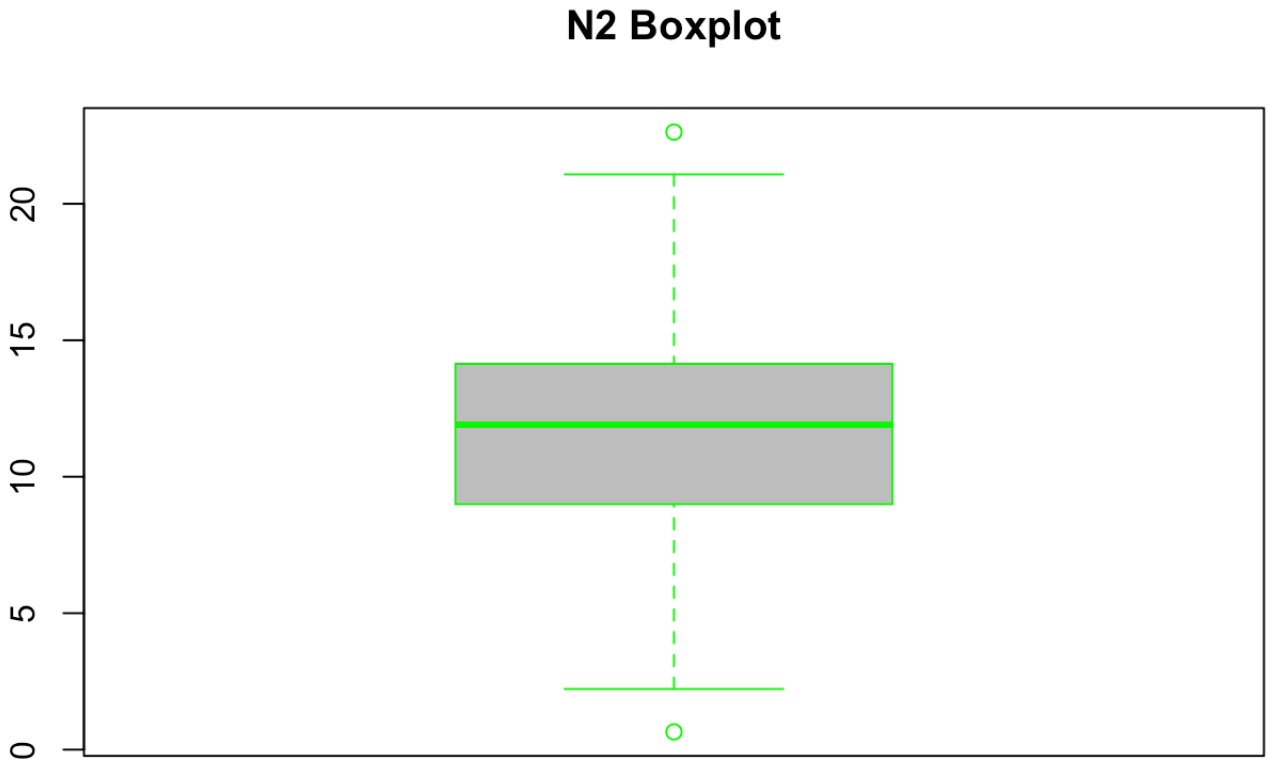
```
plot(N2, main="N2 Scatter Plot", col="blue")
```




```
hist(N2[,1])
```




```
boxplot(N2, data=N2, main="N2 Boxplot", col="gray",border="green",fill="white")
```



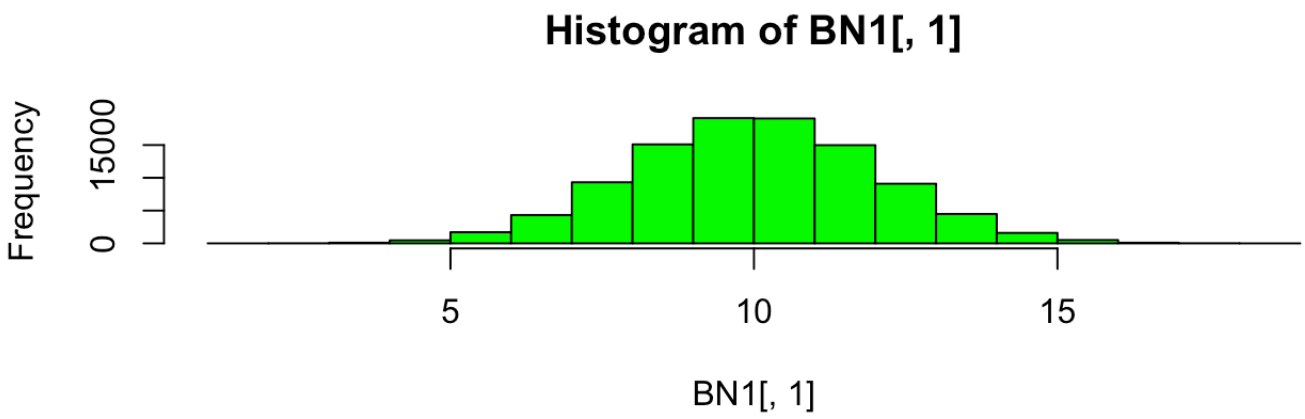

```
#BN1 and BN2 (Part 1): Comparative Statistics + Visualizations
cov(BN1,BN2) # Covariance-how variables move when compared to eachother.
```

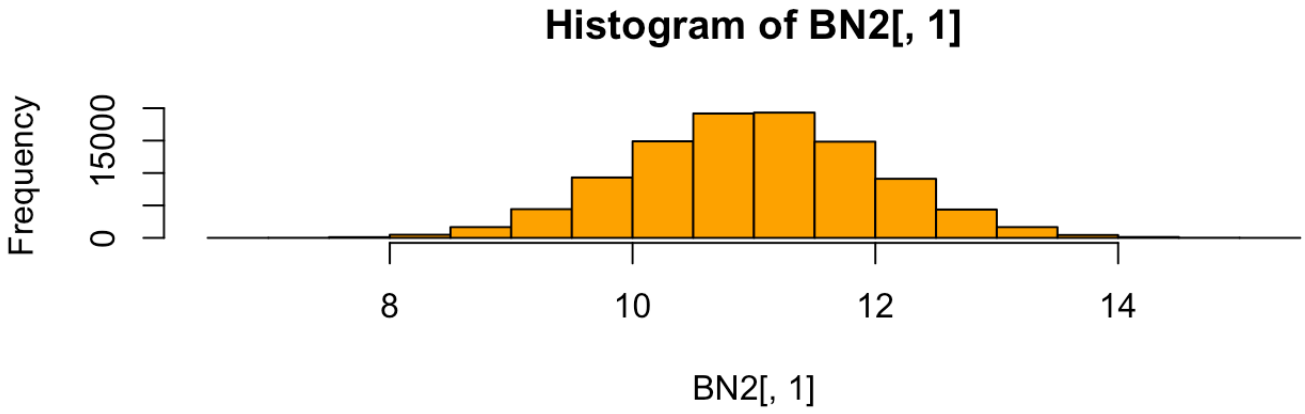
```
##             V1
## V1 0.001284024
```

```
cor(BN1,BN2) #correlation
```

```
##              V1
## V1 0.0006422369
```

```
par(mfrow=c(2,1))
hist(BN1[,1], col="green")
hist(BN2[,1], col="orange")
```






```
#BN1 and BN2 (Part 2): Statistics + Data Frame Manipulation
BN_cbind <- cbind(BN1,BN2)
colnames(BN_cbind) <- c("BN1_col","BN2_col")
names(BN_cbind)
```

```
## [1] "BN1_col" "BN2_col"
```

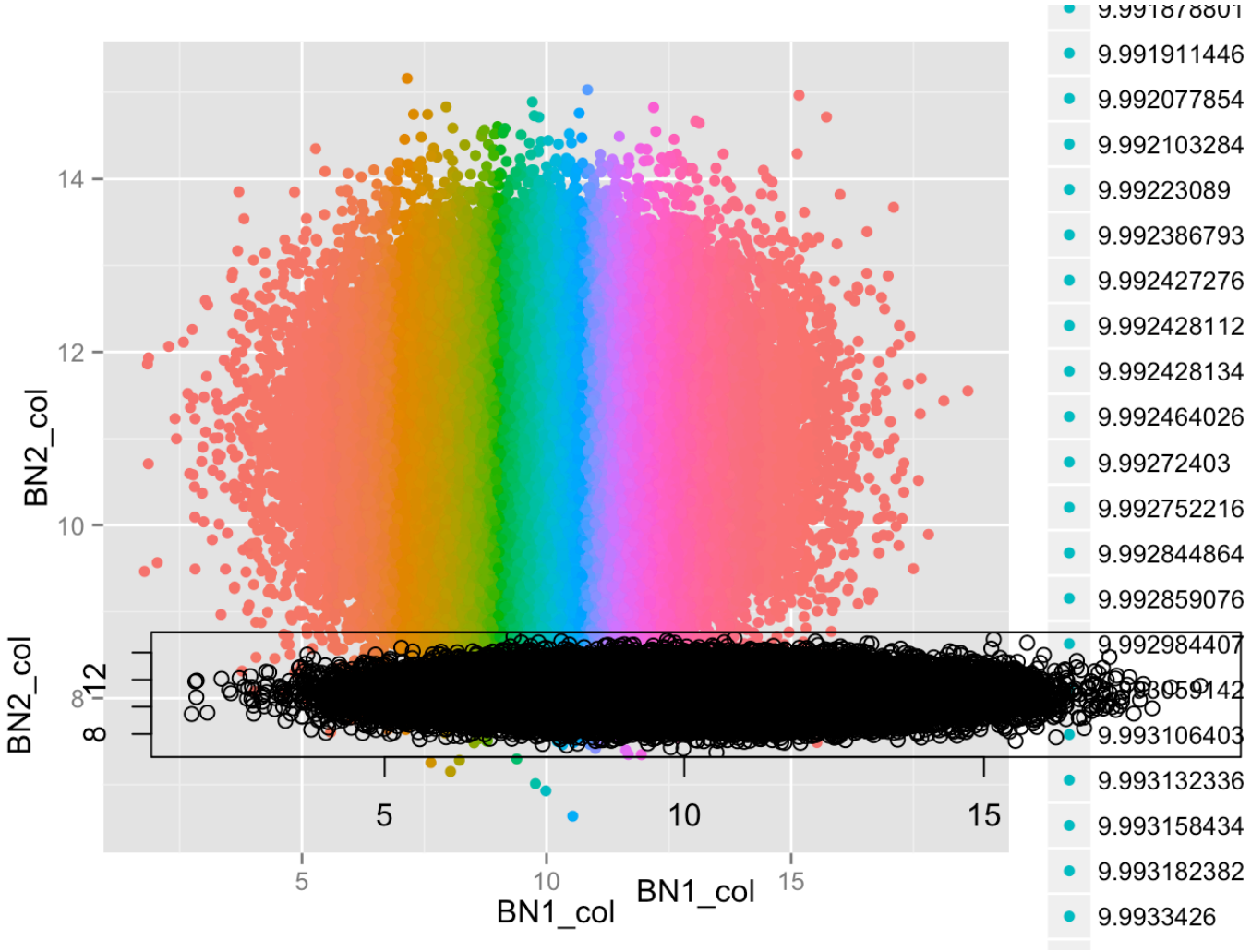
```
plot(BN1mean,BN2mean)

t.test(BN_cbind$BN1_col,BN_cbind$BN2_col,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  BN_cbind$BN1_col and BN_cbind$BN2_col
## t = -141.8112, df = 99999, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0163924 -0.9886801
## sample estimates:
## mean of the differences
##               -1.002536
```

```
library(ggplot2)
library(gridExtra)
```

```
## Loading required package: grid
```

```
p1 <- ggplot(BN_cbind, aes(x=BN1_col,y=BN2_col, main="BN1 Comparison"))
p1 + geom_point(aes(color=factor(BN1_col)))

plot(BN_cbind)
```



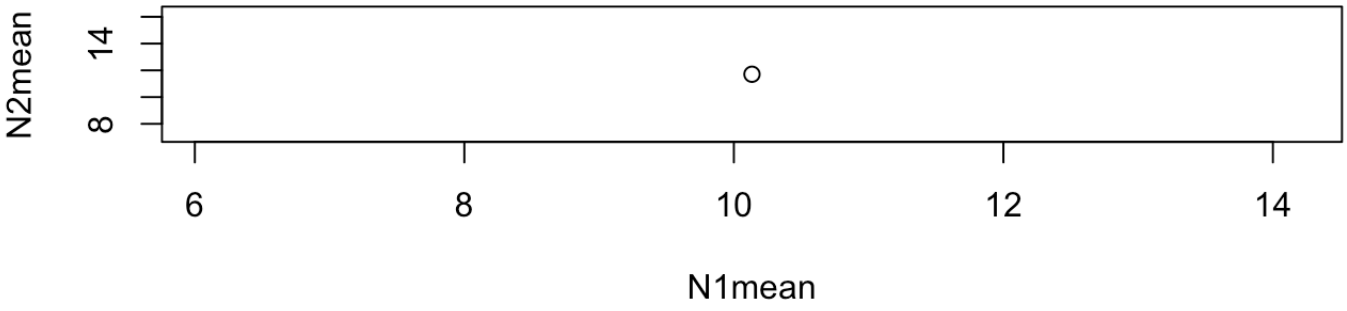

```
boxplot(BN_cbind, col=c("gold","blue"), border="orange", fill="gray88")

#N1 and N2 (Part 1): Comparative Statistics + Visualizations
cov(N1,N2) # Covariance-how variables move when compared to eachother.
```

```
##            V1
## V1 -0.6016144
```

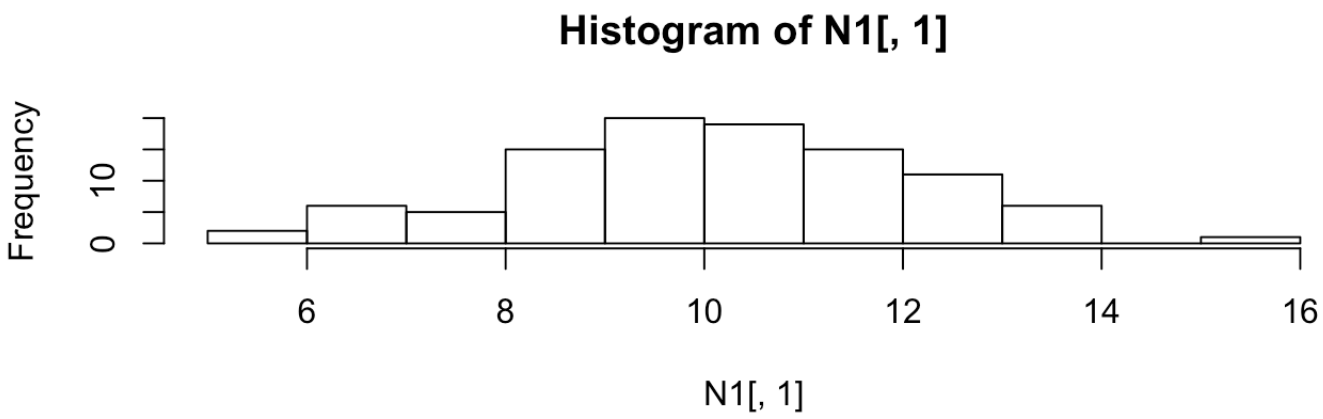
```
cor(N1,N2) #correlation
```

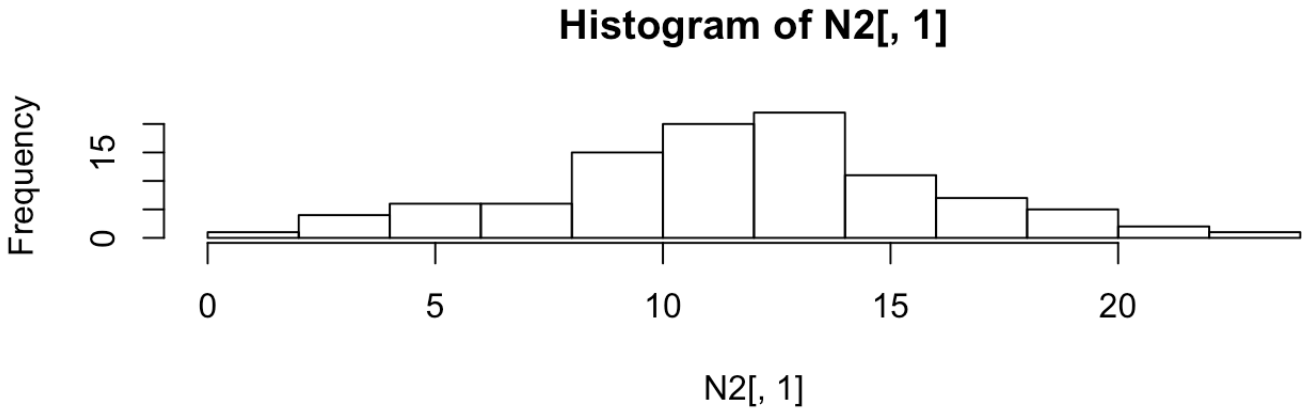
```
##             V1
## V1 -0.07024821
```

```
plot(N1mean,N2mean)
```




```
par(mfrow=c(2,1))
hist(N1[,1])
hist(N2[,1])
```

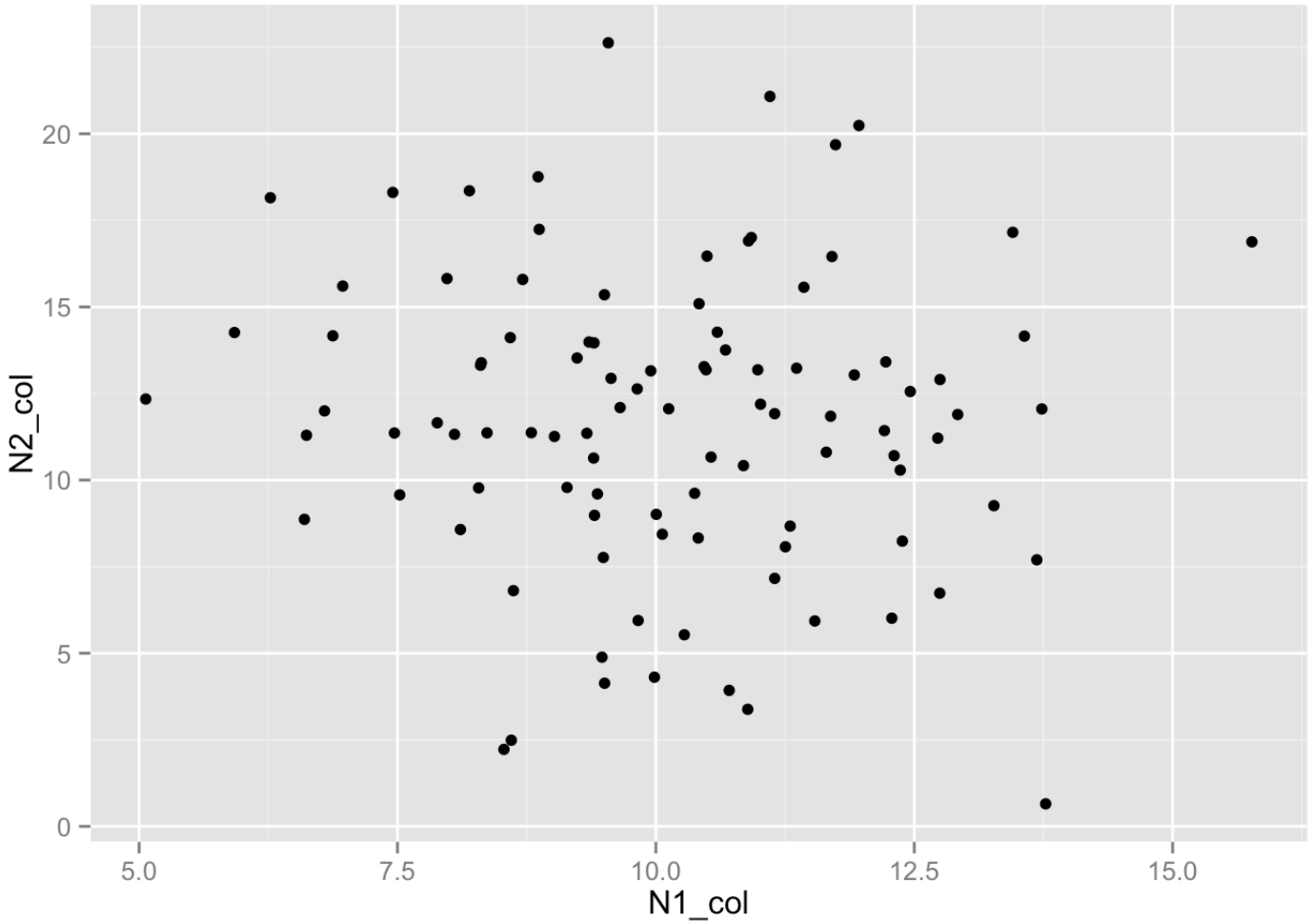





```
#N1 and N2 (Part 2): Statistics + Data Frame Manipulation
N_cbind <- cbind(N1,N2)
colnames(N_cbind) <- c("N1_col","N2_col")
names(N_cbind)
```

```
## [1] "N1_col" "N2_col"
```

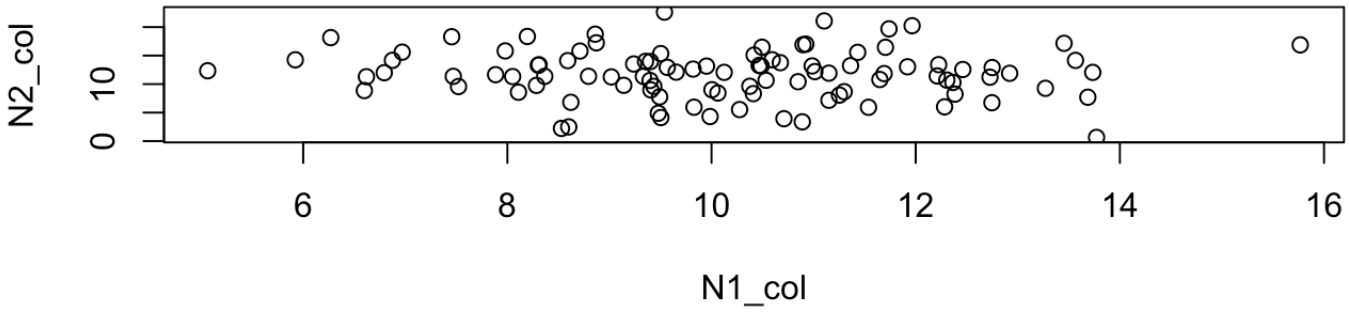
```
p2 <- ggplot(N_cbind, aes(x=N1_col,y=N2_col, main="N1 Comparison"))
p2 + geom_point()
```




```
t.test(N_cbind$N1_col,N_cbind$N2_col,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  N_cbind$N1_col and N_cbind$N2_col
## t = -3.2399, df = 99, p-value = 0.001629
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.5379347 -0.6100304
## sample estimates:
## mean of the differences
##               -1.573983
```

```
plot(N_cbind)
boxplot(N_cbind, col=c("black","blue"), border="lightgreen", fill="gray88")
```


N2_col
0
10
6
8
10
12
14
16
N1_col


0
10
N1_col
N2_col