# Template for Establishing Normality

James Long, Shyam Attiganal, Manoj Remela, Eric Young

September 26, 2019

---

## Intro

This template shows how to determine if data are Normally distributed. Three examples are included.

- The first is **not** Normally distributed (Data1).
- The second is Normally distributed (Data4).
- The third is a Normally distributed sampling distribution based on source data that are **not** Normally distributed (Data3).

File: 6304W Assignment 1 Data.xlsx
Sheet: Four Data Sets

```
# setup environment and load data
setwd(params$wd)
library(readxl)
# skewness and kurtosis module
library(moments)
# pretty table output
library(grid)
library(gridExtra)
a1df=read_excel("6304W Assignment 1 Data.xlsx",sheet="Four Data Sets")
colnames(a1df)=tolower(colnames(a1df))
```

## Data1

Column 1 data are **not** Normally distributed as illustrated by the following analysis.

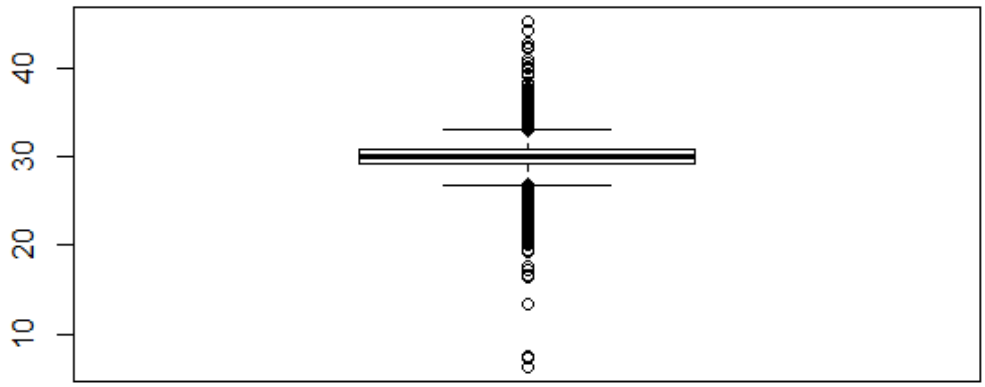
```
# compare mean and median
# these must be the same in a Normal dist.
# they might be the same in a non-Normal Dist., but if they are not, the
dist. is not Normal

# skewness should be near 0
```

```
# kurtosis should be near 3

# standard deviation alone is not informative
# comparing to range gives a sense of the (ab)normality

desc.stats <- data.frame(Mean = round(mean(a1df$data1),digits=2),
                 Median = round(median(a1df$data1),digits=2),
                 Skew = round(skewness(a1df$data1),digits=2),
                 Kurtosis = round(kurtosis(a1df$data1),digits=2),
                 SD3L = round(mean(a1df$data1)-(sd(a1df$data1)*3),digits=2),
                 Min = round(min(a1df$data1),digits=2),
                 SD3R = round(mean(a1df$data1)+(sd(a1df$data1)*3),digits=2),
                 Max = round(max(a1df$data1),digits=2),
                 row.names = "")
grid.newpage()
grid.table(desc.stats)
```

| Mean | Median | Skew | Kurtosis | SD3L | Min | SD3R | Max |
|---|---|---|---|---|---|---|---|
| 29.96 | 29.98 | -0.99 | 20.86 | 24.8 | 6.32 | 35.11 | 45.25 |

- The mean and median are nearly the same, so we must do more tests.
- Skewness is significantly below 0, indicating left skew.
- Kurtosis is significantly above 3, indicating heavy tails.

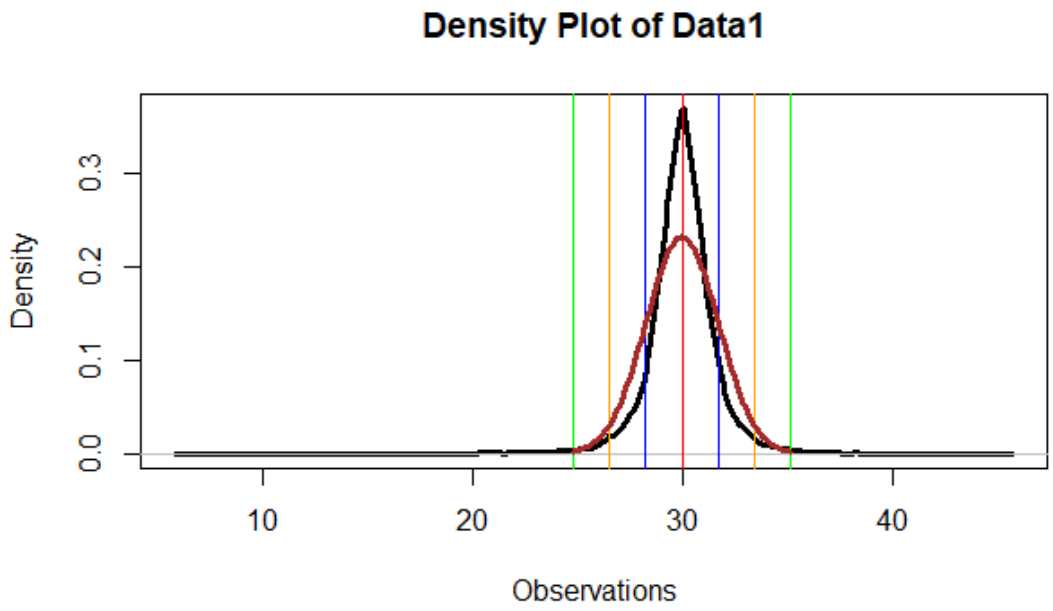
```
boxplot(a1df$data1,xlab="Data1 Boxplot")
```




Data1 Boxplot

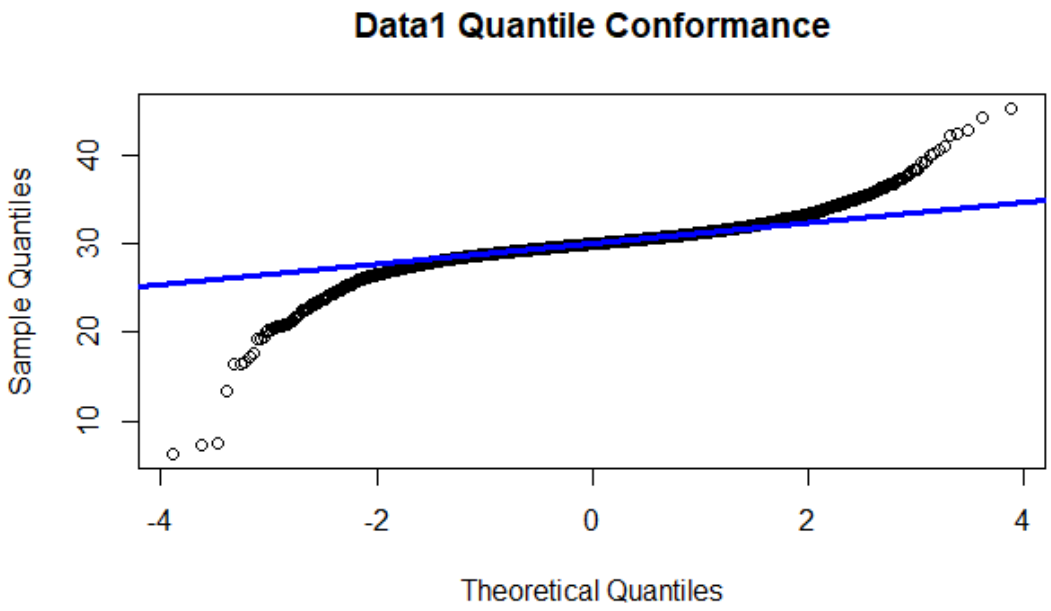
- The boxplot shows many observations outside the upper & lower whiskers (IQR x 1.5).
- The range of observations outside the whiskers is extreme relative to the range of the IQR.
- This indicates the data are likely not Normally distributed.

```
plot(density(a1df$data1),lwd=3,main="Density Plot of
Data1",xlab="Observations")
abline(v=(mean(a1df$data1)-(sd(a1df$data1)*3)),col="green")
abline(v=(mean(a1df$data1)-(sd(a1df$data1)*2)),col="orange")
abline(v=(mean(a1df$data1)-(sd(a1df$data1))),col="blue")
abline(v=mean(a1df$data1),col="red")
abline(v=(mean(a1df$data1)+(sd(a1df$data1))),col="blue")
abline(v=(mean(a1df$data1)+(sd(a1df$data1)*2)),col="orange")
abline(v=(mean(a1df$data1)+(sd(a1df$data1)*3)),col="green")
curve(dnorm(x,mean(a1df$data1),sd(a1df$data1)),from=(mean(a1df$data1)-
(sd(a1df$data1)*3)),to=(mean(a1df$data1)+(sd(a1df$data1)*3)),lwd=3,col="brown
",add=T)
```




- The density plot shows a tall, narrow curve with very long tails (black).
- Using the mean and SD of the data, a Normal curve is plotted in brown.
- Clearly, the data are not Normally distributed.

```
qqnorm(a1df$data1,main="Data1 Quantile Conformance")
qqline(a1df$data1,lwd=3,col="blue")
```




- The qqnorm plot is not an upward-sloping diagonal line.
- Clearly, the data are not Normally distributed.

```
d1m=mean(a1df$data1)
d1sd=sd(a1df$data1)
a1df.d1.quants=data.frame(c(quantile(a1df$data1,probs=c(pnorm(c(-3,-2,-
1,0,1,2,3),mean=0,sd=1)))),c(d1m-d1sd*3,d1m-d1sd*2,d1m-
d1sd,d1m,d1m+d1sd,d1m+d1sd*2,d1m+d1sd*3))
colnames(a1df.d1.quants)=c("Observed Values","Predicted Values")
rownames(a1df.d1.quants)=c("-3 sigma","-2 sigma","-1 sigma","mean","1
sigma","2 sigma","3 sigma")
grid.newpage()
grid.table(a1df.d1.quants)
```

| | Observed Values | Predicted Values |
|---|---|---|
| *-3 sigma* | 20.3259576329264 | 24.8044690844027 |
| *-2 sigma* | 26.4865688265578 | 26.5227562293821 |
| *-1 sigma* | 28.7658994919112 | 28.2410433743615 |
| *mean* | 29.9831863205593 | 29.9593305193408 |
| *1 sigma* | 31.2011271108461 | 31.6776176643202 |
| *2 sigma* | 33.2938230263753 | 33.3959048092996 |
| *3 sigma* | 38.2849822244585 | 35.114191954279 |

- We can also see the qqnorm result numerically. When we compare the observed values at each sigma with the theoretical values at each sigma, we see the observed values diverge from the predicted values in the tails.

## Data4

Column 4 data are nearly Normally distributed as illustrated by the following analysis.

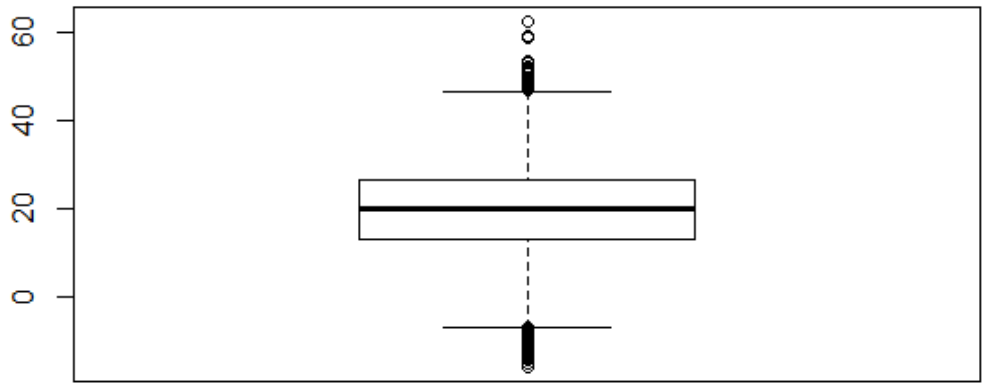
```
# compare mean and median
# these must be the same in a Normal dist.
# they might be the same in a non-Normal Dist., but if they are not, the
dist. is not Normal

# skewness should be near 0
# kurtosis should be near 3

# standard deviation alone is not informative
# comparing to range gives a sense of the (ab)normality

desc.stats <- data.frame(Mean = round(mean(a1df$data4),digits=2),
                         Median = round(median(a1df$data4),digits=2),
```

```
                Skew = round(skewness(a1df$data4),digits=2),
                Kurtosis = round(kurtosis(a1df$data4),digits=2),
                SD3L = round(mean(a1df$data4)-(sd(a1df$data4)*3),digits=2),
                Min = round(min(a1df$data4),digits=2),
                SD3R = round(mean(a1df$data4)+(sd(a1df$data4)*3),digits=2),
                Max = round(max(a1df$data4),digits=2),
                row.names = "")
grid.newpage()
grid.table(desc.stats)
```

| Mean | Median | Skew | Kurtosis | SD3L | Min | SD3R | Max |
|---|---|---|---|---|---|---|---|
| 19.85 | 19.83 | 0.03 | 3.07 | -10.03 | -15.99 | 49.74 | 62.62 |

- The mean and median are nearly the same.
- Skewness is close to 0.
- Kurtosis is close to 3.

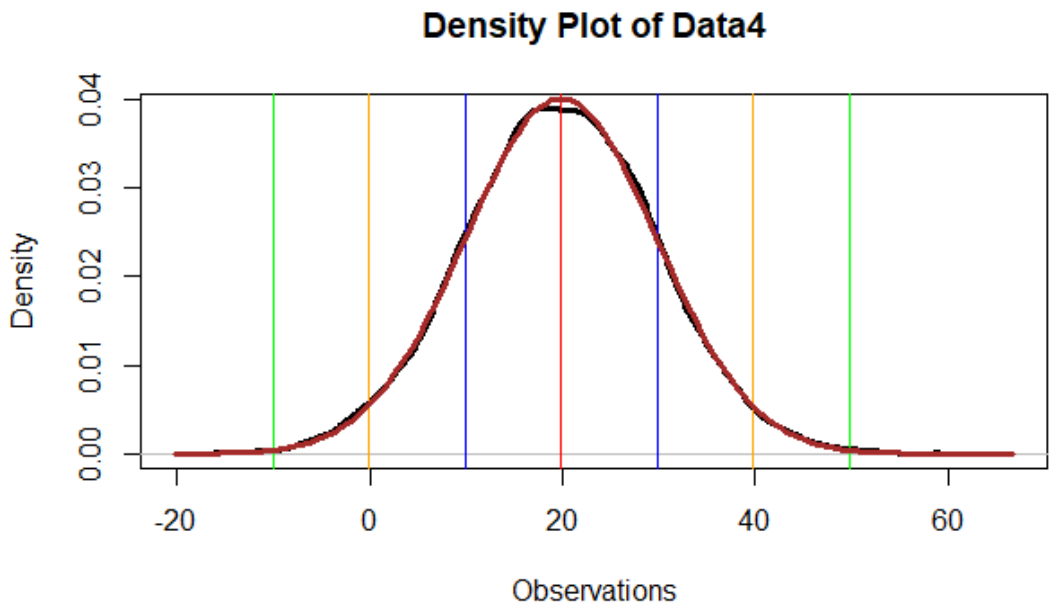
```
boxplot(a1df$data4,xlab="Data4 Boxplot")
```




Data4 Boxplot

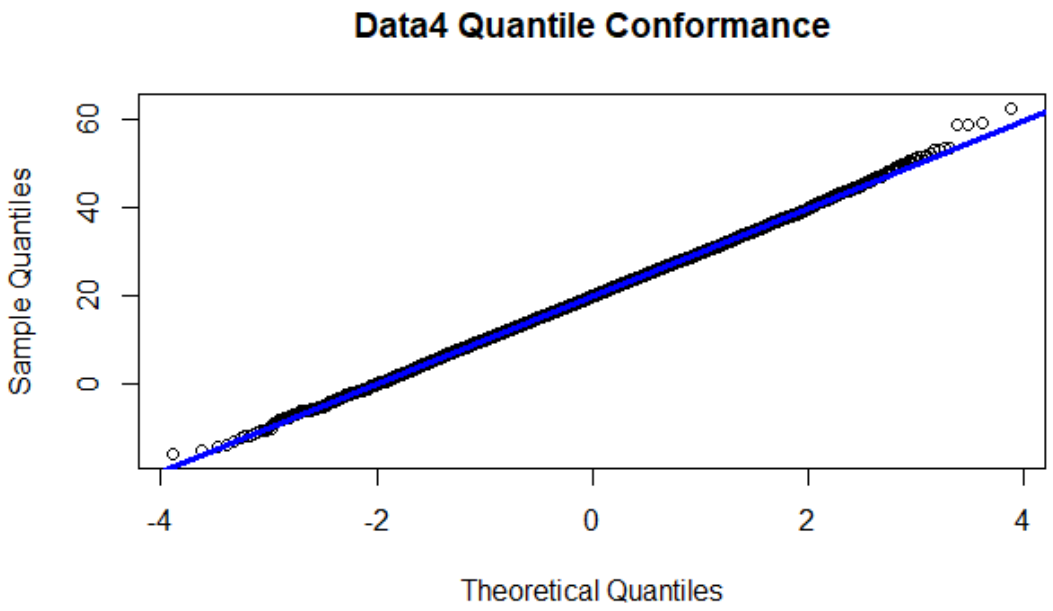
- The boxplot shows good symmetry both in the whiskers and the number of observations outside the whiskers.
- This indicates the data are Normally distributed.

```
plot(density(a1df$data4),lwd=3,main="Density Plot of
Data4",xlab="Observations")
abline(v=(mean(a1df$data4)-(sd(a1df$data4)*3)),col="green")
abline(v=(mean(a1df$data4)-(sd(a1df$data4)*2)),col="orange")
abline(v=(mean(a1df$data4)-(sd(a1df$data4))),col="blue")
abline(v=mean(a1df$data4),col="red")
abline(v=(mean(a1df$data4)+(sd(a1df$data4))),col="blue")
abline(v=(mean(a1df$data4)+(sd(a1df$data4)*2)),col="orange")
abline(v=(mean(a1df$data4)+(sd(a1df$data4)*3)),col="green")
curve(dnorm(x,mean(a1df$data4),sd(a1df$data4)),lwd=3,col="brown",add=T)
```




- The density plot shows a mostly normal curve (black).
- Using the mean and SD of the data, a Normal curve is plotted in brown.
- Clearly, the data are Normally distributed.

```
qqnorm(a1df$data4,main="Data4 Quantile Conformance")
qqline(a1df$data4,lwd=3,col="blue")
```




- The qqnorm plot is an upward-sloping diagonal line.
- Clearly, the data are Normally distributed.

```
d4m=mean(a1df$data4)
d4sd=sd(a1df$data4)
a1df.d4.quants=data.frame(c(quantile(a1df$data4,probs=c(pnorm(c(-3,-2,-
1,0,1,2,3),mean=0,sd=1)))),c(d4m-d4sd*3,d4m-d4sd*2,d4m-
d4sd,d4m,d4m+d4sd,d4m+d4sd*2,d4m+d4sd*3))
colnames(a1df.d4.quants)=c("Observed Values","Predicted Values")
rownames(a1df.d4.quants)=c("-3 sigma","-2 sigma","-1 sigma","mean","1
sigma","2 sigma","3 sigma")
grid.newpage()
grid.table(a1df.d4.quants)
```

| | Observed Values | Predicted Values |
|---|---|---|
| *-3 sigma* | -10.3233374375755 | -10.0344320721259 |
| *-2 sigma* | -0.183010205016235 | -0.071781320004586 |
| *-1 sigma* | 9.90820730219879 | 9.89086943211678 |
| *mean* | 19.8349255877311 | 19.8535201842381 |
| *1 sigma* | 29.7111338081092 | 29.8161709363595 |
| *2 sigma* | 39.7698066174126 | 39.7788216884809 |
| *3 sigma* | 51.254375033977 | 49.7414724406022 |

- We can also see the qqnorm result numerically. When we compare the observed values at each sigma with the theoretical values at each sigma, we see the observed values are close to the predicted values.

## Data3 Sampling Distribution

Column 3 data are **not** Normally distributed. However, the sampling distribution for Data3 is nearly Normal with a sample size of 50 as illustrated by the following analysis.

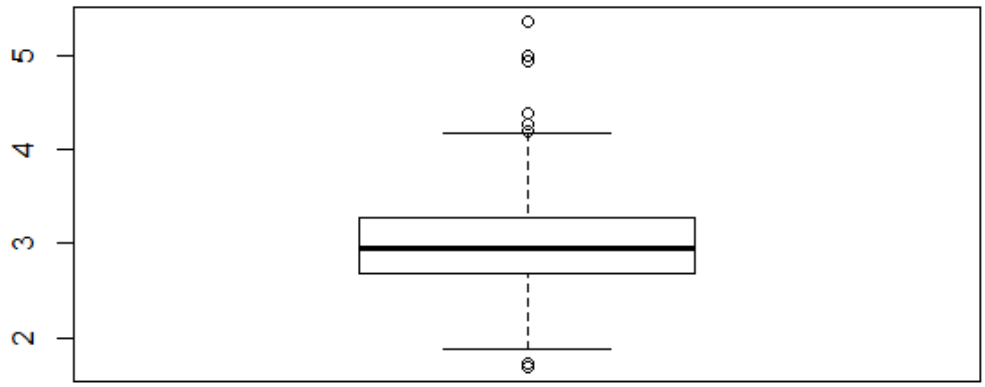
```
# create sampling distribution
df3=data.frame()
for(i in 1:1000){
  df3[1:50,i]=sample(a1df$data3,50,replace=F)
}
df3.means=colMeans(df3)

# compare mean and median
# these must be the same in a Normal dist.
# they might be the same in a non-Normal Dist., but if they are not, the
dist. is not Normal
```

```
# skewness should be near 0
# kurtosis should be near 3

# standard deviation alone is not informative
# comparing to range gives a sense of the (ab)normality

desc.stats <- data.frame(Mean = round(mean(df3.means),digits=2),
                Median = round(median(df3.means),digits=2),
                Skew = round(skewness(df3.means),digits=2),
                Kurtosis = round(kurtosis(df3.means),digits=2),
                SD3L = round(mean(df3.means)-(sd(df3.means)*3),digits=2),
                Min = round(min(df3.means),digits=2),
                SD3R = round(mean(df3.means)+(sd(df3.means)*3),digits=2),
                Max = round(max(df3.means),digits=2),
                row.names = "")
grid.newpage()
grid.table(desc.stats)
```

| Mean | Median | Skew | Kurtosis | SD3L | Min | SD3R | Max |
|---|---|---|---|---|---|---|---|
| 3 | 2.96 | 0.47 | 3.94 | 1.63 | 1.69 | 4.36 | 5.37 |

- The mean and median are nearly the same.
- Skewness is close to 0.
- Kurtosis is close to 3.

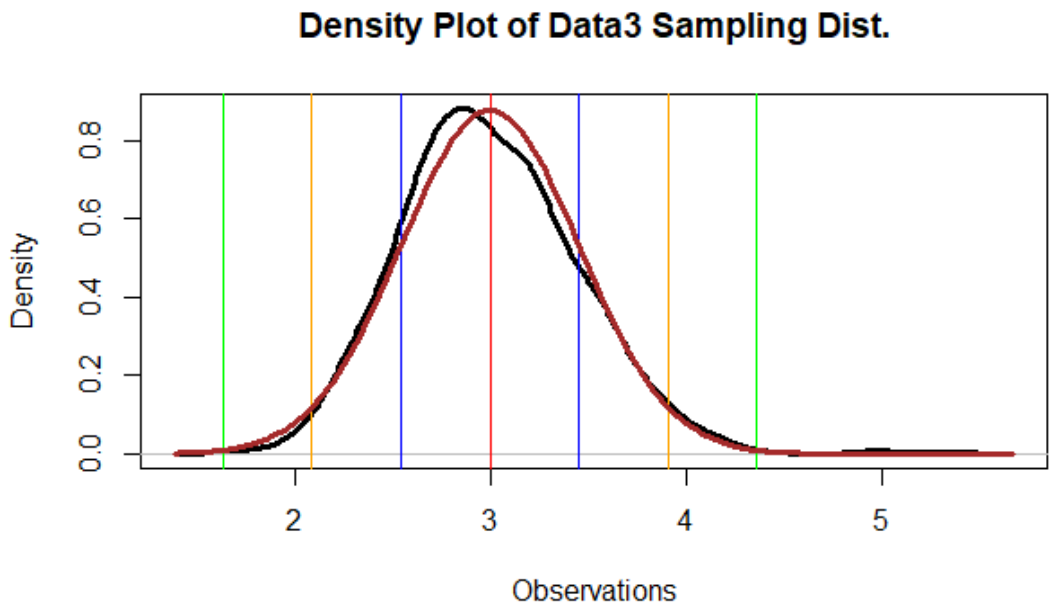
```
boxplot(df3.means,xlab="Data3 Sampling Dist. Boxplot, n = 50")
```




Data3 Sampling Dist. Boxplot, n = 50

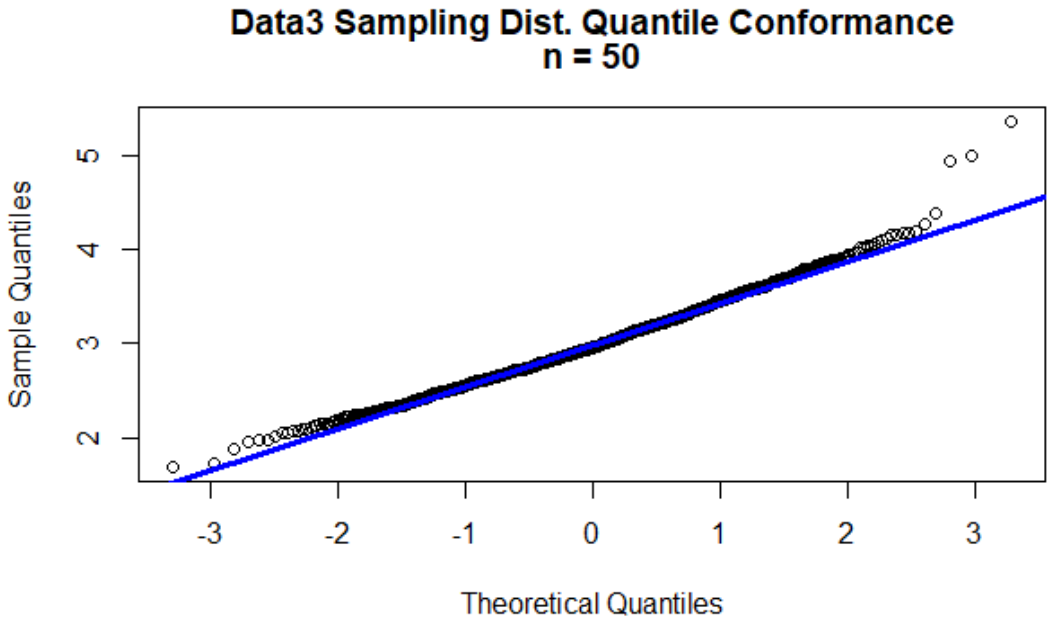
- The sampling distribution boxplot shows good symmetry.
- This indicates the data are Normally distributed.

```
plot(density(df3.means),lwd=3,main="Density Plot of Data3 Sampling
Dist.",xlab="Observations")
abline(v=(mean(df3.means)-(sd(df3.means)*3)),col="green")
abline(v=(mean(df3.means)-(sd(df3.means)*2)),col="orange")
abline(v=(mean(df3.means)-(sd(df3.means))),col="blue")
abline(v=mean(df3.means),col="red")
abline(v=(mean(df3.means)+(sd(df3.means))),col="blue")
abline(v=(mean(df3.means)+(sd(df3.means)*2)),col="orange")
abline(v=(mean(df3.means)+(sd(df3.means)*3)),col="green")
curve(dnorm(x,mean(df3.means),sd(df3.means)),lwd=3,col="brown",add=T)
```




- The sampling distribution density plot shows a mostly normal curve (black).
- Using the mean and SD of the sampling distribution, a Normal curve is plotted in brown.
- Clearly, the sampling distribution is Normally distributed.

```
qqnorm(df3.means,main=c("Data3 Sampling Dist. Quantile Conformance","n =
50"))
qqline(df3.means,lwd=3,col="blue")
```




- The sampling distribution qqnorm plot is an upward-sloping diagonal line.
- Clearly, the sampling distribution is Normally distributed.

```
mom=mean(df3.means)
sdom=sd(df3.means)
df3.quants=data.frame(c(quantile(df3.means,probs=c(pnorm(c(-3,-2,-
1,0,1,2,3),mean=0,sd=1)))),c(mom-sdom*3,mom-sdom*2,mom-
sdom,mom,mom+sdom,mom+sdom*2,mom+sdom*3))
colnames(df3.quants)=c("Observed Values","Predicted Values")
rownames(df3.quants)=c("-3 sigma","-2 sigma","-1 sigma","mean","1 sigma","2
sigma","3 sigma")
grid.newpage()
grid.table(df3.quants)
```

| | Observed Values | Predicted Values |
|---|---|---|
| *-3 sigma* | 1.78240226166271 | 1.63089025033536 |
| *-2 sigma* | 2.18244585114789 | 2.08569863683686 |
| *-1 sigma* | 2.55512175603774 | 2.54050702333836 |
| *mean* | 2.95905977678783 | 2.99531540983986 |
| *1 sigma* | 3.45750898686719 | 3.45012379634136 |
| *2 sigma* | 3.93642802028154 | 3.90493218284287 |
| *3 sigma* | 4.98840281858328 | 4.35974056934437 |

- We can also see the qqnorm result numerically. When we compare the observed values at each sigma with the theoretical values at each sigma, we see the observed values are close to the predicted values with only slight divergence in the tails.