TUTORIAL 2
STA437 WINTER 2015

AL NOSEDAL

Contents

1. Assessing the assumption of Normality

1.1. Evaluating the Normality of the Univariate Marginal Distributions.

Plots are always useful devices in any data analysis. Special plots called **Q-Q plots** can be used to assess the assumption of Normality. These plots can be made for the marginal distributions of the sample observations on each variable. They are in effect, plots of the sample quantile versus the quantile one would expect to observe if the observations actually were Normally distributed. When the points lie very nearly along a straight line, the Normality assumption remains tenable. Normality is suspect if the points deviate from a straight line. Once the reasons for the non-Normality are identified, corrective action is often possible.

To simplify notation, let $x_1, x_2, ..., x_n$ represent n observations on any single characteristic X_i. Let $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ represent these observations after they are ordered according to magnitude. For example, $x_{(n)}$ is the largest observation. The $x_{(j)}$'s are the sample quantiles. The proportion j/n of the sample at or to the left of $x_{(j)}$ is often approximated by $(j-1/2)/n$ for analytical convenience. For a standard Normal distribution, the quantiles $q_{(j)}$ are defined by the relation

$$P[Z \leq q_{(j)}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{z^2/2} dz = p_{(j)} = \frac{j - 1/2}{n}$$

Here $p_{(j)}$ is the probability of getting a value less than or equal to $q_{(j)}$ in a single drawing from a Standard Normal population. The idea is to look at the pairs of quantiles $(q_{(j)}, x_{(j)})$ with the same associated cumulative probability $(j-1/2)/n$.

If the data arise from a Normal population, the pairs $(q_{(j)}, x_{(j)})$ will be approximately linearly related, since $\sigma q_{(j)} + \mu$ is nearly the expected sample quantile.

Example 1. A sample of $n = 10$ observations gives the values in the following table:

Ordered Observations	*Probability levels*	*Standard Normal*
$x_{(j)}$	$(j-1/2)/n$	*Quantiles* $q_{(j)}$
-1	0.05	-1.645
-0.10	0.15	-1.036
0.16	0.25	-0.674
0.41	0.35	-0.385
0.62	0.45	-0.125
0.80	0.55	0.125
1.26	0.65	0.385
1.54	0.75	0.674
1.71	0.85	1.036
2.30	0.95	1.645

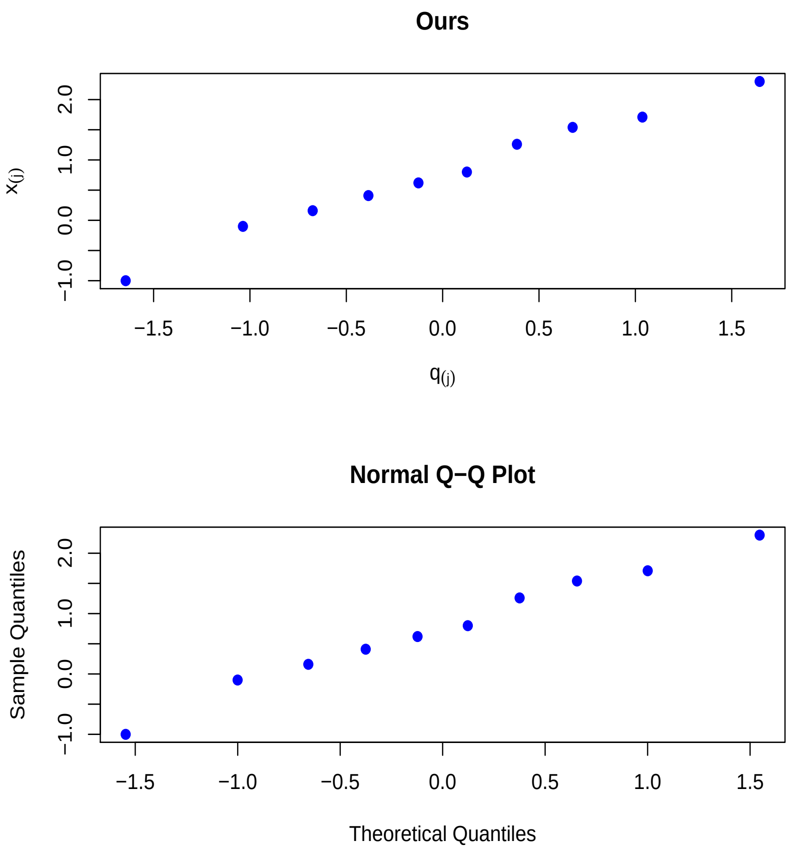
Here, for example, $P[Z \leq 0.385] = \int_{-\infty}^{0.385} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.65$.

Let us now construct the Q-Q plot and comment on its appearance. The Q-Q plot for the foregoing data, which is a plot of the ordered data $x_{(j)}$ against the normal quantiles is shown in Figure 1. The pairs of points $(q_{(j)}, x_{(j)})$ lie very nearly along a straight line, and we would not reject the notion that these data are Normally distributed-particularly with a sample size as small as $n = 10$.

R code

```
## Example
## Q-Q plot (step-by-step)

## Ordered observations

obs<-c(-1,-0.1,0.16,0.41,0.62,0.80,1.26,1.54,1.71,2.30)

n<-length(obs)

## Corresponding probability values

prob.levels<-(seq(1:n)-0.5)/n

## Standard Normal Quantiles
```

```
norm.quantiles<-qnorm(prob.levels)

## Q-Q plot

par(mfrow=c(2,1))

plot(norm.quantiles,obs,xlab=expression(q[(j)]),ylab=expression(x[(j)]),
main="Ours",col="blue",pch=19)

## Q-Q plot (using R function)

qqnorm(obs,col="blue",pch=19)
```


Ours

Normal Q−Q Plot

FIGURE 1. Q-Q plots for the data in Example 1.

Example 2 (A Q-Q plot for radiation data). The quality-control department of a manufacturer of microwave ovens is required by federal government to monitor the amount of radiation emitted when the doors of the ovens are closed. Observations of the radiation emitted through closed doors of $n = 42$ randomly selected ovens were made. The data set is available on Portal (T4-1.DAT). In order to determine the probability of exceeding a prespecified tolerance level, a probability distribution for the radiation emitted was needed. Can we regard the observations here as being Normally distributed? Use R to construct the Q-Q plot.

R code

```
radiation<-read.table(file="T4-1.DAT")

rad.vec<-matrix(unlist(radiation),ncol=1)

qqnorm(rad.vec,col="blue",pch=19)
```

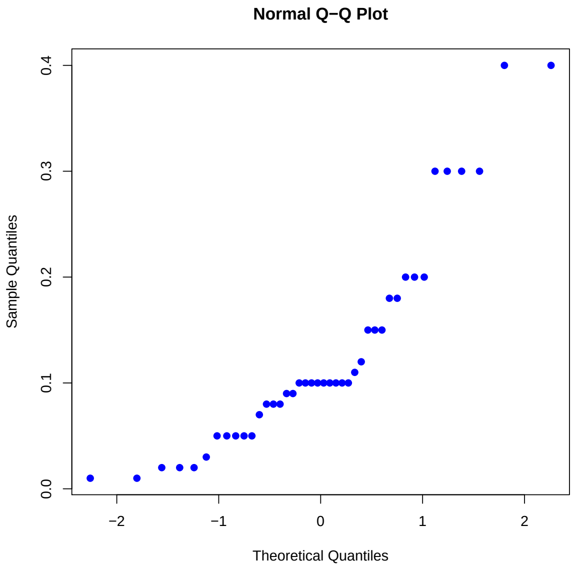


Figure 2. A Q-Q plot of the radiation data (door closed).

It appears from the plot that the data as a whole are not normally distributed.

The straightness of the Q-Q plot can be measured by calculating the correlation coefficient of the points in the plot. The correlation coefficient for the Q-Q plot is defined by

$$r_Q = \frac{\sum_{j=1}^{n}(x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n}(x_{(j)} - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(q_{(j)} - \bar{q})^2}}$$

and a powerful test of normality can be based on it. Formally, we reject the hypothesis of Normality at level of significance α if r_Q falls *below* the appropriate value in the following table.

Sample	*Significance*	*Significance*	*Significance*
size n	*level* 0.01	*level* 0.05	*level* 0.10
5	0.8299	0.8788	0.9032
10	0.8801	0.9198	0.9351
15	0.9126	0.9389	0.9503
20	0.9269	0.9508	0.9604
25	0.9410	0.9591	0.9665
30	0.9479	0.9652	0.9715
35	0.9538	0.9682	0.9740
40	0.9599	0.9726	0.9771
45	0.9632	0.9749	0.9792
50	0.9671	0.9768	0.9809
55	0.9695	0.9787	0.9822
60	0.9720	0.9801	0.9836
75	0.9771	0.9838	0.9866
100	0.9822	0.9873	0.9895
150	0.9879	0.9913	0.9928
200	0.9905	0.9931	0.9942
300	0.9935	0.9953	0.9960

Example 3 (A correlation coefficient test for Normality). Let us calculate the correlation coefficient r_Q from the Q-Q plot of Example 1 and test for Normality.

$$r_Q = \frac{8.584}{\sqrt{8.472}\sqrt{8.795}} = 0.994$$

A test of Normality at the 10% level of significance is provided by referring $r_Q = 0.994$ to the entry in our Table corresponding to $n = 10$ and $\alpha = 0.10$. This entry is 0.9351. Since $r_Q > 0.9351$, we do not reject the hypothesis of Normality.

R code

```
## A correlation coefficient test for Normality

r.q<-cor(norm.quantiles,obs)
```

1.2. **Evaluating Bivariate Normality.** We would like to check on the assumption of Normality for all distributions of $2, 3, ..., p$ dimensions. However, for practical work it is usually sufficient to investigate the univariate and bivariate distributions. We considered univariate marginal distributions earlier. It is now of interest to examine the bivariate case.

Result. Let $\mathbf{x}$ be distributed as $N_p(\mu, \mathbf{\Sigma})$ with $|\mathbf{\Sigma}| > 0$. Then

a. $(\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)$ is distributed as χ_p^2, where χ_p^2 denotes the chi-square distribution with p degrees of freedom.

b. The $N_p(\mu, \mathbf{\Sigma})$ distribution assigns probability $1 - \alpha$ to the solid ellipsoid $\{\mathbf{x} : (\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu) \leq \chi_p^2(\alpha)\}$, where $\chi_p^2(\alpha)$ denotes the upper (100α)th percentile of the χ_p^2 distribution. By this result, the set of bivariate outcomes $\mathbf{x}$ such that

$$(\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu) \leq \chi_2^2(0.5)$$

has probability 0.5. Thus, we should expect roughly the same percentage, 50%, of sample observations to lie in the ellipse given by

$$\{\text{all } \mathbf{x} \text{ such that } (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq \chi_2^2(0.5)\}$$

where we have replaced μ by its estimate $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}^{-1}$ by its estimate $\mathbf{S}^{-1}$. If not, the Normality assumption is suspect.

Example 4 (Checking bivariate Normality. Although not a random sample, data consisting of observations (x_1 = sales, x_2 = profits) for the 10 largest U.S. industrial corporations are available on Portal (us-corp.txt). These data give

$$\bar{\mathbf{x}} = \begin{pmatrix} 62309.1 \\ 2927.3 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 1000509114 & 25575600 \\ 25575600 & 1430020 \end{pmatrix}$$

so

$$\mathbf{S}^{-1} = \begin{pmatrix} 1.841298e-09 & -3.293122e-08 \\ -3.293122e-08 & 1.288259e-06 \end{pmatrix}$$

From R, $\chi_2^2(0.5) = 1.386294 \approx 1.39$. Thus, any observation $\mathbf{x}' = [x_1, x_2]$ satisfying

$$\begin{pmatrix} x_1 - 62309.1 \\ x_2 - 2927.3 \end{pmatrix}' \begin{pmatrix} 1.841298e-09 & -3.293122e-08 \\ -3.293122e-08 & 1.288259e-06 \end{pmatrix} \begin{pmatrix} x_1 - 62309.1 \\ x_2 - 2927.3 \end{pmatrix} \leq 1.39$$

is on or inside the estimated 50% contour. Otherwise the observation is outside this contour. The first pair of observations is $[x_1, x_2]' = [126974, 4224]$. In this case

$$\begin{pmatrix} 126974 - 62309.1 \\ 4224 - 2927.3 \end{pmatrix}' \begin{pmatrix} 1.841298e-09 & -3.293122e-08 \\ -3.293122e-08 & 1.288259e-06 \end{pmatrix} \begin{pmatrix} 126974 - 62309.1 \\ 4224 - 2927.3 \end{pmatrix}$$

$$= 4.342967 > 1.39$$

and this point falls outside the 50% contour. The remaining nine points have generalized distances from $\bar{\mathbf{x}}$ of $1.20, 0.59, 0.83, 1.88, 1.01, 1.02, 5.33, 0.81$, and 0.97, respectively. Since seven of these distances are less than 1.39, a proportion, 0.70, of the data falls within the 50% contour. If the observations were Normally distributed, we would expect about half, or 5, of them to be within this contour. This large a difference in proportions would ordinarily provide evidence for rejecting the notion of bivariate Normality; however, our sample size of 10 is too small to reach this conclusion.

Computing the fraction of the points within a contour and subjectively comparing it with the theoretical probability is a useful, but rather rough, procedure. A somewhat more formal method for judging the joint Normality of a data set is based on the squared generalized distances

$$d_j^2 = (\mathbf{x_j} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x_j} - \bar{\mathbf{x}}), \qquad j = 1, 2, ..., n$$

where $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ are the sample observations. The procedure we are about to describe **is not** limited to the bivariate case: it can be used for all $p \geq 2$. When the parent population is Multivariate Normal and both n and $n - p$ are greater than 25 or 30, each of the squared distances $d_1^2, d_2^2, ..., d_n^2$ should behave like a chi-square random variable. Although these distances are *not* independent or exactly chi-square distributed, it is helpful to plot them as if they were. The resulting plot is called a chi-squared plot or gamma plot, because the chi-square distribution is a special case of the more general gamma distribution. To construct the chi-square plot,

1. Order the squared distances from smallest to largest as $d_{(1)}^2 \leq d_{(2)}^2 \leq ... \leq d_{(n)}^2$.

2. Graph the pairs $\left(q_{c,p}((j - 1/2)/n), d_{(j)}^2\right)$, where $q_{c,p}((j - 1/2)/n)$ is the $100(j - 1/2)/n$ quantile of the chi-square distribution with p degrees of freedom.

The plot should resemble a straight line through the origin having slope 1. A systematic curved pattern suggests lack of Normality. One or two points far above the line indicate large distances, or outlying observations, that merit further attention.

Example 5 (Constructing a chi-square plot).

Let us construct a chi-square plot of the generalized distances Example 4. The ordered distances and the corresponding chi-square percentiles for $p = 2$ and $n = 10$ are listed in the following table:

j	$d^2_{(j)}$	$q_{c,p}((j-1/2)/n)$
1	0.59	0.10
2	0.81	0.33
3	0.83	0.58
4	0.97	0.86
5	1.01	1.20
6	1.02	1.60
7	1.20	2.10
8	1.88	2.77
9	4.34	3.79
10	5.33	5.99

A graph of the pairs $\left(q_{c,2}((j-1/2)/10), d^2_{(j)}\right)$ is shown in Figure 3. The points in Figure 3 do not lie along the line with slope 1. The smallest distances appear to be too large and the middle distances appear to be too small, relative to the distances expected from bivariate Normal populations for samples of size 10. These data do not appear to be bivariate Normal; however, the sample size is small, and it is difficult to reach a definitive conclusion. If further analysis of the data were required, it might be reasonable to transform them to observations more nearly bivariate Normal. Appropriate transformations are discussed in Section 2.

R code

```
## Example. Checking Bivariate Normality

corporations<-read.table(file="us-corp.txt",header=TRUE)

corporations

data<-corporations[ ,-1]

S<-cov(data)

x.bar<-apply(data,2,FUN=mean)

## Turning data and x.bar into column vectors

data<-matrix(unlist(data),nrow=10,ncol=2)
```

```
x.bar<-matrix(x.bar,ncol=1)

new.data<-t(data)

## First squared generalized distance

t(new.data[ ,1]-x.bar)%*%solve(S)%*%(new.data[ ,1]-x.bar)
# we should get 4.34

## Constructing a chi-square plot

# Initializing vector of squared generalized distances

d2<-numeric(10)

for (i in 1:10){

d2[i]<-t(new.data[ ,i]-x.bar)%*%solve(S)%*%(new.data[ ,i]-x.bar)

}

## Ordering squared distances from smallest to largest.

d2.j<-sort(d2)

## Quantiles from chi-square distribution with 2 df

n<-length(d2.j)

prob.levels<-(seq(1:n)-0.5)/n

chi.quantiles<-qchisq(prob.levels,2)

## Chi-square plot

## mar
## A numerical vector of the form c(bottom, left, top, right)
## which gives the number of lines of margin to be specified on the four
## sides of the plot.

par(mar=c(5,5,5,5))
```

```
plot(chi.quantiles,d2.j,xlab=expression(q[2]),
ylab=expression(d[(j)]^2),pch=19,col="blue")

title("Chi-square plot")

## Distances that fall inside the 50% contour.

q.50<-qchisq(0.5,2)

d2.j[d2.j<q.50]
```

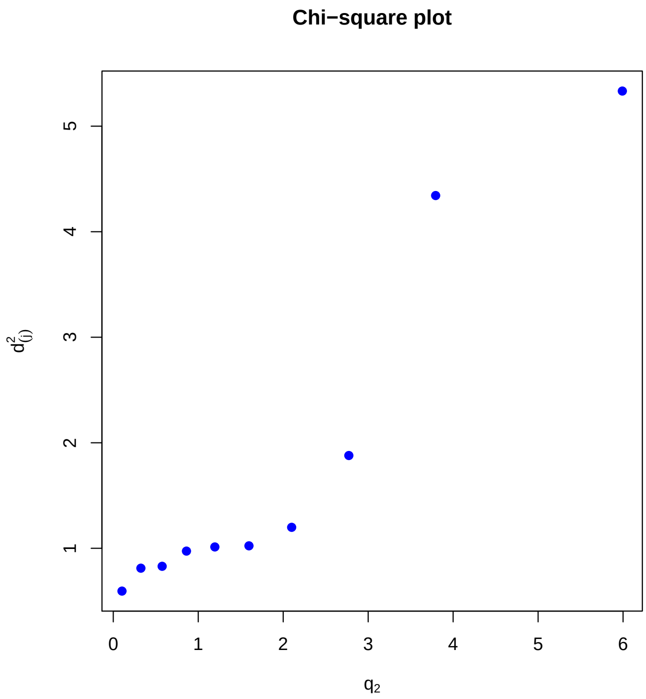


FIGURE 3. A chi-square plot of the ordered distances in Example 4.

2. TRANSFORMATIONS TO NEAR NORMALITY

If Normality is not a viable assumption, what is the next step? One alternative is to ignore the findings of a Normality check and proceed as if the data were Normally distributed. This practice is not recommended, since, in many instances, it could lead to incorrect conclusions. A second alternative is to make non-Normal data more "Normal looking" by considering **transformations** of the data. Normal-theory analyses can then be carried out with the suitably transformed data. Transformations are nothing more than a reexpression of the data in different units. For example, when a histogram of positive observations exhibits a long right-hand tail, transforming the observations by taking their logarithms or square roots will often markedly improve the symmetry about the mean and the approximation to a Normal distribution.

In many instances, the choice of a transformation to improve the approximation to normality is not obvious. For such cases, it is convenient to let the data suggest a transformation. A useful family of transformations for this purpose is the family of **power transformations**. Power transformations are defined only for positive variables. However, this not as restrictive as it seems, because a single constant can be added to each observation in the data if some of the values are negative.

Let x represent an arbitrary observation. The power family of transformations is indexed by a parameter λ. A given value for λ implies a particular transformation. For example, consider x^λ with $\lambda = -1$. Since $x^{-1} = 1/x$, this choice of λ corresponds to the reciprocal transformation. For $\lambda = 0$, we define $x^0 = ln(x)$.

A convenient analytical method is available for choosing a power transformation. Box and Cox consider the slightly modified family of power transformations

$$x^\lambda = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ ln(x) & \lambda = 0 \end{cases}$$

which is continuous in λ for $x > 0$. Given the observations $x_1, x_2, ..., x_n$, the Box-Cox solution for the choice of an appropriate power λ is the solution that *maximizes* the expression

$$L(\lambda) = -\frac{n}{2} ln \left[\frac{1}{n} \sum_{j=1}^{n} (x_j^{(\lambda)} - \bar{x_j^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{j=1}^{n} ln(x_j) \tag{1}$$

where $\bar{x_j^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)}$.

The calculation of $L(\lambda)$ for many values of λ is an easy task for a computer. It is helpful to have a graph of $L(\lambda)$ versus λ, as well as a tabular display of the pairs $(\lambda, L(\lambda))$, in order to study the behaviour near the maximizing value $\hat{\lambda}$. For instance, if either $\lambda = 0$ (logarithm) or $\lambda = 1/2$ (square root) is near $\hat{\lambda}$, one of these may be preferred because of its simplicity.

Comment. It is now understood that the transformation obtained by maximizing $L(\lambda)$ usually improves the approximation to Normality. However, there is no guarantee that even the best choice of λ will produce a transformed set of values that adequately conform to a Normal distribution. The outcomes produced by a transformation selected according to equation (1) should always be carefully examined for possible violations of the tentative assumption of Normality.

Example 6 (Determining a power transformation for univariate data). We gave readings of the microwave radiation emitted through the closed doors of $n = 42$ ovens in Example 2. The Q-Q plot of these data in Figure 2 indicates that the observations deviate from that would be expected if they were Normally distributed. Since all the observations are positive, let us perform a power transformation of the data which, we hope, will produce results that are more nearly Normal. Restricting our attention to the family of transformations we discussed above, we must find that value of λ maximizing the function $L(\lambda)$ in equation (1).

The curve of $L(\lambda)$ versus λ that allows the more exact determination $\hat{\lambda} = 0.28$ is shown in Figure 4. It is evident from the plot that a value of $\hat{\lambda}$ around 0.30 maximizes $L(\lambda)$. For convenience, we choose $\hat{\lambda} = 0.25$. The data x_j were reexpressed as

$$x_j^{(1/4)} = \frac{x_j^{1/4} - 1}{1/4} \qquad j = 1, 2, ..., 42$$

and a Q-Q plot was constructed from the transformed quantities. This is shown in Figure 5. The quantile pairs fall very close to a straight line, and we would conclude from this evidence that the $x_j^{(1/4)}$ are approximately Normal.

R code

```
## Transformations to near Normality

box.cox<-function(x,lambda){

if (lambda==0) y<-log(x)  else

y<-(x^(lambda)-1)/lambda

return(y)

}
```

```
L.lambda<-function(x,lambda){

new.x<-box.cox(x,lambda)

n<-length(new.x)

L<-(-n/2)*log( (1/n)*(sum(new.x*new.x)) - mean(new.x)^2 ) + (lambda-1)*sum(log(x))

return(L)

}


## Finding values for plot of L(lambda) vs lambda

lambda.vec<-seq(0,0.6,by=0.01)

k<-length(lambda.vec)

L.fun<-numeric(k)

for (i in 1:k){

L.fun[i]<-L.lambda(rad.vec,lambda.vec[i])

}


## Plot of L(lambda) vs lambda

par(mar=c(5,5,5,5))

plot(lambda.vec,L.fun,type="l",col="blue",xlab=expression(lambda),
ylab=expression(L(lambda)))

## Q-Q plot of transformed data

trans.rad.vec<-box.cox(rad.vec,0.25)

qqnorm(trans.rad.vec,col="blue",pch=19)
```

```
qqline(trans.rad.vec)
```

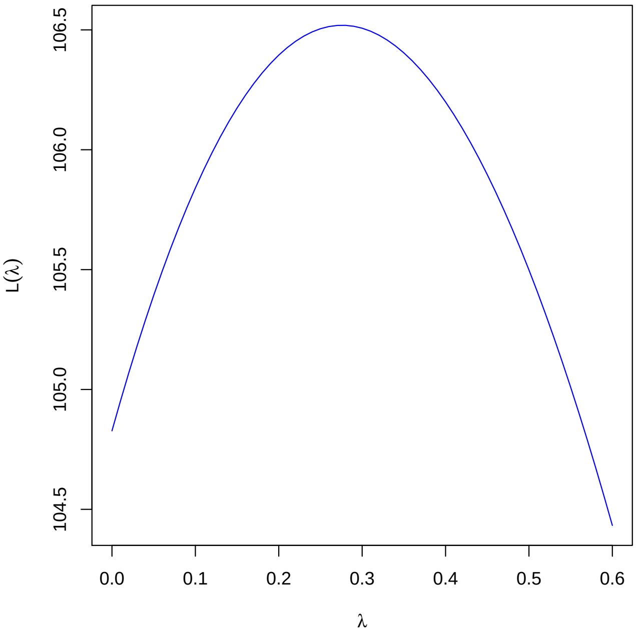


FIGURE 4. Plot of $L(\lambda)$ vs λ for radiation data (door closed).

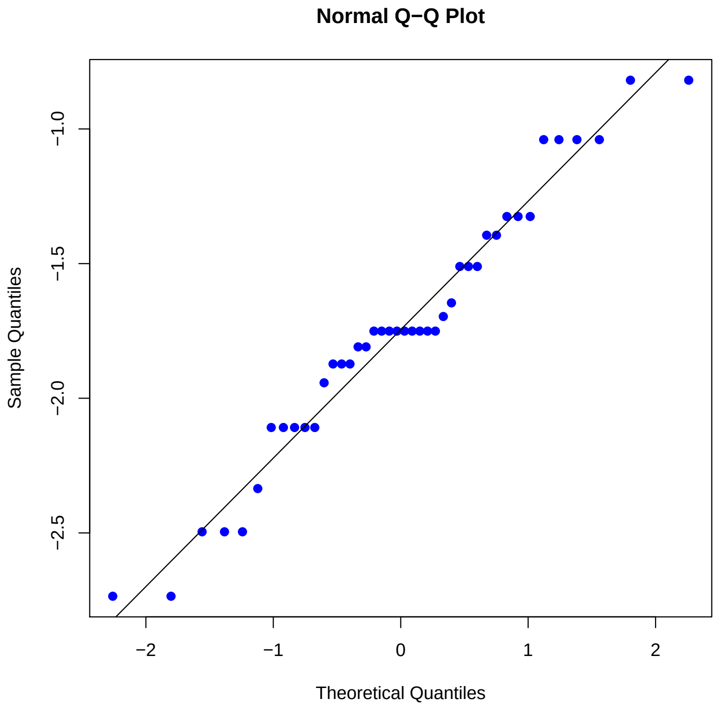


FIGURE 5. A Q-Q plot of the transformed radiation data (door closed).