Report

Xu Wang[1]

April 1, 2015

[1] I thank Yihui Xie and Yuchen Wang for providing this machine.

Assignment

The data in Table 1.5 are 42 measurements on air-pollution variables recorded at 12:00 noon in the Los Angeles area on different days. (See also the air-pollution data on the web at `www.prenhall.com/statistics`.)

1. Consider the air-pollution data given in **Table 1.5** . Construct a Q-Q plot for the solar radiation measurements .
2. Carry out a test for normality based on the correlation coefficient r_Q [see (**4-31**)]. Let $\alpha = .05$ and use the entry corresponding to $n = 40$ in **Table 4.2**.

Table 1.5


```
data <- read.csv('C:\\Users\\WangXu\\Desktop\\rapidminer\\data.csv'
,head=FALSE)
names(data) <- c('Wind', 'Solar_radiation', 'CO', 'NO', 'NO2', 'O3'
, 'HC')
head(data)

##   Wind Solar_radiation CO NO NO2 O3 HC
## 1    8              98  7  2  12  8  2
## 2    7             107  4  3   9  5  3
## 3    7             103  4  3   5  6  3
## 4   10              88  5  2   8 15  4
## 5    6              91  4  2   8 10  3
## 6    8              90  5  2  12 12  4
```

4-31

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

Table 4.2

Table 4.2 Critical Points for the Q–Q Plot Correlation Coefficient Test for Normality

Sample size n	Significance levels α .01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Figure: Table 4.2

Procedure

The steps leading to a Q-Q plot are as follows:

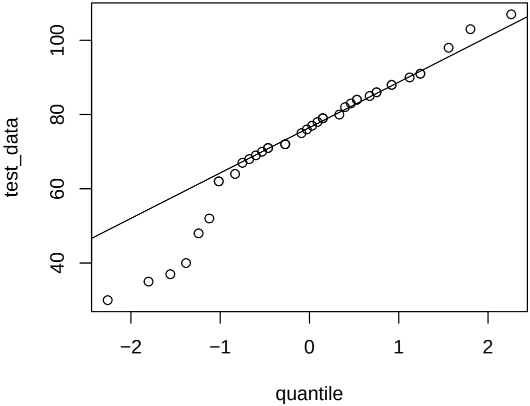
1. Order the original observations to get $x_{(1)}$, $x_{(2)}, \ldots$ $x_{(n)}$ and their corresponding probability values $(1 - \frac{1}{2})/n$, $(2 - \frac{1}{2})/n, \ldots$ $(n - \frac{1}{2})/n$;
2. Calculate the standard normal quantiles $q_{(1)}$, $q_{(2)}, \ldots q_{(n)}$
3. Plot the pairs of observations $(q_{(1)}, x_{(1)})$, $(q_{(2)}, x_{(2)}), \ldots (q_{(n)}, x_{(n)})$, and examine the "straightness" of the outcome.

Code

Text is nice but let's see what happens if we make a couple of plots in our chunk:

```
data <- read.csv('C:\\Users\\WangXu\\Desktop\\rapidminer\\data.csv'
,head=FALSE)
names(data) <- c('Wind', 'Solar_radiation', 'CO', 'NO', 'NO2', 'O3'
, 'HC')
test_data <- data$Solar_radiation
test_data <- sort(test_data)
quantile <- c()
quantile[1]  <- qnorm((1-0.5)/length(test_data))
for (i in 2:length(test_data)) {
if (test_data[i-1]<test_data[i]) {
  quantile[i] <- qnorm((i-0.5)/length(test_data))
} else {
  quantile[i] = quantile[i-1]
}
}
plot(quantile,test_data)
qqline(test_data)
```

Q-Q Plot



Normality Test

H_0: The variable "Solar radiation" is normally distributed **vs**
H_1: The variable "Solar radiation" is not normally distributed

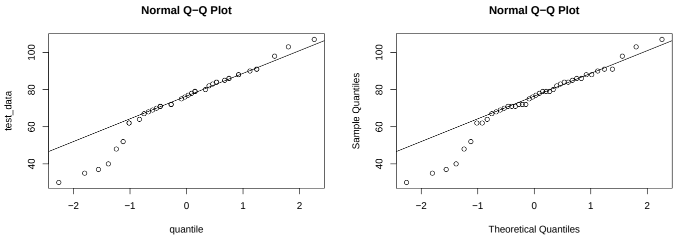
```
a <- qqnorm(test_data)
a$x <- sort(a$x)
a$y <- sort(a$y)
sum((a$x-mean(a$x))*(a$y-mean(a$y))) / ((sum((a$x-mean(a$x))^2))^0.5*(su

## [1] 0.9693258
```

Due to the r_Q is 0.9693258 . Table 4.2 shows when n=40, $\alpha = 0.05$, the r_Q is 0.9726 > the result in this question. So I reject this null hypothesis. We can get the conclusion is **The variable Solar radiation is not normally distributed**

Which is *Q-Q* Plot ?

We can see this two figures.

```
plot(quantile,test_data, main='Normal Q-Q Plot')
qqline(test_data)
qqnorm(test_data)
qqline(test_data)
```



Thanks