MultiVariate Normal Distribution

Jing Qin

01/03/2022

08/03/2022

How to assess the normality?

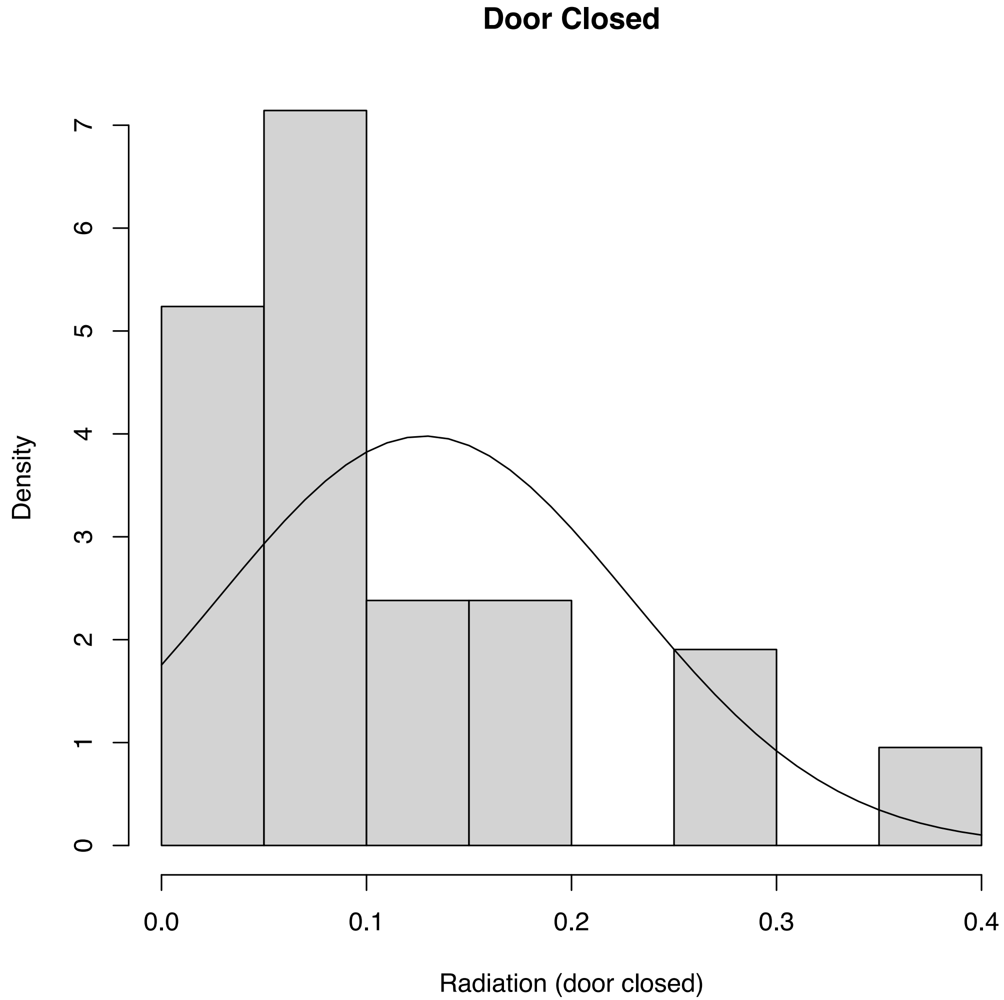
By looking at them?!

Data set: Radiation Data (door closed t4-1.dat)

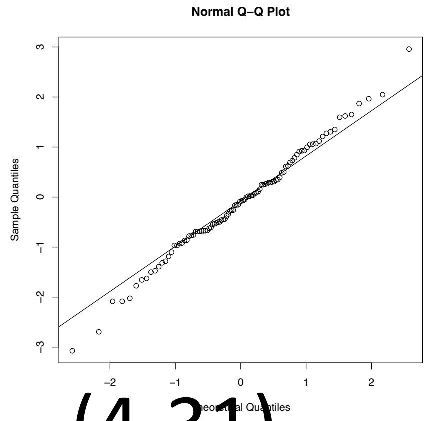
Maybe NOT

Maybe

Quantitative Analysis !!!



And the hypothesis test (of linearity)

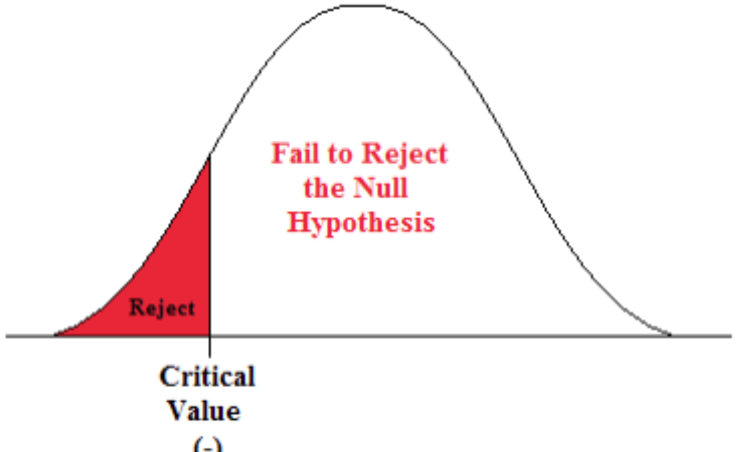


- Null hypothesis (H0): data is normally distributed
- Test statistic: r_Q = cor(x-coordinates, y-coordinates) (4-31)

```
dfc <- read.table("t4-1.dat",header=FALSE)  #dfc for closed door#
pdf("qqplotclosed2.pdf")
    qqc <- qqnorm(dfc$V1)
    qqline(dfc$V1)
dev.off()
corqq <- cor(qqc$x, qqc$y)   #corq (0.9279049) is almost the same
```

- Rejection Region/Criterion:

Test-statistic r_Q < critical value found in Table 4.2 for a significant level, then reject

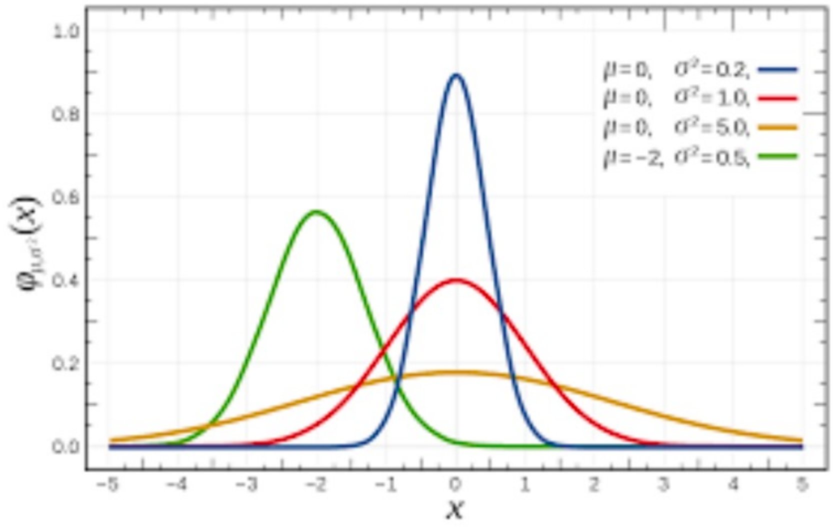


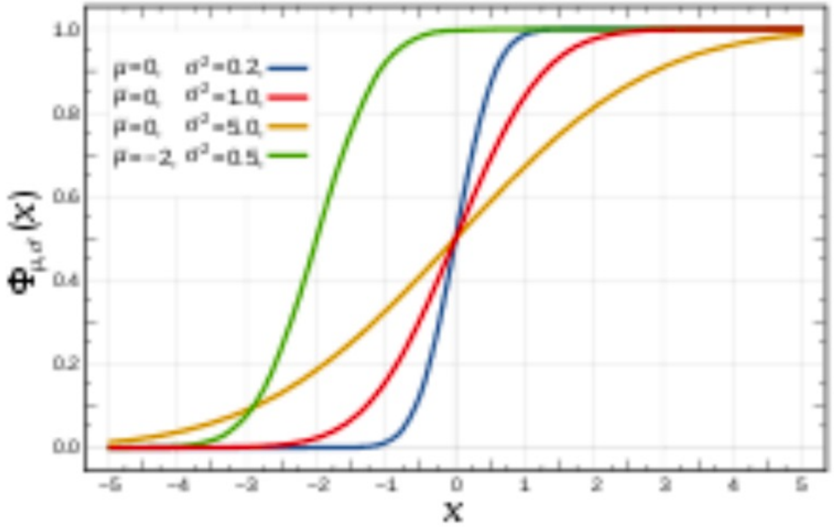
Re-cap: univariate normal distribution

Assume a r.v. X satisfies a normal/Gaussian distribution $N(\mu, \sigma^2)$, i.e.

$$X \sim N(\mu, \sigma^2)$$

- Probability density function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\}$



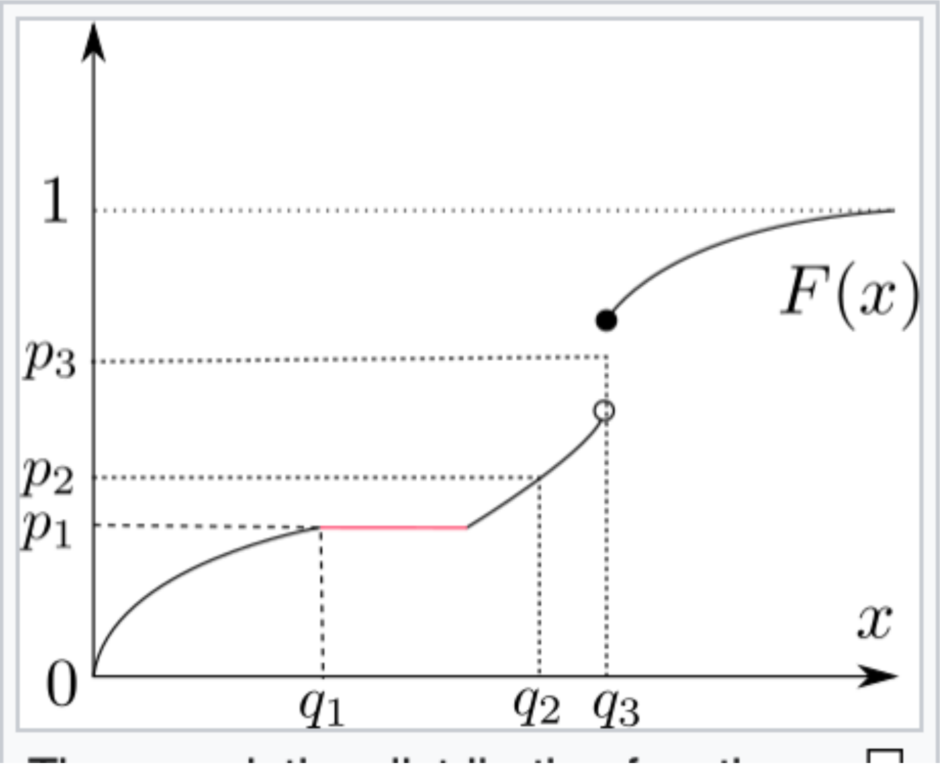


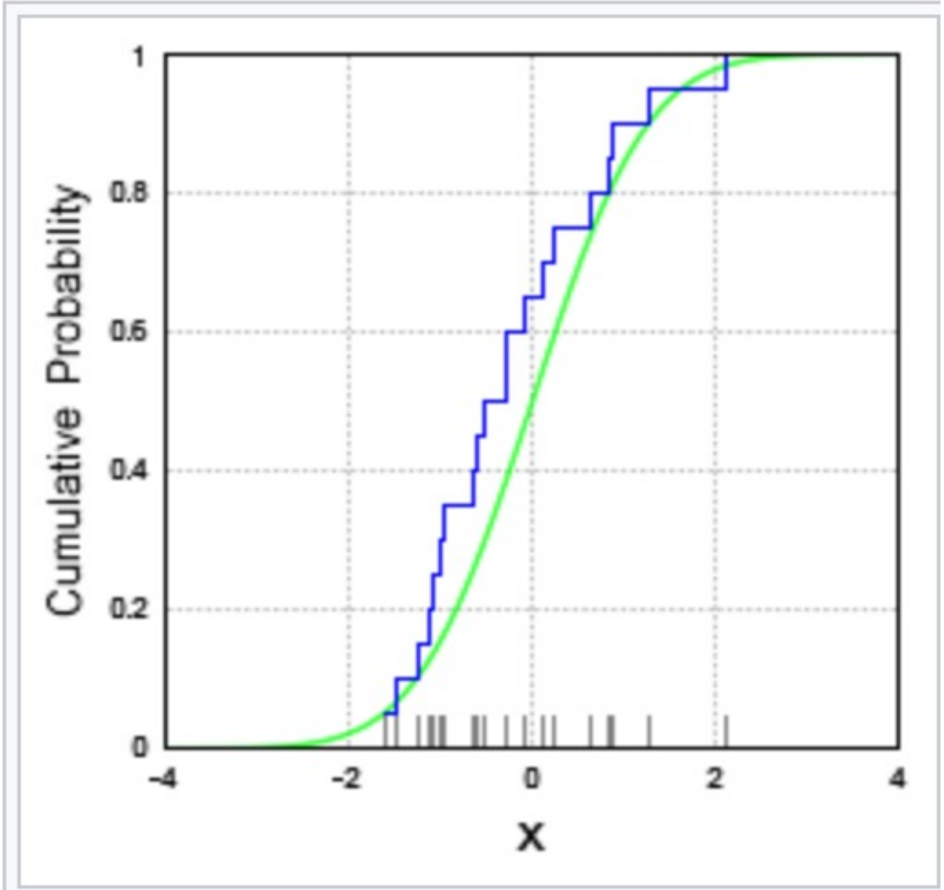
- $E(X) = \mu$ and $Var(X) = \sigma^2$

- Estimation about μ and σ with some given data: confidence interval and hypothesis test.

Quantile function

Theoretical Quantile q qxxxx()   Empirical/sample Quantile quantile()





Empirical CDF $\widehat{F_n}(x) = \dfrac{number\ of\ observations \leq x}{n}$

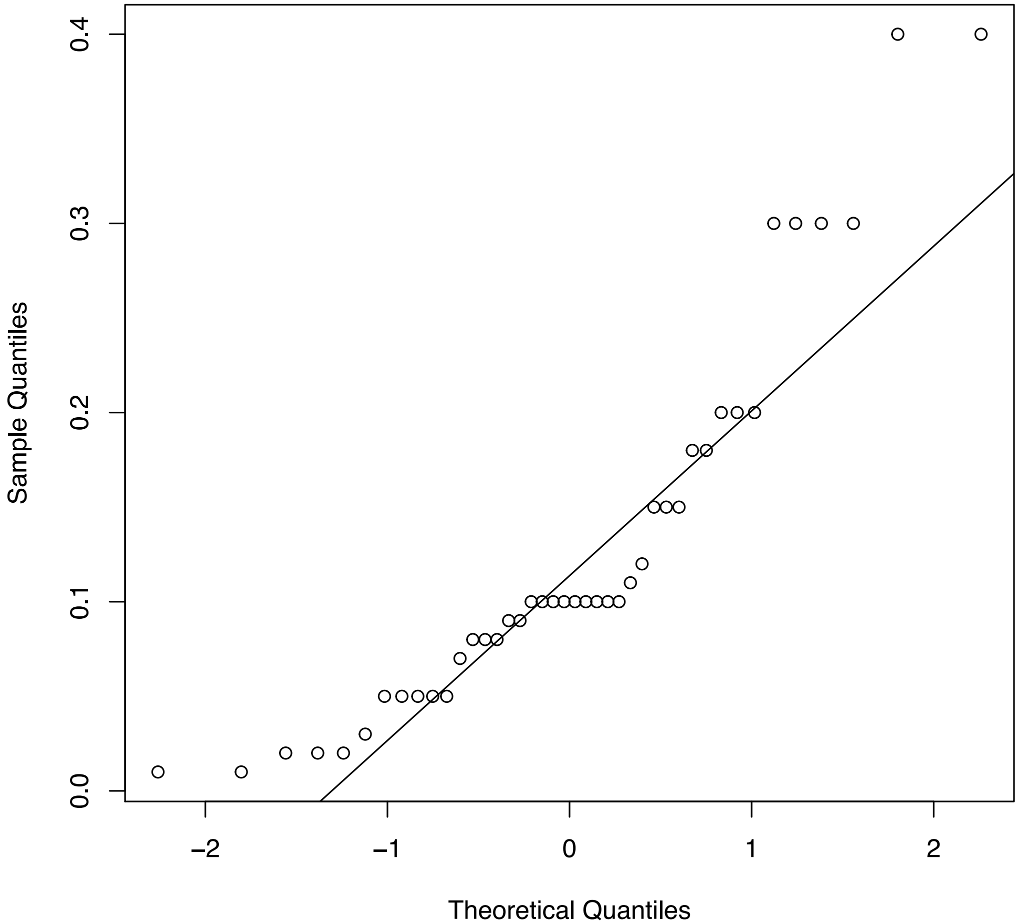
Empirical quantile $\hat{Q}_n\left[\frac{j-0.5}{n}\right] = j$-th largest observation

Quantitive method: Q-Q plot ($n \geq 20$)

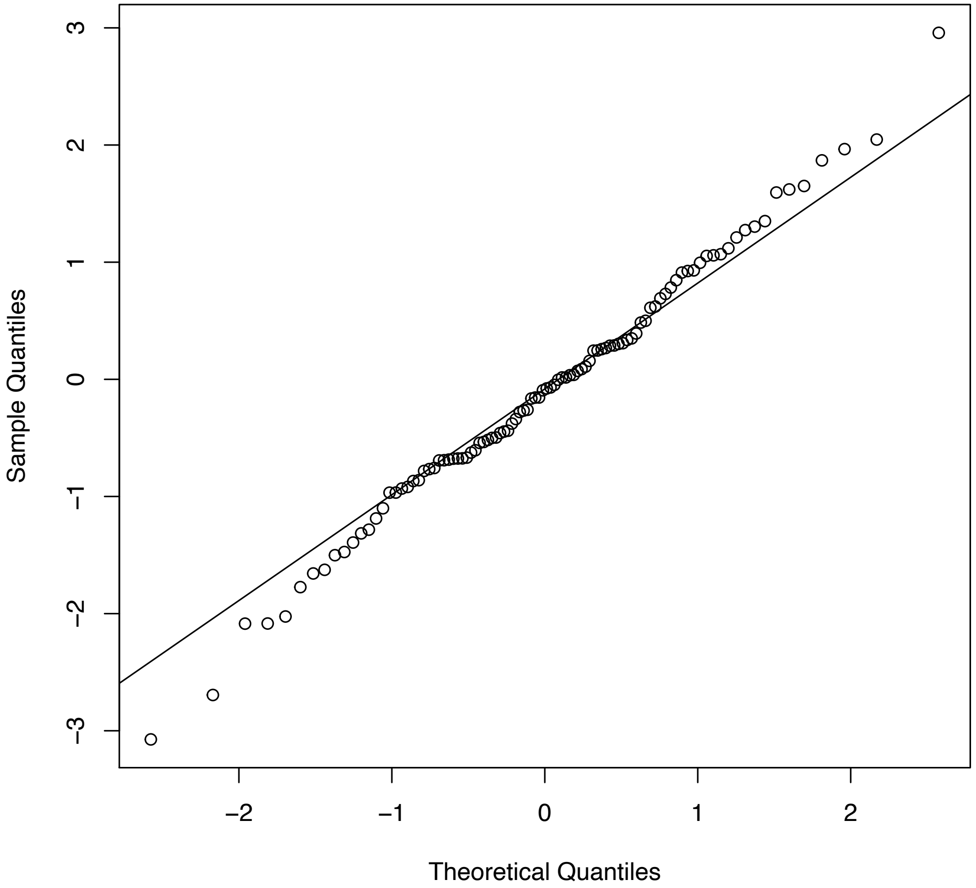
We are looking for a linear pattern here, the more linear, the better

R cmd: qqnorm() and qqline() see April22.R

Door Closed



Simulated data with rnorm



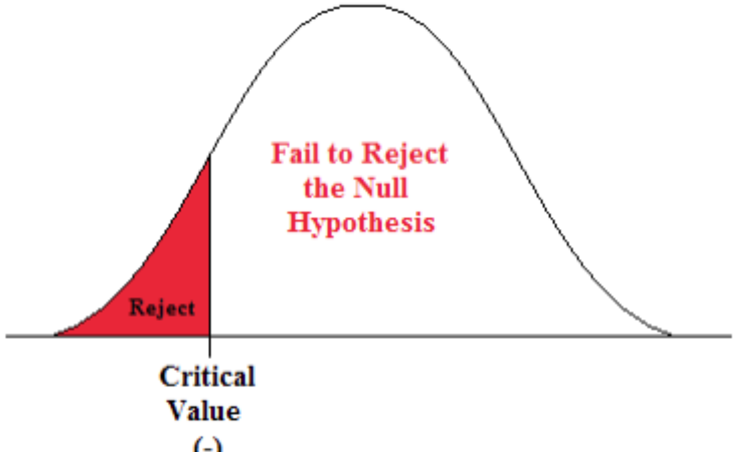
And the hypothesis test (of linearity)

- Null hypothesis (H0): data is normally distributed
- Test statistic: r_Q = cor(x-coordinates, y-coordinates) (4-31)

```
dfc <- read.table("t4-1.dat",header=FALSE)  #dfc for closed door#
pdf("qqplotclosed2.pdf")
    qqc <- qqnorm(dfc$V1)
    qqline(dfc$V1)
dev.off()
corqq <- cor(qqc$x, qqc$y)   #corq (0.9279049) is almost the same
```

- Rejection Region/Criterion:

Test-statistic r_Q < critical value found in Table 4.2 for a significant level, then reject



0.9279

Reject Null Hypothesis

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

Sample size	Significance levels α		
n	.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

This is not the end (yet)....

Data transformation (not manipulation)

- Theoretical considerations (4-33)

  Count data $Y \to$ square root transform $\sqrt{Y}$

  Proportions $\hat{P} \to$ logit transform

$$\text{logit}(\hat{P}) = \log \frac{\hat{P}}{1-\hat{P}}.$$

  Correlations $R \to$ Fisher's Z transform

$$Z(R) = \frac{1}{2} \log \frac{1+R}{1-R}.$$

Data transformation

- (4-34) Data-based transformations – BOX-COX power transformation

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

The parameter λ is determined from the data, in particular, the optimal value $\hat{\lambda}$ is the one which maximizes

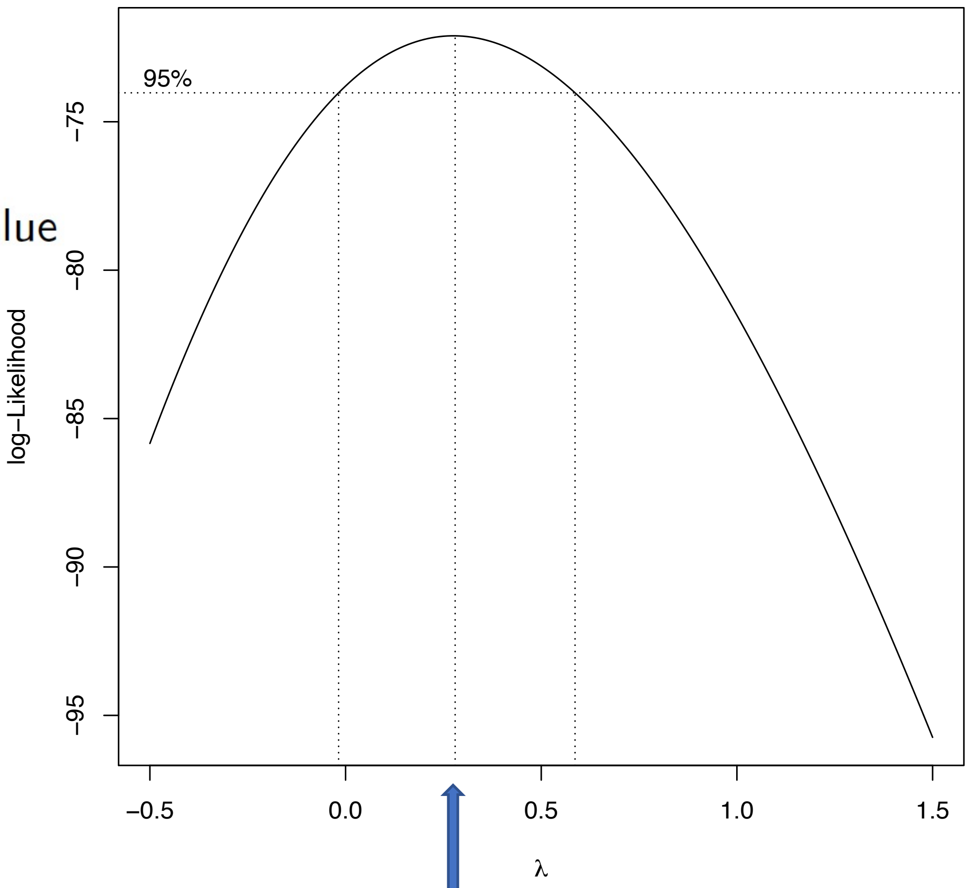
$$(4-35)\quad \ell(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^{n} \left(x_j^{(\lambda)} - \overline{x^{(\lambda)}} \right)^2 \right] + (\lambda - 1) \sum_{j=1}^{n} \ln x_j$$

where

$$\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n} \sum_{j=1}^{n} \frac{x_j^\lambda - 1}{\lambda}.$$

(Well, keep in mind not everything is normally distributed!)

(Fig 4.12)



0.28

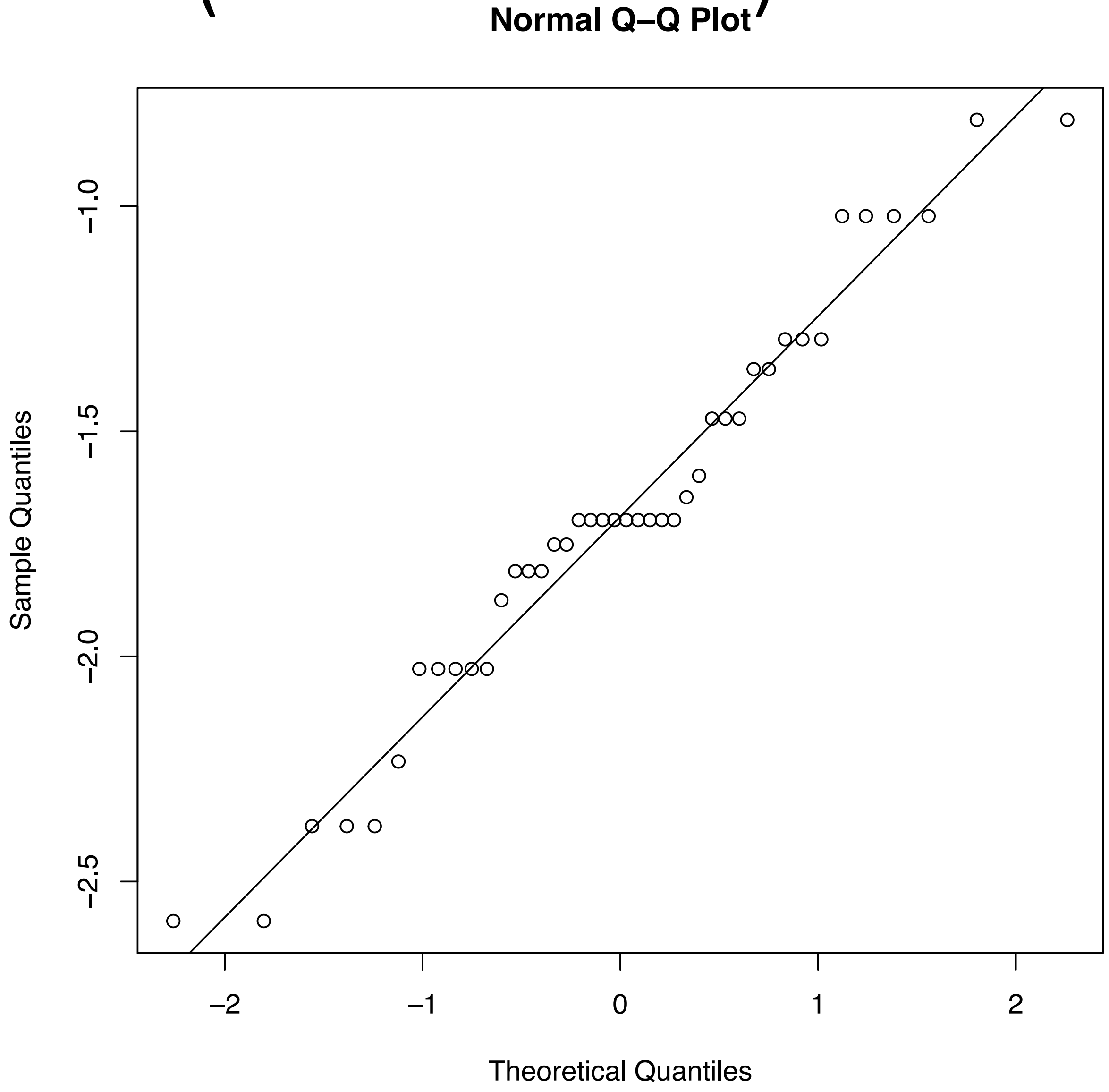
Box-Cox on radiation data (door-closed)

```
vec1 <- dfc$V1

transvec <- (vec1^.28-1)/.28  #

pdf("qqplottransclosed.pdf")
   qqts <- qqnorm(transvec)
   qqline(transvec)
dev.off()
cortrans <- cor(qqts$x, qqts$y)
```

```
> source("April22.R")
> cortrans
[1] 0.9847686
```



Now Table 4.2 again

~~0.9279~~

0.9847

<u>accept Null Hypothesis</u>

What is the hypothesis here?

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

Sample size	Significance levels α		
n	.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

<u>Not</u> all data are normally distributed!