ORIGIN ≡ 1

Assessing Data Normality

Assessing Normality of sample data is an essential part of statistical analysis. Q-Q Plots are one easy way to do this. They are also interesting at this point in our course since they demonstrate the use of the inverse cumulative probability function for the Normal Distribution.

Q-Q Plots:

Reading Anderson's Iris data:

iris := READPRN("c:/DATA/Biostatistics/iris.txt")

$SL := iris^{\langle 2 \rangle}$ **< assigning variable SL**

$n := \text{length}(SL)$ $n = 150$ **< n = number of observations X**

$i := 1 .. n$ **< constructing index variable i**

$Xbar_{SL} := \text{mean}(SL)$ $Xbar_{SL} = 5.8433$ **< mean of X**

$SD_{SL} := \sqrt{\text{Var}(SL)}$ $SD_{SL} = 0.8281$ **< sample standard deviation of X**

$SE_{SL} := \frac{SD_{SL}}{\sqrt{n}}$ $SE_{SL} = 0.0676$ **< standard error of the sample *mean* of X**

Calculating Cumultive Probability levels $\Phi_N(X)$:

We will look at variable SL here:

SL =

	1
1	5.1
2	4.9
3	4.7
4	4.6
5	5
6	5.4
7	4.6
8	5
9	4.4
10	4.9
11	5.4
12	4.8
13	4.8
14	4.3
15	5.8
16	5.7

First we sort SL:

$SL_{sort} := \text{sort}(SL)$

SL_{sort} =

	1
1	4.3
2	4.4
3	4.4
4	4.4
5	4.5
6	4.6
7	4.6
8	4.6
9	4.6
10	4.7
11	4.7
12	4.8
13	4.8
14	4.8
15	4.8
16	4.8

Now we treat each index of SL_{sort} as a quantile, and each observed value as a normal cumulative probability $\Phi_N(X)$:

$$\Phi_i := \frac{\left(i - \frac{1}{2}\right)}{n}$$

^ the 1/2 here is a correction factor

Φ =

	1
1	0.0033
2	0.01
3	0.0167
4	0.0233
5	0.03
6	0.0367
7	0.0433
8	0.05
9	0.0567
10	0.0633
11	0.07
12	0.0767
13	0.0833
14	0.09
15	0.0967
16	0.1033

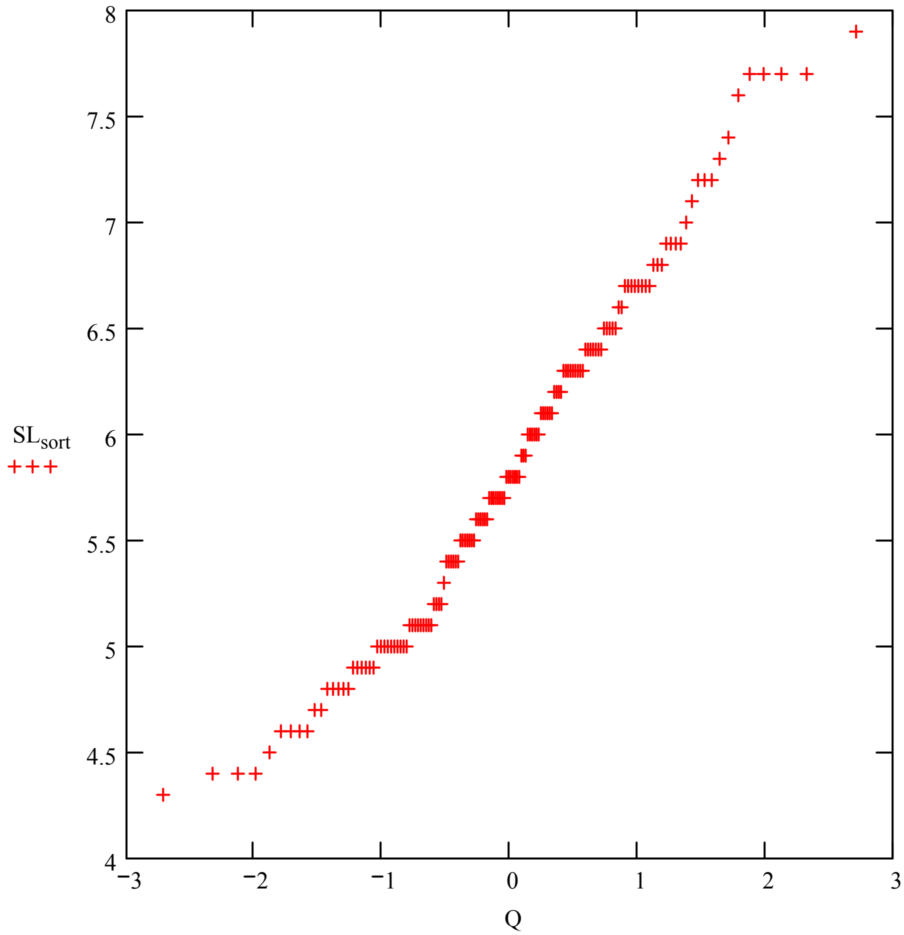
From the values of $\Phi_N(X)$, we now convert back to X

$Q_i := \text{qnorm}(\Phi_i, 0, 1)$

Plotting SL_{sort} vs Q:

Q =

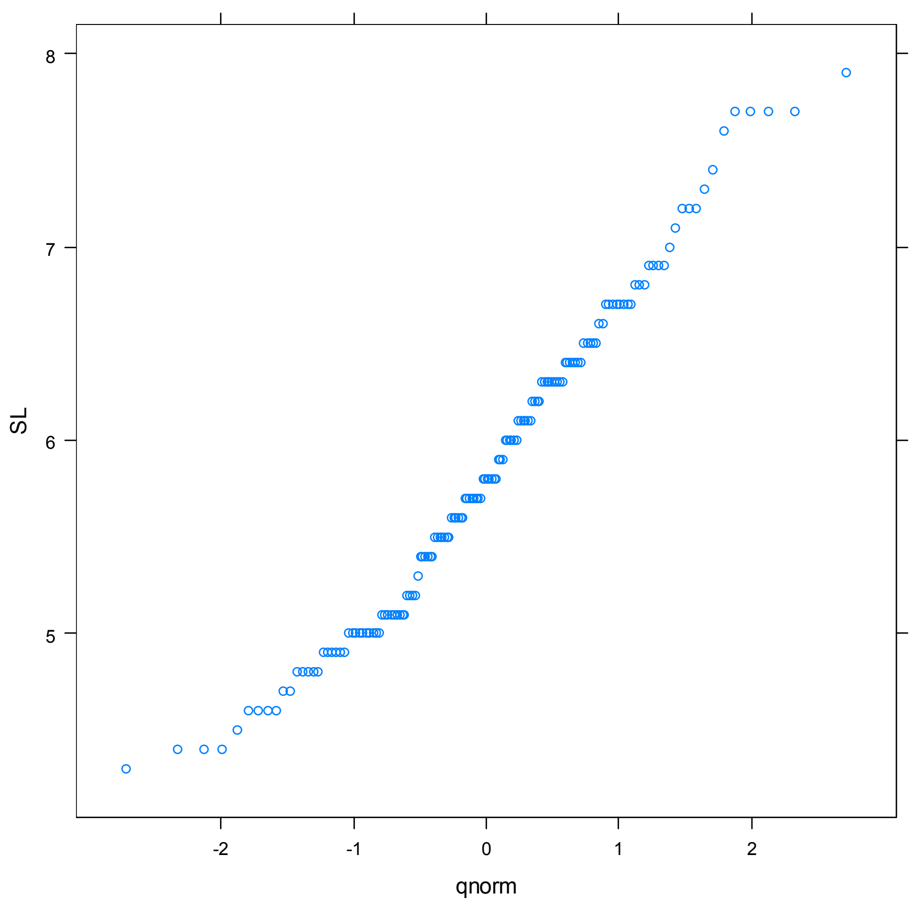
	1
1	-2.7131
2	-2.3263
3	-2.128
4	-1.9893
5	-1.8808
6	-1.7908
7	-1.7132
8	-1.6449
9	-1.5834
10	-1.5274
11	-1.4758
12	-1.4279
13	-1.383
14	-1.3408
15	-1.3008
16	-1.2628



If the sample data are distributed close to the Normal distribution, the Q-Q plot should be mostly a straight line in the center with an overall S-shaped curve towards each end.

Prototype in R:

```
#READ IRIS TABLE AND ASSIGN VARIABLE SL
K=read.table("iris.txt")
attach(K)
SL=Sepal.Length
#LOAD PACKAGE - choose "lattice" from pop-up list
local({pkg <- select.list(sort(.packages(all.available = TRUE)))
if(nchar(pkg)) library(pkg, character.only=TRUE)})
#DOCUMENTATION:
? qqmath()
#QQ PLOT IN lattice:
qqmath(SL)
```



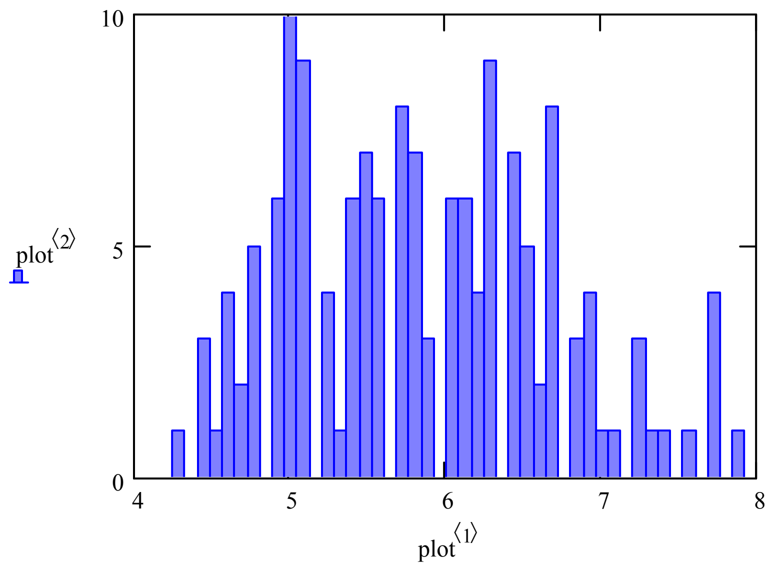
Symmetry and Kurtosis:

Zar 2010 pp. 88-91 provides an introduction into the sordid world of trying to measure departures from a Normal distribution by real data. Measures of the kind proposed are sometimes useful, but not critically important. Example calculations of two measures based on moments are prototyped here.

$X := SL \qquad X_{bar} := \text{mean}(SL)$

$plot := \text{histogram}(50, X)$

$n = 150 \qquad i := 1 .. n$



Symmetry: **Zar Eq. 6.16 p. 88**

$$SQRT_{b1} := \frac{\sqrt{n} \cdot \left[\sum_i \left(X_i - X_{bar}\right)^3\right]}{\sqrt{\left[\sum_i \left(X_i - X_{bar}\right)^2\right]^3}}$$

$SQRT_{b1} = 0.3117531$

^ This value is:
zero for perfectly symmetrical samples;
positive for "right skewed" = long tail to right
negative for "left skewed" = long tail to left

Kurtosis: **Zar Eq. 6.17 p. 89**

$$b_2 := \frac{n \cdot \left[\sum_i \left(X_i - X_{bar}\right)^4\right]}{\left[\sum_i \left(X_i - X_{bar}\right)^2\right]^2}$$

$b_2 = 2.426432$

^ A perfectly normal distribution will have b_2 = 3 termed *mesokurtic*.

b_2 < 3 is *platykurtic* - there is a lower/wider peak at at the mean and thinner tails.

b_2 > 3 is *leptokurtic* - there is a more acute peak at at the mean and fatter tails.

Prototype in R:

```
#SKEWNESS & KURTOSIS IN R:
LOAD PACKAGE - choose "moments" from pop-up list
library(moments)
#SKEWNESS
skewness(SL)
#KURTOSIS
kurtosis(SL)
```