Order Statistics 1.3

Heather Ann Dye

11/8/2022

Order Statistics 1.3

Example 1: Definition of Order Statistics Let X_i be the observed value of the ith trial from a sequence of n trials. Let

- $Y_1 = min\{X_1, X_2, \dots X_n\}$
- $Y_2 =$ second smallest $\{X_1, X_2, \dots X_n\}$
- $Y_3 =$ third smallest $\{X_1, X_2, \dots X_n\}$
- $Y_n = max\{X_1, X_2, \dots X_n\}$

Now,

$$Y_1 \leq Y_2 \leq Y_3 \ldots \leq Y_n.$$

The variables $Y_1 \leq Y_2 \leq Y_3 \leq Y_4 \leq Y_5$ are order statistics. In this example, each independent trial is a random variable X_i with the distribution

$$f(x_i) = \begin{cases} 2x_i & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

Then

$$F(a_i) = \begin{cases} 0 & a_i \leq 0 \\ a_i^2 & 0 \leq a_i < 1 \\ 1 & 1 \leq a_i \end{cases}$$

Now, $P(X_i < 1/2) = 1/4$ so that $P(Y_4 < 1/2) = \binom{5}{4}(1/4)^4(3/4) + (1/4)^5$. Note that if Y_4 is less than 1/2 then Y_1, Y_2, Y_3 are also less than 1/2.
If Y_5 is less than 1/2 then all five order statistics are less than 1/2.

Hence, we can write the following CDF for Y_4.

$$G(a) = P(Y_4 < a) = \begin{cases} 0 & a < 0 \\ \binom{5}{4}(a^2)^4(1-a^2) + (a^2)^5 & 0 \leq a < 1 \\ 1 & 1 \leq a \end{cases}$$

This is the cumulative distribution function and the associated pdf is:

$$G'(a) = g(a) = \begin{cases} 0 & a \leq 0 \\ \binom{5}{4}(4(a^2)^3(2a)(1-a^2) - 2a(a^2)^4) + 5(a^2)^4(2a) & 0 \leq a < 1 \\ 0 & 1 \leq a \end{cases}$$

Example 2: Calculational Formula for the rth order statistic Let X_i be the ith sample of n with the pdf $f(x)$ and the cdf $F(X)$. Then the CDF of Y_r is

$$G_r(a) = Pr(Y_r < a) = \sum_{i=r}^{n} \binom{n}{i} F(a)^i (1 - F(a))^{n-i}.$$

with the appropriate domains. Then, we take the derivative and compute the pdf:

$$g_r(a) = \sum_{i=r}^{n} \binom{n}{i} (iF(a)^{i-1} f(a)(1 - F(a))^{n-i} + (n-i)F(a)^i (1 - F(a))^{n-i-1}(-f(a))).$$

Example 3: Computing an order statistic. Let $X_i \sim N(0, 1)$ be the ith sample of 6. Note that $f(x)$ is described by dnorm and $F(X)$ is described by pnorm in R.

We want to calculate the probability that $P(Y_5 < 1)$.

Note that

$$P(Y_6 < a) = F(a)^6$$

and

$$P(Y_5 < a) = 6F(a)^5(1 - F(a)) + F(a)^6$$

. We can easily compute this in R by writing a function.

```
myorderstat<-function(a){6*pnorm(a)^5 *(1- pnorm(a)) + pnorm(a)^6}
myorderstat(1)
```

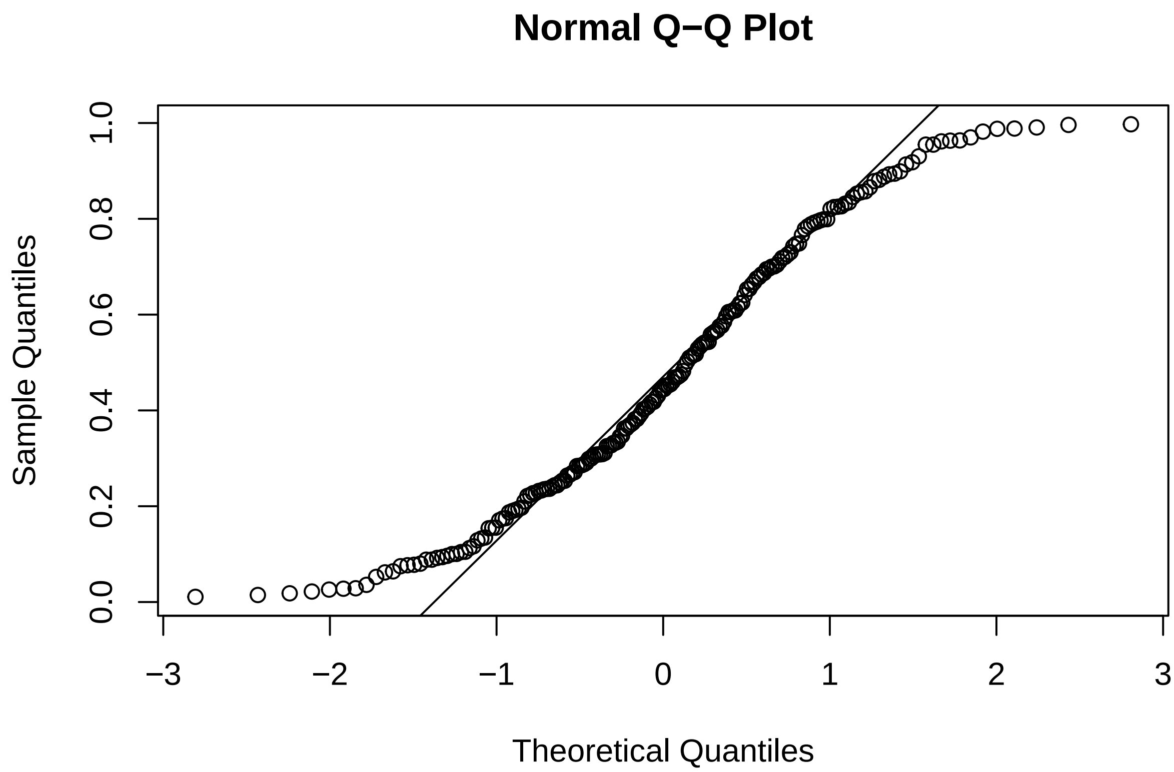
```
## [1] 0.7559919
```

Example 4: Setting up a Q-Q plot A Q-Q plot is a quantile-quantile plot and compares order statistics with a given distribution. Note that Y_r is the sample quantile of order $r/(n+1)$ and π_r is the percentile with a Q-Q plot (π_r, Y_r). In a quantile-quantile plot, the quantile of the statistic Y_r is computed. Then the corresponding statistic in the comparison distribution is computed. If the two distributions are of the same type, the ordered pairs will form a straight line since it is essentially a change of variables.

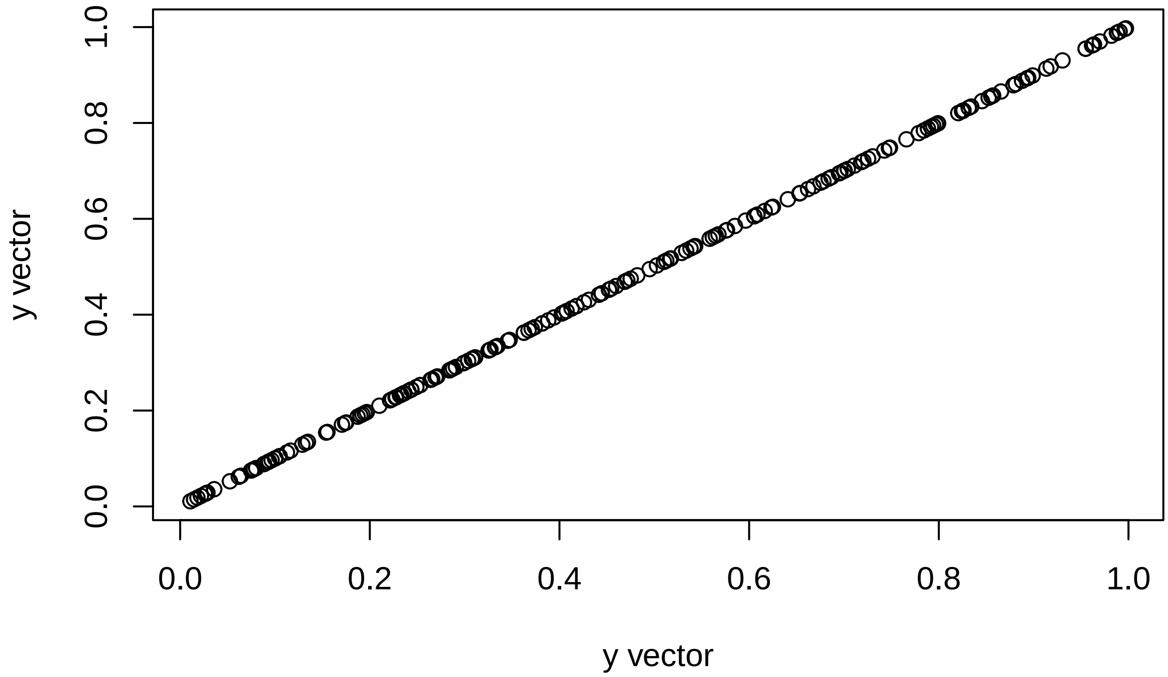
Example 5: Comparing Data with the normal distribution In this example, data from a uniform distribution is compared to a normal distribution and to itself. We first compare data from a uniform distribution to the normal distribution.

```
y=runif(200)
qqnorm(y, main = "Normal Q-Q Plot",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
       plot.it = TRUE, datax = FALSE)

qqline(y,  distribution = qnorm)
```

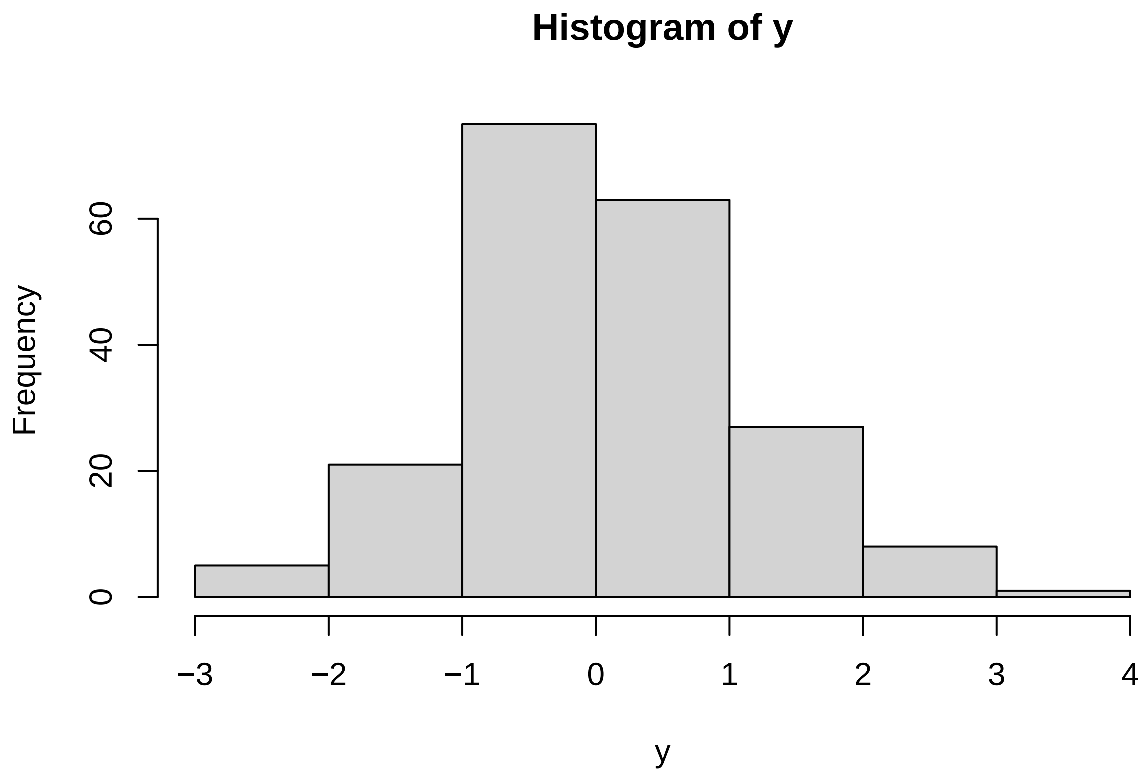



```
qqplot(y, y, xlab = "y vector", ylab = "y vector")
```



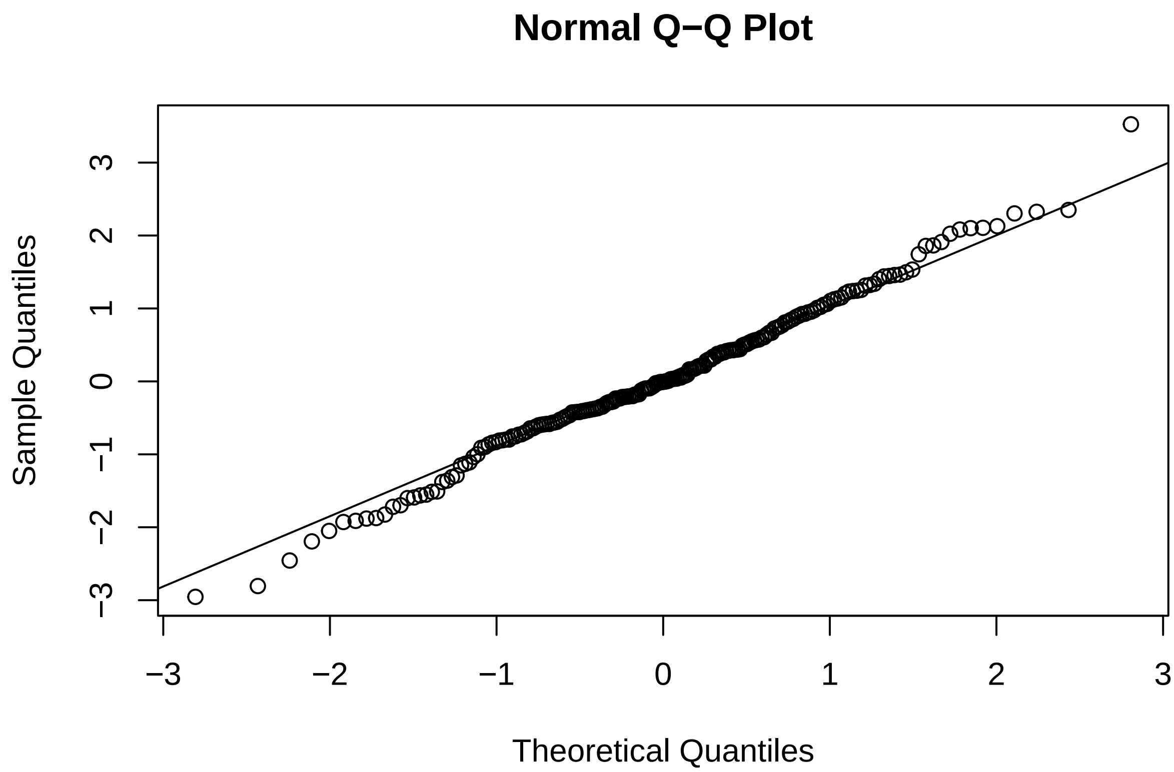
Take 2: a comparison with a normal distribution.

```
y=rnorm(200)
hist(y)
```



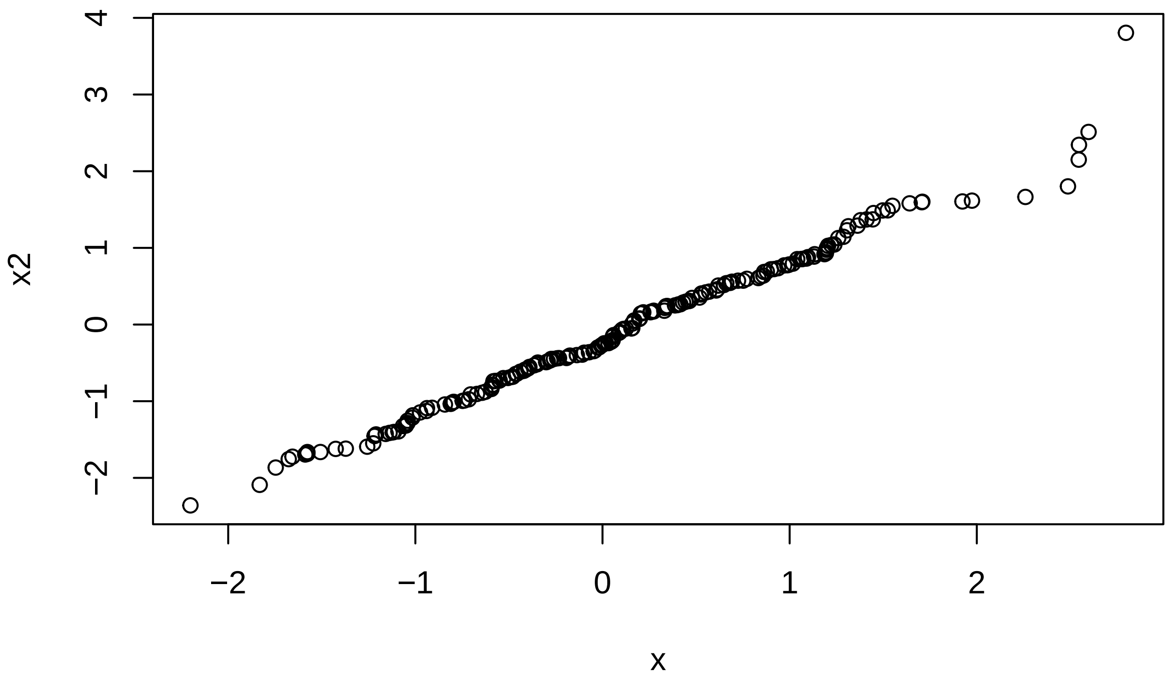

```
qqnorm(y, main = "Normal Q-Q Plot",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",plot.it = TRUE, datax = FALSE)

qqline(y,  distribution = qnorm)
```



Example 6: Comparison of normal data with qqplot We now compare normal data using the qqplot command in R. This is an alternative command, which will let us compare two data sets of our choice.

```
x=rnorm(200)
x2=rnorm(200)
qqplot(x,x2)
```



Hand computation of Q-Q Plots

- Compute the size of the data set.
- Rank order the data - the vector x
- For data in position i, compute $p = i/(n+1)$. The vector: $x2$
- Find z such that $P(Z < z) = p$. The vector: $x3$
- Plot the sorted data versus the $x3$ vector

The idea If Y_r is $N(\mu, \sigma^2)$ then

$$z = \frac{y_r - \mu}{\sigma}$$

and

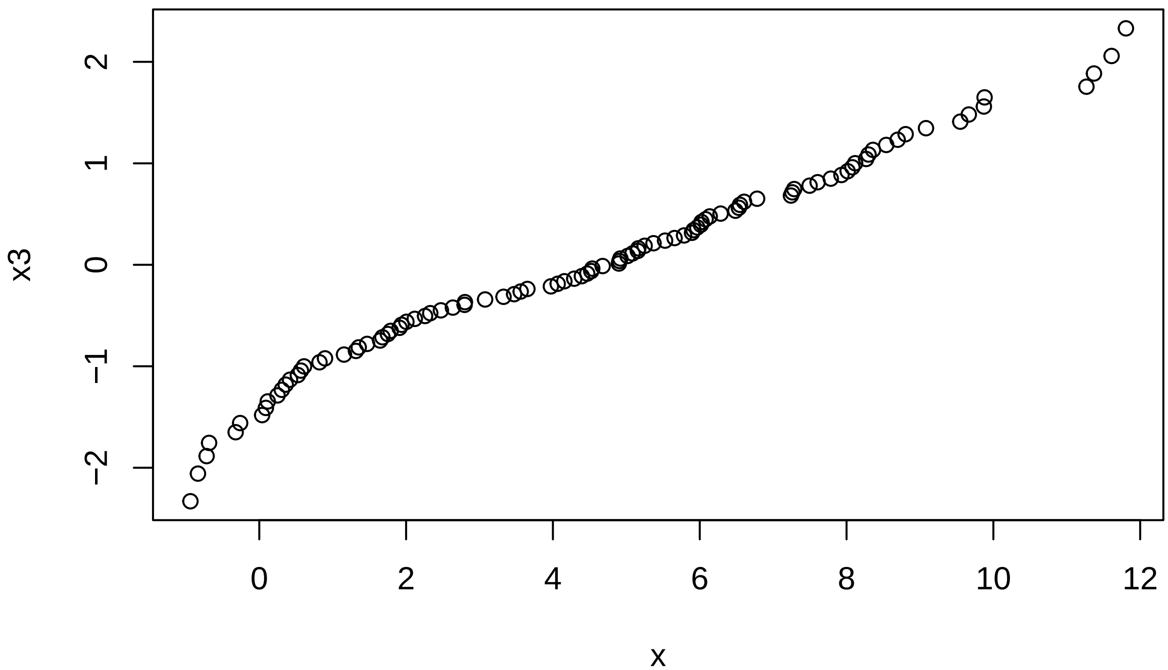
$$y_r = \mu + \sigma z.$$

If Y_r represents the kth percentile then $Y_r = \mu + \sigma z_{1-r}$ where $P(Z \leq z_{1-r}) = r$.

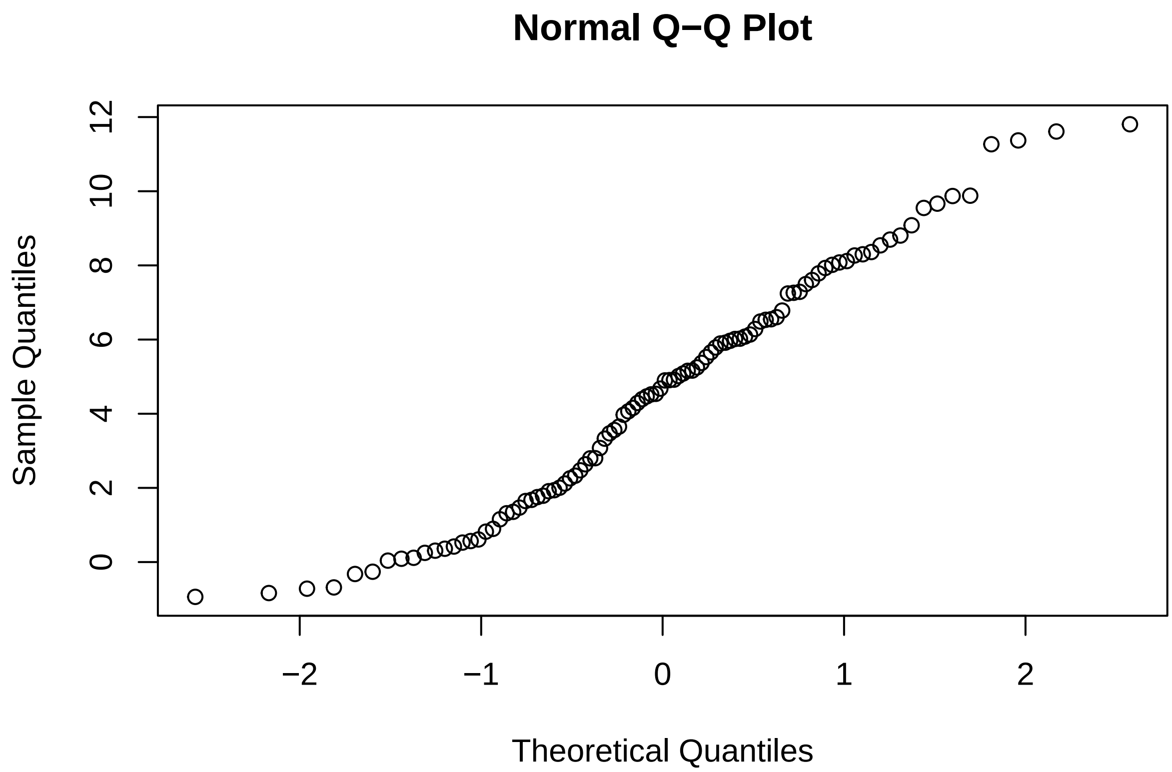
Example 7: Normal data We compute the q-q plot by hand in this example.

```
x<-sort(rnorm(100, mean=5, sd=3))

x2<-seq(1/101, 100/101, by=1/101)
x3<-qnorm(x2)
plot(x,x3)
```



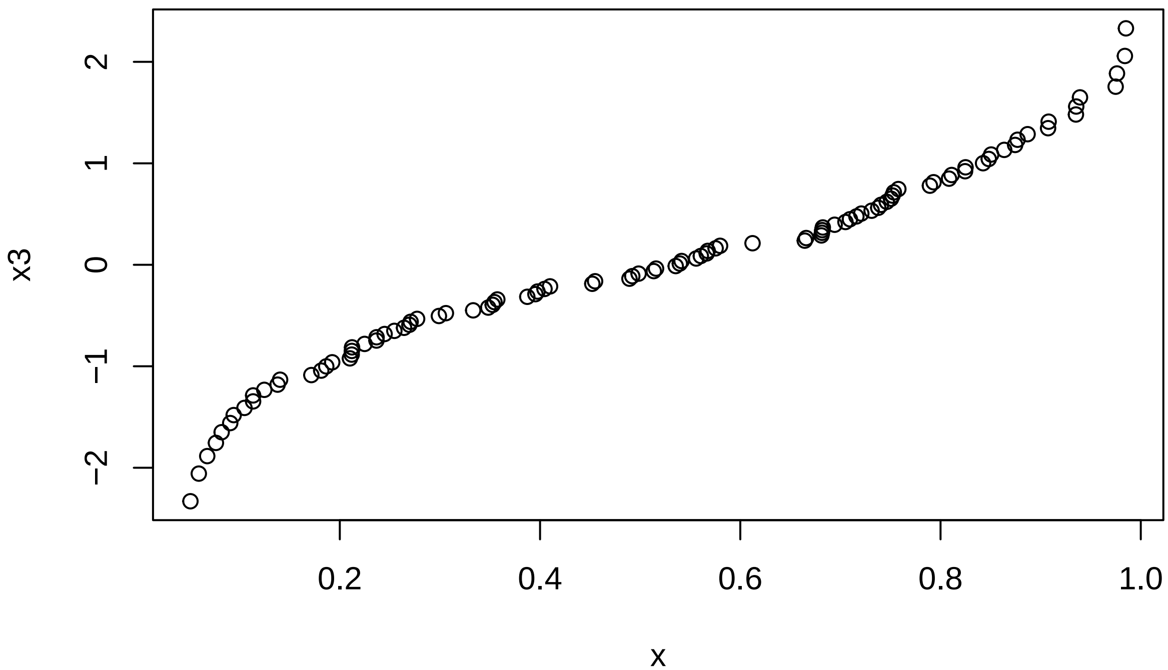

```
qqnorm(x)
```



Example 8: Uniform data - Not normally distributed.

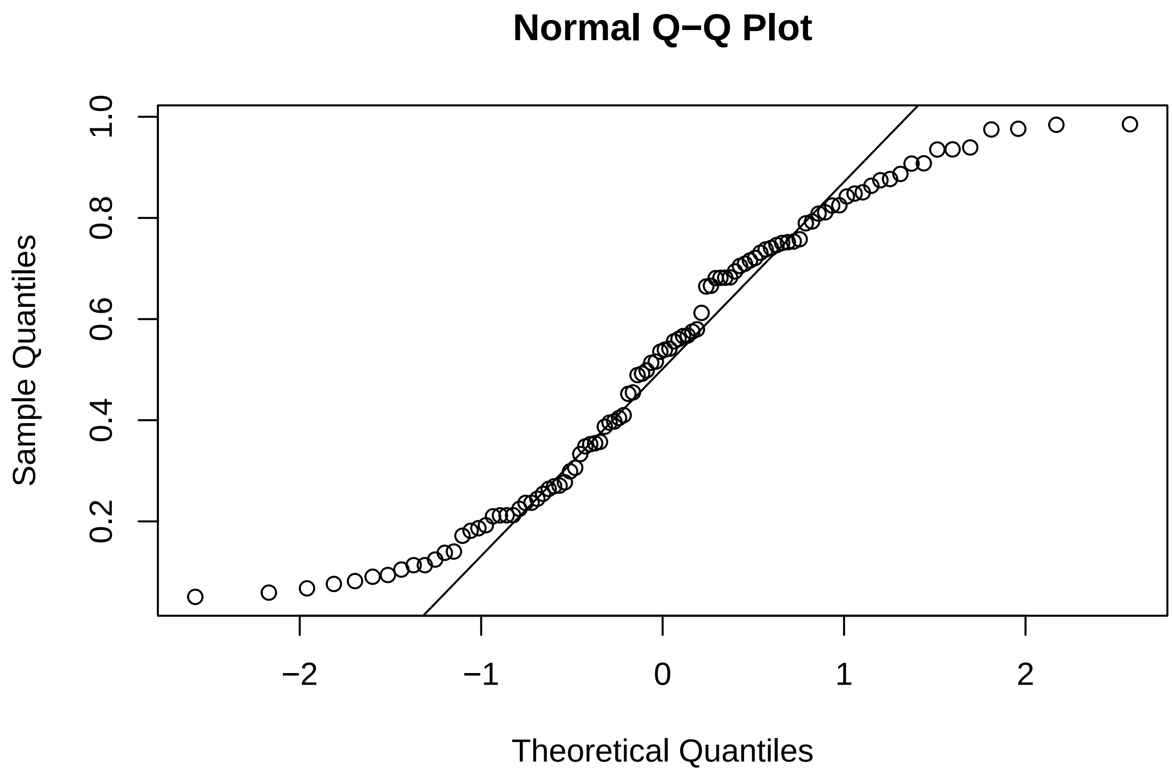
We compute the qq plot of uniform data.

```
x=sort(runif(100))
x2=seq(1/101, 100/101, by=1/101)
x3=qnorm(x2)
plot(x,x3)
```



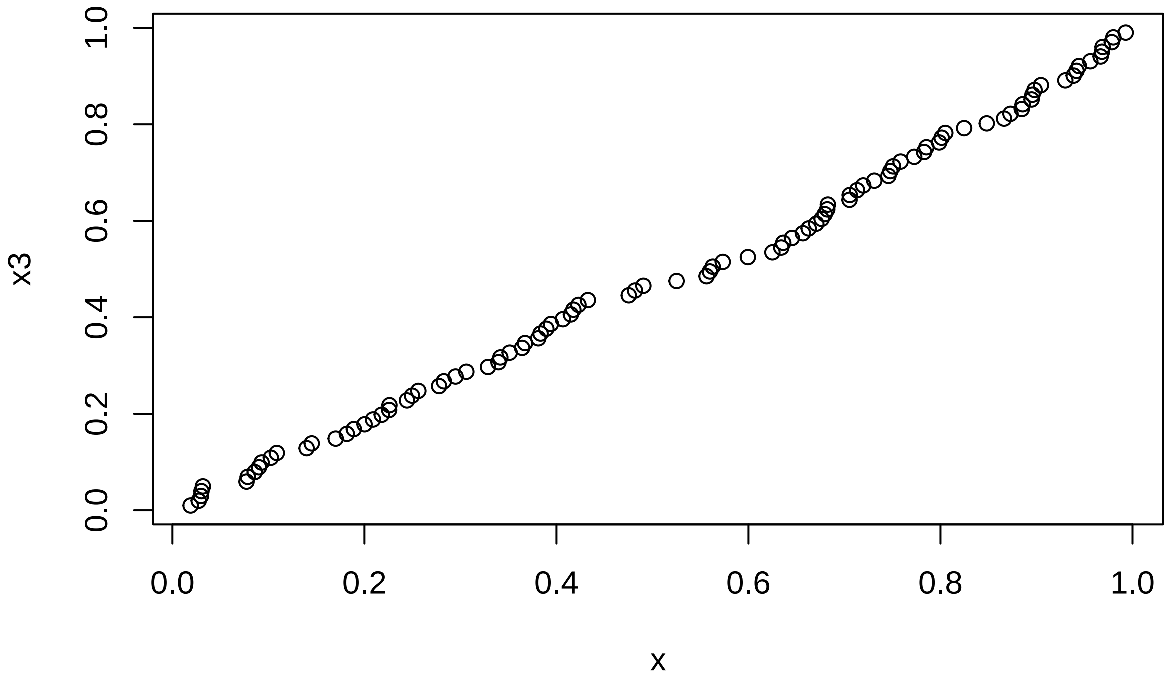
Compare your results to the output from qqnorm and qqline.

```
qqnorm(x); qqline(x)
```



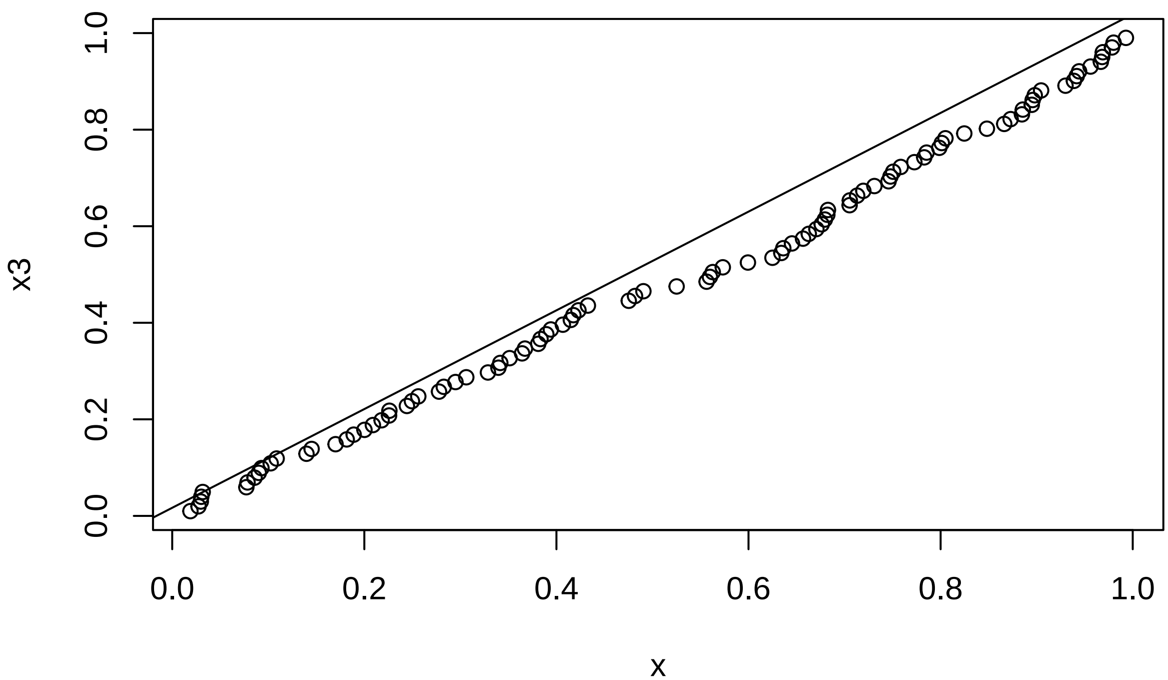
Example 9: Uniform Data compared to a uniform distribution We compare data from the uniform distribution to the uniform distribution.

```
x=sort(runif(100))
x2=seq(1/101, 100/101, by=1/101)
x3=qunif(x2)
plot(x,x3)
```



Compare your calculation to the output of qqplot.

```
qqplot(x, x3); qqline(x, distribution = qunif)
```



Basic Descriptive Statistics for samples

Beyond percentiles and histograms. We are work with a sample of size n. Assume that the population has size N. The theoretical distribution is unknown. Let x_i denote the ith sample. The term *population parameter* refers to the unknown parameters of the population. The *sample statistics* are calculated from the sample in an effort to estimate the parameters.

Example 10 - sample mean The mean of a sample of size n is denoted

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}.$$

The notation μ is used for the mean of the population.

Example 11 - the sample variance and the sample standard deviation The sample standard deviation is denoted as s and the sample variance, s^2, is the square of the sample standard deviation. The sample variance is

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

The variance is

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}.$$

Example 12: The mode The mode is the most frequently occurring value.

Example 13: The empirical rule The empirical rule is used to assess if a data set is bell shaped (or normal). For a normal distribution:

- approximately 68% of the data falls in the interval $(\bar{x} - s, \bar{x} + s)$
- approximately 95% of the data falls in the interval $(\bar{x} - 2s, \bar{x} + 2s)$
- approximately 99.7% of the data falls in the interval $(\bar{x} - 3s, \bar{x} + 3s)$

Example 14: Simpson's paradox

	$PlayerA : AB$	$Hits$	$Average$	$PlayerB : AB$	$Hits$	$Average$
$Season1$	500	126	0.252	300	75	0.250
$Season2$	300	90	0.300	500	145	0.290
$Total$	800	216	0.270	800	220	0.275

Notice that each season, Player A has a better average than Player B. However, over the two cumulative seasons, Player B has the better average. This is different from averaging the averages from each season.

```
set.seed(45)
mydata<-rbinom(40, 20, 0.5)
handmean<-sum(mydata)/length(mydata)
builtinmean<-mean(mydata)


handvariance<-sum((mydata-handmean)*(mydata - handmean))/(length(mydata)-1)
builtin<-sd(mydata)


emprule<-function(dvec, n){mean(dvec)+n*sd(dvec)*c(-1,1)}
emprule(mydata,2)
```

Example 16: R Demonstration of Descriptive Statistics

```
## [1]  4.859996 14.540004
```

```
std1<-emprule(mydata,1)
instd1<-mydata[mydata>std1[1] & mydata<std1[2]]

length(instd1)/length(mydata)
```

```
## [1] 0.775
```

```
mytable=table(mydata)
mytable
```

```
## mydata
##  4  6  7  8  9 10 11 12 13 14 18
##  1  1  2  6 13  7  3  2  2  2  1
```

```
names(mytable)[mytable==max(mytable)]
```

```
## [1] "9"
```