Stat 427/527: Advanced Data Analysis I

Chapter 4: Checking Assumptions

Instructor: Yan Lu

THE UNIVERSITY OF NEW MEXICO

- Statistical methods make assumptions about the data collection process and the shape of the population distribution.
- If you reject the null hypothesis in a test, then a reasonable conclusion is that the null hypothesis is false, provided all the distributional assumptions made by the test are satisfied.
  —-If the assumptions are not satisfied then that alone might be the cause of rejecting H_0.
  —-Additionally, if you fail to reject H_0, that could be caused solely by failure to satisfy assumptions also.
- Hence, you should always check assumptions to the best of your abilities.

Three basic assumptions

- Data are a random sample.
- The population frequency curve is normal.
- For the pooled variance two-sample test the population variances are also required to be equal.

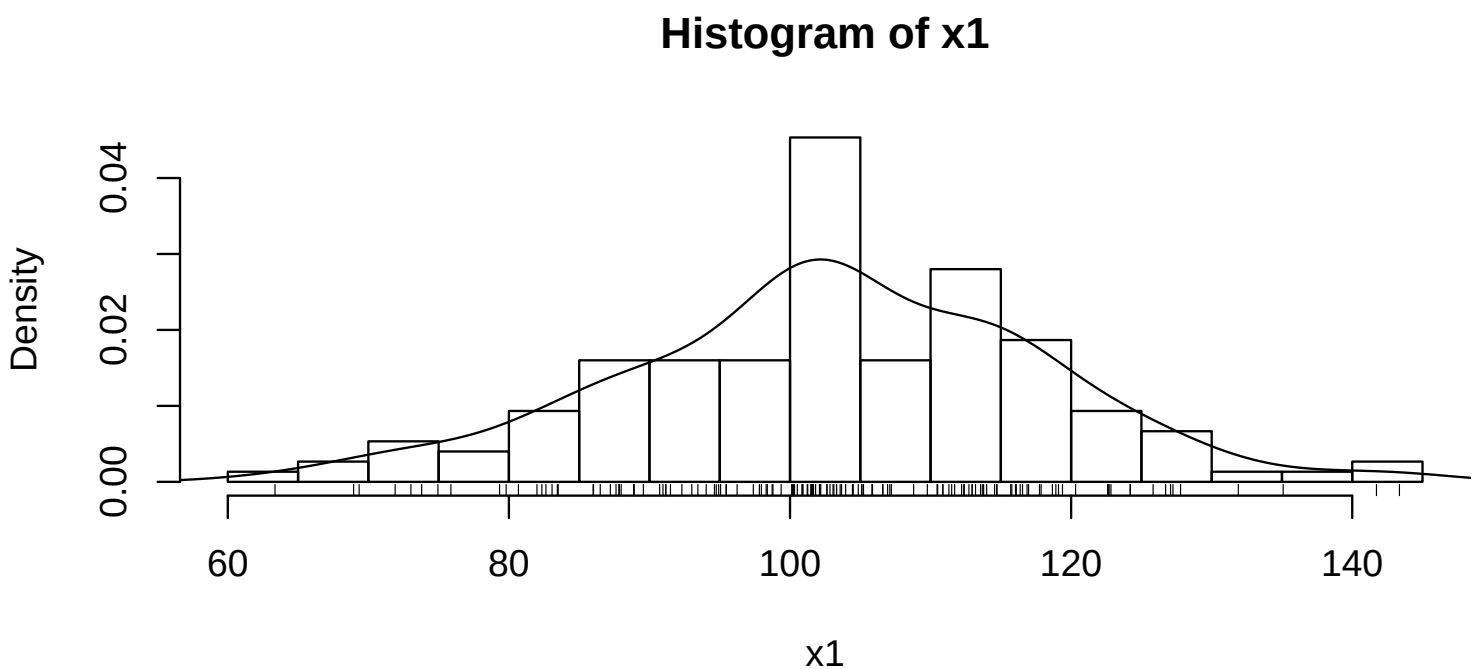
Testing Normality

An informal test of normality can be based on a *normal scores plot*, sometimes called a *rankit plot* or a *normal probability plot* or a *normal QQ plot* (QQ = quantile-quantile).

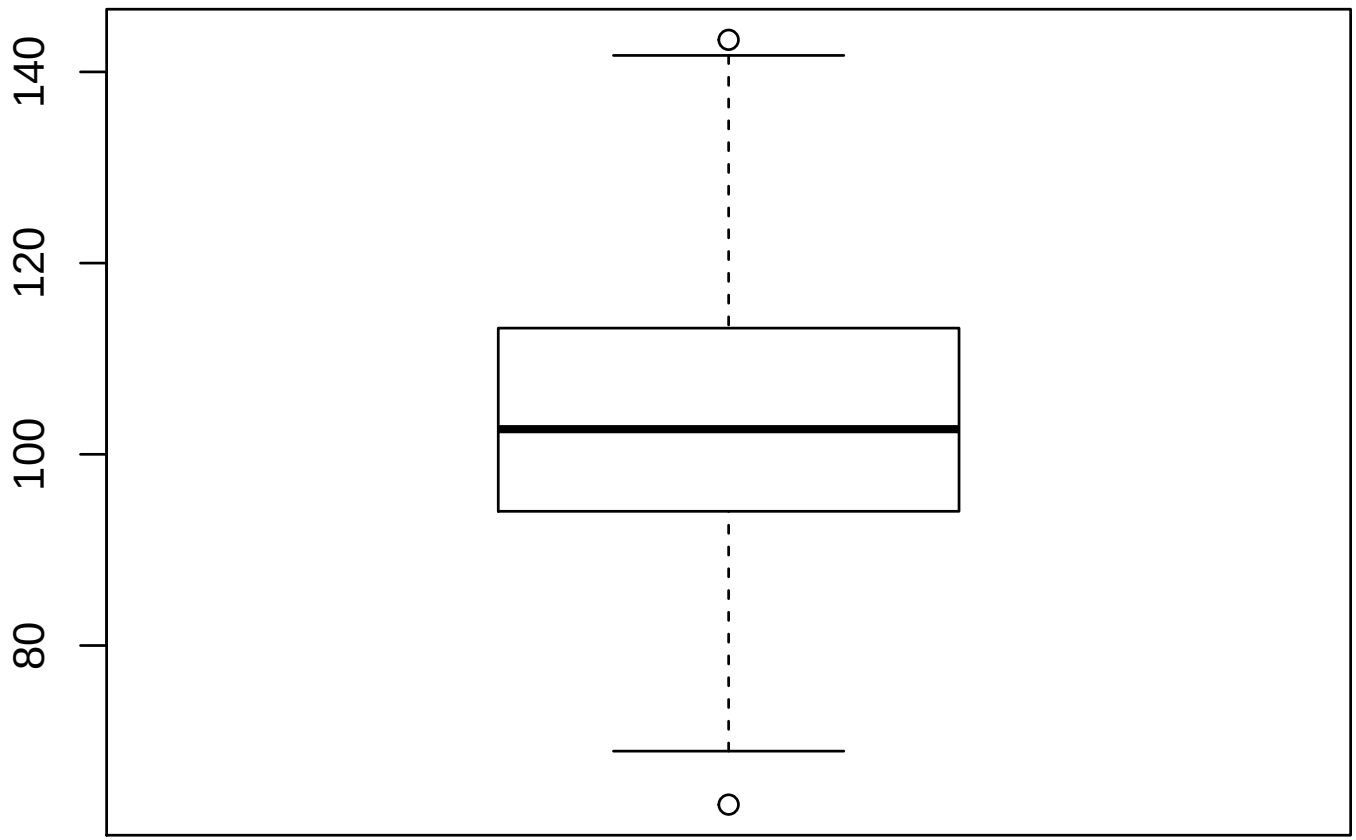
- plot the quantiles of the data against the quantiles of the normal distribution, or **expected normal order statistics** (in a standard normal distribution) for a sample with the given number of observations.
- The normality assumption is plausible if the plot is fairly linear.

- Start with simulated data from a normal distribution.

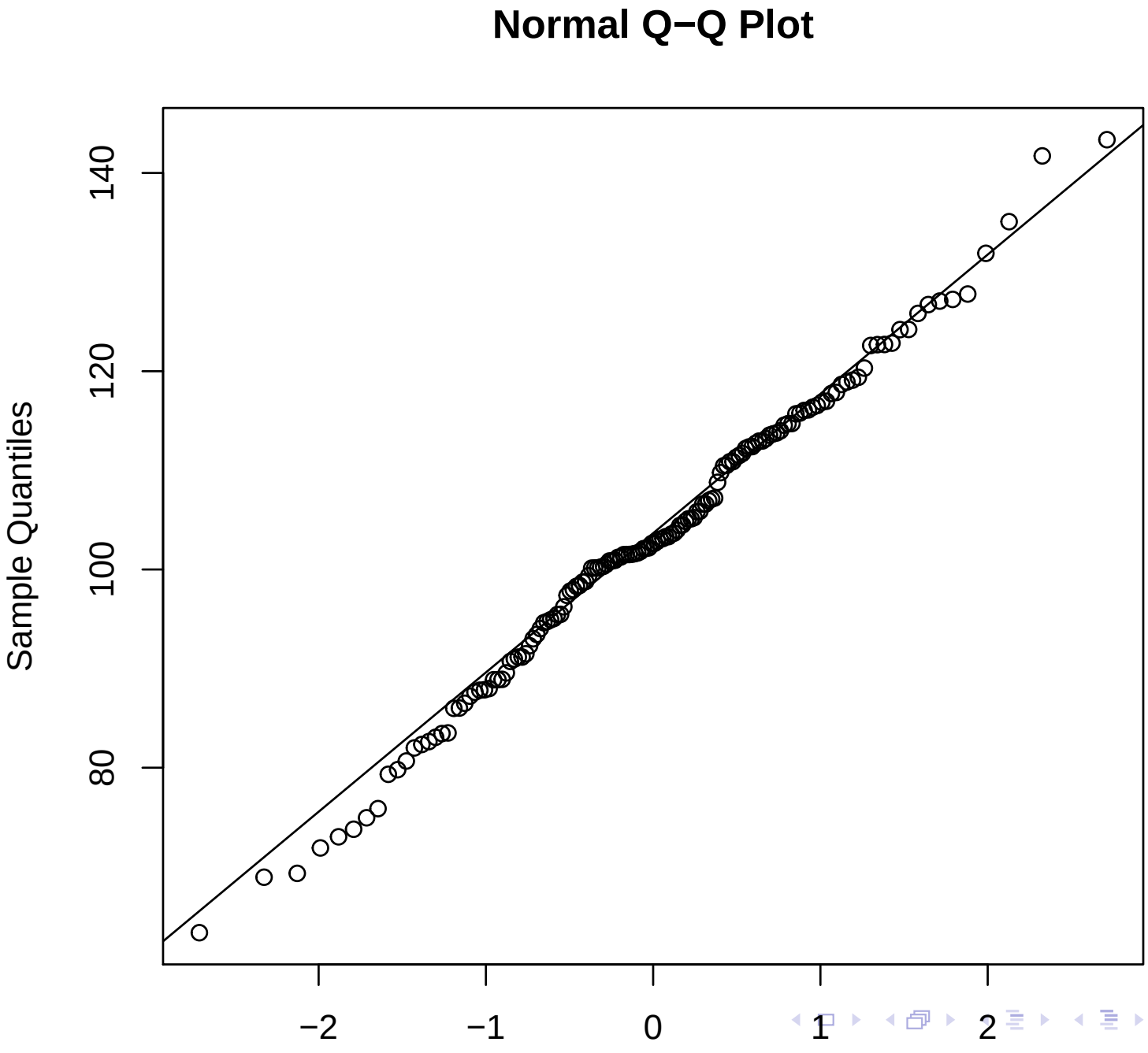
```
#### sample from normal distribution
x1 <- rnorm(150, mean = 100, sd = 15)

par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(x1, freq = FALSE, breaks = 20)
points(density(x1), type = "l")
rug(x1)
```



```r
# boxplot
boxplot(x1)
```

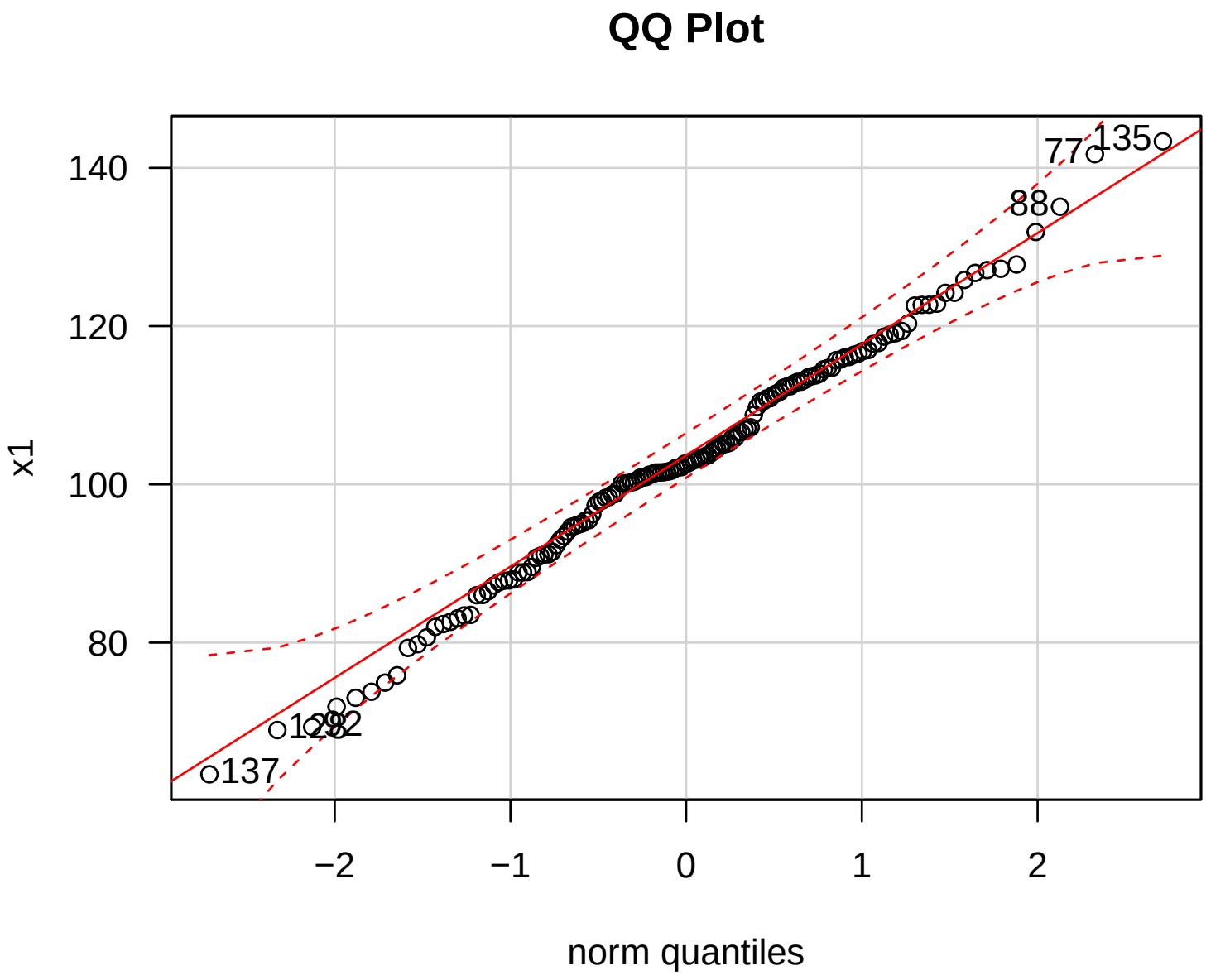


```
#### QQ plots
# R base graphics
par(mfrow=c(1,1))
# plots the data vs their normal scores
qqnorm(x1)
# plots the reference line
qqline(x1)
```


Normal Q−Q Plot
Sample Quantiles
140
120
100
80
−2
−1
0
1
2

- add confidence intervals (point-wise)

```
par(mfrow=c(1,1))
# Normality of Residuals
#install.packages("car")
library(car)
# qq plot for studentized resid
#   las    =  1 : turns labels on y-axis to
#               read horizontally
#   id.n   =  n : labels n most extreme observations,
#             and outputs to console
#   id.cex =  1 : is the size of those labels
#   lwd    =  1 : line width
qqPlot(x1, las = 1, id.n = 6, id.cex = 1, lwd = 1,
       main="QQ Plot")
```


QQ Plot
x1
norm quantiles
140
120
100
80
−2
−1
0
1
2
135
77
88
137

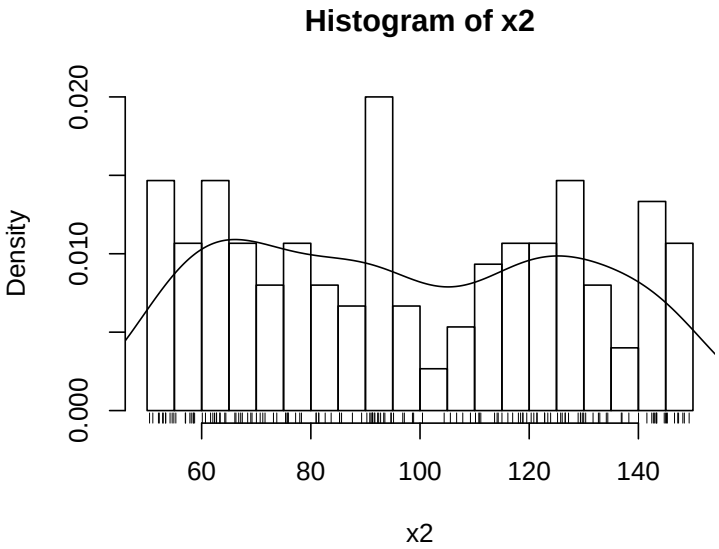
```
## 135 137  77 128  92  88
## 150   1 149   2   3 148
```

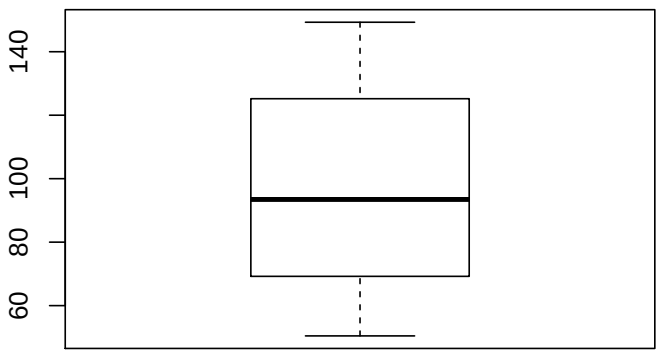
The above are 6 most extreme observations, with the corresponding values

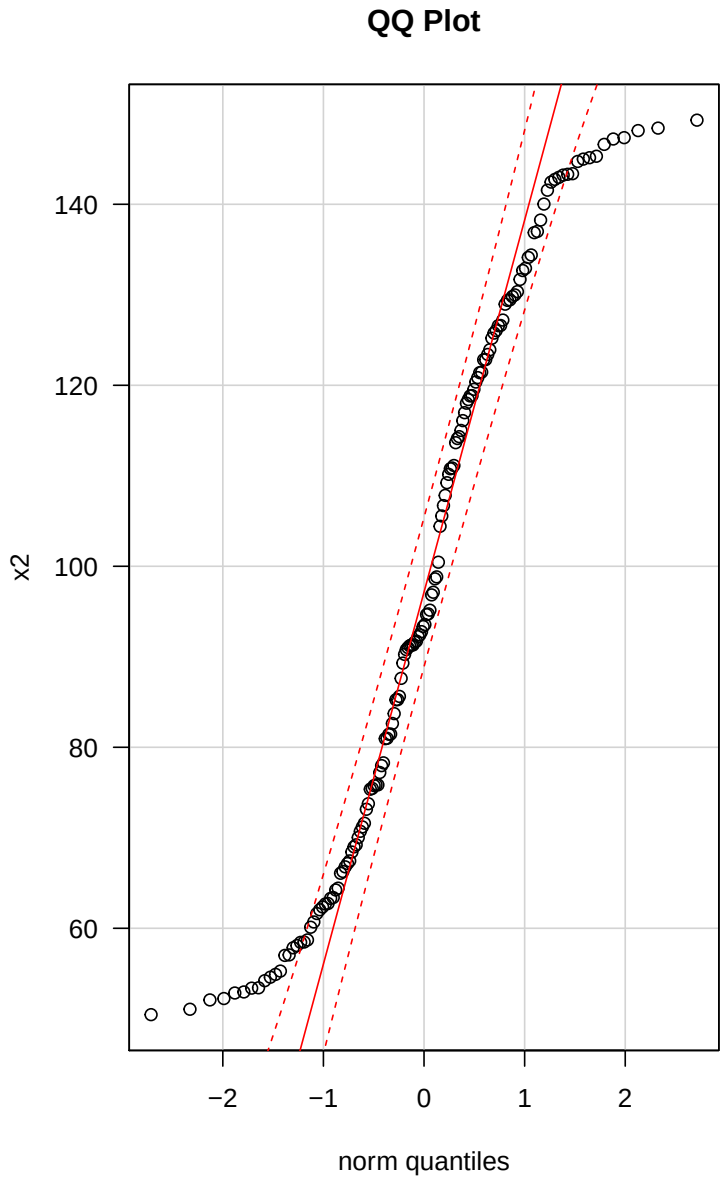
- x-axis is labelled "norm quantiles".
- This is the same graph as before, but with the normal scores identified with the percentiles to which they correspond.
- Only see a couple of data values outside the limits (in the tails, where it usually happens). Expect around 5% outside the limits, so there is no indication of non-normality here.
- We *did* sample from a normal population.

Light-tailed symmetric (Uniform)

```
#### Light-tailed symmetric (Uniform)
# sample from uniform distribution
x2 <- runif(150, min = 50, max = 150)

par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(x2, freq = FALSE, breaks = 20)
points(density(x2), type = "l")
rug(x2)

# boxplot
boxplot(x2)

par(mfrow=c(1,1))
qqPlot(x2, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot")
```

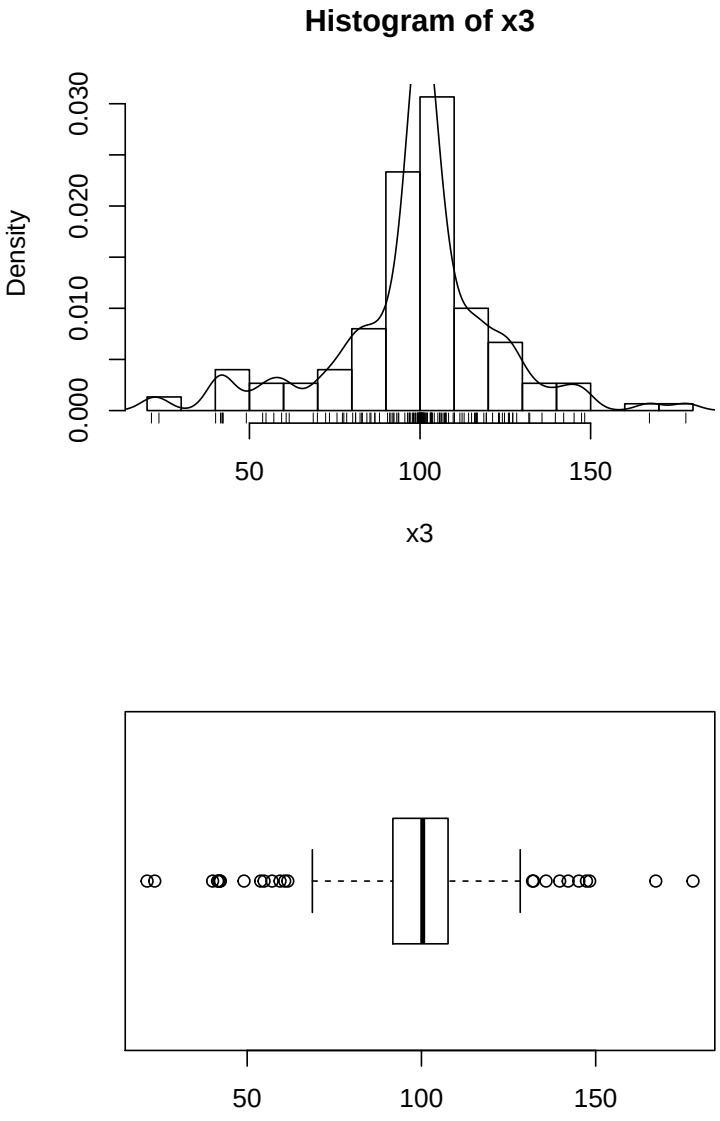

Histogram of x2
Density
0.000
0.010
0.020
60
80
100
120
140
x2

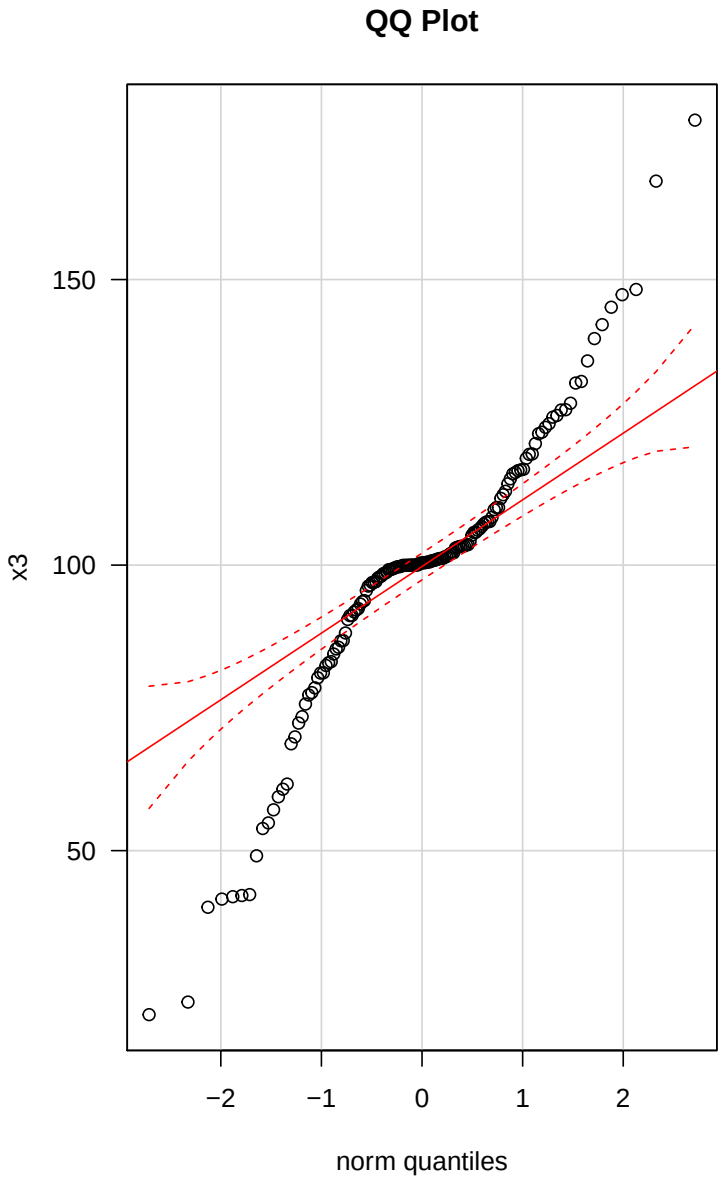

60
80
100
140


QQ Plot
x2
60
80
100
120
140
−2
−1
0
1
2
norm quantiles

Heavy-tailed (fairly) symmetric (Normal-squared)

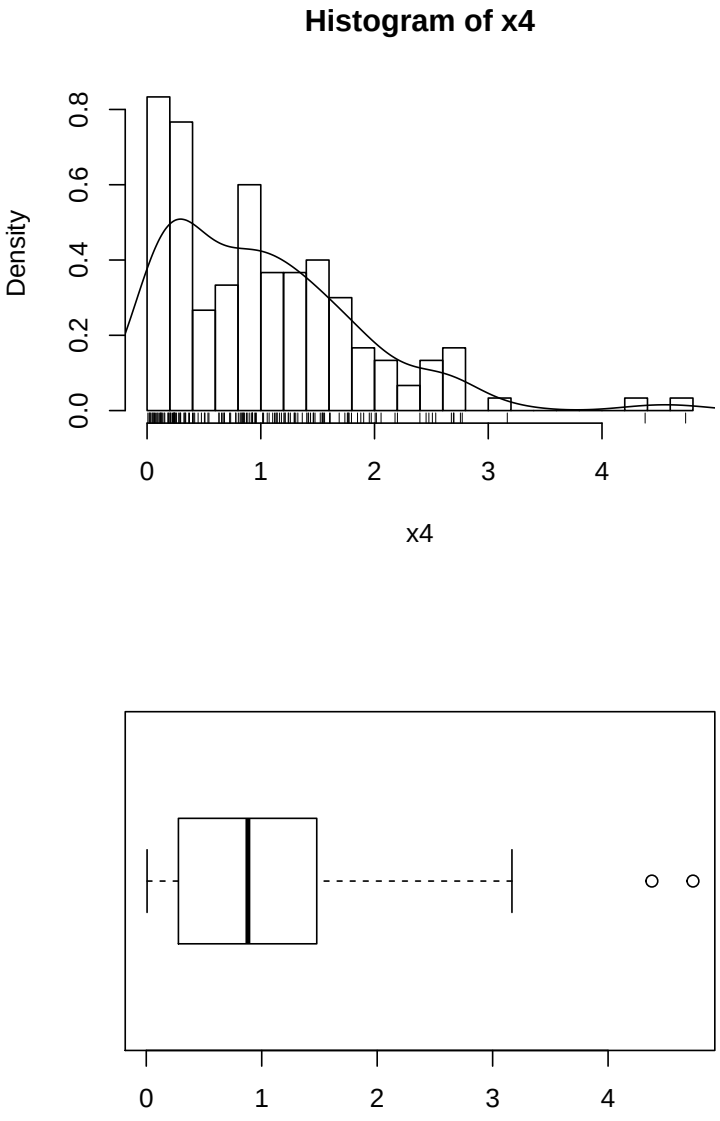
```r
#### Heavy-tailed (fairly) symmetric (Normal-squared)
# sample from normal distribution
x3.temp <- rnorm(150, mean = 0, sd = 1)
x3 <- sign(x3.temp)*x3.temp^2 * 15 + 100

par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(x3, freq = FALSE, breaks = 20)
points(density(x3), type = "l")
rug(x3)

# boxplot
boxplot(x3, horizontal=TRUE)


par(mfrow=c(1,1))
qqPlot(x3, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot")
```

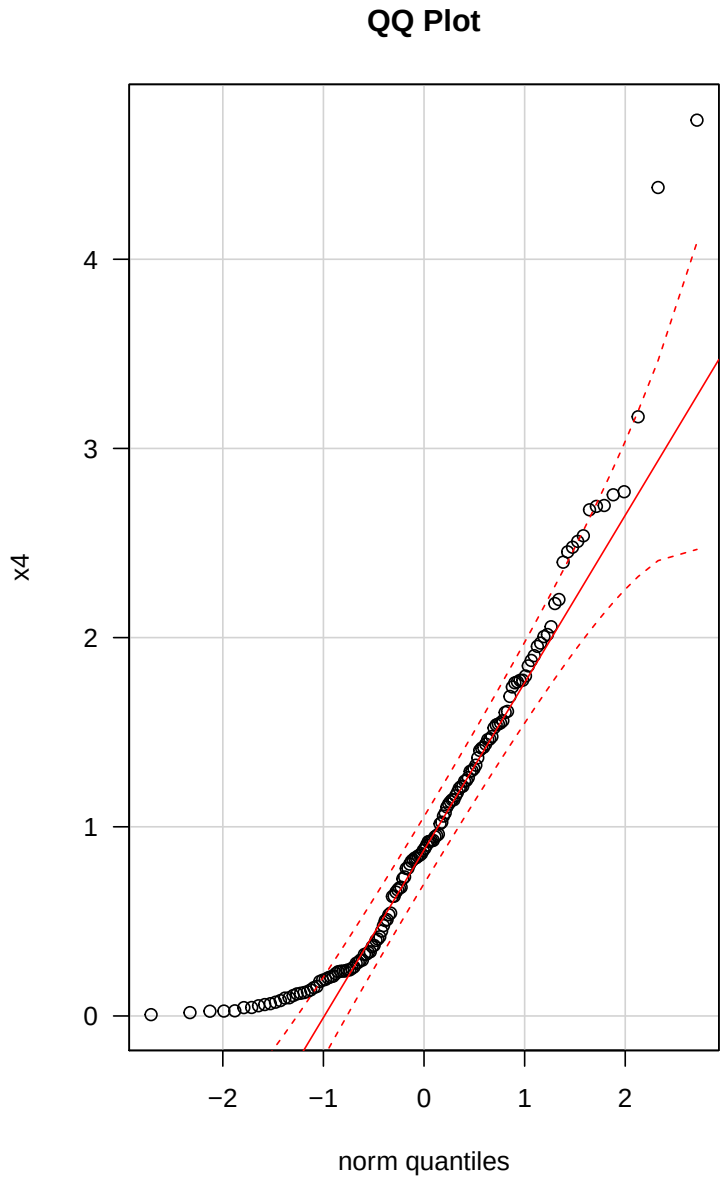

Histogram of x3
Density
0.000
0.010
0.020
0.030
50
100
150
x3
50
100
150


QQ Plot
x3
150
100
50
−2
−1
0
1
2
norm quantiles

Right-skewed (Exponential)

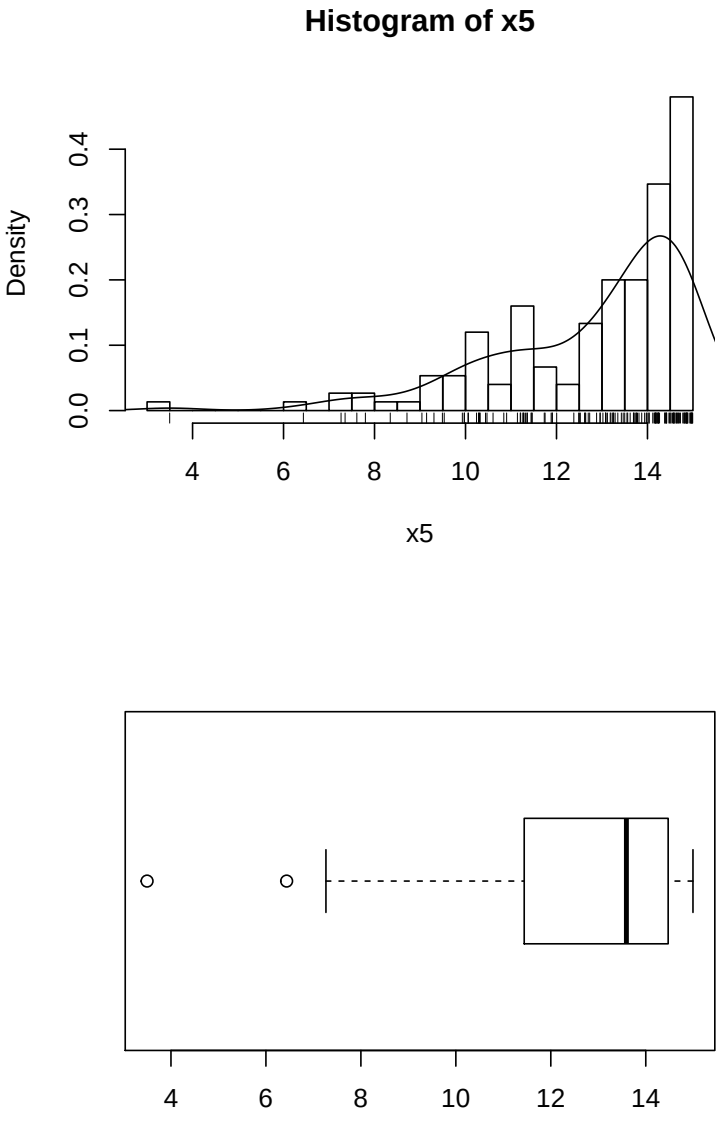
```
#### Right-skewed (Exponential)
# sample from exponential distribution
x4 <- rexp(150, rate = 1)

par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(x4, freq = FALSE, breaks = 20)
points(density(x4), type = "l")
rug(x4)

# boxplot
boxplot(x4, horizontal=TRUE)

par(mfrow=c(1,1))
qqPlot(x4, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot")
```

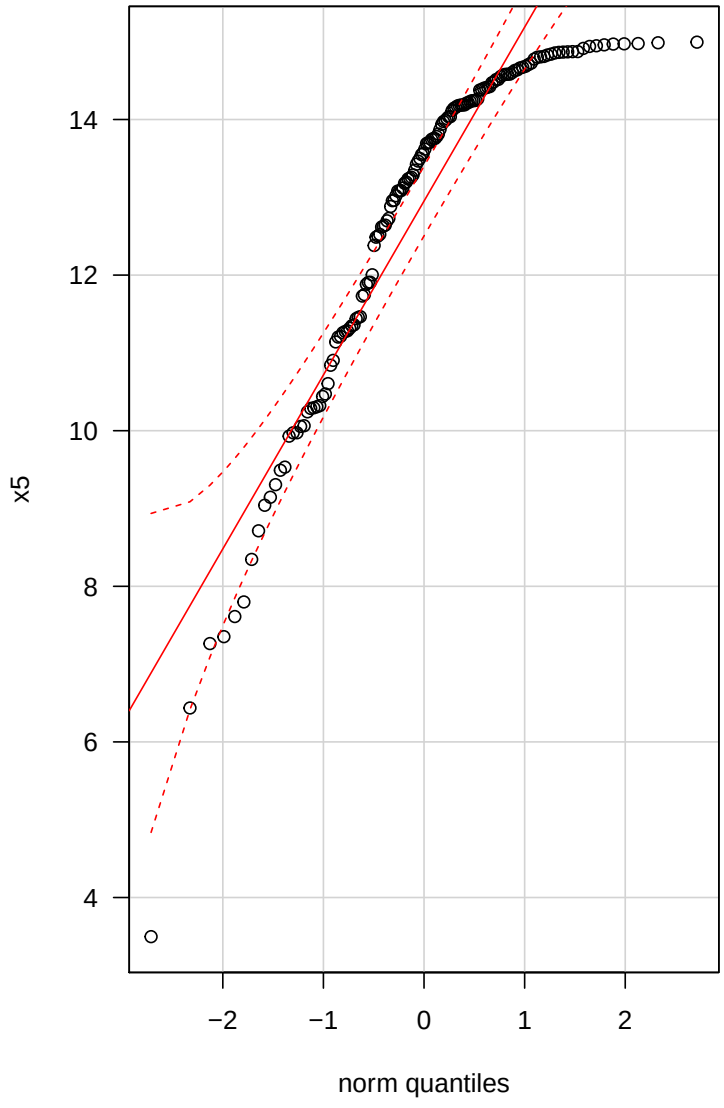

Histogram of x4
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
x4


QQ Plot
x4
0
1
2
3
4
−2
−1
0
1
2
norm quantiles

Left-skewed (Exponential, reversed)

```r
#### Left-skewed (Exponential, reversed)
# sample from exponential distribution
x5 <- 15 - rexp(150, rate = 0.5)

par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(x5, freq = FALSE, breaks = 20)
points(density(x5), type = "l")
rug(x5)

# boxplot
boxplot(x5, horizontal=TRUE)


par(mfrow=c(1,1))
qqPlot(x5, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot")
```


Histogram of x5
Density
0.0
0.1
0.2
0.3
0.4
4
6
8
10
12
14
x5


QQ Plot
x5
4
6
8
10
12
14
−2
−1
0
1
2
norm quantiles

Comments

Notice how striking is the lack of linearity in the QQ plot for all the non-normal distributions

—-The boxplot of the symmetric light-tailed distribution looks fairly good, however the QQ plot show the deviations.

—-The QQ plot is a sensitive measure of normality.

Let us summarize the patterns we see regarding tails in the plots:

	Tail	
Tail Weight	Left	Right
Light	Left side of plot points left	Right side of plot points right
Heavy	Left side of plot points down	Right side of plot points up

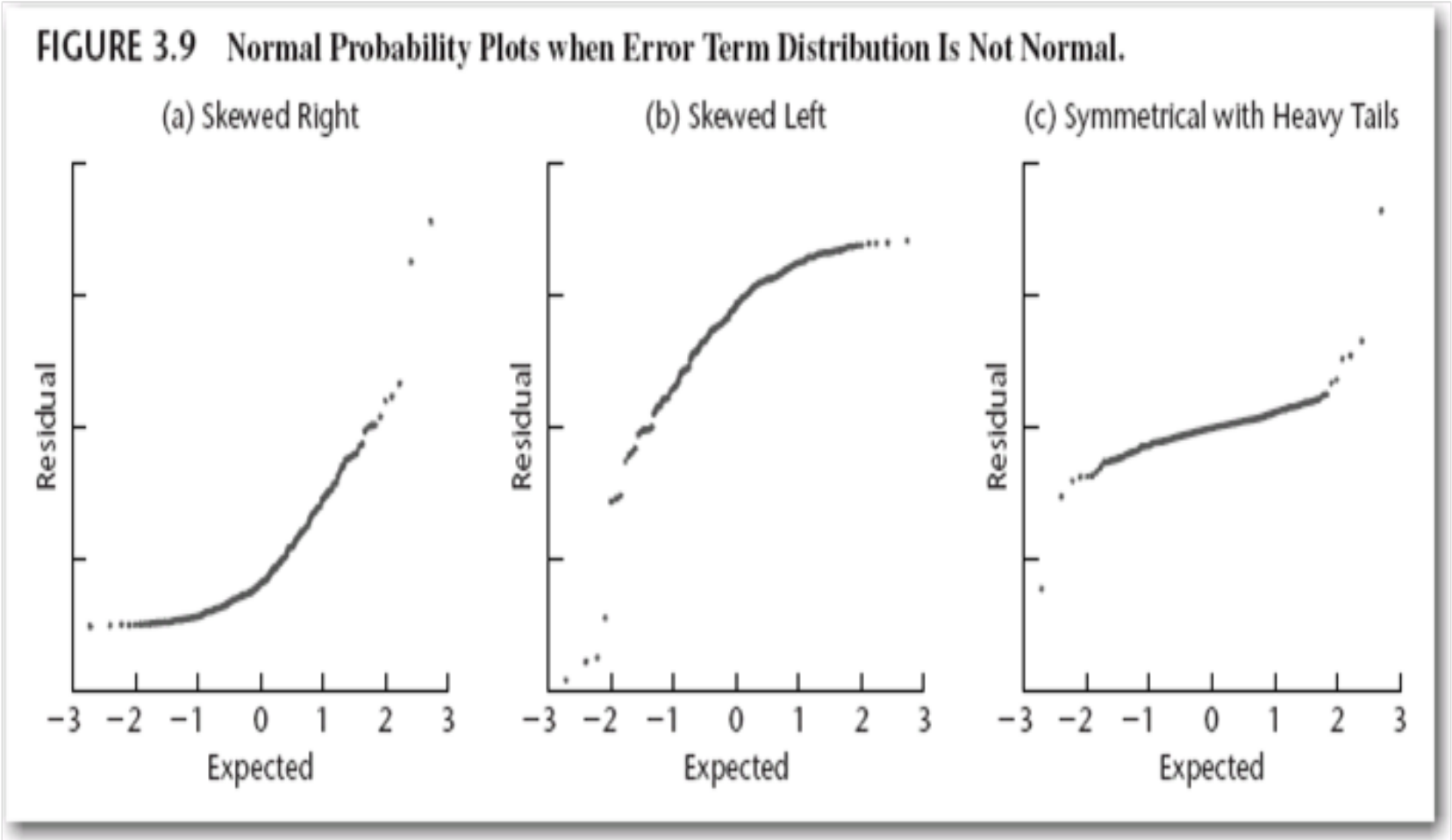


Figure 1 : Normal probability plots when error term distribution is not normal

Formal Tests of Normality

- A formal test of normality is based on the **correlation** between the data and the normal scores.
- The correlation is a measure of the strength of a linear relationship,
  —-with the sign of the correlation indicating the direction of the relationship (that is, $+$ for increasing relationship, and $-$ for decreasing).
- The correlation varies from -1 to $+1$.
  ——In a normal scores plot, you are looking for a correlation close to $+1$.
- Normality is rejected if the correlation is too small.

R has several tests of normality.

- **Shapiro-Wilk** test shapiro.test() is a base function.
- The R package nortest has some others:
  —-the **Anderson-Darling** test ad.test() is useful, related to the Kolmogorov-Smirnov
  —-the **Cramer-von Mises** test cvm.test(),

Extreme outliers and skewness have the biggest effects on standard methods based on normality.

- The Shapiro-Wilk (SW) test is better at picking up these problems than the Kolmogorov-Smirnov (KS) test.
- Tests for normality may have low power in small to moderate sized samples. Visual assessment of normality is often more valuable than a formal test.

Normal distribution

```
shapiro.test(x1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x1
## W = 0.99316, p-value = 0.6978
```

```
library(nortest)
ad.test(x1)
```

```
##
##  Anderson-Darling normality test
##
## data:  x1
## A = 0.41369, p-value = 0.3328
```

```
cvm.test(x1)
```

```
##
##  Cramer-von Mises normality test
##
## data:  x1
## W = 0.080059, p-value = 0.2052
```

Light-tailed symmetric

```
shapiro.test(x2)

##
##  Shapiro-Wilk normality test
##
## data:  x2
## W = 0.93579, p-value = 2.566e-06

library(nortest)
ad.test(x2)

##
##  Anderson-Darling normality test
##
## data:  x2
## A = 2.724, p-value = 6.826e-07

cvm.test(x2)

##
##  Cramer-von Mises normality test
##
## data:  x2
## W = 0.40862, p-value = 2.101e-05
```

Heavy-tailed (fairly) symmetric

```
shapiro.test(x3)

##
##  Shapiro-Wilk normality test
##
## data:  x3
## W = 0.92729, p-value = 6.459e-07

library(nortest)
ad.test(x3)

##
##  Anderson-Darling normality test
##
## data:  x3
## A = 4.7926, p-value = 6.396e-12

cvm.test(x3)

##
##  Cramer-von Mises normality test
##
## data:  x3
## W = 0.97764, p-value = 1.809e-09
```

Right-skewed

```
shapiro.test(x4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x4
## W = 0.89513, p-value = 7.087e-09
```

```
library(nortest)
ad.test(x4)
```

```
##
##  Anderson-Darling normality test
##
## data:  x4
## A = 3.1839, p-value = 5.132e-08
```

```
cvm.test(x4)
```

```
##
##  Cramer-von Mises normality test
##
## data:  x4
## W = 0.44715, p-value = 8.559e-06
```

Left-skewed

```
shapiro.test(x5)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x5
## W = 0.8547, p-value = 7.248e-11
```

```
library(nortest)
ad.test(x5)
```

```
##
##  Anderson-Darling normality test
##
## data:  x5
## A = 6.5439, p-value = 4.016e-16
```

```
cvm.test(x5)
```

```
## Warning in cvm.test(x5):  p-value is smaller than 7.37e-10,
cannot be computed more accurately
```

```
##
##  Cramer-von Mises normality test
##
## data:  x5
## W = 1.1491, p-value = 7.37e-10
```

Example: Paired Differences on Sleep Remedies

The following data give the amount of sleep

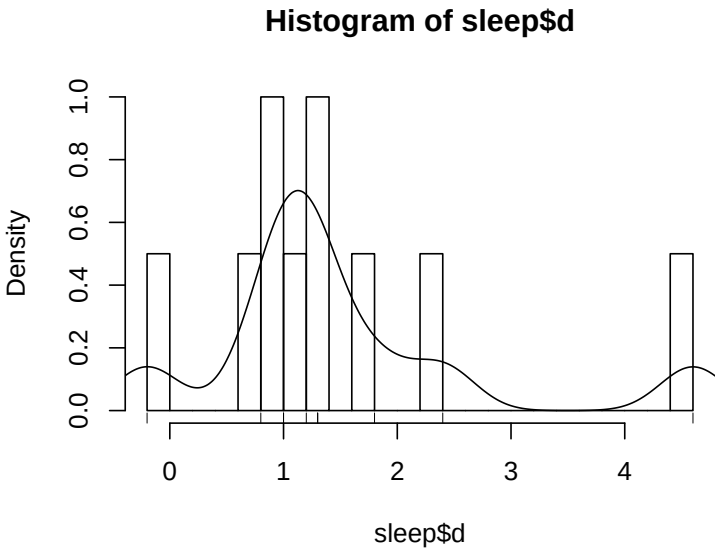
- gained in hours from two sleep remedies, A and B,
- 10 individuals who have trouble sleeping an adequate amount were observed
- Negative values imply sleep loss.
- In 9 of the 10 individuals, the sleep gain on B exceeded that on A.

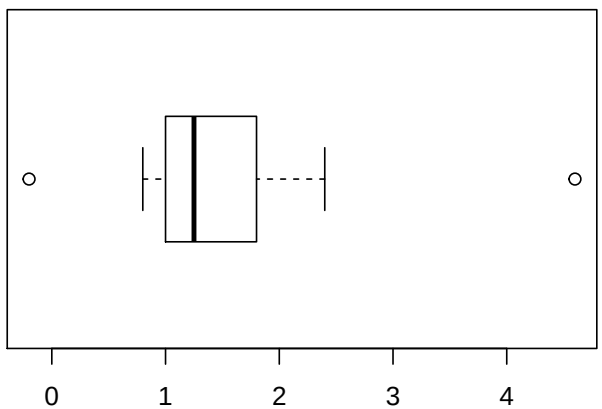
```
#### Example: Paired Differences on Sleep Remedies
# Data and numerical summaries
a <- c( 0.7, -1.6, -0.2, -1.2,  0.1,  3.4,  3.7,
        0.8,  0.0,  2.0)
b <- c( 1.9,  0.8,  1.1,  0.1, -0.1,  4.4,  5.5,
        1.6,  4.6,  3.0)
d <- b - a;
sleep <- data.frame(a, b, d)
sleep
```

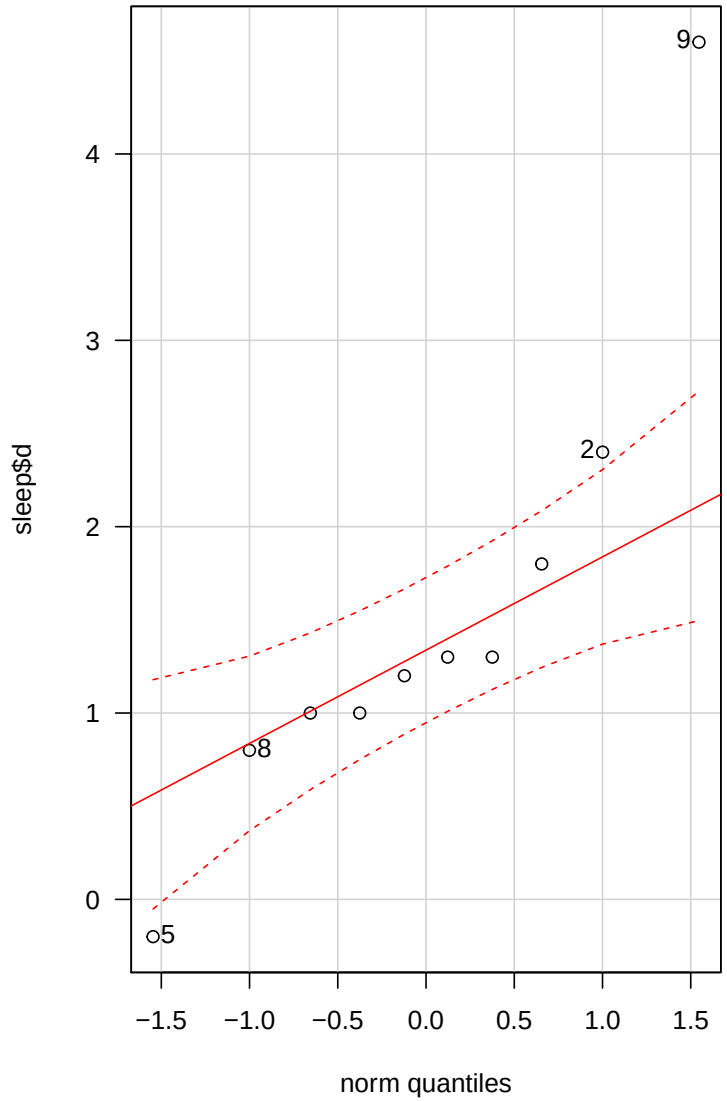
```
##        a    b    d
## 1    0.7  1.9  1.2
## 2   -1.6  0.8  2.4
## 3   -0.2  1.1  1.3
## 4   -1.2  0.1  1.3
## 5    0.1 -0.1 -0.2
## 6    3.4  4.4  1.0
## 7    3.7  5.5  1.8
## 8    0.8  1.6  0.8
## 9    0.0  4.6  4.6
## 10   2.0  3.0  1.0
```

```r
# plot of data
par(mfrow=c(2,1))
# Histogram overlaid with kernel density curve
hist(sleep$d, freq = FALSE, breaks = 20)
points(density(sleep$d), type = "l")
rug(sleep$d)

# boxplot
boxplot(sleep$d, horizontal=TRUE)



# QQ plot
par(mfrow=c(1,1))
qqPlot(sleep$d, las = 1, id.n = 4, id.cex = 1, lwd = 1,
       main="QQ Plot")
```

```
##  9  5  2  8
## 10  1  9  2
```


Histogram of sleep$d
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
sleep$d


0
1
2
3
4


QQ Plot
sleep$d
9
2
8
5
0
1
2
3
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
norm quantiles

```
# Normality tests
shapiro.test(sleep$d)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sleep$d
## W = 0.83798, p-value = 0.04173
```

```
library(nortest)
ad.test(sleep$d)
```

```
##
##  Anderson-Darling normality test
##
## data:  sleep$d
## A = 0.77378, p-value = 0.02898
```

```
# lillie.test(sleep£d)
cvm.test(sleep$d)
```

```
##
##  Cramer-von Mises normality test
##
## data:  sleep$d
## W = 0.13817, p-value = 0.02769
```

Summary of findings:

- The boxplot and normal scores plots suggest that the underlying distribution of differences for the paired sleep data is reasonably symmetric, but heavy tailed.
- The p-value for the SW test of normality is 0.042, and for the AD test is 0.029, both of which may call into question a normality assumption. Look further, SW test has a p-value of 0.042. This is not a strong rejection of the null hypothesis. Normality should still be an operational assumption.
- A non-parametric test comparing the sleep remedies (one that does not assume normality) is probably more appropriate here.

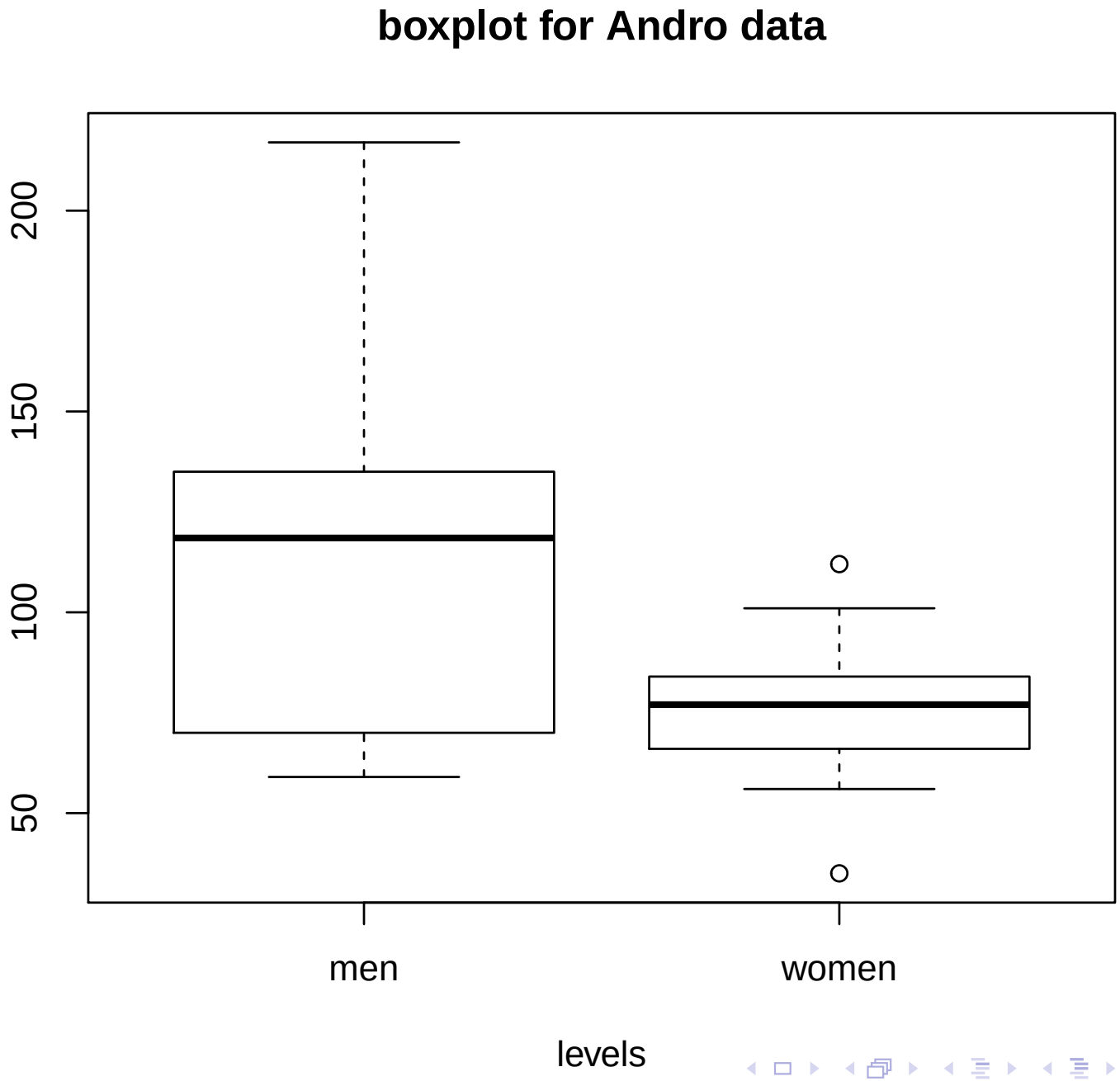
Example: Androstenedione Levels

This is an independent two-sample problem, so you must look at normal scores plots for the two groups: males and females.

```
#### Example: Androstenedione Levels
# Data and numerical summaries
men   <- c(217, 123,  80, 140, 115, 135,  59, 126,  70,  63,
           147, 122, 108,  70)
women <- c( 84,  87,  77,  84,  73,  66,  70,  35,  77,  73,
            56, 112,  56,  84,  80, 101,  66,  84)
level <- c(men, women)
sex <- c(rep("men", length(men)), rep("women", length(women)))
andro <- data.frame(level, sex)
head(andro)
```

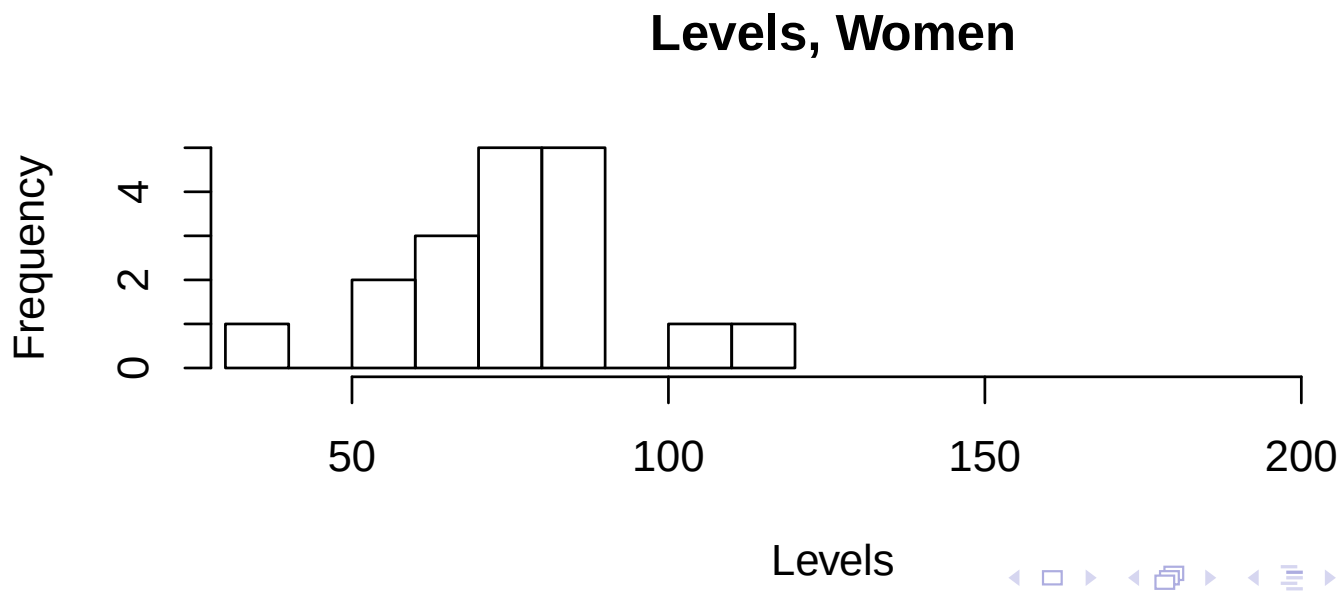
```
##   level sex
## 1   217 men
## 2   123 men
## 3    80 men
## 4   140 men
## 5   115 men
## 6   135 men
```

```r
# boxplot using R base graphics

boxplot(level ~ sex, method = "stack", data = andro,
  horizontal = FALSE,
  main = "boxplot for Andro data", xlab = "levels")
```

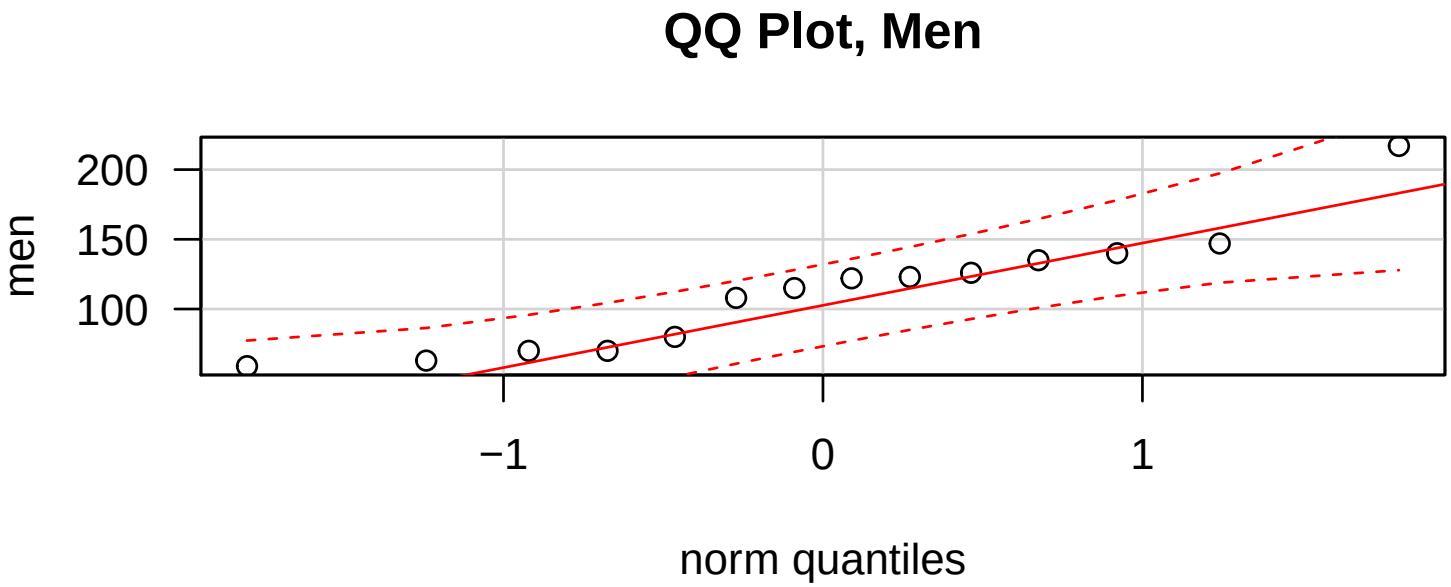

boxplot for Andro data
200
150
100
50
men
women
levels

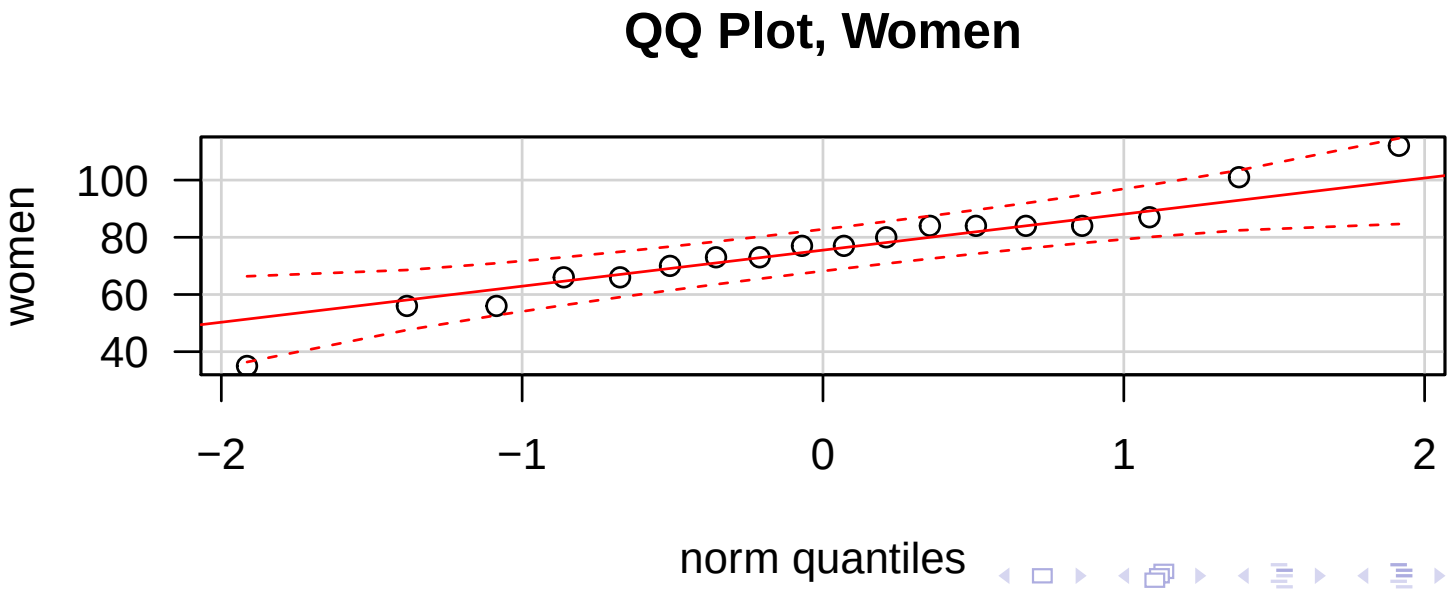
```r
# common x-axis limits based on the range of the entire data set
par(mfrow=c(2,1))
hist(andro$level[(andro$sex == "men")],
     xlim = range(andro$level),
  main = "Levels, Men", xlab = "Levels")
hist(andro$level[(andro$sex == "women")],
     xlim = range(andro$level),
  main = "Levels, Women", xlab = "Levels")
```


Levels, Men
Frequency
0
4
8
50
100
150
200
Levels


Levels, Women
Frequency
0
2
4
50
100
150
200
Levels

```
# QQ plot
par(mfrow=c(2,1))
qqPlot(men, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot, Men")
qqPlot(women, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       main="QQ Plot, Women")
```


QQ Plot, Men
men
200
150
100
−1
0
1
norm quantiles


QQ Plot, Women
women
100
80
60
40
−2
−1
0
1
2
norm quantiles

- The women's boxplot contains two mild outliers, which is unusual when sampling from a normal distribution.
  ——The tests are possibly not powerful enough to pick up this type of deviation from normality in such a small sample.
  —-In practice, this may not be a big concern. The two mild outliers probably have a small effect on inferences in the sense that non-parametric methods would probably lead to similar conclusions here.
- Histogram plots didn't show extreme skewness
- QQ plot look fine.

Tests: Men

```
shapiro.test(men)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  men
## W = 0.90595, p-value = 0.1376
```

```
library(nortest)
ad.test(men)
```

```
##
##  Anderson-Darling normality test
##
## data:  men
## A = 0.4718, p-value = 0.2058
```

```
cvm.test(men)
```

```
##
##  Cramer-von Mises normality test
##
## data:  men
## W = 0.063063, p-value = 0.3221
```

Women

```
shapiro.test(women)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  women
## W = 0.95975, p-value = 0.5969
```

```
library(nortest)
ad.test(women)
```

```
##
##  Anderson-Darling normality test
##
## data:  women
## A = 0.39468, p-value = 0.3364
```

```
cvm.test(women)
```

```
##
##  Cramer-von Mises normality test
##
## data:  women
## W = 0.065242, p-value = 0.3057
```

- Both the AD test p-value and the SW test p-value for testing normality exceeds 0.10 in each sample.
- Thus, given the sample sizes (14 for men, 18 for women), we have insufficient evidence (at $\alpha = 0.05$) to reject normality in either population.
- Most statisticians use graphical methods (boxplot, normal scores plot) to assess normality, and do not carry out formal tests.

In the independent two sample t-test, we want to test

$$H_0 : \sigma_1^2 = \sigma_2^2$$

- to decide between using the pooled-variance procedure or Satterthwaite's methods.
  ——suggest the pooled t-test and CI if H_0 is not rejected, and Satterthwaite's methods otherwise.
- number of well-known tests for equal population variances, of which Bartlett's test and Levene's test are probably the best known.
- Bartlett's test assumes the population distributions are normal
  ——check normality prior to using Bartlett's test.
- Levene's test is more robust to departures from normality than Bartlett's test; it is in the car package.

Bartlett's test

- let $n^* = n_1 + n_2 + \cdots + n_k$, where the n_is are the sample sizes from the k groups, and define

$$v = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n^* - k} \right).$$

- Bartlett's statistic for testing $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ is given by

$$B_{obs} = \frac{2.303}{v} \left\{ (n-k) \log(s^2_{\text{pooled}}) - \sum_{i=1}^{k} (n_i - 1) \log(s_i^2) \right\},$$

where s^2_{pooled} is the pooled estimator of variance and s_i^2 is the estimated variance based on the i^{th} sample.

- Large values of B_{obs} suggest that the population variances are unequal.

——For a size α test, we reject H_0 if $B_{\text{obs}} \geq \chi^2_{k-1,\text{crit}}$, where $\chi^2_{k-1,\text{crit}}$ is the upper-α percentile for the χ^2_{k-1} (chi-squared) probability distribution with $k-1$ degrees of freedom.

——A p-value for the test is given by the area under the chi-squared curve to the right of B_{obs}.

Example: Androstenedione Levels continued

The sample standard deviations and samples sizes are:

$$s_1 = 42.8, \text{ and } n_1 = 14 \text{ for men}$$

$$s_2 = 17.2, \text{ and } n_2 = 18 \text{ for women}$$

- The sample standard deviations appear to be very different
- Expect the test of equal population variances is highly significant.
- The output below confirms this: the p-values for Bartlett's test, Levene's Test are both much smaller than 0.05. An implication is that the standard pooled-CI and test on the population means is inappropriate.

```
#### Testing Equal Population Variances
# numerical summaries
c(mean(men), mean(women), sd(men), sd(women))

## [1] 112.50000  75.83333  42.75467  17.23625

c(IQR(men), IQR(women), length(men), length(women))

## [1] 60.25 17.00 14.00 18.00

## Test equal variance
# assumes populations are normal
bartlett.test(level ~ sex, data = andro)

##
##  Bartlett test of homogeneity of variances
##
## data:  level by sex
## Bartlett's K-squared = 11.199, df = 1, p-value = 0.0008183
```

```
# does not assume normality, requires car package
library(car)
leveneTest(level ~ sex, data = andro)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  7.2015 0.01174 *
##       30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary:

- Always Plot, Plot, Plot first.
- Tests for normality may have low power in small to moderate sized samples. Visual assessment of normality is often more valuable than a formal test.
- Tests for equal variance assumption have the same problem. We will discuss using residuals to assess the equal variance assumption later.