Testing the Assumption of Normality

- One of the assumptions of many statistical procedures (including the **t-test**) is that the population from which you are sampling is normally distributed.
- The t-test is said to be rather **robust** in terms of this assumption, which means that reality can deviate from this assumption a fair amount without seriously affecting the validity of the analysis.

Testing the Assumption of Normality

- This is particularly true when the size of the samples is large (thanks to the Central Limit Theorem).
- Some deviations from normality can pose a problem for the t-test, specifically those that involve getting extreme scores more frequently than you would if the distribution were normally distributed.

Testing the Assumption of Normality

Using Computer Software

- Statistical Software Packages provides two statistical tests for deviation from normality, the "Kolomogorov-Smirnov" family of tests and the "Shapiro-Wilk" test.
- The 'Kolomogorov-Smirnov' test can be used to test if two data sets are distributed according to the same distribution.
- It can also be used to test if one data set comes from a specified distribution, such as the normal distribution.

Testing the Assumption of Normality

- For the purposes of this module, we will only use a special case of the 'Kolomogorov-Smirnov' test, known as the "**Anderson-Darling**" test of normality.
- The Shapiro-Wilk Test can be implemented using the `shapiro.test()` command in `R`.

Testing the Assumption of Normality

Implementing the Tests

- The "Anderson-Darling" test can not be implemented immediately with `R`.
- Using the Anderson Darling Test requires the installation of the **nortest** package.
- Then the test can be implemented using the `ad.test()` command.
- (There are actually more procedures. We will look at R packages in greater detail on an ongoing basis.)

Testing the Assumption of Normality

Implementing the Tests

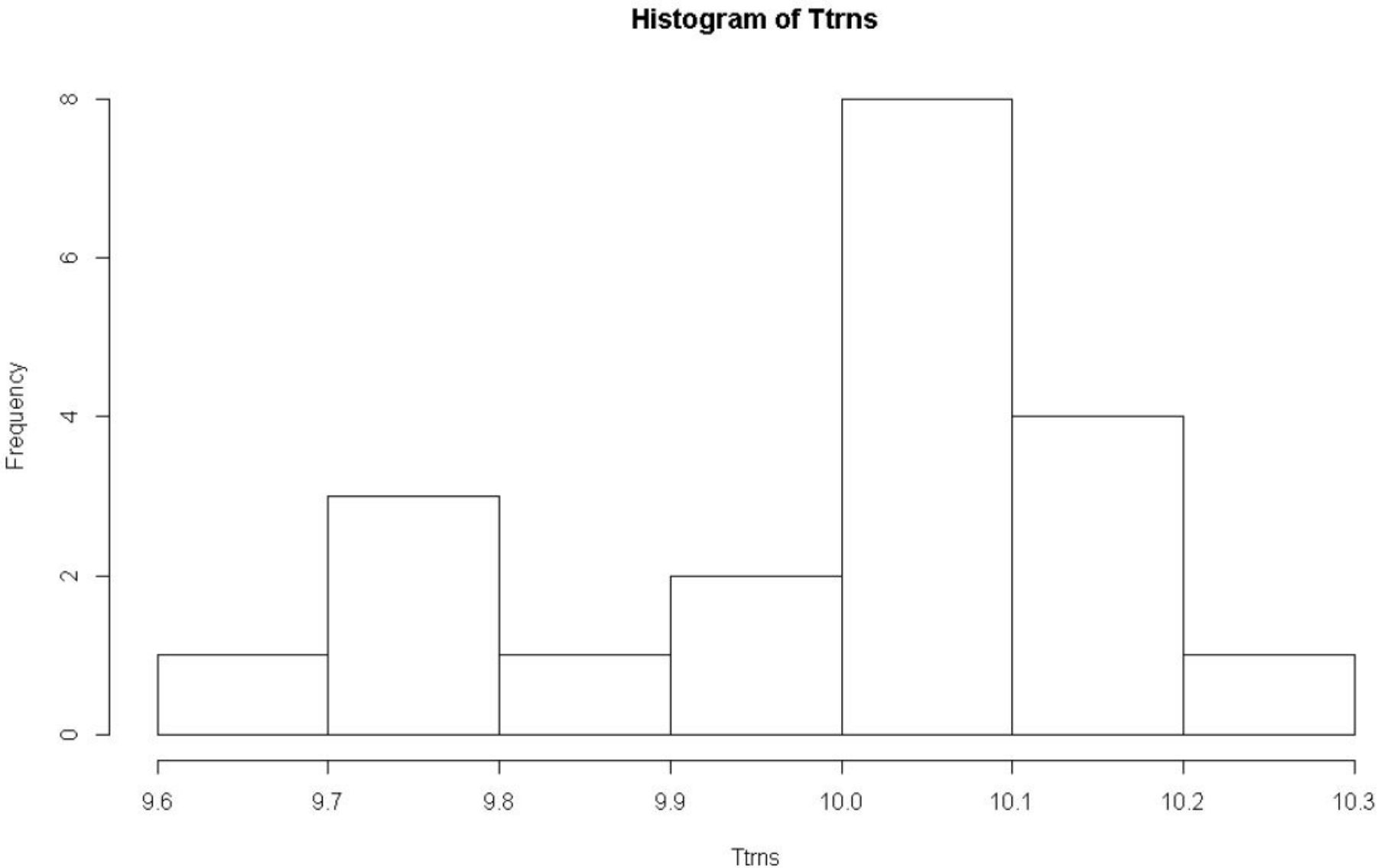
- IMPORTANT The null hypothesis of both the '**Anderson-Darling** and '**Shapiro-Wilk** tests is that the population, from which the sample is drawn, is normally distributed
- The alternative hypothesis is that the population, from which the sample is drawn, is not normally distributed.

Testing the Assumption of Normality

Implementing the Tests

- Let us use both tests to assess whether an example data set is normally distributed. (This data set is one that we are going to use in Lab Classes later on)
- Judging by this histogram on the next slides do you think the data set is normally distributed?

Testing the Assumption of Normality



Testing the Assumption of Normality

Using the Shapiro-Wilk Test

```
> shapiro.test(Ttrns)

        Shapiro-Wilk normality test

data:  Ttrns
W = 0.9188, p-value = 0.09394
```

Testing the Assumption of Normality

Using the Anderson Darling Test

```
> library(nortest)
> ad.test(Ttrns)

        Anderson-Darling normality test

data:  Ttrns
A = 0.6961, p-value = 0.0583
```

Testing the Assumption of Normality

Conclusion

- In both cases we fail to reject the null hypothesis that the data set is normally distributed.
- Just to clarify, we are not explicitly stating that the population, from which the sample is drawn, is Normally Distributed. (See next section.)

Limitations of Tests

- There are some important limitations to the usefulness of these tests.
- If you reject H_0 you can conclude that the population is not normally distributed, but if you don't reject H_0 then you only conclude that you failed to show the population is not normally distributed.
- In other words, you can prove the population is not normally distributed but you can't prove it is normally distributed.

Testing the Assumption of Normality

Limitations of Tests

- Rejecting H_0 means that the population is not normally distributed, but it doesn't tell you whether it is because it is a fat-tailed distribution, a thin-tailed distribution, a skewed distribution, or something else.
- Also - the tests are influenced by power. If you have a small sample the test may not have enough power to detect non-normality in the population.

Testing the Assumption of Normality

Q-Q plot

- The quantile-quantile (Q-Q) plot is an excellent way to see whether the data deviate from normal.
- The process used for creating a QQ plot involves determining what proportion of the 'observed' scores fall below any one score.
- Then the z- score that would fit that proportion if the data were normally distributed is calculated,

Testing the Assumption of Normality

Q-Q plot

- Finally that z- score that would cut off that proportion (the 'expected normal value') is translated back into the original metric to see what raw score that would be.
- A scatter plot is then created that shows the relationship between the actual 'observed' values and what those values would be 'expected' to be if the data were normally distributed.

Testing the Assumption of Normality

Interpreting Q-Q plots

- If the data is normally distributed then the circles on the resulting plot (each circle representing a score) will form a straight line.
- A trend line can be added to the plot to assist in determining whether or not this relationship is linear.

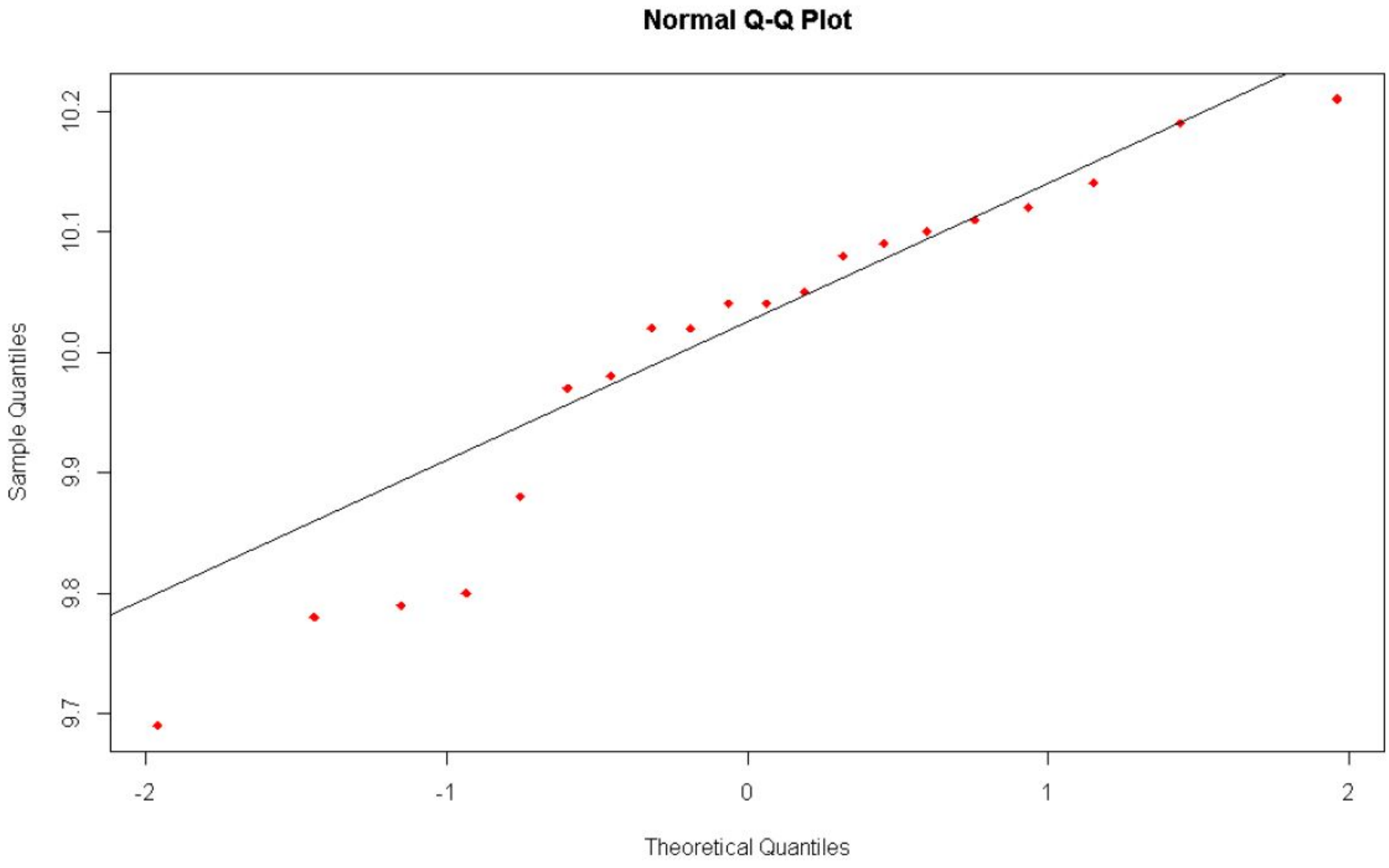
Testing the Assumption of Normality

Implenting Q-Q plots in R

```
> qqnorm(Ttrns)
> qqline(Ttrns)
```

How well do the covariates follow the trendline? Compare your conclusion to the p-values of the formal tests.

Testing the Assumption of Normality

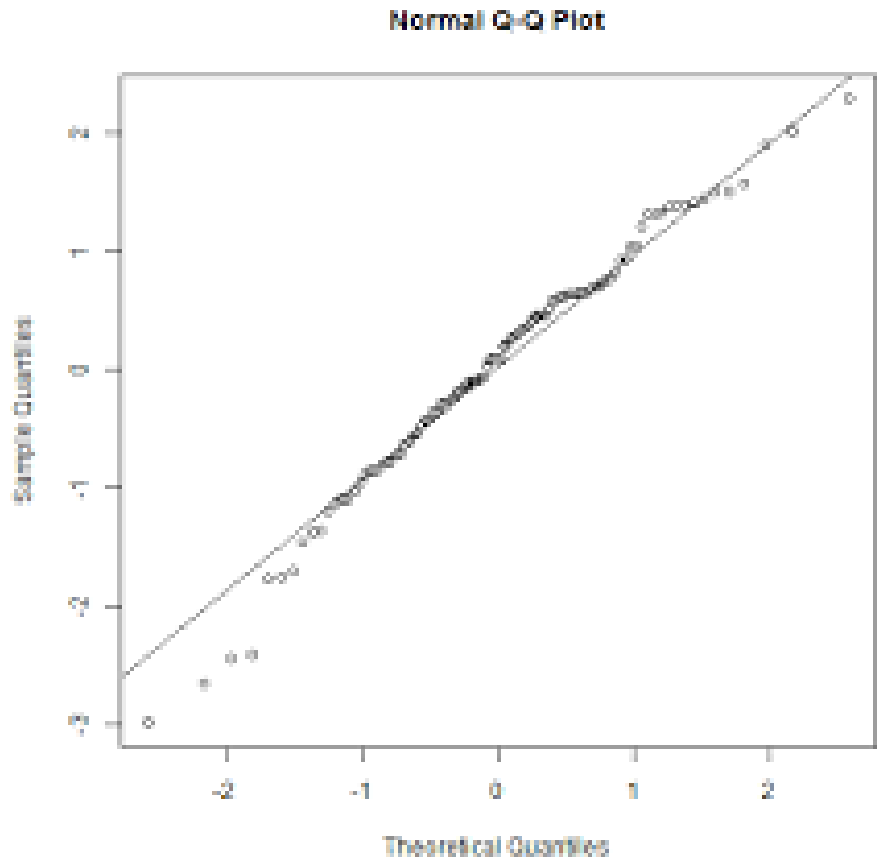


Interpreting Q-Q plots

- Most of the points follows the trend-line, but there are several observations that are fairly fair far away in the bottom left
- We will look at transformations next class

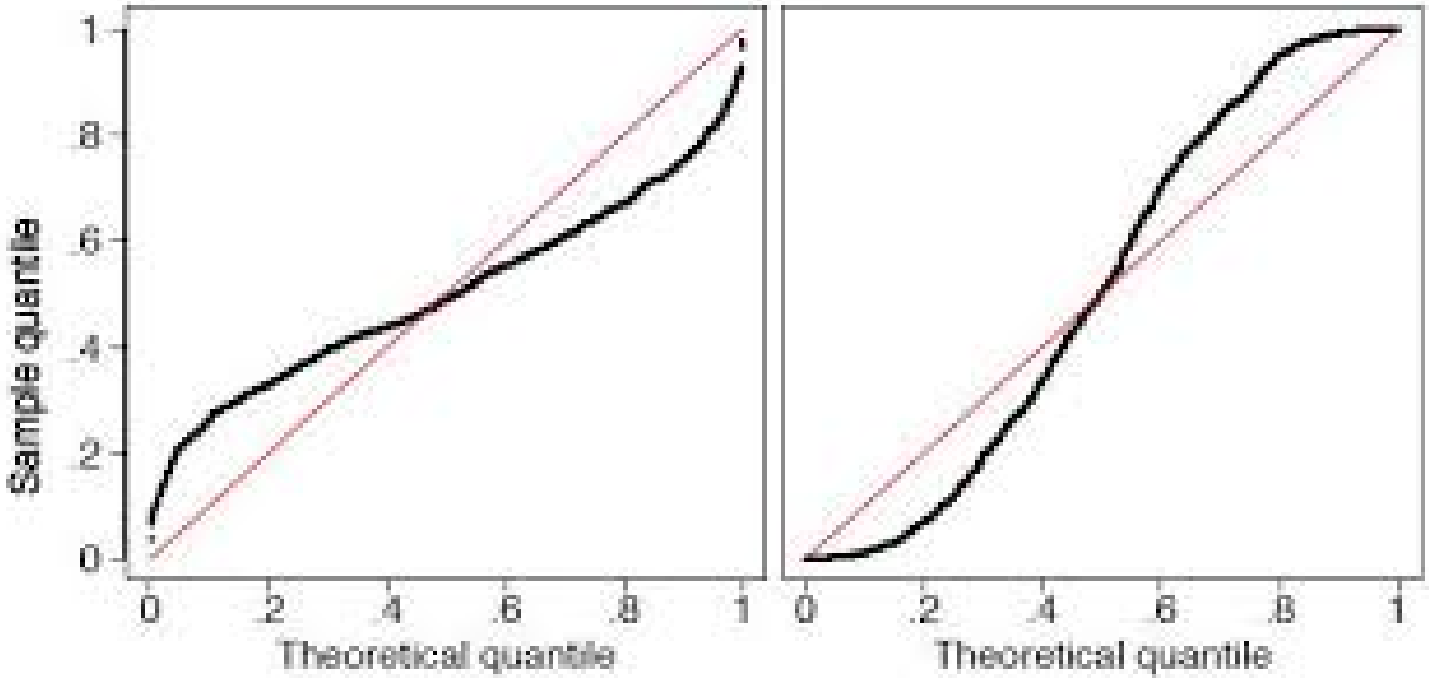
Interpreting Q-Q plots

- This is the QQplot for an unseen data set.
- Conclusion : Safe to assume normal distribution, despite some outliers



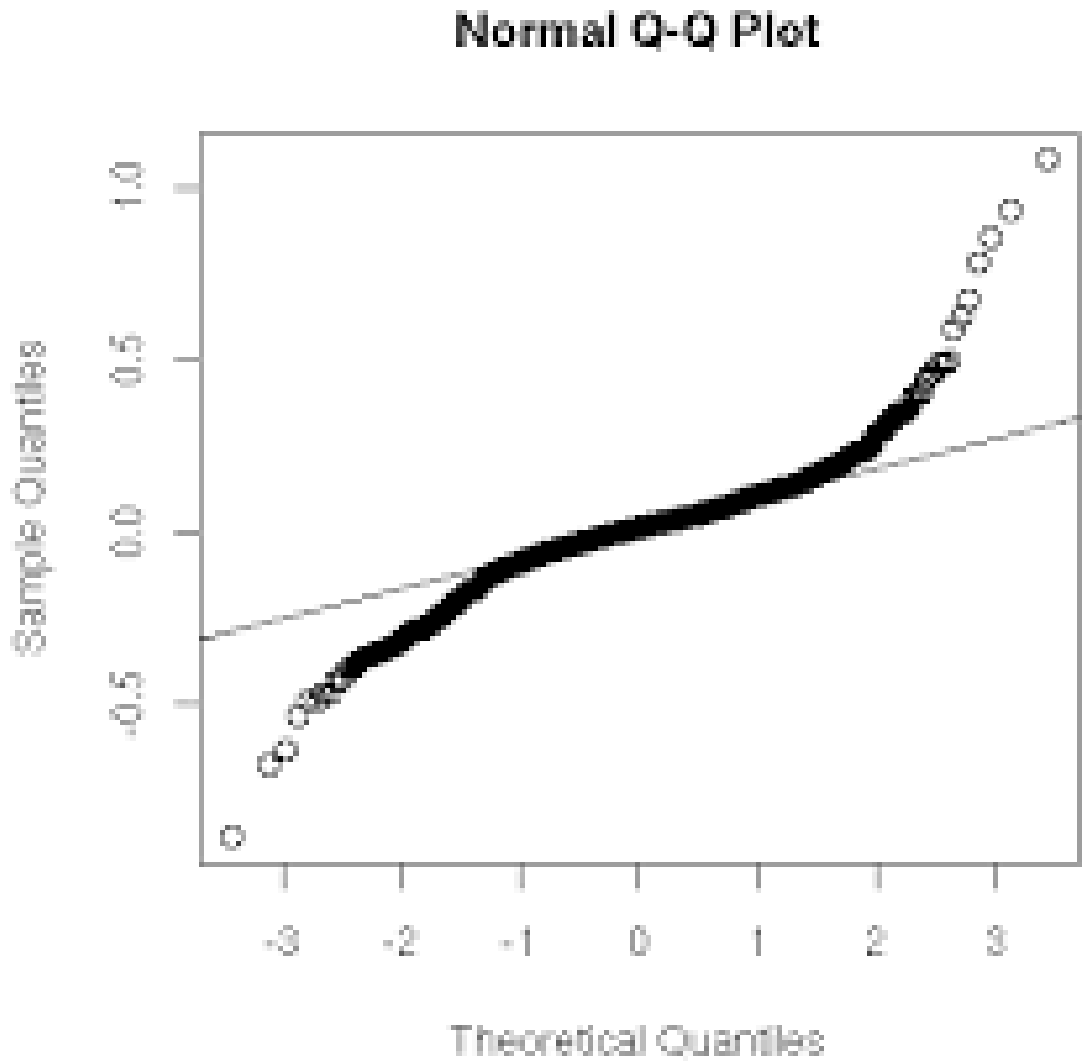
Interpreting Q-Q plots

Assumption of normal distribution clearly not valid in these cases



Interpreting Q-Q plots

Assumption of normal distribution clearly not valid in this case.



Review

- Know the null and alternative hypothesis for formal hypothesis tests for normality.
- Be able to interpret R code output.
- Discuss the limitations of these tests
- Know how to interpret Q-Q plots (in conjection with formal tests)
- (Some material will be added when we get to Statistical Process Control Section.)