Grubbs test

This test detects outliers from normal distributions. The tested data are the minimum and maximum values. The result is a probality that indicates that the data belongs to the core population. If the investigated sample has some other, especially assymmetric distribution (e.g. lognormal) then these tests give false results!

The test is based on the difference of the mean of the sample and the most extreme data considering the standard deviation (Grubbs, 1950, 1969; DIN 32645; DIN 38402).

The test can detect one outlier at a time with different probablities (see table below) from a data set with assumed normal distribution. If n>25 then the result is just a coarse approximation.

$$T_{\max} = \frac{x_n - X_{mean}}{s} \qquad T_{\min} = \frac{X_{mean} - X_1}{s}$$

where

X_1 or X_n = the suspected single outlier (max or min)
s = standard deviation of the whole data set
X_{aver} = mean

The reproducibility of a method for the determination of a pollutant in water was investigated by taking twelve samples from a single batch of water and determining the concentration of pollutant in each. The following results were obtained:

x=c(5.98, 8.80, 6.89, 8.49, 8.48, 7.47, 7.97, 5.94, 7.32, 6.64, 6.94, 3.51)

It is expected that from this sample a 95% confidence interval for the concentration of pollutant will be obtained.

There is a concern that the data may contain an outlier. Thus the following procedure has been performed on the data:

```
> grubbs.test(x, two.sided=T)

Grubbs test for one outlier
data: x
G = 2.4180, U = 0.4202, p-value = 0.04811
alternative hypothesis: lowest value 3.51 is an outlier
```

Describe what is the purpose of this procedure?

This is a statistical test called Grubbs test that allows for identification of a value in the data that appears unusually large or unusually small. Such a data point is called an outlier.

Write a conclusion that follows from it.

Since the p-value reported by the program is smaller than the standard 5% significance level, we conclude that there an outlier is detected in the data. The outlier indicated by the program is the lowest value in the data which is 3.51.

What should be done to the data in a consequence?

Since the data point 3.51 is an outlier, i.e. a value which can not be explained by the normal variability in the sample, it is recommended to omit (but not permanently delete) this point from the data and perform further analysis on the remaining datapoints.