1 Outliers

- In laboratory sciences, it is quite often the case that an outlier measurement is the result of faulty or unclean equipment.
- **Important:** Care must be take to assess that the measurement is an outlier, rather than an unusual result that is in fact genuine.
- It is good practice not to remove outliers from an overall analysis.
- However you may omit suspected outliers and run the analysis a second time, then present all of the obtained results, with and without the outliers.

There may also be an outlier , or multiple outliers, present in the data. There are several formal hypothesis tests to determine presence of an outlier. The main ones that we will use are

- Grubbs' Test
- Dixon Test

Remark : There are several variants of the Grubb's Test. Later We will look at three variants in particular.

1.1 Grubbs Test

- **(Important)**- This family of tests detects outliers from normal distributions.
- If the investigated sample has some other distribution (especially assymmetric distributions such as the lognormal distribution) then these tests give false results.
- The tested data are the minimum and maximum values, or both.
- The result is a probality that indicates that the data belongs to the core population.

1.2 Grubbs Test - First Variant

- The first test is used to detect if the sample dataset contains one outlier, statistically different than the other values.
- Important - Unless explicitly stated otherwise - Assume that this variant is being used. This is the only variant we will use with `R`.
- An alternative method is calculating ratio of variances of two datasets , the full dataset and the dataset without the outlier.

The test is based by calculating score of this outlier G (outlier minus mean and divided by standard deviation) and comparing it to appropriate critical values.

$$G = \frac{\max|Y_i - \bar{Y}|}{S}$$

- $\bar{Y}$ is the sample mean
- S is sample standard deviation

It is very similar (identical even) to the Student one sample $t-$test The critical value is based on the sample size n and the $t-$distribution quantiles.

- The test is based on the difference of the mean of the sample and the most extreme data considering the standard deviation.
- The test can detect one outlier at a time with different probablities from a data set with assumed normal distribution.
- If the sample size $n > 25$ then the result is just a coarse approximation.

1.3 Grubbs' Test - Second and Third Variants

- The second variant is used to check if **lowest and highest value are two outliers** on opposite tails of sample.
- It is based on calculation of ratio of range to standard deviation of the sample.

- The third variant is used to detect if dataset contains two outliers on the **same tail**.
- The Third variant calculates ratio of variance of full sample and sample without two extreme observations.

1.4 Review for Grubbs' Procedure

- Be able to describe the three variants of Grubbs's Test.
- For first variant, describe an algorithm used to perform the test. *No need to perform the test.*
- Be able to specify null and alternative hypothesis in each case.
  (For all three variants, the null hypothis is that there are no outliers present.)
  The alternative Hypotheses are described above also.
- Describe required assumptions and limitations.

1.5 Dixon Q Test

- The Dixon's Q test, or simply the Q test, is used for identification and rejection of outliers.
- **(Important)** - This test assumes normal distribution. Also this test should be used sparingly and never more than once in a data set.

To apply a Q test for suspicious data, arrange the data in order of increasing values and calculate Q as defined:

$$Q = \frac{\text{gap}}{\text{range}}$$

Where gap is the absolute difference between the outlier in question and the closest number to it.

- If $Q_{Test} > Q_{CV}$, where Q_{CV} is a critical value corresponding to the sample size and confidence level, then reject the questionable data point.
- Again, note that only one point may be rejected from a data set using a Q test.

1.6 Dixon Q Test: Example

Consider the data set:

0.189, 0.167, 0.187, 0.183, 0.186,
0.182, 0.181, 0.184, 0.181, 0.177

Now rearrange in increasing order:

0.167, 0.177, 0.181, 0.181, 0.182,
0.183, 0.184, 0.186, 0.187, 0.189

We hypothesize 0.167 is an outlier.
Calculate The Test Statistic Q_{Test}:

$$Q_{Test} = \frac{\text{gap}}{\text{range}}$$

$$Q_{Test} = \frac{0.177 - 0.167}{0.189 - 0.167} = 0.455.$$

Here: N is the sample size.

- With 10 observations and at 90% confidence, $Q_{Test} = 0.455 > 0.412 = Q_{CV}$, so we conclude 0.167 is an outlier.

N	Q_{crit} (CL:90%)	Q_{crit} (CL:95%)	Q_{crit} (CL:99%)
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

Figure 1:

- However, at 95% confidence, $Q_{Test} = 0.455 < 0.466 = Q_{CV}$ 0.167 is not considered an outlier.
- This means that for this example we can be 90% sure that 0.167 is an outlier, but we cannot be 95% sure.
- (Remark 95% confidence is equivalent to 5% signifificance)

1.7 Testing Outliers with `R`

- The tests for outliers come in a contributed package called **outliers**.
- In order to use it one has to download the package to the computer.
- It can be done for the command line by using `install.package("outliers")`, otherwise by using a convenient interface of the software (Rstudio).

Hypothesis for (main variant of) Grubbs' Test and the Dixon Test.

H_0 No Outlier Present in Data

H_1 There is an Outlier Present in Data

(**Important** - Only main variant of Grubbs' Test will be considered when using R)

1.8 Grubbs Test

```
x=c(0.403,0.410,0.401,0.380,
0.400,0.413,0.408)
grubbs.test(x)
# Grubbs test for one outlier
# data: x
# G = 1.4316, U = 0.0892
#       p-value = 0.09124
# alternative hypothesis:
lowest value 0.38 is an outlier
```

1.9 Dixon Test

```
x=c(0.403,0.410,0.401,0.380,
0.400,0.413,0.408)
dixon.test(x)

# Dixon test for outliers
#data: x
#Q = 0.7, p-value = 0.1721
#alternative hypothesis:
lowest value 0.38 is an outlier
```

1.10 Limitations

- Most of the test in the "outliers" package are designed for small samples.