Part X

TESTS ON OUTLIERS

Outliers are a phenomenon which inevitably occurs in the analysis of statistical data. Although there is no generally accepted unique definition of the term outlier they are commonly understood as observations which somehow do not fit into the data set. In Barnett and Lewis (1994, p. 7) we read 'We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data'. The same authors further proceed by putting emphasis on the fact that outliers are to be described as observations which are extreme as well as surprising for the observer. Whether an extreme value should be declared as an outlier depends on what we think about the main population from which we sample. Here a distinction has to be made from contaminants in the sense of observations originating from some other population, which might or might not be extreme with respect to the remaining observations and hence may or may not be outliers. How we generally decide to deal with the question of handling outliers is beyond the scope of this book, whether it is better to accommodate them by using robust methods, to detect them as they are of interest in themselves or just an undue influence on the applied analysis. Here we just present some well known discordancy tests from the toolkit of statistical methods to handle outliers. In tests of discordancy we aim at a decision on whether or not an extreme observation is to be seen as belonging to the main population or not. The main population is usually characterized by assuming some statistical distribution, which defines the null hypothesis. This assumption is sufficient to set up a statistical test. However, the choice of a reasonable test statistic as well as the assurance of desirable properties of the test depends on the existence of a meaningful alternative model. Often such models are formulated as an outlier-generating model (Barnett and Lewis 1994, p. 43).

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

15

Tests on outliers

In this chapter we present so-called discordancy tests (Barnett and Lewis 1994) for univariate samples. We consider outlier situations, where the basic sample follows some null distribution with a continuous distribution function. The general alternative hypothesis of this kind of test is that one (or more) observations are sampled from a different, maybe just shifted, distribution. Considering the ordered sample determines the extremes. Whether or not they are judged as outliers depends on their relation to the assumed null model. It is therefore common to use some spread/range test statistics which compare the extreme values with the center of the dataset or other extreme. In the tests introduced in this chapter the question usually is, if the lowest and/or highest value of the ordered sample is an outlier. However, if for instance the second highest value is an extreme as well a test might not detect the outlier as this value is masked by the other outlier. Most tests are prone to masking, which needs to be kept in mind. In Section 15.1 the null distribution is assumed to be a Gaussian distribution and in the Section 15.2 we deal with exponential and uniform distributions.

15.1 Outliers tests for Gaussian null distribution

In this section the assumed null distribution for the main population is the Gaussian distribution with unknown parameters. Most of the discussed tests can also be formulated for known parameters; please refer to Barnett and Lewis (1994) for details as well as further tests.

15.1.1 Grubbs' test

Description: Tests if there is an extreme outlier in a univariate Gaussian sample.

Assumptions:
- Data are measured on a metric scale.
- A univariate random sample $X_1 \ldots, X_n$ is given. $X_{(1)}, \ldots, X_{(n)}$ is the ordered sample.
- The null distribution is that of a Gaussian distribution. Mean and standard deviation are unknown.

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

Hypotheses: (A) $H_0 : X_1, \ldots, X_n$ belong to a Gaussian distribution
vs H_1 : Sample contains an extreme outlier
(B) $H_0 : X_1, \ldots, X_n$ belong to a Gaussian distribution
vs $H_1 : X_{(n)}$ is an upper outlier
(C) $H_0 : X_1, \ldots, X_n$ belong to a Gaussian distribution
vs $H_1 : X_{(1)}$ is a lower outlier

Test statistic: (A) $G = \max\left\{\frac{X_{(n)}-\overline{X}}{S}, \frac{\overline{X}-X_{(1)}}{S}\right\}$
(B) $G = \frac{X_{(n)}-\overline{X}}{S}$
(C) $G = \frac{\overline{X}-X_{(1)}}{S}$
with $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Test decision: Reject H_0 if for the observed value g of G
(A) $g > g_{n;1-\alpha/2}$
(B) $g > g_{n;1-\alpha}$
(C) $g > g_{n;1-\alpha}$
Critical values $g_{n;1-\alpha}$ can be found in Grubbs and Beck (1972).

p-values: Approximate formulas for p-values from Barnett and Lewis (1994):
(A) $p = 2n\left(1 - F_{n-2}\left(\sqrt{\frac{n(n-2)g^2}{(n-1)^2-ng^2}}\right)\right)$
(B) $p = n\left(1 - F_{n-2}\left(\sqrt{\frac{n(n-2)g^2}{(n-1)^2-ng^2}}\right)\right)$
(C) $p = n\left(1 - F_{n-2}\left(\sqrt{\frac{n(n-2)g^2}{(n-1)^2-ng^2}}\right)\right)$
where F_{n-2} denotes the cumulative distribution function of the t-distribution with $n-2$ degrees of freedom.
These p-values refer to $g \geq \sqrt{(n-1)(n-2 < 9/2n}$ otherwise they are upper bounds.

Annotations:

- This test is named after Frank Grubbs (1950, 1969). However, sources go back to William Thompson (1935). Thompson and Grubbs (in his earlier paper of 1950) used the sample standard deviation $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$. This leads to different critical values that can be found in Grubbs (1950). As noted by Pearson and Sekar (1936), these values can also be retrieved from a presentation of the test statistic as function of a t-distributed random variable with $n-2$ degrees of freedom.
- The test relates the difference between sample mean and maximum (or minimum) observation to the standard deviation.
- The test statistics can also be expressed as ratios of the sum of squares of deviations from mean values.

Let $SS^2 = \sum_{i=1}^{n} \left(X_{(i)} - \frac{1}{n} \sum_{i=1}^{n} X_{(i)} \right)^2$,
$SS_n^2 = \sum_{i=1}^{n-1} \left(X_{(i)} - \frac{1}{n-1} \sum_{i=1}^{n-1} X_{(i)} \right)^2$, and
$SS_1^2 = \sum_{i=2}^{n} \left(X_{(i)} - \frac{1}{n-1} \sum_{i=2}^{n} X_{(i)} \right)^2$ be the sum of squares from the complete random sample, without $X_{(n)}$ and without $X_{(1)}$. Then, the above test statistics are equivalent to:
(A) $G = \min \left(\frac{S_1^2}{SS}, \frac{SS_n^2}{S} \right)$, (B) $G = \frac{SS_n^2}{S}$, and (C) $G = \frac{SS_1^2}{S}$ (Barnett and Lewis 1994, p. 221).

Example: To test if there is an extreme outlier in a sample of height measurements of 20 students (dataset in Table A.6).

SAS code

```
* Calculate basic statistics, like maximum and minimum;
proc summary data=students;
 var height;
 output out=grupps n=n min=x_min max=x_max mean=x_mean
                                                 std=x_std;
run;

data grubbs_test;
 set grupps;
 format p_value_A p_value_B p_value_C pvalue.;

 * Calculate the test statistics;
 g_B=(x_max-x_mean)/x_std;
 g_C=(x_mean-x_min)/x_std;
 g_A=max(g_B,g_C);

 * Calculate p-values;
 t_A=sqrt((n*(n-2)*g_A**2)/((n-1)**2-n*g_A**2));
 t_B=sqrt((n*(n-2)*g_B**2)/((n-1)**2-n*g_B**2));
 t_C=sqrt((n*(n-2)*g_C**2)/((n-1)**2-n*g_C**2));

 p_value_A=2*n*(1-probt(t_A,n-2));
 p_value_B=n*(1-probt(t_B,n-2));
 p_value_C=n*(1-probt(t_C,n-2));
run;

* Output results;
proc print split='*' noobs;
 var g_A  p_value_A g_B p_value_B g_C p_value_C;
 label g_A='Test Statistic g_A*------------------'
       p_value_A='p-value A*---------'
       g_B='Test Statistic g_B*------------------'
    p_value_B='p-value B*---------'
    g_C='Test Statistic g_C*------------------'
    p_value_C='p-value C*---------';
```

```
 title 'Grubbs" Test';
run;
```

SAS output

```
          Grubbs' Test

Test Statistic g_A    p-value A
------------------    ---------
       2.39027          0.1962

Test Statistic g_B    p-value B
------------------    ---------
  1.94109              0.4285

Test Statistic g_C    p-value C
------------------    ---------
     2.39027           0.0981
```

Remarks:

- There is no SAS procedure available to calculate Grubbs' test directly.

R code

```
# Calculate basic statistics, like maximum and minimum
x_max<-max(students$height)
x_min<-min(students$height)
x_mean<-mean(students$height)
x_sd<-sd(students$height)
n<-length(students$height)

# Calculate the test statistics
g_B<-(x_max-x_mean)/x_sd
g_C<-(x_mean-x_min)/x_sd
g_A<-max(g_B,g_C)

# Calculate p-values
t_A<-sqrt((n*(n-2)*g_A^2)/((n-1)^2-n*g_A^2))
t_B<-sqrt((n*(n-2)*g_B^2)/((n-1)^2-n*g_B^2))
t_C<-sqrt((n*(n-2)*g_C^2)/((n-1)^2-n*g_C^2))

p_value_A<-2*n*(1-pt(t_A,n-2))
p_value_B<-n*(1-pt(t_B,n-2))
p_value_C<-n*(1-pt(t_C,n-2))

# Output results

"Two-sided test on extreme outlier"
g_A
p_value_A
```

```
"One-sided test on maximum is outlier"
g_B
p_value_B
"One-sided test on minimum is outlier"
g_C
p_value_C
```

R output

```
> "Two-sided test on extreme outlier"
[1] "Two-sided test on extreme outlier"
> g_A
[1] 2.390268
> p_value_A
[1] 0.1962342
> "One-sided test on maximum is outlier"
[1] "One-sided test on maximum is outlier"
> g_B
[1] 1.94109
> p_value_B
[1] 0.428505
> "One-sided test on minimum is outlier"
[1] "One-sided test on minimum is outlier"
> g_C
[1] 2.390268
> p_value_C
[1] 0.0981171
```

Remarks:

- There is no basic R function to calculate this test directly.

15.1.2 David–Hartley–Pearson test

Description: Tests if the minimum and the maximum are outliers in a random sample.

Assumptions:
- Data are measured on a metric scale.
- A univariate random sample $X_1 \ldots, X_n$ is given. $X_{(1)}, \ldots, X_{(n)}$ is the ordered sample.
- The null distribution is that of a Gaussian distribution. Mean and standard deviation are unknown.

Hypotheses: $H_0 : X_1, \ldots, X_n$ belong to a Gaussian distribution
vs $H_1 :$ $X_{(1)}$ and $X_{(n)}$ are outliers

Test statistic: $$Q = \frac{X_{(n)} - X_{(1)}}{S} \quad \text{with} \quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Test decision: Reject H_0 if for the observed value q of Q
$q > q_{n;1-\alpha}$
Critical values $q_{n;1-\alpha}$ can be found in Pearson and Hartley (1966).

p-values: Approximate formula for the p-value from Barnett and Lewis (1994):
$p = n(n-1)(1 - F_{n-2}\left(\sqrt{\frac{(n-2)q^2}{2n-2-q^2}}\right)$
where F_{n-2} denotes the cumulative distribution function of the F-distribution with $n-2$ degrees of freedom.
This p-value is for $g \geq \sqrt{\frac{3}{2}(n-1)}$, otherwise it is an upper bound.

Annotations:
- This test goes back to David *et al.* (1954).
- The test statistic is grounded on the relation of the range to the standard deviation.

Example: To test if there is a pair $X_{(1)}, X_{(n)}$ of extreme outliers in a sample of height measurements of 20 students (dataset in Table A.6).

SAS code

```
* Calculate basic statistics, like range
  and standard deviation;
proc summary data=students;
 var height;
 output out=dhp n=n range=x_range std=x_std;
run;

data dhp_test;
 set dhp;
 format p_value pvalue.;

 * Calculate the test statistic;
 q=x_range/x_std;

 * Calculate the p-value;
 t=sqrt(((n-2)*q**2)/(2*n-2-q**2));

 p_value=n*(n-1)*(1-probt(t,n-2));
run;

* Output results;
proc print split='*' noobs;
 var q p_value;
 label q='Test statistic*-------------*'
       p_value='p-value*---------*';
 title 'David-Hartley-Pearson Test';
run;
```

SAS output

```
   David-Hartley-Pearson Test

Test statistic    p-value
-------------    ---------
   4.33136        0.1047
```

Remarks:

- There is no SAS procedure available to calculate this test directly.

R code

```
# Calculate basic statistics, like maximum and minimum
x_max<-max(students$height)
x_min<-min(students$height)
x_sd<-sd(students$height)
n<-length(students$height)

# Calculate the test statistic
q<-(x_max-x_min)/x_sd

# Calculate the p-value
t<-sqrt(((n-2)*q^2)/(2*n-2-q^2))

p_value=n*(n-1)*(1-pt(t,n-2))

# Output results
"David-Hartley-Pearson Test"
q
p_value
```

R output

```
[1] "David-Hartley-Pearson Test"
> q
[1] 4.331358
> p_value
[1] 0.1046679
```

Remarks:

- There is no basic R function to calculate this test directly.

15.1.3 Dixon's tests

Description: Tests if there is an extreme outlier in a univariate Gaussian sample.

Assumptions:

- Data are measured on a metric scale.
- A univariate random sample $X_1 \ldots, X_n$ is given. $X_{(1)}, \ldots, X_{(n)}$ is the ordered sample.
- The null distribution is that of a Gaussian distribution with unknown standard deviation.

Hypotheses:

(A)–(D) $H_0 : X_1, \ldots, X_n$ belong to a Gaussian distribution
vs $H_1 : X_{(1)}$ is a lower outlier
(E)–(G) $H_0 : X_1, \ldots, X_n$ belong to a Gaussian distribution
vs $H_1 : X_{(n)}$ is an upper outlier

Test statistic:

$$\text{(A) } R^l_{10} = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}} \qquad \text{(E) } R^u_{10} = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}}$$

$$\text{(B) } R^l_{11} = \frac{X_{(2)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \qquad \text{(F) } R^u_{11} = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}}$$

$$\text{(C) } R^l_{20} = \frac{X_{(3)} - X_{(1)}}{X_{(n)} - X_{(1)}} \qquad \text{(G) } R^u_{20} = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(1)}}$$

$$\text{(D) } R^l_{22} = \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}} \qquad \text{(H) } R^u_{20} = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}}$$

where $X_{(1)} < X_{(2)} < \cdots < X_{(n-1)} < X_{(n)}$ are the order statistics.

Test decision:

Reject H_0 if for the observed value r^l_{ij} of R^l_{ij} or r^u_{ij} of R^u_{ij}
(A)–(D) $r^l_{ij} > r_{ij;n;\alpha}$
(E)–(F) $r^u_{ij} > r_{ij;n;\alpha}$
for $i, j = 1, 2$.
Critical values $r_{ij;n;\alpha}$ can be found in Dixon (1951).

p-values:

(A)–(D) $p = P(R^l_{ij} > r^l_{ij})$
(E)–(F) $p = P(R^u_{ij} > r^u_{ij})$

Annotations:

- These tests were introduced by Wilfrid Dixon (1950).
- The various R_{ij} differ in whether only the potential outlier or also other extremes are omitted in the calculation of the test statistic. Omitting further extreme observations avoids masking effects while at the same time giving away information.
- Dixon (1951) reported gives critical values for small sample sizes. Rorabacher (1991) gives extended tables for two-tailed tests and corrects some typographical errors in Dixon's tables.

Example: To test if there is an extreme outlier in a sample of height measurements of 20 students (dataset in Table A.6).

SAS code

```
* Calculate the necessary values;
proc summary data=students;
 var height;
 output out=dixon n=n
                  idgroup(max(height) out[3](height)=max)
                  idgroup(min(height) out[3](height)=min);
run;

* Output dataset includes following variables;
* max_1 = x_(n), max_2=x_(n-1), max_3=x_(n-2);
* min_1 = x_(1), min_2=x_(2), min_3=x_(3);

data dixon_test;
 set dixon;

 * Calculate the test statistics r10 and r22;
 r10=(min_2-min_1)/(max_1-min_1);
 r22=(min_3-min_1)/(max_3-min_1);

 * Calculate p-values;
 if (r10<=0.300)                then p_value_r10=">=0.0500";
 if (r10> 0.300 and r10<=0.391) then p_value_r10=" <0.0500";
 if (r10>=0.391)                then p_value_r10=" <0.0100";

 if (r22<=0.535)                then p_value_r22=">=0.0500";
 if (r22> 0.535 and r22<=0.450) then p_value_r22=" <0.0500";
 if (r22>=0.450)                then p_value_r22=" <0.0100";
run;

* Output results;
proc print split='*' noobs;
 var r10  p_value_r10 r22 p_value_r22;
 label r10='Test on lower outlier*
            avoiding x(1)*
            ----------------------------------'
       p_value_r10='p-value r10*-----------'

       r22='Test on lower outliers*
       avoiding x(2), x(n), X(n-1)*
       ----------------------------------'
       p_value_r22='p-value r22*-----------';
 title 'Dixon"s Tests';
run;
```

SAS output

```
                  Dixon's Tests

      Test on lower outlier
          avoiding x(1)              p-value r10
---------------------------------   -----------
             0.18519                  >=0.0500
```

```
      Test on lower outliers
   avoiding x(2), x(n), X(n-1)       p-value r22
---------------------------------   -----------
             0.43902                  >=0.0500
```

Remarks:

- There is no SAS procedure available to calculate these tests directly.
- In this example only the tests for hypotheses (A) and (D) are calculated. The other tests are performed accordingly.
- The p-values are approximated using the tables provided by Dixon (1951).
- To calculate the three highest/lowest values of the sample `proc summary` is used. The option of the command `output` to do that is `idgroup(max(time) out[3](time)=max)`. The command `max(time)` indicates that the maximum value of the variable *time* should be calculated. The command `out[3](time)=max` tells SAS to return the three highest values and name these values *max*. The highest value will be named *max_1* , the second highest *max_2* and the third highest *max_3*. A similar approach is used to calculate the three lowest values of the sample.

R code

```
# Calculate sample size
n<-length(students$height)

# Sort height
x<-sort(students$height)

# Calculate test statistics r10 and r22
r10<-(x[2]-x[1])/(x[n]-x[1])
r22<-(x[3]-x[1])/(x[n-2]-x[1])

# Calculate p-values
if (r10<=0.300)             p_value_r10<-">=0.0500"
if (r10>0.300 & r10<=0.391) p_value_r10<-" <0.0500"
if (r10>=0.391)             p_value_r10<-" <0.0100"

if (r22>=0.535)             p_value_r22<-">=0.0500"
if (r22>0.535 & r10<=0.450) p_value_r22<-" <0.0500"
if (r22>=0.450)             p_value_r22<-" <0.0100"

# Output results
"Test on lower outlier avoiding x(1)"
r10
p_value_r10

"Test on lower outliers avoiding x(2), x(n), x(n-1)"
r22
p_value_r22
```

R output

```
[1] "Test on lower outlier avoiding x(1)"
> r10
[1] 0.1851852
> p_value_r10
[1] ">=0.0500"
>
[1] "Test on lower outliers avoiding x(2), x(n), x(n-1)"
> r22
[1] 0.4390244
> p_value_r22
[1] ">=0.0500"
```

Remarks:

- There is no basic R function available to calculate these tests directly.
- In this example only the tests for hypotheses (A) and (D) are calculated. The other tests are performed accordingly.
- The p-values are approximated using the tables provided by Dixon (1951).

15.2 Outlier tests for other null distributions

In this section the assumed null distribution for the main population is the exponential or uniform distribution with unknown parameters. Most of the discussed tests can also be formulated for known parameters; please refer to Barnett and Lewis (1994) for details as well as further tests.

15.2.1 Test on outliers for exponential null distributions

Description: Tests if there is an extreme outlier in a univariate exponential sample.

Assumptions:
- Data are measured on a metric scale.
- A univariate random sample $X_1, \ldots, X_n$ is given. $X_{(1)}, \ldots, X_{(n)}$ is the ordered sample.
- The null distribution is that of an exponential distribution with unknown parameter.

Hypotheses:
(A) $H_0 : X_1, \ldots, X_n$ belong to an exponential distribution
vs $H_1 : X_{(n)}$ is an upper outlier
(B) $H_0 : X_1, \ldots, X_n$ belong to an exponential distribution
vs $H_1 : X_{(1)}$ is a lower outlier

Test statistic:

$$\text{(A)}\ E = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}}$$

$$\text{(B)}\ E = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}$$

Test decision: Reject H_0 if for the observed value e of E
(A) $e_A > e^u_{n;\alpha}$
(B) $e_B > e^l_{n;\alpha}$
Critical values $e^u_{n;\alpha}$ and $e^l_{n;\alpha}$ are given in Barnett and Lewis (1994, pp. 475–477) as well as in Likeš (1966).

p-values: Based on cumulative distribution functions of the test statistics from Barnett and Lewis (1994, p.199):
(A) $p = (n-1)(n-2)B((2-e)/(1-e), n-2)$
(B) $p = 1-(n-2)B((1+(n-2)e)/(1-e), n-2)$
where $B(a, b)$ is the beta function with parameters a and b.

Annotations:
- This test was proposed by Likeš (1966).
- This test relates the excess to the range and is of Dixon's type (see Test 15.1.3) but for exponential distributions.

Example: To test if there is an upper (lower) outlier in a sample of waiting times at a ticket machine (dataset in Table A.10).

SAS code

```
* Calculate the sample statistics;
proc summary data=waiting;
 var time;
 output out=expo n=n idgroup(max(time) out[2](time)=max)
                     idgroup(min(time) out[2](time)=min);
run;

* Output dataset includes following variables;
* max_1 = x_(n), max_2=x_(n-1);
* min_1 = x_(1), min_2=x_(2);

data expo_test;
 set expo;
 format p_value_B p_value_C pvalue.;

 * Calculate the test statistics;
 e_A=(max_1-max_2)/(max_1-min_1);
 e_B=(min_2-min_1)/(max_1-min_1);

 * Calculate p-values;
 p_value_A=(n-1)*(n-2)*beta((2-e_A)/(1-e_A),n-2);
```

```
 p_value_B=1-(n-2)*beta((1+(n-2)*e_B)/(1-e_B),n-2);
run;

* Output results;
proc print split='*' noobs;
 var e_A  p_value_A;
 label e_A='Test statistic e_A*------------------'
       p_value_A='p-value A*---------';
 title 'Test on an upper outlier in an exponential sample';
run;

proc print split='*' noobs;
 var   e_B p_value_B;
 label e_B='Test statistic e_B*-------------------'
       p_value_B='p-value B*---------';
 title 'Test on a lower outlier in an exponential sample';
run;
```

SAS output

```
Test on an upper outlier in an exponential sample

Test statistic e_A    p-value A
------------------    ---------
      0.16438          0.70481

Test on a lower outlier in an exponential sample

Test statistic e_B    p-value B
-------------------   ---------
     0.013699          0.2800
```

Remarks:

- There is no SAS procedure available to calculate this test directly.
- To calculate the highest, second highest, lowest and second lowest value of the sample `proc summary` is used. The option of the command `output` to do that is `idgroup(max(time) out[2](time)=max)`. The command `max(time)` indicates that the maximum value of the variable *time* should be calculated. The command `out[2](time)=max` tells SAS to return the two highest values and name these values *max*. The highest value will be named *max_1* and the second highest *max_2*. A similar approach is used to calculate the two lowest values of the sample.
- To calculate the p-values the `beta` function must be used.

R code

```
# Calculate sample size
n<-length(waiting$time)
```

```
# Sort waiting time
x<-sort(waiting$time)

# Calculate test statistic
e_A<-(x[n]-x[n-1])/(x[n]-x[1])
e_B<-(x[2]-x[1])/(x[n]-x[1])

# Calculate p-values
p_value_A<-(n-1)*(n-2)*beta((2-e_A)/(1-e_A),n-2)
p_value_B<-1-(n-2)*beta((1+(n-2)*e_B)/(1-e_B),n-2)

# Output results
"Test on an upper outlier in an exponential sample"
e_A
p_value_A

"Test on a lower outlier in exponential sample"
e_B
p_value_B
```

R output

```
[1] "Test on an upper outlier in an exponential sample"
> e_A
[1] 0.1643836
> p_value_A
[1] 0.704813
>
[1] "Test on a lower outlier in exponential sample"
> e_B
[1] 0.01369863
> p_value_B
[1] 0.2800093
```

Remarks:

- There is no basic R function to calculate this test directly.
- To calculate the p-values the `beta` function must be used.

15.2.2 Test on outliers for uniform null distributions

Description: Tests if there are h lower and k upper outliers in a univariate uniform sample.

Assumptions:

- Data are measured on a metric scale.
- A univariate random sample $X_1 \ldots, X_n$ is given. $X_{(1)}, \ldots, X_{(n)}$ is the ordered sample.

- The null distribution is that of a uniform distribution with unknown lower and upper bounds.

Hypotheses: (A) $H_0 : X_1, \ldots, X_n$ belong to a uniform distribution vs $H_1 : X_{(1)}, \ldots, X_{(h)}$ are lower outliers and $X_{(n-k)}, \ldots, X_{(k)}$ are upper outliers for given $h \geq 0$ and $k \geq 0$ with $h + k > 0$.

Test statistic:

$$U = \frac{X_{(n)} - X_{(n-k)} + X_{(h+1)} - X_1}{X_{(n-k)} - X_{(h+1)}} \times \frac{n - k - h - 1}{k + h}$$

Test decision: Reject H_0 if for the observed value u of U
$u > f_{1-\alpha;2(k+h),2(n-k-h-1)}$

p-values: $p = P(U \geq u)$

Annotations:

- The test statistic U follows an F- distribution with $2(k + h)$ and $2(n - k - h - 1)$ degrees of freedom (Barnett and Lewis 1994).
- $f_{1-\alpha;2(k+h),2(n-k-h-1)}$ is the $1 - \alpha$-quantile of the F-distribution with $2(k + h)$ and $2(n - k - h - 1)$ degrees of freedom.
- For more information on this test and modifications in the case of known upper or lower bounds see Barnett and Lewis (1994, p. 252).

Example: To test if there is an upper and a lower outlier in a sample of p-values of 20 t-tests (dataset in Table A.11).

SAS code

```
* Calculate the necessary values;
proc summary data=pvalues;
 var pvalue;
 output out=uniform n=n idgroup(max(pvalue)
                                out[2](pvalue)=max)
                     idgroup(min(pvalue)
                                out[2](pvalue)=min);
run;

* Output dataset includes following variables;
* max_1 = x_(n), max_2=x_(n-1);
* min_1 = x_(1), min_2=x_(2);

data uniform_test;
 set uniform;
 format p_value pvalue.;
```

```
 * Calculate the test statistic;
 u=((max_1-max_2+min_2-min_1)/(max_2-min_2))*((n-3)/2)   ;

 * Calculate p-values;
 p_value=1-probf(u,4,2*(n-3));

run;

* Output results;
proc print split='*' noobs;
 var u p_value;
 label u='Test statistic*------------'
       p_value='p-value B*---------';
 title 'Test on lower and upper outlier in a
                               univariate sample';
run;
```

SAS output

```
  Test on lower and upper outlier in a univariate sample

              Test statistic    p-value
              --------------   ---------
                 0.66878        0.6181
```

Remarks:

- There is no SAS procedure available to calculate this test directly.
- To calculate the highest, second highest, lowest and second lowest value of the sample `proc summary` is used. The option of the command `output` to do that is `idgroup(max(time) out[2](time)=max)`. The command `max(time)` indicates that the maximum value of the variable *time* should be calculated. The command `out[2](time)=max` tells SAS to return the two highest values and name these values *max*. The highest value will be named *max_1* and the second highest *max_2*. A similar approach is used to calculate the two lowest values of the sample.

R code

```
# Set parameter for testing of lower and upper outliers
h=1
k=1

# Read dataset and sort it
x<-sort(pvalues$pvalue)
n<-length(x)

# Calculate test statistic
```

```
u<-((x[n]-x[n-k]+x[h+1]-x[1])/
                        (x[n-k]-x[h+1]))*((n-k-h-1)/(k+h))

# Calculate p-value
p_value<-1-pf(u,2*(k+h),2*(n-k-h-1))

# Output results
"Test on lower an upper outlier in a univariate sample"
u
p_value
```

R output

```
[1] "Test on lower an upper outlier in a univariate sample"
> u
[1] 0.6687817
> p_value
[1] 0.6181188
```

Remarks:

- There is no basic R function to calculate this test directly.

References

Barnett V. and Lewis T. 1994 *Outliers in Statistical Data*., 3rd edn. John Wiley & Sons, Ltd.

David H.A., Hartley H.O. and Pearson E.S. 1954 The distribution of the ratio, in a single normal sample, of range and standard deviation. *Biometrika* **41**, 482–493.

Dixon W.J. 1950 Analysis of extreme values. *The Annals of Mathematical Statistics* **21**, 488–506.

Dixon W.J. 1951 Ratios involving extreme values. *The Annals of Mathematical Statistics* **22**, 68–78.

Grubbs F.E. 1950 Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics* **21**, 27–58.

Grubbs F.E. 1969 Procedures for detecting outlying observations in samples. *Technometrics* **11**, 1–21.

Grubbs F.E. and Beck G. 1972 Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics* **14**, 847–854.

Likeš J. 1966 Distribution of Dixon's statistics in the case of an exponential population. *Metrika* **11**, 46–54.

Pearson E.S. and Hartley H.O. 1966 *Biometrika Tables for Statisticians*, 3rd edn. Cambridge University Press.

Pearson E.S. and Sekar C.C. 1936 The efficiency of statistical tools and a criterion for the rejecting of outlying observations. *Biometrika* **28**, 308–320.

Rorabacher D.B. 1991 Statistical treatment for rejection of deviant values: Critical values of Dixon's 'Q' parameter and related subrange ratios at the 95% confidence level. *Analytical Chemistry* **63**, 139–146.

Thompson W.R. 1935 On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics* **6**, 214–219.