REGRESSION ANALYSIS AND
ANALYSIS OF VARIANCE

P. L. Davies

Eindhoven, February 2007

Reading List


Contents

1 Location and Scale in $\mathbb{R}$ 5
  1.1 Measures of location ................................. 5
    1.1.1 The mean ....................................... 5
    1.1.2 The median ..................................... 6
    1.1.3 The shorth ..................................... 7
    1.1.4 What to use? .................................... 8
  1.2 Measures of scale or dispersion ..................... 9
    1.2.1 The standard deviation ......................... 9
    1.2.2 The MAD ....................................... 9
    1.2.3 The length of the shorth ....................... 10
    1.2.4 The interquartile range IQR ................... 10
    1.2.5 What to use? ................................... 11
  1.3 Boxplots ............................................. 11
  1.4 Empirical distributions .............................. 13
    1.4.1 Definition ..................................... 13
    1.4.2 Empirical distribution function ............... 14
    1.4.3 Transforming data .............................. 14
    1.4.4 Statistics and functionals ................... 16
    1.4.5 Fisher consistency ............................ 17
  1.5 Equivariance considerations ......................... 18
    1.5.1 The group of affine transformations .......... 18
    1.5.2 Measures of location .......................... 18
    1.5.3 Measures of scale ............................. 19
  1.6 Outliers and breakdown points ....................... 19
    1.6.1 Outliers ...................................... 19
    1.6.2 Breakdown points and equivariance .......... 20
    1.6.3 Identifying outliers ........................... 24
  1.7 M-measures of location and scale ................... 30
    1.7.1 M-measures of location ....................... 30
    1.7.2 M-measures of scatter ....................... 35
    1.7.3 Simultaneous M-measures of location and scale 38
  1.8 Analytic properties of M–functionals ............... 40
    1.8.1 Local boundedness and breakdown ............. 42
  1.9 Redescending M–estimators I ....................... 47
  1.10 Differentiability and asymptotic normality .......... 51
  1.11 Redescending M–estimators II ..................... 62
  1.12 Confidence intervals .............................. 64
1 Location and Scale in \( \mathbb{R} \)

1.1 Measures of location

1.1.1 The mean

We start with measures of location for a data set \( \mathbf{x}_n = (x_1, \ldots, x_n) \) consisting of \( n \) real numbers. By far the most common measure of location is the arithmetic mean

\[
\text{mean}(\mathbf{x}_n) = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1.1}
\]

**Example 1.1.** 27 measurements of the amount of copper (milligrams per litre) in a sample of drinking water.

\[
\begin{align*}
2.16 & \quad 2.21 & \quad 2.15 & \quad 2.05 & \quad 2.06 & \quad 2.04 & \quad 1.90 & \quad 2.03 & \quad 2.06 \\
2.02 & \quad 2.06 & \quad 1.92 & \quad 2.08 & \quad 2.05 & \quad 1.88 & \quad 1.99 & \quad 2.01 & \quad 1.86 \\
1.70 & \quad 1.88 & \quad 1.99 & \quad 1.93 & \quad 2.20 & \quad 2.02 & \quad 1.92 & \quad 2.13 & \quad 2.13
\end{align*}
\]

The mean is given by

\[
\bar{x}_{27} = \frac{1}{27} (2.16 + \ldots + 2.13) = \frac{1}{27} 54.43 = 2.015926
\]

In \( \mathbb{R} \) the command is

\[
> \text{mean(copper)} \\
[1] 2.015926
\]

**Example 1.2.** Length of a therapy in days of 86 control patients after a suicide attempt.

\[
\begin{align*}
1 & \quad 1 & \quad 1 & \quad 5 & \quad 7 & \quad 8 & \quad 8 & \quad 13 & \quad 14 & \quad 14 & \quad 17 & \quad 18 & \quad 21 \\
21 & \quad 22 & \quad 25 & \quad 27 & \quad 30 & \quad 30 & \quad 31 & \quad 31 & \quad 32 & \quad 34 & \quad 35 & \quad 36 \\
37 & \quad 38 & \quad 39 & \quad 39 & \quad 40 & \quad 49 & \quad 49 & \quad 54 & \quad 56 & \quad 56 & \quad 62 & \quad 63 & \quad 65 \\
65 & \quad 67 & \quad 75 & \quad 76 & \quad 79 & \quad 82 & \quad 83 & \quad 84 & \quad 84 & \quad 90 & \quad 91 & \quad 92 \\
93 & \quad 93 & \quad 103 & \quad 103 & \quad 111 & \quad 112 & \quad 119 & \quad 122 & \quad 123 & \quad 126 & \quad 129 & \quad 134 & \quad 144 \\
147 & \quad 153 & \quad 163 & \quad 167 & \quad 175 & \quad 228 & \quad 231 & \quad 235 & \quad 242 & \quad 256 & \quad 256 & \quad 257 & \quad 311 \\
314 & \quad 322 & \quad 369 & \quad 415 & \quad 573 & \quad 609 & \quad 640 & \quad 737
\end{align*}
\]

The mean is given by

\[
\bar{x}_{86} = \frac{1}{86} (1 + \ldots + 737) = \frac{1}{86} 10520 = 122.3256.
\]
Example 1.3. Measurements taken by Charles Darwin on the differences in growth of cross- and self-fertilized plants. Heights of individual plants (in inches)

<table>
<thead>
<tr>
<th></th>
<th>Pot I</th>
<th></th>
<th>Pot II</th>
<th></th>
<th>Pot III</th>
<th></th>
<th>Pot IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cross</td>
<td>23.5</td>
<td>Cross</td>
<td>22.0</td>
<td>Cross</td>
<td>22.2</td>
<td>Cross</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>Self</td>
<td>17.4</td>
<td>Self</td>
<td>20.0</td>
<td>Self</td>
<td>18.6</td>
<td>Self</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.0</td>
<td></td>
<td>19.2</td>
<td></td>
<td>20.4</td>
<td></td>
<td>22.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>21.0</td>
<td></td>
<td>21.5</td>
<td></td>
<td>18.3</td>
<td></td>
<td>23.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17.4</td>
<td></td>
<td>20.0</td>
<td></td>
<td>18.6</td>
<td></td>
<td>12.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20.0</td>
<td></td>
<td>18.4</td>
<td></td>
<td>15.2</td>
<td></td>
<td>12.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20.0</td>
<td></td>
<td>18.6</td>
<td></td>
<td>16.5</td>
<td></td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18.0</td>
<td></td>
<td>18.0</td>
<td></td>
<td>18.0</td>
<td></td>
<td>18.0</td>
<td></td>
</tr>
</tbody>
</table>

The differences in heights are $23.5 - 17.4, \ldots, 12.0 - 18.9$.

6.1, $-8.4$, 1.0, 2.0, 0.8, 2.9, 3.6, 5.2, 1.8, 3.6, 7.0, 3.0, 9.3, 7.5, $-6.0$

The mean of the differences is 2.626667.

1.1.2 The median

The second most common measure of location is the median. To define the median we order the observations $(x_1, \ldots, x_n)$ according to their size

$$x(1) \leq x(2) \ldots \leq x(n) \quad (1.2)$$

and define the median to be the central observation of the ordered sample. If the size of the sample $n$ is an odd number $n = 2m+1$ then this is unambiguous as the central observation is $x_{(m+1)}$. If the number of observations is even $n = 2m$ then we take the mean of the two innermost observations of the ordered sample. This gives

$$\text{med}(x_n) = \begin{cases} x_{(m+1)} & \text{if } n = 2m+1 \\ (x_{(m)} + x_{(m+1)})/2 & \text{if } n = 2m. \end{cases} \quad (1.3)$$

Example 1.4. We take the first 5 measurements of Example 1.1 These are

$$2.16, 2.21, 2.15, 2.05, 2.06$$

so the ordered sample is

$$2.05, 2.06, 2.15, 2.16, 2.21.$$
The number of observation is odd, 5 = \( n = 2m + 1 \) with \( m = 2 \). The median is therefore

\[
\text{med}(2.16, 2.21, 2.15, 2.05, 2.06) = x_{(m+1)} = x_{(3)} = 2.15.
\]

We take the first 6 measurements of Example 1.1. These are

\[
2.16, 2.21, 2.15, 2.05, 2.06, 2.04
\]

so the ordered sample is

\[
2.04, 2.05, 2.06, 2.15, 2.16, 2.21.
\]

The number of observation is odd, 6 = \( n = 2m \) with \( m = 3 \). The median is therefore

\[
\text{med}(2.16, 2.21, 2.15, 2.05, 2.06, 2.04) = (x_{(m)} + x_{(m+1)})/2
\]

\[
= (x_{(3)} + x_{(4)})/2 = (2.06 + 2.15)/2
\]

\[
= 4.21/2 = 2.105.
\]

In R the command is

\[
> \text{median(copper)}
\]

[1] 2.03

1.1.3 The shorth

A less common measure of location is the so called shorth (shortest half) introduced by Tukey. We require some notation

(a) \([x]\) spoken “floor \( x \)” is the largest integer \( n \) such that \( n \leq x \). For example \([2.001] = 2\), \([5] = 5\), \([-2.3] = -3\).

(b) \([x]\) spoken “ceiling \( x \)” is the smallest integer \( n \) such that \( x \leq n \).

(c) For any number \( x \), \([x] + 1\) is the smallest integer \( n \) which is strictly greater than \( x \).

(d) The R command is floor:

\[
> \text{floor}(2.39)
\]

[1] 2
For a sample \((x_1, \ldots, x_n)\) of size \(n\) we consider all intervals \(I = [a, b]\) which contain at least \(\lfloor \frac{n}{2} \rfloor + 1\) of the observations, that is at least half the number of observations. Let \(I_0\) denote that interval which has the smallest length amongst all intervals containing at least \(\lfloor \frac{n}{2} \rfloor + 1\) observations. The shorth is defined to be the mean of those observations in the shortest interval. There is a degree of indeterminacy in this definition as there may be more than one such interval. One way of trying to break the indeterminacy is to take that interval with the largest number of observations. However this may not be sufficient as the next example shows.

**Example 1.5.** We consider the data of Example 1.1. The first step is to order the data.

\[
1.70 \ 1.86 \ 1.88 \ 1.88 \ 1.90 \ 1.92 \ 1.92 \ 1.93 \ 1.99 \\
1.99 \ 2.01 \ 2.02 \ 2.02 \ 2.03 \ 2.04 \ 2.05 \ 2.05 \ 2.06 \\
2.06 \ 2.06 \ 2.08 \ 2.13 \ 2.13 \ 2.15 \ 2.16 \ 2.20 \ 2.21
\]

The sample size is \(n = 27\) so \(\lfloor \frac{n}{2} \rfloor + 1 = \lfloor \frac{13.5}{2} \rfloor + 1 = 13 + 1 = 14\). We look for the shortest interval which contains at least 14 observations. The intervals \([1.70, 2.03], [1.86, 2.04], [1.88, 2.05], \ldots, [2.03, 2.21]\) all contain at least 14 observations. Their lengths are 0.33, 0.18, 0.17, 0.17, \ldots, 0.18. The smallest intervals have length 0.14, namely \([1.92, 2.06], [1.99, 2.13], [2.01, 2.15]\) and \([2.02, 2.16]\). The intervals \([1.92, 2.06]\) and \([1.99, 2.13]\) both contain 15 observations. Can you think of any way of breaking the indeterminacy? Is there always a means of obtaining a unique value for the shorth?

\[
1.92 \ 1.92 \ 1.93 \ 1.99 \ 1.99 \ 2.01 \ 2.02 \ 2.02 \\
2.03 \ 2.04 \ 2.05 \ 2.05 \ 2.06 \ 2.06 \ 2.06 \ 2.06
\]

The mean of these observations is 2.048.

**Exercise 1.6.** Write an R function to calculate the shortest half.

### 1.1.4 What to use?

We now have three measures of location and there are more to come. A natural question is which one to use. The natural answer is that it depends on the data and subject matter. What we can do is to analyse the behaviour of the different measures both analytically and practically when applied to different data sets. As an example we point out that the mean and median are always uniquely defined but not the shorth. This is a disadvantage of the shorth but it may have advantages which, for particular data sets, outweigh this. We shall examine the three measures in greater depth below but in the meantime the student can try them out on different data sets and form an own judgment.
1.2 Measures of scale or dispersion

1.2.1 The standard deviation

The most common measure of dispersion is the standard deviation. Unfortunately there are two competing definitions with no universal agreement. They are

\[
\text{sd}(x_n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \tag{1.4}
\]

and

\[
\text{sd}(x_n) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}. \tag{1.5}
\]

**WARNING**

In these notes I shall use definition (1.4). In \texttt{R} the definition (1.5) is used.

**Example 1.7.** We consider the data of Example 1.1. The standard deviation according to (1.4) is 0.1140260. If you use \texttt{R} you get

\[
> \text{sd(copper)}
\]

\[
[1] 0.1161981
\]

1.2.2 The MAD

The second most common measure of dispersion but one which is nevertheless not widely known is the MAD (Median Absolute Deviation) and is defined as follows. We form the absolute deviations from the median

\[
|x_1 - \text{med}(x_n)|, \ldots, |x_n - \text{med}(x_n)|.
\]

The MAD is now simply the median of these absolute deviations.

\[
\text{MAD}(x_n) = \text{median}(|x_1 - \text{med}(x_n)|, \ldots, |x_n - \text{med}(x_n)|). \tag{1.6}
\]

**Example 1.8.** We consider the data of Example 1.1. The median of the data is 2.03 which gives the following absolute deviations
The median of these values is 0.1 so \( \text{MAD}(x_n) = 0.1 \). If you use R you get

```r
> mad(copper)
[1] 0.14826
```

which is not the same. If instead you use

```r
> mad(copper,const=1)
[1] 0.1
```

which is the same.

I shall use \( \text{mad}(\cdot,\text{const}=1) \) in the lecture notes apart from the R command.

The factor 1.4826 in the R version of the MAD is chosen so that the MAD of a \( \mathcal{N}(\mu,\sigma^2) \)-random variable is precisely \( \sigma \). To see this we transfer the definition of the MAD to a random variable \( X \) as the median of \( |X - \text{median}(X)| \).

**Exercise 1.9.** Let \( X \) be a \( \mathcal{N}(\mu,\sigma^2) \)-random variable. Show that \( \text{MAD}(X) = 0.6744898\sigma \) so that \( 1.482602 \text{MAD}(X) = \sigma \).

### 1.2.3 The length of the shorth

The third measure of dispersion is the length of the shortest interval in the definition of the shorth. We denote this by \( \text{Ishorth} \).

**Example 1.10.** We consider the data of Example 1.1 and use the results of Example 1.5. The length of the shortest interval is 0.14 and hence \( \text{Ishorth}(x_n) = 0.14 \).

### 1.2.4 The interquartile range IQR

The fourth measure of dispersion is the so called interquartile range. For this we require the definitions of the first Q1 and third Q3 quartiles of a data set. Again unfortunately there is no agreed definition and the one I give here differs from them all. An argument in its favour will be given below. We order the data and define Q1 and Q3 as follows

\[
Q_1 = \frac{x_{\left\lceil \frac{n}{4} \right\rceil} + x_{\left\lfloor \frac{n}{4} \right\rfloor + 1}}{2}, \quad Q_3 = \frac{x_{\left\lceil \frac{3n}{4} \right\rceil} + x_{\left\lfloor \frac{3n}{4} \right\rfloor + 1}}{2},
\]

The interquartile range is then simply \( \text{IQR} = Q_3 - Q_1 \).
Example 1.11. We consider the data of Example 1.1. We have 27 observations and so
\[ \lceil n/4 \rceil = \lceil 27/4 \rceil = 6.75 = 7, \quad \lfloor n/4 \rfloor + 1 = \lfloor 6.75 \rfloor + 1 = 6 + 1 = 7. \]
This gives \( Q_1 = x_{(7)} = 1.82 \). Similarly \( Q_3 = x_{(21)} = 2.08 \) so that the interquartile range \( IQR = 2.08 - 1.82 = 0.16 \).

1.2.5 What to use?
We now have four measures of scale or dispersion and there are more to come. The comments made in Section 1.1.4 apply here as well. One particular problem with the MAD is that it is zero if more than 50% of the observations are the same. This can happen for small data sets with rounding but can also occur in counting data where there is for example a large atom at zero.

Exercise 1.12. Think of an improved version of the MAD which is less susceptible to the rounding of measurements.

1.3 Boxplots
Even for small samples it is difficult to obtain a general impression of the data sets, for example location, scale, outliers and skewness, from the raw numbers. A boxplot is a graphical representation of the data which in most cases can provide such information at a glance. Figure 1 shows from left to right the boxplots of the copper data (Example 1.1), the suicide data (Example 1.2) and the Darwin data (Example 1.3 with the differences taken over all plots). We now describe the construction of a boxplot with the proviso that there are different definitions in the literature.

The line through the box is the median. The lower edge of the box is the first quartile \( Q_1 \) and the upper edge is the third quantile \( Q_3 \). The “whiskers” extending from the bottom and top edges of the box start at the edge of the box and are extended to that observation which is furthest from the edge but at a distance of at most 1.5 IQR from the edge. Observations outside the whiskers are displayed individually.

Although the boxplot can be used for individual data sets it is most useful for comparing data sets.

Example 1.13. The data displayed in Table 1 were taken from an interlaboratory test and give the measurements returned by 14 laboratories of the amount of mercury (micrograms per litre) in a water sample. Each laboratory was required to give four measurements. The boxplot shown in Figure
Figure 1: The panels from left to right are the boxplots of the copper data (Example 1.1), the suicide data (Example 1.2) and the Darwin data (Example 1.3) respectively.

Table 1: The results of an interlaboratory test involving 14 laboratories each of which was required to return four measurements. The results give the quantity of mercury in a sample of drinking water. The units are micrograms per litre. The different precisions are because the laboratories were using different instruments and there was no instruction to round to a certain number of significant figures.
Figure 2: The boxplots of the interlaboratory test data of Example 1.13.

2 gives an immediate impression of the outlying laboratories and also the precision of the measurements they returned. The fact that the measurements have different numbers of significant figures is due to the returns of the individual laboratories. The R command is

\[ \text{> boxplot(data.frame(t(mercury)))} \]

**Exercise 1.14.** Interpret the boxplot of Figure 2.

### 1.4 Empirical distributions

#### 1.4.1 Definition

Given a real number \( x \) we denote by \( \delta_x \) the measure with unit mass at the point \( x \), the so called Dirac measure. For any subset \( B \) of \( \mathbb{R} \) we have

\[ \delta_x(B) = \{ x \in B \} \]

where we have used here the de Finetti notation of identifying a set with its indicator function. These means \( \{ x \in B \} = 1 \) if \( x \in B \) and \( \{ x \in B \} = 0 \) if
Given a sample \((x_1, \ldots, x_n)\) we define the corresponding empirical measure \(P_n\) by
\[
P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.
\] (1.8)

The empirical distribution \(P_n\) is a probability measure and for any subset \(B\) of \(\mathbb{R}\) we have
\[
P_n(B) = \frac{1}{n} \times \text{number of points } x_i \text{ in } B.
\]
Furthermore for any function \(f : \mathbb{R} \to \mathbb{R}\) we have
\[
\int f(u) \, dP_n(u) = \frac{1}{n} \sum_{i=1}^{n} f(x_i).
\] (1.9)

### 1.4.2 Empirical distribution function

Given data \(x_n = (x_1, \ldots, x_n)\) the empirical distribution function \(F_n\) is defined by
\[
F_n(x) = \mathbb{P}_n((-\infty, x]) = \frac{1}{n} \sum_{i=1}^{n} \{x_i \leq x\}
\] (1.10)

Figure 1.8 shows the empirical distribution functions of the data sets of the Examples 1.1 and 1.2.

### 1.4.3 Transforming data

First of all we recall some standard notation from measure and integration theory. It has the advantage of unifying definitions and concepts just as measure and integration theory allows in many situations a unified approach to discrete (say Poisson) and continuous (say Gaussian) probability models.

**Definition 1.15.** *Given a probability distribution* \(P\) on (the Borel sets of) \(\mathbb{R}\) and a (measurable) transformation \(f : \mathbb{R} \to \mathbb{R}\) we define the transformed distribution \(P^f\) by
\[
P^f(B) := P(\{x : f(x) \in B\}) = P(f^{-1}(B)).
\] (1.11)
Exercise 1.16. Show that $P_f$ is indeed a probability distribution.

In many situations we have a natural family $F$ of transformations of the sample space (here $\mathbb{R}$) which forms a group. Examples are the affine group on $\mathbb{R}$ which corresponds to changes in units and origin and will be investigated in the following section. In the case of directional data in two and three dimensions one is lead to the group of rotations or orthogonal transformations. We refer to Davies and Gather (2005) for further examples.

If we have a data set $x_n = (x_1, \ldots, x_n)$ and transform it using a function $f : \mathbb{R} \to \mathbb{R}$ we obtain the data set $x'_n = (x'_1, \ldots, x'_n)$ with $x'_i = f(x_i), i = 1, \ldots, n$. If the empirical distribution of $x_n$ is $P_n$ then the empirical dis-
tribution of $x'_n$ is given by

$$\mathbb{P}'_n(B) = \frac{1}{n} \sum_{i=1}^{n} \{x'_i \in B\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \{f(x_i) \in B\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \{x_i \in f^{-1}(B)\}$$

$$= \mathbb{P}_n(f^{-1}(B)) = \mathbb{P}^f_n(B)$$

so that $\mathbb{P}'_n = \mathbb{P}^f_n$.

### 1.4.4 Statistics and functionals

A statistic may be regarded simply as a function of the data so that the mean, median and standard deviation of a data set are all statistics. There is however another way of looking at many but not all statistics and that is to regard them as a function of a probability distribution. Functions on large spaces are usually called functionals and so this approach is a functional analytic approach. The advantage of doing this is that we are able to apply the ideas and techniques of functional analysis to statistical problems. If we denote the identity function on $\mathbb{R}$ by $\iota$, that is $\iota(x) = x$, then we see that for a data set $x_n$ with empirical distribution $\mathbb{P}_n$

$$\langle x_n \rangle = \int \iota(x) \, d\mathbb{P}_n(x) = \int x \, d\mathbb{P}_n(x) =: T_{\text{ave}}(\mathbb{P}_n).$$

If we set $\mathcal{P}' = \{P : \int |x| \, dP(x) < \infty\}$ then we define the mean functional by

$$T_{\text{ave}}(P) = \int x \, dP(x) \quad \text{for all} \quad P \in \mathcal{P}'. \quad (1.12)$$

The mean is not defined for all distributions but the median can be. We have

**Theorem 1.17.** For any probability distribution $P$ on the Borel sets of $\mathbb{R}$ the set

$$I_{\text{med}} = \{x : P((-\infty, x]) \geq 1/2, \ P([x, \infty)) \geq 1/2\}$$

is a non-empty closed interval.

**Definition 1.18.** For any distribution $P$ we define the median to be the mid-point of the interval $I_{\text{med}}$ of Theorem 1.17.
Exercise 1.19. Show that this definition agrees with the old one for a data set \( x_n \) in the following sense. If the data \( x_n \) have empirical distribution function \( \mathbb{P}_n \) then
\[
\text{med}(x_n) = P_{\text{med}}(\mathbb{P}_n)
\]
To define the MAD we proceed as follows. For any probability measure \( P \) let \( f \) denote the function \( f(x) = |x - \text{med}(P)| \). Then
\[
T_{\text{mad}}(P) := P_{\text{med}}(P^f).
\]
Exercise 1.20. Show that this definition agrees with the old one for a data set \( x_n \).

1.4.5 Fisher consistency

Suppose we have a parametric family \( \mathcal{P}_\Theta = \{ P_\theta : \theta = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k \} \) of distributions on \( \mathbb{R} \).

Definition 1.21. A statistical function \( T : \mathcal{P}' \to \mathbb{R} \) with \( \mathcal{P}_\Theta \subset \mathcal{P}' \) is called Fisher consistent for \( \theta_j \) if
\[
T(P_\theta) = \theta_j \quad \text{for all} \quad \theta \in \Theta.
\]
In some cases Fisher consistency does not hold but can be made to do so by a simple transformation.

Example 1.22. For the MAD and the \( \mathcal{N}(\mu, \sigma^2) \) distribution we have
\[
T_{\text{mad}}(\mathcal{N}(\mu, \sigma^2)) = 0.6745\sigma
\]
so that the MAD is not Fisher consistent for \( \sigma \). It can be made so by multiplying by 1.4826.

A Fisher consistent functional is not necessarily unbiased.

Example 1.23. Consider the family of binomial distributions \( \mathcal{B}(n, p) \) and define \( T(P) = \sqrt{T_{\text{ave}}(P)} \) for any distribution \( P \) concentrated on \( \mathbb{R}_+ = [0, \infty) \). Then \( T(\mathcal{B}(n, p)) = \sqrt{p} \) but there is no unbiased estimate for \( \sqrt{p} \).

Exercise 1.24. Prove the claim made in Example 1.23 that there is no unbiased estimator for \( \sqrt{p} \).

A Fisher consistent estimator which is in some sense continuous is however consistent. If \( T \) is Fisher consistent for \( \theta_j \) and \( (X(\theta)_n)_1^\infty \) are i.i.d. random variables with distribution \( P_\theta \) and empirical distributions \( (\mathbb{P}_n)_1^\infty \) then the Glivenko-Cantelli theorem tells us that \( \mathbb{P}_n \to P_\theta \). If now \( T \) is continuous at \( P_\theta \) then
\[
\lim_{n \to \infty} T(\mathbb{P}_n) = T(P_\theta) = \theta_j.
\]
1.5 Equivariance considerations

1.5.1 The group of affine transformations

A real data set \( x_n = (x_1, \ldots, x_n) \) often consists of readings made in certain units such as heights measured in centimeters. If the same heights are measured in inches \( x'_n = (x'_1, \ldots, x'_n) \) then to a first approximation we have

\[
x_i = 2.54x'_i, \quad i = 1, \ldots, n.
\]

The basis from which the heights are measured may also change. Instead of being measured from sea level or, in Holland, NAP they may be measured with respect to the floor of a house. If the units of measurement are the same the two samples will be related by

\[
x_i = x'_i + b, \quad i = 1, \ldots, n
\]

for some real number \( b \). In general we can consider a change of units and a change of basis which leads to transformations \( A : \mathbb{R} \to \mathbb{R} \) of the form

\[
A(x) = ax + b \quad \text{with } a, b \in \mathbb{R}, \ a \neq 0.
\]

Such a transformation is called an affine transformation and the set \( \mathcal{A} \) of affine transformations \( A : \mathbb{R} \to \mathbb{R} \) forms a group, the affine group. An affine transformation may be regarded as a change of units and origin.

**Exercise 1.25.** Show that \( \mathcal{A} \) is indeed a group.

1.5.2 Measures of location

If we apply the affine transformation \( A(x) = ax + b \) to the data set \( x_n = (x_1, \ldots, x_n) \) to give \( x'_n = (x'_1, \ldots, x'_n) \) with \( x'_i = A(x_i) \) then this relationship transfers to the means

\[
\text{mean}(x'_n) = \bar{x}'_n = A(\bar{x}_n) = a\bar{x}_n + b = a \text{mean}(x_n) + b. \tag{1.13}
\]

This relationship holds also for the median and the shorth. Indeed we may take it to be the defining property of a location measure. In other words a location measure \( T_L \) is one for which

\[
T_L(A(x_n)) = A(T_L(x_n)). \tag{1.14}
\]

Such measures are called location equivariant. For this to be mathematically correct the domain \( \mathcal{P}' \) of \( T_L \) must be invariant under the group of affine transformations, that is \( P^A \in \mathcal{P}' \) for all \( P \in \mathcal{P}' \) and for all affine transformations \( A \).

**Exercise 1.26.** Show that the Definition 1.18 is the only one which makes the median affine equivariant.
1.5.3 Measures of scale

Similar considerations apply to scale measures although the definition of equivariance is different. Indeed if we transform the data set \( x_n = (x_1, \ldots, x_n) \) using the affine transformation \( A(x) = ax + b \) to give the data set \( x'_n = (x'_1, \ldots, x'_n) \) then the relationship between the standard deviations is

\[
sd(x'_n) = |a|sd(x_n).
\]

This applies to the MAD, the length of the shorth and the interquartile range IQR. We call a measure \( T \) scale equivariant if

\[
T_S(A(x_n)) = |a|(T_L(x_n)), \quad A(x) = ax + b. \tag{1.15}
\]

Here again we require \( P^A \in \mathcal{P}' \) for all \( P \in \mathcal{P}' \) and for all affine transformations \( A \).

1.6 Outliers and breakdown points

1.6.1 Outliers

Outliers are important for two reasons. Firstly an outlier is an observation which differs markedly from the majority of the other observations. This may be simply due to a mistake, an incorrectly placed decimal point, or it may be because of particular circumstances which are themselves of interest. Secondly, an undetected outlier can completely invalidate a statistical analysis. In one–dimensional data outliers can be detected by visual inspection. Nevertheless there is an interest in automatic detection. There are national and international standards for analysing the results of interlaboratory tests (Example 1.13). The main problem when devising the statistical part of such a standard is how to deal with outliers which are very common in such data. Another reason for the interest in automatic methods is the large amount of data available which makes it impossible to have every data set visually inspected. The effect of outliers can be illustrated as follows.

Example 1.27. We consider the copper data of Example 1.1. The mean, standard deviation, median and MAD of the data are, as returned by R, respectively

\[
2.015926, \quad 0.1140260, \quad 2.03, \quad 0.1.
\]

Suppose we replace the last observation 2.13 by 21.3. The corresponding results are

\[
2.725926, \quad 3.644391, \quad 2.03, \quad 0.1.
\]
The standard deviation has been most effected followed by the mean. The values of the median and MAD have not changed. Even if we change the value of the second last observation also from 2.13 to 21.3 the median and the MAD do not change. The values returned by R for this altered sample are 3.435926, 5.053907, 2.03, 0.1.

We look at this phenomenon more closely in the next section.

1.6.2 Breakdown points and equivariance

We wish to measure the resilience of a statistical measure $T$ with respect to outliers. A simple and useful concept is that of the breakdown point introduced by Hampel (1968). The version we use is the so called finite sample breakdown point due to Donoho and Huber (1983). To define this we consider a sample $x_n = (x_1, \ldots, x_n)$ and to allow for outliers we consider samples which can be obtained from $x_n$ by replacing at most $k$ of the $x_i$ by other values. We denote a generic replacement sample by $x^k_n = (x^k_1, \ldots, x^k_n)$ and its empirical distribution by $P^k_n$. We have

$$\sum_{i=1}^n \{x^k_i \neq x_i\} \leq k.$$ 

We define the finite sample breakdown point of a statistical functional $T$ at the sample $x_n$ by

$$\epsilon^*(T, x_n) = \min \{k/n : \sup_{x^k_n} |T(x_n) - T(x^k_n)| = \infty\} \quad (1.16)$$

**Example 1.28.** The mean has a breakdown point $\epsilon^*(\text{mean}, x_n) = 1/n$ for any data $x_n$. We need only move one observation say $x_1$ to $\infty$ and it is clear that the mean of this sample will also tend to infinity.

**Example 1.29.** The median has a breakdown point $\epsilon^*(\text{med}, x_n) = [(n + 1)/2]/n$ for any data $x_n$. We consider the case that $n = 2m$ is even. If we move $k = m - 1$ observations all to $\infty$ or all to $-\infty$ (worst case) then the two central values of the ordered replacement sample will come from the original sample $x_n$. This implies

$$\sup_{x^k_n} |\text{med}(x^k_n) - \text{med}(x_n)| < \infty, \quad k \leq m - 1.$$ 

If we now move $m$ points all to $\infty$ then only one of the two central points will come from the original sample. This implies

$$\sup_{x^k_n} |\text{med}(x^k_n) - \text{med}(x_n)| = \infty, \quad k = m$$

20
and hence
\[ \varepsilon^*(\text{med}, x_n) = m/2m = 1/2 = \lfloor (n + 1)/2 \rfloor/n. \] (1.17)

**Exercise 1.30.** Show that (1.17) also holds if the sample size \( n = 2m + 1 \) is an odd number.

This is the highest possible finite sample breakdown point for measures of location as we shall now show.

**Theorem 1.31.** Let \( T_L \) be a measure of location, that is \( T \) satisfies (1.14). Then for any sample \( x_n \) we have
\[ \varepsilon^*(T_L, x_n) \leq \lfloor (n + 1)/2 \rfloor/n. \]

**Proof.** We restrict ourselves to the case where \( n = 2m \) is even. We consider two replacement samples \( x'_n \) and \( x''_n \) with \( k = m \) defined by
\[
\begin{align*}
x'_i &= x_i, \ i = 1, \ldots, m, \\
x'_{m+i} &= x_i + \lambda, \ i = 1, \ldots, m, \\
x''_i &= x_i - \lambda, \ i = 1, \ldots, m, \\
x''_{m+i} &= x_i, \ i = 1, \ldots, m.
\end{align*}
\]
The samples \( x'_n \) and \( x''_n \) are related by
\[ x''_i = x'_i - \lambda, \ i = 1, \ldots, n \]
and hence, as \( T_L \) is affinely equivariant \( |T_L(x''_n) - T_L(x'_n)| = \lambda \). Both are replacement samples with \( k = m \) and on letting \( \lambda \to \infty \) we see that not both \( |T_L(x'_n) - T_L(x_n)| \) and \( |T_L(x''_n) - T_L(x_n)| \) can remain bounded. Therefore
\[ \sup_{x''_n} |T_L(x''_n) - T_L(x_n)| = \infty \]
and hence
\[ \varepsilon^*(T_L, x_n) \leq m/2m = 1/2 = \lfloor (n + 1)/2 \rfloor/n. \] \( \Box \)

**Remark 1.32.** In Theorem 1.31 we did not make full use of the affine equivariance. All we required was the translation equivariance.

**Remark 1.33.** Theorem 1.31 connects the idea of breakdown point and equivariance which at first sight are unrelated. However if no conditions are placed upon a functional \( T \) then the breakdown point can be 1. A trivial way of attaining a breakdown point of 1 is simply to put \( T(x_n) = 0 \) for all \( x_n \). An apparently less trivial example is to restrict \( T \) to an interval \([a, b]\) for which it is is a priori known that the correct value must lie in this interval. It was argued in Davies and Gather (2005) that equivariance considerations are required in order that a non-trivial bound for the breakdown point exists.
We now apply the above ideas to measures of scale. One problem is the definition of breakdown. Clearly breakdown will be said to occur if arbitrarily large values are possible. As measures of scale are non-negative it is not possible to have the values tending to $-\infty$. The question is whether values tending to zero should be regarded as breakdown. This is often the case and zero values of scale due to rounding the measurements are one of the problems when evaluating interlaboratory tests. To take this into account we use the metric
\[ d(s_1, s_2) = |\log(s_1) - \log(s_2)| \] (1.18)
on $\mathbb{R}_+$.

The finite sample breakdown point for measures of scale is defined by
\[ \varepsilon^*(T_S, x_n) = \min \{ k/n : \sup_{x_n} |\log(T_S(x_{n}^k)) - \log(T_S(x_{n}))| = \infty \}. \] (1.19)

with the convention that $\varepsilon^*(T_S, x_n) = 0$ if $T_S(x_{n}) = 0$.

In order to state the theorem for measures of scale corresponding to Theorem 1.31 we require the following definition.

**Definition 1.34.** For any probability measure $P$ on $\mathbb{R}$ we define
\[ \Delta(P) = \max_{x} P(\{x\}) \]
and for a sample $x_n$ we sometimes write
\[ \Delta(x_n) := \Delta(P_n) = \max_{x} P_n(\{x\}) = \max_{x} |\{j : x_j = x\}|/n. \]

In other words $\Delta(P)$ is the largest mass concentrated at a single point. For a sample it is the largest proportion of observations with the same value.

**Theorem 1.35.** Let $T_S$ be a measure of scale, that is $T_S(A(x_n)) = |a|T_S(x_n)$ for $A(x) = ax + b$. Then for any sample $x_n$ we have
\[ \varepsilon^*(T_S, x_n) \leq \left\lfloor \frac{(n - \Delta(x_n)) + 1}{2} \right\rfloor /n. \]

**Proof.** Let $\Delta(x_n) = l/n$. We restrict ourselves to the case where $n - l = 2m$ is even. At least one value in the sample $x_n$ is present $l$ times and we take this to be $x_1, \ldots, x_l$. Let $A$ be an affine transformation with $A(x_1) = ax_1 + b = x_1$ and $a \neq 1$. We consider two replacement samples $x_n'$ and $x_n''$ with $k = m$ defined by
\[ x_i' = x_i, \hspace{1em} 1 \leq i \leq l, \hspace{1em} x_{l+i}' = x_{l+i}, \hspace{1em} 1 \leq i \leq m, \hspace{1em} x_{l+m+i}' = A^{-1}(x_{l+i}), \hspace{1em} 1 \leq i \leq m, \]
and
\[ x_i'' = A(x_i) = x_i, \quad 1 \leq i \leq l, \quad x_i'' = A(x_{l+i}), \quad 1 \leq i \leq m, \]
\[ x_{l+m+i}'' = x_{l+i}, \quad 1 \leq i \leq m. \]

The samples \( x'_n \) and \( x''_n \) are related by
\[ x''_i = A(x'_i), \quad i = 1, \ldots, n \]
and hence, as \( T_S \) is affinely equivariant \( T_S(x''_n) = |a|T_S(x'_n) \). Both are replacement samples with \( k = m \) and on letting \( a \to \infty \) or \( a \to 0 \) we see that not both \( \log(T_S(x'_n)) - \log(T_S(x_n)) \) and \( \log(T_S(x''_n)) - \log(T_S(x_n)) \) can remain bounded. Therefore
\[ \sup_{x''_n} |\log(T_S(x''_n)) - \log(T_S(x_n))| = \infty \]
and hence
\[ \varepsilon^*(T_S, x_n) \leq \frac{m}{n} = \frac{(n-l)/2}{n} = \left\lfloor \frac{n(1 - \Delta(P_n)) + 1}{2} \right\rfloor / n. \]

Exercise 1.36. Prove Theorem 1.35 for the case \( n - l = 2m + 1 \).

Example 1.37. The standard deviation functional \( \text{sd} \) has a breakdown point \( \varepsilon^*(\text{sd}, x_n) = 1/n \) for any data \( x_n \). We need only move one observation say \( x_1 \) to \( \infty \) and it is clear that the standard deviation of this sample will also tend to infinity.

Example 1.38. We calculate the finite sample breakdown point of the MAD. We note firstly that if \( \Delta(x_n) > 1/2 \) then \( \text{MAD}(x_n) = 0 \) and hence \( \varepsilon^*(\text{MAD}, x_n) = 0 \). This shows that the MAD does not attain the upper bound of Theorem 1.35. Suppose now that \( \Delta(x_n) = l/n \leq 1/2 \) and that \( n = 2m \) is even. If the value \( x_1 \) is taken on \( l \) times and we now alter \( m + 1 - l \) of the remaining sample values to \( x_1 \) we see that the MAD of this replacement sample is zero. This gives
\[ \varepsilon^*(\text{MAD}, x_n) \leq \frac{m}{n} = \frac{(n-l)/2}{n} = \left\lfloor \frac{n(1 - \Delta(P_n)) + 1}{2} \right\rfloor / n - \Delta(x_n). \]

This upper bound also holds for \( n \) odd. So far we have altered the sample so that the MAD of the replacement sample is zero. We can also try and make the MAD breakdown by becoming arbitrarily large. A short calculation
shows that in order to do this we must send off $\lfloor (n+1)/2 \rfloor$ observations to $\infty$ and hence
\[
\varepsilon^*(\text{MAD}, x_n) \leq \min\{\lfloor n/2 + 1 \rfloor/n - \Delta(x_n), [(n+1)/2]/n\}
\]
as $\Delta(x_n) \geq 1/n$. An examination of the arguments shows that the MAD can only breakdown under the conditions we used for the upper bound and hence
\[
\varepsilon^*(\text{MAD}, x_n) = \lfloor n/2 + 1 \rfloor/n - \Delta(x_n).
\]
As already mentioned this is less than the upper bound of Theorem 1.35. It is not a simple matter to define a scale functional which attains the upper bound but it can be done.

### 1.6.3 Identifying outliers

This section is based on Davies and Gather (1993). We take the copper data of Example 1.1 and we wish to identify any possible outliers. There are many ways of doing this and much misguided effort was put into devising statistical tests for the presence of outliers in one-dimensional data sets (see Barnett and Lewis (1994)). The problem is that identifying outliers does not fit well into standard statistical methodology. An optimal test for say one outlier may be devised but, as we shall see, it performs badly if more than one outlier is present. The Bayesian approach is no better requiring as it does the parameterization of the number of outliers and their distribution together with an appropriate prior. We describe the so called ‘two-sided discordancy test’ (Barnett and Lewis (1994, pages 223-224)). For a data set $x_n$ the test statistic is
\[
TS(x_n) = \max_i |x_i - \text{mean}(x_n)|/\text{sd}(x_n)
\]
with critical value $c(TS, n, \alpha)$ for a test of size $\alpha$. The distribution of the test statistic $TS$ is calculated under the normal distribution. The affine invariance of $TS$ means that it is sufficient to consider the standard normal distribution. Simulations show that for $n = 27$ and $\alpha = 0.05$ the critical value is $c(TS, 27, 0.05) = 2.84$ where we have used the R version of the standard deviation. The attained value for the copper data is 2.719 so that the test accepts the null hypothesis of no outliers. Suppose we now replace the largest observation 2.21 by 22.1. The value of the test statistic is now 5.002 and the observation 22.1 and only this observation is identified as an outlier. We now replace the second largest observation, 2.20 by 22. The value of the
test statistic is now 3.478 and the observations 22 and 22.2 and only these observations are identified as outliers. We continue and replace the third largest observation 2.16 by 21.6. The value of the test statistic is now 2.806 and we come to the rather remarkable conclusion that there are no outliers present. The source of the failure is easily identified. The values of the mean as we move from no outlier to three are 2.016, 2.752, 3.486 and 4.206. The corresponding values of the standard deviation are 0.116, 3.868, 5.352 and 6.376. We conclude that the mean and standard deviation are so influenced by the very outliers they are meant to detect that the test statistic drops below its critical value. The following exercises explain this behaviour.

**Exercise 1.39.** Let \( X \) be a \( N(0, 1) \) random variable. Show that for all \( x > 0 \)
\[
P(|X| \geq x) \leq \sqrt{\frac{2}{\pi}} \frac{1}{x} \exp(-x^2/2).
\]

**Exercise 1.40.** Let \( X_i, i = 1, \ldots, n \) be i.i.d. \( N(0, 1) \) random variables. Show that
\[
\lim_{n \to \infty} P(\max_{1 \leq i \leq n} |X_i| \leq \sqrt{\tau \log(n)}) = \begin{cases} 0, & \tau < 2, \\ 1, & \tau \geq 2. \end{cases}
\]

**Exercise 1.41.** Show that the critical values \( c(TS, n, \alpha) \) of the two-sided discordancy test satisfy
\[
\lim_{n \to \infty} \frac{c(TS, n, \alpha)}{\sqrt{2 \log(n)}} = 1
\]
for any \( \alpha, 0 < \alpha < 1 \).

**Exercise 1.42.** Show that for large \( n \) the two-sided discordancy test can fail to detect \( n/(\tau \log(n)) \) arbitrary large outliers for any \( \tau < 2 \).

The failure of the two-sided discordancy test to identify large outliers is related to the low finite sample breakdown points of \( 1/n \) of the mean and standard deviation. The remedy is simply to replace the mean and standard deviations by other measures with a high finite sample breakdown point. The obvious candidates are the median and the MAD. We replace the \( TS \) of (1.20) by
\[
TH(x_n) = \max_i |x_i - \text{med}(x_n)|/\text{MAD}(x_n).
\]
(1.21)
The ‘H’ in (1.21) stands for Hampel who proposed cleaning data by removing all observations \( x_i \) for which \( |x_i - \text{med}(x_n)| \geq 5.2\text{MAD}(x_n) \) and then
calculating the mean of the remainder (see Hampel 1985). Again we calculate the cut-off value for $TH$ using the normal distribution. Repeating the simulations as for $TS$ we obtain a value of $c(TH, 27, 0.05) = 5.78$ where we have used the $const = 1$ version of the MAD in R. The values for none, one, two and three outliers as before are 3.3, 200.7, 200.7 and 200.7 so that all outliers are easily identified. How far can we go? If we replace the largest 13 observations by ten times their actual value the value of $TH$ is 60.82 and all outliers are correctly identified. However if we replace the largest 14 observations by ten times their actual value the value of $TH$ is 10.94 and the smallest 13 observations are now identified as outliers. The sample is now

$$
\begin{array}{cccccccccccc}
1.70 & 1.86 & 1.88 & 1.88 & 1.90 & 1.92 & 1.92 & 1.93 & 1.99 \\
1.99 & 2.01 & 2.02 & 2.02 & 2.03 & 20.4 & 20.5 & 20.5 & 20.6 \\
\end{array}
$$

In the absence of further information this is perfectly reasonable as the large observations are now in the majority. The fact that the switch over takes place when about half the observations are outliers is related to the finite sample breakdown point of the median and the MAD, namely $1/2$.

The cut-off value $c(TH, 27, 0.05) = 5.78$ was obtained by simulations. Simulations are a very powerful technique for evaluating integrals in high dimensions which is what we are essentially doing but they are computationally expensive. Furthermore it is not possible to simulate the value of $c(TH, n, \alpha)$ for every $n$ and $\alpha$ and thought must be given to how best to calculate the $c(TH, n, \alpha)$ at least approximately. The following method is often appropriate.

1. Determine a simple asymptotic formula which applies for large $n$.
2. Use simulations to obtain a sample size $n_1$ for which the asymptotic formula (1) is sufficiently accurate for $n \geq n_1$.
3. Choose a $n_0 < n_1$ such that it is possible to approximate the values for $n_0 \leq n \leq n_1$ by a simple curve to be determined empirically.
4. Simulate the values for $n \leq n_0$ and store them in a table.

We restrict ourselves to say 3 values of $\alpha$, say 0.1, 0.05 and 0.01. For small values of $n$, say 3:19, that is $n_0 = 19$, we calculate the simulated values and store them. These are given in Table 2 and Figure 4 shows a plot of the values of $c(TH, n, 0.1)$ for $n = 3(1)19$.

**Exercise 1.43.** Explain the odd-even sample size behaviour of $c(TH, n, 0.1)$ evident in Figure 4.
For large values of $n$ we try and determine the asymptotic behaviour of $c(TH, n, \alpha)$ and then consider the differences to the actual behaviour for large $n$. With luck these will tend to zero and exhibit a regular behaviour which allows them to be approximated by a simple function of $n$. This can be calculated using the linear regression module of R and we finally obtain a simple formula for $c(TH, n, \alpha)$. The first step is to calculate the asymptotic behaviour of $c(TH, n, \alpha)$. If the sample $X_n$ is i.i.d. $\mathcal{N}(0, 1)$ then $\text{med}(X_n) \approx 0$ and $\text{MAD}(X_n) \approx 0.67449$. This gives

$$TH(X_n) \approx \max_i |X_i|/0.67449.$$  

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\alpha = 0.1$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>15.40</td>
<td>32.54</td>
<td>158.97</td>
</tr>
<tr>
<td>4</td>
<td>7.11</td>
<td>10.19</td>
<td>21.82</td>
</tr>
<tr>
<td>5</td>
<td>7.84</td>
<td>11.58</td>
<td>26.42</td>
</tr>
<tr>
<td>6</td>
<td>6.18</td>
<td>8.07</td>
<td>13.53</td>
</tr>
<tr>
<td>7</td>
<td>6.78</td>
<td>9.17</td>
<td>16.28</td>
</tr>
<tr>
<td>8</td>
<td>5.84</td>
<td>7.23</td>
<td>11.98</td>
</tr>
<tr>
<td>9</td>
<td>5.99</td>
<td>7.45</td>
<td>12.16</td>
</tr>
<tr>
<td>10</td>
<td>5.58</td>
<td>6.77</td>
<td>10.35</td>
</tr>
<tr>
<td>11</td>
<td>5.68</td>
<td>6.92</td>
<td>11.06</td>
</tr>
<tr>
<td>12</td>
<td>5.37</td>
<td>6.47</td>
<td>9.34</td>
</tr>
<tr>
<td>13</td>
<td>5.52</td>
<td>6.58</td>
<td>9.38</td>
</tr>
<tr>
<td>14</td>
<td>5.28</td>
<td>6.16</td>
<td>8.46</td>
</tr>
<tr>
<td>15</td>
<td>5.44</td>
<td>6.40</td>
<td>8.66</td>
</tr>
<tr>
<td>16</td>
<td>5.25</td>
<td>6.12</td>
<td>8.17</td>
</tr>
<tr>
<td>17</td>
<td>5.36</td>
<td>6.20</td>
<td>8.42</td>
</tr>
<tr>
<td>18</td>
<td>5.21</td>
<td>6.01</td>
<td>7.91</td>
</tr>
<tr>
<td>19</td>
<td>5.26</td>
<td>6.02</td>
<td>8.00</td>
</tr>
</tbody>
</table>

Table 2: The cut-off values $c(TH, n, \alpha)$ of $TH$ of (1.21) for $n = 3(1)19$ and for $\alpha = 0.1, 0.05$ and $0.01$.

Now if $x$ is chosen so that

$$1 - \alpha = P(\max_i |X_i| \leq x) = P(|X_1| \leq x)^n$$

we see that

$$2P(X_1 \leq x) - 1 = P(|X_1| \leq x) = (1 - \alpha)^{1/n}$$

so that

$$c(TH, n, \alpha) \approx \Phi^{-1}\left((1 + (1 - \alpha)^{1/n})/2\right)/0.67449. \quad (1.22)$$
Figure 4: The values of $c(TH, n, 0.1)$ of $TH$ of (1.21) for $n = 3(1)19$. This is the second column of Table 2.
This is the asymptotic formula we wished to derive. On putting
\[
\gamma(TH, n, \alpha) = 0.67449 \frac{c(TH, n, \alpha)}{\Phi^{-1}((1 + (1 - \alpha)^{1/n})/2)}
\]
we expect that \(\lim_{n \to \infty} \gamma(TH, n, \alpha) = 1\). In fact due to the variability of the median and MAD we expect that for finite \(n\) the value of \(\gamma(TH, n, \alpha)\) should exceed one. It seems reasonable therefore that \(\lim_{\infty} \log(\gamma(TH, n, \alpha) - 1) = 0\). The R command for \(\Phi^{-1}(p)\) is
\[
> \text{qnorm}(p)
\]
so that \(\Phi^{-1}((1 + (1 - \alpha)^{1/n})/2)\) can be easily calculated. The next question is which values of \(n\) to use. As we expect the dependence to be in \(\log(n)\) (experience!) we choose the values of \(n\) such that \(\log(n)\) is approximately linear. In the following we put \(n = 20 \cdot 1.25^{0(1)19}\). To test whether this is satisfactory a small simulation with 1000 simulations for each value of \(n\) can be carried out. The following results are based on 50000 simulations for each value of \(n\). The upper panel of Figure 5 shows the values of \(\gamma(TH, n, 0.1)\) plotted against those of \(\log(n)\). As expected they tend to one as \(n\) increases. The lower panel of Figure 5 shows the values of \(\log(\gamma(TH, n, 0.1) - 1)\) plotted against those of \(\log(n)\). As expected they tend to \(-\infty\) as \(n \to \infty\). Luckily the dependence on \(\log(n)\) is clearly almost linear. The line in the lower panel show the regression line
\[
\log(\gamma(TH, n, 0.1) - 1) = 1.261 - 0.867 \log(n).
\]
Putting all this together we get the following approximation for \(c(TH, n, 0.1)\) where we simplify the numerical values as the exact ones are not very important. We are only trying to identify outliers and this does not depend on the last decimal point. Bearing this in mind we have the approximation for \(n \geq 20\)
\[
c(TH, n, 0.1) \approx 1.48(1 + 4.65 n^{-0.88})\Phi^{-1}((1 + 0.95^{1/n})/2).
\]
The corresponding results for \(c(TH, n, 0.05)\) and \(c(TH, n, 0.01)\) for \(n \geq 20\) are
\[
c(TH, n, 0.05) \approx 1.48(1 + 8.10 n^{-0.93})\Phi^{-1}((1 + 0.99^{1/n})/2),
\]
\[
c(TH, n, 0.01) \approx 1.48(1 + 8.10 n^{-0.93})\Phi^{-1}((1 + 0.99^{1/n})/2).
\]

**Exercise 1.44.** Repeat the simulations and approximations for the outlier identifier based on the shorth and the length of the shortest half-interval:
\[
TSH(x_n) = \max_i |x_i - \text{shrth}(x_n)|/\text{lshrth}(x_n).
\]
Figure 5: The upper panel shows the values of $\gamma(TH,n,0.1)$ of (1.23) plotted against the values of $\log(n)$. The lower panel shows the values of $\log(\gamma(TH,n,0.1) - 1)$ plotted against the values of $\log(n)$ together with the approximating linear regression.

1.7 M-measures of location and scale

1.7.1 M-measures of location

So far we only have three measures of location, the mean the median and the shorth. We now introduce a whole class of location measures which we motivate by showing that the mean and the median are special cases. Let $\mathbb{P}_n$ denote the empirical measure of the sample $x_n$. Then if we use (1.9) with $f(x) = \iota(x) = x$ we see that

$$\text{mean}(x_n) = \int \iota(x) \, d\mathbb{P}_n(x) = \int x \, d\mathbb{P}_n(x).$$

From this it follows that

$$\int \iota(x - \text{mean}(x_n)) \, d\mathbb{P}_n(x) = 0$$

(1.27)
and in fact this defines the mean. By choosing a different function namely the sign function $\text{sgn}$ defined by

$$
\text{sgn}(x) = \begin{cases} 
-1, & x < 0, \\
0, & x = 0, \\
1, & x > 0,
\end{cases}
$$

(1.28)

we see that the median satisfies

$$
\int \text{sgn}(x - \text{med}(x_n)) \, d\mathbb{P}_n(x) = 0.
$$

(1.29)

This leads to a general class of statistical functionals $T_L$, known as M–functionals defined by

$$
\frac{1}{n} \sum_{i=1}^{n} \psi(x - T_L(\mathbb{P}_n)) = \int \psi(x - T_L(\mathbb{P}_n)) \, d\mathbb{P}_n(x) = 0.
$$

(1.30)

It follows from the definition of $T_L$ that it depends only on the data $x_n$ through the empirical distribution $\mathbb{P}_n$ and hence we can regard it as a statistical functional. Clearly there is no reason why M–functionals should be restricted to empirical distributions and we may define $T_L(P)$ for any probability distribution $P$ by

$$
\int \psi(x - T_L(P)) \, dP(x) = 0.
$$

(1.31)

M–estimators were first introduced by Huber (1964) as a generalization of maximum likelihood estimates by breaking the link between the function $\psi$ and the density $f$ of the proposed model (see also Huber (1981)). To see this consider the standard location problem. Let $X_i, i = 1, \ldots, n$ be i.i.d. random variables with density $f(x - \mu), \mu \in \mathbb{R}$, and we wish to estimate $\mu$ by maximum likelihood. The estimate is

$$
\hat{\mu}_n = \arg\max_{\mu} \sum_{i=1}^{n} \log f(X_i - \mu).
$$

If $f$ is differentiable then the first derivative of the the right hand side must be zero at $\hat{\mu}_n$ and this give

$$
\sum_{i=1}^{n} \frac{f'(X_i - \hat{\mu}_n)}{f(X_i - \hat{\mu}_n)} = 0
$$

which is (1.30) with $\psi(x) = f'(x)/f(x)$. 

31
Not all estimators are M-estimators. An example is the class of L-estimators which are of the form
\[ L(x_N) = \sum_{i=1}^{n} a_{ni} x_{in} \]
where the ordered sample is \( x_{1:n} \leq \ldots \leq x_{n:n} \).

**Example 1.45.** Consider an ordered sample \( x_{1:n} \leq \ldots \leq x_{n:n} \) and for \( 0 < \alpha < 1/2 \) put \([\alpha n] =: h\). Define an estimate as follows. Replace \( x_{in} \) for \( 1 \leq i \leq h \) by \( x_{h+1:n} \) and \( x_{n-i+1:n} \) for \( 1 \leq i \leq h \) by \( x_{n-h:n} \). The mean of the modified sample is known as the \( \alpha \)-Winsorized mean. Show that it is an L-estimator.

Under appropriate conditions (1.30) always has a unique solution.

**Theorem 1.46.** Let \( \psi : \mathbb{R} \to \mathbb{R} \) be an odd, continuous, bounded and strictly increasing function. Then the equation
\[ \int \psi(x - m) \, dP(x) = 0 \]
has a unique solution \( m =: T_L(P) \) for every distribution \( P \).

**Proof.** The conditions placed on \( \psi \) guarantee that \( \lim_{x \to \pm\infty} \psi(x) = \psi(\pm\infty) \) exist with \(-\infty < \psi(-\infty) < 0 < \psi(\infty) < \infty \). The function
\[ \Psi(m) = \int \psi(x - m) \, dP(x) \]
is continuous and strictly monotone decreasing with \( \Psi(-\infty) = \psi(\infty) > 0 \) and \( \Psi(\infty) = \psi(-\infty) < 0 \). It follows that there exist a uniquely defined \( m \) with \( \Psi(m) = 0 \) and this proves the theorem. \( \square \)

**Remark 1.47.** Although it is generally assumed that \( \psi \) is an odd function the theorem still holds is this is replaced by \( \psi(-\infty) < 0 < \psi(\infty) \).

The statistical functional \( T_L \) is as it stands not affine equivariant. A change of scale of the observations will not be correctly reflected in the value of the \( T_L \).

**Exercise 1.48.** Show that if \( \psi \) satisfies the conditions of Theorem 1.46 then \( T_L \) of (1.30) is translation equivariant but not affine equivariant.
To make $T_L$ affine equivariant we must augment (1.30) by an auxiliary measure of scale $T_S(P)$. This leads to solving
\[
\int \psi \left( \frac{x - T_L(P)}{T_S(P)} \right) dP(x) = 0
\]
for $T_L(P)$. In general (but not always) a good auxiliary measure of scale is the MAD which leads to
\[
\int \psi \left( \frac{x - T_L(P)}{T_{mad}(P)} \right) dP(x) = 0.
\]
(1.33)

To be of use one must to calculate the M-measure for data sets. In other words it must be possible to solve
\[
\Psi(m) = \frac{1}{n} \sum_{i=1}^{n} \psi \left( \frac{x_i - m}{s_0} \right) = 0
\]
where $s_0$ is the auxiliary scale functional. The following procedure can be shown to work.

(a) Put $m_0 = \text{med}(x_n)$ and $s_0 = \text{MAD}(x_n)$.

(b) If $\Psi(m_0) = 0$ put $T_L(P_n) = m_0$ and terminate.

(c) If $\Psi(m_0) < 0$ put $m_2 = m_0$, choose $k$ so that $\Psi(m_0 - 2^k s_0) > 0$ and put $m_1 = m_0 - 2^k s_0$.

(d) If $\Psi(m_0) > 0$ put $m_1 = m_0$, choose $k$ so that $\Psi(m_0 + 2^k s_0) < 0$ and put $m_2 = m_0 + 2^k s_0$.

(e) Put $m_3 = (m_1 + m_2)/2$. If $\Psi(m_3) < 0$ put $m_2 = m_3$. If $\Psi(m_3) > 0$ put $m_1 = m_3$.

(f) Iterate (e) until $m_2 - m_1 < 10^{-4} s_0 / \sqrt{n}$.

(g) Put $T_L(P_n) = (m_1 + m_2)/2$ and terminate.

We do of course require an explicit function $\psi$. In the following we shall restrict ourselves to the $\psi$–functions of the next example. We can control their behaviour by an appropriate choice of the tuning constant $c$.

**Example 1.49.** The function $\psi_c$ given by
\[
\psi_c(x) = \frac{\exp(cx) - 1}{\exp(cx) + 1}, \quad x \in \mathbb{R}, \quad c > 0, \quad (1.34)
\]

satisfies the conditions of Theorem 1.46.
Exercise 1.50. Write an R function to calculate $T_L$ for the $\psi$–function of (1.49).

Example 1.51. If we put $c = 3$ and use $T_S = T_{mad}$ then the value of the M-measure of location for the copper data of Example 1.1 is 2.0224 which lies between the mean and the median.

That this is generally the case can be seen by noting

$$\lim_{c \to 0} \psi_c(x)/c = x$$

and

$$\lim_{c \to \infty} \psi_c(x) = \text{sgn}(x).$$

It follows that by letting $c$ increase from a small to a large value we can extrapolate between the mean and the median. The question arises as to which value to choose. To help us in this decision we can investigate the behaviour of the different functionals under well–controlled situations by simulating data with different distributions. The ones we choose are the standard normal distribution and the t-distributions with 20, 10, 5, 4, 3, 2 and 1 degrees of freedom. For each of the eight functionals the median, the M–functional with $c=40, 10, 5, 3, 2, 1$ and the mean and we calculate the mean square error

$$MSE_i = \frac{1}{10000} \sum_{j=1}^{10000} T_i(\mathbb{P}_{nj})^2, \quad i = 1, \ldots, 8$$

based on 10000 simulations for each situation. In all cases we use the MAD as an auxiliary scale measure. The results are given in Table expressed in terms of their percentage efficiency with respect to the most efficient method that is $100 \min(MSE_{1:8})/MSE_i, i = 1, \ldots, 8$. For the normal distribution the mean is optimal.

Exercise 1.52. Repeat the simulations using the standard deviation as auxiliary measure of scale and interpret your results.

The results indicate that a values of $c$ between 2 and 3 are a reasonable choice. They represent a good compromise between being efficient for normal data whilst still being able to accommodate larger values from the $t$–distributions. Here is Tukey on the compromises which lead to the milk bottle:
Table 3: Relative efficiencies of the mean and median and the M–functional with ψ–function given by (1.34) for different values of c. The MAD was used as an auxiliary measure of scale.

<table>
<thead>
<tr>
<th></th>
<th>median</th>
<th>c = 40</th>
<th>c = 10</th>
<th>c = 5</th>
<th>c = 3</th>
<th>c = 2</th>
<th>c = 1</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>N(0, 1)</td>
<td>64</td>
<td>66</td>
<td>73</td>
<td>81</td>
<td>88</td>
<td>93</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td>t20</td>
<td>68</td>
<td>70</td>
<td>77</td>
<td>85</td>
<td>91</td>
<td>96</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>t10</td>
<td>72</td>
<td>74</td>
<td>82</td>
<td>89</td>
<td>95</td>
<td>98</td>
<td>100</td>
<td>94</td>
</tr>
<tr>
<td>t5</td>
<td>79</td>
<td>80</td>
<td>88</td>
<td>95</td>
<td>99</td>
<td>100</td>
<td>96</td>
<td>80</td>
</tr>
<tr>
<td>t4</td>
<td>81</td>
<td>83</td>
<td>91</td>
<td>97</td>
<td>100</td>
<td>99</td>
<td>93</td>
<td>70</td>
</tr>
<tr>
<td>t3</td>
<td>83</td>
<td>85</td>
<td>92</td>
<td>98</td>
<td>100</td>
<td>98</td>
<td>88</td>
<td>49</td>
</tr>
<tr>
<td>t2</td>
<td>87</td>
<td>89</td>
<td>97</td>
<td>100</td>
<td>97</td>
<td>92</td>
<td>76</td>
<td>3</td>
</tr>
<tr>
<td>t1</td>
<td>92</td>
<td>95</td>
<td>100</td>
<td>95</td>
<td>84</td>
<td>72</td>
<td>49</td>
<td>0</td>
</tr>
</tbody>
</table>

Given that our concern is with engineering, it follows that we are likely to be concerned with additional attributes beyond those which arise naturally in a science-flavored approach to the underlying problems. The history of the milk bottle is an illuminating example. Maximum milk for minimum glass would lead to spherical shape; adding a bound on cross-sectional dimensions would lead to circular cylinders with hemispherical ends. The following steps of practicality occurred, historically:

- a flat bottom to rest stably,
- a reduced top to allow a small cap,
- a rounded square cross section to pack better in the refrigerator,
- pushed in sides to make the bottles easier to grasp,
- replacement of glass by paper restarted evolution with new considerations.

Those who had been concerned with the initial optimization could logically (but not reasonably) have argued that all these additional considerations had dirtied up a clean problem. As a user of milk bottles, I am glad to see the additional attributes brought in.

1.7.2 M-measures of scatter

M-measures of scale $T_S$ are defined in a similar manner as follows.
Definition 1.53. For a function $\chi$ and a sample $x_n$ with empirical distribution $P_n$ any solution $T_S(x_n)$ of

$$\frac{1}{n} \sum_{i=1}^{n} \chi \left( \frac{x_i}{T_S(x_n)} \right) = \int \chi \left( \frac{x}{T_S(x_n)} \right) dP_n(x) = 0.$$  \hspace{1cm} (1.35)

is called an $M$-measure or $M$-estimator of scale. For a general probability distribution $P$ we define $T_S(P)$ to be the solution of

$$\int \chi \left( \frac{x}{T_S(P)} \right) dP(x) = 0.$$  \hspace{1cm} (1.36)

Example 1.54. Suppose that the sample $x_n$ has mean zero and put $\chi(x) = x^2 - 1$. Show that $T_S$ is the standard deviation of the sample. Repeat for (1.36).

Example 1.55. Suppose that the sample $x_n$ has median zero and put $\chi(x) = \text{sgn}(\lvert x \rvert - 1)$. Show that $T_S$ is the MAD of the sample. Repeat for (1.36).

Corresponding to Theorem 1.46 we have

Theorem 1.56. Let $\chi : \mathbb{R} \to \mathbb{R}$ be an even, continuous, bounded and strictly increasing function on $\mathbb{R}_+$ with $\chi(0) = -1$ and $\chi(\infty) = 1$. Then the equation

$$\int \chi(x/s) dP(x) = 0$$

has a unique solution $s =: T_S(P)$ with $0 < T_S(P) < \infty$ for every probability distribution $P$ with $P(\{0\}) < 1/2$.

Proof. As $P(\{0\}) < 1/2 < 1$ the function $\Xi(s) = \int \chi(x/s) dP(x)$ is strictly monotone decreasing and continuous. Furthermore $\lim_{s \to \infty} \Xi(s) = -1$ and $\lim_{s \to 0} \Xi(s) = -P(\{0\}) + (1 - P(\{0\})) > 0$ as $P(\{0\}) < 1/2$. Hence there exist a unique $s =: T_S(P)$ with $\Xi(s) = 0$. \hspace{1cm} \Box

Example 1.57. The function

$$\chi(x) = \frac{x^4 - 1}{x^3 + 1}$$  \hspace{1cm} (1.37)

satisfies the assumptions of the Theorem 1.56 and it will be our standard $\chi$-function.
Exercise 1.58. Explain why is there no point in using a tuning constant $c$ and putting
\[ \chi_c(x) = \frac{(cx)^4 - 1}{(cx)^4 + 1}. \]
Does the same hold if we put
\[ \chi_p(x) = \frac{|x|^p - 1}{|x|^p + 1} \]
with $p > 0$?

Usually scale is a nuisance parameter and efficiency is not as important as robustness. This explains to some extent the choice of $\chi$ in Example 1.37.

Exercise 1.59. Plot the $\chi_p$ of Example 1.58 for various values of $p$ and compare with $\text{sgn}(|x| - 1)$.

The calculation of an M-measure of scale follows the lines of that for M-measures of location. We write
\[ \Xi(s) = \frac{1}{n} \sum_{i=1}^{n} \chi(x_i/s). \]
(a) Put $s_0 = \text{MAD}(x_n)$.
(b) If $\Xi(s_0) = 0$ put $T_S(P_n) = s_0$ and terminate.
(c) If $\Xi(s_0) < 0$ put $s_2 = s_0$, choose $k$ so that $\Xi(s_0 2^{-k}) > 0$ and put $s_1 = s_0 2^{-k}$.
(d) If $\Xi(s_0) > 0$ put $s_1 = s_0$, choose $k$ so that $\Psi(s_0 2^k) < 0$ and put $m_2 = s_0 2^k$.
(e) Put $s_3 = \sqrt{s_1 s_2}$. If $\Xi(s_3) < 0$ put $s_2 = s_3$. If $\Xi(s_3) > 0$ put $s_1 = s_3$.
(f) Iterate (e) until $s_2/s_1 - 1 < 10^{-4}/\sqrt{n}$
(g) Put $T_S(P_n) = \sqrt{s_1 s_2}$ and terminate.

Exercise 1.60. Write an R function to calculate $T_S$ for the $\chi$–function of Exercise 1.37.

Example 1.61. The measure $T_S$ with the $\chi$–function (1.37) is not Fisher consistent for $\sigma$ of the $\mathcal{N}(0, \sigma^2)$ distribution. We have $T_S(\mathcal{N}(0, 1)) = 0.646409$ so to make it Fisher consistent we must multiply by 1.547005. The same holds for the exponential distribution $\mathcal{E}(\lambda)$ which is a scale family. We have $T_S(\mathcal{E}(\lambda)) = 0.6756551$ so to make it Fisher consistent at the exponential distribution we should multiply by 1.480045.
Example 1.62. If we treat the suicide data as a pure scale problem then the value of $T_S$ with the $\chi$-function (1.37) is 71.49496. If we make $T_S$ Fisher consistent at the exponential distribution (look at the lower panel of Figure 1.37) then we should multiply by 1.547008. This results in 105.81 compared with a mean of 122.32.

1.7.3 Simultaneous M-measures of location and scale

In general location and scale functionals have to be estimated simultaneously. This leads to the following system of equations

$$\frac{1}{n} \sum_{i=1}^{n} \psi \left( \frac{x_i - m}{s} \right) = \int \psi \left( \frac{x - m}{s} \right) dP_n(x) = 0 \quad (1.38)$$

$$\frac{1}{n} \sum_{i=1}^{n} \chi \left( \frac{x_i - m}{s} \right) = \int \chi \left( \frac{x - m}{s} \right) dP_n(x) = 0 \quad (1.39)$$

or for a general probability distribution $P$,

$$\int \psi \left( \frac{x - m}{s} \right) dP(x) = 0 \quad (1.40)$$

$$\int \chi \left( \frac{x - m}{s} \right) dP(x) = 0 \quad (1.41)$$

In order to guarantee that a solution exists we must place conditions on the functions $\psi$ and $\chi$ as well as on the probability distribution $P$.

$(\psi 1)$ \hspace{1em} $\psi(-x) = -\psi(x)$ for all $x \in \mathbb{R}$.

$(\psi 2)$ \hspace{1em} $\psi$ is strictly increasing.

$(\psi 3)$ \hspace{1em} $\lim_{x \to \infty} \psi(x) = 1$.

$(\psi 4)$ \hspace{1em} $\psi$ is differentiable with a continuous derivative $\psi^{(1)}$.

$(\chi 1)$ \hspace{1em} $\chi(-x) = \chi(x)$ for all $x \in \mathbb{R}$.

$(\chi 2)$ \hspace{1em} $\chi : \mathbb{R}_+ \to [-1, 1]$ is strictly increasing.

$(\chi 3)$ \hspace{1em} $\chi(0) = -1$.

$(\chi 4)$ \hspace{1em} $\lim_{x \to \infty} \chi(x) = 1$.

$(\chi 5)$ \hspace{1em} $\chi$ is differentiable with a continuous derivative $\chi^{(1)}$.

$(\psi \chi 1)$ \hspace{1em} $\chi^{(1)}/\psi^{(1)} : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing.

$(P)$ \hspace{1em} $\Delta(P) < 1/2$ where $\Delta(P) := \max_x P(\{x\})$ which agrees with (1.34) when $P = P_n$.  

38
The existence of joint M-measures of location and scale is not trivial but the following theorem holds.

**Theorem 1.63.** Under the above assumptions \((\psi_i), i = 1, \ldots, 4, (\chi_i), i = 1, \ldots, 5, (\psi \chi_1)\) and \((P)\) the equations (1.38) and (1.39) have a unique solution \((m, s) =: (T_L(P), T_S(P))\).

**Proof.** A proof may be found in Huber (1981). \(\square\)

**Example 1.64.** We consider the \(\psi\)- and \(\chi\)-functions

\[
\psi(x) = \psi(x, c) = \frac{(\exp(cx) - 1)}{(\exp(cx) + 1)} \quad \text{(1.42)}
\]

\[
\chi(x) = \frac{x^4 - 1}{x^4 + 1}. \quad \text{(1.43)}
\]

They clearly fulfill all the conditions of Theorem 1.63 for all values of \(c > 0\) apart possibly for the condition \((\psi \chi_1)\). Direct calculations show that this holds for \(c > 2.56149\). This does not exclude the existence and uniqueness if \(c < 2.56149\) as the condition is a sufficient one.

**Exercise 1.65.** Verify the calculation of Example 1.64

An algorithm for calculating the solutions for a data set may be derived from the proof of the theorem given in Huber (1981). The following simple iteration scheme seems to always converge although I have no proof:

- Put \(s = s_1 = T_{mad}(P_n)\) in (1.38) and solve for \(m = m_1\).
- Put \(m = m_1\) in (1.39) and solve for \(s = s_2\).
- Put \(s = s_2\) in (1.38) and solve for \(m = m_2\).
- Put \(m = m_2\) in (1.39) and solve for \(s = s_3\).
- Iterate as indicated and terminate when

\[
|m_{k+1} - m_k| + |s_{k+1} - s_k| < 10^{-4} s_k / \sqrt{n}.
\]

**Example 1.66.** We calculate the simultaneous location and scale functionals \(T_L\) and \(T_S\) for the copper data using the functions

\[
\psi(x) = \psi_c(x) = \frac{(\exp(cx) - 1)}{(\exp(cx) + 1)}
\]

\[
\chi(x) = \frac{x^4 - 1}{x^4 + 1}.
\]
Table 4: Relative efficiencies of the mean and median and the location part of the M–functional defined by (1.38) and (1.39) with $\psi$–function given by (1.34) for different values of $c$ and $\chi$–function given by (1.37).

<table>
<thead>
<tr>
<th></th>
<th>median</th>
<th>$c = 40$</th>
<th>$c = 10$</th>
<th>$c = 5$</th>
<th>$c = 3$</th>
<th>$c = 2$</th>
<th>$c = 1$</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(0, 1)$</td>
<td>64</td>
<td>65</td>
<td>70</td>
<td>76</td>
<td>82</td>
<td>87</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>$t_{20}$</td>
<td>69</td>
<td>69</td>
<td>74</td>
<td>80</td>
<td>86</td>
<td>92</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>$t_{10}$</td>
<td>72</td>
<td>73</td>
<td>78</td>
<td>84</td>
<td>90</td>
<td>95</td>
<td>100</td>
<td>94</td>
</tr>
<tr>
<td>$t_{5}$</td>
<td>79</td>
<td>80</td>
<td>85</td>
<td>91</td>
<td>96</td>
<td>99</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>$t_{4}$</td>
<td>81</td>
<td>82</td>
<td>88</td>
<td>93</td>
<td>97</td>
<td>100</td>
<td>98</td>
<td>70</td>
</tr>
<tr>
<td>$t_{3}$</td>
<td>83</td>
<td>84</td>
<td>89</td>
<td>95</td>
<td>98</td>
<td>100</td>
<td>95</td>
<td>49</td>
</tr>
<tr>
<td>$t_{2}$</td>
<td>87</td>
<td>88</td>
<td>94</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>86</td>
<td>3</td>
</tr>
<tr>
<td>$t_{1}$</td>
<td>93</td>
<td>95</td>
<td>100</td>
<td>99</td>
<td>94</td>
<td>84</td>
<td>61</td>
<td>0</td>
</tr>
</tbody>
</table>

with tuning parameter $c = 3$ we obtain $T_L = 2.021713$ (mean 2.015926, median 2.03) and $T_S = 0.0674$ (standard deviation 0.1140260, MAD 0.1). If we make $T_S$ Fisher consistent at the normal model we must multiply $T_S$ by 1.547005. The scale then becomes 0.1043.

**Exercise 1.67.** Verify the calculation of Example 1.66.

We now repeat the simulations of Table 3 for the simultaneous M–functional. Care should be taken when making a direct comparison as the location estimates are based on different scales and therefore the values of $c$ cannot be compared directly. To see this we note if we use the $\psi$–function (1.34) with $c = c_0$ and the auxiliary scale measure $T_S$ when solving (1.32) the result is exactly the same as solving (1.32) with $c = 2c_0$ and auxiliary scale measure $2T_S(P)$. In our particular case Table 3 was based on the default version of the MAD which is Fisher consistent whereas the results of Table 4 use the $T_S$ of the simultaneous M–functional. This is approximately a factor 0.6464 smaller which means that a value of $c$ in Table 3 corresponds roughly to a values of 1.547$c$ in Table 4.

### 1.8 Analytic properties of M–functionals

We now take advantage of the functional analytical approach and investigate two analytical properties, namely local boundedness and differentiability, which are of relevance for statistical applications. I know of no statistical application of continuity alone. Local boundedness is useful for explorative data analysis whereas differentiability is useful for the more formal inference part of statistics. Concepts such as continuity and differentiability require a
topology. We denote the set of all probability distributions on the Borel sets of $\mathbb{R}$ by $\mathcal{P}$ and turn $\mathcal{P}$ into a metric space by introducing the Kolmogorov metric

**Definition 1.68.**

$$d_{ko}(P, Q) = \sup_{x \in \mathbb{R}} \left\{ |P((-\infty, x]) - Q((-\infty, x]| \right\}. \quad (1.44)$$

A second useful metric is the Kuiper metric

**Definition 1.69.**

$$d_{ku}(P, Q) = \sup_{a < b} \left\{ |P((a, b]) - Q((a, b])| \right\}. \quad (1.45)$$

**Exercise 1.70.** Show that $d_{ko}$ and $d_{ku}$ define the same topology on $\mathcal{P}$.

**Exercise 1.71.** Show that the metric space $(\mathcal{P}, d_{ko})$ is neither separable nor complete.

A real–valued statistical functional $T$ is now a mapping from some subset $\mathcal{P}_T$ into $\mathbb{R}$

$$T: (\mathcal{P}_T, d_{ko}) \to \mathbb{R}. \quad \text{In general } \mathcal{P}_T \text{ will be a strict subset of } \mathcal{P} \text{ but will contain the subset } \mathcal{E}_n \text{ of all empirical measures } \mathbb{P}_n \text{ of the form}$$

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

where $\delta_{x}$ is the Dirac measure

$$\delta_{x}(B) = \{x \in B\}.$$

**Example 1.72.** For the mean functional $T(P) = \int x \, dP(x)$ we have

$$\mathcal{P}_T = \{P: \int |x| \, dP(x) < \infty \}.$$

Clearly $\mathcal{E}_n \subset \mathcal{P}_T$.

**Example 1.73.** For the median functional defined as in Definition (1.18) we have $\mathcal{P}_T = \mathcal{P}$. 
1.8.1 Local boundedness and breakdown

We now investigate how sensitively $T(P)$ depends on $P$. A useful concept is the bias function $b(T,P,\varepsilon,d)$ of $T$ at the point $P$. It is defined by

**Definition 1.74.**

$$b(T,P,\varepsilon) := b(T,P,\varepsilon,d_{ko}) = \sup \{ |T(Q) - T(P)| : d_{ko}(P,Q) < \varepsilon \}$$

(1.46)

where $T(Q) := \infty$ if $Q \notin \mathcal{P}_T$. The functional $T$ is called locally bounded at $P$ if $b(T,P,\varepsilon) < \infty$ for some $\varepsilon > 0$.

**Example 1.75.** For the mean functional $T(P) = T_{ave}(P) = \int x \, dP(x)$ we have

$$b(T_{ave},P,\varepsilon) = \infty$$

for all $P \in \mathcal{P}_T$ and $\varepsilon > 0$. The mean is not locally bounded.

**Example 1.76.** For the median functional $T(P) = P_{med}(P)$ we have

$$b(P_{med},P,\varepsilon) < \infty$$

for all $P \in \mathcal{P}_T$ and $\varepsilon < 1/2$. The median is locally bounded.

We prove a stronger result than that stated in Example 1.76.

**Definition 1.77.** We call a functional $T$ translation equivariant if

- $P^\mu \in \mathcal{P}_T$ for all $\mu \in \mathbb{R}$ and $P \in \mathcal{P}_T$ where $P^\mu(B) = P(B - \mu)$ for all Borel sets $B$.

- $T(P^\mu) = T(P) + \mu$ for all $\mu$ and $P \in \mathcal{P}_T$

Clearly translation equivariance is a weakening of affine equivariance for location functionals.

**Theorem 1.78.** Let $T : \mathcal{P} \to \mathbb{R}$ be a translation equivariant functional. Then

$$b(P_{med},N(0,1),\varepsilon,d_{ko}) = \Phi^{-1}(1/2 + \varepsilon)$$

and

$$b(P_{med},N(0,1),\varepsilon,d_{ko}) \leq b(T,N(0,1),\varepsilon,d_{ko})$$

for all $\varepsilon$, $0 \leq \varepsilon < 1/2$. 

42
Proof. Let \( \Phi \) denote the distribution function of the standard normal distribution. The Kolmogorov ball with centre \( \Phi \) and radius \( \varepsilon \) is given by

\[
K(\Phi, \varepsilon) = \{ F : \Phi(x) - \varepsilon < F(x) < \Phi(x) + \varepsilon, \, x \in \mathbb{R} \}.
\]

This is a tube with upper boundary \( F_u \) and lower boundary \( F_l \) given by

\[
F_u(x) = \min\{1, \Phi(x) + \varepsilon\}, \quad F_l(x) = \max\{0, \Phi(x) - \varepsilon\}
\]

We can regard \( F_u \) as a distribution function with a point mass of \( \varepsilon \) at \(-\infty\) and \( F_l \) as a distribution function with a point mass of \( \varepsilon \) at \(+\infty\).

We now move \( F_u \) to the right so that it just remains in the ball \( K(\Phi, \varepsilon) \). This is shown in Figure 6. Some calculation shows that we can translate \( F_u \) by an amount \( \delta \) where

\[
F_u(-x_1 - \delta) = F_l(x_1) = 1/2 \quad \text{where} \quad F_u(-x_1) = 1/2.
\]

If now \( 0 < \varepsilon < 1/2 \) we have \( F_u(x) = \Phi(x) + \varepsilon \) and inserting this into (1.47) leads to

\[
\Phi(-x_1 - \delta) + \varepsilon = 1/2
\]

so that

\[
\delta = x_1 - \Phi^{-1}(1/2 - \varepsilon) = 2\Phi^{-1}(1/2 + \varepsilon)
\]

on noting that \( \Phi(-x_1) + \delta = F_u(-x_1) = 1/2 \). As \( T \) is translation equivariant we have

\[
T(F_u(\cdot - \delta)) - T(F_u) = 2\Phi^{-1}(1/2 + \varepsilon)
\]

and as both \( F_u \) and \( F_u(\cdot - \delta) \) lie in the Kolmogorov ball this implies

\[
|T(\Phi) - T(F_u)| \geq \Phi^{-1}(1/2 + \varepsilon)
\]

or

\[
|T(\Phi) - T(F_u(\cdot - \delta))| \geq \Phi^{-1}(1/2 + \varepsilon)
\]

In either case we have

\[
b(\varepsilon, T, N(0, 1), d_{ko}) \geq \Phi^{-1}(1/2 + \varepsilon).
\]

It remains to show that

\[
b(\varepsilon, T_{med}, N(0, 1), d_{ko}) = \Phi^{-1}(1/2 + \varepsilon).
\]

We note that the median respects the stochastic ordering of distributions. That is if \( F \leq G \) then \( T_{med}(F) \geq T_{med}(G) \). As \( F_l \leq F \leq F_u \) for all \( F \) in \( K(\Phi, \varepsilon) \) we have

\[
T_{med}(F_u) \leq T_{med}(F) \leq T_{med}(F_l)
\]

43
Figure 6: The figure shows the distribution function $\Phi$ of the standard normal distribution, the upper bound $F_u$ and the lower bound $F_l$ of the Kolmogorov ball $K(\Phi, \varepsilon)$ with $\varepsilon = 0.15$. It also shows the upper bound $F_u$ translated to the right so that it just remains in the $K(\Phi, \varepsilon)$.

and as $T_{\text{med}}(F_u) = -x_1 = -\Phi^{-1}(1/2 + \varepsilon)$ and $T_{\text{med}}(F_u) = x_1$ we see in combination with $T_{\text{med}}(\Phi) = 0$ that (1.48) holds.

We define the breakdown point of a functional $T$ at the distribution $P$ as follows.

**Definition 1.79.** For a statistical functional $T$ we define

$$
\varepsilon^*(T, P, d_{k0}) = \inf \{ \varepsilon > 0 : \sup_{Q, d_{ko}(Q, P) < \varepsilon} |T(Q) - T(P)| = \infty \}.
$$

(1.49)

Corresponding to Theorem 1.31 we state without proof

**Theorem 1.80.** Let $T : \mathcal{P}_T \rightarrow \mathbb{R}$ be a translation equivariant functional. Then for any $P \in \mathcal{P}_T$

$$
\varepsilon^*(T, P, d_{k0}) \leq 1/2
$$

and

$$
\varepsilon^*(\text{median}, P, d_{k0}) = 1/2.
$$

The calculation of breakdown points $\varepsilon^*(T, P, d_{k0})$ is somewhat more technically complicated than the calculation of finite sample breakdown points $\varepsilon^*(T, x_n)$. Indeed this was one of the motivations for introducing finite sample breakdown points. We now calculate the finite sample breakdown point
of the location M-functional $T$ defined by
\[ \frac{1}{n} \sum_{i=1}^{n} \psi(x_i - T(x_n)) = 0 \]
where $\psi$ satisfies the usual conditions for a $\psi$–function with $\psi(\infty) = 1$.

Consider now replacement samples $x_n^k$ for which $T(x_n^k)$ tends to infinity. For the sample points $x_i$ which have not been replaced $\psi(x_i - T(x_n^k))$ tends to -1. For the $k$ remaining points $x_n^k$ we have $\psi(x_n^k - T(x_n^k)) \leq 1$. From this and
\[ \frac{1}{n} \sum_{i=1}^{n} \psi(x_n^k - T(x_n^k)) = 0 \]
we deduce $-(n - k) + k \geq 0$ and hence $k \geq n/2$. This implies $\varepsilon^*(T, x_n) \geq \lfloor (n + 1)/2 \rfloor$. By Theorem 1.31 however $\varepsilon^*(T, x_n) \leq \lfloor (n + 1)/2 \rfloor$ and hence $\varepsilon^*(T, x_n) = \lfloor (n + 1)/2 \rfloor$. We note that if we choose the $\psi$–function
\[ \psi_c(x) = \frac{\exp(cx) - 1}{\exp(cx) + 1} \]
then the breakdown point does not depend on $c$. If we denote the functional with $\psi = \psi_c$ by $T_c$ then the bias function $b(\varepsilon, T_c, P, d_{ko})$ does depend on $c$. To show this we calculate
\[ T_c((1 - \varepsilon)N(0, 1) + \varepsilon \delta_{100}) \]
which represents a standard normal sample contaminated by $100\varepsilon\%$ of outliers all situated at 100. Figure 7 shows the bias curves of the mean (uppermost) and the median (lowest) together with those of the functionals $T_c$ for $c = 1, 2, 3, 4, 5, 10, 40$ in decreasing order of bias. These indicate that a value of $c \geq 3$ has a bias behaviour for degrees of contamination of up to $30\%$ which is not greatly inferior to that of the median.

We look at the breakdown behaviour of the joint M-functional (1.38) and (1.39). For a sample $x_n$ the we set $(m, s) = (T_L(P_n), T_S(P_n))$ so the equations read
\[ \sum_{i=1}^{n} \psi \left( \frac{x_i - m}{s} \right) = 0 \]  
\[ \sum_{i=1}^{n} \chi \left( \frac{x_i - m}{s} \right) = 0. \]

The situation is more complicated. The condition on $\Delta(P) < 1/2$ for the existence and uniqueness of a solution is only sufficient. This makes it difficult to
Figure 7: The figure shows the bias curves of the mean (uppermost) and the median (lowest) together with those of the functionals $T_c$ for $c = 1, 2, 3, 4, 5, 10, 40$ in decreasing order of bias.

decide whether a solution exists even if $\Delta(\mathbb{P}) \geq 1/2$. Because of this we consider only the breakdown of the location part $T_L(\mathbb{P}_n)$. Suppose we can alter $k$ observations in such a manner than we obtain a sequence of values $((m_l, s_l))_1^\infty$ of solutions of (1.50) and (1.51) with $\lim_{l \to \infty} m_l = \infty$. We consider two cases.

**Case 1.** $s_l$ is bounded from $\infty$.
For the $n-k$ observations not altered $(x_i - m_l)/s_l$ tends to $-\infty$. This implies $-(n-k) + k \geq 0$ and hence $k \geq n/2$.

**Case 2.** $\limsup_{l \to \infty} s_l = \infty$.
We can find a subsequence with $\lim_{l' \to \infty} m_{l'}/s_{l'} = \gamma$ with $0 \leq \gamma \leq \infty$. For the $n-k$ observations not altered $(x_i - m_{l'})/s_{l'}$ tends to $-\gamma$. This implies $(n-k)\psi(-\gamma) + k \geq 0$ and $(n-k)\chi(-\gamma) + k \geq 0$. As $\psi$ is an odd function and $\chi$ an even function we deduce $\psi(\gamma) \leq k/(n-k)$ and $\chi(\gamma) \geq -k/(n-k)$. As $\gamma \geq 0$ this implies

$$\chi^{-1}(-k/(n-k)) \leq \gamma \leq \psi^{-1}(k/(n-k))$$

Let $\varepsilon_0$ be the solution of

$$\chi^{-1}(-\varepsilon_0/(1 - \varepsilon_0)) = \psi^{-1}(\varepsilon_0/(1 - \varepsilon_0)).$$  \tag{1.52}
Then \( k/n \geq \varepsilon_0 \). As \( \varepsilon_0 < 1/2 \) we see that the finite sample breakdown point is at least \( \varepsilon^* \geq [n\varepsilon_0]/n \).

We conclude \([n\varepsilon_0]/n \leq \varepsilon^*(T_L, P_n) \leq ([n+1]/2)/n\). In fact one can go further and show that
\[
[n\varepsilon_0]/n \leq \varepsilon^*(T_L, P_n) \leq ([n\varepsilon_0 + 1])/n
\]
with \( \varepsilon_0 \) given by (1.52).

We now calculate the finite sample breakdown point for the simultaneous M–estimator given by (1.38) and (1.39) with
\[
\psi_c(x) = \exp(cx) - 1 \\
\chi(x) = \frac{x^4 - 1}{x^4 + 1}
\]
A quick calculation shows
\[
\psi_c^{-1}(u) = \frac{1}{c} \log \left( \frac{1 + u}{1 - u} \right) \\
\chi^{-1}(u) = \left( \frac{1 + u}{1 - u} \right)^{1/4}
\]
which, when inserted into (1.52), leads to
\[
-\log(1 - 2\varepsilon^*) \left( \frac{1 - 2\varepsilon^*}{1 - \varepsilon^*} \right)^{1/4} = c
\]
which can be solved numerically. Figure 8 shows a plot of the breakdown point as a function of the tuning constant \( c \). It is clear that the breakdown point for the simultaneous M–estimator does depend on the tuning constant \( c \) which is in contrast to the simple location functional with auxiliary scale taken to be the MAD. The reason why the breakdown does not occur in the simulation results of Table 4 is that all simulations were done with symmetric distribution whereas breakdown occurs only when the outliers or contamination is one–sided. In the next section we discuss the case of one–sided outliers and how they can be dealt with.

### 1.9 Redescending M–estimators I

As a first step we investigate the bias behaviour of the shorth. We consider the simple point mass contamination model
\[
P_{\varepsilon,x} = (1 - \varepsilon)N(0, 1) + \varepsilon\delta_x
\]
which represents 100\( \varepsilon \)% contamination of the \( N(0, 1) \) distribution by a point mass of \( \varepsilon \) situated at the point \( x \). The bias of the shorth is \( b(x) = |\text{shorth}(P_{\varepsilon,x})|\)
Figure 8: The figure shows the finite sample breakdown point of the simultaneous M-estimator with $\psi = \psi_c$ given by (1.34) and $\chi$ given by (1.37) as a function of the tuning parameter $c$.

Figure 9: The figure shows the bias of the shorth and that of the median for the contamination model (1.54) as a function of $x$. 
and is shown in Figure 9 together with the bias of the median. The largest bias of the shorth exceeds that of the median but when $x$ is large the bias of the shorth drops to zero in contrast to that of the median. The reason is that the shorth completely discounts outliers when they are sufficiently large whereas their influence on the median is bounded but is never zero. This is true for all M–estimators whose $\psi$–function is monotone and bounded. The influence of large outliers can only be decreased to zero if the monotonic property is dropped and $\lim_{x \to \pm \infty} \psi(x) = 0$. Such a $\psi$–function is called re-descending. Figure 10 shows the Hampel three part re-descending $\psi$–function with hinges at the points 1.31, 2.039 and 4 as in Hampel et al (1986) page 166. It is an odd function of $x$ and for $x > 0$ it is defined by

$$
\psi_{hamp}(x) = \begin{cases} 
  x, & 0 \leq x \leq 1.31, \\
  1.31, & 1.31 < x \leq 2.039, \\
  1.31 - 0.6680265(x - 2.039), & 2.039 < x \leq 4, \\
  0, & x > 4.
\end{cases}
$$

(1.55)

We require an auxiliary scale measure which we take to be the MAD and are lead to solving the equation

$$
\Psi_n(m) = \frac{1}{n} \sum_{i=1}^{n} \psi \left( \frac{x_i - m}{\text{MAD}(x_n)} \right) = 0
$$

(1.56)

where $\psi$ is now no longer monotonic. Figure 11 shows a plot of the values of $\Psi_n(m)$ against $m$ for a sample of size $n = 100$ of which 60 were $N(0,1)$ and the remaining 40 were $N(6.5,0.8^2)$. It is apparent that the solution of (1.56) is not unique. There are several ways of overcoming this problem and the one we choose is to take that solution which is closest to the median of the sample as follows. Evaluate $\Psi_n(m_0)$ with $m_0 = T_{med}(x_n)$ and then move left and right in steps say of 0.1MAD$(x_n)$ so that $m^+_k = m_0 + 0.1k\text{MAD}(x_n)$ and $m^-_k = m_0 - 0.1k\text{MAD}(x_n)$, $k = 1, 2, \ldots$ until the first time that $\Psi_n(m^-_k)$ or $\Psi_n(m^+_k)$ has a different sign from $\Psi_n(m_0)$. We now have two points where $\Psi_n$ has different signs and we can use the bisection method to find the zero with a prescribed accuracy.

Figure 12 repeats Figure 9 for the Hampel re-descending M–estimator with $\psi$ given by (1.55). It is seen that the largest bias is greater than the largest bias of the median as was to be expected. The difference is however not substantial and the Hampel estimator has the advantage that the influence of large outliers is down–weighted to zero. If the Hampel estimator is included in the results given in Table 3 then its relative efficiencies for the $N(0,1)$ and the $t_k$–distributions for $k = 20, 10, 5, 4, 3, 2, 1$ are

$$
90, 91, 93, 96, 97, 93, 95 \text{ and } 101
$$

(1.57)

49
Figure 10: The figure shows the plot of the Hampel three part redescending \( \psi \)-function given by (1.55).

Figure 11: The figure shows the plot of \( \Psi_n \) of (1.56) for a sample of size \( n = 100 \) as described in the text.
Figure 12: The figure shows the bias of the Hampel redescending $M$-estimator with $\psi$--function given by (1.55) and that of the median for the contamination model (1.54) as a function of $x$.

respectively. In other words the overall performance of the Hampel estimator is better than that of all the estimators included in Table 3. A redescending estimator will be our estimator of choice when considering the one-way analysis of variance.

1.10 Differentiability and asymptotic normality

If the data $X_n$ are normally distributed then the mean $\bar{X}_n$ is also normally distributed. If the data are not normally distributed but have a finite second moment then it follows from the Central Limit Theorem that $\bar{X}_n$ is asymptotically normally distributed. If the distribution of the $X_i$ has long tails then the normal approximation for small values of $n$ may be poor. If the $X_i$ do not have a finite second moment then $\bar{X}_n$ may not even be asymptotically normally distributed. This is the case for the Cauchy distribution where $\bar{X}_n$ has the same distribution as $X_1$. One advantage of $M$--estimators with smooth $\psi$-- and $\chi$--functions is that they are asymptotically normally distributed for almost any distribution $P$ of the sample $X_n$. Indeed the only restriction is $\Delta(P) < 1/2$. The upper panel of Figure 13 shows the distribution of $\bar{X}_n$ for $n = 10$ for $N(0,1)$ data. The figures is based on 50000 simulations and
shows a histogram of the different values of $\bar{X}_n$ together with the density of the normal distribution. The lower panel shows the same for the redescending $M$–estimator to be defined in Section 1.11 below. The simulated standard deviations of the two estimates were 1.00 and 1.06 and respectively. Figure 14 shows the results for Cauchy data where the non-normality of the mean is apparent. The standard deviations over the 50000 simulations were 427.51 and 2.02 respectively. If the distribution of the sample is skewed then the normal approximation deteriorates. Figure 15 shows the results corresponding to Figure 13 for data generated according to the Gamma distribution of Figure 16 which has shape parameter 3 and scale parameter 1. The median is a location $M$–estimator but its $\psi$–function is not even continuous. Depending on the distribution $P$ of the data the median may converge in a finite number of steps, it may be asymptotically normal or it may not even be consistent. The same applies to the MAD as a scale–estimator. Even if we just use the MAD as an auxiliary scale functional it may still influence the asymptotic behaviour of the location estimate. We show an example with data generated according to

$$P = 0.5U([0, 0.8]) + 0.5U([1, 1.3]).$$

The MAD is not continuous at this $P$. Figure 17 shows the result of 5000 simulations for the $M$–estimator defined by (1.33) with $\psi$ given by (1.42) with $c = 3$. The sample size was $n = 100$ and the data were generated according to $P$ of (1.58). The deviation from normality is clear. Figure 18 shows exactly the same simulation but using simultaneous estimator defined by (1.38) and (1.39) with $\psi$ and $\chi$ given respectively by (1.42) with $c = 3$ and (1.43). Here the normal approximation is clear.

Consider a statistical functional $T$ defined on a family $\mathcal{P}_T$ of probability distributions on the Borel sets of $\mathbb{R}$. We suppose that $\mathcal{P}_T$ is convex, that is, $P_1$ and $P_2$ in $\mathcal{P}_T$ imply that $\alpha P_1 + (1 - \alpha)P_2$ also belongs to $\mathcal{P}_T$ for all $\alpha$, $0 \leq \alpha \leq 1$. We further suppose that $\mathcal{P}_T$ contains all empirical measures. We wish to quantify how much influence an observation at the point $x$ has on the value of the functional $T$. To motivate the definition consider an empirical distribution $\mathbb{P}_n$ and then add an additional observation at the points $x$. The new sample has sample size $n + 1$ and its empirical distribution is

$$\mathbb{P}_{n+1} = (1 - 1/(n + 1))\mathbb{P}_n + \delta_x/(n + 1).$$

The difference in the value of $T$ is given by $T(\mathbb{P}_{n+1}) - T(\mathbb{P}_n)$ which will in general tend to 0 as $n \rightarrow \infty$. We therefore consider the ‘derivative’. More formally we define the influence function of a statistical functional $T$ at the distribution $P$ as follows.
Figure 13: The upper panel shows the histogram of 50000 values of the mean taken from a $N(0,1)$ sample of size $n = 10$ together with the density of the normal distribution with the same median and . The lower panel shows the results for the redescending $M$–estimator defined in Section 1.11).
Figure 14: As for Figure 13 but this time with Cauchy distribution of Figure 16.

Figure 15: As for Figure 13 but this time with Gamma distribution data.
Figure 16: The density of the gamma distribution with shape parameter 3 and scale parameter 1.

Figure 17: The results of 5000 simulations of the M-estimator defined by (1.33) with $c = 3$ and data of size $n = 100$ distributed according to (1.58). The density of the normal distribution with the same mean and variance is superimposed.
Figure 18: The results of 5000 simulations of the M–estimator defined by
the simultaneous M–estimator defined by (1.38) and (1.39) with \( \psi \) and \( \chi \) given respectively by (1.42) with \( c = 3 \) and (1.43) and data of size \( n = 100 \) distributed according to (1.58). The density of the normal distribution with
the same mean and variance is superimposed.

**Definition 1.81.** If for \( P \in \mathcal{P}_T \)

\[
I(x, T, P) = \lim_{\varepsilon \to 0} \frac{T((1 - \varepsilon)P + \varepsilon \delta_x) - T(P)}{\varepsilon}.
\]

(1.59)

exists for all \( x \in \mathbb{R} \) then \( I(x, T, \mathbb{P}) \) is called the influence function of \( T \) at the
measure \( P \).

The influence function was first introduced by Hampel (1968, 1974) (see
also Hampel, Ronchetti, Rousseeuw and Stahel (1986)).

**Example 1.82.** We set \( T = T_{\text{ave}} \) and \( \mathcal{P}_T = \{ P : \int |u| \, dP(u) < \infty \} \). We have

\[
T_{\text{ave}}((1 - \varepsilon)P + \varepsilon \delta_x)) = \int u \, d((1 - \varepsilon)P + \varepsilon \delta_x)(u)
\]

\[
= (1 - \varepsilon) \int u \, dP(u) + \varepsilon \int u \, d\delta_x(u)
\]

\[
= (1 - \varepsilon)T_{\text{ave}}(P) + \varepsilon x
\]

and hence

\[
I(x, T_{\text{ave}}, P) = x - T_{\text{ave}}(P).
\]
Example 1.83. We take $T = T_{\text{med}}$ and $\mathcal{P}_T$ to be the set of all probability measures with a density $f_P$ which is continuous and strictly positive at the median of $P$. On writing $T_{\text{med}}(P) = m$ we have for $x > m$ that the median of $(1 - \varepsilon)P + \varepsilon \delta_x$ must move an amount $\gamma$ to the right where $\gamma f_P(m) \approx \varepsilon/2$. This gives

$$T_{\text{med}}((1 - \varepsilon)P + \varepsilon \delta_x)) - T_{\text{med}}(P) \approx \varepsilon/(2f_P(m))$$

and hence $I(x, T_{\text{med}}, P) = 1/f_P(m)$ if $x > m$. A corresponding result holds for $x < m$ and we finally obtain

$$I(x, T_{\text{med}}, P) = \frac{\text{sgn}(T_{\text{med}}(P) - x)}{2f_P(T_{\text{med}}(P))}. \quad (1.60)$$

The influence function of (1.59) is a so called Gâteaux derivative. It has an intuitive interpretation and aids the understanding of the behaviour of statistical functionals but is too weak to be of use in itself. A stronger form of differentiability is the so called Fréchet differentiability which is defined as follows.

Definition 1.84. Let $\mathcal{P}'$ be an open subset of $(\mathcal{P}, d_{ko})$. A statistical functional $T : (\mathcal{P}', d_{ko}) \rightarrow (\mathbb{R}, |||)$ is Fréchet differentiable at $P \in \mathcal{P}'$ if there exists a bounded measurable function $u : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$|T(Q) - T(P) - \int u(x) d(Q - P)(x)| = o(d_{ko}(Q, P)). \quad (1.61)$$

This means that for every $\varepsilon > 0$ there exists a $\gamma > 0$ such that

$$|T(Q) - T(P) - \int u(x) d(Q - P)(x)| \leq \varepsilon d_{ko}(Q, P)$$

for all $Q$ in $B(P, \gamma) = \{Q : d_{ko}(Q, P) < \gamma\}$. Consider now $\eta > 0$ and set

$$Q = Q_\eta = (1 - \eta)P + \eta \delta_x.$$ 

Then $d_{ko}(Q_\eta, P) \leq \eta$ and if $\eta \leq \gamma$ we have

$$|T((1 - \eta)P + \eta \delta_x) - T(P) - \eta u(x) + \eta \int u dP| \leq \varepsilon d_{ko}(Q_\eta, P) \leq \varepsilon \eta.$$

On letting $\eta$ tend to zero and remembering that $\varepsilon$ can be chosen to be arbitrarily small we see that

$$\lim_{\eta \rightarrow 0} \left( T((1 - \eta)P + \eta \delta_x) - T(P) \right) / \eta = u(x) - \int u dP.$$ 

57
This is however the definition of the influence function of (1.59). It follows that if $T$ is differentiable at $P$ in the sense of (1.61) then the influence function of $T$ exists at $P$ and

$$|T(Q) - T(P) - \int I(x, T, P) d(Q - P)(x)| = o(d_{ko}(Q, P)).$$

(1.62)

Although the mean and median functionals have an influence function at certain distributions $P$ they are nowhere Fréchet differentiable. We now show that M–functionals with a smooth and bounded $\psi$–function are Fréchet differentiable.

**Example 1.85.** Let $\psi$ be bounded, strictly monotone increasing with a continuous and bounded first and second derivatives and with $\psi^{(1)}$ everywhere positive. Let $T = T(P)$ be the solution of

$$\int \psi(x - T(P)) dP(x) = 0.$$

Theorem 1.46 guarantees that $T(P)$ is well–defined for all $P \in \mathcal{P}$. We first show that $T$ is continuous at $P$. If not we can find a sequence $Q_n$ of distributions with $\lim_{n \to \infty} d_{ko}(Q_n, P) = 0$ but $|T(Q_n) - T(P)| > \delta$ for some $\delta > 0$. As

$$\Psi(m) := \int \psi(x - m) dP(x)$$

is a strictly decreasing and differentiable function of $m$ and $\Psi(T(P)) = 0$ this implies that for some $\gamma > 0$

$$\gamma < |\Psi(T(Q_n))| = \left| \int \psi(x - T(Q_n)) dP(x) \right|$$

$$= \left| \int \psi(x - T(Q_n)) d(P - Q_n)(x) \right|$$

$$= \left| \int \psi^{(1)}(x - T(Q_n))(F(x) - F_n(x)) dx \right|$$

$$\leq \|\psi^{(1)}\|_{\infty} d_{ko}(Q_n, P) \to 0$$

where $F_n$ and $F$ are the distribution functions of $Q_n$ and $P$ respectively. This contradiction shows that $T$ is continuous at every distribution $P$. Let now $Q$
be such that \( d_{ko}(Q, P) \) is small. Then \(|T(Q) - T(P)|\) is small and we have

\[
0 = \int \psi(x - T(Q)) \, dQ(x) - \int \psi(x - T(P)) \, dP(x) \\
= \int \psi(x - T(P) + (T(Q) - T(P))) \, dQ(x) - \int \psi(x - T(P)) \, dP(x) \\
= \int \psi(x - T(P)) + (T(Q) - T(P)) \psi^{(1)}(x - T(P)) \, dQ(x) \\
\quad - \int \psi(x - T(P)) \, dP(x) + o(|T(Q) - T(P)|) \\
= \int \psi(x - T(P)) \, d(Q - P)(x) \\
\quad + (T(Q) - T(P)) \int \psi^{(1)}(x - T(P)) \, dQ(x) + o(|T(Q) - T(P)|).
\]

Arguing as above we have

\[
\left| \int \psi^{(1)}(x - T(P)) \, dQ(x) - \int \psi^{(1)}(x - T(P)) \, dP(x) \right| \leq \|\psi^{(2)}\|_{\infty} d_{ko}(Q, P)
\]

and hence

\[
(T(Q) - T(P))(1 + o(1)) \int \psi^{(1)}(x - T(P)) \, dP(x) = \int \psi(x - T(P)) \, d(Q - P)(x)
\]

This implies as above

\[
|(T(Q) - T(P))(1 + o(1)) \int \psi^{(1)}(x - T(P)) \, dP(x)| \leq \|\psi^{(1)}\|_{\infty} d_{ko}(Q, P)
\]

and hence as \( \psi^{(1)}(x) > 0 \) we finally conclude

\[
\left| T(Q) - T(P) - \frac{\int \psi(x - T(P)) \, d(Q - P)(x)}{\int \psi^{(1)}(x - T(P)) \, dP(x)} \right| = o(d_{ko}(Q, P)).
\]

This implies that \( T \) is Fréchet differentiable at each distribution \( P \) with influence function

\[
I(x, T, P) = \frac{\psi(x - T(P))}{\int \psi^{(1)}(u - T(P)) \, dP(u)}. \tag{1.63}
\]

**Example 1.86.** A corresponding result holds for the simultaneous M–functional of (1.40) and (1.41). If \( \psi, \chi \) and \( P \) satisfy the conditions of Theorem 1.63 and
in addition $\psi$ and $\chi$ have bounded and continuous first and second derivatives then $T_L$ and $t_S$ are Fréchet differentiable at each distribution $P$ with $\Delta(P) < 1/2$. The influence functions satisfy the following equations

\[
\psi \left( \frac{u - T_L(P)}{T_S(P)} \right) T_S(P) = I(x, T_L, P) \int \psi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) + I(x, T_S, P) \int \left( \frac{u - T_L(P)}{T_S(P)} \right) \psi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) \quad (1.64)
\]

\[
\chi \left( \frac{u - T_L(P)}{T_S(P)} \right) T_S(P) = I(x, T_L, P) \int \chi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) + I(x, T_S, P) \int \left( \frac{u - T_L(P)}{T_S(P)} \right) \chi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) \quad (1.65)
\]

which can be solved for $I(x, T_L, P)$ and $I(x, T_S, P)$. If $P$ is symmetric then the expressions simply. In this case

\[
\int \left( \frac{u - T_L(P)}{T_S(P)} \right) \psi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) = 0
\]

and

\[
\int \chi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) = 0
\]

so that

\[
I(x, T_L, P) = T_S(P) \frac{\psi \left( \frac{x - T_L(P)}{T_S(P)} \right)}{\int \psi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u)} \quad (1.66)
\]

and

\[
I(x, T_S, P) = T_S(P) \frac{\chi \left( \frac{u - T_L(P)}{T_S(P)} \right)}{\int \left( \frac{u - T_L(P)}{T_S(P)} \right) \chi^{(1)} \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u)} \quad (1.67)
\]

**Exercise 1.87.** Solve (1.64) and (1.65) for $I(x, T_L, P)$ and $I(x, T_S, P)$.
Fréchet differentiability can be used to prove central limit theorems as follows. Let \((X_n)_{1}^{\infty} = (X_n(P))_{1}^{\infty}\) be a sequence of i.i.d. random variables with distribution \(P\). We denote the empirical distribution by \(\mathbb{P}_n(P)\) and note the following inequality (see Massart (1990))

\[
\mathbb{P}(\sqrt{n}d_{ko}(\mathbb{P}_n(P), P) \geq \lambda) \leq 2 \exp(-2\lambda^2)
\]

which holds uniformly for all \(P\). In particular this implies that \(d_{ko}(\mathbb{P}_n(P), P) = O(1/\sqrt{n})\). If we use this in (1.62) with \(Q = \mathbb{P}_n(P)\) we obtain after multiplying by \(\sqrt{n}\)

\[
\sqrt{n}(T(\mathbb{P}_n(P)) - T(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (I(X_i, T, P) - \mathbf{E}(I(X_i, T, P))) + o(1) \quad (1.68)
\]

It follows from (1.61) that \(u\) is defined only up to an additive constant. Without loss of generality we can therefore assume that \(\mathbf{E}(I(X(P), T, P)) = 0\). The influence function \(I(x, T, P)\) is bounded by assumption and it follows from the Central Limit Theorem that

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(X_i, T, P) \Rightarrow N \left(0, \int I(x, T, P)^2 dP(x) \right).
\]

As the \(o(1)\) term in (1.68) tends to zero in probability we obtain

\[
\sqrt{n}(T(\mathbb{P}_n(P)) - T(P)) \Rightarrow N \left(0, \int I(x, T, P)^2 dP(x) \right). \quad (1.69)
\]

**Example 1.88.** Let \(T_L(\mathbb{P}_n(P))\) and \(T_S(\mathbb{P}_n(P))\) be the solutions of (1.38) and (1.39) where the data \(X_i(P), 1 = 1, \ldots\) are i.i.d. with common distribution \(P\) with \(\Delta(P) < 1/2\). Then

\[
\sqrt{n}(T_L(\mathbb{P}_n(P)) - T_L(P)) \Rightarrow N \left(0, \int I(x, T_L, P)^2 dP(x) \right) \quad (1.70)
\]

where \(I(x, T_L, P)\) is given by (1.64) and (1.65). This implies that \(T_L(\mathbb{P}_n(P))\) is asymptotically for all \(P\) except for those with a large atom. This contrasts with the mean which requires a finite second moment.

**Example 1.89.** It is not necessary that a functional \(T\) be Fréchet differentiable for (1.69) to hold. The mean functional \(T_{ave}\) is not Fréchet differentiable but if \(P\) has a finite second moment then (1.69) holds as \(I(x, T_{ave}, P) = x - T_{ave}(P)\) so that \(\int I(x, T, P)^2 dP(x)\) is the variance of the observations.
Example 1.90. The median is also a functional which is nowhere Fréchet differentiable but for which (1.69) holds for all $P$ with a density continuous and strictly positive at $T_{med}(P)$. We then have

$$\sqrt{n} (T_{med}(\mathbb{P}_n(P)) - T_{med}(P)) \Rightarrow N \left( 0, 1/(4f_P(T_{med}(P))^2) \right)$$

which agrees with (1.60) and (1.69). In the particular case that $P = N(0, 1)$ we have

$$\sqrt{n} T_{med}(\mathbb{P}_n(N(0, 1))) \Rightarrow N(0, \pi/2).$$

This implies that its efficiency relative to the mean is $100 \cdot 2/\pi \% = 63.7\%$ which is consistent with the first line of Table 3.

### 1.11 Redescending M–estimators II

The Hampel redescending M–estimator is not Fréchet differentiable for two reasons. Firstly the $\psi$–function of (1.55) is not differentiable and secondly it uses the MAD as an auxiliary measure of scale. The MAD is not in general continuous so that the Hampel redescending M–estimator will not in general be asymptotically normally distributed. We can over come both problems as follows. Firstly we define a redescending $\psi$–function which is similar to (1.55) but which has a continuous and second derivative. Given $0 < c_1 < c_2 < c_3$ we construct a redescending $\psi$–function as follows. Firstly we note that

$$p_1(x) = 15(1 - (x/c_1)^2)^2/(8c_1)$$

defines a polynomial whose first derivative $p_1^{(1)}$ at $x = c_1$ is zero. If we define

$$p_2(x) = \int_0^x p_1(u) \, du = 15(x/c_1 - 2x^3/(3c_1^3) + x^5/(5c_1^5))$$

then $p_0$ is an odd function of $x$ with

$$p_2(0) = 0, \quad p_2(c_1) = 1, \quad p_2^{(1)}(c_1) = p_2^{(2)}(c_1) = 0.$$

We define a third polynomial by

$$p_3(x) = 15(x^2 - c_2^2)^2 * (c_3^2 - x^2)^2$$

which satisfies

$$p_3(c_2) = p_3(c_3) = p_3^{(1)}(c_2) = p_3^{(1)}(c_3) = 0.$$

We put

$$p_4(x) = 1 - c \int_{c_2}^x p_3(u) \, du$$

62
Figure 19: The figure shows the smooth redescending $\psi$–function defined by \((1.75)\) with \(c_1 = 3, c_2 = 6, c_3 = 9\).

where

\[ c = \frac{1}{\int_{c_2}^{c_3} p_3(u) \, du} \]

from which it follows

\[ p_4(c_2) = 1, \, p_4(c_3) = 0, \, p_4^{(1)}(c_2) = p_4^{(1)}(c_3) = p_4^{(2)}(c_2) = p_4^{(2)}(c_3) = 0. \]

The redescending $\psi$–function is now defined by

\[
\psi(x) = \begin{cases} 
  p_2(x), & |x| \leq c_1, \\
  1, & c_1 < |x| \leq c_2, \\
  p_4(x), & c_2 < |x| \leq c_3, \\
  0, & |x| > c_3.
\end{cases} \tag{1.75}
\]

The default values we use are \(c_1 = 3, c_2 = 6, c_3 = 9\). Figure 19 shows the resulting function. We require a Fréchet differentiable auxiliary measure of scale. The one we choose is scale measure of the joint location and scale functionals of \((1.40)\) and \((1.41)\) with $\psi$ and $\chi$ given by \((1.42)\) and \((1.43)\) with \(c = 5\). Corresponding to \((1.57)\) we have the following relative efficiencies

\[ 92, 94, 97, 100, 101, 97, 94 \text{ and } 81. \tag{1.76} \]
The asymptotic efficiencies can be calculated exactly using (1.66). For the 
\( P = \Phi = N(0, 1) \) distribution we have
\[
\sqrt{n} T_L(P_n(\Phi)) \Rightarrow N \left( 0, \int I(x, T_L, \Phi)^2 d\Phi(x) \right)
\]
(1.77)
with
\[
I(x, T_L, \Phi) = T_S(\Phi) \frac{\psi \left( \frac{x}{T_S(\Phi)} \right)}{\int \psi^{(1)} \left( \frac{u}{T_S(\Phi)} \right) d\Phi(u)}.
\]
(1.78)
We have \( T_S(\Phi) = 0.647155 \) and evaluating the integral gives
\[
\int \psi \left( \frac{x}{T_S(\Phi)} \right)^2 \varphi(x) dx / \left( \int \psi^{(1)} \left( \frac{x}{T_S(\Phi)} \right) \varphi(x) dx \right)^2 = 2.56.
\]
(1.79)
On noting that \( 0.647155 \sqrt{2.56} = 1.07436 \) it follows that
\[
\sqrt{n} T_L(P_n(\Phi)) \Rightarrow N(0, 1.07436)
\]
(1.80)
which implies an asymptotic efficiency of \( 1/1.07436 = 0.931 \) which is consistent with (1.76).

1.12 Confidence intervals

Point estimates such as 2.016 for the copper data of Example 1.1 are in themselves of little use unless they are accompanied by some indication of the possible range of values. This is accomplished by giving a confidence interval. The standard confidence interval for data such as that of the copper data is the one derives from the t–distribution. A \( \alpha \)–confidence interval is given by
\[
\left[ \text{mean}(x_n) - qt((1 + \alpha)/2, n - 1)\text{sd}(x_n)/\sqrt{n}, \right.
\]
\[
\left. \text{mean}(x_n) + qt((1 + \alpha)/2, n - 1)\text{sd}(x_n)/\sqrt{n} \right]
\]
If we set \( \alpha = 0.95 \) then the confidence interval for the copper content of the water sample is \([1.970, 2.062]\). The confidence interval based on the t–distribution is however susceptible to outliers and deviations from normality. If we replace the last observation 2.13 by 21.3 the confidence interval is now \([1.257, 4.1950]\). A simple but effective way of avoiding this is to use the Hampel 5.2MAD rule. All observation which are further than 5.2MAD (3.5MAD with the R–default version of the MAD) from the median are eliminated and
Figure 20: The figure shows the bias of the smooth redescending M–estimator with \( \psi \)-function given by (1.75) with \( c_1 = 3 \), \( c_2 = 6 \), \( c_3 = 9 \) and that of the median for the contamination model (1.54) as a function of \( x \).

the confidence interval is calculated using the remaining observations. If we do this for the sample with 2.13 replaced by 21.3 the confidence interval is \([1.965, 2.058]\) which agrees well with confidence interval for the initial sample. Eliminating larger observations and then ignoring this will cause the coverage probability to be somewhat lower than the nominal one of 0.95. This can be taken into account by increasing the factor \( q_t((1 + \alpha)/2, n - 1)sd(x_n)/\sqrt{n} \). Simulations for \( n = 27 \) show that it must be multiplied by 1.05. Table 5 shows the results for the standard confidence interval based on the t–distribution for a sample size \( n = 27 \). The first lines show the coverage probability for the distributions \( t_k, k = 1, \ldots, 5 \) and the \( N(0,1) \) distribution. the second lines shows the mean lengths of the confidence intervals. Lines 3 and 4 show the corresponding results for the Hampel 5.2 method calibrated for the normal distribution, \( n = 27 \) and a coverage probability of 0.95. Lines 5 and 6 show the corresponding results for the joint M–estimator describes in the last section, again calibrated for the normal distribution, \( n = 27 \) and a coverage probability of 0.95. Having to perform the simulations to calibrate the confidence interval for every value of \( n \) can be avoided by using the asymptotic behaviour which can be calculated and then performing simulations to decide from which value of \( n \) onwards the asymptotic result is sufficiently accurate.
Table 5: The first line shows the coverage probabilities for the standard confidence interval based on the t–distribution with a nominal covering probability of 0.95. The second line shows the mean lengths of the intervals. Lines 3 and 4 show the corresponding results for the Hampel 5.2 method calibrated for the normal distribution, n = 27 and a coverage probability of 0.95. Lines 5 and 6 show the corresponding results for the joint M–estimator described in the last section, again calibrated for the normal distribution, n = 27 and a coverage probability of 0.95.

This was done in Section 1.6.3 for developing a procedure for identifying outliers. We repeat it here for confidence intervals. We know from (1.77), (1.78) and (1.79) that a 0.95 confidence interval for a sample of size n will be of the form

\[
\left[ T_L(\mathbb{P}_n) - \text{fac}(n) \cdot 1.96 \cdot 1.6 T_S(\mathbb{P}_n)/\sqrt{n}, \right.
\]

\[
T_L(\mathbb{P}_n) + \text{fac}(n) \cdot 1.96 \cdot 1.6 T_S(\mathbb{P}_n)/\sqrt{n},
\]

where \( \lim_{n \to \infty} \text{fac}(n) = 1 \). The factor 1.96 is \( \Phi^{-1}(0.975) \) and the factor 1.6 is simpl\[ \sqrt{2.56} \] of (1.79). As we expect the behaviour to be somewhat worse than that based on the t–distribution we replace 1.96 by \( q_t(0.975, n-1) \).

\[
\left[ T_L(\mathbb{P}_n) - \text{fac}(n) \cdot q_t(0.975, n-1) \cdot 1.6 T_S(\mathbb{P}_n)/\sqrt{n}, \right.
\]

\[
T_L(\mathbb{P}_n) + \text{fac}(n) \cdot q_t(0.975, n-1) \cdot 1.6 T_S(\mathbb{P}_n)/\sqrt{n},
\]

With this alteration Figure 21 shows the values of \( \log(\text{fac}(n) - 1) \) plotted against those of \( \log(n) \). We approximate the values by a line of the form

\[ \log(\text{fac}(n) - 1) = a + b \log n + c/\log n \]

and determine the values of \( a, b \) and \( c \) using the linear model fit provided by R.
> b<-lm(log(tmp$res-1)~I(log(tmpn))+I(1/log(tmpn)))
> summary(b)

Call:
  lm(formula = log(tmp$res - 1) ~ I(log(tmpn)) + I(1/log(tmpn)))

Residuals:
  Min      1Q  Median      3Q     Max
-0.143204 -0.011041  0.002116  0.025020  0.083480

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.60421    0.11072  -5.457  1.01e-05 ***
I(log(tmpn)) -0.84637    0.02148 -39.405 < 2e-16 ***
I(1/log(tmpn)) 3.14815    0.12931  24.345 < 2e-16 ***
---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04544 on 26 degrees of freedom
Multiple R-Squared: 0.9988,  Adjusted R-squared: 0.9987
F-statistic: 1.115e+04 on 2 and 26 DF,  p-value: < 2.2e-16

The Figure 21 shows the resulting fit. The approximation is given by

\[
\text{fac}(n, 0.95) = 1 + \exp(-0.601 - 0.846 \log n + 3.148 / \log n). \tag{1.81}
\]

The resulting 0.95–confidence interval for the copper data is \([1.984, 2.076]\) compared to that based on the \(t\)-distribution which is \([1.970, 2.062]\). If we replace the last observation by 21.3 the confidence interval becomes \([1.982, 2.078]\) compared with \([1.257, 4.1950]\).

**Remark 1.91.** From (1.70) and (1.66) we know that for an i.i.d. sample \(X_n\) with common distribution \(P\)

\[
\sqrt{n} (T_L(P_n(P)) - T_L(P)) \Rightarrow N \left( 0, \int I(x, T_L, P)^2 dP(x) \right)
\]
Figure 21: The figure shows the estimated values of \( \log(\text{fac}(n) - 1) \) obtained by simulations plotted against the values of \( \log(n) \). The lines show the approximating function of (1.81).

where

\[
I(x, T_L, P) = T_S(P) \frac{\psi \left( \frac{x - T_L(P)}{T_S(P)} \right)}{\int \psi(1) \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u)}
\]

As \( T_S(P_n(P)) \) converges to \( T_S(P) \) it follows that

\[
\sqrt{n} \frac{(T_L(P_n(P)) - T_L(P))}{T_S(P_n(P))} \Rightarrow N(0, \Sigma(P)^2)
\]

where

\[
\Sigma(P)^2 = \frac{\int \psi \left( \frac{u - T_L(P)}{T_S(P)} \right)^2 dP(u)}{\left( \int \psi(1) \left( \frac{u - T_L(P)}{T_S(P)} \right) dP(u) \right)^2}
\]

which of course depends on the underlying \( P \). One possibility is to replace \( \Sigma(P) \) by \( \Sigma(P_n(P)) \) which will then automatically adapt to the underlying \( P \) as \( d_{ko}(P_n(P), P) \to 0 \). Fortunately the dependence of \( \Sigma(P) \) on \( P \) is not very strong as long as \( P \) is close to the normal distribution. For the redescending \( \psi \)-function the values of \( \Sigma(P) \) for the uniform, the Gaussian, the Laplace
and the Cauchy distribution are respectively

\[ 2.071, 2.563, 2.972, 2.766 \]

and their square roots deviate by even less

\[ 1.439, 1.601, 1.724, 1.663. \]

In consequence we simply use the value which derives from the Gaussian distribution.
2 The One-Way Table

2.1 The two sample \( t \)-test

The following example is Exercise 5 on page 65 of Miller (1986). Sodium citrate is added to blood to prevent coagulation (blood transfusions etc.) By adding the protein thromboplastin the blood can be made to coagulate again. In standardized tests the average time for the blood with added sodium nitrate to coagulate after the addition of thromboplastin is about 38 seconds. In a study to investigate the effectiveness of streptokinase to dissolve blood clots the the partial thromboplastin times (PTT) for patients whose clot dissolved (R) and those whose clot did not dissolve (NR) were measures with the following results.

<table>
<thead>
<tr>
<th></th>
<th>R: 41 86 90 74 146 57 78 55</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R: 105 46 94 26 101 72 119 88</td>
</tr>
<tr>
<td></td>
<td>NR: 34 23 36 25 35 23 87 48</td>
</tr>
</tbody>
</table>

Table 6: PTT times for patients whose clot dissolved (R, \( n_1 = 16 \)) and those whose clot did not dissolve (NR, \( n_2 = 8 \))

![Figure 22: The boxplot of the data of Table 6.](image)

We may use a \( t \)-test to test the hypothesis that there is no significant difference between the means of the two sample. Assuming independent
normal samples and equal variances the test statistic is

\[
T = \frac{(\text{mean}(x_{1n_1}) - \text{mean}(x_{2n_2}))}{\sqrt{\left(\frac{(n_1-1)sd(x_{1n_1})^2 + (n_2-1)sd(x_{2n_2})^2}{n_1+n_2-2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}
\]  

(2.1)

which follows the \(t\)-distribution with \(n_1 + n_2 - 2 = 16 + 8 - 2 = 24\) degrees of freedom. With \(x_{1n_1}\) denoting the R-sample the value of the test statistics is 3.612347 which for a two-sided test on 24 degrees of freedom gives a p-value of

\[2\text{pt}(3.612347, 24) = 0.0014\]

which is highly significant. Of course one might wish to test assumption of the equality of the variances. If we use the \(F\)-test

\[F = \frac{sd(x_{1n_1})^2}{sd(x_{2n_2})^2}\]  

(2.2)

which under the assumption of independent normal samples and equality of variances follows the \(F\)-distribution with \((n_1 - 1, n_2 - 1)\)-degrees of freedom. For the data for Table 6 the value of the test statistic is 0.4356. The 0.025- and 0.975-quantiles of the \(F\)-distribution are 0.2453 and 3.1248 respectively so that the hypothesis of equality of variances is not rejected. A glance at the boxplot of Figure 22 shows however a large observation 87. If we use the Hampel 5.2 rule it would just not be eliminated as an outlier but remove it in any case to see its effect on the result. The new sample has \(n_2 = 7\) and a standard deviation of 8.484 against 14.826 for the original sample. If we test for equality of variances the value of the test statistic (2.2) is now 0.07948 on \((7, 16)\) degrees of freedom. The corresponding p-value is

\[2\text{pf}(0.07948, 7, 16) = 0.00235\]

so that now the assumption of equality of variances is not consistent with the data. There is no general agreement about how to proceed when the variances are different. One popular procedure is to use

\[t = \frac{\text{mean}(x_{1n_1}) - \text{mean}(x_{2n_2})}{\sqrt{\frac{sd(x_{1n_1})^2}{n_1} + \frac{sd(x_{2n_2})^2}{n_2}}}\]  

(2.3)

and treat it as a \(t\)-statistic with \(\nu\) degrees of freedom where

\[\min(n_1 - 1, n_2 - 1) \leq \nu \leq n_1 + n_2 - 2\]

72
with one suggestion being
\[
\nu = \frac{\left(\frac{\text{sd}(x_{1n_1})^2}{n_1} + \frac{\text{sd}(x_{2n_2})^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{\text{sd}(x_{1n_1})^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{\text{sd}(x_{2n_2})^2}{n_2}\right)^2}.
\] (2.4)

After eliminating the observation 88 the value of the statistic \(T\) of (2.3) is 5.886 and the value of \(\nu\) is 20.52. This gives a \(p\)-value of 4.198707 \(\cdot 10^{-6}\). The average time for the blood to coagulate is about \(\mu = 38\) seconds. We may wish to test the hypothesis that the NR–patients are no different from the average. The test statistic is
\[
t = \frac{\sqrt{n} (\text{mean}(x_n) - \mu)}{\text{sd}(x_n)}.
\] (2.5)

The values for the NR–data are 0.0671 and -1.875 with and without the observation 88 respectively and the degrees of freedom are 8 and 7 respectively. In neither case is the result significant so that we can conclude that the NR–data at least in this respect do not differ from average data. There are several aspects of this analysis which are worth mentioning. Firstly we have now carried out six tests which although not excessive and in this particular case not problematic does in general carry the risk of finding a spurious effect. The more tests one conducts the more likely it is that a significant result will be found just by chance. Somehow we must explicitly allow for multiple testing. Secondly when dealing with more than one sample we must allow for different variances and also to some extent different distributions. Finally we must take outliers into account. These are the main topics of the following sections.

### 2.2 The \(k\) sample problem

We are given \(k\) samples of possibly different sizes and which were obtained under different conditions, for example, different forms of treatment for some disease or for purifying water. We are interested in differences in the location values of the samples and, in particular, in quantifying the differences and deciding whether or not they can be explained by chance. In the past \(k\) was a relatively small number but with the advent of gene expression data \(k\) may now be of the order of tens of thousands. The problem is that it becomes more difficult to recognize a signal the larger the number of samples is. Thus is shown by Figure 23. We consider \(k\) \(N(0, 1)\)–samples of size \(n = 5\) but add a signal of \(\mu = 1\) to the values of the first sample. For each sample we calculate the absolute value of the \(t\)–statistic \(T_i = \sqrt{nX_{i,5}}/\text{sd}(X_{i,5})\).
Figure 23: The upper panel shows the percentage of simulations for which the first sample had the highest \( t \)-statistic as a function of the number of samples \( k \). The lower panel shows the proportion of samples with a higher absolute \( t \)-value than first sample expressed as a percentage.

The upper panel of Figure 23 shows the percentage of simulations for which the first sample had the highest absolute \( t \)-statistic plotted against the number of sample. For \( k = 50 \) the first \( t \)-statistic has the largest absolute value in only 20% of the cases. The lower panel shows the proportion of samples with a higher absolute \( t \)-value than first sample expressed as a percentage. For \( k = 50 \) it shows that on average about 7 of the samples have a \( t \)-statistic whose absolute value is larger than that for the first sample. Clearly if \( k = 50000 \) the signal will have to be very large before it can be distinguished from the noise. Clearly even for small \( k \) the number of samples has to be taken into account when analysing the data.

2.3 The classical method

The Analysis of Variance goes back to R. A. Fisher who developed the idea whilst working at the Rothamsted Experimental Station in the 1920s.

http://www.rothamsted.bbsrc.ac.uk/

His concern was yields of wheat and their dependence on factors such as the field, the fertilizer and the variety of wheat. We consider the simplest
situation where the same wheat with the same fertilizer was grown on different fields. The question to be answered is whether the different location, perhaps as a proxy for different soils or exposure to the sun had a significant effect on the yield. We have \( k \) samples corresponding to the \( k \) fields and we denote them by \( x_{i,n_i} = (x_{i,1}, \ldots, x_{i,n_i}), \ i = 1, \ldots, k \). One can imagine the \( i \)th field being split up into \( n_i \) small plots and the yield in each plot being measured. Fisher’s idea was to decompose the sums of squares of the yields into three components. The first represents the average yield over all fields, the second the differences in the yields between the fields and the third the differences in yields between the plots. On writing \( n = \sum_{i=1}^{k} n_i \) Table 7 is

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sums of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average yield</td>
<td>1</td>
<td>( n \bar{x}_n^2 )</td>
</tr>
<tr>
<td>Between fields</td>
<td>( k - 1 )</td>
<td>( \sum_{i=1}^{k} n_i (\bar{x}_{i,n_i} - \bar{x}_n)^2 )</td>
</tr>
<tr>
<td>Within fields</td>
<td>( n - k )</td>
<td>( \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,n_i})^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( n )</td>
<td>( \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j}^2 )</td>
</tr>
</tbody>
</table>

Table 7: The classical one-way table

just a mathematical identity which may be proved as follows.

\[
\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_n)^2 + n \bar{x}_n^2 \\
\quad = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,n_i} + (\bar{x}_{i,n_i} - \bar{x}_n))^2 + n \bar{x}_n^2 \\
\quad = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,n_i})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_{i,n_i} - \bar{x}_n)^2 + n \bar{x}_n^2. 
\]

Even though this is an identity it can give insight into the behaviour of the wheat on different fields. Fisher however went further and proposed modelling such data as follows

\[
X_{i,j} = \mu_i + \sigma \epsilon_{i,j}, \ j = 1, \ldots, n_i, \ i = 1, \ldots, k \]  

(2.6)
where the $\epsilon_{i,j}$ are standard Gaussian white noise. Under these assumptions it can be shown that the random variables
\[
nX^2_n, \quad \sum_{i=1}^{k} n_i (\overline{X}_{i,n_i} - \overline{X}_n)^2, \quad \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_{i,n_i})^2
\]
are independently distributed according to non-central $\chi^2$–distributions. Suppose now that the means $\mu_i$ are all equal. In this case
\[
\sum_{i=1}^{k} n_i (\overline{X}_{i,n_i} - \overline{X}_n)^2 \quad \text{and} \quad \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_{i,n_i})^2
\]
obey a central $\chi^2$–squared distribution with $k-1$ and $n-k$ degrees of freedom respectively. It follows that
\[
F_{k-1,n-k} = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{X}_{i,n_i} - \overline{X}_n)^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_{i,n_i})^2}
\]
obyes an $F$–distribution with $k - 1$ and $n - k$ degrees of freedom. This enables us to reject the null hypothesis of equality of means if
\[
\frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{X}_{i,n_i} - \overline{X}_n)^2 \geq qf (0.95, k-1, n-k).
\]

**Example 2.1.** We now apply the procedure to three data sets to be found on page 114 from Miller (1986). They are as follows:

\[\begin{align*}
x_{1,23} &= (5.37, 5.80, 4.70, 5.70, 3.40, 8.60, 7.48, 5.77, 7.15, 6.49, 4.09, 5.94, \\
&\quad 6.38, 9.24, 5.66, 4.53, 6.51, 7.00, 6.20, 7.04, 4.82, 6.73, 5.26),
\end{align*}\]

\[\begin{align*}
x_{2,17} &= (3.96, 3.04, 5.28, 3.40, 4.10, 3.61, 6.16, 3.22, \\
&\quad 7.48, 3.87, 4.27, 4.05, 2.40, 5.81, 4.29, 2.77, 4.40),
\end{align*}\]

\[\begin{align*}
x_{3,28} &= (5.37, 10.60, 5.02, 14.30, 9.90, 4.27, 5.75, 5.03, \\
&\quad 5.74, 7.85, 6.82, 7.90, 8.36, 5.72, 6.00, 4.75, 5.83, 7.30, 7.52, 5.32, \\
&\quad 6.05, 5.68, 7.57, 5.68, 8.91, 5.39, 4.40, 7.13)
\end{align*}\]

(2.7)

The data give plasma bradykininogen levels measured in control patients $x_{1,23}$, in patients where Hodgkin’s disease is active $x_{2,17}$ and in patients where the disease is inactive $x_{3,28}$.
The means and standard deviations of the three samples are respectively 6.081, 4.242, 6.791 and 1.362, 1.302, 2.176. The overall mean is 5.914. From this we see that

\[ \sum_{i=1}^{k} n_i (\bar{x}_{i,n_i} - \bar{x}_n)^2 = 69.73 \quad \text{and} \quad \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,n_i})^2 = 195.88. \]

and hence

\[ F_{k-1,n-k} = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_{i,n_i} - \bar{x}_n)^2 \frac{n}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,n_i})^2 = 34.87/3.01 = 11.57. \]

In this situation we are only concerned with large values of the test statistic so we use a one–sided test. The 0.95 quantiles of the \(F\)–distribution with 2 and 65 degrees is 3.138142. Clearly the attained value lies outside of this range so that the null hypothesis of equality of means is rejected.

The \(F\)–distribution is an \(L^2\)–statistic, that is, it is based on quadratic differences which result from optimality considerations deriving from the proposed model (2.6). The use of \(L^2\)–methods was criticized in Chapter 1 and these criticisms remain valid here, namely that if the data have a somewhat heavier tail than the normal distribution the power of the test to detect differences in the means can decline sharply. Even if the data are normally
distributed it is not clear what to do if the variance differs from sample to sample. Finally even if we stick to the model (2.6) the test is not satisfactory. If the null hypothesis is accepted then there is little to be said at this level, but if the null hypothesis is rejected then all we can conclude is that the means are different without being able to specify which is different from what. We have just seen this with the data on Hodgkin’s disease. In many cases we are explicitly interested on quantifying the differences, say between the control patients and those with active Hodgkin’s disease, and the $F$-test does not accomplish this. This leads naturally to the idea of multiple comparisons.

### 2.4 Multiple comparisons

#### 2.4.1 Tukey: Honest Significant Difference

We suppose initially that the design is balanced, that is $n_1 = \ldots = n_k = m$. The direct comparison of the means $\mu_i$ and $\mu_j$ leads naturally to the statistic

$$
\sqrt{m/2} \frac{(\overline{X}_{i,m} - \overline{X}_{j,m})}{S_{n-k}}
$$

where

$$
S_{n-k}^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{i,j} - \overline{X}_{i,n_i})^2.
$$

Under the model (2.6) the statistic of (2.8) follows the $t$ distribution with $n - k$ degrees of freedom. As we are interested in all pairwise comparisons this would lead to the statistic

$$
\sqrt{m/2} \max_{1 \leq i < j \leq k} \frac{|\overline{X}_{i,m} - \overline{X}_{j,m}|}{S_{n-k}}
$$

(2.9)

However in the case of pairwise comparisons the factor $2$ in $m/2$ is dropped and incorporated into the definition of the distribution. The so called studentized range distribution $tr_{k,\nu}$ is defined as the distribution of

$$
\frac{\max_{1 \leq i \leq k} X_i - \min_{1 \leq j \leq k} X_j}{\sqrt{\chi_{\nu}^2/\nu}}
$$

(2.10)

where $X_i, 1 \leq i \leq k$ are i.i.d $N(0,1)$ and $\chi_{\nu}^2$ an independent $\chi^2$-random variable with $\nu$ degrees of freedom. The quantiles of this distribution are available in $\mathbb{R}$ as

$$
qtukey(p, k, \nu).
$$

78
Under the null hypothesis of equal means in model (2.6) it follows that

\[
\sqrt{m} \max_{1 \leq i < j \leq k} \frac{|\overline{X}_{i,m} - \overline{X}_{j,m}|}{S_{n-k}} = \frac{\max_{1 \leq i < j \leq k} \sqrt{m} \overline{X}_{i,m} - \min_{1 \leq i < j \leq k} \sqrt{m} \overline{X}_{j,m}}{S_{n-k}}
\]

follows the \( t_{k,n-k} \) distribution. The null hypothesis of no difference between the means is rejected at level \( \alpha \) if

\[
\sqrt{m} \max_{1 \leq i < j \leq k} \frac{|\overline{X}_{i,m} - \overline{X}_{j,m}|}{S_{n-k}} > \text{qtukey}(1 - \alpha, k, n - k). \tag{2.11}
\]

We note that this test allows us to identify which pairs of sample differ in their means. It is precisely those pairs \((i, j)\) for which

\[
\sqrt{m} \frac{|\overline{X}_{i,m} - \overline{X}_{j,m}|}{S_{n-k}} > \text{qtukey}(\alpha, k, n - k). \tag{2.12}
\]

So far we have assumed that the sample sizes of the \( k \) samples are equal. If this is not so we replace (2.12) by

\[
\sqrt{m} \frac{|\overline{X}_{i,m} - \overline{X}_{j,m}|}{S_{n-k}} > \text{qtukey}(1 - \alpha, k, n - k) \left( \frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right) \right)^{1/2}. \tag{2.13}
\]

It has been shown that this test is conservative, that is the probability that the null hypothesis will not be rejected if true is at least \( \alpha \). The tests (2.12) and (2.13) can be inverted to give confidence intervals for the differences \( \mu_i - \mu_j \). For the test (2.12) we obtain as a 100\( \alpha \)%-confidence interval for \( \mu_i - \mu_j \)

\[
\left[ \overline{X}_{i,m} - \overline{X}_{j,m} - S_{n-k} \text{qtukey}(\alpha, k, n - k)/\sqrt{m}, \right. \\
\left. \overline{X}_{i,m} - \overline{X}_{j,m} + S_{n-k} \text{qtukey}(\alpha, k, n - k)/\sqrt{m} \right]. \tag{2.14}
\]

A corresponding interval for the case of unequal sample sizes can be derived from (2.13).

Sometimes more complicated hypotheses than equality of means may be tested. If for example the sample sizes are equal and two groups \( G_1 \) and \( G_2 \) of samples have special interest we may be interested in testing whether their overall means differ. This leads to the null hypothesis

\[
\frac{1}{|G_1|} \sum_{i \in G_1} \mu_i = \frac{1}{|G_2|} \sum_{j \in G_2} \mu_j
\]
which may be written in the form
\[
\frac{1}{|G_1|} \sum_{i \in G_1} \mu_i - \frac{1}{|G_2|} \sum_{j \in G_2} \mu_j = 0.
\]

Even more generally we may write
\[
\sum_{i=1}^{k} c_i \mu_i = 0 \quad \text{where} \quad \sum_{i=1}^{k} c_i = 0.
\]

**Definition 2.2.** A contrast is a set of \( k \) numbers \( c_1, \ldots, c_k \), not all zero, such that \( \sum_{i=1}^{k} c_i = 0 \).

**Theorem 2.3.** Let \((c_i)_1^n\) be real numbers with \( \sum_{i=1}^{n} c_i = 0 \). Then for any real numbers \((x_i)_1^n\) we have
\[
\left| \sum_{i=1}^{n} c_i x_i \right| \leq \frac{1}{2} \max_{1 \leq i < j \leq n} |x_j - x_i| \sum_{i=1}^{n} |c_i|.
\]

**Proof.** We have
\[
\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i^+ x_i - \sum_{i=1}^{n} c_i^- x_i.
\]
where \( c_i^+ = \max(0, c_i) \) and \( c_i^- = \max(0, -c_i) \). As \( \sum_{i=1}^{n} c_i = 0 \) we have \( \sum_{i=1}^{n} c_i^+ = \sum_{i=1}^{n} c_i^- \). Let \( c = \min\{c_i^+ > 0, c_i^- > 0\} \) and suppose without loss of generality that this is \( c_j^+ \). Then we choose \( c_k^- > 0 \) and we can write
\[
\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i^+ x_i - \sum_{i=1}^{n} c_i^- x_i = c_j^+ (x_j - x_k) + \sum_{i=1}^{n} c_i^+ x_i - \sum_{i=1}^{n} c_i^- x_i
\]
where
\[
\sum_{i=1}^{n} \tilde{c}_i^+ = \sum_{i=1}^{n} \tilde{c}_i^- = \sum_{i=1}^{n} c_i^+ - c_i^- < \sum_{i=1}^{n} c_i^+
\]
We can now repeat this process with the \( \tilde{c}_i \) and finally we have
\[
\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} \gamma_i (x_i - x_{j(i)})
\]
for some non-negative numbers \( \gamma_i \) with \( \sum_{i=1}^{n} \gamma_i = \sum_{i=1}^{n} c_i^+ \). This implies
\[
\left| \sum_{i=1}^{n} c_i x_i \right| = \left| \sum_{i=1}^{n} \gamma_i (x_i - x_{j(i)}) \right| \leq \max_{1 \leq i < j \leq n} |x_j - x_i| \sum_{i=1}^{n} \gamma_i = \max_{1 \leq i < j \leq n} |x_j - x_i| \sum_{i=1}^{n} c_i^+.
\]
The claim follows on noting
\[ \sum_{i=1}^{n} |c_i| = \sum_{i=1}^{n} c_i^+ + \sum_{i=1}^{n} c_i^- = 2 \sum_{i=1}^{n} c_i^+. \]

Using this result we can adapt Tukey’s test to contrasts as follows. We have
\[ \sum_{i=1}^{k} c_i \bar{X}_{i,m} = \sum_{i=1}^{k} c_i (\bar{X}_{i,m} - \bar{X}_{j,m}) \]
for any \( j \) as \( \sum_{i=1}^{k} c_i = 0 \). This gives
\[ \left| \sum_{i=1}^{k} c_i \bar{X}_{i,m} \right| \leq \sum_{i=1}^{k} |c_i| |\bar{X}_{i,m} - \bar{X}_{j,m}| \leq \frac{1}{2} \max_{1 \leq i < j \leq k} |\bar{X}_{i,m} - \bar{X}_{j,m}| \sum_{i=1}^{k} |c_i|. \]

From this we may deduce that a conservative \( \alpha \)-test
\[ \sqrt{m/2} \left| \sum_{i=1}^{k} c_i \bar{X}_{i,m} \right| > q_{\text{tukey}}(1 - \alpha, k, n - k) \sum_{i=1}^{k} |c_i|/2 \tag{2.15} \]
or a \( \alpha \)-confidence interval for \( \sum_{i=1}^{k} c_i \mu_i \)
\[ \left[ \sum_{i=1}^{k} c_i \bar{X}_{i,m} - S_{n-k} q_{\text{tukey}}(\alpha, k, n - k) \frac{1}{2} \sum_{i=1}^{k} |c_i|/\sqrt{m}, \right. \]
\[ \left. \sum_{i=1}^{k} c_i \bar{X}_{i,m} + S_{n-k} q_{\text{tukey}}(\alpha, k, n - k) \frac{1}{2} \sum_{i=1}^{k} |c_i|/\sqrt{m} \right]. \tag{2.16} \]

We note that in the case \( c_i = 1, c_j = -1 \) and the rest of the \( c_i = 0 \) the test reduces to testing \( \mu_i = \mu_j \) using Tukey’s test.

### 2.4.2 Bonferroni and Holm

If \( m \) statistical tests are carried out under the corresponding null hypotheses and all at the level \( \alpha \) then for large \( m \) about \( m\alpha \) of these tests will result in a significant value of the test statistic and the null hypotheses will be rejected. If sufficiently many tests are carried out at a fixed level then there will be
one significant result just by chance. A general method of adjusting for the number of tests carried is based on the simple inequality

\[ P(\bigcup_{i=1}^{m} F_i) \leq \sum_{i=1}^{m} P(F_i) \]

which holds without restrictions on the \( F_i \). Let \( F_i \) denote the event that the \( i \)th null hypothesis \( H_i \) is rejected given that all the null hypotheses are true. Then we can guarantee that the probability of falsely rejecting at least one is at most \( \alpha \) by testing each individual hypothesis at the level \( \alpha/m \). We then have

\[ P(\bigcup_{i=1}^{m} F_i) \leq \sum_{i=1}^{m} P(F_i) \leq \sum_{i=1}^{m} \alpha/m = \alpha. \]

As an example we consider Tukey’s least honest difference test. If we test all pairs of means being equal we have \( k(k - 1)/2 \) null hypotheses. Each single test is at the level \( \beta \) is of the form

\[ \sqrt{\frac{m}{2}} \frac{|\bar{X}_{i,m} - \bar{X}_{j,m}|}{S_{n-k}} > q_t(1 - \beta/2, n - k) \]

and by setting \( \beta = \alpha/(k(k - 1)/2) = 2\alpha/(k(k - 1)) \). This results in the tests

\[ \sqrt{\frac{m}{2}} \frac{|\bar{X}_{i,m} - \bar{X}_{j,m}|}{S_{n-k}} > q_t(1 - \alpha/(k(k - 1)), n - k). \]

Just as with Tukey’s least honest difference test it is guaranteed that the probability of at least one type 1 error is at most \( \alpha \). It can be shown that the critical points for the test statistics are larger than those of the Tukey test which is therefore more powerful. However the Bonferroni method is more general.

Holm proposed a sequential procedure which leads to more powerful tests than the Bonferroni method and is equally generally applicable. For each of the test statistics \( T_i \) we calculate the attained \( p \)-value \( p_i \) which is defined as follows. Suppose that under the \( i \)th null hypothesis the test statistic \( T_i \) has the distribution function \( F_i \). Then we define \( p_i = 1 - F_i^{-1}(T_i) \) which is the attained level of the tests. We order the \( p_i \) from smallest to largest \( p_1(p_2) \leq \ldots \leq p_m \). To ease the notation we suppose the tests are labeled in this order. The procedure is as follows. If \( p_{(1)} > \alpha/m \) then all null hypotheses are accepted. If \( p_{(1)} \leq \alpha/m \) then the corresponding null hypothesis is rejected and we consider the next value \( p_{(2)} \) which is compared with \( \alpha/(m - 1) \). If \( p_{(2)} > \alpha/(m - 1) \) then all remaining null hypotheses are
accepted. If \( p(2) < \alpha/(m - 1) \) then the corresponding hypothesis is rejected and we move on to \( p(3) \) which is compared with \( \alpha/(m - 2) \). If \( p(3) > \alpha/(m - 2) \) then all remaining null hypotheses are accepted. If \( p(3) < \alpha/(m - 2) \) then the corresponding hypothesis is rejected and we move on to \( p(4) \) which is compared with \( \alpha/(m - 3) \). This is continued until either all null hypotheses are rejected or for some \( i \) we have \( p(i) > \alpha/(m - i + 1) \) in which case this and all remaining hypotheses are accepted. It can be shown that this method also guarantees that the probability of at least one type 1 error is at most \( \alpha \) but it is clearly more powerful than the Bonferroni method. One point in favour of the Bonferroni method is that it can provide confidence intervals whereas the Holm method cannot.

2.4.3 Scheffé’s method

Tukey’s test for contrasts is only exact in the case that all the \( c_i \) are zero apart from two values which are +1 and -1. In all other cases the test is conservative but loses power. Scheffé’s method was specially developed to deal with contrasts. We restrict ourselves to the case of a balanced design where all the sample have the same sample size \( m \). We consider a contrast \( \mathbf{c} = (c_1, \ldots, c_k) \) which, without loss of generality we may take to have length 1, \( \|\mathbf{c}\| = \sum_{i=1}^{k} c_i^2 = 1 \). The test statistic we use is

\[
\sum_{i=1}^{k} c_i \bar{X}_{i,m} = \sum_{i=1}^{k} c_i (\bar{X}_{i,m} - \bar{X}_m).
\]

If we now maximize this over all possible contrasts of length 1 we have

\[
\max_{\mathbf{c}, \|\mathbf{c}\|=1} \left| \sum_{i=1}^{k} c_i (\bar{X}_{i,m} - \bar{X}_m) \right| = \left( \sum_{i=1}^{k} (\bar{X}_{i,m} - \bar{X}_m)^2 \right)^{1/2}
\]

with the maximizing contrast being given by

\[
c_i = \frac{\bar{X}_{i,m} - \bar{X}_m}{\left( \sum_{i=1}^{k} (\bar{X}_{i,m} - \bar{X}_m)^2 \right)^{1/2}}
\]

which is indeed a contrast as

\[
\sum_{i=1}^{k} (\bar{X}_{i,m} - \bar{X}_m) = 0.
\]
From this we see that

\[
P\left( \max_{\|c\|=1} \left| \sum_{i=1}^{k} c_i (\bar{X}_{i,m} - \bar{X}_m) \right| / S_{n-k} \geq K \right)
= P\left( \sum_{i=1}^{k} \frac{(\bar{X}_{i,m} - \bar{X}_m)^2}{S_{n-k}^2} \geq K^2 \right).
\]

Under the null hypotheses of equality of means we have

\[
\frac{m}{k-1} \sum_{i=1}^{k} \frac{(\bar{X}_{i,m} - \bar{X}_m)^2}{S_{n-k}^2} \overset{D}{=} F_{k-1,n-k}
\]

where \( F_{n_1,n_2} \) denotes the \( F \)-distribution with \( n_1 \) and \( n_2 \) degrees of freedom. This leads to the following test for contrasts of size \( \alpha \)

\[
\left| \sum_{i=1}^{k} c_i \bar{X}_{i,m} \right| \geq S_{n-k} \sqrt{\frac{k-1}{m} F_{k-1,n-k}(1-\alpha)}.
\]

The test guarantees that with probability of at least \( 1 - \alpha \) no type 1 error will be committed non matter how many contrasts are tested. Confidence intervals for \( \sum_{i=1}^{k} c_i \mu_i \) can be obtained by inverting the tests. The 100\( \alpha \)% confidence interval is given by

\[
\left[ \sum_{i=1}^{k} c_i \bar{X}_{i,m} - S_{n-k} \sqrt{\frac{k-1}{m} F_{k-1,n-k}(\alpha)} \right]
\]

\[
\sum_{i=1}^{k} c_i \bar{X}_{i,m} + S_{n-k} \sqrt{\frac{k-1}{m} F_{k-1,n-k}(\alpha)} \right].
\]

We note that all these confidence intervals hold simultaneously. Scheffé’s method can be generalized to the case that the samples sizes are not equal without difficulty. The resulting tests for any contrast \( c \) are given by

\[
\left| \sum_{i=1}^{k} c_i \bar{X}_{i,m} \right| \geq S_{n-k} \sqrt{(k-1) F_{k-1,n-k}(1-\alpha)} \left( \sum_{i=1}^{k} \frac{c_i^2}{n_i} \right)^{1/2} \quad (2.17)
\]

where we now no longer assume that \( \sum_{i=1}^{k} c_i^2 = 1 \). The corresponding confidence intervals are

\[
\left[ \sum_{i=1}^{k} c_i \bar{X}_{i,m} - S_{n-k} \sqrt{(k-1) F_{k-1,n-k}(\alpha)} \left( \sum_{i=1}^{k} \frac{c_i^2}{n_i} \right)^{1/2} \right]
\]

\[
\sum_{i=1}^{k} c_i \bar{X}_{i,m} + S_{n-k} \sqrt{(k-1) F_{k-1,n-k}(\alpha)} \left( \sum_{i=1}^{k} \frac{c_i^2}{n_i} \right)^{1/2} \right].
\]

(2.18)
2.5 Simultaneous confidence intervals

The methods considered so far are all based on the normal distribution. Some of them allow for different samples size but none are explicitly designed for different variances even if the distribution is normal. Moreover we do not require that the different samples differ only in their location and scale values; the underlying distributions may differ as well and they may be slightly skewed. Outliers occur in real data and have to be accounted for as well. Contrasts are not the only form of test which may be considered. The samples may be ordered according to some measurement, say severity of a disease, and one possible test is whether the means are increasing or decreasing. Finally confidence intervals may also be required, both for individual means as well as contrasts. We now offer an approach which allows for deviations from the model and permits any possible test to be performed or confidence interval to be calculated whilst still controlling the probability of a type 1 error and the levels of the confidence intervals.

We consider the redescending M-estimator \((T_L, T_S)\) as described in the Section 1.11. For each sample \(\mathbf{X}_i\) we calculate a confidence interval \(I_i, i = 1, \ldots, k\). We want the confidence intervals to be simultaneous confidence intervals of level \(\alpha\), that is we require

\[
P(T_L(\mathbb{P}_i) \in I_i, i = 1, \ldots, k) = \alpha
\]

where \(\mathbb{P}_i\) denotes the distribution of the \(i\)th sample. We shall assume that the samples are independent, an assumption which is not always plausible. However given this we have

\[
P(T_L(\mathbb{P}_i) \in I_i, i = 1, \ldots, k) = \prod_{i=1}^{k} P(T_L(\mathbb{P}_i) \in I_i) = \alpha
\]

and if we require all confidence intervals to be of the same level then we must have

\[
P(T_L(\mathbb{P}_i) \in I_i) = \alpha^{1/k}, i = 1, \ldots, k.
\]

Alternatively we can use Bonferroni bounds which leads to replacing \(\alpha\) by \(1 - (1 - \alpha)/k\) which is always close to \(\alpha^{1/k}\) so that the independence of the samples is almost a worst case situation. The construction of confidence intervals was described in Section 1.12 but there only conventional sizes \(\alpha = 0.9, 0.95\) and 0.99 were used. In the present case much higher values of \(\alpha\) will be encountered. If for example \(\alpha = 0.95\) and \(k = 5\) then \(\alpha^{1/k} = 0.990\) whereas for \(k = 10\) and \(k = 50\) the values are 0.998996 and 0.9998 respectively. We therefore require the quantiles of the statistic

\[
\sqrt{n}(T_L(\mathbb{P}_n) - T_L(\mathbb{P}))
\]
for a much larger range. This can be done by more extensive simulations and for the redescending $\psi$–estimator and we indicate how this is done. We consider firstly the case that we base the asymptotic behaviour on the normal distribution as described in Remark 1.91. The asymptotic value of the $\alpha$–quantile is $1.60qnorm(\alpha)T_S(P)$ which for finite $n$ we approximate by

$$\text{fac}(\alpha, n)qt(\alpha, n - 1)1.60T_S(P_n).$$  \hspace{1cm} (2.19)

The values of the correction factors $\text{fac}(\alpha, n)$ for $3 \leq n \leq 100$ are given by

$$\text{fac}(\alpha, n) = \exp(a(n - 2) + b(n - 2)\beta^2 + c(n - 2)\beta)$$

where $\beta = \log(-\log(1 - \alpha))$ and the constants $a(n - 2)$, $b(n - 2)$ and $c(n - 2)$ are determined from simulations. For $n \geq 100$ we put

$$\text{fac}(\alpha, n) = \exp((a(98) + b(98)\beta^2 + c(98)\beta) \exp(-1.13 \log(0.01 \log n)))$$

For example for $n = 3$ we have $a(1) = 1.872$, $b(1) = 1.862$ and $c(1) = -1.689$. We use the same form of approximation if we replace the factor $1.6$ in (2.19) by the empirical version

$$\left(\frac{\frac{1}{n} \sum_{i=1}^{n} \psi \left(\frac{x_i - T_L(E_n)}{T_S(E_n)}\right)^2}{\left(\frac{1}{n} \sum_{i=1}^{n} \psi^{(1)} \left(\frac{x_i - T_L(E_n)}{T_S(E_n)}\right)\right)^2}\right)^{1/2}$$

but of course with different values for the constants $a$, $b$ and $c$. For $n = 3$ we have $a(1) = 1.708$, $b(1) = 2.014$ and $c(1) = -2.216$. For such small values of $n$ there is however little point in using the empirical version as it is impossible to say anything about the distribution of the samples for such a small sample size. For values of $n \geq 20$ the empirical version does have some advantages.

We now apply the methodology to the data of Example 2.1 using the default redescending M–estimator. The values of $T_L$ are 6.037, 4.041, 6.360 respectively and those of $T_S$ are 0.771, 0.593, 1.023. The simultaneous confidence intervals with $\alpha = 0.95$ are

$$I_1 = [5.226, 6.849], I_2 = [3.226, 4.856], I_3 = [5.433, 7.288].$$  \hspace{1cm} (2.20)

These are shown graphically in Figure 25 which can be compared with the boxplot 24. We see that the interval for the data set $x_{2.17}$ is disjoint from the others so that its location is significantly different. The other two intervals overlap so that the data are consistent with a common location values. A
95% confidence interval for this common value is simply the intersection of the two intervals $I_1 \cap I_3 = [5.433, 6.849]$.

Once the simultaneous confidence intervals have been calculated all questions concerning the relationship between the population values can be answered by reference to the intervals. One example is the question of contrasts. Suppose we wish to know whether $\sum_{i=1}^{k} c_i T_L(P_i) = 0$ is consistent with the data. To consider this we calculate an upper and a lower bound for $\sum_{i=1}^{k} c_i T_L(P_i)$. We denote the $i$th interval by $[a_i, b_i]$. A confidence interval for the contrast is

$$\left[ \sum_{i=1}^{k} c_i d_i, \sum_{i=1}^{k} c_i e_i \right]$$

where

$$d_i = \begin{cases} a_i & c_i > 0 \\ b_i & c_i < 0 \end{cases}, \quad e_i = \begin{cases} b_i & c_i > 0 \\ a_i & c_i < 0 \end{cases}$$

If this interval covers the origin then $\sum_{i=1}^{k} c_i T_L(P_i) = 0$ is consistent with the data. The special case $c_i = 1, c_j = -1$ with the remainder of the $c_i$s zero gives a confidence bound for $T_L(P_i) - T_L(P_j)$. We note that the length of the confidence interval is

$$\sum_{i=1}^{k} c_i (e_i - d_i) = \sum_{i=1}^{k} c_i^+ (b_i - a_i) - \sum_{i=1}^{k} c_i^- (a_i - b_i) = \sum_{i=1}^{k} |c_i|(b_i - a_i).$$
A different question may arise if the samples are in a certain order such as the degree of severity of a disease or the temperature at which the observations were taken. In such a case it may be of interest to ask whether the population values are increasing. This will be accepted if there exist a sequence \( \mu_j \) with \( \mu_j \in I_j \) and \( \mu_1 \leq \mu_2 \leq \ldots \leq \mu_k \). It is possible to give confidence bounds for such increasing sequences as follows. If we ask for an increasing sequence then an upper bound is given by

\[
u_i = \min\{b_i, b_{i+1}, \ldots, b_k\}, \quad i = 1, \ldots, k
\]

and the corresponding lower bound by

\[
l_i = \max\{a_1, a_2, \ldots, a_i\}, \quad i = 1, \ldots, k.
\]

If at any point the lower bound \( l_i \) exceeds the upper bound \( u_i \) then there is no increasing sequence. Otherwise the tube is a confidence bound for an increasing sequence in the sense that any increasing sequence \( \mu_i, i = 1, \ldots, k \) with \( \mu_i \) acceptable for the \( i \)th sample then this sequence must lie in the tube.

Small sample with rounding of data can cause problems with M-estimators where the restriction that the largest atom is less than 0.5 does not hold. This the case for the data of example 1.13 where the first laboratory returned the data 2.2, 2.215, 2.2, 2.1 giving an atom of size 0.5 at the point 2.2. There is no correct method of dealing with this problem and any method will be to some extent ad hoc. For the present we use the following. If the size of the largest atom is at most 0.4 then we use the usual M-estimator. If it larger than 0.4 then we consider the strictly positive absolute deviations from the atom. If there are \( m \) of them we use as scale the \( \lceil m/2 \rceil \)th largest of these values. The location part which is calculated separately using this auxiliary scale. Using this Figure 26 shows the simultaneous confidence. The result is not very satisfactory. There are negative values which in the present context do not make sense and some of the confidence intervals are very large. This is because of the large value of \( \alpha = 0.95^{1/15} = 0.99634 \) and the fact that there are only 4 observations per laboratory. One way of overcoming the negative values is to transform the data by taking logarithms, performing the analysis and transforming back. This would not however reduce the size of the confidence intervals. How this may be done is the subject of the next section. makes no sense for such data.

2.6 Borrowing strength

If the measurements of the different samples have been made under similar conditions then there may be good reasons to suppose that the scales of the
samples should be more or less the same. This is the case for the inter-
laboratory data of Example 1.13 where the techniques of measurements are
similar so there is reason to suspect that the scales should also be similar.
Whether there is any justification for this is an empirical question and not
one that can be decided by statistics. If there is reason to suppose that the
scales are similar than advantage can be taken of this by pooling the various
scale estimates to give one which is more reliable. At the moment there is
no generally accepted way of doing this and the following is just a suggestion
which still needs to be investigated in more detail. Whether it is satisfactory
or not can only be decided by applications to real data. Let $S_i = T_S(x_{i,n_i})$
denote the scale of the $i$th sample. We replace this by

$$S_i^* = \frac{\sum_{j=1}^{k} S_j \exp(-2(\log(S_i/S_j))^2)}{\sum_{j=1}^{k} \exp(-2(\log(S_i/S_j))^2)}$$  \hfill (2.21)$$

and in the expression for the confidence interval we replace

$$\text{fac}(n_i, \alpha)qt(\alpha, n_i - 1)S_i \quad \text{by} \quad \text{fac}(n_i^*, \alpha)qt(\alpha, n_i^* - 1)S_i^*$$

where

$$n_i^* = \sum_{j=1}^{k} n_j \exp(-2(\log(S_i/S_j))^2).$$  \hfill (2.22)$$
We do not perform a formal test of equality of the different scales but use a weighted average of those scale values which are close to that of the sample under consideration. This is the case for the data of (2.7). We calculate the residuals from each sample to give a total of 68 observations. The scale estimate for this sample is 0.8670 and the location estimates for the three samples are 6.038, 4.099 and 6.286 respectively. The confidence intervals are

\[ I_1 = [5.395, 6.680], I_2 = [3.350, 4.847], I_3 = [5.704, 6.868]. \]

Only the third interval has become shorter which is due to the fact that the samples sizes are so large that pooling brings little benefit. If the sample sizes are small the effect can be considerable. Figure 27 shows the result for the data of Example 1.13. For this data the scales with borrowing strength

![Figure 27: The figure shows the simultaneous confidence intervals for the data of Example 1.13 but borrowing strength.](image)

are

\[
\begin{array}{cccccccccc}
0.0123 & 0.0179 & 0.2383 & 0.0079 & 0.2203 & 0.1336 & 0.0385 \\
0.0576 & 0.0127 & 0.0349 & 0.2111 & 0.0057 & 0.0080 & 0.0119 \\
\end{array}
\]

compared to the original scales of

\[
\begin{array}{cccccccccc}
0.0150 & 0.0250 & 0.3848 & 0.0078 & 0.2609 & 0.1520 & 0.0510 \\
0.0800 & 0.0158 & 0.0460 & 0.2422 & 0.0052 & 0.0080 & 0.0142. \\
\end{array}
\]

The differences are not large. What has lead to the considerable reduction in the lengths of the confidence intervals is the increase in the number of
degrees of freedom. Without borrowing strength the number of degrees of freedom is 3 for all samples. With borrowing strength the numbers as given by (2.22) are

\[ 18, 15, 10, 16, 13, 11, 13, 11, 18, 13, 13, 10, 16, 18. \]

## 2.7 An example

We now apply the different methods to the data set of Example 2.1. As we have already seen Fisher’s $F$-test leads to the conclusion that the means are significantly different.

### 2.7.1 Tukey: Honest Significant Difference

We have $k = 3$ and $n = 68$. Direct calculation gives

\[
\begin{align*}
s_{n-k}^2 &= \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{i,n_i})^2 \\
&= \frac{1}{65} (40.81718 + 27.15985 + 127.8995) = 3.013485
\end{align*}
\]

so that $s_{n-k} = 1.735939$. For $\alpha = 0.05$ the $1 - \alpha$-quantile of Tukey’s studentized range with $k = 3$ and $n - k = 65$ is

\[
> \text{qtukey}(0.95, 3, 65)
\]

[1] 3.392061

The means are 6.08087, 4.241765 and 6.791429 and the confidence intervals are given by $100\alpha\%$-confidence interval for $\mu_i - \mu_j$

\[
\begin{align*}
\left[ \bar{x}_{i,n_i} - \bar{x}_{j,n_j} - s_{n-k} \text{qtukey}(1 - \alpha, k, n - k) \sqrt{\frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}, \\
\bar{x}_{i,n_i} - \bar{x}_{j,n_j} + s_{n-k} \text{qtukey}(1 - \alpha, k, n - k) \sqrt{\frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \right]
\end{align*}
\]

We have

\[
\begin{align*}
s_{n-k} \text{qtukey}(1 - \alpha, k, n - k) \sqrt{\frac{1}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \\
&= 1.735939 \cdot 3.392061 \sqrt{0.5(0.04347826 + 0.05882353)} \\
&= 1.331756.
\end{align*}
\]
Consequently the confidence interval for $\mu_1 - \mu_2$ is

$$[6.08087 - 4.241765 - 1.331756, 6.08087 - 4.241765 + 1.331756] = [0.507349, 3.170861].$$

Similarly the confidence intervals for $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$ are respectively $[-1.8822833, 0.4611653]$ and $[-3.8298881, -1.2694396]$. As the interval for $\mu_1 - \mu_3$ contains 0 it follows that the difference in the means for the control and inactive patients is not significant.

### 2.7.2 Bonferroni

We use the $t$-tests

$$\left(\sqrt{\frac{n_i n_j}{n_i + n_j}}\right) \frac{|\bar{x}_{i,n_i} - \bar{x}_{j,n_j}|}{s_{k,n-k}}$$

based on $n - k$ degrees of freedom. As we perform three tests we must replace $\alpha = 0.05$ by $\alpha^* = \alpha/3 = 0.01667$. The critical value we must use is

$$qt(1 - \alpha^*/2, n - k) = qt(0.991667, 65) = 2.457531.$$

For the samples 1 and 2 the values of the test statistic is

$$\left(\sqrt{\frac{23 \cdot 17}{23 + 17}}\right) \frac{|6.08087 - 4.241765|}{1.735939} = 3.312306$$

which is larger than 2.457531 so the means are significantly different. The values of the test statistic for samples 1 and 3 and samples 2 and 3 are respectively 1.454533 and 4.776895. We therefore conclude that then means of samples 1 and 3 are not significantly different but those of samples 2 and 3 are. The confidence intervals are given by

$$\left[\bar{x}_{i,n_i} - \bar{x}_{j,n_j} - s_{k,n-k}qt(1 - \alpha^*/2, n - k)\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \right.$$  

$$\left.\bar{x}_{i,n_i} - \bar{x}_{j,n_j} - s_{k,n-k}qt(1 - \alpha^*/2, n - k)\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \right]$$

For $\mu_1 - \mu_2$ we get

$$6.08087 - 4.241765 - 1.735939 \cdot 2.457531 \sqrt{1/23 + 1/17} = 0.4746002$$

and

$$6.08087 - 4.241765 + 1.735939 \cdot 2.457531 \sqrt{1/23 + 1/17} = 3.20361$$

which gives a confidence interval $[0.4746002, 3.20361]$. For $\mu_1 - \mu_3$ we obtain $[-1.911096, 0.4899782]$ and for $\mu_2 - \mu_3$ we get $[-3.861369, -1.237959]$. This can be compared with the interval for Tukey’s method.
2.7.3 Holm

We use the \( t \)-test as in the last section. The attained values of the test statistic are 3.312306, \((H_{12} : \mu_1 = \mu_2), 1.454533 (H_{13} : \mu_1 = \mu_3)\) and 4.776895 \((H_{23} : \mu_2 = \mu_2))\). The attained \( p \)-values are

\[2(1 - \text{pt}(3.312306, 65)), 2(1 - \text{pt}(1.454533, 65)), 2(1 - \text{pt}(4.776895, 65))\]

which are respectively

\[0.001514134, 0.1506107, 0.0000153655.\]

On ordering these values we see that the hypotheses are to be tested in the order \( H_{23}, H_{12}, H_{13} \). The associated values of \( \alpha \) are \( \alpha/3, \alpha/2 \) and \( \alpha \). For \( \alpha = 0.05 \) these are 0.016667, 0.025 and 0.05. As 0.0000153655 < 0.016667 we reject \( H_{23} \) and consider \( H_{12} \). As 0.001514134 < 0.025 we reject \( H_{12} \) and consider \( H_{13} \). As 0.1506107 > 0.05 we accept \( H_{23} \).

2.7.4 Scheffé

We use (2.17). The value of \( S_{n-k} \) is, as before, 1.735039 and the value of \( F_{k-1,n-k}(1 - \alpha) \) for \( k = 3, n = 68 \) and \( \alpha = 0.05 \) is

\[> \text{qf}(0.95, 2, 65)\]

[1] 3.138142

The null hypothesis \( H_{12} : \mu_1 = \mu_2 \) can be written as \( H_{12} : \mu_1 - \mu_2 = 0 \) which corresponds to the contrast \( c_1 = 1, c_2 = -1 \) and \( c_3 = 0 \). The critical value of the test statistic of (2.17) is therefore

\[S_{n-k} \sqrt{(k-1)F_{k-1,n-k}(1 - \alpha)} \left( \sum_{i=1}^{k} c_i^2/n_i \right)^{1/2} = 1.735039 \sqrt{(3-1) \cdot 3.138142 \sqrt{1^2/23 + (-1)^2/17}} = 1.390280.\]

The value of the tests statistic is \(|6.08087 - 4.241765| = 1.839105\) so that the null hypothesis \( h_{12} : \mu_1 = \mu_2 \) is rejected. For the hypothesis \( H_{13} : \mu_1 = \mu_3 \) the critical value is

\[1.735039 \sqrt{(3-1) \cdot 3.138142 \sqrt{1^2/23 + 1^2/17}} = 1.223215\]

and the value of the test statistic is \(|6.08087 - 6.791429| = 0.710559\) so that \( H_{13} : \mu_1 = \mu_3 \) is not rejected. Finally the critical value for \( H_{23} : \mu_2 = \mu_2 \) is

\[1.735039 \sqrt{(3-1) \cdot 3.138142 \sqrt{1^2/17 + 1^2/28}} = 1.336483\]

93
and the value of the test statistic is $|4.241765 - 6.791429| = 2.549664$ so that the hypothesis $H_{23} : \mu_2 = \mu_3$ is also rejected. The confidence intervals are calculated according to (2.18). For $\mu_1 - \mu_2$ we get

$$[6.08087 - 4.241765 - 1.390280, 6.08087 - 4.241765 + 1.390280] = [0.448825, 3.229385].$$

The corresponding intervals for $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$ are

$$[-1.933774, 0.512656] \quad \text{and} \quad [-3.886147, -1.213181]$$

respectively.

2.7.5 Simultaneous confidence intervals

The analysis is based on the simultaneous confidence intervals for $\mu_1, \mu_2$ and $\mu_3$ of (2.20)

$$I_1 = [5.226, 6.849], \quad I_2 = [3.226, 4.856], \quad I_3 = [5.433, 7.288].$$

As $I_1 \cap I_2 = \emptyset$ the hypothesis $H_{12} : \mu_1 = \mu_2$ is rejected. Similarly as $I_2 \cap I_3 = \emptyset$ the hypothesis $H_{23} : \mu_2 = \mu_3$ is rejected. As $I_1 \cap I_3 \neq \emptyset$ the hypothesis $H_{13} : \mu_1 = \mu_3$ is accepted and $I_1 \cap I_3 = [5.433, 6.849]$ is a confidence interval for the common value. The confidence intervals for $\mu_1 - \mu_2$, $\mu_1 - \mu_3$ and $\mu_2 - \mu_3$ are respectively

$$[5.226 - 4.856, 6.849 - 3.226] = [0.370, 3.623]$$

3 The Two-way table

3.1 The Two-Way Table

3.1.1 The problem

The one-way table takes only one factor, for example the variety of wheat, into consideration. The two-way table takes two factors into consideration, for example the variety of the wheat and the fertilizer. The problem is to separate the effects of the two factors which is done, at least initially in assuming that their joint effect is simply additive. Interactions can also be taken into account as it may be that a particular fertilizer is particularly suitable for a particular variety of wheat. Clearly more complicated relationships between the two factors which go beyond additivity and individual interactions can be imagined but we shall consider only the simplest model.
3.1.2  The classical method

We consider two factors each of which is at several levels. In the wheat/fertilizer example the levels are the number of varieties of wheat and the number of different fertilizers. We have $n_{ij}$ observation with Factor 1 at level $i$ and Factor 2 at level $j$. The purely additive model for this situation is

$$X_{ijk} = a_i + b_j + \sigma \epsilon_{ijk}, \quad k = 1, \ldots, K_{ij}, \ i = 1, \ldots, I, \ j = 1, \ldots, J,$$  \hspace{1cm} (3.1)

where the $\epsilon_{ijk}$ are standard white noise. The parameter $a_i$ represents the effect of the first factor at level $i$ and $b_j$ represents the effect of the second factor at level $j$. The model assumes that the effects are additive. The first thing to notice is that the parameters $a_i$ and $b_j$ are not identifiable. If we increase all the $a_i$ by $d$ and decrease all the $b_j$ by $d$ than the model remains exactly as before. What can be estimated are differences in the $a_i$s and $b_j$s and they are the relevant quantities for such questions as whether one variety of wheat is better than another. For this reason the model is often written in the form

$$X_{ijk} = \mu + a_i + b_j + \sigma \epsilon_{ijk}, \quad k = 1, \ldots, K_{ij}, \ i = 1, \ldots, I, \ j = 1, \ldots, J.$$  \hspace{1cm} (3.2)

with the following restrictions imposed on the $a_i$s and $b_j$s,

$$\sum_{i=1}^{I} a_i = \sum_{j=1}^{J} b_j = 0.$$  \hspace{1cm} (3.3)

The restrictions (3.3) are often presented as a technical condition to enable the parameters to be identified and one which is neutral in terms of data analysis. This is not the case. The restrictions (3.3) are equivalent to the means of the rows and column effects being zero. If there is one large value, say $a_{i_0}$, then all the other $a_i$s may be zero giving a perhaps false impression of being worse than average. It may be more sensible to obtain identifiability by requiring that the medians of the row and column effects be zero.

The classical method of estimating the parameters $a_i$ and $b_j$ in (3.1) and one that may be sold as maximum likelihood using the Gaussian model is to minimize

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K_{ij}} (X_{ijk} - \mu - a_i - b_j)^2.$$  \hspace{1cm} (3.4)

subject to the restrictions (3.3). This is only a simple task if the design is balance, that is if $K_{ij} = K$ for all $i$ and $j$. In the unbalanced case the
calculations can be done as a particular form of linear regression analysis as we shall indicate later. In the balanced case we have the following Table 8 which corresponds to Table 7 for the one-way analysis of variance.

Under the model (3.1) the sums of squares in the third column are independently distributed and, when divided by \( \sigma^2 \), they follow non-central \( \chi^2 \)–distributions with the number of degrees of freedom given in the second column. If there are no row or column effects then the distributions become central \( \chi^2 \)–distributions and these considerations lead to an \( F \)–test for the null hypotheses

\[
H^a_0 : a_1 = \ldots = a_I, \quad H^b_0 : b_1 = \ldots = b_J = 0. \tag{3.5}
\]

Thus the test statistic for \( H^a_0 \) is

\[
\frac{TJK \sum_{i=1}^I (\bar{X}_{i.-} - \bar{X}_{.-})^2}{\frac{TJK-I-J+1}{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (X_{ijk} - \bar{X}_{i.-} - \bar{X}_{.-} + \bar{X}_{..})^2}}
\]

and a test of size \( \alpha \) rejects the null hypothesis \( H^a_0 \) if the test statistics exceeds \( F_{I-1,IJK-I-J+1}(1-\alpha) \). The test for the null hypothesis \( a_i = a_j \) is based on the test statistic

\[
\frac{\sqrt{TJK} (\bar{X}_{i.-} - \bar{X}_{j.-})}{\sqrt{\frac{TJK-I-J+1}{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (X_{ijk} - \bar{X}_{i.-} - \bar{X}_{.-} + \bar{X}_{..})^2}}}
\]
which under the null hypothesis follows a $t$-distribution with $IJK - I - J + 1$ degrees of freedom. The standard methods for multiple testing in described for the one-way table can be applied here.

### 3.1.3 Interactions

Interactions terms are often introduced into the model which then becomes

$$X_{ijk} = \mu + a_i + b_j + c_{ij} + \sigma \epsilon_{ijk}, \quad k = 1, \ldots, K, \ i = 1, \ldots, I, \ j = 1, \ldots, J.$$  \hfill (3.6)

Once again the parameters are not identifiable and the restrictions (3.3) are no longer sufficient. To see this we can simply put $c_{ij} = a_i + b_j$ and then replace the initial $a_i$ and $b_j$ by zero. The conventional answer and one that is justified by the requirement of identifiability is to require that the marginal sums of the interactions are zero,

$$c_i = \sum_{j=1}^{J} c_{ij} = c_j = \sum_{i=1}^{I} c_{ij} = 0$$ \hfill (3.7)

for all $i$ and $j$. The variance decomposition of Table 8 can be extended to cover interactions as shown in Table 9. Once again under the model (3.6) and the usual assumptions of normality the sums of squares on the third column are independently distributed random variables with non-central $\chi^2$-distributions if the relevant parameter values are non-zero. This allows the construction of $F$-tests for the hypotheses and confidence intervals can be obtained as was done in the one-way table. Care must be taken however when interpreting the null hypotheses if the interaction terms are non-zero. Thus the null hypothesis $H_0: a_1 = a_2$ will only mean the the levels 1 and 2 of the first factor are more or less equivalent if the interactions are zero. If they are not zero the null hypothesis expresses no difference only after the interactions have been factored out. There is however a deeper reason why it is difficult to interpret the effect of interactions and we now discuss it in some detail.

As mentioned above the restrictions (3.7) are often regarded simply as a way of allowing the model to be identified. With very few exceptions no mention is made of the fact that (3.7) is not data-analytically neutral. To see this we simply note that the restrictions imply that the smallest number of non-zero interactions is four. The next smallest number is six and this will only occur if one row or column has three interactions. If this is not the case then eight interactions will be required. In particular a single interaction in
<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sums of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand mean</td>
<td>1</td>
<td>$IJK \bar{x}^2$</td>
</tr>
<tr>
<td>Rows</td>
<td>$I - 1$</td>
<td>$JK \sum_{i=1}^I (\bar{x}<em>{i.} - \bar{x}</em>{..})^2$</td>
</tr>
<tr>
<td>Columns</td>
<td>$J - 1$</td>
<td>$IK \sum_{j=1}^J (\bar{x}<em>{.j} - \bar{x}</em>{..})^2$</td>
</tr>
<tr>
<td>Interactions</td>
<td>$(I - 1)(J - 1)$</td>
<td>$K \sum_{i=1}^I \sum_{j=1}^J (\bar{x}<em>{ij} - \bar{x}</em>{i.} - \bar{x}<em>{.j} + \bar{x}</em>{..})^2$</td>
</tr>
<tr>
<td>Error</td>
<td>$IJ(K - 1)$</td>
<td>$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (x_{ijk} - \bar{x}_{ij})^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$IJK$</td>
<td>$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K x_{ijk}^2$</td>
</tr>
</tbody>
</table>

Table 9: The classical balanced two-way table with interactions.

a single cell is not allowed. It is not clear why nature should be so obliging to statisticians. It is easier to understand the problem in the no noise situation of an exact model. We start with the data shown in the left tableau of Table 10. The right tableau shows the same data expressed in the form corresponding to (3.6), namely

$$x_{ij} = \mu + a_i + b_j + c_{ij}$$

with the restrictions (3.3) and (3.7) in force. The bottom right entry shows the grand mean, the left most column shows the row effects and the bottom row shows the column effects. The remainder in the main body shows the interaction terms. Only a professional statistician would prefer the right-hand tableau.

For reasons of parsimony it seems reasonable to choose the row and column effects so as to minimize the number of interactions required. This may be expressed in the form

$$\text{minimize } a_i, b_j \mid \{(i, j) : |x_{ij} - a_i - b_j| > 0\}.$$  \hspace{1cm} (3.8)

In general the solution of (3.8) will not be unique in the sense that there may be several minimal sets of pairs $(i, j)$. A simple example is shown Table
Table 10: A single interaction.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2/9</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2/9</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2/9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>-2/9</th>
<th>-2/9</th>
<th>2/9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>-1/9</td>
</tr>
<tr>
<td>0</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>-1/9</td>
</tr>
<tr>
<td>0</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>-1/9</td>
</tr>
</tbody>
</table>

Table 11: Two different minimal sets.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

In some cases (3.8) may have a unique solution but this may depend on the particular values of the interactions even if these interactions are located in the same cells. This leads to the following definition. Given a set of interactions \( c_{ij} \) we define the interaction pattern as the matrix

\[
C = \{c_{ij} \neq 0\}_{1 \leq i \leq I, 1 \leq j \leq J}.
\]

In other words \( C_{ij} = 1 \) if and only if \( c_{ij} \neq 0 \) and \( C_{ij} = 0 \) otherwise. Consider now numbers \( x_{ij} \) generated by \( x_{ij} = a_i + b_j + c_{ij} \) where the \( c_{ij} \) have the interaction pattern \( C \). We call \( C \) unconditionally identifiable if

\[
x_{ij} - \tilde{a}_i - \tilde{b}_j = c_{ij}
\]

for all \( c_{ij} \) with interaction pattern \( C \) where \( \tilde{a}_i \) and \( \tilde{b}_j \) are any solution of (3.8). We have the following theorem

**Theorem 3.1.** Suppose \( x_{ij} = a_i^* + b_j^* + c_{ij} \) where the \( c_{ij} \) have an unconditionally identifiable interaction pattern \( C \). Then the \( L^1 \) solution

\[
\text{minimize } \sum_{i=1}^{I} \sum_{j=1}^{J} |x_{ij} - a_i - b_j|
\]

is unique and if the solution is \( a_i^1, i = 1, \ldots, I \) and \( b_j^1, j = 1, \ldots, J \) then

\[
x_{ij} - a_i^1 - b_j^1 = c_{ij}.
\]

The theorem would be of little use if there were hardly any unconditionally identifiable interaction patterns but it turns out that there are many. Here are two sufficient conditions.
Theorem 3.2.

(a) An interaction pattern $C$ in which the majority of the rows and a majority of the columns contain no interactions is unconditionally identifiable.

(b) An interaction pattern $C$ such that in each row and column less than one quarter of the cells contain interactions is unconditionally identifiable.

It is clear that if $C$ is an unconditionally identifiable interaction pattern then so is any pattern $C'$ obtained from $C$ by permuting rows and columns and by interchanging rows and columns. It follows from Theorem 3.2 (a) and (b) respectively that

$$
\begin{pmatrix}
1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\quad
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

are both unconditionally identifiable. there is an elegant method of deciding whether an interaction pattern is unconditionally identifiable.

Theorem 3.3. Let $C$ be an interaction pattern and $C'$ be any other interaction pattern obtainable from $C$ by adding row and columns of 1s mod (2). If $|C|$ denotes the number of 1s in $C$ then $C$ is unconditionally identifiable if and only if $|C| < |C'|$ for any $C' \neq C$.

There are essentially $2^{\min(I,J)}$ possibilities to check using a simple computer programme. As long as say $\min(I, J) \leq 30$ this is possible in essentially finite time. using Theorem 3.3 it is possible to show that the interaction pattern

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 1
\end{pmatrix}
$$

is unconditionally identifiable, a fact which does not follow from Theorem 3.2.

When analysing a two-way table we will require an estimate of scale based on the residuals. We want this estimate to be robust and to be able to withstand large outliers or interactions in certain cells. In order to be able to
avoid these we need to know how many there can be. An upper bound is provided by the next theorem.

**Theorem 3.4.** Suppose the $c_{ij}$ have an unconditionally identifiable interaction pattern. Then at most

$$
\min \left\{ \left( J - \left\lfloor \frac{J-1}{2} \right\rfloor \right) \left\lfloor \frac{I-2}{2} \right\rfloor , \left( I - \left\lfloor \frac{I-1}{2} \right\rfloor \right) \left\lfloor \frac{J-2}{2} \right\rfloor \right\} + \left\lfloor \frac{I-1}{2} \right\rfloor \left\lfloor \frac{J-1}{2} \right\rfloor \right) \right\} (3.9)
$$

of the interactions are non-zero.

If $I = J = 3$ then the maximum number is 1. If $I = J = 5$ then the maximum number is at most 7. An unconditionally identifiable pattern with 6 interactions was given just after Theorem 3.3.

So far we have considered exact data but the results carry over to noisy data in a sense made precise by the following theorem.

**Theorem 3.5.** Let $x_{ij}$ be data of the form

$$
x_{ij} = a_i^0 + b_j^0 + r_{ij}, \ 1 \leq i \leq I, \ 1 \leq j \leq J
$$

where the $r_{ij}$ are arbitrary but fixed. Consider all possible data sets of the form

$$
x_{ij}^c = x_{ij} + c_{ij}, \ 1 \leq i \leq I, \ 1 \leq j \leq J
$$

where the $c_{ij}$ are arbitrary but have an unconditionally identifiable interaction pattern. If $\tilde{a}_i^c$ and $\tilde{b}_j^c$ are $L^1$–solutions of

$$
\minimize \sum_{i=1}^{I} \sum_{j=1}^{J} |x_{ij}^c - a_i - b_j|
$$

normalized to have $\tilde{a}_i^c = 0$, then

$$
sup_c (||\tilde{a}^c|| + ||\tilde{b}^c||) < \infty.
$$

The theorem say that if the interactions are sufficiently large then then they can be identified using the $L^1$–solution. We have to decide what we mean by “sufficiently large” and again we have recourse to simulations. We use the Gaussian distribution as our target model and it is sufficient to consider data of the form

$$
X_{ij} = Z_{ij}, \ 1 \leq i \leq I, \ 1 \leq j \leq J
$$

101
where the $Z_{ij}$ are Gaussian white noise. We calculate the residuals from an $L^1$ solution and now wish to define a robust scale functional based on these residuals. It is a property of an $L^1$ solution that $I + J - 1$ of these residuals may be zero, and indeed will be zero if the solution is unique. Clearly our estimate must be based on the remaining $IJ - I - J + 2$ residuals. Of these according to Theorem 3.4 at most $N(I, J)$ can have been effected by outliers where $N(I, J)$ is the number defined in (3.9). Let $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(IJ)}$ denote the ordered residuals. We set

$$s_{IJ} = r_{(L)}, \quad L = \left\lfloor \frac{IJ + I + J - 1}{2} \right\rfloor.$$  \hspace{1cm} (3.10)

It is easy to show that if the interactions are sufficiently large then this order statistic cannot come from an interaction cell. We use the scale to standardize the residuals and by means of simulations we can obtain the say 95% quantile of $\max_{i,j} |r_{ij}| / s_{IJ}$. This is then the bound for outliers or interactions. For any data set $x_{ij}$ the residual in the cell $(i, j)$ will be declared to be an outlier or an interaction if

$$|r_{ij}| \geq \text{qtw}(0.95, I, J) S_{IJ}.$$ 

The values of qtw$(0.95, I, J)$ can be approximated by simple functions just as in the one-way table. There is a problem however which must first be overcome. In general the $L^1$ solution is not unique and although the sum of the absolute residuals is the same for different solutions the values of $s_{IJ}$ are not. This means that the procedure is not invariant to a permutation of rows and columns. There is no easy solution to this problem but it is only serious for small values of $I$ and $J$. To reduce the effect we randomize the rows and columns, calculate the new $L^1$ residuals and define $s_{IJ}$ to be the smallest value attained. We illustrate the method using a simple data set shown in the leftmost tableau of Table 12. The centre tableau shows the residuals of the standard least squares analysis whilst the rightmost tableau shows the residuals of an $L^1$ analysis. If we calculate the sum of squares of the least squares residuals and divide by the number of degrees of freedom $9-5=4$ then the estimated standard deviation is 34.61. The largest standardized residual, that in cell (3,3) is 1.300 which is in no manner suspicious.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>449</td>
<td>413</td>
<td>326</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>409</td>
<td>358</td>
<td>291</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>341</td>
<td>278</td>
<td>312</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: The first table shows the original data, the second the residuals after a standard $L^2$ fit and the third the residuals of an $L^1$ fit.
seems clear when one looks at the $L^1$ residuals. The large residual in cell (3,3) does indeed look like an outliers or an interaction. However matters are not that simple. It is not easy to to recognize an outlier with only nine observations when 5 parameters have been estimated. However we can simulate by taking $3 \times 3$-tables of white noise, standardizing the residuals by dividing by the scale $s_{33}$ of (3.10) and then estimate say the 95%-quantile of this statistic. If we do this the result is about 5.12. For the data of Table 12 we have $s_{33} = 12$ and consequently the largest standardized residual is $89/12 = 7.4167$ so that there are grounds for regarding this observation as an outlier. If we eliminate it and perform a least squares analysis the outlying value is now about 94 which suggests perhaps a transcription error.

The $L^1$ solution is related to Tukey’s median polish. We start with the rows (or columns) and subtract the median. We then subtract the median from the columns and continue in this fashion until hopefully convergence. Table 13 shows the result of applying the median polish to the data of Table 12.

<table>
<thead>
<tr>
<th></th>
<th>36</th>
<th>0</th>
<th>-87</th>
<th>413</th>
<th>0</th>
<th>0</th>
<th>-20</th>
<th>413</th>
<th>0</th>
<th>0</th>
<th>-20</th>
<th>413</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>0</td>
<td>0</td>
<td>-67</td>
<td>358</td>
<td>-15</td>
<td>0</td>
<td>0</td>
<td>358</td>
<td>-15</td>
<td>0</td>
<td>0</td>
<td>358</td>
</tr>
<tr>
<td>29</td>
<td>-34</td>
<td>0</td>
<td>312</td>
<td>36</td>
<td>0</td>
<td>-67</td>
<td>36</td>
<td>0</td>
<td>-67</td>
<td>36</td>
<td>0</td>
<td>-67</td>
</tr>
</tbody>
</table>

Table 13: The first three sweeps of Tukey’s median polish for the data of Table 12. Convergence is attained after the third sweep.

12. Convergence is reached after three sweeps and at each stage, if the median of the row or column in non-zero the sum of the absolute values of the residuals is decreased. However even if convergence takes place the the result does not converge to an $L^1$ solution. This is shown by the rightmost tableau of Table 13. The sum of the absolute residuals is 136 whereas the sum for the $L^1$ solution, the rightmost tableau of Table 12 is 121. The median polish may not converge and as just shown, even if it does converge it does not in general converge to an $L^1$ solution. Finally, in common with the $L^1$ solution, it is not in general permutation invariant.

As shown above, the $L^1$ solution can be used to identify outliers or interactions and in this respect it is optimal as in can identify all sufficiently large interactions whose interaction pattern is unconditionally identifiable. It is however not suitable for providing confidence intervals of parameters of
interest. The $L^1$ solution minimizes
\[
\sum_{i=1}^{I} \sum_{j=1}^{J} |x_{ij} - a_i - b_j|
\]
and the obvious generalization would be to minimize
\[
\sum_{i=1}^{I} \sum_{j=1}^{J} \rho(x_{ij} - a_i - b_j) \tag{3.11}
\]
for some function $\rho$. If $\rho$ is strictly convex and if we standardize by putting for instance $a(1) = 0$ then the solution is unique. If moreover $\rho(u) \propto A|u|$ for some constant $A > 0$ and for large $|u|$ then it can be shown that Theorem 3.5 holds for the solution. For such a $\rho$ it is possible to calculate the solution by solving
\[
\sum_{j=1}^{J} \psi(x_{ij} - b_j - a_i) = 0 \tag{3.12}
\]
for $a_i$ with the corresponding equations for the $b_j$. At each stage the sum in (3.11) will be decreased and so tend to a limit. For the $a_i$ and $b_j$ at the limit (3.12) will hold and the strict convexity now implies that this is the solution. The solution is however not scale equivariant and to make it so must (3.11) be replaced by
\[
\text{minimize } \sum_{i=1}^{I} \sum_{j=1}^{J} \rho((x_{ij} - a_i - b_j)/s_{IJ}) \tag{3.13}
\]
where $s_{IJ}$ is a robust scale estimator. Ideally one would like to have a well-defined robust (against unconditionally identifiable interaction patterns) location functional which can be easily calculated. Unfortunately $s_{IJ}$ of (3.10) is in general not well defined due to the non-uniqueness of the $L^1$ solutions. Theoretically one could minimize $s_{IJ}$ over all $L^1$ solutions but it is not clear how to due this. Strategies such as taking differences of rows and columns can theoretically be made to work, or at least provide a scale functional which does not explode, but it may well implode in the presence of interactions. The problem is that taking differences may well increase the number of outliers: the difference of two rows each with one outlier may have two outliers. Finally even if such a scale exists it is not clear how to obtain confidence intervals in the presence of interactions.
Addendum

The following seems to be a reasonable and practical approach to the problem sketched above. Calculate the $L^1$–norm of the residuals of an $L^1$–solutions which we denote by $ml^1$. These are all the same up to machine accuracy. Let $s_{IJ}$ be as in (3.10). Consider the $\rho$–function with derivative $\psi_c(x) = \frac{(\exp(cx) - 1)}{(\exp(cx) + 1)}$ as in (1.34). We start with $c = 1$, calculate the solution of (3.13) and then calculate the $L^1$–norm of the residuals which we denote by $l^1$. If $l^1 > ml^1 + 10^{-3} IJs_{IJ}$ then we increase $c$ by a factor of 1.5 and repeat. This is continued until $l^1 \leq ml^1 + 10^{-3} IJs_{IJ}$. As $c$ increases the function $\rho$ will approach $A|u|$ for some $A > 0$ and the solutions will tend to an $L^1$–solutions. Thus for some $c$ sufficiently large $l^1 \leq ml^1 + 10^{-3} IJs_{IJ}$ will eventually hold and the solution will be unique. There is still a slight conceptual problem in the non-uniqueness of $s_{IJ}$ but as it is only used as a measure of accuracy possibly different $s_{IJ}$ resulting from different $L^1$–solutions will only lead to small differences in the final solution. The choice $10^{-3}$ is dictated by problems of machine accuracy. If there are different $L^1$–solutions the function to minimize in (3.13) will be very flat at the minimum making it difficult to locate the minimum itself in spite of the strict convexity. The factor $10^{-3}$ seems to work very well. An example is shown in Table 14.

<table>
<thead>
<tr>
<th></th>
<th>-0.15</th>
<th>-1.56</th>
<th>0.83</th>
<th>1.46</th>
<th>-0.26</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.03</td>
<td>-0.47</td>
<td>1.21</td>
<td>1.85</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>-0.31</td>
<td>-0.28</td>
<td>-0.70</td>
<td>1.09</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>0.84</td>
<td>-2.01</td>
<td>-1.22</td>
<td>-0.20</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>-2.20</td>
<td>0.62</td>
<td>1.40</td>
<td>0.48</td>
<td>-1.95</td>
</tr>
<tr>
<td></td>
<td>-0.53</td>
<td>0.00</td>
<td>0.51</td>
<td>0.52</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
<td>-1.60</td>
<td>-0.43</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>2.87</td>
<td>0.00</td>
<td>-0.39</td>
<td>0.00</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td>-0.86</td>
<td>1.94</td>
<td>1.54</td>
<td>0.00</td>
<td>-1.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>-0.15</th>
<th>-1.56</th>
<th>0.83</th>
<th>1.46</th>
<th>-0.26</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.29</td>
<td>-0.71</td>
<td>0.00</td>
<td>0.51</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>-0.96</td>
<td>0.00</td>
<td>0.00</td>
<td>0.52</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.43</td>
<td>-1.68</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>2.44</td>
<td>0.00</td>
<td>-0.90</td>
<td>0.00</td>
<td>2.19</td>
</tr>
<tr>
<td></td>
<td>-1.29</td>
<td>1.94</td>
<td>1.03</td>
<td>0.00</td>
<td>-1.24</td>
</tr>
</tbody>
</table>

Table 14: The original data are shown in the top-left tableau. The residuals from two different $L^1$–solutions are shown top-right and bottom-left. The bottom-right tableau show the result of the procedure described in the addendum.
4 Linear Regression

4.1 The classical theory

4.1.1 The straight line

Figure 28 shows the simplest situation of a cloud of points in $\mathbb{R}^2$ which may reasonably be approximated by a straight line and indeed were generated as such. The classical model is

$$Y(i) = a + bx(i) + \epsilon(i), 1 \leq i \leq n$$

(4.1)

where the $\epsilon(t)$ are Gaussian white noise with variance $\sigma^2$. The word “linear” refers to the coefficients $a$ and $b$ in (4.1) rather than the term $x$. Thus

$$Y(t) = a + b\log(x(t)) + \sigma\epsilon(t), 1 \leq t \leq n$$

(4.2)

is also a linear model and Figure 29 shows data generated by this model.

![Figure 28: A cloud of point in $\mathbb{R}^2$ with an approximating straight line in black and the generating line according to (4.1) in red.](image)

The theory of the model (4.1) is relatively simple. the least squares or, in the Gaussian case, maximum likelihood estimator is

$$\argmin_{a,b} \sum_{i=1}^{n} (y(i) - a - x(i))^2$$

(4.3)
Figure 29: A cloud of point in $\mathbb{R}^2$ with an approximating log–linear function in black and the generating function according to (4.2) in red.

and the solution is given by

\[
\hat{b} = \frac{\frac{1}{n} \sum_{i=1}^{n} x(i) Y(i) - \frac{1}{n} \sum_{i=1}^{n} x(i) \frac{1}{n} \sum_{i=1}^{n} Y(i)}{\frac{1}{n} \sum_{i=1}^{n} x(i)^2 - \left( \frac{1}{n} \sum_{i=1}^{n} x(i) \right)^2} \tag{4.4}
\]

\[
\hat{a} = \frac{1}{n} \sum_{i=1}^{n} Y(i) - \frac{1}{n} \sum_{i=1}^{n} x(i) \tag{4.5}
\]

as can be quickly verified. The results for the second model (4.2) follow on simply replacing $x(i)$ by $\log(x(i))$. On substituting (4.1) in (4.4) and (4.5) we obtain

\[
\hat{b} = b + \frac{\frac{1}{n} \sum_{i=1}^{n} x(i) \varepsilon(i) - \frac{1}{n} \sum_{i=1}^{n} \varepsilon(i)}{\frac{1}{n} \sum_{i=1}^{n} x(i)^2 - \left( \frac{1}{n} \sum_{i=1}^{n} x(i) \right)^2}
\]

\[
\hat{a} = a + \frac{\frac{1}{n} \sum_{i=1}^{n} \varepsilon(i) \frac{1}{n} \sum_{i=1}^{n} x(i)^2 - \frac{1}{n} \sum_{i=1}^{n} x(i) \frac{1}{n} \sum_{i=1}^{n} x(i) \varepsilon(i)}{\frac{1}{n} \sum_{i=1}^{n} x(i)^2 - \left( \frac{1}{n} \sum_{i=1}^{n} x(i) \right)^2}
\]

As the $\varepsilon(i)$ are Gaussian white noise with variance $\sigma^2$ it follows that $\hat{a} \sim \sigma$.
\( N(a, \Sigma(a)^2) \) and \( \hat{b} \sim N(b, \Sigma(b)^2) \) where
\[
\Sigma(a)^2 = \mathbb{E}((\hat{a} - a)^2) = \frac{\sigma^2}{n} \left( 1 + \frac{1}{n} \sum_{i=1}^{n} x(i)^2 - \left( \frac{1}{n} \sum_{i=1}^{n} x(i) \right)^2 \right) \tag{4.6}
\]
\[
\Sigma(b)^2 = \mathbb{E}((\hat{b} - b)^2) = \frac{\sigma^2}{n} \left( \frac{1}{n} \sum_{i=1}^{n} x(i)^2 - \left( \frac{1}{n} \sum_{i=1}^{n} x(i) \right)^2 \right) \tag{4.7}
\]
The residuals \( r(i), i = 1, \ldots, n \) are given by
\[ R(i) = Y(i) - \hat{a} - \hat{b} x(i) \]
which suggests
\[ \hat{\sigma}^2 = \frac{1}{n - 2} \sum_{i=1}^{n} R(i)^2 \tag{4.8} \]
as an estimate of \( \sigma \). If the factor \( n - 2 \) in (4.8) is replaced by \( n \) then we get the maximum likelihood estimator for \( \sigma \) based on the Gaussian model. It can be shown that \( (n - 2)\hat{\sigma}^2/\sigma^2 \) has a \( \chi^2 \)-squared distribution with \( n - 2 \) degrees of freedom. Moreover \( \hat{\sigma} \) is distributed independently of \( \hat{a} \) and \( \hat{b} \). From this it follows that \( \Sigma(a)^{-1}(\hat{a} - a)/\hat{\sigma} \) and \( \Sigma(b)^{-1}(\hat{b} - b)/\hat{\sigma} \) both have a \( t \)-distribution with \( n - 2 \) degrees of freedom. This allows simple test for various values of \( a \) and \( b \). Often one is interested in whether the co-variable \( b \) has any influence at all. Under the null hypothesis of not influence \( H_0 : b = 0 \) it follows that \( \Sigma(b)^{-1}\hat{b}/\hat{\sigma} \) has a \( t \)-distribution with \( n - 2 \) degrees of freedom. For a test of size 0.05 this leads to the critical value of \( qt(0.975, n - 2) \) and the null hypothesis is rejected if
\[
|\hat{b}| \geq qt(0.975, n - 2) \Sigma(b)\hat{\sigma}.
\]

### 4.1.2 The general case

Apart from generality and greater flexibility the general case has the advantage of clarity: it is easier to understand and interpret. We have observations \((y_i, \mathbf{x}_i^\top), I = 1, \ldots, n\), where \( \mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ik}) \) is the row vector of \( k \) covariates. The model is
\[
y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i, \quad 1 \leq i \leq n, \tag{4.9}
\]
where again \( \varepsilon_i \) denote Gaussian white noise with variance \( \sigma^2 \). This can be written in matrix form as
\[
y = X\beta + \varepsilon \tag{4.10}
\]
where

\[
y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}
\]

and

\[
\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.
\]

The simple straight line model of Section 4.1.1 can be written in this form by putting \( k = 2 \), \( x_{1j} = 1 \) and \( x_{2j} = x_j, j = 1, \ldots, n \). The least squares estimator which is, as always, the maximum likelihood estimate in the Gaussian model can be written as

\[
\text{minimize}_\beta \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x_{ij} \beta_j)^2 = \|y - X\beta\|^2 = (y - X\beta)^\top (y - X\beta).
\]

By differentiating we obtain

\[
\beta_{LS} = (X^\top X)^{-1} X^\top y. \tag{4.11}
\]

This requires that \( X^\top X \) be a non-singular matrix which is the case if the columns of \( X \) are linearly independent. If this is not so, some column depend linearly on the remaining and may be removed. That the solution to the least squares problem is given by (4.11) can also be seen as follows

\[
(y - X\beta)^\top (y - X\beta)
= (y - X\beta_{LS} - X(\beta - \beta_{LS}))^\top + (y - X\beta_{LS} - X(\beta - \beta_{LS}))
= (y - X\beta_{LS}^\top)(y - X\beta_{LS}) - 2(y - X\beta_{LS}^\top)X(\beta - \beta_{LS})
+ X(\beta - \beta_{LS})^\top X(\beta - \beta_{LS}) \tag{4.12}
\]

Now

\[
(y - X\beta_{LS})^\top X = (y - X(X^\top X)^{-1}X^\top y)^\top X = y^\top (X - X(X^\top X)^{-1})
= y^\top (I - X(X^\top X)^{-1}X^\top )X = Y^\top (X - X) = 0
\]

and on using this in (4.12) we obtain

\[
\|y - X\beta\|^2 = \|y - X\beta_{LS}\|^2 + \|X(\beta - \beta_{LS})\|^2 \tag{4.13}
\]
from which it clearly follows that the sum of squared deviations is minimized by \( \beta_{LS} \). This has a geometric interpretation. We have

\[
X\beta_{LS} = X(X^\top X)^{-1}Xy = Py
\]

with \( P = X(X^\top X)^{-1}X^\top \). Now \( P^\top = P \) and

\[
P^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = P
\]

so that \( P \) is a projection operator. In fact \( P = X(X^\top X)^{-1}X^\top \) projects onto the \( k \) dimensional linear subspace of \( \mathbb{R}^n \) spanned by the columns of \( X \). We have

\[
y = (I - P)y + Py
\]

and as \( (I - P)^\top P = (I - P)P = P - P^2 = P - P = 0 \) we see that (4.13) is nothing more than Pythagorüs’s theorem. On substituting \( y = X\beta + \epsilon \) in (4.11) we obtain

\[
\beta_{LS} = \beta + (X^\top X)^{-1}X\epsilon
\]

and using the fact that the \( \epsilon \) are white Gaussian noise with variance \( \sigma^2 \) we see

\[
\beta_{LS} \sim N(\beta, \sigma^2(X^\top X)^{-1}).
\]

(4.15)

The residuals are given by

\[
r = y - X\beta_{LS} = y - X(X^\top X)^{-1}Xy = (I - P)y = (I - P)\epsilon
\]

and hence

\[
\|r\|^2 = \epsilon^\top (I - P)\epsilon.
\]

(4.16)

It can be shown that \( \|r\|^2/\sigma^2 \) has a \( \chi^2 \)-squared distribution with \( n - k \) degrees of freedom and is distributed independently of \( \beta_{LS} \). This allows the usual tests of hypotheses and the construction of confidence intervals. In particular we have

\[
\Sigma_{ii}^{-1}(\beta_{LS,i} - \beta_i)/S_{n-k} \sim t_{n-k}
\]

(4.17)

where \( \Sigma_{ii} \) is the \( i \)th diagonal element of \( (X^\top X)^{-1} \) and \( S_{n-k}^2 = \|r\|^2/(n - k) \).

If

\[
|\beta_{LS,i} - b_i| \geq \Sigma_{ii}S_{n-k}qt(0.975, n - k)
\]

(4.18)

then the null hypothesis \( H_0 : \beta_i = b_i \) is rejected. A 0.95–confidence interval for \( \beta_i \) is given by

\[
[\beta_{LS,i} - \Sigma_{ii}S_{n-k}qt(0.975, n - k), \beta_{LS,i} + \Sigma_{ii}S_{n-k}qt(0.975, n - k)].
\]

(4.19)
We can also conclude that
\[ (\beta_{LS} - \beta)^T (X^T X) (\beta_{LS} - \beta) \sim \sigma^2 \chi^2_k \] (4.20)
so that
\[ \frac{1}{k} (\beta_{LS} - \beta)^T (X^T X) (\beta_{LS} - \beta)/S^2_{n-k} \sim F_{k,n-k}. \] (4.21)
In particular the null hypothesis of zero influence of the co-variates \( H_0 : \beta = 0 \) can be tested using the \( F \)-test
\[ \frac{1}{k} \beta_{LS}^T (X^T X) \beta_{LS}/S^2_{n-k} \geq F_{k,n-k}(0.95). \] (4.22)

4.1.3 The two-way table

The two-way analysis of variance can be treated as a linear regression problem as follow. We consider the case without interactions but with an arbitrary number of observations in each cell. To put it in the form \( y = X \beta \) we must first decide on some way of ordering the \( y_i \). We start with \( y_{ijk} \) and let the third index change fastest followed by the second. The order is \( y_{111}, y_{112}, \ldots, y_{11K}, y_{121}, y_{122}, \ldots, y_{12K}, \ldots, y_{IJK}, \ldots \).

As already mentioned in the section on the two-way table the model is not identifiable and one simple way of forcing identifiability is simple to eliminate \( a_1 \) by setting \( a_1 = 0 \). The first column of the \( X \)-matrix corresponding to \( a_2 \), the second to \( a_3 \), the \( I \)th column to \( b_1 \) and the \((I + J - 1)\)th column to \( b_J \). If \( y_{ijk} \) occurs in row \( t \) then the corresponding \( t \)th row of \( X \) is as follows. If \( i = 1 \) then the \((i + j - 1)\)th element of the row is 1 and the remainder are zero. If \( i \geq 2 \) the the \((i - 1)\)th and the \((I + j - 1)\)th elements of the row are 1 and the remainder are zero. This guarantees that the matrix \( X^T X \) is non-singular so that all coefficients can by estimated. Under this convention the vector \( \beta \) is given by
\[ \beta^T = (a_2, \ldots, a_I, b_1, \ldots, b_J). \]

With somewhat more effort interactions as defined in the classical manner can be included as follows. To ensure that \( X^T X \) is non-singular we put \( c_{11}, c_{21}, \ldots = c_{I1} = 0 \) and \( c_{11} = c_{12} = \ldots = c_{1J} = 0 \). The \( X \) matrix is extended so that the \((I + J - 1 + i - 1 + (j - 2)(I - 1))\)th column refers to \( c_{ij} \). If \( y_{ijk} \) is on row \( t \) then for \( 2 \leq i \leq I \) and \( 2 \leq j \leq J \) the first \( I + J - 1 \) values are as before and additionally the \((I + J - 1 + i - 1 + (j - 2)(I - 1))\)th element is 1, the remainder being 0. The vector \( \beta \) is given by
\[ \beta^T = (a_2, \ldots, a_I, b_1, \ldots, b_J, c_{22}, c_{32}, \ldots, c_{12}, c_{23}, \ldots, c_{IJ}). \]

All tests and confidence intervals can now be carried out as in the general linear regression model.

111
4.2 Breakdown and equivariance

4.2.1 Outliers and least-squares

Generate data according to

\[ Y(i) = 5 + 0.1x_1(i) + 5x_2(i) + Z(i), \quad i = 1, \ldots, 100, \]  

(4.23)

with \( x_1 = (1, 2, \ldots, 100)^\top \) and \( x_2 \) 100 i.i.d. \( \mathcal{N}(0,1) \) random variables. The standard least-squares estimates are

\[ 4.747357, 0.104378, 4.915645 \]

respectively. The plot of the residuals is shown in Figure 30.

Figure 30: Plot of least-squares residuals for data generated according to (4.23).

We now generate outliers by putting

\[ x_1(5(2:10)) = 2x_1(5(2:10)), \quad x_2(5(2:10)) = 25x_2(5(2:10)). \]  

(4.24)

The least squares estimates are

\[ 4.91767, 0.10334, 0.31907 \]

respectively. The plot of the residuals is shown in Figure 31.

Although the standard deviation of the residuals has increased quite considerably to 4.91895 the residual plot does not show any outliers. This is
Figure 31: Plot of least-squares residuals for data generated according to (4.23) and outliers generated as in (4.24).

known as the masking effect. the outliers alter the least-squares coefficients to such a degree that they are no longer detectable from an outliers plot. If we calculate the residuals based on the coefficients used to generate the data (4.23) then the outliers are clearly visible. This is shown in Figure 32.

4.2.2 Equivariance

Consider the linear regression problem with data \((y, X)\) with \(y \in \mathbb{R}^n\) and \(X\) an \(n \times k\)-matrix. The ordinary least-squares regression functional \(T_{LS}\) is given by

\[
T_{LS}((y, X)) = (X^\top X)^{-1}X^\top y.
\]

Suppose now \(y \to \alpha y\) with \(\alpha \neq 0\). Then

\[
T_{LS}((\alpha y, X)) = (X^\top X)^{-1}X^\top \alpha y = \alpha (X^\top X)^{-1}X^\top y = \alpha T_{LS}((y, X)).
\]

Similarly if \(y \to y + X\gamma\) then

\[
T_{LS}((y + X\gamma, X)) = (X^\top X)^{-1}X^\top (y + X\gamma) = (X^\top X)^{-1}X^\top y + \gamma = T_{LS}((y, X)) + \gamma.
\]
Figure 32: Plot of the residuals calculated using the coefficients in (4.23) but with the outliers in the outliers generated as in (4.24).
Finally if $\Gamma$ is a non-singular $k \times k$–matrix then
\[
T_{LS}((y, X\Gamma)) = ((X\Gamma)^\top X\Gamma)^{-1}(X\Gamma)^\top y \\
= (\Gamma^\top X^\top X\Gamma)^{-1}\Gamma^\top X^\top y \\
= \Gamma^{-1}(X^\top X)^{-1}(\Gamma^\top)^{-1}\Gamma^\top X^\top y \\
= \Gamma^{-1}(X^\top X)^{-1}X^\top y \\
= \Gamma^{-1}T_{LS}((y, X)).
\]

In general
\[
(y, X) \rightarrow (\alpha y + X\gamma, X\Gamma) \quad (4.25)
\]
then
\[
T_{LS}(\alpha y + X\gamma, X\Gamma) = \gamma + (\alpha + \Gamma^{-1})T_{LS}(y, X). \quad (4.26)
\]
The set of operations (4.25) form a group and the corresponding change in the parameter $\beta$ as in (4.26) is the corresponding equivariance structure.

**Definition 4.1.** A regression functional $T_R$ is called regression equivariant if

\[
T_R(\alpha y + X\gamma, X\Gamma) = \gamma + (\alpha + \Gamma^{-1})T_R(y, X)
\]

for all $\alpha \neq 0$, for all $\gamma \in \mathbb{R}^k$ and for all non-singular $k \times k$–matrices $\Gamma$.

The $L_1$–regression functional
\[
T_{L_1}(y, x) = \text{argmin}_\beta \sum_{i=1}^n |y_i - (X\beta)_i| \quad (4.27)
\]
is regression equivariant.

**Remark 4.2.** We have defined the regression functionals on the sample space and not on the space of probability distributions. This can be done as follows. Let $\mathcal{P}'$ be a family of probability measures on $\mathbb{R} \times \mathbb{R}^k$. The group $G$ of transformations on the sample space is as before
\[
g((y, x^\top)\top) = (\alpha y + x^\top \gamma, (\Gamma x)^\top)^\top \quad (4.28)
\]
where as before $\alpha \neq 0, \gamma \in \mathbb{R}^k$ and $\Gamma$ a non-singular $k \times k$–matrix. We suppose that $\mathcal{P}'$ is closed under the group operation, $P^g \in \mathcal{P}'$ for all $P \in \mathcal{P}'$. A regression functional $T_R$ is a mapping $T_R : \mathcal{P}' \rightarrow \mathbb{R}^k$ which is equivariant under the group operation: if $g \in G$ is given by (4.28) then for all $P \in \mathcal{P}'$
\[
T_R(P^g) = \gamma + (\alpha + \Gamma^{-1})T_R(P). \quad (4.29)
\]
Given data \((y_i, x_i^\top)^\top, i = 1, \ldots, n\), we define the corresponding empirical measure as before by

\[
\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(y_i, x_i^\top)^\top}
\]

and the definition of equivariance reduces to what we had before.

**Remark 4.3.** The least-squares estimator \(T_{LS}\) can be defined as

\[
T_{LS}(\mathbb{P}) = \arg\min_{\beta} \int (y - x^\top \beta)^2 \, d\mathbb{P}(y, x).
\]

This requires

\[
\mathbb{P}' = \{\mathbb{P} : \int \|z\|^2 \, d\mathbb{P}(z) < \infty\}.
\]

Similarly the \(L_1\) functional \(T_{L_1}\) is defined by

\[
T_{L_1}(\mathbb{P}) = \arg\min_{\beta} \int |y - x^\top \beta| \, d\mathbb{P}(y, x)
\]

and requires

\[
\mathbb{P}' = \{\mathbb{P} : \int \|z\| \, d\mathbb{P}(z) < \infty\}.
\]

### 4.2.3 Breakdown

We consider only the finite sample breakdown point. Given a sample \((y_i, x_i^\top)^\top, i = 1, \ldots, n\), of size \(n\) we denote the corresponding empirical measure by \(\mathbb{P}_n\). We are now allowed to alter \(m\) of the observations to give a sample \((y'_i, x_i'^\top)^\top, i = 1, \ldots, n\), with

\[
|\{i : (y'_i, x_i'^\top)^\top \neq (y_i, x_i^\top)^\top\}| = m.
\]

We denote the corresponding empirical distribution by \(\mathbb{P}_n^m\) and define the finite sample breakdown point \(\epsilon^*(T_L, \mathbb{P}_n)\) by

\[
\epsilon^*(T_L, \mathbb{P}_n) = \min_m \left\{ \frac{m}{n} : \sup_{\mathbb{P}_n^m} \|T_R(\mathbb{P}_n^m) - T_R(\mathbb{P}_n)\| = \infty \right\}. \tag{4.30}
\]

We note that in the definition of breakdown point the covariate \(x_i\) can also be changed and not only the values of the \(y_i\).

We have the following theorem.
**Theorem 4.4.** For a data set \((y_i, x_i^\top)^\top, i = 1, \ldots, n\), with empirical measure \(P_n\) we define

\[
\Delta(P_n) = \max_L P_n(L)
\]

where \(L\) is any linear subspace of \(\mathbb{R}^{k+1}\) of dimension at most \(k\) then for an equivariant regression functional \(T_R\) we have

\[
\epsilon^*(T_R, P_n) \leq \left\lfloor \frac{n(1 - \Delta(P_n)) + 1}{2} \right\rfloor / n.
\]

**Remark 4.5.** We always have \(\Delta(P_n) \geq k/n\). If \(\Delta(P_n) = k/n\) then the points are said to be “in general position”.

**Example 4.6.** For \(T_R = T_{LS}\) we clearly have \(\epsilon^*(T_{LS}, P_n) = 1/n\). This follows from the representation \(T_{LS}(P_n) = (X^\top X)^{-1}X^\top y\). We keep \(X\) fixed and let \(y_1\) for example tend to infinity.

**Example 4.7.** For \(T_R = T_{L_1}, k = 1\), we also have \(\epsilon^*(T_{L_1}, P_n) = 1/n\). This is perhaps not so obvious as in the location case minimizing the \(L_1\)-norm gives the median which has the highest possible breakdown point. We write

\[
s(\beta) = |y_1' - \beta x_1'| + \sum_{i=2}^{n} |y_i - \beta x_i|
\]

where we alter the observation first observation. Suppose now that \(x_1' > \sum_{i=2}^{n} |x_i|\) and that \(\beta\) is such that \(y_1' > \beta x_1'\). Then for \(\eta > 0\) with \(y_1' > (\beta + \eta)x_1'\) we have

\[
s(\beta + \eta) - s(\beta) \leq -\eta x_1' + \sum_{i=2}^{n} \left| y_i - (\beta + \eta)x_i \right| - \left| y_i - \beta x_i \right| \leq -\eta x_1' + \eta \sum_{i=2}^{n} |x_i| < 0.
\]

This implies that the solution \(\beta^*\) must satisfy \(y_1' \leq \beta^* x_1'\). Similarly it must also satisfy \(y_1' \geq \beta^* x_1'\) and hence the \(L_1\)-norm minimizer is \(\beta^* = y_1'/x_1'\).

This leads to the question as to whether there exists a regression functional with a finite sample breakdown point given by the upper bound in Theorem 4.4.
4.3 The Least Median of Squares (LMS)

The genesis of this idea is Tukey’s shorth. It was applied to the regression problem by Hampel who proposed the following. Given data \((y_i, x_i^\top)^\top, i = 1, \ldots, n\) and for a given \(\beta\) we form the residuals

\[ r_i(\beta) = y_i - x_i^\top \beta, \quad i = 1, \ldots, n. \]

Hampel proposed determining \(\beta\) as the solution of the problem

\[ \arg\min_{\beta} \text{med}\{ |r_1(\beta)|, \ldots, |r_n(\beta)| \}. \]  \(\text{(4.31)}\)

The connection with Tukey’s shorth is clear. In the case of simple regression \(y = a + bx\) we take parallel lines a vertical distance of \(h\) apart and then choose the lines to minimize \(h\) under the restriction that at least \(\lfloor n/2 \rfloor + 1\) of the data points \((y_i, x_i)\) lie on or between the two lines. The method was propagated by Rousseeuw who introduced the name “least median of squares”. Clearly (4.31) is equivalent to

\[ \arg\min_{\beta} \text{med}\{ |r_1(\beta)|^2, \ldots, |r_n(\beta)|^2 \}. \]  \(\text{(4.32)}\)

The use of squares was to relate it to the usual least-squares method. We denote Hampel’s functional by \(T_{LMS}\). We have the following theorem.

**Theorem 4.8.** The finite sample breakdown point of Hampel’s regression functional \(T_{LMS}\) is

\[ \epsilon^*(T_{LMS}, \mathbb{P}_n) = (\lfloor n/2 \rfloor - n \Delta(\mathbb{P}_n) + 2)/n. \]  \(\text{(4.33)}\)

**Remark 4.9.** The finite sample breakdown point of \(t_{LMS}\) as defined by (4.31) is slightly lower than the upper bound of Theorem 4.4. The upper bound can however be attained by a slight variation of \(T_{LMS}\). We define it by

\[ T_{LMS} = \arg\min_{\beta} d_q(\beta) \]  \(\text{(4.34)}\)

where

\[ d_i(\beta) = |r_i(\beta)| = |y_i - x_i^\top \beta| \]

\[ d(1)(\beta) \leq d(2)(\beta) \leq \ldots \leq d(n)(\beta) \]

\[ q = \lfloor n/2 \rfloor + \lfloor (n \Delta(\mathbb{P}_n) + 1)/2 \rfloor. \]

With this choice of quantile we have

\[ \epsilon^*(T_{LMS}, \mathbb{P}_n) = \left[ \frac{n(1 - \Delta(\mathbb{P}_n)) + 1}{2} \right]/n. \]
No functional is of use unless it can be calculated at least approximately. We now turn to this problem for the case of simple regression $y = a + bx$. Suppose the solution is $a_{\text{LMS}}$ and $b_{\text{LMS}}$ and let $\text{med}\{|r_1(\beta)|, \ldots, |r_n(\beta)|\} = h$. Then at least $\lfloor (n/2) \rfloor + 1$ observations lie between the two lines

$$y = a_{\text{LMS}} + b_{\text{LMS}}x + h, \quad y = a_{\text{LMS}} + b_{\text{LMS}}x - h.$$ 

Suppose that no data point lies on any of the two lines. then it is clear that we can reduce $h$ until one data point is either on the upper or lower line. If there is one on the upper lines but none on the lower than we can move the lower lines upwards which will decrease the perpendicular distance between the two lines. we continue until we hit a data point. at this stage we have a solution with at least one point on the upper line and one point on the lower line. We now rotate both lines through the same angle. This does not alter the perpendicular distance. We keep doing this until one of the lines hits another data point. Whether this point was initially between the lines or is a point from outside the distance between the two lines has not decreased. From this it is clear that a solution can always be put into this form, at least two data points on one line and at least one data point on a parallel line. This suggests the following algorithm. Select all pairs of different points and for each pair calculate the line through them. Calculate the residuals from this line and then work out how much the line must be moved upwards or downwards so the $\lfloor n/2 \rfloor + 1$ data points lie between the two lines. This gives the vertical distance between the lines. Finally that pair which minimize the vertical distance can be used to calculate the LMS line. This algorithm is of order $O(n^3)$ because we have $O(n^2)$ pairs of points and calculating the required parallel lines is of order $O(n)$. This is one of the weaknesses of the LMS. In $k$ dimensions the complexity is of order $O(n^{k+1})$ which makes an exact solution impractical for all but small $k$ and $n$. There are good heuristic algorithms which are based on a random section of points and also more sophisticated ideas but in the last analysis they cannot be guaranteed to give the correct answer. Other weakness of the LMS and other available high-breakdown functionals is that they are inherently unstable. Just as the shorth may not have a unique solution, neither does LMS. Even if the solution is unique and continuous at some distribution it will be discontinuous at some other distribution arbitrarily close. All of which means that such high breakdown methods are not ideally suited to the construction of confidence intervals or other more delicate forms of statistical analysis. On the other hand they are extremely useful in a data analytical sense. It is good statistical practice to always run a robust regression alongside a least squares regression. If the results are similar then one may feel somewhat happier about the linear model. if they are dissimilar, then the data exhibit
Figure 33: Plot of the residuals calculated from the LMS fit with the data generated as in (4.23) and with the outliers generated as in (4.24).

a structure not captured by the linear model. R has a robust library:

```r
> library(lqs)
> tmp <- lmsreg(tmpy ~ I(tmpx1) + I(tmpx2))
> tmp$coef
(Intercept)   I(tmpx1)   I(tmpx2)
4.76729825  0.09963781  5.07685616
> plot(tmp$res)
```

Figure 33 shows the residuals from the LMS fit. The outliers are clearly visible and the coefficients close to the one used to generate the data.
References


