MEASURES OF LOCATION AND SCALE IN R

P. L. Davies

Eindhoven, August 2007

Reading List


Contents

1 Location and Scale in \( \mathbb{R} \) ............................................. 2
  1.1 Measures of location ............................................. 2
    1.1.1 The mean ............................................. 2
    1.1.2 The median ............................................. 3
    1.1.3 The shorth ............................................. 4
    1.1.4 What to use? ............................................. 5
  1.2 Measures of scale or dispersion .................................. 6
    1.2.1 The standard deviation ................................... 6
    1.2.2 The MAD ............................................. 6
    1.2.3 The length of the shorth ................................ 7
    1.2.4 The interquartile range IQR ............................... 7
    1.2.5 What to use? ............................................. 8
  1.3 Boxplots ......................................................... 9
  1.4 Empirical distributions .......................................... 11
    1.4.1 Definition ............................................. 11
    1.4.2 Empirical distribution function ......................... 11
  1.5 Equivariance considerations .................................... 13
    1.5.1 Transforming data ....................................... 13
    1.5.2 The group of affine transformations .................... 13
    1.5.3 The group of affine transformations .................... 14
    1.5.4 Measures of location ................................... 14
    1.5.5 Measures of scale ....................................... 15
  1.6 Outliers and breakdown points ................................... 15
    1.6.1 Outliers ............................................. 15
    1.6.2 Breakdown points and equivariance ..................... 16
    1.6.3 Identifying outliers ................................... 20
  1.7 M-measures of location and scale ................................ 25
    1.7.1 M-measures of location ................................... 25
    1.7.2 M-measures of scale ..................................... 28
    1.7.3 Simultaneous M-measures of location and scale .......... 30
1 Location and Scale in \( \mathbb{R} \)

1.1 Measures of location

1.1.1 The mean

We start with measures of location for simple data sets \( x_n = (x_1, \ldots, x_n) \) consisting of \( n \) real numbers. By far the most common measure of location is the arithmetic mean

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

**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.

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

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>Pot</th>
<th>Cross</th>
<th>Self</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>23.5</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>17.4</td>
<td>20.4</td>
</tr>
<tr>
<td>II</td>
<td>22.0</td>
<td>19.2</td>
</tr>
<tr>
<td></td>
<td>20.0</td>
<td>18.4</td>
</tr>
<tr>
<td>III</td>
<td>Cross</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>Self</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.0</td>
</tr>
<tr>
<td>IV</td>
<td>Cross</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>Self</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.0</td>
</tr>
</tbody>
</table>

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

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

The mean of the differences is 2.62667.

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

$$x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)}$$

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

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

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

$$\text{med}(2.16, 2.21, 2.15, 2.05, 2.06, 2.04) = \frac{(x_{m}) + x_{(m+1)})}{2} \quad = \frac{(x_{(3)} + x_{(4)})}{2} = \frac{(2.06 + 2.15)}{2} \quad = \frac{4.21}{2} = 2.105.$$  

In R the command is

```
> 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) $\lfloor x \rfloor$ spoken “floor $x$” is the largest integer $n$ such that $n \leq x$. For example $\lfloor 2.001 \rfloor = 2, \lfloor 5 \rfloor = 5, \lfloor 2.3 \rfloor = 2$.

(b) $\lceil x \rceil$ spoken “ceiling $x$” is the smallest integer $n$ such that $x \leq n$.

(c) For any number $x$, it holds that $\lceil x/2 \rceil + 1$ is the smallest integer $n$ which is strictly greater than $x/2$.

(d) The R command is floor:

```
> floor(2.39)
[1] 2
```
For a sample \((x_1, \ldots, x_n)\) of size \(n\) we consider all intervals \(I = [a, b]\) that contain at least \(\lceil n/2 \rceil + 1\) 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 \(\lceil n/2 \rceil + 1\) 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. Firstly there may be more than one such interval. Secondly there may be only one such interval but the number of observations in this interval is not clear. This is the case if at least one of the endpoints is a repeated observation. We cannot remove the first indeterminacy but the fact should be reported. The second can be removed by including all observations which lie in the interval according to multiplicity.

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

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

The sample size is \(n = 27\) so \(\lceil n/2 \rceil + 1 = \lceil 14 \rceil = 14\). 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 is of length 0.14, namely \([1.92, 2.06]\) and contains the 15 observations

\begin{align*}
1.92 & \quad 1.92 & \quad 1.93 & \quad 1.99 & \quad 1.99 & \quad 2.01 & \quad 2.02 & \quad 2.02 \\
2.03 & \quad 2.04 & \quad 2.05 & \quad 2.05 & \quad 2.06 & \quad 2.06 & \quad 2.06
\end{align*}

The mean of these observations is 2.048.

**Exercise 1.6.** Write an \texttt{R}-programme 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 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 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 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 MAD($x_n$) = 0.1. If 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 `mad(., 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 MAD($X$) = 0.6744898$\sigma$ so that 1.482602 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 lshorth.

**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 lshorth($x_n$) = 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

$$Q1 = (x(\lceil n/4 \rceil) + x(\lfloor n/4 \rfloor + 1))/2, \quad Q3 = (x(\lceil 3n/4 \rceil) + x(\lfloor 3n/4 \rfloor + 1))/2,$$

(1.7)

The interquartile range is then simply $\text{IQR} = Q3 - Q1$. 

7
**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 \( \text{IQR} = 2.08 - 1.82 = 0.16 \). If you use R you get

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

[1] 0.145

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.** Within a production line the deviation of target of the length of newly produced nails are measured. A small sample consist of the following observed values:

\[
1.3 \quad 2.4 \quad -0.5 \quad 0.2 \quad -1.3 \quad -1.8 \quad 0.6 \quad 1.3 \quad 2.1 \quad -1.6
\]

(i) Estimate the standard deviation by the following statistics:

\[
\hat{\sigma} = \sqrt{\frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})^2}
\]

\[
\tilde{\sigma} = 1.483 \text{MAD}
\]

\[
\sigma^* = \frac{x_{(10)} - x_{(1)}}{d_{10}}
\]

where \( x_{(1)}, \ldots, x_{(10)} \) are the ordered (from small to large) values of the sample and \( 1/d_{10} = 0.325 \).

(ii) Assume that instead the following values had been observed.

\[
1.3 \quad 2.4 \quad -0.5 \quad 0.2 \quad -1.3 \quad -1.8 \quad 6.0 \quad 1.3 \quad 2.1 \quad -1.6
\]

Repeat the calculations from 1) and discuss the results.
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). 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 $1.5 \text{IQR}$ 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 following data are taken from an interlaboratory test and show the measurements returned by 14 laboratories of the amount of mercury (micrograms per litre) in a water sample. Each laboratory was re-
Figure 2: The boxplots of the interlaboratory test data of Example 1.13.

quired to give four measurements. The boxplot 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.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab. 1</td>
<td>2.2</td>
<td>2.215</td>
<td>2.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Lab. 2</td>
<td>2.53</td>
<td>2.505</td>
<td>2.505</td>
<td>2.47</td>
</tr>
<tr>
<td>Lab. 3</td>
<td>2.3</td>
<td>3.2</td>
<td>2.8</td>
<td>2.0</td>
</tr>
<tr>
<td>Lab. 4</td>
<td>2.33</td>
<td>2.35</td>
<td>2.3</td>
<td>2.325</td>
</tr>
<tr>
<td>Lab. 5</td>
<td>2.1</td>
<td>2.68</td>
<td>2.46</td>
<td>1.92</td>
</tr>
<tr>
<td>Lab. 6</td>
<td>2.692</td>
<td>2.54</td>
<td>2.692</td>
<td>2.462</td>
</tr>
<tr>
<td>Lab. 7</td>
<td>2.292</td>
<td>2.049</td>
<td>2.1</td>
<td>2.049</td>
</tr>
<tr>
<td>Lab. 8</td>
<td>2.120</td>
<td>2.2</td>
<td>1.79</td>
<td>2.12</td>
</tr>
<tr>
<td>Lab. 9</td>
<td>1.967</td>
<td>1.899</td>
<td>2.001</td>
<td>1.956</td>
</tr>
<tr>
<td>Lab. 10</td>
<td>3.8</td>
<td>3.751</td>
<td>3.751</td>
<td>3.797</td>
</tr>
<tr>
<td>Lab. 11</td>
<td>2.696</td>
<td>2.22</td>
<td>2.01</td>
<td>2.57</td>
</tr>
<tr>
<td>Lab. 12</td>
<td>2.191</td>
<td>2.213</td>
<td>2.21</td>
<td>2.228</td>
</tr>
<tr>
<td>Lab. 13</td>
<td>2.329</td>
<td>2.252</td>
<td>2.252</td>
<td>2.26</td>
</tr>
<tr>
<td>Lab. 14</td>
<td>2.37</td>
<td>2.42</td>
<td>2.33</td>
<td>2.36</td>
</tr>
</tbody>
</table>
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

$$\delta_x(B) = 1_{\{x \in B\}}.$$  

For any function $f : \mathbb{R} \to \mathbb{R}$ we have

$$\int_{\mathbb{R}} f(u) \, d\delta_x(u) = f(x).$$

Given a sample $(x_1, \ldots, x_n)$ we define the corresponding empirical measure $\mathbb{P}_n$ by

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}. \tag{1.8}$$

The empirical distribution $\mathbb{P}_n$ is a probability measure and for any subset $B$ of $\mathbb{R}$ we have

$$\mathbb{P}_n(B) = \frac{1}{n} \times \text{number of points } x_i \text{ in } B.$$  

Furthermore for any function $f : \mathbb{R} \to \mathbb{R}$ we have

$$\int_{\mathbb{R}} f(u) \, d\mathbb{P}_n(u) = \frac{1}{n} \sum_{i=1}^{n} f(x_i). \tag{1.9}$$

1.4.2 Empirical distribution function

Given data $x_n = (x_1, \ldots, x_n)$ the empirical distribution function $F_n$ is defined by

$$F_n(x) = \mathbb{P}_n((-\infty, x])$$

$$= \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \leq x\}} \tag{1.10}$$

Figure 1.8 shows the empirical distribution functions of the data sets of the Examples 1.1 and 1.2.
Figure 3: The upper panel shows the empirical distribution function of the data of Example 1.1. The lower panel shows the empirical distribution function of the data of Example 1.2.

**Exercise 1.15.** Let \( g : \mathbb{R} \to \mathbb{R} \) be defined by \( g(x) := x^2 1_{[0,10]}(x) \), where

\[
1_{[0,10]}(x) = \begin{cases} 
1 & \text{for } x \in [0, 10] \\
0 & \text{elsewhere}
\end{cases}
\]

Calculate \( \int g \, dP \) if \( P \) is

(i) the uniform distribution on \([8, 12]\)

(ii) Dirac measure at \( x = 5 \)

(iii) binomial distribution with \( n = 2 \) and \( p = 0.25 \)

(iv) a probability measure on \( M = \{1, 2, \ldots, 15\} \) defined by \( P(B) = \frac{|B\cap M|}{|M|} \), where \(|.|\) denotes the cardinality.
1.5 Equivariance considerations

1.5.1 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 1.16.** Given a probability distribution \( P \) on (the Borel sets of) \( \mathbb{R} \) and a (measurable) transformation \( f : \mathbb{R} \rightarrow \mathbb{R} \) we define the transformed distribution \( P^f \) by

\[
P^f(B) := P(\{x : f(x) \in B\}) = P(f^{-1}(B)). \tag{1.11}
\]

**Exercise 1.17.** Show that \( P^f \) is a probability distribution.

In many situations we have 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.

1.5.2 The group of affine transformations

If we have a data set \( x_n = (x_1, \ldots, x_n) \) and transform it using a function \( f : \mathbb{R} \rightarrow \mathbb{R} \), then 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 distribution of \( x'_n \) is given by

\[
P'_n(B) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x'_i \in B\}}
\]

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

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

\[
= P_n(f^{-1}(B)) = P^f_n(B)
\]

so that \( P'_n = P^f_n \).
1.5.3 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 centimetres. If the same heights are measured in inches \( x'_n = (x'_1, \ldots, x'_n) \), then to a first approximation we have

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

The basis from which the heights are measured may also change. Instead of being measured from, for example, sea level 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 \( A : \mathbb{R} \rightarrow \mathbb{R} \) of the form

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

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

**Exercise 1.18.** Show that \( \mathcal{A} \) is indeed a group.

1.5.4 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

\[
\text{mean}(x'_n) = \bar{x}'_n = A(\bar{x}_n) = a\bar{x}_n + b = a \text{mean}(x_n) + b. \tag{1.12}
\]

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

\[
T_L(A(x_n)) = A(T_L(x_n)). \tag{1.13}
\]

Such measures are called location equivariant.
1.5.5 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

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

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

\[
T_S(A(x_n)) = |a|(T_L(x_n)), \quad A(x) = ax + b. \tag{1.14}
\]

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 anal-
ysis. 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 1.19. We consider the copper data of Example 1.1. The mean,
standard deviation, median and MAD of the data are 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 for this altered sample are

\[ 3.435926, \quad 5.053907, \quad 2.03, \quad 0.1. \]

We look as 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 in 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

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

We define the finite sample breakdown point of \( T \) at the sample \( x_n \) by

\[ \epsilon^*(T, x_n) = \min_{k} \left\{ \frac{k}{n} : \sup_{x^k_n} |T(x_n) - T(x^k_n)| = \infty \right\} \quad (1.15) \]

**Example 1.20.** 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 1.21.** The median has a breakdown point \( \epsilon^*(\text{med}, x_n) = \left\lfloor \frac{(n+1)/2}{n} \right\rfloor \) 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

\[ \sup_{x^k_n} |\text{med}(x^k_n) - \text{med}(x_n)| = \infty \]

and hence

\[ \epsilon^*(\text{med}, x_n)) = m/2m = 1/2 = \left\lfloor \frac{(n+1)/2}{n} \right\rfloor. \quad (1.16) \]
**Exercise 1.22.** Show that (1.16) 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 1.23.** Let $T_L$ be a measure of location, that is affinely equivariant as defined by (1.13). 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

$$
x'_i = x_i, \ i = 1, \ldots, m, \quad x'_{m+i} = x_i + \lambda, \ i = 1, \ldots, m,
$$

$$
x''_i = x_i - \lambda, \ i = 1, \ldots, m, \quad x''_{m+i} = x_i, \ i = 1, \ldots, m.
$$

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

$$
\varepsilon^*(T_L, x_n) \leq m/2m = 1/2 = [(n + 1)/2]/n.
$$

□

**Remark 1.24.** In Theorem 1.23 we did not make full use of the affine equivariance. All we required was the translation equivariance.

**Remark 1.25.** Theorem 1.23 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 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
\[
d(s_1, s_2) = |\log(s_1) - \log(s_2)|
\] (1.17)
on \mathbb{R}_+.

The finite sample breakdown point for measures of scale is defined by
\[
\varepsilon^*(T_S, x_n) = \min \{k/n : \sup_{x_n^k} |\log(T_S(x_n^k)) - \log(T_S(x_n))| = \infty\}. \tag{1.18}
\]
with the convention that \(\varepsilon^*(T_S, x_n) = 0\) if \(T_S(x_n) = 0\).

**Exercise 1.26.** The following two samples are given:

(i) 1.0 1.8 1.3 1.4 1.9 1.1 2.0 1.6 1.7 1.2 0.9
(ii) 1.0 1.8 1.3 1.3 1.9 1.1 1.3 1.6 1.7 1.3 1.3

Calculate the MAD for both samples. Then replace individual observations by different values and repeat the calculation. How many observations must be replaced to get arbitrarily close to the boundary of the parameter space \([0, \infty)\)?

In order to state the theorem for measures of scale corresponding to Theorem 1.23 we require
\[
\Delta(x_n) := \max_x \mathbb{P}_n(\{x\}) = \max_x |\{j : x_j = x\}|/n. \tag{1.19}
\]

**Theorem 1.27.** 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
\[
\varepsilon^*(T_S, x_n) \leq \left[(n(1 - \Delta(x_n)) + 1)/2\right]/n.
\]

**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
\[
x'_i = x_i, \quad 1 \leq i \leq l,
x'_{i+l} = x_{i+l}, \quad 1 \leq i \leq m,
x'_{i+m+l} = A^{-1}(x_{i+l}), \quad 1 \leq i \leq m,
\]
and
\[ x''_i = A(x_i) = x_i, \quad 1 \leq i \leq l, \quad x''_{l+i} = A(x_{l+i}), \quad 1 \leq i \leq m, \]
\[ x''_{l+m+i} = x_{l+i}, \quad 1 \leq i \leq m. \]
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 \( a \to \infty \) and \( a \to 0 \) 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
\[ \varepsilon^*(T_s, x_n) \leq m/2m = m/2 = [(n(1 - \Delta(P_n)) + 1)/2]/n. \]

Exercise 1.28. Prove Theorem 1.27 for the case \( n - l = 2m + 1 \).

Example 1.29. The standard deviation functional \( sd \) has a breakdown point \( \varepsilon^*(sd, x_n) = 1/n \) 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 1.30. 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 1.27. 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
\[ \varepsilon^*(\text{MAD}, x_n) \leq (m + 1 - l)/n = [n/2 + 1]/n - \Delta(x_n). \]
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 \([n+1]/2\) observations to \(\infty\) and hence
\[
\epsilon^*(\text{MAD}, x_n) \leq \min\{[n/2+1]/n - \Delta(x_n), [(n+1)/2]/n\} = [n/2 + 1]/n - \Delta(x_n)
\]
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
\[
\epsilon^*(\text{MAD}, x_n) = [n/2 + 1]/n - \Delta(x_n).
\]
As already mentioned this is less than the upper bound of Theorem 1.27. 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
\[
T_S(x_n) = \max_i |x_i - \text{mean}(x_n)|/\text{sd}(x_n) \quad (1.20)
\]
with critical value \(c(T_S, n, \alpha)\) for a test of size \(\alpha\). The distribution of the test statistic \(T_S\) is calculated under the normal distribution. The affine invariance of \(T_S\) 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 22.0. The value of the
test statistic is now 3.478 and the observations 22.0 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, 3.686, 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 1.31.** Let $X_i, i = 1, \ldots, n$ be i.i.d. $\mathcal{N}(0, 1)$ random variables.
Show that
\[
\lim_{n \to \infty} P \left( \max_{1 \leq i \leq n} |X_i| \leq \sqrt{\tau \log(n)} \right) = \begin{cases} 
0, & \tau < 2, \\
1, & \tau \geq 2.
\end{cases}
\]

**Exercise 1.32.** 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 1.33.** 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 remov-
ing 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 calcu-
late 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 $\text{const} = 1$ version of the $\text{mad}$ 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

\[
\begin{align*}
1.70 & \quad 1.86 & \quad 1.88 & \quad 1.88 & \quad 1.90 & \quad 1.92 & \quad 1.92 & \quad 1.93 & \quad 1.99 \\
1.99 & \quad 2.01 & \quad 2.02 & \quad 2.02 & \quad 2.03 & \quad 20.4 & \quad 20.5 & \quad 20.5 & \quad 20.6 \\
20.6 & \quad 20.6 & \quad 20.8 & \quad 21.3 & \quad 21.3 & \quad 21.5 & \quad 21.6 & \quad 22.0 & \quad 22.1.
\end{align*}
\]

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. 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, we calculate the simulated values and store them. The values are given in Table 1 and Figure 4 shows a plot of the values of $c(TH, n, 0.1)$ for $n = 3(1)19$.

**Exercise 1.34.** 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 $c(TH, n, \alpha)$ for large $n$. 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

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

Now if $x$ is chosen so that

$$1 - \alpha = P(\max_i |X_i| \leq x) = P(|X_1| \leq x)^n$$
\[ \alpha = 0.1, 0.05, 0.01 \]

\begin{table}[h]
\centering
\begin{tabular}{|c|ccc|}
\hline
n & $\alpha = 0.1$ & $\alpha = 0.05$ & $\alpha = 0.01$ \\
\hline
3 & 15.40 & 32.54 & 158.97 \\
4 & 7.11 & 10.19 & 21.82 \\
5 & 7.84 & 11.58 & 26.42 \\
6 & 6.18 & 8.07 & 13.53 \\
7 & 6.78 & 9.17 & 16.28 \\
8 & 5.84 & 7.23 & 11.98 \\
9 & 5.99 & 7.45 & 12.16 \\
10 & 5.58 & 6.77 & 10.35 \\
11 & 5.68 & 6.92 & 11.06 \\
12 & 5.37 & 6.47 & 9.34 \\
13 & 5.52 & 6.58 & 9.38 \\
14 & 5.28 & 6.16 & 8.46 \\
15 & 5.44 & 6.40 & 8.66 \\
16 & 5.25 & 6.12 & 8.17 \\
17 & 5.36 & 6.20 & 8.42 \\
18 & 5.21 & 6.01 & 7.91 \\
19 & 5.26 & 6.02 & 8.00 \\
\hline
\end{tabular}
\caption{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.}
\end{table}

we see that

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

so that

\[ c(TH, n, \alpha) \approx \Phi^{-1} \left( \frac{(1 + (1 - \alpha)^{1/n})/2}{0.67449} \right) \]  \hfill (1.22)

On putting

\[ \gamma(TH, n, \alpha) = 0.67449 c(TH, n, \alpha)/\Phi^{-1} \left( \frac{(1 + (1 - \alpha)^{1/n})/2}{0.67449} \right) \]  \hfill (1.23)

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 $\lim_{n \to \infty} \log(\gamma(TH, n, \alpha) - 1) = 0$. The R command for $\Phi^{-1}(p)$ is

\[ \texttt{> qnorm(p)} \]

so that $\Phi^{-1} \left( \frac{(1 + (1 - \alpha)^{1/n})/2}{0.67449} \right)$ can be easily calculated. The next question is which values of $n$ to use. As we expect the dependence to be in $\log(n)$
We choose the values of $n$ such that $\log(n)$ is approximately linear. In the following we put $n = 20 \cdot 1.25^{0.19}$. 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

$$c(TH, n, 0.1) \approx 1.48(1 + 3.53 n^{-0.87}) \Phi^{-1} \left(\frac{(1 + 0.91/n)}{2}\right).$$  \hspace{1cm} (1.24)
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.

The corresponding results for $c(TH, n, 0.05)$ and $c(TH, n, 0.01)$ are

\[
c(TH, n, 0.05) \approx 1.48(1 + 4.65 n^{-0.88}) \Phi^{-1} \left(1 + 0.95^{1/n}/2\right), \quad (1.25)
\]
\[
c(TH, n, 0.01) \approx 1.48(1 + 8.10 n^{-0.93}) \Phi^{-1} \left(1 + 0.99^{1/n}/2\right). \quad (1.26)
\]

**Exercise 1.35.** Repeat the simulations and approximations for the outlier identifier based on the shorth and the length of the shortest half-interval:

\[
TSH(x_n) = \max_i |x_i - \text{shrh}(x_n)|/\text{lshrh}(x_n).
\]

### 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 \( P_n \) denote the empirical measure of the sample \( x_n \). Then if we use (1.9) with \( f(x) = \iota(x) = x \) we see that

\[
\text{mean}(x_n) = \int_{\mathbb{R}} \iota(x) \, dP_n(x) = \int_{\mathbb{R}} x \, dP_n(x).
\]

From this it follows that

\[
\int_{\mathbb{R}} \iota(x - \text{mean}(x_n)) \, dP_n(x) = 0 \tag{1.27}
\]

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

\[
\text{sgn}(x) = \begin{cases} 
-1, & x < 0, \\
0, & x = 0, \\
1, & x > 0
\end{cases} \tag{1.28}
\]

we see that the median satisfies

\[
\int_{\mathbb{R}} \text{sgn}(x - \text{med}(x_n)) \, dP_n(x) = 0. \tag{1.29}
\]

This leads to a general class of location measures \( T_L \) defined by

\[
\int_{\mathbb{R}} \psi(x - T_L(x_n)) \, dP_n(x) = 0. \tag{1.30}
\]

**Definition 1.36.** Any location measure \( T_L \) defined as a solution of (1.30) is called an M-measure or M-estimator of location.

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

**Theorem 1.37.** Let \( \psi: \mathbb{R} \to \mathbb{R} \) be an odd, continuous, bounded and strictly increasing function. Then the equation

\[
\int_{\mathbb{R}} \psi(x - m) \, dP_n(x) = 0
\]

has a unique solution \( m =: T_L(x_n) \) for every sample \( x_n \).
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

\[
\Psi(m) = \int_{\mathbb{R}} \psi(x - m) \, d\mathbb{P_n}(x)
\]

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

Example 1.39. The function \( \psi_c \) given by

\[
\psi_c(x) = \frac{\exp(cx) - 1}{\exp(cx) + 1}, \quad x \in \mathbb{R}, \quad c > 0,
\]

satisfies the conditions of Theorem 1.37. We note

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

and

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

This will be our standard \( \psi \)-function where we have the freedom to choose the tuning constant \( c \).

To be of use one must to calculate the M-measure for data sets. 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 \( T_L(x_n) = m_0 \) and terminate.

(c) If \( \Psi(m_0) < 0 \) put \( m_2 = m_0 \), choose \( k \) so that \( \Psi(m_0 - ks_0) > 0 \) and put \( m_1 = m_0 - ks_0 \).

(d) If \( \Psi(m_0) > 0 \) put \( m_1 = m_0 \), choose \( k \) so that \( \Psi(m_0 + ks_0) < 0 \) and put \( m_2 = m_0 + ks_0 \).

(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 $T_L(x_n) = (m_1 + m_2)/2$ and terminate.

**Exercise 1.40.** Write an R-program to calculate $T_L$ for the $\psi$-function of (1.39).

**Exercise 1.41.** Show that if $\psi$ satisfies the conditions of Theorem 1.37 then $T_L$ is translation equivariant but not affine equivariant.

**Exercise 1.42.** Suppose that $\psi$ satisfies the conditions of Theorem 1.37. Calculate the finite sample breakdown point $\varepsilon^* (T_L, x_n)$ of $T_L$.

To make $T_L$ affine equivariant we must augment (1.30) by an auxiliary measure of scale $T_S$. This leads to solving

$$
\int_{\mathbb{R}} \psi\left(\frac{x - T_L(x_n)}{T_S(x_n)}\right) d\mathbb{P}_n(x) = 0
$$

(1.31)

for $T_L(x_n)$. A good auxiliary measure of scale is the MAD which leads to

$$
\int_{\mathbb{R}} \psi\left(\frac{x - T_L(x_n)}{\text{MAD}(x_n)}\right) d\mathbb{P}_n(x) = 0.
$$

(1.32)

**Example 1.43.** If we put $c = 3$ and use $T_S = \text{MAD}$, 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.

### 1.7.2 M-measures of scale

M-measures of scale $T_S$ are defined in a similar manner as follows.

**Definition 1.44.** For a function $\chi$ and a sample $x_n$ with empirical distribution $\mathbb{P}_n$ any solution of

$$
\int_{\mathbb{R}} \chi\left(\frac{x}{T_S(x_n)}\right) d\mathbb{P}_n(x) = 0.
$$

(1.33)

is called an M-measure or M-estimator of scale.

**Exercise 1.45.** 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.

**Exercise 1.46.** Suppose that the sample $x_n$ has median zero and put $\chi(x) = \text{sgn}(|x| - 1)$. Show that $T_S$ is the MAD of the sample.

Corresponding to Theorem 1.37 we have
Theorem 1.47. 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

$$\int_{\mathbb{R}} \chi(x/s) \, d\mathbb{P}_n(x) = 0$$

has a unique solution $s =: T_S(\mathbb{P})$ with $0 < T_S(\mathbb{P}_n) < \infty$ for every sample $\mathbf{x}_n$ with $\mathbb{P}_n(\{0\}) < 1/2$.

Proof. As $\mathbb{P}_n(\{0\}) < 1/2 < 1$ the function $\Xi(s) = \int_{\mathbb{R}} \chi(x/s) \, d\mathbb{P}_n(x)$ is strictly monotone decreasing and continuous. Furthermore $\lim_{s \to \infty} \Xi(s) = -1$ and $\lim_{s \to 0} \Xi(s) = -\mathbb{P}_n(\{0\}) + (1 - \mathbb{P}_n(\{0\})) > 0$ as $\mathbb{P}_n(\{0\}) < 1/2$. Hence, there exists a unique $s =: T_S(\mathbb{P}_n)$ with $\Xi(s) = 0$. □

Example 1.48. The function

$$\chi(x) = \frac{x^4 - 1}{x^4 + 1}$$

satisfies the assumptions of the theorem and it will be our standard $\chi$-function.

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 $s_0 = \text{MAD}(\mathbf{x}_n)$.

(b) If $\Xi(s_0) = 0$ put $T_S(\mathbb{P}_n) = s_0$ and terminate.

(c) If $\Xi(s_0) < 0$ put $s_2 = s_0$, choose $k$ so that $\Xi(s_02^{-k}) > 0$ and put $s_1 = s_02^{-k}$.

(d) If $\Xi(s_0) > 0$ put $s_1 = s_0$, choose $k$ so that $\Psi(s_02^k) < 0$ and put $m_2 = s_02^k$.

(e) Put $s_3 = \sqrt{s_1s_2}$. 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 $T_S(\mathbb{P}_n) = \sqrt{s_1s_2}$ and terminate.
Exercise 1.49. Write an R-programme to calculate $T_S$ for the $\chi$-function of Example 1.48.

In order to make $T_S$ affine equivariant we require an measure of location $T_L$ and then solve

$$\int_{\mathbb{R}} \chi \left( \frac{x - T_L(x_n)}{T_S(x_n)} \right) d\mathbb{P}_n(x) = 0.$$  

Rather than do this in an ad hoc method we now consider joint M-measures of location and scatter.

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

$$\int_{\mathbb{R}} \psi \left( \frac{x - m}{s} \right) d\mathbb{P}_n(x) = 0 \quad (1.34)$$
$$\int_{\mathbb{R}} \chi \left( \frac{x - m}{s} \right) d\mathbb{P}_n(x) = 0 \quad (1.35)$$

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

(\psi_1) $\psi(-x) = -\psi(x)$ for all $x \in \mathbb{R}$.
(\psi_2) $\psi$ is strictly increasing.
(\psi_3) $\lim_{x \to \infty} \psi(x) = 1$.
(\psi_4) $\psi$ is continuously differentiable with derivative $\psi^{(1)}$.

(\chi_1) $\chi(-x) = \chi(x)$ for all $x \in \mathbb{R}$.
(\chi_2) $\chi : \mathbb{R}_+ \to [-1, 1]$ is strictly increasing.
(\chi_3) $\chi(0) = -1$.
(\chi_4) $\lim_{x \to \infty} \chi(x) = 1$.
(\chi_5) $\chi$ is continuously differentiable with derivative $\chi^{(1)}$.

(\psi\chi_1) $\chi^{(1)}/\psi^{(1)} : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing.

($x_n$) $\Delta(x_n) < 1/2$ where $\Delta(x_n)$ is defined by (1.19)

The existence of joint M-measures of location and scale is not trivial but the following theorem holds.
**Theorem 1.50.** Under the above assumptions $(\psi_i), i = 1, \ldots, 4$, $(\chi_i), i = 1, \ldots, 5$, $(\psi \chi_1)$ and $(x_n)$ the equations (1.34) and (1.35) have a unique solution $(m, s) =: (T_L(x_n), T_S(x_n))$.

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

**Example 1.51.** We consider the $\psi$– and $\chi$–functions

$$
\psi(x) = \psi(x, c) = \frac{(\exp(cx) - 1)}{(\exp(cx) + 1)}
$$

$$
\chi(x) = \frac{x^4 - 1}{x^4 + 1}.
$$

They clearly fulfill all the conditions of Theorem 1.50 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 1.52.** Verify the calculation of Example 1.51

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:

1. Put $s = s_1 = \text{MAD}(P_n)$ in (1.34) and solve for $m = m_1$.
2. Put $m = m_1$ in (1.35) and solve for $s = s_2$.
3. Put $s = s_2$ in (1.34) and solve for $m = m_2$.
4. Put $m = m_2$ in (1.35) and solve for $s = s_3$.
5. Iterate as indicated and terminate when

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

**Example 1.53.** We calculate the simultaneous location and scale functionals $T_L$ and $T_S$ for the copper data using the functions

$$
\psi(x) = \psi(x, c) = \frac{(\exp(cx) - 1)}{(\exp(cx) + 1)}
$$

$$
\chi(x) = \frac{x^4 - 1}{x^4 + 1}.
$$

with tuning parameter $c = 3$ We obtain $T_L = 2.026956$ (mean 2.015926, median 2.03) and $T_S = 0.0674$ (standard deviation 0.1140260, MAD 0.1).
Exercise 1.54. Verify the calculation of Example 1.53.

We look at the breakdown behaviour of the joint M-functional (1.34) and (1.35). For a sample \( x \) the we set \((m, s) = (T_L(\mathbb{P}_n), T_S(\mathbb{P}_n))\) so the equations read

\[
\sum_{i=1}^{n} \psi \left( \frac{x_i - m}{s} \right) = 0 \quad (1.36)
\]
\[
\sum_{i=1}^{n} \chi \left( \frac{x_i - m}{s} \right) = 0. \quad (1.37)
\]

The situation is more complicated. The condition on \( \Delta(\mathbb{P}) < 1/2 \) for the existence and uniqueness of a solution is only sufficient. This makes it difficult to 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 \(((m_l, s_l))_{l=1}^{\infty}\) of solutions of (1.36) and (1.37) 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

\[
\chi^{-1}(-\varepsilon_0/(1-\varepsilon_0)) = \psi^{-1}(\varepsilon_0/(1-\varepsilon_0)). \quad (1.38)
\]

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 \( [n\varepsilon_0]/n \leq \varepsilon^*(T_L, \mathbb{P}_n) \leq [(n+1)/2]/n \). In fact one can go further and show that

\[
[n\varepsilon_0]/n \leq \varepsilon^*(T_L, \mathbb{P}_n) \leq ([n\varepsilon_0 + 1]/n \quad (1.39)
\]

with \( \varepsilon_0 \) given by (1.38).
References


