Summarizing Data: One Variable

The Centre of the Data and the Effects of Extreme Values

In this section we will continue to consider some summary measures we can use when exploring a dataset. We will use the skeleton data provided by our anthropologist and work with the age estimates according to the method of Di Gangi *at al.* Since we are interested in how well the Di Gangi method works, we will consider the difference between the estimated age and the actual age at death for the 400 skeletons in our dataset.

Age estimation error = age according to Di Gangi method - actual age

First, here is a boxplot of these data.

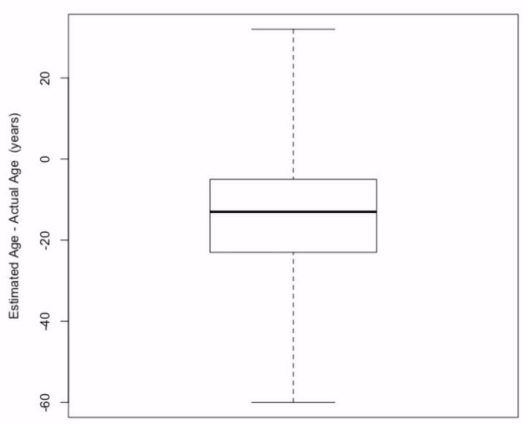

Figure 1: Boxplot of skeleton age estimation errors using the Di Gangi method

We see that the Di Gangi method tends to underestimate the age at death since more than 75% of the age estimation errors are negative. Note that the minimum value is -60, meaning that for one observation the estimate was 60 years younger than the actual age.

A **modified boxplot** clearly displays unusual observations. To draw a modified boxplot:

1. Draw a box from the 1st to the 3rd quartiles.
2. Indicate the median with a line in the box.
3. Calculate:
   interquartile range (IQR) = 3rd quartile - 1st quartile

lower inner fence = 1st quartile $- 1.5 \times$ IQR
upper inner fence = 3rd quartile $+ 1.5 \times$ IQR

4. The whiskers extend to the largest data value that is less than or equal to the upper inner fence and the smallest data value that is greater than or equal to the lower inner fence.

5. Data values outside the fences are indicated separately and may be worthy of attention.

EXAMPLE 1
Let's use the skeleton data to illustrate these concepts. Here is the five number summary for the dataset:

Minimum	1st Quartile	Median	3rd Quartile	Maximum
-60	-23	-13	-5	32

Table 1: Five number summary for skeleton data

$$\begin{aligned}\text{IQR} &= \text{3rd quartile - 1st quartile} \\ &= -5 - (-23) \\ &= 18\end{aligned}$$

$$\begin{aligned}\text{lower inner fence} &= \text{1st quartile} - 1.5 \times \text{IQR} \\ &= -23 - 1.5 \times 18 \\ &= -23 - 27 \\ &= -50\end{aligned}$$

$$\begin{aligned}\text{upper inner fence} &= \text{3rd quartile} + 1.5 \times \text{IQR} \\ &= -5 + 1.5 \times 18 \\ &= -5 + 27 \\ &= 22\end{aligned}$$

Figure 2 shows the modified boxplot. We can see that our minimum value of -60 is outside the inner fences and that there are also other unusual observations beyond the fences.

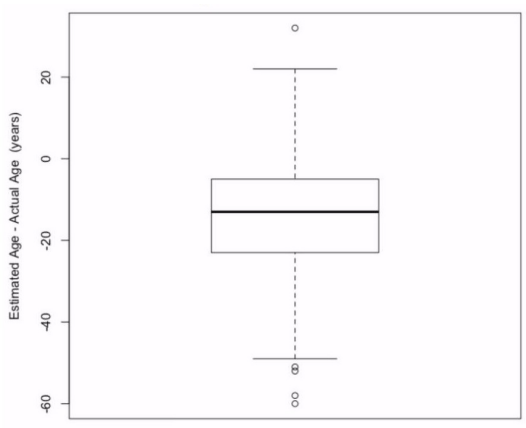

Figure 2: Modified boxplot of skeleton age estimation errors using the Di Gangi method

Summary of some measures of centre

Median: the middle data value

Mean: the average. Given n observations $x_1, x_2, ..., x_n$, the mean of the variable x is noted as $\bar{x}$ and is calculated as:

$$\text{Mean} = \bar{x} = \frac{\sum \text{data values}}{\text{number of data points}} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Trimmed mean: the $z\%$ trimmed mean is the mean of the remaining data values after the k largest and k smallest data values have been removed where $k = \frac{a}{100} \times n$. (If k is not an integer, the number of data values to remove from each end of the ordered data is the integer less than k.)

We say that the mean is not a **robust** statistic since it is not resistant to extreme observations. In contrast, the median and the trimmed mean are robust statistics.

For the skeletons example, the median $= -13$, the 10% trimmed mean $= -13.8$ and mean $= 14.2$. Notice that the mean is more negative than the median for this data. This happens because the mean is pulled down by the unusually large negative observations we observed in the modified boxplot above.

Example 2
As a further illustration of the robustness of the median and trimmed mean versus the mean, let's look at the New York Red Bulls salary data.

The salaries ordered from smallest to largest are shown in Table 2.

Rank	NY Red Bulls 2012 Salaries
1	33,750.00
2	33,750.00
3	33,750.00
4	33,750.00
5	44,000.00
6	44,000.00
7	44,000.00
8	44,000.04
9	45,566.67
10	65,000.00
11	95,000.00
12	103,500.00
13	112,495.50
14	138,188.00
15	141,666.67
16	181,500.00
17	185,000.00
18	190,000.00
19	194,375.00
20	195,000.00
21	205,000.00
22	292,500.00
23	301,999.00
24	4,600,000.00
25	5,600,000.00

Table 2: NY Red Bulls 2012 Salaries

As is typical in the MLS league, a few players have very large salaries, while most of the players have salaries more in line with the average American. You can see how unusual the top two salaries are compared to the rest of the team by looking at a modified box plot (Figure 3).

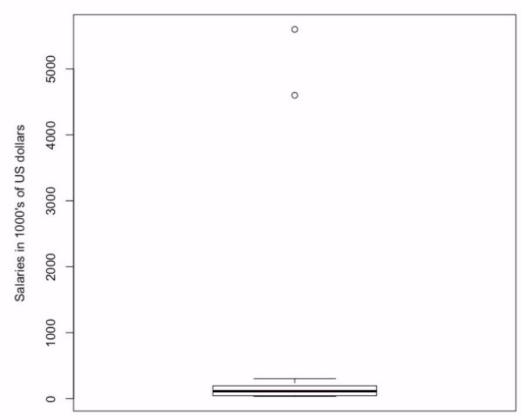

Figure 3: Modified boxplot of NY Red Bulls 2012 Salaries

Since there are 25 salaries, the median is the 13th value in the ordered salaries which is $\$112,495.50$.

The mean salary is:

$$\text{mean} = \frac{33,750 + 33,750 + ... + 4,600,800 + 5,600,000}{25} = 518,311.60$$

The mean is much larger than the median because it is inflated by the top two salaries. To remove the influence of the top players we can use a trimmed mean. To calculate the 8% trimmed mean salary, we must exclude the top 8% of salaries and the bottom 8% of salaries. 8% of 25 is 2 so we will exclude the top 2 salaries and the bottom 2 salaries:

$$8\% \text{ trimmed mean} = \frac{33,750 + 33,750 + ... + 292,500 + 301,999}{21} = 128,109.10$$

The trimmed mean is a value that is much more representative of the salaries of the team than the mean.