Summarizing Data: One Variable

The Spread of the Data

Let's consider the life expectancy example in order to introduce concepts about the spread of data. The boxplot below shows us various summary statistics: the minimum, the first quartile, the median, the third quartile and the maximum.

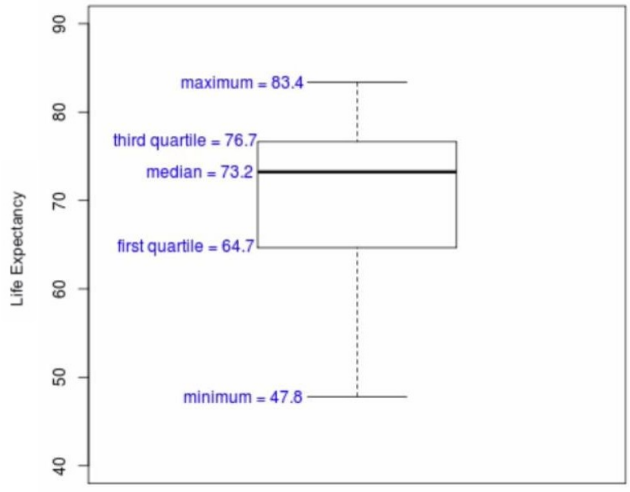

Figure 1: Boxplot of the life expectancy data for 197 countries and territories

The simplest way to think about the spread of the data is to look at its **range**.

Range = maximum value - minimum value = $83.4 - 47.8 = 35.6$

The range indicates that all of the data can fit into an interval of length 35.6 years.

To understand the spread of these data we can also look at the **interquartile range** (IQR):

IQR = 3rd quartile - 1st quartile = $76.7 - 64.7 = 12.0$

In this case, the interquartile range indicates that the middle half of the data can fit into an interval which is of length 12.0 years.

Other important measures of the spread of data are the **variance** and **standard deviation** (SD). Let $x_1, x_2, ..., x_n$ represent the data and $\bar{x}$ be the sample mean. Then,

$$\text{variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\text{SD} = \sqrt{\text{variance}}$$

Example 1
Let's calculate range, IQR, variance, and standard deviation for the following data:

$$68, 69, 74, 76, 79, 87, 88, 90, 93$$

range = maximum − minimum = 93 − 68 = 25
IQR = 3rd quartile − 1st quartile = 88 − 74 = 14

For the variance we must first calculate the mean:

$$\bar{x} = \frac{68 + 69 + ... + 93}{9} = 80.4$$

Using the variance formula:

$$\text{variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(68 - 080.4)^2 + (69 - 080.4)^2 + \cdots + (93 - 080.4)^2}{8} = 87.3$$

$$\text{SD} = \sqrt{87.3} = 9.3$$

Example 2
Let's consider the 2012 Red Bulls Salary data again. In a previous lecture we determined the mean and the median (measures of centre) for these data. We also considered trimming the data by removing the two highest and the two lowest salaries. Table 1 below shows various measures of centre and spread for the original data and the trimmed data. Since the range and standard deviations change tremendously if we remove a few outlying values, these two measures of spread are not robust towards outliers. In contrast, the IQR is \$150,000 for the full data and \$146,000 for the trimmed data, demonstrating that the IQR is a robust measure of spread.

	Original data	8% Trimmed data	Robust?
median	112,000	112,000	Yes
mean	518,000	128,000	No
range	5,566,000	268,000	No
IQR	150,000	146,000	Yes
SD	1,389,000	84,000	No

Table 1: Measures of centre and spread for the NY Red Bulls 2012 salaries data