Section 1.1. Boxplots

Note. In this section we repeat several ideas covered in Introduction to Probability and Statistics (MATH 1530). In particular, we consider the five number summary and boxplots; see my online notes for Introduction to Probability and Statistics on Chapter 2. Describing Distributions with Numbers. The material of this section is only applicable to univariate data.

Definition. For a collection of real numbers ("data"), $x_1, x_2, \ldots, x_n$, the *order statistics* involve an arrangement of the data as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. The *median* of this data is

$$M = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd,} \\ \{x_{(n/2)} + x_{(n/2+1)}\}/2 & \text{if } n \text{ is even.} \end{cases}$$

The *lower quartile* F_L of this data is the median of the data $x_{(1)}, x_{(2)}, \ldots, x_{(\lfloor (n+1)/2 \rfloor)}$, and the *upper quartile* F_U if the median of the date $x_{(\lceil (n+1)/2 \rceil)}, x_{(\lceil (n+1)/2 \rceil + 1)}, \ldots, x_{(n)}$. (Here, $\lfloor x \rfloor$ is the "rounding down" function and $\lceil x \rceil$ is the "rounding up" function.) The *F-spread* (or *interquartile range*) d_F is $d_F = F_U - F_L$. The *outside bars* are $F_D + 1.5d_F$ and $F_L - 1.5d_F$. A data value which is either greater than $F_U + 1.5d_F$ or less than $F_L - 1.5d_F$ is an *outlier*. The maximum and minimum data values are the *extremes*. The *five number summary* of the data is: minimum, F_L, M, F_U, maximum. The *mean* is $\overline{x} = n^{-1} \sum_{i=1}^n x_i$.

Note. Notice that when n is odd, the median $M = x_{((n+1)/2}$ is included in both $x_{(1)}, x_{(2)}, \ldots, x_{(\lfloor (n+1)/2 \rfloor)}$ and $x_{(\lceil (n+1)/2 \rceil)}, x_{(\lceil (n+1)/2 \rceil + 1)}, \ldots, x_{(n)}$. When n is even, these two collections of data are disjoint (but together they include all of the data).

In the event that n is odd, the approach here differs from that of Introduction to Probability and Statistics where the median is excluded from both collections of data (see Example S.2.3 in my notes on Chapter 2. Describing Distributions with Numbers).

Example 1.1.A. We reproduce Table 1.1 from the text book here.

Table 1.1. The 15 largest world cities in 2006.

City	**Country**	**Pop. (10,000)**	**Order Statistics**
Tokyo	Japan	3,420	$x_{(15)}$
Mexico City	Mexico	2,280	$x_{(14)}$
Seoul	South Korea	2,230	$x_{(13)}$
New York	USA	2,190	$x_{(12)}$
Sao Paulo	Brazil	2,020	$x_{(11)}$
Bombay	India	1,985	$x_{(10)}$
Delhi	India	1,970	$x_{(9)}$
Shanghai	China	1,815	$x_{(8)}$
Los Angeles	USA	1,800	$x_{(7)}$
Osaka	Japan	1,680	$x_{(6)}$
Jakarta	Indonesia	1,655	$x_{(5)}$
Calcutta	India	1,565	$x_{(4)}$
Cairo	Egypt	1,560	$x_{(3)}$
Manila	Philippines	1,495	$x_{(2)}$
Karachi	Pakistan	1,430	$x_{(1)}$

Since $n = 15$ is odd, then the median is $M = x_{((n+1)/2)} = x_{((15+1)/2)} = x_{(8)} = 1{,}815$.

The lower quartile F_L is the median of the data $x_{(1)}, x_{(2)}, \ldots, x_{(8)}$, since $x_{(\lfloor (n+1)/2 \rfloor)} = x_{(\lfloor (15+1)/2 \rfloor)} = x_{(8)}$. Since 8 is even, the lower quartile is $\{x_{(8/2)} + x_{(8/2+1)}\}/2 = \{x_{(4)} + x_{(5)}\}/2 = (1{,}565 + 1{,}655)/2 = 1{,}610$. The upper quartile F_U is the median of the data $x_{(8)}, x_{(9)}, \ldots, x_{(15)}$, since $x_{(\lceil (n+1)/2 \rceil)} = x_{(\lceil (15+1)/2 \rceil)} = x_{(8)}$. Since 8 is even, the upper quartile is $\{x_{(11)} + x_{(12)}\}/2 = (2{,}020 + 2{,}190)/2 = 2{,}105$. So the five number summary of the data is: minimum $= 1{,}430$, $F_L = 1{,}610$, $M = 1{,}815$, $F_U = 2{,}105$, maximum $= 3{,}420$.

Note. For a given collection of data $x_1, x_2, \ldots, x_n$, Härdle and Simar describe the construction of a boxplot in the following steps (see page 7).

1. Draw a box with borders at F_L and F_U (notice that half of the data lies in this box).

2. Draw the median as a solid line and the mean as a dotted line.

3. Draw "whiskers" from each end of the box to the most remote (i.e., least and greatest) data value that is not an outlier.

4. Show outliers that lie outside the interval $[F_L - 3d_F, F_U + 3f_F]$ with a $\bullet$, and show the remaining outliers that lie outside the interval $[F_L - 1.5d_F, F_U + 1.5f_F]$ with a $\star$.

Example 1.1.A (continued). The boxplot for the large cities data of Example 1.1.A is given in Figure 1.2.

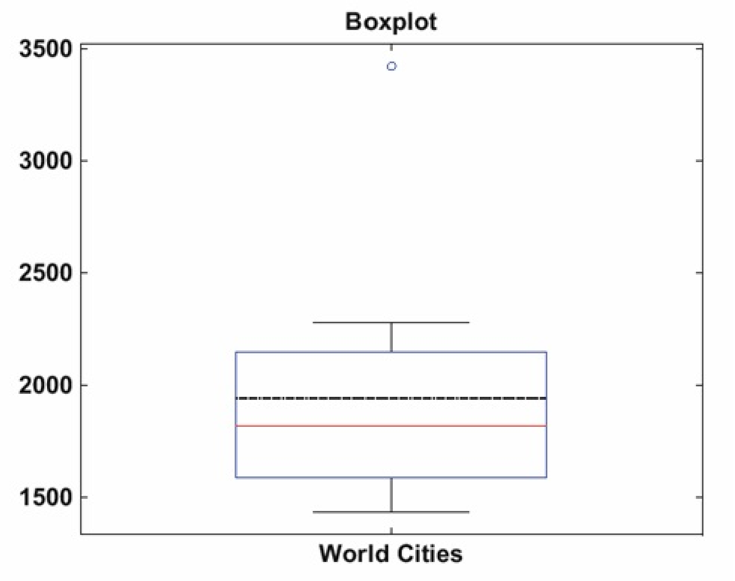

Figure 1.2. Boxplot for world cities.

Revised: 4/25/2021