

Box Plots and Five-Number Summaries

A box plot is a graphical summary of data that requires few numerical calculations. Five numbers are used to summarize the data: 1) the smallest value; 2) the first quartile (Q1); 3) the median (or second quartile, Q2); 4) the third quartile (Q3); 5) the largest value. This set of numbers is called the five-number summary.

Definitions

Q1: first quartile = 25th percentile

Q2: median or second quartile = 50th percentile

Q3: third quartile = 75th percentile

Interquartile range (IQR): = $Q3 - Q1$

Lower inner fence: $Q1 - 1.5(IQR)$

Upper inner fence: $Q3 + 1.5(IQR)$

Lower outer fence: $Q1 - 3(IQR)$

Upper outer fence: $Q3 + 3(IQR)$

Percentile: a measure of relative standing of a value in a data set. The p th percentile is a value such that approximately p percent of the data have values less than that value. For example, if your SAT results ranks your math test score at the 75th percentile, then about 75% of the test-takers got *lower* scores than you did.

Quartiles divide the data into *quarters* (chunks of 25%). So in general, 25% of the data are between the minimum and Q1; 25% are between Q1 and the median; 25% are between the median and Q3; and 25% are between Q3 and the maximum.

Calculating the five numbers

1. Rank order the data in ascending order.
2. The smallest value is the minimum.
3. To find the quartiles, first calculate the *index* for each percentile:

$$\text{index: } i = \left(\frac{p}{100} \right) n \quad \text{where } p \text{ is the percentile and } n \text{ is the number of data values}$$

If the index i is not an integer, round up to the nearest integer, e.g., round 3.25 up to 4. i denotes the *position* of the value corresponding to that percentile, e.g., the 4th number.

If the index i is an integer, average the value in *position* i with the value in the *position* $(i + 1)$ to get the value for that percentile.

$$\begin{array}{ll} \text{for Q1, } p = 25 \text{ and } (p/100) = \frac{1}{4} & \text{for Q2, } p = 50 \text{ and } (p/100) = \frac{1}{2} \\ \text{for Q3, } p = 75 \text{ and } (p/100) = \frac{3}{4} & \end{array}$$

4. The largest value is the maximum.

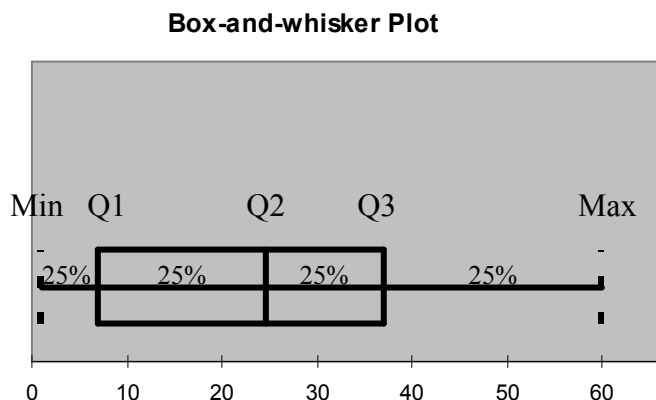
Example: Determine the five-number summary for the data set {1, 2, 4, 7, 11, 18, 24, 25, 32, 36, 37, 38, 41, 60}, n = 14

1. The minimum is 1.
2. The index for Q1 is: $i = \frac{1}{4} * 14 = 3.5 \rightarrow$ round up to 4. The number in the fourth position is 7.
3. The index for Q2 (median) is: $i = \frac{1}{2} * 14 = 7$. Since this is an integer, average the number in the 7th position with the number in the 8th position. $(24 + 25)/2 = 24.5$.
4. The index for Q3 is: $i = \frac{3}{4} * 14 = 10.5 \rightarrow$ round up to 11. The number in the 11th position is 37.
5. The maximum is 60.

5-Number Summary	
Minimum	1
Q1	7
Q2 (median)	24.5
Q3	37
Maximum	60

Drawing the basic box plot

The “box” part of the box plot is bounded by the values for Q1 and Q3 with the Q2 value shown as a vertical line inside the box. Lines or “whiskers” extend from the ends of the box to the minimum and maximum (or alternatively to a fence – more on fences later). The 5-number summary above results in the following box plot (using PHStat in Excel):



PHStat also produces a 5-number summary:

Box-and-whisker Plot	
Five-number Summary	
Minimum	1
First Quartile	7
Median	24.5
Third Quartile	37
Maximum	60

Interquartile range (IQR)

The IQR, or length of the box, is a measure of variability in the sample and can be used to compare two samples.

Whiskers

If one whisker is very much longer than the other, it indicates the data is probably skewed in the direction of the longer whisker – like having a long tail stretching out to one side of the distribution.

Fences

Fences are used to examine the data for possible **outliers**. Some software programs stop the whiskers at the inner fences and plot data beyond the fences using discrete symbols. PHStat in Excel draws the whiskers all the way to the extreme values.

Fewer than 5% of the data should fall beyond the inner fences – these are potential outliers. Data beyond the outer fences are probably outliers, and bear scrutiny if further analysis is to be done on the sample.

Example: Construct a boxplot with fences on the following data set. $n = 20$

.553	.606	.654	.690
.570	.609	.662	.693
.576	.611	.668	.749
.601	.615	.670	.844
.606	.628	.672	.933

1. The minimum is .553
2. The index for Q1 is: $i = \frac{1}{4} * 20 = 5$. Since this is an integer, average the number in the 5th position with the number in the 6th position. $(.606 + .606)/2 = .606$
3. The index for Q2 (median) is: $i = \frac{1}{2} * 20 = 10$. Since this is an integer, average the number in the 10th position with the number in the 11th position. $(.628 + .654)/2 = .641$
4. The index for Q3 is: $i = \frac{3}{4} * 20 = 15$. Since this is an integer, average the number in the 15th position with the number in the 16th position. $(.672 + .690)/2 = .681$
5. The maximum is .933

5-Number Summary	
Minimum	.553
Q1	.606
Q2 (median)	.641
Q3	.681
Maximum	.933

$$\begin{aligned} \text{IQR} &= Q3 - Q1 = .681 - .606 = .075 \\ 1.5(\text{IQR}) &= .1125 \qquad 3(\text{IQR}) = .225 \end{aligned}$$

$$\text{Lower inner fence} = Q1 - 1.5(\text{IQR}) = .606 - .1125 = .4935$$

$$\text{Upper inner fence} = Q3 + 1.5(\text{IQR}) = .681 + .1125 = .7935$$

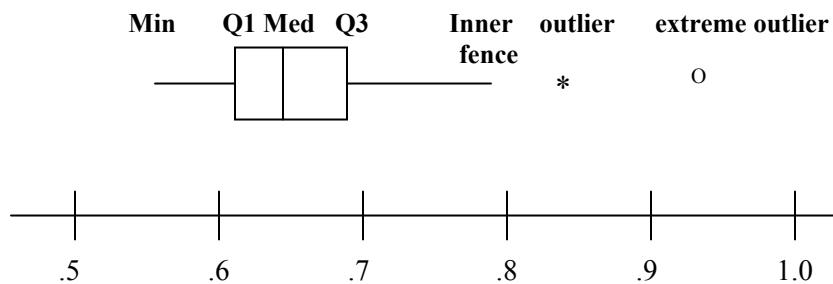
There are no values below the lower inner fence; there are two values above the upper inner fence.

Lower outer fence: does not need to be calculated for this sample, since all values fall within the lower inner fence.

$$\text{Upper outer fence} = Q3 + 3(\text{IQR}) = .681 + .225 = .906$$

There is one value above the upper outer fence – this is an extreme outlier.

Construct the boxplot:



CAUTION!!!

- 1) There is no one standard way to draw the boxplot. You could draw the whisker to the fence or to the max or min value within the fence. You could draw the whiskers to the extreme values without showing fences. Follow your teacher's instructions.
- 2) There is no one standard way to determine quartile values or even percentile values in general. If you use Excel functions, a graphing calculator, or other software, you may get different values than if you use the method in this worksheet. If your instructor gives you a certain method to do these calculations, use that method!