Purplemath

http://www.purplemath.com/modules/boxwhisk.htm

Box-and-Whisker Plots: Quartiles, Boxes, and Whiskers

Sections: Quartiles, boxes, and whiskers, Five-number summary, Interquartile ranges and outliers

Statistics assumes that your data points (the numbers in your list) are clustered around some central value. The "box" in the box-and-whisker plot contains, and thereby highlights, the middle half of these data points.

To create a box-and-whisker plot, you start by ordering your data (putting the values in numerical order), if they aren't ordered already. Then you find the median of your data. The median divides the data into two halves. To divide the data into quarters, you then find the medians of these two halves. Note: If you have an even number of values, so the first median was the average of the two middle values, then you include the middle values in your sub-median computations. If you have an odd number of values, so the first median was an actual data point, then you do not include that value in your sub-median computations. That is, to find the sub-medians, you're only looking at the values that haven't yet been used.

You have three points: the first middle point (the median), and the middle points of the two halves (what I call the "sub-medians"). These three points divide the entire data set into quarters, called "quartiles". The top point of each quartile has a name, being a "Q" followed by the number of the quarter. So the top point of the first quarter of the data points is "Q_1", and so forth. Note that Q_1 is also the middle number for the first half of the list, Q_2 is also the middle number for the whole list, Q_3 is the middle number for the second half of the list, and Q_4 is the largest value in the list.

Once you have these three points, Q_1, Q_2, and Q_3, you have all you need in order to draw a simple box-and-whisker plot. Here's an example of how it works.

- **Draw a box-and-whisker plot for the following data set:**

  4.3, 5.1, 3.9, 4.5, 4.4, 4.9, 5.0, 4.7, 4.1, 4.6, 4.4, 4.3, 4.8, 4.4, 4.2, 4.5, 4.4

  My first step is to order the set. This gives me:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

  The first number I need is the median of the entire set. Since there are seventeen values in this list, I need the ninth value:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

  The median is $Q_2 = 4.4$.

  The next two numbers I need are the medians of the two halves. Since I used the "4.4" in the middle of the list, I can't re-use it, so my two remaining data sets are:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4 and 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

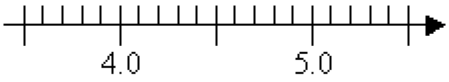
The first half has eight values, so the median is the average of the middle two:

$Q_1 = (4.3 + 4.3)/2 = 4.3$

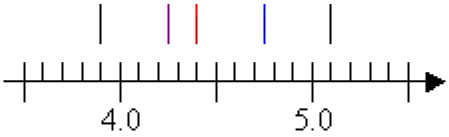
The median of the second half is:

$Q_3 = (4.7 + 4.8)/2 = 4.75$

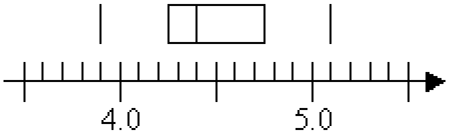
Since my list values have one decimal place and range from 3.9 to 5.1, I won't use a scale of, say, zero to ten, marked off by ones. Instead, I'll draw a number line from 3.5 to 5.5, and mark off by tenths.



Now I'll mark off the minimum and maximum values, and Q_1, Q_2, and Q_3:



The "box" part of the plot goes from Q_1 to Q_3:



By the way, box-and-whisker plots don't have to be drawn horizontally as I did above; they can be vertical, too.

More terminology: The top end of your box may also be called the "upper hinge"; the lower end may also be called the "lower hinge". The lower hinge is also called "the 25th percentile"; the median is "the 50th percentile"; the upper hinge is "the 75th percentile". This means that 25%, 50% and 75% of the data, respectively, is at or below that point. The distance between the hinges may be referred to as the "H-spread" or, as you will see on the following page, the "Interquartile Range", abbreviated "IQR". ("Hinge" actually has a different technical definition, but the term is sometimes used informally.)

Also, some books and software will include the overall median (Q_2) when computing Q_1 and Q_3 for data sets with an odd number of elements. The Texas Instruments calculators do *not* include Q_2 in this case, so you may encounter a book answer that doesn't match the calculator answer. And different software packages use all different sorts of formulas. Be careful to use the formula from *your* book when doing *your* homework!

Additionally, the box-and-whisker plot may include a cross or an "X" marking the mean value of the data, in addition to the line inside the box that marks the median. The difference between the "X" and the median line can then be used as a measure of "skew".

Please don't ask me to explain "skew".

- **Draw the box-and-whisker plot for the following data set:**
77, 79, 80, 86, 87, 87, 94, 99

My first step is to find the median. Since there are eight data points, the median will be the average of the two middle values: $(86 + 87) \div 2 = 86.5 = Q_2$

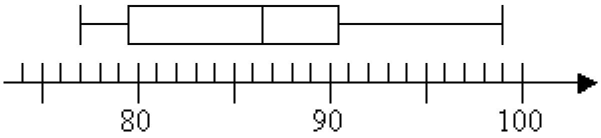
This splits the list into two halves: 77, 79, 80, 86 and 87, 87, 94, 99. Since the halves of the data set each contain an even number of values, the sub-medians will be the average of the middle two values.

$Q_1 = (79 + 80) \div 2 = 79.5$
$Q_3 = (87 + 94) \div 2 = 90.5$

The minimum value is 77 and the maximum value is 99, so I have:

min: 77, Q_1: 79.5, Q_2: 86.5, Q_3: 90.5, max: 99

Then my plot looks like this:



As you can see, you only need the five values listed above (min, Q_1, Q_2, Q_3, and max) in order to draw your box-and-whisker plot. This set of five values has been given the name "the five-number summary".

- **Give the five-number summary of the following data set:**
  79, 53, 82, 91, 87, 98, 80, 93

The five-number summary consists of the numbers I need for the box-and-whisker plot: the minimum value, Q_1 (the bottom of the box), Q_2 (the median of the set), Q_3 (the top of the box), and the maximum value (which is also Q_4). So I need to order the set, find the median and the sub-medians, and then list the required values in order.

ordering the list: 53, 79, 80, 82, 87, 91, 93, 98, so the minimum is 53 and the maximum is 98

finding the median: $(82 + 87) \div 2 = 84.5 = Q_2$

lower half of the list: 53, 79, 80, 82, so $Q_1 = (79 + 80) \div 2 = 79.5$

upper half of the list: 87, 91, 93, 98, so $Q_3 = (91 + 93) \div 2 = 92$

five-number summary: 53, 79.5, 84.5, 92, 98

Part of the point of a box-and-whisker plot is to show how spread out your values are. But what if one or another of your values is way out of line? For this, we need to consider "outliers"....

The "interquartile range", abbreviated "IQR", is just the width of the box in the box-and-whisker plot. That is, $\mathrm{IQR} = Q_3 - Q_1$. The IQR can be used as a measure of how spread-out the values are. Statistics assumes that your values are clustered around some central value. The IQR tells how spread out the "middle" values are; it can also be used to tell when some of the other values are "too far" from the central value. These "too far away" points are called "outliers", because they "lie outside" the range in which we expect them.

The IQR is the length of the box in your box-and-whisker plot. An outlier is any value that lies more than one and a half times the length of the box from either end of the box. That is, if a data point is below $Q_1 - 1.5\times\mathrm{IQR}$ or above $Q_3 + 1.5\times\mathrm{IQR}$, it is viewed as being too far from the central values to be reasonable. Maybe you bumped the weigh-scale when you were making that one measurement, or maybe your lab partner is an idiot and you should never have let him touch any of the equipment. Who knows? But whatever their cause, the outliers are those points that don't seem to "fit".

(Why one and a half times the width of the box? Why does that *particular* value demark the difference between "acceptable" and "unacceptable" values? Because, when John Tukey was inventing the box-and-whisker plot in 1977 to display these values, he picked $1.5\times\text{IQR}$ as the demarkation line for outliers. This has worked well, so we've continued using that value ever since.)

- **Find the outliers, if any, for the following data set:**

  10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7,
  14.7, 14.7, 14.9, 15.1, 15.9, 16.4

  To find out if there are any outliers, I first have to find the IQR. There are fifteen data points, so the median will be at position $(15 + 1) \div 2 = 8$. Then $Q_2 = 14.6$. There are seven data points on either side of the median, so Q_1 is the fourth value in the list and Q_3 is the twelfth: $Q_1 = 14.4$ and $Q_3 = 14.9$. Then $\text{IQR} = 14.9 - 14.4 = 0.5$.

  Outliers will be any points below $Q_1 - 1.5\times\text{IQR} = 14.4 - 0.75 = 13.65$ or above $Q_3 + 1.5\times\text{IQR} = 14.9 + 0.75 = 15.65$.

  Then the outliers are at 10.2, 15.9, and 16.4.

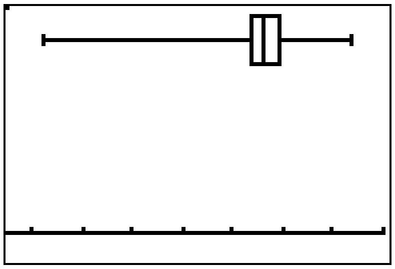
The values for $Q_1 - 1.5\times\text{IQR}$ and $Q_3 + 1.5\times\text{IQR}$ are the "fences" that mark off the "reasonable" values from the outlier values. Outliers lie outside the fences.

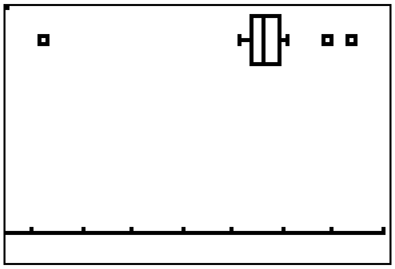
If your assignment is having you consider outliers and "extreme values", then the values for $Q_1 - 1.5\times\text{IQR}$ and $Q_3 + 1.5\times\text{IQR}$ are the "inner" fences and the values for $Q_1 - 3\times\text{IQR}$ and $Q_3 + 3\times\text{IQR}$ are the "outer" fences. The outliers (marked with asterisks or open dots) are between the inner and outer fences, and the extreme values (marked with whichever symbol you didn't use for the outliers) are outside the outer fences.

By the way, your book may refer to the value of "$1.5\times\text{IQR}$" as being a "step". Then the outliers will be the numbers that are between one and two steps from the hinges, and extreme value will be the numbers that are more than two steps from the hinges.

Looking again at the previous example, the outer fences would be at $14.4 - 3\times0.5 = 12.9$ and $14.9 + 3\times0.5 = 16.4$. Since 16.4 is right on the upper outer fence, this would be considered to be only an outlier, not an extreme value. But 10.2 is fully below the lower outer fence, so 10.2 would be an extreme value.

Your graphing calculator may or may not indicate whether a box-and-whisker plot includes outliers. For instance, the above problem includes the points 10.2, 15.9, and 16.4 as outliers. One setting on my graphing calculator gives the simple box-and-whisker plot which uses only the five-number summary, so the furthest outliers are shown as being the endpoints of the whiskers:





A different calculator setting gives the box-and-whisker plot with the outliers specially marked (in this case, with a simulation of an open dot), and the whiskers going only as far as the highest and lowest values that aren't outliers:

Note that my calculator makes no distinction between outliers and extreme values.

If you're using your graphing calculator to help with these plots, make sure you know which setting you're supposed to be using and what the results mean, or the calculator may give you a perfectly correct but "wrong" answer.

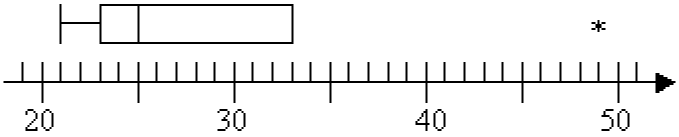
- **Find the outliers and extreme values, if any, for the following data set, and draw the box-and-whisker plot. Mark any outliers with an asterisk and any extreme values with an open dot.**

  21, 23, 24, 25, 29, 33, 49

  To find the outliers and extreme values, I first have to find the IQR. Since there are seven values in the list, the median is the fourth value, so $Q_2 = 25$. The first half of the list is 21, 23, 24, so $Q_1 = 23$; the second half is 29, 33, 49, so $Q_3 = 33$. Then IQR = 33 – 23 = 10.

  The outliers will be any values below 23 – 1.5×10 = 23 – 15 = 8 or above 33 + 1.5×10 = 33 + 15 = 48. The extreme values will be those below 23 – 3×10 = 23 – 30 = –7 or above 33 + 3×10 = 33 + 30 = 63.

  So **I have an outlier at 49 but no extreme values**, I won't have a top whisker because Q_3 is also the highest non-outlier, and my plot looks like this:



It should be noted that the methods, terms, and rules outlined above are what I have taught and what I have most commonly seen taught. However, your course may have different specific rules, or your calculator may do computations slightly differently. You may need to be somewhat flexible in finding the answers specific to your curriculum.