Describing Data 2:

Percentiles, Quartiles & Boxplots

The Describing Data 1 worksheet describes common measures of centre and the most common measure of variation: the standard deviation. Another way to measure variation is by using **quartiles**. This is a more accurate description of variability than the standard deviation if a data set has strong **outliers** (values that lie so far away from the rest of the data as to be unusual) or is strongly skewed.

Recall that to calculate a median value, you first want to find its locator, the position within the list (in increasing order!) where it can be found. Use this formula:

$$L = \frac{n + 1}{2}$$

Once you know the position, find the data point. If the calculation leads to a fraction answer, like 11.5, take the mean of the two numbers surrounding it (so for the #11.5 observation, average the #11 and the #12). It's a common mistake to stop once you've filled out a formula and assume you have the answer, but with the median (and the quartiles momentarily) there's always an additional step. For the data set {112, 116, 118, 122, 123}, L = 3, but it's not reasonable to say that 3 is a typical value for these observations. The 3rd value, 118, is the median.

We can expand on this idea. Instead of cutting the data set into halves, and using the halfway mark as the median, we can chop a data set into whatever fraction we like. Two of the most common ways to do this is to cut into 100 pieces to form **percentiles**, and to cut into four pieces to form **quartiles**.

The method for calculating these values will vary from source to source. The old MATH 1111 textbook had one way; the business stats course Downtown had a second, and neither one was the way your author was taught in university. The new MATH 1111 textbook has yet a fourth way, but it's no matter — all these methods do the same job.

As with the median, when finding a dividing point for a percentile, quartile or what-have-you, the first step is the find the locator. With any of these "tiles" the formula for the locator is basically the same: Take the fraction and multiply by the size of the data set:

Percentiles

$$\text{For } P_k,\ L = \frac{k}{100} \cdot n$$

Quartiles

$$\text{For } Q_k,\ L = \frac{k}{4} \cdot n$$

You might have noticed that the median formula doesn't follow the same pattern—the median locator will be slightly above ½n. Triola's rules for percentiles and quartiles also add a little something onto the locator before you use it:

If L is a whole number, average the #L and #(L+1) observations. If not, round L up and use the resulting observation.

Authored by Darren Rigby

Example 1: For the data set below (n = 50), find (a) P_{30} and (b) Q_3.

{8.13, 15.62, 16.10, 17.12, 17.90, 18.62, 18.70, 20.41, 21.04, 21.08, 21.44, 21.47, 21.74, 22.33, 22.40, 22.88, 23.07, 24.04, 24.29, 25.01, 25.15, 26.05, 26.45, 26.47, 26.56, 26.71, 26.84, 27.58, 28.20, 28.47, 28.55, 30.58, 31.49, 32.18, 33.06, 33.57, 33.68, 33.86, 34.06, 34.47, 35.84, 35.86, 35.99, 36.07, 36.84, 37.18, 38.40, 43.54, 47.51, 52.51}

Solution: a) The locator for P_{30} (the thirtieth percentile) is $(30/100) \cdot 50 = \#15$. This is a whole number, so P_{30} is the mean of #15 and #16: $(22.40 + 22.88) \div 2 = 22.64$.

b) The locator for Q_3 (the third quartile) is $(3/4) \cdot 50 = \#37.5$. This is not a whole number so we round up to #38: 33.86.

It should be obvious that $Q_1 = P_{25}$ and $Q_3 = P_{75}$. If a question ever asks for P_{50} or Q_2, use the median calculation.

The **interquartile range** (IQR) is the difference between the third quartile and first quartile: $Q_3 - Q_1$. This range will give the spread of the middle 50% of the values of the data set. It's another measure of variation, and one that's not affected as much by skewness or outliers as the standard deviation is.

Example 2: For the data set in Example 1, find the IQR.

Solution: The locator for Q_1 is $(1/4) \cdot 50 = \#12.5$, which we round up to #13. The 13th observation is 21.74, so, using our Q_3 found in Example 1, the IQR is $33.86 - 21.74 = 12.12$.

We've mentioned "outliers" now a couple of times with a vague definition about being "too far away" from the other values. We use IQR to determine what's too far. If a data point from the set is 1.5 × IQR or more above Q_3 or 1.5 × IQR or more below Q_1 then that data point is an outlier.

Example 3: Determine whether the data set in Example 1 contains any outliers.

Solution: The easiest way to do that is to find the critical values — the numbers that separate outliers from non-outliers* on each side. To be an outlier on the high side, a data point must be equal to or greater than:

$Q_3 + 1.5 \times IQR = 33.86 + 1.5 \cdot 12.12 = 52.04$

To be an outlier on the low side, a data point must be equal to or less than:

$Q_1 - 1.5 \times IQR = 21.74 - 1.5 \cdot 12.12 = 3.56$

Therefore the data set has an outlier: 52.51.

A good summary of a data set, especially one with a skewed or otherwise asymmetrical distribution, is called a **five-number summary**. It includes the minimum value of the set, the first quartile, the median, the third quartile and the maximum value. The five-number summary gives a good idea of the spread of data as well as the presence of any outliers in the data set.

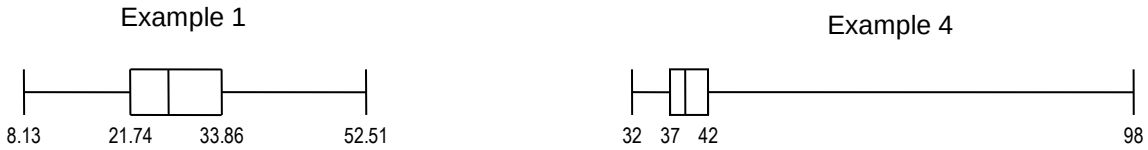
[*] Not surprisingly, called an **inlier**, but this is an obscure term.

Example 4: Give the five-number summary of: {32, 33, 37, 37, 39, 40, 42, 45, 98}.

Solution: The minimum value of the data set is 32; the maximum is 98. There are 9 observations in the data set, so n = 9. The data is already in numerical order.
The median value is at position (9 + 1) ÷ 2 = #5. The data point in position #5 is 39. This is the median.

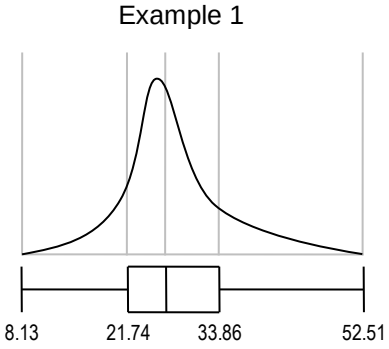
The locators for Q_1 and Q_3 are (1/4) · 9 = #2.25 and (3/4) · 9 = #6.75. Neither of these is a whole number, so we round them both up to #3 and #7. The 3rd observation is 37, and 7th is 42. Therefore the five-number summary is {32, 37, 39, 42, 98}.

We can use a five-number summary to get a rough visualization of the data. If we imagine a horizontal number line, we can make vertical marks at the five numbers in the summary and connect them with horizontal lines to make a **boxplot** (sometimes a box-and-whisker plot) of the data:



(The medians have not been labelled only for lack of space.) The two distributions are pretty similar except for the extreme outlier in Example 4's data. Both medians are slightly to the left compared to Q_1 and Q_3, which suggests some skew to the right even without the outliers, but we would never have seen this trend without the boxplot.

It's even possible to draw a curve representative of the histogram for the data using a boxplot. Imagine creating four spaces divided by walls where the vertical lines in the boxplot are. Then imagine pouring sand into those spaces so that (1) the same amount of sand goes into each space, (2) the sand has to be level with the floor at the extreme left and right of the diagram, and (3) the level of sand should be one smooth curve everywhere, meaning where there's a wall between two spaces, the height of the sand must be the same on each side of the wall. All this implies that narrow spaces have sand piled high, and wide spaces have lower piles of sand. For the data in Example 1, that would look something like this:



EXERCISES

A. For the following sets of data, calculate the (a) first quartile, (b) third quartile, (c) interquartile range, and (d) the five-number summary. Round to one extra decimal place compared to the source data.

1) {8, 24, 9, 6, 10, 18, 7, 14, 16, 21, 13, 24}

2) {3, 6, 5, 4, 6, 5, 9, 10, 11, 7, 9}

3) {41, 39, 38, 42, 43, 39, 40, 43, 26, 42, 42, 41, 41, 42, 27, 55, 60}

B. Identify the data set (1, 2, or 3 according to the numbered exercises above) with the greatest variability based on (a) standard deviation, (b) range and (c) IQR. (If you've done "Describing Data 1" these are the same data sets as Exercise A there, so you may already know the standard deviation.)

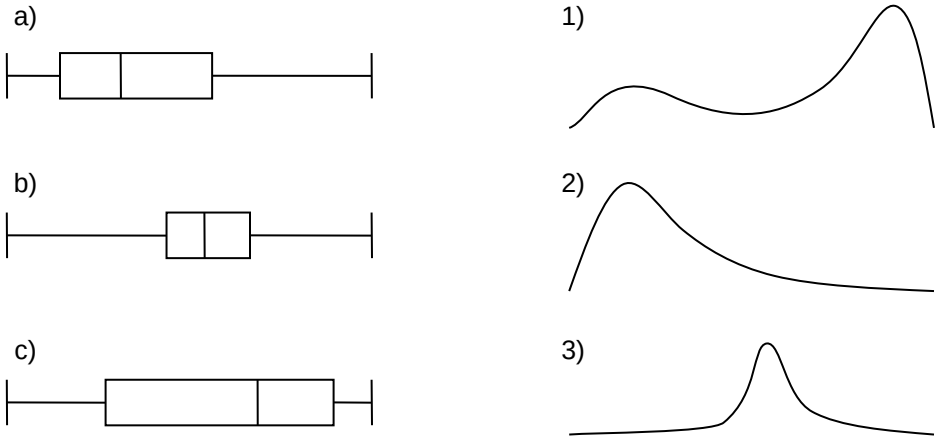
C. Explain why the answers to B(a) and B(b) are different from B(c).

D. a) Similar to a quartile or percentile, a quintile divides a data set into five equal sections. Find the third quintile of the data set in Example 1.

b) A tercile (or tertile) divides a data set into three equal sections. Find the second tercile of the data set in Example 1.

c) What percent of the observations should be below (1) the third quintile? (2) the second tercile?

E. Match the three boxplots to the curves.



SOLUTIONS

A. 1) (a) 8.5 (position #3.5) (b) 19.5 (position #9.5) (c) 11 (d) 6, 8.5, 13.5, 19.5, 24
2) (a) 5 (position #3) (b) 9 (position #9) (c) 4 (d) 3, 5, 6, 9, 11
3) (a) 39 (position #4.5) (b) 42.5 (position #13.5) (c) 3.5 (d) 26, 39, 41, 42.5, 60

B. a) Data set #3 has the greatest standard deviation.
b) Data set #3 has the greatest range. c) Data set #1 has the greatest IQR.

C. Data set #3 has outliers at both ends of the data. Because of this, data set #3 has a high standard deviation and range. However, the IQR is less sensitive to outliers. For this reason, the IQR of data set #3 is much lower than that of data set #1.

D. a) L = #30, so the third quintile is the mean of #30 and #31: 28.51
b) L = #33.3, so the second tercile is #34: 32.18
c) 1) 60% 2) 66⅔%

E. a) 2 b) 3 c) 1