1 EXPLORING DATA

1.08 Example

Say I live in a city with 8 high schools. I want to know what, per high school, the average grade for chemistry is. The lowest possible grade is a 0 and the highest possible grade is a 10. *This* is the data matrix. You can see that the cases studied here are not individual students, but schools. The variable of interest is the average grade for chemistry. Now, imagine you want to know a couple of things. First, you want to know what the distribution of the variable "average grade for chemistry" looks like. Second, you want to know what the center of the distribution is. Third you want to know more about the variability of the distribution. Fourth, you want to make a box plot to visually represent center, variability and outliers. And fifth, you want to know what the z-score of school three tells you.

Let's start with the first question. You want to know what the distribution looks like. We're dealing with a quantitative variable here and a small sample size, so the best way to display the distribution is with a dot plot. The possible grades range from 0 to 10 so you mark these values on the horizontal line of your dot plot. Next, you add the dots for all the observations. The first school got a 7.4 so you place the dot here. You do that for all 8 schools. This is the result. You can see that there is one outlier. School 3 has an average grade of only a 4.1.

The second thing you want to know is what the center of the distribution is. Well, you know that we have 3 measures of central tendency: the mode, the median and the mean. Let's start with the mode. There is one value that occurs twice: 7.4. So the mode is 7.4. The median is the middle value when we order our values from low to high. *This* is the order of the values. We have two middle values: 7.1 and 7.4. The mean of these two values is 7.1 plus 7.4 divided by 2 equals 7.25. That's our median. To compute the mean, we use *this* formula. First, we add up all values and then we divide that outcome by the size of our sample. That makes 54.9 divided by 8 equals 6.86. This shows you that the relatively extreme score of school 3 here causes the mean to be lower than the median.

Thirdly, we want to know how far the values of the distribution are spread out. We know three measures of dispersion. The first one is the range. The range is equal to the largest value minus the smallest value. That is 8.1 minus 4.1 equals 4. The interquartile range is the difference between the first and the third quartile. So, first we have to compute Q1 and Q3. *This* was how we computed the median. We do the same for the left side of the distribution. Q1 is the average of 6.2 and 6.7. That is 6.45. And for the right side. That's the average of 7.4 and 7.9 equals 7.65. Q3 minus Q1 equals 7.65 minus 6.45 equals 1.2. Now the third measure of dispersion: the standard deviation. It's a bit more work to compute it. We need *this* formula. First we subtract the mean from every individual score. So, that's 7.4 minus 6.86 equals 0.54. We do that for all values. *This* is the result. Next, we square all these values. 0.54 squared equals 0.2916. We do that again for all values and add these scores up. That makes 11.3388. We have now finished *this* part of the formula. The next step is to divide by n minus 1. That's 8 minus 1 is 7. 11.3388 divided by 7 makes about 1.6. As a final step we have to take the square root of this outcome. That's about 1.27. That's our standard deviation.

The fourth thing we wanted to do with our data was making a boxplot. We already have all the information we need. We have Q1 and Q3. They determine the borders of our box. So our box goes from *here* (6.45) to *here* (7.65). We display the median (or, in other words, Q2) with a horizontal line here at 7.25. Do we have outliers? Outliers are values that are more than 1.5 IQR below Q1 or above Q3. 1.5 IQR equals 1.5 multiplied with 1.2. That equals 1.8. 1.8 *below* 6.45 means that values below 6.45 minus 1.8 equals 4.65 are outliers. We have one such value: 4.1. We display the outlier with a dot. The end of the lower whisker is the minimum score that is not an outlier. That's 6.2. That's *here*.

1.8 *above* 7.65 means that values above 7.65 plus 1.8 equals 9.45 are outliers as well. We don't have values this high, so there are no outliers on this side of the box. The end of the whisker is the maximum value, which is 8.1. And *here's* our boxplot. The boxplot shows at a glance that our observations lie between approximately 6 and 8, and that we have one clear outlier.

The fifth and final thing we want to know is the z-score of school 3. *This* is the formula. So that makes 4.1 minus 6.86 divided by 1.27. That makes -2.17. This indicates that this value lies 2.17 standard deviations below the mean and can therefore be conceived of as a rather exceptional value. So what can we conclude? If your plan was to send your children to school number 3, think about it twice! Although there might be a good reason why the grade is so low, you'd at least want to know why!