1.4 – Descriptive Data Story of a Data Point

So far, we've learned how to calculate the values of center and spread and interpret what they mean. But what happens if we want to focus on a single data point instead? Can we determine whether that data point is typical, exceptional, or otherwise unusual compared to the rest?

1 Measure of Relative Standing

The descriptive measures of the relationship of a data point to the rest of the data set is called *relative standing* [2]. The specific measure of relative standing we will use for this course is the z-score. The z-score uses the mean, $\bar{x}$, and standard deviation, s, of the data set to specify the location of a record in terms of the number of standard deviations away from the mean [2, 1].

Definition 1.4.1 (Z-Score [1])
The sample *z-score* for a data point i of variable x is

$$z_i = \frac{x_i - \bar{x}}{s}$$

For example, remember the analogy of 20 cadets deadlifting and then the Olympic lifter coming in and lifting 700 lbs? Let's examine the Olympic lifter and two other cadets to see how they compare. First we need to find the mean and standard deviation of the data, including the value for the Olympic lifter. Then let's select two other cadets to compare, one with a deadlift of 150 lbs and the other who lifts 350 lbs. Problem 1 shows the calculations for the z-scores for all three.

If a z-score is near 0, then the measure is close to the mean of the sample. Therefore, the farther a z-score is from zero, the more unusual it is. A large positive z-score means the value is far above average; a large negative one means it is far below average. [2]. Z-scores are also called *standardized values* [1]. From the solution of Problem 1, we see that the Olympic lifter's lift is over three standard deviations from the mean, while Cadet 1 who only lifts 150 lbs is under the mean by just over one standard deviation. Cadet 2, in comparison to the other two lifters, is close to the mean.

The question remains, is the Olympic lifter's result unusual or exceptional? The best way to determine this is to locate where the data is within the histogram. Does it appear to be an outlier or does the data have high variability? If it does appear to be an outlier, and the z-score supports it being far away from the mean, then we can consider that point unusual. This is where the interpret step of the modeling process comes into play.

Problem 1.4.1: Deadlift Z-Scores

Given two cadets who lift, 150 and 350 lbs respectively, and one Olympic lifter who lifts 700 lbs, calculate how many standard deviations away from the mean each lift is if $s = 114.9$ and $\bar{x} = 271.9$.

Solution:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Olympic Lifter:

$$z = \frac{700 - 271.9}{114.9}$$

$$z = 3.73$$

Cadet 1 (lifts 150 lbs):

$$z = \frac{150 - 271.9}{114.9}$$

$$z = -1.06$$

Cadet 2 (lifts 350 lbs):

$$z = \frac{350 - 271.9}{114.9}$$

$$z = 0.68$$

2 Comparing Across Variables

We've looked at how to interpret a single z-score to understand how a lifter would compare to the rest of the group in one event, but what if we want to compare across different events?

For example: Who is more fit overall - Cadet 2 or an Olympic lifter?

Suppose Cadet 2 runs the two-mile in 12.15 minutes and deadlifts 350 lbs, while the Olympic lifter deadlifts 700 lbs, but runs the two-mile in 19.42 minutes. We can see that the Cadet runs a faster two-mile and the Olympic lifter lifts more weight, but these events are on different scales. How can we make a fair comparison as to who is more fit overall relative to the mean?

First let's look at determining who is farthest from the mean in each event. From Problem 1 we know that the Olympic lifter has a z-score of 3.73, making him stronger than Cadet 2 who had a z-score of 0.68 for deadlift. We can similarly calculate the z-scores their two-mile run times. From Problem 2 we can see that the lifter is slower than the mean by two standard deviations, while Cadet 2 is almost two standard deviations faster than the average. We still don't know who is more fit overall.

Because our z-scores are *standardized*, meaning they're on the same scale, the easiest way to determine who is the best overall would be to add the z-scores together [1]. However, in this example we have a slight problem. When analyzing weight a larger positive z-score is better,

while for run time a larger negative z-score is better. The direction of our z-scores are different. Let's put the run time z-score into the same direction as the deadlift by finding $-z$ [1]. Now the Olympic lifter has a run time score of -2.00 and Cadet 2 has a score of 1.75.

Problem 1.4.2: Run Time Z-Scores

Cadet 2 runs the two-mile in 12.15 minutes and the Olympic lifter runs it in 19.42 minutes. If $s = 1.9$ and $\bar{x} = 15.5$ determine the standardized value of their run times.

Solution:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Olympic Lifter:

$$z = \frac{19.42 - 15.5}{1.9}$$

$$z = 2.00$$

Cadet 2:

$$z = \frac{12.15 - 15.5}{1.9}$$

$$z = -1.75$$

If we add the respective variable z-scores together we see that Cadet 2 has a total relative standing of $z = 2.43$ whereas the Olympic lifter has a relative standing of $z = 1.72$. This suggests that Cadet 2 is more fit overall relative to the average cadet. Despite the lifter's exceptional deadlift, Cadet 2's strong showing in both events gives them a higher composite standing.

3 Finding Outliers

We can also use z-scores as a tool to identify potential outliers within a data set. Remember from the previous lesson that an outlier is a point that falls outside the main cluster of the data. Without context to know what is typical or unusual for a dataset, it can sometimes be difficult to determine whether a point is truly an outlier. You will learn how to formally test for outliers in MA206, but for the purposes of this course, we will use the following rule of thumb: a data point *may be* considered an outlier if its z-score is greater than 2.5 or less than –2.5. This is not a strict rule; it's simply a guideline to help you assess the validity of your data and decide whether further investigation is needed.

Let's pull up the cleaned data that we first saw in the *1.2 – Modeling with Descriptive Statistics* reading, shown again in Table 1. We can calculate the z-scores of every measurement and then sort each event to quickly identify outliers.

Table 2 provides the standardized values for each record for both the run time and deadlift. From this table, we can see that record `A011` has a potentially unusual run time. We had previously flagged this as a possible outlier, and now we have numerical evidence to support that concern. Similarly, `A012`'s deadlift is 2.30 standard deviations above the mean. This raises a new question: Was our assumption that the original entry was a typo (changing 4000 to 400) correct? Or is 400 itself an extreme but valid lift? These standardized values don't give us final answers, but they help us decide what's worth a second look.

Table 1: Sample Cadet Fitness Data

ID	Run Time	Deadlift	Age
A001	14.8	300	20
A002	13.5	275	19
A003	15.2	320	21
A004	13.9	290	20
A005	14.1	310	19
A010	14.0	295	20
A011	10.7	280	
A012	15.5	400	22

Table 2: Sample Cadet Fitness Data with Standardized Values for Each Event

ID	Run Time	Std. Run Time	Deadlift	Std. Deadlift
A001	14.8	0.56	300	-0.22
A002	13.5	-0.33	275	-0.84
A003	15.2	0.84	320	0.28
A004	13.9	-0.05	290	-0.47
A005	14.1	0.09	310	0.03
A010	14.0	0.02	295	-0.34
A011	10.8	-2.5	280	-0.72
A012	15.5	1.04	400	2.30

Now that we've identified potentially unusual values using z-scores, what should we do with that information? In some cases, an outlier might signal an error in the data, such as a typo or an incorrect unit, and warrant a closer look or correction. In other cases, it might represent a real but rare event, like an exceptionally strong or fast cadet. Identifying outliers is not about removing data automatically; it's about asking good questions. Does this value make sense given the context? Could it affect the conclusions of my analysis?

Ultimately, recognizing outliers helps us reflect on the quality of the data, evaluate assumptions we've made, and decide whether further investigation or model adjustment is necessary.

References

[1] Richard D. De Veaux, Paul F. Velleman, and David E. Bock. *Intro Stats*. Pearson, 2009.

[2] James T. McClave and Frank H. Dietrick III. *Statistics*. Dellen Publishing Company, 1979.