Data Analysis of 2016 Rio Olympics Decathlon

Johnny Weng

San Diego State University Sports MBA

The following data analysis will review results of the 2016 decathlon event that took place at the Summer Olympics in Rio De Janeiro, Brazil in 2016. We will first analyze the events by identifying the event with the most and least variance and show our discoveries as far as what inferences can be made from our findings. We will then go into detail in analyzing relationships amongst the events, such as correlation in performance, any event(s) that we found to be unique against the others, which event best predicts the final standings, and a recommendation of one event to be eliminated based on our analysis. Additionally, we will do a couple of comparisons of athletes by finding the two athletes that had the most similar performance and honing in on Olympic decathlon champion Ashton Eaton's performance in 2012 in comparison to 2016. We hope to provide an in-depth analysis through answering questions with the applicable data.

Analyzing the Events

In wanting to identify the events with the most and least variance, we first tracked down our data set of all competitors of the decathlon in the 2016 Summer Olympics. Although there were 32 initial competitors, we found 23 to best fit our data set, as the remaining competitors either did not start or finish each event, or were penalized in one or more events, resulting in a score of 0/DNF. At this point, we calculated the mean, standard deviation, and coefficient of variation for each of the ten decathlon events, as shown in Table 1. We kept in mind that these values may be quite significantly different, as each event is particularly different in their scales. For example, there was roughly one second in between the first-place finisher and the last place finisher in the 100m event, whereas the javelin throw separates first and last place by 20 meters.

Table 1

Descriptive Data and coefficient of variation for Rio 2016 Decathlon

Event	*Mean*	*Standard Deviation*	*Coefficient of Variation*
100m	10.903	0.248	2.28%
Long Jump	7.314	0.316	4.32%
Shotput	14.145	1.049	7.41%
High Jump	2.022	0.093	4.62%
400m	48.866	1.178	2.41%
110m HH	14.617	0.633	4.33%
Discus Throw	44.217	4.139	9.36%
Pole Vault	4.848	0.304	6.28%
Javelin Throw	61.311	6.421	10.47%
1500m	273.696	10.029	3.66%

After calculating the mean, standard deviation, and coefficient of variation for each of the events, we can conclude that the 1500 meter race had the highest variance, and the high jump had the lowest variance. This shows us that there was very little difference between the scores in the high jump, and that there were a wide range of different finishing times for the 1500m. As previously stated, since each event has varying levels of qualifying metrics, such as feet, time, etc., we identified that we could get a better glimpse of the results by finding the coefficient of variation; this shows us the ratio of the standard deviation to the mean. After evaluating the coefficients of variation, we found that the javelin throw led the pack with a 10.47% ratio and could therefore be considered as having the highest variation in the data series. The lowest coefficient of variation, the 100m race, shows that it had the lowest variation, which makes sense as we previously mentioned that roughly one second separated the first and last place finishers.

From this data, we can infer that identifying the coefficient of variation gives us a better glimpse into finding the variation between each event, since the majority of the events are measured and calculated (and therefore scored) differently. This means we are better off evaluating the variance of each event by utilizing the coefficient of variation rather than the

variance. We can also infer that the 100 meter event among the competitors is one of the more competitive events, as the level of skill may be closer and therefore the finishing results would be close as well. Looking at the Javelin numbers, we can see that the level in skill or distance is extremely varied, possibly pertaining to the fact that this is one of the more difficult events to compete in. Because of this fact, we can see even after weighing the numbers in terms of ratios, Javelin still has the highest variance.

Analyzing the Relationship Among the Events

In attempting to analyze some relationships amongst the decathlon events, we felt that producing a correlation matrix would give us a prime opportunity to do so. In Table 2, we ran a correlation matrix which shows us the correlation between each set of events. The results are shown below:

Table 2
Correlation Matrix of 2016 Rio Olympics Decathlon

	100m	*Long Jump*	*Shotput*	*High Jump*	*400m*	*110m HH*	*Discus Throw*	*Pole Vault*	*Javelin Throw*	*1500m*
100m	1									
Long Jump	-0.429	1								
Shot Put	-0.055	-0.073	1							
High Jump	0.072	0.443	-0.091	1						
400m	0.630	-0.644	-0.058	-0.243	1					
110m HH	0.409	-0.468	-0.084	-0.378	0.708	1				
Discus Throw	-0.005	-0.094	0.430	0.050	0.273	0.263	1			
Pole Vault	0.205	0.568	-0.025	0.387	-0.155	-0.225	0.032	1		
Javelin Throw	0.116	0.048	0.404	0.296	-0.139	-0.330	0.423	0.005	1	
1500m	0.315	-0.608	0.166	-0.272	0.628	0.442	0.437	-0.169	0.078	1

Note. N = 23.

After we ran this unadjusted correlation matrix, we needed to find out what results were significant or not. Since our N = 23, we utilized alpha to identify a critical value of r at .413. This means we would deem all results within -.413 through .413 as insignificant. To show only our significant results, we created an adjusted correlation matrix shown in Table 3:

Table 3

Adjusted Correlation Matrix of 2016 Rio Olympics Decathlon

	100m	*Long Jump*	*Shotput*	*High Jump*	*400m*	*110m HH*	*Discus Throw*	*Pole Vault*	*Javelin Throw*	*1500m*
100m	1									
Long Jump	-0.429	1								
Shot Put			1							
High Jump		0.443		1						
400m	0.630	-0.644			1					
110m HH		-0.468			0.708	1				
Discus Throw			0.430				1			
Pole Vault		0.568						1		
Javelin Throw							0.423		1	
1500m						0.442	0.437			1

Note: N =23. R = .413 & -4.13

After creating both of these matrices, it allowed us to create better conclusions in regards to analyzing specific relationships between decathlon events. The two events that seemed to be most correlated in terms of performance were the 110m HH and 400m running events, evidenced by a significance factor of .708. The closer a value is to 1, the more correlated the value is. As far as uniqueness goes, we found similarities between the three throwing events; the discus throw, the javelin throw, and shotput. However, we decided that the javelin throw was most unique for a few reasons. The first reason being that it only showed significant correlation to one other event, and furthermore the least amount of significance per correlation with a value of .423. Additionally, we found this to be unique because the javelin throw also had the highest coefficient of variation; which in our opinion, just shows how different the event is from the others.

To find the event that best predicts the final standings, we utilized another correlation matrix, but this time including the final standings. By integrating each finisher's final amount of points in the matrix, we were able to create Table 4 below:

Table 4

Correlation of events in relation to finishing score

	100m	*Long Jump*	*Shotput*	*High Jump*	*400m*	*110m HH*	*Discus Throw*	*Pole Vault*	*Javelin Throw*	*1500m*	*Finishing Score*
100m	1										
Long Jump	-0.429	1									
Shot Put	-0.055	-0.073	1								
High Jump	0.072	0.443	-0.091	1							
400m	0.630	-0.644	-0.058	-0.243	1						
110m HH	0.409	-0.468	-0.084	-0.378	0.708	1					
Discus Throw	-0.005	-0.094	0.430	0.050	0.273	0.263	1				
Pole Vault	0.205	0.568	-0.025	0.387	-0.155	-0.225	0.032	1			
Javelin Throw	0.116	0.048	0.404	0.296	-0.139	-0.330	0.423	0.005	1		
1500m	0.315	-0.608	0.166	-0.272	0.628	0.442	0.437	-0.169	0.078	1	
Finishing Score	-0.372	0.757	0.321	0.608	-0.681	-0.694	0.230	0.503	0.531	-0.468	1

Note. Yellow highlight shows correlation of finishing score to each event. Green highlight shows that Long Jump has highest correlation to finishing score.

By integrating the finishing score as a variable, we were easily able to identify which events had significant correlation to the final standings, which included the high jump, the 400m, the 110m HH, pole vault, javelin throw, 1500m, and the event which had the most correlation, the long jump.

If we had to eliminate one event from the decathlon, we would eliminate the discus throw event. The primary reason is that it has the lowest correlation to final standings, as shown in Table 4. Since it has a small correlation with the final results, we believe that if it were removed from the decathlon, the standings wouldn't change as much as if a higher correlated event, such as the long jump, were removed.

Comparing Athletes

In order to find the two most correlated athletes, we again ran a correlation matrix between all of the athletes with their adjusted z-scores for each event. Although all of the athletes are extremely correlated since the Olympics typically involves the most spectacular athletes at their craft, the two athletes that showed most correlation were Adam Helcelet and Kurt Felix, with a correlation of .999924. In order to specifically identify how correlated their results were,

we calculated z-scores of each of their results and created a graph for better visualization as seen in Figure 1. Ashton Eaton, winner of the decathlon event in both the 2012 and 2016 summer games, is only the third Olympian to achieve back-to-back gold medals in the decathlon. In order to compare his results from the 2012 to the 2016 games, we created a similar comparison in calculating z-scores for each of his events. A visual representation, shown in Figure 2, provides an interesting look at his results; he fared better in seven of the events in the 2012 games than in the 2016 games. Probably seeing that his success would be tough to repeat in four years, Eaton promptly declared his retirement after the 2016 games, likely going down as one of the best decathletes to ever live.

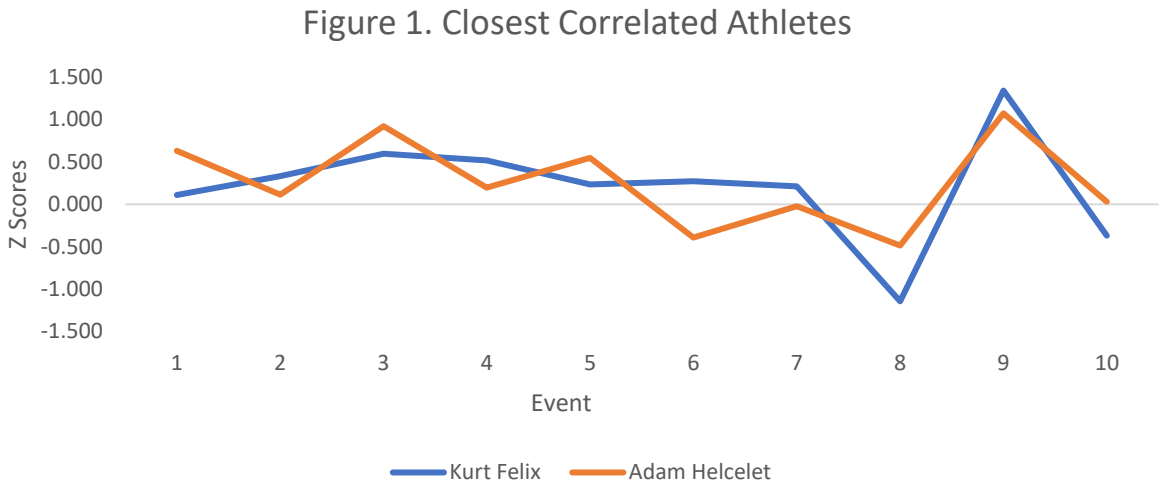

Figure 1. Events are described as follows: 1 = 110m; 2 = Long Jump; 3 = Shotput; 4 = High Jump; 5 = 400m; 6 = 110 HH; 7 = Discus Throw; 8 = Pole Vault; 9 = Javelin Throw; 10 = 1500m.

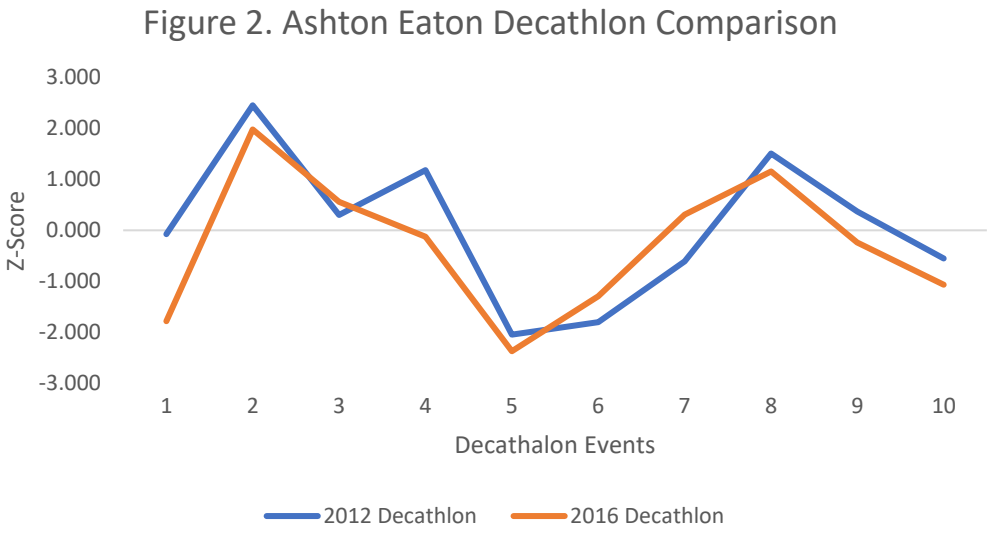

Figure 2. Events are described as follows: 1 = 110m; 2 = Long Jump; 3 = Shotput; 4 = High Jump; 5 = 400m; 6 = 110 HH; 7 = Discus Throw; 8 = Pole Vault; 9 = Javelin Throw; 10 = 1500m.

Final Comments

In utilizing data, it gives us a much better perspective on how we can analyze specific happenings, events, and circumstances, such as the 2016 Olympic decathlon. Instead of looking at the standings and taking a fledgling guess at what event is the most or least important, or what event is most equal, we can instead utilize analytical skills to come up with almost precise answers to those questions. Identifying the coefficient of variation allowed us to better analyze the variance of one event to another. Running multiple correlation matrices allowed us to gain a better understanding of the importance (or unimportance) of specific decathlon events in relation the others. Ultimately, by using the aforementioned skills, we were able to perform an in-depth analysis on some interesting byproducts of the Olympic decathlon.