Ryan Long
Bellevue University
DSC-530 Winter 2019-2020 - Term Project

Statistical/Hypothetical Question
The statistical questions I wanted to answer were, 1) Do most runners pace with negative splits and 2) does negative splits result in a faster marathon pace?

Negative splitting is the practice of finishing a race faster than the initial pace [1]. There is no defined method for the practice in terms of how a distance can be segmented. A distance can be divided equally, with the second half completed faster than the first. For example the first half a mile can be run in 2:55 (175 seconds) and the second half in 2:35 (155), for a total time of 5:30 (330). Alternatively, the distance can be quartered, with quarters subsequent the first completed faster than the previous. Example quarter miles could be covered in 1:30 (90), 1:25 (85), 1:20 (80), and 1:15 (75), again for a total time of 5:30 (330).

For purposes of my exploratory data analysis project, Negative split (NS) will mean a runner who ran faster in the second half of the race than the first (increased speed). Positive split (PS) will mean a runner performed slower in the second half of their race than the first (decreased speed).

Outcome of your EDA
Results from the 2019 IMT Des Moines Marathon were used for analysis. The dataset contained a wide variety of variables including 10k, Half Marathon, 20 mile, and Last 10K times along with Total time and Average Pace. Additional variables were calculated to refine the analysis into roughly quarter segments. In addition to recoding additional variables, pacing metrics were converted from HH:MM:SS format to a seconds per mile integer value.

I plotted a PMF using the first half and second half pace for all runners. Based on the plot, it showed runners are more likely to run slower in the second half of the race. I also calculated a CDF using the first half and second half pace. It showed the fastest runners and slower runners were likely to run near their average paces given the plotted overall CDF.

Of the 1,276 finishers will complete, data 139 NS and the remaining 1,137 PS. Mean race pace for the entire population was 609 seconds per mile (10:09), while NS were 529 (8:49), and PS were 619 (10:19). The standard deviation for PS was 115 seconds, while NS was 94 seconds. The NS split group exhibited a lower and tighter execution of pace.

To model the null hypothesis, which is the distribution of average pace of the negative and positive split groups are the same, a permutation or difference of means test was performed. The resulting p-value from the test was 0, thus the conclusion that the difference in average pace of the two groups is significant. Regression analysis was performed on the two groups, Negative Splitters (NS) and Positive Splitters (PS). Average Pace for both groups was used as the dependent variable, and first and second half paces were chosen as the independent variables.

Based on the R-squared calculations First (0.994) and Second (0.993) half pace provide a strong relationship to the average pace of negative splitters. A positive relationship between positive splitters first half pace exists (0.910), but the second half pace is stronger (0.950). This information combined with the histograms, PMF, and CDF analysis indicate though fewer runners negative split, it results in a significantly faster average race pace.

What was missed during the analysis?
Experience in and knowledge of execution. The number of negative splitters represented roughly 11% of the population. This could indicate runners are inexperienced, unknowledgeable, or incapable (undertrained) of executing

negative splits as it clearly results in a lower average pace. Being able to quantify these attributes for so many individuals would require many more data points.

Were there any variables which could have helped in the analysis?
I was unable to sufficiently incorporate average gradient for each of the segments into the analysis. It would have been useful to understand if the second half of the race had a net gain in elevation compared to the first half or if it was flat. Temperature during the day and throughout the day has an impact on races lasting more than 2 hours. The NS group averaged 8:49 per mile, which is roughly an average time of 3:51:00. With a race time of early morning spanning close to 11AM or 12PM temperatures can swing 20-40 degrees and negatively impact the second half of the race. Knowing and being able to quantify the impacts of external variables such as gradient and temperature on self selected pace could have provided context to the analysis.

Were there any assumptions made you felt were incorrect?
As noted above, assumptions made which may be incorrect were runners were experienced, knowledgeable, and capable of negative splitting. Some runners in the PS group may be looking for a one-and-done experience, didn't train, and didn't know they should have started slower or equal to the pace they would hope to achieve.

What challenges did you face, what did you not fully understand?
I initially had trouble navigating the data and determining how to best represent the paces and split out the data into two groups of PS and NS. I felt challenged, but it was not impossible to learn about formats and code within pandas which we did not cover directly in our class exercises. What I have not fully thought about, but should be considered is incorporating data from multiple events and performing a more detailed analysis of age group categories.

References
Negative Split. (n.d.) Retrieved November 27, 2019 from Wikipedia https://en.wikipedia.org/wiki/Negative_split

http://onlineraceresults.com/race/view_race.php?race_id=70421&submit_action=select_result&order_by=default&group_by=default#results