**Introduction**

For this activity, you will be exploring data from the 2019 Boston Marathon by analyzing the results of the finishing runners. Focusing on the single quantitative variable of result times in seconds, you will examine both visualizations and summary statistics to make key conclusions. The incorporation of z-scores will allow for comparisons to be made between two subsets of the data by determining performance of top finishers.

**Learning Goals**

By the end of the activity, you will be able to:

* Analyze structure and distributions of histograms and boxplots
* Use basic summary statistics to assess center and spread
* Calculate further relevant metrics for quantitative data
* Determine which metrics are most relevant to your data
* Calculate and compare z-scores for individual cases

**Methods**

For this activity, students will primarily use basic concepts of histograms and boxplots to analyze distributions. Students will also require formulas for deviation, IQR, standard deviation, fences, and z-score. For r users, summary statistics can be calculated using the summary() command.

**Single Quantitative Variable**

A picture containing diagram, line, technical drawing, plot

Description automatically generated

**The histogram of race results times (seconds) has 16 bins. How would it change if we had 5 bins? What about 100 bins?**

**How would you describe the shape of the histogram? (Left-Skewed/ Right-Skewed/ Approximately Symmetric) (unimodal / multimodal)**

*Summary statistics for all finishers of the race.*

**A picture containing text, font, white, calligraphy

Description automatically generated**

**Calculate the deviation of a result time of 16000 seconds.**

**The variance of this sample is 7,278,396. Provide an interpretation for this number in context.**

**Calculate the standard deviation for the sample.**

**Find the IQR (Interquartile range).**

**Using your answer to the previous question, determine the values of the upper and lower fence.**

**Is a time of 21000 seconds an outlier? What about a time of 7500 seconds?**

**Do you think the mean or median is higher for the distribution? Explain your reasoning.**

**A picture containing screenshot, line, rectangle, text

Description automatically generated**

**Using the histogram, boxplot, and summary statistics you calculated, determine whether the mean or median is a more representative statistic to describe the center?**

**Is IQR or standard deviation a better assessment of spread in this instance?**

***A picture containing text, font, receipt, screenshot

Description automatically generated***

**Using the mean and standard deviation of the Men’s field results, calculate the z-score of the top male finisher who completed the race in 7677 seconds.**

**A picture containing text, font, screenshot, receipt

Description automatically generated**

**Using the mean and standard deviation of the Women’s field results, calculate the z-score of the top female finisher who completed the race in 8611 seconds.**

**Based on your answers to the previous two questions, determine which top finisher had the more remarkable result in relation to their respective field.**