**Introduction**

For this activity, you will be exploring data from the 2023 Boston Marathon by analyzing the result times of the finishing runners. Focusing on the single quantitative variable of result times in minutes, 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 performances of top finishers.

**Learning Goals**

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

* Analyze structures and distributions of histograms
* 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

**Data**

The data below shows the top 6 finishers of the 2023 Boston Marathon. In addition to the **name** of the runner, the **finish\_net\_minutes** variable shows the race result time of the runner that is timed from when they cross the starting gate at the beginning of the race to the finish line. The full data (boston\_marathon\_2023.csv) is available on the GitHub repo associated with this module and contains many more variables associated with the runner information and times.

**A picture containing text, font, screenshot

Description automatically generated**

**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() and var() commands.

**Exercises**

**A picture containing screenshot, diagram, plot, pixel

Description automatically generated**

1. a. Consider the histogram above comprised of race results times in seconds with 20 bins. How would it change if it had 5 bins? What about 100 bins?

If it had 5 bins, it would condense to 5 columns that group together thousands more times. With 100, we would have 100 different time groups and the histogram would be far less condensed.

1. How would you describe the shape of the histogram?

Slightly right skewed and unimodal.

*Summary statistics for result times (minutes) of all finishers of the race.*

****

1. Calculate the deviation of a result time of 295 minutes.

222.5 – 295 = 2649

1. The variance of the sample is 1914.45. Calculate the standard deviation and provide an interpretation.

Standard Deviation: = 43.754

Interpretation: On average, the race times vary by 43.754 minutes around the mean race time.

1. What is the value of the interquartile range (IQR)?

244.4 – 191.1 = 53.3

1. Using your answer to the previous question, calculate the values of the upper and lower fence for determining mild and extreme outliers.

Mild:

Q3 + 1.5 \* IQR = Upper fence

Q1 – 1.5 \* IQR = Lower fence

244.4 + 1.5\*53.3 = 324.35 (minutes)

191.1 – 1.5\*53.3 = 111.15 (minutes)

Extreme:

Q3 + 3 \* IQR = Upper fence

Q1 – 3 \* IQR = Lower fence

244.4 + 3\*53.3 = 404.3 (minutes)

191.1 – 3\*53.3 = 31.2 (minutes)

1. Is a result time of 405 minutes an outlier? If it is an outlier, state if it is mild or extreme.

Yes, a time of 330 minutes is an extreme outlier because it exceeds the upper fence.

1. What about a result time of 115 minutes? Is there anything interesting about this conclusion? Explain your reasoning.

While a time of 115 minutes is not considered an outlier in this case, it is 10.9 minutes faster than that of the first places finisher. This is due to the nature of the race that produces a higher density of finishers in the front pack, while other running times taper off with greater deviation. This can be seen in the right skew of the distribution.

1. The mean and standard deviation of only men’s results can be found below. Using this information, calculate the z-score of the top male finishers who completed the race in 125.9 minutes. Provide an interpretation of the result.

**Mean Standard Deviation**

212.43 43.66

Z-score = (125.9 - 212.43) / 43.66 = -1.98

The result time of the top male finishers is -1.98 standard deviations away from the mean time for all men.

1. The mean and standard deviation of only women’s results can be found below. Using this information, calculate the z-score of the top female finishers who completed the race in 142 minutes. Provide an interpretation of the result.

**Mean Standard Deviation**

235.93 40.14

Z-score = (141.63 - 235.93) / 40.14 = -2.34

The result time of the top female finishers is -2.35 standard deviations away from the mean time for all women.

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

The top female finishers had the more remarkable result in relation to their field because their z-score was comparatively larger than that of the top male finishers.