**Understanding Variability in Boston Marathon Finish Times**

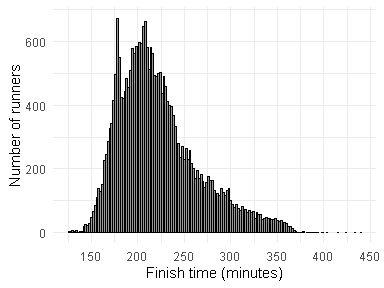
The Boston Marathon is an annual marathon traditionally held on Patriots’ Day, the third Monday of April. Begun in 1897, the event is one of the world’s oldest annual marathons and ranks as one of the world’s most prestigious road racing events. The Boston Marathon attracts runners from all over the world, featuring thousands of participants and extensive spectator support, embodying the spirit of community and endurance.

We will be investigating the results from the 2023 Boston Marathon - which consists of 26,598 runners that completed the race. In addition to the name and gender 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 at [SCORE Data Repo](https://data.scorenetwork.org/running/boston_marathon_2023.html) and contains many more variables associated with the runner information and times.

suppressPackageStartupMessages(library(tidyverse))  
boston <- read.csv("https://data.scorenetwork.org/data/boston\_marathon\_2023.csv")

1. The histogram below displays the distribution of finish times for all 26,598 participants. Use it to help answer the following questions.

hist\_all <- boston |>  
 ggplot(aes(x = finish\_net\_minutes,  
 y = ifelse(after\_stat(count) > 0, after\_stat(count), NA))  
 ) +   
 geom\_histogram(color = "black", fill = "grey60", binwidth = 2) +  
 theme\_minimal() +  
 scale\_x\_continuous(breaks=seq(100,500, by = 50)) +  
 labs(x = "Finish time (minutes)", y = "Number of runners")  
  
hist\_all



1. Describe the shape, center, and spread of the distribution.

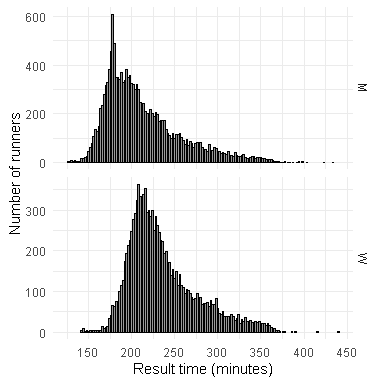
*The distribution of finish times (in minutes) for 2023 Boston Marathon participants is seemingly bimodal and skewed towards higher values. (i.e., right skewed). There is a mode for runners that typically takes approximately 175 minutes, another mode around 210 minutes, and most runners finish in the 140 minute to 375 minute range.*

1. Identify some potential confounding variables that might help explain my their is variability in finish times.

*Answers will vary. A few potential confounding variables that might explain variability in marathon finish times include gender/sex, age, training regimen, previous marathon experience, and individual health conditions or injuries.*

1. The output below displays the distribution of finish times participants separated by gender (M = Men, W = Women) as well as some summary statistics. Use it to help answer the following questions.

hist\_by\_gender <- boston |>  
 ggplot(aes(x = finish\_net\_minutes,  
 y = ifelse(after\_stat(count) > 0, after\_stat(count), NA))) +   
 geom\_histogram(color = "black", fill = "grey60", binwidth = 2) +  
 theme\_minimal() +  
 labs(x = "Result time (minutes)", y = "Number of runners") +  
 scale\_x\_continuous(breaks=seq(100,500, by = 50)) +  
 facet\_grid(vars(gender), scales = "free\_y")  
hist\_by\_gender



sumstat\_df <- function(x, probs = c(0,0.25, 0.5, 0.75,1)) {  
 tibble(  
 val = c(mean(x, na.rm = TRUE), sd(x, na.rm = TRUE),   
 quantile(x, probs, na.rm = TRUE)  
 ),  
 quant = c("Mean","SD","Min", "Q1", "Med", "Q3", "Max")  
 )  
}  
boston |>   
 select(gender,finish\_net\_minutes) |>  
 reframe(sumstat\_df(finish\_net\_minutes), .by = gender) |>  
 tidyr::pivot\_wider(names\_from = quant, values\_from = val) |>  
 knitr::kable()

| gender | Mean | SD | Min | Q1 | Med | Q3 | Max |
| --- | --- | --- | --- | --- | --- | --- | --- |
| M | 212.4253 | 43.65522 | 125.9000 | 179.5583 | 200.9333 | 233.7417 | 433.6500 |
| W | 235.9280 | 40.13792 | 141.6333 | 207.6750 | 226.0500 | 255.8667 | 440.5167 |

1. Compare and contrast the distributions for men and women finish times. Be sure to comment on how the centers, spreads, and shapes compare.

*Center: Men’s times are centered around approximately 175 minutes while Women’s times are centered closer to 210 minutes.*

*Spread: The spread of Men’s times is slightly larger ranging from 125 minutes to approximately 375. (There are a few runners beyond that, up to about 440 minutes.) Women’s times range from about 140 minutes to 375 minutes. (With a few runners again having larger times.)*

*Shape: Both distributions are unimodel and right skewed with Men’s times exhibiting slightly more skew.*

1. Evans Chebet, from Kondabilet, Kenya, was the fastest male to complete the race clocking in at 125.9 minutes. (Evans also won the 2022 Boston and 2022 New York marathons) Hellen Obiri, from Kisii, Kenya, won the women’s competition with a time of 141.63 minutes. (Hellen also won the 2023 New York marathon and is a two-time silver medalist in the 5000m event.) Relative to their competition, which runner, Evans or Hellen, had the better finishing time? Justify your answer numerically.

*Answers will likely vary. Instructors are encouraged to discuss how there may not be a clear way to answer to this question. First, students likely need to decide on what “competition” means. (They will likely go to just comparing within gender - but possibly think about incorporating age groups too.)*

*Some potential routes students may go are:*

* *checking if either time is an outlier (according to the 1.5IQR rule, neither is an outlier)*
* *Distance from a center point or Q1 (likely concluding that Hellen’s time is relatively better)*
* *Using z-scores: while*

1. What other information might be useful to collect about that could be incorporated into an analysis like this? What type of issues might occur with getting this information?

*An obvious choice would be age of the runner. We have age groups, so we could partially incorporate this information, but not perfectly so. Other pieces of information would be much harder to track for 26,000 runners.*