CAN STATISTICAL ANALYSIS HELP ME ARRIVE AT SCHOOL ON TIME?

BACKGROUND AND OBJECTIVE OF THE STUDY

Since I began attending my current high school in September 2024, I've realized that I frequently arrive late. The school bell rings at 8:10 AM, and I need to be there by 8:05 AM to make sure that I'm on time. My parents drive me to school each day, but due to other commitments, they can't leave home too early. This has made me wonder: what's the latest time I can leave home and still arrive on time most days?

I've been learning about statistics and how it can help people make better decisions using data. This got me wondering further: can statistics help me figure out the best time to leave home for school?

With these questions in mind, I decided to learn more about statistics and collect data to solve the problem for me and my parents!

OVERVIEW OF THE APPROACH AND DATA COLLECTION

Daily travel times from my home to school were collected in two phases. The first phase was from October 10 to December 20, 2024. On each school day during this period, I recorded the time I left home (Start time) and the time I arrived at school (Arrival time). Additionally, I tracked the waiting times at three red lights along my route to school.

After phase one was completed, I analyzed the collected data during the Christmas break using basic statistics. and developed a strategy according to the analysis to adjust my Start time so that I could arrive at school on time.

Since January 6, 2025, my parents and I have been following the strategy. Daily travel times were recorded from January 6 to February 4, 2025, marking the second phase of the study. The data collected during this phase were used to evaluate the effectiveness of my strategy.

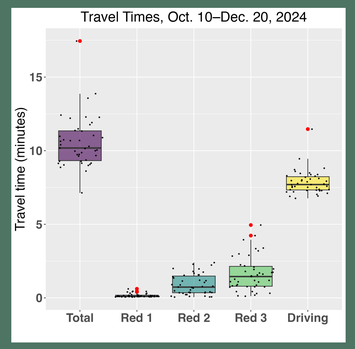
MAIN STATISTICAL CONCEPTS USED

- **Percentile**: A kth percentile of a dataset is a value such that k% of the data points are smaller than that value and the other (100-k)% of the data points are greater than the value. The 50th percentile is also called the ***median***.
- **Interquartile range (IQR)**: The difference between the 75th percentile and the 25th percentile. The IQR reflects the variability of the dataset.
- **Box plot**: A figure displaying some summary statistics. It includes:
  - A box that represents the IQR of the data, which spans vertically from the 25th percentile (Q1) to the 75th percentile (Q3).
  - A horizontal line inside the box that represents the median of the data.
  - Two vertical lines, known as whiskers, that extend from the bottom and top edges of the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3. Data points outside of this range are considered outliers—observations that deviate significantly from the rest of the data..
- **Mean**: The average value of the dataset.
- **Standard deviation (SD)**: A quantity that measures the variability of the data relative to its mean.

ANALYSIS OF DATA COLLECTED IN PHASE ONE

Travel time data were collected over 45 school days in phase one (Oct. 10–Dec. 20, 2024). During this period, I arrived after the targeted arrival time of 8:05 AM on 31 days, and after the bell time of 8:10 AM on 14 days.

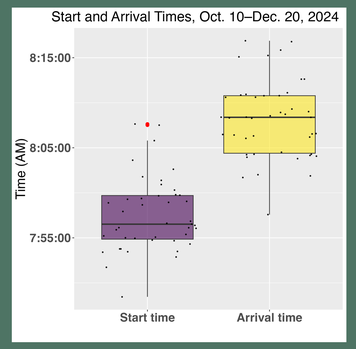
Box plots and summary statistics of daily total travel times from my home to school (Total), waiting times at the three red lights along the route (Red 1, Red 2, Red 3), and driving times excluding waiting times at red lights (Driving) are presented. Daily total travel times exhibit a large variability, primarily due to fluctuations in Red 3 and Driving. An outlier is observed in daily total travel times. This outlier coincides with the outlier in Driving and the most extreme outlier in Red 3. The mean of daily total travel times is slightly greater than the median, indicating that longer travel times tend to be significantly prolonged.



		Total	Red 1	Red 2	Red 3	Driving
Percentiles	90th	12.3	0.4	1.9	3.3	8.6
	75th	11.4	0.2	1.5	2.1	8.2
	50th	10.2	0.1	0.7	1.4	7.7
	25th	9.3	0.1	0.3	0.8	7.3
	10th	9.0	0.1	0.1	0.3	7.1
Mean		10.6	0.2	0.9	1.6	7.9
SD		1.7	0.1	0.7	1.2	0.8
IQR		2.1	0.1	1.2	1.3	0.9

Summary Statistics of Travel Times (unit: minute), Oct. 10–Dec. 20, 2024

The Start times show considerable variation, with two outliers, which was caused by my parents' unpredictable morning schedules. The Arrival times exhibit slightly greater variability (in terms of SD and IQR) than the Start times. This was due to the combined effect of the variations in both Start times and daily total travel times.



		Start time	Arrival time
Percentiles	90th	08:02:27	08:14:36
	75th	07:59:42	08:10:47
	50th	07:56:31	08:08:23
	25th	07:54:51	08:04:24
	10th	07:53:27	08:03:50
Mean		07:57:32	08:08:06
SD (minutes)		4.1	4.3
IQR (minutes)		4.8	6.4

Summary Statistics of Start and Arrival Times, Oct. 10–Dec. 20, 2024

CONCLUSIONS FROM PHASE ONE

1. A primary cause for my late arrivals in phase one is the late Start times. The mean Start time was 7:57:32 AM, and the mean daily total travel time was 10.6 minutes, resulting in an average Arrival time of 8:08:06 AM.
2. The large variability in both the Start times and travel times, which were driven by uncertain factors, led to unpredictable Arrival times.

STRATEGY TO AVOID LATE ARRIVALS

To arrive at school before 8:05 AM, I set a target Start time using the following formula, based on the data collected in phase one:

8:05 AM − (90th percentile of total travel times) − (IQR of Start times) ≈ 7:48 AM

The 90th percentile of daily total travel times is used in the formula instead of the median or mean to account for potential unexpectedly long travel times due to day-to-day variations. The IQR of Start times is incorporated to mitigate possible delays caused by uncertainty in the Start times.

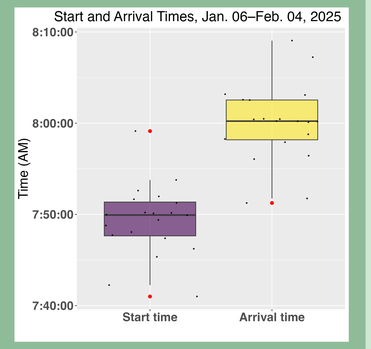
EVALUATING THE STRATEGY USING NEW DATA

Following my strategy, my parents and I have been aiming to leave home for school at 7:48 AM since January 2025. In phase two (Jan. 6–Feb. 4, 2025) of the study, the same time variables were recorded over 20 days. I arrived after 8:05 AM on two occasions, but never after 8:10 AM.

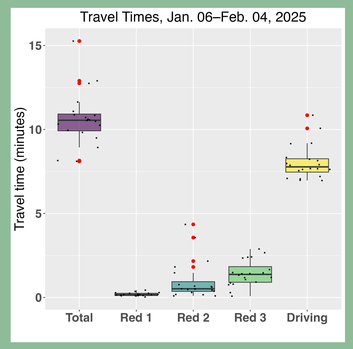
During phase two, there was one outlier in Start times, which was caused by an unexpected event at home. This outlier led to one of my two late arrivals.

		Start time	Arrival time
Percentiles	90th	07:52:44	08:03:35
	75th	07:51:21	08:02:32
	50th	07:49:56	08:00:14
	25th	07:47:39	07:58:11
	10th	07:45:02	07:55:38
Mean		07:49:21	08:00:01
SD (minutes)		4.0	4.3
IQR (minutes)		3.7	4.4

Summary Statistics of Start and Arrival Times, Jan. 6–Feb. 4, 2025



The mean Start time was 7:49:21 AM, which is later than the targeted Start time. However, this wasn't a problem because variation in Start times was accounted in my strategy. The SD of Start times was 4.0 minutes, which is similar to that observed in phase one.



A few outliers are observed in daily total travel times, Red 2 and Driving, most of which were caused by a temporary construction that slowed down the traffic near red light 2. This disruption caused my other late arrival in phase two.

Overall, the rate of my late arrivals in phase two was significantly reduced compared to phase one. The two late arrivals in phase two were due to uncontrollable factors. Given my current situation, further reducing the average Start time seems difficult. However, with effort, the variation in Start times may be reduced so that the Arrival times can be more consistent in the future.

FINAL CONCLUSION

1. By collecting and analyzing data using statistics in phase one, I came up a strategy to minimize late arrivals. This strategy is proven to be effective based on the data collected in phase two.
2. I learned some useful statistics, and, more importantly, was convinced that even simple statistics can help us solving problems in daily lives.
3. I learned that plans will be disrupted by uncertain events. By incorporating uncertainty into the planning process, we can enhance the resilience of plans against potential disruptions.