Math 341 Project 1

Isaac Boaz

March 5, 2023

Probability Questions

1. Both roles rely heavily on data analysis, with the ability to crunch vast amounts of information to make informed decisions being a vital aspect of each. In the fast-paced world of Formula One racing, Computer Science majors can utilize their skills in programming and data analysis to process the mountains of data generated by a race.
   Another similarity between these roles is the need for algorithmic thinking. In Formula One racing, strategists must break down complex variables like tire wear, fuel consumption, and track conditions into manageable components to develop effective race strategies. Similarly, Computer Science majors use algorithmic thinking to break down complex problems into smaller, more manageable pieces.

2. As binomial only checks two outcomes (success or failure), we can assign a safety car leading as a success and otherwise as a failure. One thing to note is that these laps aren't necessarily independent, as each lap may have an impact on the safety car's deployments further down the line.

3. The Poisson distribution measures the # of occurrences of an event within a fixed time/space. A requirement of this distribution is that $n \to \infty$, $p \to 0$, $\lambda = np$. As n represents the number of laps, we can assume that it should be relatively high (given a timespan of a few seasons for example). Similarly, we can assume the safety car's deployment rate should be relatively low.

4. The # of safety car deployments can be modeled as a Poisson distribution, and thus, the interval between each deployment can be represented by an exponential distribution.

5. We can assume that the two time periods are independent of each other as they are disjoint.

6.

$$P(X \geq t_1 + t_2 \mid X \geq t_1) = P(X \geq t_2), t_1 \geq 0, t_2 \geq 0$$
$$t_1 = 3$$
$$t_2 = 5$$
$$P(X \geq 3 + 5 \mid X \geq 3) = P(X \geq 8 \mid X \geq 3)$$
$$= P(X \geq 5)$$

As the memoryless' property name implies, the probability of an event occurring at a time t is independent of the time that has passed since the event occurred. Thus, the probability of an event occurring at time $t_1 + t_2$ is independent of the probability of the event occurring at time t_1. Therefore, the probability of an event occurring at time $t_1 + t_2$ (given t_1 has passed) is equal to the probability of the event occurring at time t_2.

Statistics Questions

1. Judging by the best Poisson line fit, the distributions seem to follow a Poisson distribution.

2. Since the Poisson distribution's equation is $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, we can see that as λ increases, the probability of an event occurring increases. Since Poisson takes in the average amount of times an event occurs within a given timespan, it makes sense to use the mean # of accidents for λ.

3. In general it seems that the interval between safety car deployments follows an exponential distribution.

4. Since the exponential distribution's equation is $P(X \leq t) = 1 - e^{-\lambda t}$, we can see that as λ increases, the probability of an event occurring increases. Since exponential takes in the average amount of time between events, it makes sense to use the mean interval between safety car deployments (ie the inverse of # of deployments within a set timespan) for λ.

5.

Means	
first_half	0.7291667
interval1	1.328827
second_half	0.6960784
interval2	1.421543

Going by the given means, we can see that each interval is approximately the inverse of the # of safety car deployments per half (i.e: interval $\approx \frac{1}{\#ofdeployments}$). This makes sense as the exponential distribution (time between deployments) is the distribution of the time between events, and the Poisson distribution (number of deployments within a timespan), are inversely related.

6. This dataset is not necessarily independent or identical. Taking two random samples from any time between the two data groups may result in comparing a near-beginning season race to a near-end season race.

7. It can be assumed that both datasets are distributed similarly, and are similar enough to be considered independent.

8. In the case of exponential distribution, since the variance is equal to the mean it would not make sense to assume common variance.

9. t.test() results:
   Two Sample t-test

   data: interval1 and interval2 t = -0.39129, df = 139, p-value = 0.6962
   alternative hypothesis: true difference in means is not equal to 0
   95 percent confidence interval: -0.5612108 0.3757772
   sample estimates: mean of x, mean of y / 1.328827, 1.421543

   Welch Two Sample t-test

   data: interval1 and interval2 t = -0.39133, df = 139, p-value = 0.6962
   alternative hypothesis: true difference in means is not equal to 0
   95 percent confidence interval: -0.5611611 0.3757275
   sample estimates: mean of x, mean of y / 1.328827, 1.421543

10. Based purely on the given intervals by the tests, both intervals encapsulate 0 (having negative and positive on either end). As such, there is no statistically significant change in the distribution between these two time periods.

12. Looking at the Q-Q plot, we can see that the first dataset follows the exponential distribution more closely than after the introduction of the VSC.

Exp. Q-Q Plot (First Half) **Exp. Q-Q Plot (Second Half)**
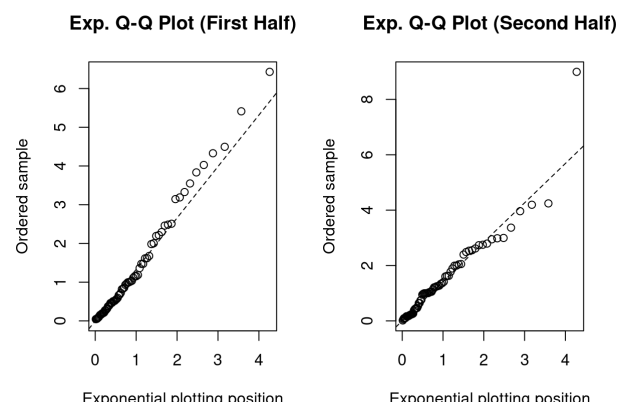
13. The ACF plot shows us only one significant spike near the beginning of the dataset, which could be attributed to an outlier. Overall, there is a lack of evidence against independence.

ACF of Interval (First Half)

ACF of Interval (Second Half)
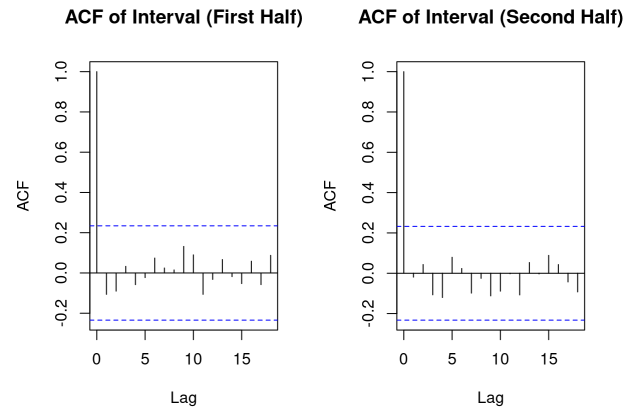
14. This project introduced a lot of different statistical and probabilistic ideas. I think having a more thorough introduction of the formulas and equations covered (such as t tests) would more solidly introduce me to these ideas. Overall, this project taught me a lot about probability and statistics.

Team Description

Members Isaac Boaz

Project Title Formula One Racing Data Analysis