Will Mora

Probability and Applied Statistics

Solving Statistical Problems with Formula One Datasets

**TEXTBOOK**

The textbook used in class that was used to obtain problems is *Mathematical Statistics with Applications – 7th Edition (Wackerly, Mendenhall, Schaeffer)*. The questions answered here are based on problems found at the end of the sections.

**DATASETS**

The dataset used to solve statistical problems consists of results from every Formula One race in the 2022 season. It features several different pieces of information, such as the tracks, driver names, driver car numbers, starting and finishing position, points earned, and fastest lap. The dataset can be found on GitHub at:

<https://github.com/toUpperCase78/formula1-datasets/blob/master/Formula1_2022season_raceResults.csv>.

A supplemental dataset that consists of all the teams that competed in the 2022 Formula One season and the specifications of their cars can be found at:

<https://github.com/toUpperCase78/formula1-datasets/blob/master/Formula1_2022season_teams.csv>.

**BACKGROUND**

Formula One is the highest class of racing for single-seater cars, as well as one of the most popular racing leagues in the world. Governed by the International Automobile Federation (FIA), the first season of Formula One took place in 1950 where there were seven races.

In 2022, there were a total of twenty-two races with twenty cars in every race. Each of the ten teams had two cars in every race. After every race, the top ten finishers are rewarded points with the winner receiving twenty-five points. The driver with the fastest lap of the race is given one bonus point only if they finished in the top ten.

**STATISTICAL PROBLEMS**

The questions below are broken down into categories and are based on the questions at the end of the sections of the textbook.

1. **PROBABILITY**

* Set Notation
* Sample Points and Sample Spaces
* Tools for Counting Sample Points
* Conditional Probability
* Two Laws of Probability
* Event Composition
* Bayes’ Theorem

1. **PROBABILITY DISTRIBUTIONS**

* Binomial Probability Distribution
* Geometric Probability Distribution
* Hypergeometric Probability Distribution
* Poisson Probability Distribution
* Chebyshev’s Theorem
* Uniform Probability Distribution

1. **MULTIVARIATE PROBABILITY DISTRIBUTIONS**

* Bivariate and Multivariate Probability Distributions
* Marginal and Conditional Probability Distributions

**STATISTICAL PROBLEMS**

**I. PROBABILITY**

**Set Notation**

Set notation is a part of set theory. The union of two elements, denoted by , is the combination of two sets. The resultant set would feature all of the elements from the previous sets with duplicates omitted. The intersection, written as , are all the elements that can be found in both sets with elements found in only one set omitted. A set with no elements is called an empty set. The complement of a set represents all of the elements not in itself, but are in the set of elements being considered, which is called the universal set. Elements can represent a number of things, such as numbers, items, or events.

Suppose the top two finishers of any given race can feature Ferraris or Red Bulls with Ferrari denoted as *F* and Red Bull denoted as *R* with a combination like *RF* meaning Red Bull finishing first with Ferrari second. The four points in the set *S* of possible observations are:

*S* = { *FF*, *FR*, *RF*, *RR*}

Let *A* denote the set with no Ferraris, *B* denote the set with two Ferraris, and *C* denote the set with at least one Ferrari. List the elements in: and .

**Sample Points and Sample Spaces**

A sample point is a possible outcome in an experiment. The set of all sample points is called the sample space.

A car in a race in finish in 1st through 5th, 6th through 10th, 11th through 15th, or 16th through 20th.

1. List the sample space for this experiment.
2. Assuming that all sample points are equally likely, find the probability that the car finishes in one of the first ten positions.
3. where

= 1st through 5th, = 6th through 10th, = 11th through 15th, and = 16th through 20th are the sample points

**Tools for Counting Sample Points**

A combination is a group of distinct objects taken at some constant rate. Unlike combinations, order does matter with permutations, which are also groups of distinct objects taken at some constant rate.

In Formula One, there are twenty individual cars, ten teams, and four power units. How many cars would be needed for there to be one of each combination of car-team-power unit?

If the selection of exactly one car from each team was desired (given two cars per team), how many different ways can the ten selected cars be ordered?

**Conditional Probability**

The conditional probability of an event happening depends on whether a separate event has occurred. If an event is unaffected by the occurrence of another event, the events are independent of each other. Otherwise, they are dependent.

At the 2022 Canadian Grand Prix, Red Bull driver Max Verstappen won the race, followed by Ferrari driver Carlos Sainz, and then followed by Mercedes driver Lewis Hamilton.

a. What is the probability that the other Ferrari driver, Charles Leclerc, finished in the fourth position?

b. What is the probability that McLaren driver, Lando Norris, finished in the fourth position?

c. What is the probability that both McLaren drivers finish in fourth and fifth?

a. Since there is one Ferrari driver left out of the remaining seventeen cars,

b. Since both McLaren drivers did not finish in the top three,

c. If one McLaren driver finishes in fourth and then the other McLaren driver finishes in fifth,

**Event Composition**

Two laws of probability that are important to event composition are the multiplicative and additive laws. They help show if two events are independent or dependent. The multiplicative law says that if the probability of the intersection of two events is the same as the product of their individual probabilities, the events are independent. The additive law states that if the probability of the union of two events is equal to the sum of their individual probabilities, they are mutually exclusive.

Max Verstappen won 68.2% of the races in the 2022 season. Suppose that the probability of him winning the next five races is also 68.2%. If the races are independent, what is the probability of him winning all five races?

**Bayes’ Theorem**

Bayes’ Theorem is one of the most important theorems in data science. It is the conditional probability of an event based on previous knowledge of the outcome of different events with similar conditions.

In a sample size of sixteen races, Charles Leclerc started in first nine times (56.25%) and Max Verstappen started in first seven times (43.75%). Leclerc won two races (9.09%) and Verstappen won six races (27.27%). A race is chosen at random. Find the conditional probability that this race was won by Max Verstappen.

Leclerc, Verstappen, Races

**II. PROBABILITY DISTRIBUTIONS**

**Binomial Probability Distribution**

In a binomial experiment, each trial results in one of two possible outcomes: success or failure. Every trial is independent and identical. The goal is to find the number of successes, or desired outcomes, in a given number of trials.

Of the 40 finishing positions for the Williams, 22 were 16th or worse (55%). Five finishes are chosen at random. Find the probability for each of the following events:

a. All five positions are 16th or worse

b. At least four positions are 16th or worse

a.

b.

**Geometric Probability Distribution**

Similar to a binomial probability distribution, a geometric probability distribution has independent and identical trials that can result in one of the two outcomes of success or failure. Instead of finding the number of successes, though, the goal is to find when the first success occurs. So, unlike binomial probability distributions, the experiment can end at any number of trials.

36.36% of the drivers at the 2022 Monaco Grand Prix had a faster lap than 1:17.00. If drivers are chosen at random, find the probability that the first driver with a faster lap than 1:17.00 is the fifth driver chosen.

**Hypergeometric Probability Distribution**

The hypergeometric probability distribution is used when finding the number of successes in a given number of trials without replacement. This is different from the binomial probability distribution which has replacement.

A garage holds five cars, of which two are Aston Martins, two are Alfa Romeos, and one is a AlphaTauri. Two are chosen without replacement. What is the probability that both cars will be Aston Martins?

**Poisson Probability Distribution**

The Poisson probability distribution is used to model count data. It is used when trying to determine the probability of the number of successes, or occurrences, on a per-unit basis such as per-unit time, per-unit area, per-unit volume, and so on. This distribution is suitable when analyzing situations where the number of trials is very large and the probability of success is very small.

During the 2022 British Grand Prix, cars cross the start-finish line according to a Poisson distribution at an average of four per second. During a given second, what are the probabilities that

a. no more than two cars cross the start-finish line?

b. at least one car crosses the start finish line?

a.

b.

**Chebyshev’s Theorem**

Chebyshev’s Theorem describes the minimum number of data points that must be within a certain number of standard deviations from the mean. This theorem can be used with probability distributions and all data sets.

Sebastien Vettel’s fastest lap during the 2022 United States Grand Prix was 1:41.5. His speed, on average, is seconds per lap. Let’s say he wants his lap times Y to be within .5 seconds of at least 75% of the time. What is the largest value of , the standard deviation of *Y*, that can be tolerated if his objective is to be met?

**Uniform Probability Distribution**

In a uniform probability distribution, all outcomes are equally likely. Experiments with this kind of distribution include a card being drawn from a deck of cards, a marble being drawn from a bag, or the flipping of a coin.

Lewis Hamilton’s fastest lap during the 2022 Austrian Grand Prix was 1:09.0. Suppose his speeds were uniformly distributed between 1:07.0 and 1:11.0. Find the probability that his next lap is

a. lower than 1:08.0

b. higher than 1:10.5

a.

b.

**III. MULTIVARIATE PROBABILITY DISTRIBUTIONS**

**Bivariate and Multivariate Probability Distributions**

A bivariate probability distribution is a type of joint distribution that has two independent variables. It shows the probability of the occurrence of the two variables. Similarly, multivariate probability distributions have two or more independent variables.

Of the ten Formula One teams, six are based in the United Kingdom, two are based in Italy, and two are based elsewhere (Switzerland and the United States). Two teams are to be selected. Let *X* denote the number of teams based in the United Kingdom and *Y* denote the number of teams based in Italy. Find the joint probability function of *X* and *Y*.

**Marginal and Conditional Probability Distributions**

A marginal probability distribution is the distribution and values of one of the variables while ignoring the other variables. In other words, it is the probability of an event to occur, independent of others. In retrospect, a conditional probability distribution takes into account the outcomes of other events when calculating the likelihood of another event.

From the joint probability function above, find the marginal probability distribution of *X*, the number of teams based in the United Kingdom among the three selected.