Homework 2

Statistical Inference, Spring 1400

1- True/False Problems. Circle either True or False (and explain/correct if False):

a. If a fair coin is tossed many times and the last eight tosses are all heads, then the chance that the next toss will be heads is somewhat less than 50%.

b. If events X and Y are independent, then they are also mutually exclusive.
c. The General Addition Rules states that the probability of either event X or event Y occurring is the sum of P(X) with P(Y).
d. P(A | B) = P(B | A).
e. A probability tree is one way to represent a sample space.
f. If P(A and B) = 0, then either P(A) or P(B) must also be equal to zero.
g. The mean of a Poisson distribution is not always equal to its variance.
h. If the binomial probability of success is near 0.5, then the distribution is approximately symmetric.
i. The probability of a student randomly guessing answers to a true/false exam is best modeled with a binomial distribution.

2- In France, Ministry of Health reports the daily number of covid-19 cases. The 2020 France final report indicates that 14.6% of French people are infected to covid-19, 20.7% speak a language other than French at home, and 4.2% fall into both categories.

a. Are the rate of infected people to covid-19 and people who speak a language other than French at home disjoint?
b. What percentage of French people are infected to covid-19 and only speak French at home?
c. What percentage of French people are infected to covid-19 or speak a language other than French at home?
d. What percentage of French people are not infected to covid-19 and only speak French at home?
e. Is the event that someone gets infected independent of the event that the person speaks a language other than French at home?

3- There are 64 teams in a football league. These teams compete in pairs in the first round, and the winners go to the second round, and again, they compete in pairs. In the same way, these competitions will continue until there is only one team left. Considering the 64 teams in the league, the games of this tournament will take place within 6 rounds. Someone has taken part in a challenge to predict the outcome of this league. He must present his prediction for all of the games at the beginning of the season and, at the end of the season, he earns points based on the predictions he made at the beginning of that season. The rule of challenge is that for every correct prediction in the first round, he gets 1 point. In each subsequent rounds of the tournament, the score is doubled from the previous round (if you correctly predict a match in the second you get 2

points and so on). This person has no information about these teams, so he decides to predict all the games by tossing a fair coin. Calculate the expected value, E(X) for how much he will earn ultimately.

4- To battle against spam emails, Bob installs two anti-spam programs. An email arrives, which is either legitimate (event L) or spam (event L^C), and which program j marks as legitimate (event M_j) or marks as spam (event M_j^C) for j ∈ {1,2}. Assume that 10% of Bob's email is legitimate and that the two programs are each "90% accurate" in the sense that $P(M_j|L) = P(M_j^C|L^C) = 9/10$. Also, assume that if an email is spam, the outputs of these two programs are conditionally independent.

   a. Find the probability that the email is legitimate, given that the 1st program marks it as legitimate.
   b. Find the probability that the email is legitimate, given that both programs mark it as legitimate.
   c. Bob runs the first program and M_1 occurs. He updates his probabilities and then runs the 2nd program. Let $\tilde{p}(A) = P(A|M_1)$ be the updated probability function after running the first program. Explain briefly in words whether or not $\tilde{p}(L|M_2) = P(L|M_1 \cap M_2)$: is conditioning on $M_1 \cap M_2$ in one step equivalent to first conditioning on M_1, then updating probabilities, and then conditioning on M_2?

5- One pharmaceutical company makes cancer pills in boxes of 100. If QC (quality control) says that 0.5% of the cancer pills are damaged, then what percent of the boxes will:

   a. Have no damaged pills?
   b. Have 2 or more damaged pills?
   c. **(R)** Do the calculations in the form of an R script.

6- A travel agency knows that over the long run, 90% of passengers who reserve seats will show up for their trip. On a particular trip with 300 seats, the travel agency accepts 324 reservations. Use the normal approximation to answer the following questions:

   a. What is the chance that the trip will be overbooked? The event of a passenger showing up is independent of one another.
   b. Regarding the previous question, now assuming that passengers always trip in pairs "and trip only if both people in the pair show up". Check whether your answer to the previous question is consistent with this result.

Homework 2

Statistical Inference, Spring 1400

7- **(R)** A student who took the statistical inference course scored 620 on the exams and 670 on the projects. The mean score for the exam section was 462 with a standard deviation of 119, and the mean score for the project section was 584 with a standard deviation of 151. Suppose that both distributions are nearly normal.

a. Write down the shorthand for both of these distributions
b. What is Z score on the exam section and project section?
c. Draw a standard normal distribution curve and mark these two Z scores.
d. What do these Z scores tell you?
e. Relative to others, in which sections did she perform better?
f. Find her percentile scores for the two sections (exam section, project section)
g. What percentage of the students performed better in the exams than her?
h. What percentage of the students performed better in the projects than her?
i. Below are listed the final exam scores of 20 introductory statistics students.
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94
Do these data appear to follow a normal distribution? Use histogram, QQ-plot and boxplot diagrams in order to justify your answer.

8- N guests gave their raincoats and their umbrellas to the doorman at the entrance of the "Marlinspike" mansion in order to attend the party. At the end of the night, the doorman was completely plastered. So, he gave the leaving guests a random raincoat and a random umbrella, in a way that each person got a pair of raincoat and umbrella in a uniformly random manner.

a. What is the probability that nobody gets back his own raincoat and his umbrella?
b. What is the probability that everybody gets at least his own raincoat or his own umbrella? Calculate this probability when N goes to infinity.
c. If each guest can find his/her own raincoat with probability p=0.2 and find his/her umbrella with probability p=0.1, then what is the probability that N/2 persons neither get their raincoats nor their umbrella right, given that the other N/2 persons have already received their belongings correctly. (N is divisible by 2).

9- Suppose you desperately need to access Telegram and the VPN service you were buying subscriptions from has been blocked entirely. You now have to choose from one of the two free VPN services that are still working. But, as you may know, the free VPN services are very slow and do not necessarily connect whenever you request a connection. The first service connects successfully 10% of the time, and the other one connects 30% of the time. You want to use the service that has more chance of establishing a VPN connection. However, you don't know which one is better. So, initially, you assume that both VPN services are equally likely to be the better ones. Then:

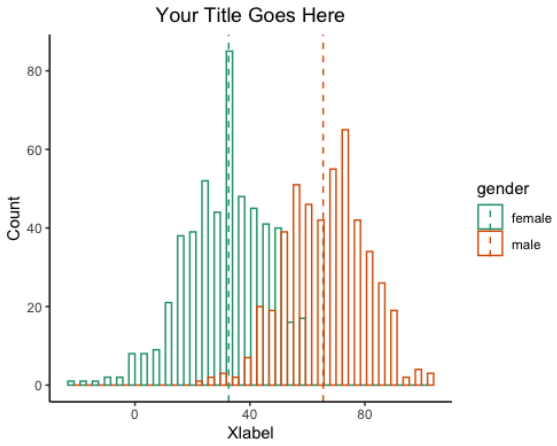
a. You randomly try one of the services. Given that you managed to establish a connection, what is the probability that the VPN service you've tried was the better one?
b. **(R)** You start randomly trying these two services and after each trial, you update your belief about the better service using the Bayesian rule. How many trials on average (among 100 experiments) does it take to be 90% certain about the better service?

Homework 2

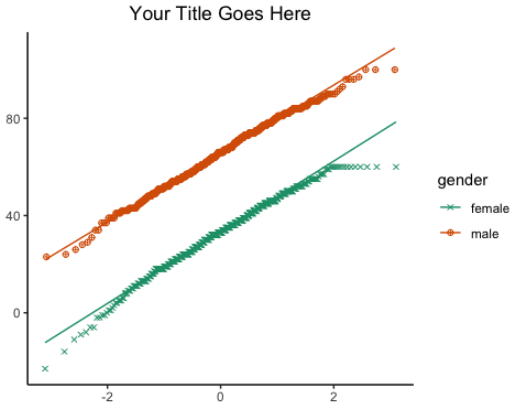
Statistical Inference, Spring 1400

10- (R, ggplot) In this question, you are going to use the "Students Performance" dataset. This dataset comprises of some details about 1,000 students as well as their scores in mathematics, reading, and writing exams. ***Note that you must use ggplot library in order to draw the diagrams.***

a. Load the `StudentsPerformance.csv` dataset. Draw the histogram of "writing_score" variable for each gender. Your output must be similar to the following image. Please pay attention to all of the details you can see in the image and choose a bin size that is right for the data.



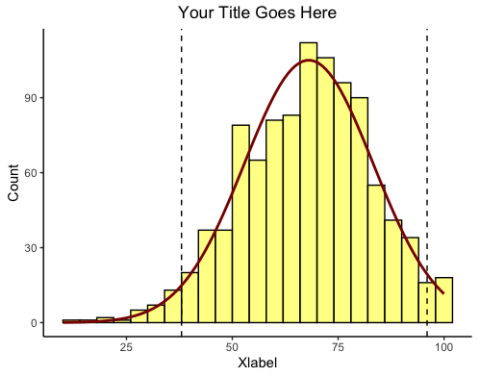
b. Inspect the distributions of "reading_score", "writing_score", and "math_score" by drawing their QQ-Plot for each gender. Your plots must look like the following image. Can we sensibly assume that these variables are each coming from a gaussian distribution?



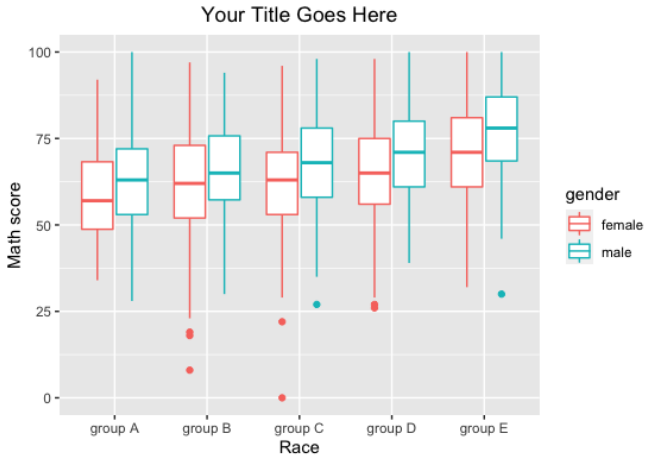
c. Assuming that the probability distributions of these three variables are gaussian, find the maximum likelihood estimation of these distributions. Then, draw the probability density function of each variable on top of its histogram. You also have to mark the 2.5% and the 97.5% percentiles on the diagram. Your result for each variable must be similar to the following diagram.

Homework 2

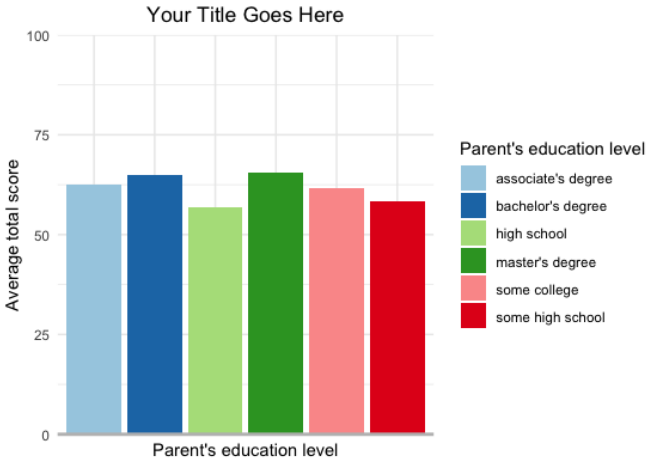
Statistical Inference, Spring 1400



d. Draw separate boxplots of "math_score" variable for students of different races and genders. Your diagram must look like the following image.



e. Calculate the total score (average of math, reading, and writing scores) of the individual students and store them in a new column in the dataset. Then, draw a barplot such that every bar represents the average of the total score of the students whose parents have the same level of education. Your final result must look like this:



Homework 2

Statistical Inference, Spring 1400

11- **(Bonus)** In one of the ancient tribes of the Maya civilization, there was a popular game called “Buga-Uga”. In this game, the player picks a knife and stands beside a special table designed for this game. The player then tosses the knife to the air in a way that the knife would equally likely land anywhere on the table but it never lands on the edges (for simplicity, you can assume that the table is very large). Whenever the knife lands on the table, it leaves a trace with the shape of a line of length K. The surface of the table is made of wooden strips of two types. Below, you can see a small portion of the table from the top. The width of each wooden strip is W (W > K).

If the player tosses the knife and it lands in such a way that the trace is left only in a single strip, the player wins. What is the probability that the player wins three successive rounds?