Problem Set 2

Statistics 104

Due February 21, 2020 at 11:59 pm

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, be sure to include the code in your solution.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Problem 1.

Suppose that there are 100 students entering the Master's of Business Administration (MBA) degree program at Harvard Business School (HBS). Of these students, 20 have two years of work experience, 30 have three years of work experience, 15 have four years of work experience, and 35 have five or more years of work experience.

a) One of the students is selected at random. What is the probability that this student has at least three years of work experience?

b) The selected student has at least three years of work experience. What is the probability the student has four years of work experience?

c) Consider the population of students who receive an admissions offer from the HBS MBA program (in a particular year). Three students are selected at random from the students accepted to the HBS MBA program. Calculate the probability that all three students have five or more years of work experience. Describe a key assumption required to make the calculation and comment on whether the assumption is reasonable.

d) Would it be reasonable to use the probability calculated in part a) as an estimate of the proportion of students entering the MBA degree program at MIT (Sloan Graduate School of Management) who have at least three years of work experience? Explain your answer. Limit your explanation to at most five sentences.

Problem 2.

A breath analyzer is used by the police to estimate blood alcohol content (BAC) from a breath sample. If a person's BAC is above (or equal to) the legal limit, they can be arrested for suspicion of driving while impaired from alcohol. However, the only way to accurately measure BAC level is through taking a blood sample; a number of factors are known to influence the results of breath analyzers, such as hypoglycemia.

A particular brand of breath analyzer is accurate about 80% of the time. That is, if an individual actually has BAC equal to or above the legal limit, the device indicates a positive result with probability 0.80, and if an individual actually has BAC below the legal limit, the device indicates a negative result with probability 0.80.

a) Suppose that a police officer tests 5 drivers who are *correctly* suspected of driving under the influence (i.e., driving with BAC equal to or above the legal limit).

   i. Assume independence between the drivers' results and use algebraic methods to find the probability that at least two of the drivers will correctly test positive.

   ii. Verify your answer to part i. via simulation, using a `for( )` loop to replicate the five tests 100,000 tests.

b) Suppose that on any particular Saturday night, about 5% of drivers are known to be driving under the influence.

   i. Calculate the probability that a driver who tests positive actually has BAC level equal to or above the legal limit.

   ii. How accurate would the device need to be for the probability in part i. to be 0.80?

   iii. In language accessible to someone who has not taken a statistics course, explain why the probability in part i. is much lower than the accuracy of the breath analyzer. Limit your answer to at most five sentences.

Problem 3.

Mensa is a high IQ society that admits people as members if they can score at the 98^{th} percentile or above on certain standardized IQ (intelligence quotient) tests. On one such test, the Stanford Binet, the qualifying score is 132. The test consists of n questions, each with m choices.

a) On any given test question, the person taking the test knows the answer with probability p. Assume that when the person does not know the answer, the person guesses an option completely at random. Calculate the probability a person knew the answer to a question, given that they answered it correctly.

b) If a person receives a score of 132 or higher on the test, they are considered to have an IQ of 132 or higher. However, individuals with IQ less than 132 can also receive such scores about 0.1% of the time due to lucky guessing. Given that a person is labeled as having IQ of 132 or higher, what is the probability they actually have IQ below 132? Assume that all individuals with IQ of 132 or higher receive an accurate score 95% of the time.

Problem 4.

A deck of 100 cards is numbered from 1 through 100. The deck is shuffled and three cards are drawn without replacement. Let X be the value of the first card dealt, Y be the value of the second card dealt, and Z be the value of the third card dealt.

Calculate $P(X < Y < Z)$. Explain your reasoning.

Problem 5.

Suppose that two fair six-sided dice are rolled. Run the following code to simulate the results for 100 sets of (two) dice rolls. The simulation is written to estimate the probability that at least one of the two dice rolls was a 1, given that the sum of the rolls equals 5.

```
#define parameters
number.rolls = 2
replicates = 100

#create empty vector to store results
successes.5 = vector("numeric", replicates)
successes.1 = vector("numeric", replicates)

#set seed
set.seed(2020)

#simulate the draws
for(k in 1:replicates){

  rolls = sample(1:6, number.rolls, replace = TRUE)

  if(sum(rolls) == 5){

    successes.5[k] = 1

    if(rolls[1] == 1 | rolls[2] == 1){

      successes.1[k] = 1
    }
  }
}
```

a) Explain the purpose of `set.seed( )`.

b) Explain the code used to generate `rolls`.

c) Examine the structure of the nested `if` statements.

   i. Explain the condition necessary for a 1 to be recorded in the `successes.5` vector.

   ii. Explain the conditions necessary for a 1 to be recorded in the `succesess.1` vector.

d) From the simulation results, estimate the probability that at least one of the two dice rolls was a 1, given that the sum of the rolls equals 5.

e) Use algebraic methods to calculate the probability from part d).

f) Explain the apparent discrepancy between the results from part d) versus part e).

Problem 6.

A standard 52-card deck of French playing cards consists of four suits: hearts, spades, clubs, and diamonds. There are 13 cards of each suit; each suit has cards of rank 2 through 10, along with an ace, king, queen, and jack. Typically, hearts and diamonds are the red suits, while spades and clubs are the black suits.

Four cards are drawn from the deck, one at a time, without replacement.

a) The second card drawn is from a red suit. Based on this information, what is the probability it is a heart?

b) Calculate the probability of drawing exactly one heart (out of the four cards).

c) Using simulation, estimate the probability of drawing exactly one card of each suit. Be sure to clearly comment your code.

Problem 7.

Twins can be either "identical" or "fraternal". About 30% of human twins are identical. Identical twins develop from the same fertilized egg that split, and so are necessarily the same sex; half of identical twin pairs are males and half are females. Of fraternal twin pairs, half are male-female twins, one quarter are female-female twins, and one quarter are male-male twins.

A mother has given birth to a pair of twin girls. What is the probability they are identical twins?

Problem 8.

A genetic test is used to determine if people have a predisposition for thrombosis, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people have this predisposition. The test is 95% accurate for those who have the predisposition, and 97% accurate for those who do not have the predisposition.

Simulate the results for administering this test to a population of 100,000 individuals.

a) How many individuals in this hypothetical population are expected to test positive for the predisposition?

b) Estimate the probability that an individual who tests positive has the predisposition.

c) Suppose that two new tests have been developed. Test A improves accuracy for those who have the predisposition to 98% (while accuracy for those who do not have the predisposition remains at 97%), while Test B improves accuracy for those who do not have the predisposition to 99% (while accuracy for those who do have the predisposition remains at 95%).

   Which test offers a higher increase in the probability that a person who tests positive actually has the predisposition? Explain the reasoning behind your answer. Limit your answer to at most seven sentences.