Module 2 Homework

Cody S

This assignment will be reviewed by peers based upon a given rubric. Make sure to keep your answers clear and concise while demonstrating an understanding of the material. Be sure to give all requested information in markdown cells. It is recommended to utilize Latex.

Problem 1

What does it mean for one event C to cause another event E — for example, smoking (C) to cause cancer (E)? There is a long history in philosophy, statistics, and the sciences of trying to clearly analyze the concept of a cause. One tradition says that causes raise the probability of their effects; we may write this symbolically is

$$P(E|C) > P(E). \qquad (1)$$

Part a)

Does equation (1) imply that $P(C|E) > P(C)$? If so, prove it. If not, give a counter example.

Answer

The probability of an event given a cause being higher than the probability of the event occuring at all does imply that the probability of the cause is also higher given the event.

Given that all probabilities P must be $0 \leq P \leq 1$, Baye's theorem

$P(E|C) = \frac{P(C|E) \cdot P(E)}{P(C)}$

Can be rearranged to

$\frac{P(E|C)}{P(E)} = \frac{P(C|E)}{P(C)}$

and given that

$P(E|C) > P(E)$

Then the left side of the rearranged equation must be

$\frac{P(E|C)}{P(E)} > 1$

Since the two side are equal then the right side must also be

$\frac{P(C|E)}{P(C)} > 1$

Meaning that

$P(C|E) > P(C)$

In some cases this makes intuitive sense in that the equations don't consider in what order P and C occurred, but only that they are observed. Observing that somebody smokes means that they are more likely to have cancer, meaning that the population of people with cancer will contain a higher proportion of smokers than the entire population.

Part b)

Another way to formulate a probabilistic theory of causation is to say that

$$P(E|C) > P(E|C^C). \qquad (2)$$

Show that equation (1)

$$P(E|C) > P(E). \qquad (1)$$

implies equation (2).

Answer

If the probability of an event given a causative condition is higher than the total probability of an event (Equation 1) then the probability of the event occuring given that condition is necessarily higher than the probability of the event given the condition does not occur (Equation 2). This can be seen from the law of total probability.

$$P(E) = P(E|C) \cdot P(C) + P(E|C^C) \cdot P(C^C)$$

If $P(E|C) > P(E)$ then the $P(E|C)$ term sets the highest possible value for $P(E)$ across all values of $P(C)$, and $P(C^C)$ sets the lowest possible value for $P(E)$. Since $P(E)$ is the average of $P(E|C)$ and $P(E|C^C)$ (weighted by the given value of $P(C)$) then $P(E|C^C)$ must be lower than $P(E|C)$ to "pull" $P(E)$ down from the upper bound of $P(E|C)$.

Part c)

Let C be the drop in the level of mercury in a barometer and let E be a storm. Briefly describe why this leads to a problem with using equation (1) (or equation (2)) as a theory of causation.

Answer

Neither equation 1 nor 2 account for the possibility of a third condition that influences E and C while C and E don't influence each other. If C were strictly causal for E then it would be possible to induce a storm by artificially lowering the mercury in the barometer.

Part d)

Let A, C, and E be events. If $P(E|A \cap C) = P(E|C)$, then C is said to screen A off from E. Suppose that $P(E \cap C) > 0$. Show that screening off is equivalent to saying that $P(A \cap E|C) = P(A|C)P(E|C)$. What does this latter equation say in terms of independence?

Answer

$P(E|A \cap C) = P(E|C)$ means that the occurence of A doesn't change the probability of E given C, knowing if A occurred or not provides no additional information about the probability of E if C has occurred. It can be rewritten as

$$\frac{P(E \cap A \cap C)}{P(A \cap C)} = P(E|C)$$

which is equivalent to

$$\frac{P(A \cap E|C)P(C)}{P(A|C)P(C)} = P(E|C)$$

$P(C)$ in the numerator and denominator cancel out and the remaining terms can be rearranged to

$$P(A \cap E|C) = P(A|C)P(E|C)$$

This final equation shows that the probability of A and E together given C is the same as the combined probabilities of A given C and E given C, meaning that A and E don't affect each other and are independent, but are both affected by C.

Part e)

Now let A be a the drop in the level of mercury in a barometer, E be a storm, and C be a drop in atmospheric pressure. Does the result from part (d) help fix the problem suggested in part (c)?

Answer

Yes, this shows how the drop in the barometer can provide information about the probability of a storm occurring without being the cause of the storm by introducing the third condition (the drop in atmospheric pressure) that affects the barometer and the weather similarly.

Problem 2

Suppose you have two bags of marbles that are in a box. Bag 1 contains 7 white marbles, 6 black marbles, and 3 gold marbles. Bag 2 contains 4 white marbles, 5 black marbles, and 15 gold marbles. The probability of grabbing the Bag 1 from the box is twice the probability of grabbing the Bag 2.

If you close your eyes, grab a bag from the box, and then grab a marble from that bag, what is the probability that it is gold?

Part a)

Solve this problem by hand. This should give us a theoretical value for pulling a gold marble.

Answer

- Bag 1 - 7 white, 6 black, 3 gold = 16 total
- Bag 2 - 4 white, 5 black, 15 gold = 24 total
- P(Bag1) = 2/3
- P(Bag2) = 1/3

$$P(\text{Gold}) = (P(\text{Bag1})P(\text{Gold1})) + (P(\text{Bag2})P(\text{Gold2}))$$

$$P(\text{Gold}) = \left(\frac{2}{3}\right)\left(\frac{3}{16}\right) + \left(\frac{1}{3}\right)\left(\frac{15}{24}\right) = \frac{1}{3}$$

Part b)

Create a simulation to estimate the probability of pulling a gold marble. Assume you put the marble back in the bag each time you pull one out. Make sure to run the simulation enough times to be confident in your final result.

Note: To generate n random values between [0,1], use the `runif(n)` function. This function generates n random variables from the Uniform(0,1) distribution, which we will learn more about later in this course!

```r
# set number of simulations
sims = 1000000
# create vector to hold results of gold marble pulls
gold = rep(0, sims)

for (i in 1:sims) {
    # simulate choose one of the bags
    bag_choice = runif(1)

    # define and simulate gold pull probability for bag 2
    if (bag_choice <= 1/3) {
        pull_bag_2 = runif(1)
        if (pull_bag_2 <= 15/24) {
            gold[i] = 1
        }
    # define and simulate gold pull probability for bag 1
    } else {
        pull_bag_1 = runif(1)
        if (pull_bag_1 <= 3/16) {
            gold[i] = 1
        }
    }
}

result <- mean(gold)
result
```

0.332855

```r
expected_value <- 1 / 3
result_error <- abs(1 - (result/expected_value)) * 100

cat(sprintf("The average of %d simulations is %.4f which is
\nwithin %.3f%% of the expected result of 1/3 ", sims, result, result_error))
```

```
The average of 1000000 simulations is 0.3329 which is

within 0.143% of the expected result of 1/3
```

Problem 3

Suppose you roll a fair die two times. Let A be the event "the sum of the throws equals 5" and B be the event "at least one of the throws is a 4".

Part a)

By hand, solve for the probability that the sum of the throws equals 5, given that at least one of the throws is a 4. That is, solve $P(A|B)$.

Answer

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

$P(A) = \frac{1}{9}$ (4 ways to roll a sum 5 out of 36 possible rolls)

$P(B) = 1 - (\frac{5}{6})^2 = 11/36$

$P(B|A) = \frac{1}{2}$ (4 ways to roll a sum of 5, 2 of which contain a 4)

$P(A \cap B) = P(A) \cdot P(B|A)$

$P(A \cap B) = \frac{1}{9} \cdot \frac{1}{2} = \frac{1}{18}$

$$P(A|B) = \frac{\frac{1}{18}}{\frac{11}{36}} = \frac{2}{11} \approx 0.1818$$

Part b)

Write a simple simulation to confirm our result. Make sure you run your simulation enough times to be confident in your result.

Hint: Think about the definition of conditional probability.

```
# set number of simulations
sims <- 1000000

# create vector to count the number of rolls that contain a 4
contains_4 <- rep(0, sims)

# create a vector to count the number of rolls that contain a 4 and sum to 5
sum_5_given_4 <- rep(0,sims)

# create a dice numbered 1 through 6
x <- c(1:6)

for (i in 1:sims) {
  # roll the dice twice
  rolls = sample(x, 2, replace = TRUE)

    # if the roll contains a 4, count it in the contain_4 vector
    if (4 %in% rolls){
      contains_4[i] = 1
    }

    # if there is a 4 and the sum is 5, count it in the sum vector
    if ((4 %in% rolls) && sum(rolls) == 5){
      sum_5_given_4[i] = 1
    }
}

# determine the proportion of rolls that contain a 4 that summed to 5
result <- sum(sum_5_given_4) / sum(contains_4)
result
```

```
0.18261966386004
```

```
expected_value <- 2 / 11
result_error <- abs(1 - (result/expected_value)) * 100

cat(sprintf("The average of %d simulations is %.4f which
 \nis within %.3f%% of the expected result of 2/11.", sims, result, result_error))
```

```
The average of 1000000 simulations is 0.1826 which

is within 0.441% of the expected result of 2/11.
```