Problem set 1: Probability (solutions)

Due October 7, 2024, at 10am

(Your name here)

NOTE: Start with the file `ps1_2024_probability.qmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F24/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the* `Render` *button). Submit both the qmd file and the knitted PDF via Canvas.*

Problem 1: joint probability, conditional probability, independence

Let A and B be two events that could result from a random process.

Suppose the joint probabilities are:

Result	Probability
$A \cap B$	1/4
$A^C \cap B$	1/12
$A \cap B^C$	1/6
$A^C \cap B^C$	1/2

(1a) What are $P(A)$ and $P(B)$, i.e. the marginal probabilities of A and B?

Answer:

By the law of total probability, $P(A)$ is $P(A \cap B) + P(A \cap B^C) = 1/4 + 1/6 = 3/12 + 2/12 = 5/12$.

$P(B)$ is $P(A \cap B) + P(A^C \cap B) = 1/4 + 1/12 = 3/12 + 1/12 = 4/12 = 1/3$.

(1b) What is $P(A|B)$, i.e. the conditional probability of A given B?

Answer:

By definition, $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{1/3} = 3/4$.

(1c) Suppose the random process is the selection of a student at random from a school. Considering the joint probability table above, what might events A and B be? Consider, for example, "the student eats carrots in his/her lunch", "the student can read", "the student has at least one sibling", "the student plays with Lego every day", "the student can tie his/her shoes", "the student wears a hat to school". You may want to specify what kind of school you have in mind.

Answer:

Inspecting the joint probability table, we note that A and B usually occur together – i.e. $P(A|B) > P(A)$ and $P(B|A) > P(B)$. So we need events A and B to be things that would tend to either both be true or both be not true of students at the school. We also need A to be true of about half of the students and B to be true of about a third. From the list given, "the student can read" "the student can tie his/her shoes" would both tend to be true of older kids and not true of younger kids. Depending on the age and the definition of "can read", it could be the case that more kids can read than can tie their shoes. So A could be reading and B could be shoe tying. If this is a school for children of ages 4-8 or so, the marginal probabilities could be roughly correct.

(1d) Suppose events A and B had the same marginal probabilities you reported in (1a), but the events were independent. What then would be the joint probabilities? Fill out the table below. Make sure the probabilities add up to 1.

Answer:

Result	Probability
$A \cap B$	$= P(A)P(B) = 5/12 \times 1/3 = 5/36$
$A^C \cap B$	$= (1 - P(A))P(B) = 7/12 \times 1/3 = 7/36$
$A \cap B^C$	$= P(A)(1 - P(B)) = 5/12 \times 2/3 = 10/36$
$A^C \cap B^C$	$= (1 - P(A))(1 - P(B)) = 7/12 \times 2/3 = 14/36$

Problem 2: `R` coding

First set your seed to 123 so that our answers are comparable.

```
set.seed(123)
```

(The seed determines the output of "random" processes in R, so that if two students use the same function after setting the same seed, they will get the same answer. Set the seed once at the beginning of your code; do not set it again later in the code.)

(2a) Create a vector of length 1000 that could be a sample from the marginal distribution of event A in the previous question, where "A" indicates that event A occurred and "!A" indicates that it did not occur. Store this vector in a variable called `A_vec`. Report the proportion of times that event A occurred in this sample.

Answer:

```
A_vec <- sample(x = c("A","!A"), size = 1000, replace = T, prob = c(5/12, 7/12))
mean(A_vec == "A") # one way to get the proportion of "A"
```

```
[1] 0.412
```

```
length(A_vec[A_vec == "A"])/length(A_vec) # another way
```

```
[1] 0.412
```

```
sum(A_vec == "A")/length(A_vec) # another way
```

```
[1] 0.412
```

The proportion should be approximately $5/12 = .41\overline{6}$, and it is indeed close. It is not exactly $5/12$ because of randomness.

(2b) Do the same for event B. That is, create a vector of length 1000 that could be a sample from the marginal distribution of event B in the previous question, where "B" indicates that event B occurred and "!B" indicates that it did not occur. Store this vector in a variable called `B_vec`. Report the proportion of times that event B occurred in this sample.

Answer:

```
B_vec <- sample(x = c("B", "!B"), size = 1000, replace = T, prob = c(1/3, 2/3))
mean(B_vec == "B") # one way to get the proportion
```

```
[1] 0.318
```

```
length(B_vec[B_vec == "B"])/length(B_vec) # another way
```

```
[1] 0.318
```

```
sum(B_vec == "B")/length(B_vec) # another way
```

```
[1] 0.318
```

The proportion should be approximately 1/3, and it is indeed close.

(2c) Given how `A_vec` and `B_vec` are created, A and B are independent events. This implies that $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, $\mathrm{P}(B \mid A) = \mathrm{P}(B)$, and $\mathrm{P}(A \cap B) = \mathrm{P}(A) \times P(B)$. Confirm that this is approximately true in the samples `A_vec` and `B_vec`.

Answer:

```
#One way:
# Get the elements of A_vec where B occurred
A_vec2 <- A_vec[B_vec == "B"]
# and then compute the proportion of those where A occurred
mean(A_vec2 == "A")
```

```
[1] 0.3930818
```

```
# Another way
mean(A_vec[B_vec == "B"] == "A") # all in one line, same thing
```

```
[1] 0.3930818
```

```
# Could also output a table and compute from that
(tab <- table(A_vec, B_vec))
```

```
     B_vec
A_vec  !B   B
   !A 395 193
   A  287 125
```

```
tab["A", "B"]/sum(tab[, "B"])
```

```
[1] 0.3930818
```

```r
# should be about 5/12 = .41666667

# Same thing for B:
mean(B_vec[A_vec == "A"] == "B")
```

```
[1] 0.3033981
```

```r
# should be about 1/3
```

(2d) Repeat (2a)-(2c) where `A_vec` and `B_vec` have length 1 million. You should find that the equivalencies in (2c) are more nearly true in these larger samples, which we will later investigate in the law of large numbers. Is it the case here?

Answer:

```r
A_vec <- sample(x = c("A","!A"), size = 1000000, replace = T, prob = c(5/12, 7/12))
B_vec <- sample(x = c("B","!B"), size = 1000000, replace = T, prob = c(1/3, 2/3))
# this
mean(A_vec[B_vec == "B"] == "A")
```

```
[1] 0.4162201
```

```r
# should be even closer to 5/12

# this
mean(B_vec[A_vec == "A"] == "B")
```

```
[1] 0.3334366
```

```r
# should be even closer to 1/3
```

Indeed, we find they are closer.