Problem set 1: Probability

Due October 2, 2023, at 9pm

(Your name here)

NOTE: Start with the file `ps1_2023_probability.qmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F23/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the qmd file and the knitted PDF via Canvas.*

Problem 1: joint probability, conditional probability, independence

Suppose a speech is chosen at random from a collection of politicians' speeches. Let A denote the event that the speech contains the word "congratulate". Let B denote the event that the speech contains the word "shameful". Let $P(A)$ denote the (marginal) probability of event A and $P(B)$ denote the (marginal) probability of event B .

(1a) Suppose events A and B are independent. Fill out the joint probability table below using $P(A)$ and $P(B)$.

Result	Probability
$A \cap B$	$P(A)P(B)$
$A^C \cap B$	$(1 - P(A))P(B)$
$A \cap B^C$	$P(A)(1 - P(B))$
$A^C \cap B^C$	$(1 - P(A))(1 - P(B))$

(1b) Do you think events A and B really would be independent, given a collection of speeches from e.g. the US Congress or the British House of Commons? Say why or why not, perhaps with reference to the definition of independence.

Answer: They probably are not independent. A speech in which "shameful" appears is less likely to have language praising someone, like "congratulate". Of course, it's possible that politicians who congratulate someone like to also condemn their opponents (e.g. for not appreciating that person) in the same speech.

Now suppose the true joint probabilities are:

Result	Probability
$A \cap B$	1/8
$A^C \cap B$	1/4
$A \cap B^C$	7/24
$A^C \cap B^C$	1/3

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

Answer:

By the law of total probability, we have

$P(A) = P(A \cap B) + P(A \cap B^C) = 1/8 + 7/24 = 10/24 = 5/12$

$P(B) = P(A \cap B) + P(A^C \cap B) = 1/8 + 1/4 = 3/8.$

(1d) What is $P(A|B)$, i.e. the conditional probability of A given B? What would it be if the events were independent?

Answer: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{1/8+1/4} = \frac{1}{3}$. If the events were independent, then $P(A|B) = P(A) = 5/12$.

(1e) How much more or less likely is a speech to contain both "congratulate" and "shameful" than would be the case if the events were independent?

Answer: If they were independent, then $P(A \cap B) = 5/12 \times 3/8 = 5/32$. This is $1/32$ larger than the actual $P(A \cap B)$, which is $1/8 = 4/32$.

Problem 2: `R` coding

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

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

(2a) Create a vector of length 1000 that resembles a sample of speeches in Problem 1, where 1 indicates that the word "congratulate" appears in the speech and 0 indicates that it does not. Use the marginal probability from problem (1c). Store this vector in a variable called `A_vec`. Report `mean(A)`.

Answer:

```
A_vec <- sample(x = c(0, 1), size = 1000, prob = c(1 - 5/12, 5/12), replace = T)
mean(A_vec)
```

```
[1] 0.412
```

```r
# An alternative using rbinom
set.seed(123)
A_vec <- rbinom(n = 1000, size = 1, prob = 5/12)
mean(A_vec)
```

```
[1] 0.412
```

(2b) Do the same for the event that the word "shameful" appears in the speech. That is, create a vector of length 1000, where 1 indicates that the word "shameful" appears in the speech and 0 indicates that it does not. Use the marginal probability from problem (1c). Store this vector in a variable called `B_vec`. Report `mean(B_vec)`. (Do not reset your seed.)

Answer:

```r
B_vec <- sample(x = c(0, 1), size = 1000, prob = c(1 - 3/8, 3/8), replace = T)
mean(B_vec)
```

```
[1] 0.356
```

Suppose that your vectors `A` and `B` are measurements relating to the same 1000 speeches – that is, if the first element of both is 1, that means both "shameful" and "congratulate" appeared in the first speech.

(2c) Restricting to speeches in which "shameful" appears, in what proportion did "congratulate" appear?

Answer:

```r
mean(A_vec[B_vec == 1])
```

```
[1] 0.3960674
```

(2d) Answer (2a)-(2c) again, but this time for a sample of 1 million speeches. Compare your answers with the smaller and larger sample.

Answer:

```r
A_vec_longer <- sample(x = c(0, 1), size = 1000000, prob = c(1 - 5/12, 5/12), replace = T)
mean(A_vec_longer)
```

```
[1] 0.416253
```

```r
B_vec_longer <- sample(x = c(0, 1), size = 1000000, prob = c(1 - 3/8, 3/8), replace = T)
mean(B_vec_longer)
```

```
[1] 0.37505
```

```r
mean(A_vec_longer[B_vec_longer == 1])
```

```
[1] 0.4159365
```

Answer: In each case the average is closer to the true probability of getting a 1. Later we will discuss this as a feature of the law of large numbers.

(2e) If our sample was infinitely large, what would the answer to (2c) be? How does this relate to independence and conditional probability?

Answer: If the sample were infinitely large, the answer to (2c) should be 5/12, the marginal probability of A. Problem (2c) is asking us to compute something that, if the sample were infinite, would equal the conditional probability $P(A \mid B)$. Because A and B are independent (given the way we were instructed to create them), $P(A \mid B) = P(A) = 5/12$.