Statistics 210 FC Assignment 2 Due: 4PM, April 12, 2019

1 Let Ω denote the sample space; let $A, B, C \subset \Omega$ be three events such that $\mathbb{P}(C) > 0$. The events A and B are called *conditionally independent given* C if

$$\mathbb{P}(A \cap B|C) \;=\; \mathbb{P}(A|C)\mathbb{P}(B|C).$$

Answer the following questions:

a (6 marks) Assume that the sample space is $\Omega = \{s_0, s_1, \ldots, s_{10}\}$ and $\mathbb{P}(s_0) = \mathbb{P}(s_1) = \cdots = \mathbb{P}(s_{10}) = 1/11$. We define the events:

$$\begin{aligned} A &= \{s_0, s_7\}, \\ B &= \{s_3, s_4, s_6, s_7\}, \\ C &= \{s_0, s_1, \ldots, s_7\}. \end{aligned}$$

Prove that

$$\begin{aligned} \mathbb{P}(A \cap B|C) &= \mathbb{P}(A|C)\mathbb{P}(B|C), \\ \mathbb{P}(A \cap B) &\neq \mathbb{P}(A)\mathbb{P}(B). \end{aligned}$$

Hint: Calculate all the probabilities that are involved,
$\mathbb{P}(A \cap B|C)$, $\mathbb{P}(A|C)$, $\mathbb{P}(B|C)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(A)$, $\mathbb{P}(B)$,
and then use the results in order to show that the relationships expressed by the two equations are correct. When computing the probabilities, have a look at the content of Section 1.3 of the course book.

b (6 marks) Now we assume that $\Omega = \{s_0, s_1, \ldots, s_7\}$ and $\mathbb{P}(s_0) = \mathbb{P}(s_1) = \cdots = \mathbb{P}(s_7) = 1/8$. We define the events:

$$\begin{aligned} A &= \{s_0, s_7\}, \\ B &= \{s_2, s_3, s_6, s_7\}, \\ C &= \{s_0, s_2, s_3, s_6, s_7\}. \end{aligned}$$

Prove that

$$\begin{aligned} \mathbb{P}(A \cap B|C) &\neq \mathbb{P}(A|C)\mathbb{P}(B|C), \\ \mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B). \end{aligned}$$

Comment: These simple examples show that, in general, the conditional independence of A and B given C neither implies, nor is implied by, the statistical independence of A and B.

2 Answer the following questions. Show your working for each part.

a (6 marks) Suppose that X is a random variable that represents the number shown when we roll the die D_1 which has 9 sides. The probability function of X is given by

$$\mathbb{P}(X = x) = \log_{10}\left(\frac{x+1}{x}\right), \text{ for } x \in \{1, 2, \ldots, 9\},$$

where $\log_{10}(\cdot)$ denotes the logarithm base 10. Show that the cumulative distribution function of X is:

$$F_X(x) = \begin{cases} 0, & \text{if } x \in (-\infty, 1), \\ \log_{10}(\lfloor x \rfloor + 1), & \text{if } x \in [1, 9], \\ 1, & \text{if } x \in (9, \infty), \end{cases}$$

where $\lfloor \cdot \rfloor$ is the greatest integer less than or equal to the real number in the argument. For example, $\lfloor 2.0 \rfloor = 2$, $\lfloor 2.5 \rfloor = 2$, $\lfloor 2.999999 \rfloor = 2$.

b (5 marks) Let Y be a random variable that represents the number shown when we roll the 9-sided die D_2 which is known to be fair. Hence, we have:

$$\mathbb{P}(Y = y) = \frac{1}{9}, \text{ for } y \in \{1, 2, \ldots, 9\}.$$

Show that the cumulative distribution function of Y is given by

$$F_Y(y) = \begin{cases} 0, & \text{if } y \in (-\infty, 1), \\ \frac{1}{9}\lfloor y \rfloor, & \text{if } y \in [1, 9], \\ 1, & \text{if } y \in (9, \infty), \end{cases}$$

3 Suppose that a family has $n = 6$ children. Assume that the probability that any child will be a girl is $1/2$ and that all births are independent. Let X be a random variable that represents the number of girls in the family.

Answer the following questions:

a (2 marks) State the distribution of X, with parameters.

b (2 marks) Express the following events by using the random variable X:

- The family has at least one girl.
- The family has at least one boy.

c (5 marks) Given that the family has at least one girl, determine the probability that the family has at least one boy.

Hint: Use the results from part (a) and from part (b).

4 Let $p \in (0,1)$. Assume that $X \sim \text{Bernoulli}(p)$, $Y \sim \text{Bernoulli}(1/2)$, and the random variables X and Y are statistically independent. The random variable Z is given by $Z = X \oplus Y$, where the operator $\oplus$ is defined in Table 1. The operation that corresponds to the symbol $\oplus$ is called *exclusive or* (often abbreviated to XOR), or *addition mod 2.*

Most importantly, for any $x, y \in \{0,1\}$, $x \oplus y = 0$ if $x = y$ and $x \oplus y = 1$ if $x \neq y$.

x	y	$z = x \oplus y$
0	0	0
0	1	1
1	0	1
1	1	0

Table 1: Definition of the operator $\oplus$.

Answer the following questions:

a (5 marks) Show that $Z \sim \text{Bernoulli}(1/2)$.

b (10 marks) In the case when $p \neq 1/2$, prove that the random variables Y and Z are *not* statistically independent. Show your working.

Hint: Use the definition of independence for random variables, which is given in Section 1.4 of the course book.

c (1 mark) Explain why the random variables Y and Z are statistically independent when $p = 1/2$.

5 After their STATS 210 lecture, Alice told Bob that, in any English text it is equally likely to have letters from the set $\mathcal{V}$ and from the set $\mathcal{C}$ that are defined below:

$\mathcal{V} = \{\text{A}, \text{a}, \text{E}, \text{e}, \text{I}, \text{i}, \text{O}, \text{o}, \text{U}, \text{u}\}$,

$\mathcal{C} = \{\text{B}, \text{b}, \text{C}, \text{c}, \ldots, \text{Y}, \text{y}, \text{Z}, \text{z}\}$.

Note that the entries of $\mathcal{C}$ are all the letters of the English alphabet (upper-case and lower-case) which are not included in $\mathcal{V}$.

Bob was not convinced that the claim is true and, because of that, he decided to investigate it by analyzing Shakespeare's Sonnet 130. More precisely, he took the text of the sonnet and replaced each letter from $\mathcal{V}$ by symbol 'V'. Then he replaced any letter from $\mathcal{C}$ by symbol 'C'. He also removed all the punctuation marks, but he maintained the spaces between the original words.

The text produced after these alterations is:

CC CVCCCVCC VCVC VCV CVCCVCC CVCV CCV CVC
CVCVC VC CVC CVCV CVC CCVC CVC CVCC CVC
VC CCVC CV CCVCV CCC CCVC CVC CCVVCCC VCV CVC
VC CVVCC CV CVCVC CCVCC CVCVC CCVC VC CVC CVVC
V CVCV CVVC CVCVC CVCVCCVC CVC VCC CCVCV
CVC CV CVCC CVCVC CVV V VC CVC CCVVCC
VCC VC CVCV CVCCVCVC VC CCVCV CVCV CVCVCCC
CCVC VC CCV CCVVCC CCVC CCVC CC CVCCCVCC CVVCC
V CVCV CV CVVC CVC CCVVC CVC CVCC V CCVC
CCVC CVCVC CVCC V CVC CVCV CCVVCVCC CVVCC
V CCVCC V CVCVC CVC V CVCCVCC CV
CC CVCCCVCC CCVC CCV CVCCC CCVVCC VC CCV CCVVCC
VCC CVC CC CVVCVC V CCVCC CC CVCV VC CVCV
VC VCC CCV CVCVVC CVCC CVCCV CVCCVCV

After counting carefully, Bob and Alice came to the conclusion that, out of the 464 letters of the text, only 171 are from $\mathcal{V}$.

Answer the following questions:

a (2 marks) Out of the 464 letters of the text, let X be the number of the letters from $\mathcal{V}$. Under the assumptions that (i) the symbols in the altered text have no influence on each other and (ii) it is equally likely to have in the altered text the symbols 'V' and 'C', state the distribution of X, with parameters.

b (2 marks) We wish to test the hypothesis that it is equally likely to have letters from the sets $\mathcal{V}$ and $\mathcal{C}$ in the English texts. Formulate the null and alternative hypotheses in terms of the distribution of X and its parameters. Specify the full distribution of X and use a two-sided alternative hypothesis.

c (2 marks) Sketch as a curve the shape of the probability function of X under the null hypothesis. Your sketch should have axes labelled x and $\mathbb{P}(X = x)$. Mark on the sketch the upper and lower limits of x, and the approximate value of x where the curve peaks under the null hypothesis. Also mark the observed value $x = 171$ so that you can see the tail probabilities required for the p-value, and shade under the curve the area represented by the p-value.

d (5 marks) Write down the `R` command required to find the p-value for the hypothesis test, and run this command in `R` to find the p-value. Interpret the result in terms of the strength of evidence against the null hypothesis.

Total: 59 marks.