Signals and Data Exercises — Week 1

We assume the following import statements.

```
import numpy as np
import matplotlib.pyplot as plt
```

First Session

Exercise 1 Sample Spaces

1. write down all members of the sample spaces and count or calculate their size for the following sampling processes,
   a) a single coin flip that can come up either heads or tails.
   b) 3 coin flips in sequence.
   c)
   d) the truth of a scientific hypothesis.
2. for the following random processes, just describe the sample space in words and calculate its size if possible — note that some of the examples below have infinite sample spaces.
   a) the roll of 5 dice with 6 sides each in a game of Yatzy. Calculate the size of this sample space.
   b) the number of defective components in a container full of an unknown number of electronic components.
   c) the trading history of a stock trading algorithm that buys or sells a stock randomly each second until it has both bought and sold at least once.

Hints

Note that for this exercise you just have to describe the different possible outcomes, not the probabilities of each event.

Exercise 2 Probabilities

One of the simplest ways to generate random values in the real world is by rolling a die. In this case, we call the random value of the die X which takes values in the sample space $S = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$, corresponding to the 6 possible sides of a normal die.

1. Use the numpy function `np.random.randint` to define a function with the name `dieroll` that simulates a single roll of a six-sided die.

```
def dieroll():
    random_roll = # your sampling code goes here
    return random_roll
```

   a) Roll the die multiple times by calling `dieroll()`. Check that it only outputs values in the event space.

   b) The probability of the die landing on any side is supposedly $1/6$ — all sides are equally probable. Simulate 6000 die rolls. Verify that each possible roll is roughly equally likely as expected. You can base your solution on the following code snippet for plotting by defining `samples_equal_to_n`, `N`, and `true_prob` appropriately,

```
running_sample_N = np.arange(1,N+1)
running_count = np.cumsum(samples_equal_to_n)
plt.plot(running_sample_N, running_count/running_sample_N,
    'r-', label='running average')
plt.plot([1,N+1], [true_prob, true_prob],
    'k--', label='probability')
plt.legend()
```

2. Now consider rolling 2 dice.

   a) build a new function using `dieroll` that rolls two dice and returns the pair.

   b) calculate the probability of rolling two sixes and check that it matches with the sampler.

As each die roll is independent, we have the general result that if we roll K dice,

$$\mathbb{P}(\bigcap_{k=1}^{K} \{\text{die } k \text{ is a } ⚅\}) = \prod_{k=1}^{K} \mathbb{P}(\text{die } k \text{ is a } ⚅). \tag{1}$$

3. Finally, consider rolling 5 dice, as in the Yatzy example before.

   a) Use the formula above to calculate the probability of rolling 5 6's (a Yatzy, in the game).

b) Extend your sampling function from before to 5 dice.

c) Use your sampler to check your calculation, but call[1]

```
np.random.seed(1)
```

before drawing 6000 samples. What went wrong?

Exercise 3 Sum and Product Rule

1. Imagine we take a random family with two children. Let the probability of both the younger and the elder child being a girl be

$$\mathbb{P}(\text{elder child is a girl}) = \mathbb{P}(\text{younger child is a girl}) = 1/2. \tag{2}$$

We also assume that that the gender of each child is independent of the other.

a) What is the sample space?

b) What is the probability of both children being girls?

c) What is the probability that there is at least one girl in the family?

d) What is the probability that both children are girls if the elder child is a girl?

e) What is the probability that both children are girls if either the younger or the elder child is a girl?

Exercise 4 Probability Rules

Recall the following axioms and definitions for sample space S and events $A, B \subset S$,

$$\begin{aligned}
\mathbb{P}(S) &= 1 && \text{(unit measure)} \\
\mathbb{P}(A|B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} && \text{(conditional)} \\
\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) && \text{(sum rule)} \\
\mathbb{P}(A \cap B) &= \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) && \text{(product rule)}
\end{aligned}$$

1. If the probability of an event is $\mathbb{P}(A)$, what is the probability of its complement $\mathbb{P}(A^c)$? Use the result to derive $\mathbb{P}(\emptyset)$ (probability of the empty set of no events).

2. Assume that we have a *partition* of the sample space S, defined as K sets A_k where,

[1] this is just to make sure that you all get the same result.

- none of the sets overlap, so that their intersection is empty $A_k \cap A_j = \emptyset$.
- the union of all the sets, denoted by $\bigcup_{k=1}^{K} A_k$, is equal to the full sample space S.

To give an example, for a die, the set A_1 could be the event of rolling 1, A_2 the event of 2, and so on.

Now prove that $\sum_{k=1}^{K} \mathbb{P}(A_k) = 1$ and $\sum_{k=1}^{K} \mathbb{P}(A_k|B) = 1$ for any B.

3. Assume a partition A_k like above. Prove that

$$\mathbb{P}(B) = \sum_{k=1}^{K} \mathbb{P}(B|A_k)\mathbb{P}(A_k). \tag{3}$$

The result is known as **the Law of Total Probability**.

4. Use the product rule to show that,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{4}$$

This is known as **Bayes Theorem**.

Hints

It can be helpful to draw Venn diagrams to build an intuition.

1. Use the unit measure and sum rule and consider the event $A \cup A^c$.
2. Try applying the sum rule to $\mathbb{P}(\bigcup_{k=1}^{K} A_k)$. Note that for the second equation, we can use that if $C \subset S$ is a third event we can generalize the sum rule to also cover

$$\mathbb{P}(A \cup B|C) = \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(A \cap B|C). \quad \text{(conditional sum rule)}$$

3. Use the product rule, the sum rule, the unit measure, and the result of the preceding exercise.

Second Session

Exercise 5 Marginals and Conditionals

Imagine that two different drugs for treating the same disease are in use at a hospital and that you set out to determine which drug is the better one.

You collect the following data on the patients treated at the hospital, indicating whether each patient recovered or remained sick after the administration of either drug,

	Cured	Sick
Drug A	273	77
Drug B	289	61

1. Consider the random process where we select one person from the study at random. You can calculate the following either by hand or using array operations in NumPy.

   a) What is the marginal probability of that person being sick?

   b) What is the joint probability of that person being both sick and in the trial group for drug A?

   c) What is the conditional probability that a person is sick given that he took drug B?

   d) Which drug would you argue is better based on the study?

It comes to your attention that the doctors submit each person to a test before treatment, and if the test is positive most of the doctors prefer treating the patient with drug A which they believe to be superior. You recover the test results, and split your survey into two,

Negative test	**Cured**	**Sick**
Drug A	81	6
Drug B	234	36

Positive test	**Cured**	**Sick**
Drug A	192	71
Drug B	55	25

2. Consider the same random process as before.

   a) Which drug works best for those with a positive test? Which drug works best for those with a negative test?

   b) Were the doctors right or wrong?

The above is an example of *Simpson's paradox* — even if one drug is better, it can appear worse if you test it on more of the difficult cases.

Exercise 6 Bayes Theorem

The following example is one that everybody learning probability and statistics has to go through.

1. Imagine that you have to help doctors identify a disease that occurs with probability 0.01 (1 out of every 100). If a patient is sick, the test is positive 90% of the time and if the patient is healthy, the test is negative 90% of the time.

   a) What is the probability of the test being positive?

   b) What is the joint probability $P(S = 1, T = 1)$ of being sick and the test being positive?

c) Finally, calculate the probability that a person is sick, conditioned on his test being positive, i.e. $P(S = 1|T = 1)$. Try to make a guess based on your intuition first.

Note that in this way, Bayes' theorem allows us to make predictions about things that we do not know based on the information that we actually have.

Hints

You will need the law of total probability, the product rule, Bayes' theorem, and the definition of a conditional distribution.

Exercise 7 A Probabilistic Classifier

Suppose that 30 percent of computer owners use a Mac, 50 percent use Windows, and 20 percent use Linux. Suppose that 65 percent of the Mac users have succumbed to a virus, with the same occurring to 82 percent of the Windows users, and 50 percent of the Linux users.

1. consider the random process where we select a computer owner completely at random. Use Bayes' theorem as you did above to calculate the probability of the user employing either of the three operating systems, conditioned on there being a virus on his system.

These probabilities quantify how likely each OS is given the observation of the virus. We can use this to make predictions, and we can use the probability of misclassification (being wrong) as the metric. If $s_{\text{prediction}}$ is our prediction and S is the operating system, the probability of misclassification (being wrong) is,

$$\mathbb{P}(S \neq s_{\text{prediction}}|V = 1)$$

2. Calculate the probability of misclassification for the three decision rules where we predict the user to have either Windows, Mac, or Linux.

You should find that the OS with the highest probability conditioned on the virus being present is the prediction with the lowest conditional risk. This is in fact the best possible classifier with the lowest probability of misclassification, known as the *Bayes optimal classifier*.

Hints

If stuck on the first part, try using the law of total probability to calculate the probability of having a virus, and then follow the steps in the previous exercise.