1 Codebreaking

In the previous section, we mostly looked at encryption and decryption of many ciphers. Now, we'll look at how to *break* some of these ciphers. It should be noted that codebreaking is not necessarily "exact science"; that is, there's not necessarily an algorithm that guarantees producing the correct plaintext from ciphertext in one shot without access to the key. Instead, these techniques can help constrain the search for the correct ciphertext.

1.1 Frequency Analysis

Frequency analysis is a powerful technique used to break simple – and sometimes also polygraphic – substutition ciphers. The idea is relatively simple.

> **Heuristic:** The relative frequencies of letters remain *roughly* stable across different samples of English texts, and `ETAOINSHRDLU` is the *approximate* order of the 12 most common letters.

We can use this heuristic to break simple substutition ciphers. Ideally, the technique works best with longer ciphertexts, but the idea is to guess the decryption key one letter at a time, doing one of the following at each step:

1. Assign the most frequent unassigned letter of ciphertext to be the most frequent unassigned letter in some sample English text (or perhaps some other letter with a similar frequency).
2. Look through the ciphertext and see if you can make any guesses about words that seem to appear there. If you see something, fill in the blanks in that word by making appropriate guesses for the key.

If, at any point, it seems like your guesses are leading to nonsense or implausible sequences of letters, backtrack and make another guess. A few comments:

- Usually, we can start with two applications of option 1. For example, we can guess that the most common letter in the ciphertext is `E` and the second most commmon letter is `T`.
- We can also note that `THE` occurs frequently in English (and other similar words like `THEY` or `THEIR` or `THEN`).
- If, after you make the `T` and `E` substutitions, you see the `T*E` pattern frequently (`*` being some *fixed* letter in ciphertext), you can make the assumption that `*` could be `H`.
  - Also, perhaps if you see `TH*T` occurring in your ciphertext after making the substutitions and with `*` fixed, you can probably assume that `*` is `A`.
- If you can't spot any possible words, you can always try using option 1 instead and match the most frequent letters.

Usually, the first few guesses after `E` and `T` are the hardest. Once you've made a few correct guesses, it becomes easy to see words.

1.2 Interlude: Probability

Notice how, in the previous observation, we made use of the Heuristic to help us mount attacks on substutition ciphers. We can use variants of this observation for other ciphers, but this requires us to first talk about **probability**.

1.2.1 Experiments and Events

In probability thoery, the word *experiment* is used to talk abstractly and heuristically about processes which generate "outcomes" and which might be rather intricate. These experiments are formally modeled by *probability spaces*. For now, we'll use the following definition.

Definition 1.1: (Discrete) Probability Space

A **(discrete) probability space** is a nonempty countable[a] set Ω called the **sample space** and whose elements are called **outcomes**. Each outcome $x \in \Omega$ is assigned a real number $\mathbb{P}[x]$ between 0 and 1 called its **probability**. The probabilities of all the outcomes must sum to 1; that is,

$$\sum_{x \in \Omega} \mathbb{P}[x] = 1.$$

[a] "Countable" means that the outcomes can be put in a list so that the summation $\sum_{x \in \Omega} \mathbb{P}[x]$ makes sense. Any finite set, and some infinite sets, are countable. For now, we'll focus on the finite case.

The probability associated to each outcome should be thought of as some measure of our "confidence" that our experiment will produce that outcome. For example, it might be the percentage of times we expect the experiment to produce that outcome if the experiment were to be releated many times.

(Example.) Rolling a dice is an example of an experiment.

- The possible outcomes of this experiment are the numbers 1 through 6; that is, the **sample space** is
$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$
- Assigning each outcome a probability of $\frac{1}{6}$, that is, for $x \in \Omega$,
$$\mathbb{P}[x] = \frac{1}{6},$$
means that the dice is "fair" and each outcome is equally likely.

Thus, we constructed a probability space; we have a finite set Ω that enumerates the possible outcomes, and we assigned a probability to each outcome.

A single experiment can also have "multiple parts," as seen in the next example.

(Example.) Flipping a fair coin twice can be thought of as a single experiment.

- Possible outcomes of this experiment might be something like "heads and then heads again" or "heads and then tails" and so on. All these outcomes taken together as a set form the sample space,
$$\Omega = \{HH, HT, TH, TT\},$$
where H means "Heads" and T means "Tails."
- We can assign each of these four outcomes probability $\frac{1}{4}$, that is for some $x \in \Omega$
$$\mathbb{P}[x] = \frac{1}{4}.$$

Here, we've modeled the situation where the coin is fair and the result of each coin flip is unrelated to the other.

Notice how both examples above have outcomes with the same probabilities. This is a common situation, and thus has a name.

Definition 1.2: Uniform Distribution

A probability space is **uniform** if all of its outcomes have equal probability.

Sometimes, we might be interested in grouping the various outcomes together. We can do so with a definition.

Definition 1.3: Event

Given a probability space, an **event** E is a subset of the sample space Ω; that is,

$$E \subset \Omega.$$

We define

$$\mathbb{P}[E] = \sum_{x \in E} \mathbb{P}[x].$$

Remark: The words "event" and "outcome" have distinct definitions in probability theory.

(Example.) Consider the example of rolling a dice again. An *event* might be something like "the dice roll is odd." Formally, if we think of the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, the event "the dice roll is odd" corresponds to the event

$$E = \{1, 3, 5\}.$$

This event is also assigned a probability, by summing together the probabilities of all outcomes that comprise the event:

$$\mathbb{P}[E] = \mathbb{P}[1] + \mathbb{P}[3] + \mathbb{P}[5] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

(Exercise.) Suppose you have 4 boxes (labeled 1, 2, 3, and 4), and you have 8 colors available (red, blue, green, yellow, pink, purple, teal, brown). Consider an experiment where each of the 4 boxes is assigned a color. For example, one possible outcome of this experiment might be the one where box 1 is colored red, box 2 is colored blue, box 3 is colored green, and box 4 is colored blue.

1. How many possible outcomes are there?

   The answer is $8^4 = 70$ outcomes. We can assign any of the 8 colors to box 1, any of the 8 colors to box 2, any of the 8 colors to box 3, and any of the 8 colors to box 4.

2. How many outcomes are in the event "no two boxes have the same color?"

   The answer is $8 \cdot 7 \cdot 6 \cdot 5 = 1680$. Once we pick a color, we can no longer use that color for the next box.

(Exercise.) Suppose you have k boxes and you have n colors available. Consider again the same experiment where each of the k boxes is assigned one of the n colors "at random" (i.e., construct a uniform probability space).

1. What is the probability of the event that no two boxes have the same color?

   Note that the number of outcomes such that no two boxes have the same color is given by $n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot (n-k+1)$. The total number of outcomes is n^k. The probability is given by

   $$\frac{n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot (n-k+1)}{n^k}.$$

2. What is the probability that there are at least two boxes of the same color?

Note that the event that at least two boxes have the same colors is the opposite of the event that no two boxes have the same colors. In other words,

$$\mathbb{P}(\geq 2 \text{ Boxes Have Same Color}) = 1 - \mathbb{P}(\text{No Two Boxes Have Same Color}).$$

This gives us

$$\mathbb{P}(\geq 2 \text{ Boxes Have Same Color}) = 1 - \frac{n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot (n-k+1)}{n^k}.$$

3. Find expressions in terms of n and k.

Notice that

$$n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot (n-k+1) = (n)_k = \frac{n!}{(n-k)!}.$$

This is known as a falling factorial. So,

(a) $\frac{\frac{n!}{(n-k)!}}{n^k}$.

(b) $1 - \frac{\frac{n!}{(n-k)!}}{n^k}$.

1.2.2 Random Variables

A common way that events show up is through **random variables**. We can think of random variables as representations of making an observation (or taking a measurement) on the outcome of an experiment. A random variable has a set of possible values that it can take. Letters like X or Y can be used to denote random variables.

Definition 1.4: Random Variable

Fix a probability space Ω. A **random variable** is a function with domain Ω and its set of *possible* values is the range of this function.

(Example.) Consider the "multi-part" experiment discussed earlier (with the coin being flipped twice). We can make the observation that the first coin flip can be thought of as a random variable, which we can call X. X can take the value "heads" or "tails." Then, we can write things like $X = H$ to refer to the event that the first coin flip landed heads. In other words, in the sample space

$$\Omega = \{HH, HT, TH, TT\},$$

the notation $X = H$ describes the event $\{HH, HT\}$ and we have

$$\mathbb{P}[X = H] = \mathbb{P}[HH] + \mathbb{P}[HT] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

(Example.) Suppose we're interested in the number of heads. We can define another random variable Y that can take values 0, 1, or 2. The notation $Y = n$ for either $n = 0, 1, 2$ describes the event that we

observe n heads out of the two coin flips. So, for $Y = 1$, we have the event $\{HT, TH\}$ and

$$\mathbb{P}[Y = 1] = \mathbb{P}[HT] = \mathbb{P}[TH] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

However, for $Y = 0$, we have the event $\{TT\}$ and

$$\mathbb{P}[Y = 0] = \mathbb{P}[TT] = \frac{1}{4}.$$

Definition 1.5: Uniform Random Variable

A random variable is **uniform** if all of its values have equal probability.

In the previous two examples, X is uniform (it can either take heads or tails, i.e., $X = H$ or $X = T$, both of which have probabilities $1/2$) whereas Y is not uniform.

Definition 1.6: Expected Value

Suppose X is a random variable whose values are real numbers. The **expected value**, known as the expectation, of X, denoted $\mathbb{E}[X]$, is defined by

$$\mathbb{E}[X] = \sum_{\text{values } a} a \cdot \mathbb{P}[X = a].$$

(Example.) In the experiment involving two coin flips, the random variable Y which counts the number of heads has real number values (0, 1, 2). Its expectation is given by

$$\mathbb{E}[Y] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

(Exercise.) Consider the experiment where you roll a pair of fair dice. Let the random variable X denote the sum of the dice rolls.

1. What are the possible values of X?

   The possible values are
   $$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

2. What is $\mathbb{P}[X = 7]$?

   Note that the pair of fair dice will have sum 7 if we get
   $$\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1).\}$$
   Therefore,
   $$\mathbb{P}[X = 7] = \frac{6}{36} = \frac{1}{6}.$$

3. What is $\mathbb{P}[X = 7 \text{ or } 11]$?

Note that the paur of fair dice will have 11 if we get

$$\{(6,5),(5,6)\}.$$

Combining this with this previous part, we have 8 possible combinations. This gives us

$$\mathbb{P}[X = 7 \text{ or } 11] = \frac{8}{36} = \frac{2}{9}.$$

4. What is $\mathbb{E}[X]$?

Note that

- For sum 2, there is only 1 possible combination.
- For sum 3, there are 2 possible combinations.
- For sum 4, there are 3 possible combinations.
- For sum 5, there are 4 possible combinations.
- For sum 6, there are 5 possible combinations.
- For sum 7, there are 6 possible combinations.
- For sum 8, there are 5 possible combinations.
- For sum 9, there are 4 possible combinations.
- For sum 10, there are 3 possible combinations.
- For sum 11, there are 2 possible combinations.
- For sum 12, there are 1 possible combinations.

Therefore, the expected value is

$$\begin{aligned}\mathbb{E}[X] &= 2\frac{1}{36} + 3\frac{2}{36} + 4\frac{3}{36} + 5\frac{4}{36} + 6\frac{5}{36} + 7\frac{6}{36} + 8\frac{5}{36} + 9\frac{4}{36} + 10\frac{3}{36} + 11\frac{2}{36} + 12\frac{1}{36} \\ &= 7.\end{aligned}$$