NUMB3RS Activity: How Hard Can It Be?

Episode: "Toxin"

Topic: Entropy **Grade Level**: 11 - 12

Objective: Using entropy to determine the complexity of a coding scheme.

Time: about 30 minutes

Introduction

In Toxin, the FBI tracks a serial poisoner who uses contaminated drugs. In this activity, the FBI intercepted an encrypted message they believe contains a phone number that could lead to an "insider" who may be helping the poisoner. Don suggests using a computer to start trying to break the code. Charlie proposes that first they gather as much information as possible and use Information Theory to help determine how long it this might take.

Information Theory, developed by Claude Shannon, is the basis behind encrypting and decrypting codes. Entropy, in cryptology, is the mathematical measure of the amount of uncertainty based on probability distributions. It measures how likely it is for someone to successfully decipher a code–the higher the entropy, the harder to break.

A code using only two symbols has the lowest level of uncertainty. After one "guess" you would know the code. Shannon defined a unit of entropy based on this fact and used a fair coin as an illustration. A fair coin has a sample space {h, t}, heads or tails, with probabilities $P(\text{h}) = P(\text{t}) = 0.5$. **Entropy** is defined as a function of H, where $H(P(\text{event})) = -\log_2(P(\text{event}))$. In this case of $H(P(\text{h}))$, $H(P(\text{h})) = -\log_2(P(\text{h})) = -\log_2(0.5) = 1$.

$H(P(\text{t}))$ is equal to $H(P(\text{h}))$. The unit for entropy is called a bit.

The **expected entropy** for the sample space is the sum of the probability of each event in the sample space times its entropy.

For the coin toss the expected entropy for sample space (C) is

$$E_C = (0.5)(-\log_2(0.5)) + (0.5)(-\log_2(0.5)) = (0.5)(-(-1)) + (0.5)(-(-1)) = 1 \text{ bit.}$$

The expected entropy of n events (E_n) would be found by:

$$E_n = P(x_1)(\text{entropy } x_1) + P(x_2)(\text{entropy } x_2) + \ldots + P(x_n)(\text{entropy } x_n) \text{ and therefore,}$$

$$E_n = P(x_1)(-\log_2(P(x_1)) + P(x_2)(-\log_2(P(x_2)) + \ldots + P(x_n)(-\log_2(P(x_n))$$

A sample space with more possible outcomes would raise the level of uncertainty and would also increase the expected entropy. For example, the expected entropy for a fair six-sided die (D) would be

$$E_D = (6)\left(\frac{1}{6}\right)\left(-\log_2\left(\frac{1}{6}\right)\right) \approx (6)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)(-(-2.585)) \approx 2.585 \text{ bits.}$$

Events with a probability of 0 are defined to have an entropy of 0. In this activity, students are introduced to entropy and how to calculate expected entropy.

Terry Souhrada, University of Montana, MT

Discuss with Students

Lead a discussion using some of the questions below.

1. How many possible numbers could be formed if each digit of a 10-digit phone number could contain any of the number 0–9?
2. If a computer could process one of the possible numbers in one second, how long would it take to process all the possible telephone numbers?
3. How would knowing that the first digit in the access code (second three digits) of the telephone number cannot be 0 or 1 effect the number of possible telephone numbers and the time it takes to process them?
4. Would a code using the letters of the alphabet be easier or harder to break than a code using the numbers 0–9? Explain.

Discuss with students answers: 1. 10^{10}*= 10,000,000,000* **2**. *about 317 yrs.* **3.** s*ample response: It would make the number of possible telephone number less (8*10^{9}= 8,000,000,000) and reduce the time to process the numbers (about 253.5 years).* **4.** *sample response: Harder. There are 26 symbols in a code using letters that have to be decoded. In a telephone number, there are only 10.*

Student page answers: 1. *1/10 or 0.1* **2**. *0.1(*$-\log_2(0.1)$*)* **3**. *about 0.332* **4**. *Since the digits are all equally likely the entropy for the last four digits would be 10 time 0.332 or about 3.32.* **5**. *0.06(*$-\log_2(0.06)$*) or about 0.244* **6.** *about 3.25* **7**. *Use the area code because the entropy is less indicating an easier encryption to break. Also other information about area codes is known, such as no area code begins with a 0, which may further help decipher the code.* **8**. *about 4.70* **9**. *about 3.63* **10**. *Greater. It is adding more coded symbols that must be decoded. For example, just adding "space" as one of the coded symbols raised the entropy from about 3.63 to about 4.08.* **11**. *The code using random generated letters would be more difficult to break, indicated by a greater entropy value.*

Terry Souhrada, University of Montana, MT

Episode: "Toxin"

Name: _____ Date: _____

NUMB3RS Activity: How Hard Can It Be?

The FBI intercepted a message containing 10 numbers they think is a phone number from the Pacific time zone. Don asks Charlie if it is feasible to break a phone number code using a computer. Charlie replies that it will take time but suggests that expected entropy can give them some indication of the difficulty involved. The expected entropy of a code is its level of "difficulty" to break. More characters or numbers in a code give it more possibilities requiring more time to break it.

1. To calculate the expected entropy, you must know the probability of each digit appearing. For the last four digits, the probability for selecting any of the digits 0–9 is equally likely. What is the probability of selecting 1 digit from 0–9?
2. The entropy uses logarithms with a base of 2. In general the expected entropy for each number appearing is $P(n) \cdot (-\log_2 P(n))$, where $P(n)$ is the probability of the digit appearing. What expression would represent the expected entropy for a 5 appearing?
3. Most calculators require base two logarithms to be converted to either base 10 (common) or base *e* (natural) logarithms. You can convert $-\log_2(0.1)$ using common (base 10) logarithms by entering $-\log(0.1)/\log(2)$ on a Texas Instruments TI-84 Plus calculator. What is the expected entropy for a 5 appearing?
4. To find the expected entropy for the encryption used on the last four digits of the phone number, add the expected entropies for each of the numbers 0–9. What is the expected entropy for the last four digits?
5. The expected entropy for the area code (first three digits in the phone number) will be different than that of the last four because not all the numbers 0–9 appear in area code with the same probability. For example, of the 36 area codes in the Pacific time zone, the probability of a 4 appearing is about 0.06. Based on this probability, what is the expected entropy for a 4 appearing in the area code?
6. The table below shows the probability for each number appearing in the 36 area codes. Determine the expected entropy for the encryption of the area code.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Probability	0.14	0.12	0.09	0.08	0.06	0.16	0.10	0.07	0.07	0.11

7. If you want to break the code for a seven-digit telephone number, would you start with the area code or the last four digits? Explain.

Terry Souhrada, University of Montana, MT

The FBI agents intercepted a second message they think is indicating where the serial poisoner and the "insider" are going to meet. Unfortunately, the message is encoded and there is no indication what the key to the code might be.

8. Assuming all the letters are equally likely to appear in the message, what is the expected entropy for the code used to create the written messages?

9. Because of the structure of words and sentences in English, the probability of each letter appearing is not equal. The table below shows the approximate probability for each letter and a space appearing in the English language. Determine the expected entropy of the letters based on these probabilities. (Hint: Use the list feature on your calculator to do the calculations more efficiently.)

Letter	Probability	Letter	Probability	Letter	Probability
a	0.0642	j	0.0008	s	0.0514
b	0.0127	k	0.0049	t	0.0796
c	0.0218	l	0.0321	u	0.0228
d	0.0317	m	0.0198	v	0.0083
e	0.1031	n	0.0574	w	0.0175
f	0.0208	o	0.0632	x	0.0013
g	0.0152	p	0.0152	y	0.0164
h	0.0467	q	0.0008	z	0.0005
i	0.0575	r	0.0484	space	0.1859

10. If punctuation marks were encoded and added to the table above, would you expect the new entropy to be greater or less than the entropy found for 9? Explain.

11. Based on the entropy from Parts 8 and 9, which code would be the more difficult to decipher?

Terry Souhrada, University of Montana, MT

The goal of this activity is to give your students a short and simple snapshot into a very extensive math topic. TI and NCTM encourage you and your students to learn more about this topic using the extensions provided below and through your own independent research.

Extensions

Computers & Entropy

- Entropy plays an important role when computer technologists try to create "unbreakable" encryptions for processes such as purchasing items over the internet using a credit card. In this type of setting, it is important that the purchaser know that the information being transmitted cannot be decoded and for the seller to be able to read the information once it is received. Research symmetric coding, asymmetric coding, one-key pads and other encrypting procedures and how each affects the entropy of the code being used. The sources below are given as a starting place.

  http://www.cs.bris.ac.uk/~bradley/publish/SSLP/chapter3.html
  http://www.bletchleypark.net/cryptology/Perfect_Secrecy.pdf

Entropy in Science

- Explore the meaning and use of entropy in other areas including but not limited to biology, physics, and chemistry. Compare your findings to the meaning of entropy in making and breaking codes.

Coding in Other Languages

- Consider that a lot of code-breaking is not conducted in English. If you're familiar with a foreign language, think about how the chart on the previous page would change for other languages. What other parts of speech or letter-usage will be more common in that language than in English?

Investigating Letter Distribution in Scrabble

- The board game Scrabble has a number of tiles with letters on them used to form words. Consider the distribution of the letters on the tiles. Based on the chart on the previous page, are the letter tiles in Scrabble distributed according to their common use in English?

Terry Souhrada, University of Montana, MT