# Wheel of Fortune: An Information Theory Approach

Peter Danshov, Dr. Johann Thiel, New York City College of Technology, CUNY – Spring 2014

## Introduction

Wheel of Fortune is a television game show with rules similar to the game Hangman. Three players are given a set of blank spaces representing letters in a short phrase. Each player takes a turn spinning a wheel for money and prizes, and then guessing a letter. If that letter appears in the puzzle, its locations in the phrase are revealed and the player can try to solve the puzzle by guessing the phrase. After several rounds, the player that has won the most prizes goes on to play in a bonus round, where the rules are a little different. Players are given a single puzzle along with the puzzle category (thing, phrase, etc.) and the locations of the letters R, S, T, L, and E (if they appear at all). They can then select three more consonants and a vowel. After all of the chosen letters are revealed, the player has ten seconds to guess the solution to the puzzle (multiple guesses are allowed and encouraged). Winners are usually awarded a prize that is much larger than the prizes given out during the earlier part of the game.

In this project, the goal is to use regular expressions and ideas from Shannon's information theory to analyze the distribution of letters and words in the bonus round puzzles of Wheel of Fortune. In doing so, we hope to discover potential patterns that players can exploit to improve their chances of solving these puzzles.

## Literature Review

**Information Theory:** Information theory is a very broad area in applied mathematics which was founded by Claude E. Shannon [1]. In this project, our concern is how it applies to the amount of information that is contained in fragments of words in the English language.

**Entropy:** Shannon [1] defines the entropy of a random process with a finite number of outcome probabilities $p_1, p_2, \ldots, p_n$ as

$$H = -\sum_{i=1}^{n} p_i \log p_i.$$

❖The value of $H$ measures the uncertainty in being able to guess the outcome of the random process. Lower values of $H$ correspond to lower uncertainty.

❖While it may initially seem counterintuitive, low uncertainty means low information content in this setting. The following common examples illustrate this concept:

▪**Coin flips**

Imagine a coin that comes up heads with probability $0 \leq p \leq 1$ and tails with probability $q = 1 - p$. If $p = 1$, then we can expect heads to be the end result of a flip of the coin. In other words, the experiment has no uncertainty and therefore carries no information. We knew what the result would be ahead of time. In this case, $H = 0$.

When $p = q = 1/2$, $H$ is maximized because this is the most uncertain we can be about the outcome of the coin flip.

▪ **Redundancy in English**

Shannon [2] conducted experiments where he measured the redundancy in the English language. He showed people the beginning of a sentence, one letter at a time, asking them to guess the next letter. In doing so, he showed that the English language appears to be at least 50% redundant. While this is helpful in making sure that the right message is communicated between two people, it also seems to say that most English messages could be shortened by at least half without any loss in meaning or content. One example of this redundancy can be seen with the letter Q. Most words in the English language that have a Q usually have a U immediately afterwards. In this case, the removal of U after a Q would very likely not affect the intended meaning of the word. The likelihood of the U after a Q is so high that the informational content of the U would be considered low.

## Research Questions

How do letter and word distributions in Wheel of Fortune bonus round puzzles differ from letter and word distributions in an average English text? Do certain letters carry more/less information content than others? Can these differences be used to create a strategy that improves a player's chance of solving these puzzles?

## Methodology

First we gathered as much data as we could find about the Wheel of Fortune bonus round puzzles. We managed to collect over 1200 puzzles used from 2007 to 2013 [3], as well as additional information, such as the puzzle category, letters chosen by players, and whether the player won or lost. We then wrote a series of programs in the PERL language [4] and employed regular expressions to count the frequency of single letters or letter combinations. Regular expressions are, in some sense, mini-programs written inside of other programs that help process text (see Figure 1).

```
#! perl
# Measures the frequencies of each letter. Converts all lower case to
# upper case.

%hash;
$count = 0;

while (<>) {
    while(length > 0) {
        $char = chop;
        $hash{uc $char}++;
    }
}

foreach $key (sort keys %hash){
  if($key =~ /[A-Z]/){
     $count+=$hash{$key};
  }
}

foreach $key (sort keys %hash){
  if($key =~ /[A-Z]/){
    print $key . " " . $hash{$key}/$count*100 . "\n";
  }
}
```

Figure 1: Sample program using regular expressions.

The occurrences of the letters and letter combinations were stored and grouped by category in double hashed tables, which are basically arrays of arrays. We were then able to calculate the occurrence probability of single letters using the formula

$$p_k = \frac{C_k}{C}$$

where $p_k$ is the probability that the letter $k$ will appear in a puzzle, $C_k$ is the observed number of times the letter $k$ appeared across all puzzles, and $C$ is the total number of letters counted across all puzzles. Note that the formula for $p_k$ can be easily changed to calculate the probability of multiple letter combinations as well.

We then repeated the above computations using Mark Twain's *The Adventures of Huckleberry Finn*, which served as a model of "normal English text" that we could use for comparison.

## Results

Here we present some of our results.

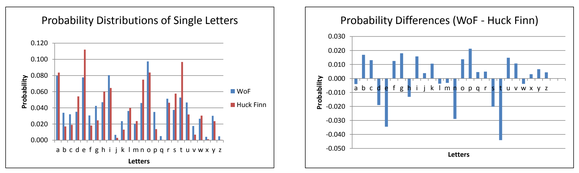


Figure 2: Single letter probability distribution computations.

The leftmost chart in Figure 2 shows the single letter probability distribution for both Wheel of Fortune bonus round puzzles (WoF) and *The Adventures of Huckleberry Finn* (Huck Finn). The rightmost chart in Figure 2 shows the difference between the two distributions per individual letter. In Figure 3 below, we compute the cumulative probability distributions for WoF and Huck Finn.

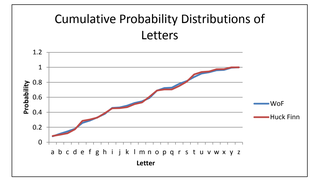


Figure 3: Single letter cumulative probability distributions.

As a first and a basic model we assume that the probability of each individual letter appearing in the puzzles or text is an independent event. With this assumption, we approximate the entropies for WoF and Huck Finn using Figure 2.

$$H_{WoF} \approx 1.327, H_{Huck\ Finn} \approx 1.259$$

## Discussion

From a practical perspective, it can be difficult to measure how different the letter distributions between the Wheel of Fortune bonus round puzzles and *The Adventures of Huckleberry Finn* really are. On one hand, Figure 2 seems to show that there is some kind of difference, but Figure 3 suggests that it is not a *large* difference. Due to the large sample sizes in both cases, a standard sample comparison test is likely to suggest that both distributions are significantly different. The entropy computation suggests that the Wheel of Fortune bonus round puzzle letters are distributed more evenly than in a typical English text, making them harder to predict. However, this is based on an assumption about individual letter probabilities (that they are independent events) that is not an accurate model of the English language. Improvements could certainly be made in this area.

## Conclusion

Further research needs to be carried out on bigrams & trigrams (letter pairs and triplets, respectively) using a more complicated model for the English language. We have collected such data, but have not yet used it. The lengths of the puzzle words has also not been taken into account. As a first approximation, the data from Figure 2 suggests that players can try to maximize their chances of discovering letters in the bonus round puzzle by selecting the letters H, G, D, and O.

## References

[1] Shannon, C.E., *A Mathematical Theory of Communication*, Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October, 1948 .

[3] Shannon, C.E., *Prediction and Entropy of Printed English*, Bell System Technical Journal, 30: 1. January 1951 pp 50-64

[2] http://www.angelfire.com/mi4/malldirectories/wheel/wheelbonus.html

[3] Tom Christiansen; brian d foy; Larry Wall; Jon Orwant, *Programming Perl: Unmatched power for text processing and scripting*, O'Reilly Media, 4th Ed, 2012