Group Exercise 1 Question E

Algorithm Description

Our algorithm samples $\lceil log(n) \rceil$ words from Alices wordlist and sends them one by one to Bob, who checks if he's got the same word in his list. If Bob can match at least one of the words he receives to his own wordlist he believes that it's a fair representation of Alices wordlist as well, and reports the Hypothesis to be *true*.

Pseudo Code

```
A = List of Alices words
B = List of Bobs words
  where |A| == |B|

n = The length of the lists (|A|)

for i -> 0 to (k*log n)
  random sample elem from A
  if elem exists in B
    return true
  end
end
return false
```

Theoretical Approach

The algorithm builds upon the idea that we are free to report anything if the number of common words x is $0 < x < \frac{1}{10} \cdot n$. As such, we will always report *true* if we get at least 1 common word. Since a *true* will only be given if a common word is found and a *false* is only reported if every sample element is unique, the algorithm will always report *false* if every element in the original lists are unique. The probability then to report *true* when both lists contains at least p parts common words would be:

$$P(X = True) = 1 - P(X = False) = 1 - (1 - p)^{k \cdot log_2(n)}$$

Where X in this case is the stochastic variable representing the algorithms output with a sample space $\Omega = true, false$.

It's easy to see that when $n \to \infty$ the probability we will answer YES when 10% of the words are equal will approach 1. The constant k can be used to increase that probability.