Recall key equivocation

\[ H(K \mid C^n) = H(K) + H(M^n) - H(C^n) \]

Generally, \( H(K) = \log(|K|) \) and \( H(C^n) \approx n \log(|C|) \)

What can we say about \( H(M^n) \)?
Entropy of a Natural Language

Can we estimate $H(M^n)$, where $M^n$ varies over $n$-grams of English language plaintext?

For any language $L$, we define

$$H_L = \lim_{n \to \infty} \frac{H(M^n)}{n}$$

if it exists.

★ The per-character entropy of the language $L$. 
Experimental results of Shannon (1951) on English plaintext.

\[
\begin{align*}
H(M^0) & \quad 4.70 \\
H(M^1) & \quad 4.14 \\
H(M^2)/2 & \quad 3.56 \\
H(M^3)/3 & \quad 3.30
\end{align*}
\]
Alternate approach:

**Theorem:** If $M_1, M_2, \ldots, M_n$ are successive characters from a language $L$ and if

$$\lim_{n \to \infty} H(M_n \mid M_1, M_2, \ldots, M_{n-1})$$

exists, then $H_L$ exists and is equal to this limit.

With $n = 15$, Shannon obtained $1.2 \leq H_E \leq 2.0$. 
One can prove that for a natural language $L$, $H_L$ always exists. Hence for sufficiently large $n$,

$$H(M^n) \approx n \cdot H_L.$$ 

For English, it is common to take $H_E \approx 1.5$. 

© 2011 Clifford Bergman
Aohtener ittrnesneig tihng aubot Elgsnih is taht as lnog as you lvaee the frsit and lsat leretts of ecah wrod alnoe, you can rgarnernae the rset of the lrettes and slitl raed the prrpgaaah.