CS/ENGRI 172, Fall 2002

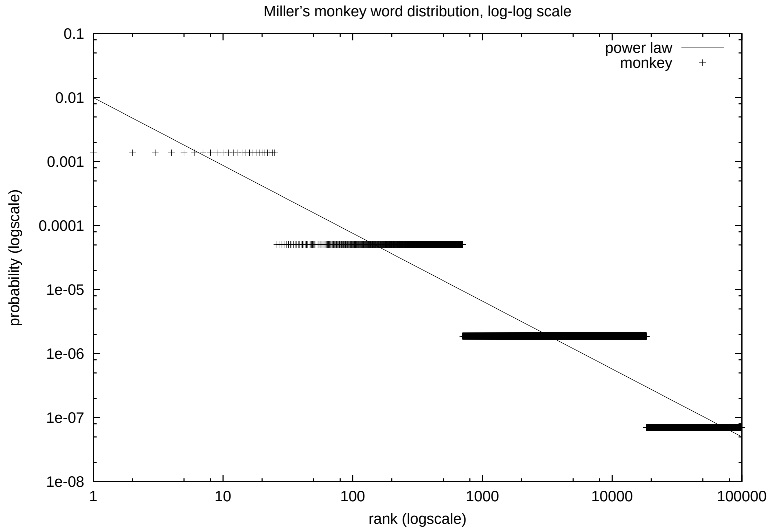
11/15/02: Lecture Thirty-Three Handout

Topics: Miller's monkey and Zipf's law; statistical authorship attribution.

Miller's monkey model

We adapt George A. Miller, "Some effects of intermittent silence", *The American Journal of Psychology* 70(2), pp. 311–314, 1957.

Assuming each of the 27 keys (26 letters plus the return key) has equal probability $p = 1/27$ of being struck, we get a probability of p^{i+1} assigned to each of the 26^i "words" of length i. We would therefore get the following distribution over words (shown on a log-log scale with a power-law (prob = $.01/(\text{rank}^{1.06})$) overlain for comparison):



The *average* probability function (not plotted explicity) appears to closely match the plotted power law; indeed, we can mathematically prove that it is a power law.

Authorship attribution and the Federalist Papers

This classic statistical study is described in Mosteller and Wallace's book *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* (Springer-Verlag 1984).

Eight of the Federalist Papers were signed by Hamilton; the remaining seventy-seven were written under a pseudonym. Of these, there is agreement that Jay wrote five, Hamilton 43, and Madison 14, with three written jointly, leaving 12 *disputed* papers.

An incomplete chronology:

1804	death of Hamilton
1807	"M"-signed list
1817	"Benson" list
1818	Madison claim

(OVER)

The test: the *log odds ratio*, which for a given document d containing indicator words $v_1, v_2, \ldots, v_n$ is defined as:

$$\begin{aligned}\log \frac{P_d(H)}{P_d(M)} &\approx \log \left[\frac{P_H(v_1) \cdots P_H(v_n)}{P_M(v_1) \cdots P_M(v_n)} \right] \\ &= \sum_{i=1}^{n} \log \frac{P_H(v_i)}{P_M(v_i)}\end{aligned}$$