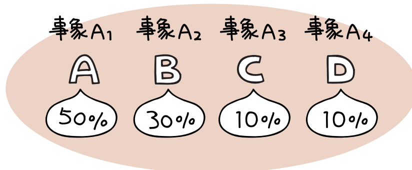
平均情報量で見るデータ量の理論値

実際の符号化方式の話へと入る前に、前節でやった平均情報量（エントロピー）で理論的な情報量の平均値を求めてみましょう。

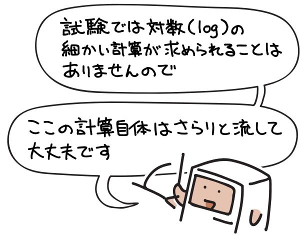


A 事象A_1の情報量 ⟶ $-\log_2 0.5 = 1$ビット

B 事象A_2の情報量 ⟶ $-\log_2 0.3 \fallingdotseq 1.74$ビット

C 事象A_3の情報量 ⟶ $-\log_2 0.1 \fallingdotseq 3.32$ビット

D 事象A_4の情報量 ⟶ $-\log_2 0.1 \fallingdotseq 3.32$ビット

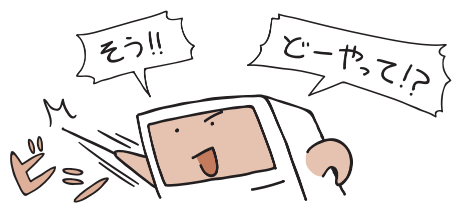


この場合の平均情報量は

$$\underbrace{(0.5\times 1)}_{A}+\underbrace{(0.3\times 1.74)}_{B}+\underbrace{(0.1\times 3.32)}_{C}+\underbrace{(0.1\times 3.32)}_{D}=1.686\text{ ビット}$$

単純に各文字2ビットずつを割り当てた場合、当然1文字あたりのデータ量は2ビットとなります。これは、ABCD各文字が均等に出現するとした場合の平均情報量（$-\log_2 0.25=2$ビット）と等しいものです。しかし、その出現頻度にバラつきがある場合は、個々の確率を元に情報量を算出することで、1文字あたりの平均的なデータ量を1.686ビットにまで圧縮できる可能性のあることがわかります。

その差は約16%。けっして小さい値ではありません。



…というところで、それでは実際の符号化方式を見てみましょう。