甘利俊一『情報理論』第１章 第１節 定義 1.3 の研究

2016.07.11 Takashi Aurues

ここでは、甘利先生の説明に沿って、また説明を補足しながら、情報量とは何か、エントロピーとは何かを読み解いていきたいと思います。なお、文庫版 023 ページの定義 1.3 は何らかの修正（または補足説明の追加）が必要だと思われます。甘利俊一『情報理論』：ちくま学芸文庫 2015 第 4 刷¥1,300+税、2011 に第 1 刷、初版はダイヤモンド社 1970

「**情報**とは何か？」に対して、「私たちの不確実な知識を確実にしてくれるものが情報である」と説明しています。

「**情報量**はどうやって測るか？」に対して、「知識の不確実性がどの程度減ったかをもって情報量を計る」と答えています。

つまり、情報と不確実性は、同じコインの表と裏の関係に置かれています。まだ数式で具体的には示されていませんが、「情報量」＝「不確実性の大きさ」となっています。

要するに、「情報」を真正面から捉えるよりも、裏から「不確実性」を捉える方が理解しやすいので、それを使って「情報」の量的変化を見ましょうと提案しています。これは、「船の大きさ（重さ）」を「排水量」で表すのと似たような考え方です。その情報が打ち消すことのできる不確実性の大きさで、その情報がもつ情報量を決めようというわけです。

ところで、「明日晴れる確率は 50%でしょう」という天気予報のように、出現が不確実な現象は、確率で出現の可能性が言い表されます。ある現象が起こるかどうか不確実だということは、その現象が確率的現象であるということです。

したがって、情報量は確率の関数として計算することができます。「情報量を表す関数 $f(\ \ )$ はどのような関数か？」という質問に対して、次のように幾つかの仮定を置きながら段階的に関数の形を求めています。

（１）まず n 個の排反事象（独立事象）が 等確率 で起こる場合を仮定

この場合、個々の事象の出現確率は等しく、全確率は１なので、確率 q と事象の数 n との関係は $q \times n = 1$ です。ゆえに $q = 1/n$ であり、$n = 1/q$ です。

確率的選択肢である事象の数 n が多いほど（＝確率 q が小さいほど）、どれが起こるかの予測は難しくなり、不確実性が大きく（＝情報量が大きく）なります。

情報量を表す関数は加法性（ここでの加法性は、積⇔和と変わる加法性）をもつと期待します。情報量を加法的に足し算・引き算で扱えた方が、計算が便利であり、また直感的に把握しやすいからです。

そこで、$n = mk\ (m, k \in \mathrm{N})$ と置くとき、$f(mk) = f(m) + f(k)$ となるように $f(\ \ )$ を決めます。

解析的手続きをとるために連続性を仮定し、微分の定義式を作り、積分して $\int f'(x)dx = c\log_e x + d$ を得ます。積⇔和と変わる加法性を持つただひとつの関数として対数が導き出されています。甘利先生は詳しく説明していますが、文系高校数学の範囲外なので、ここでは計算説明を省略します。

$x = 1$ のとき（ $n = 1$ のとき）、つまり 1 個の事象が 1 の確率で生じるとき、不確実性は「0」なので、$f(1) = 0$ 。故に $d = 0$ と決めます。

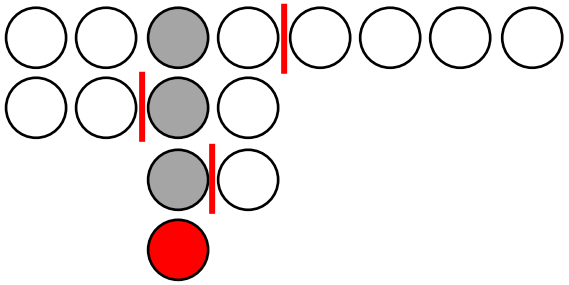
「等確率で二者択一のときの不確実性の大きさを 1 bit と決める」仮定を置くと、 $f(2) = 1$ 。故に $c = \log_2 e$ と決めます。bit（ビット）は二者択一型情報量の単位です。

以上、幾つかの仮定の下で $f(x) = \log_2 x$ bit と関数の形が決まりました。

全部で $\boldsymbol{n}$ 個から成る等確率排反事象（各確率 : $\boldsymbol{1/q}$）がもつ不確実性の大きさ（＝情報量）は $\boldsymbol{f(n) = \log_2 n\ \mathrm{bit} = f\left(\frac{1}{q}\right) = \log_2 \frac{1}{q} = -\log_2 q\ \mathrm{bit}}$ となります。

仮定したことは、「等確率の排反事象」、「積⇔和の加法性」、「二者択一のときの情報量を 1 bit」です。全確率は $\sum q_i = q_1 + q_2 + \cdots q_n = 1$ です。

等確率の簡単な例を紹介しています。 $n = 2^m$ 個の等確率排反事象があり、その中の 1 個が起こったとします。どれが起こったか分からないときの不確実性の大きさは $f(2^m) = \log_2 2^m = m$ bit と計算されます。



例えば、$n = 2^3 = 8$ のとき、8 個の事象を半分に分けて、起こった事象がどちらにあるか二者択一の質問に対する回答を得ることを 3 回繰り返すと、起こった事象を見つけることができます。8 個の事象が持つ不確実性の大きさが 3 bit とは、こういう意味を持っています。「二者択一 1 回分の不確実性＝1 bit」が 3 回分あるということです。

（2）「等確率でない場合」への一般化の準備（等確率という制限からの解放）

まず、出現確率 p がわかっている事象 A について、「事象 A が起こりました」という知らせが持つ情報量は幾らか、この知らせがどれだけの大きさの不確実性を消したのかを考えます。考え方はシンプルです。

$p = k/n$ で計算できるように、n 個の等確率事象を考え、そのうち k 個が事象 A であるとします。たとえば、$\mathrm{A} = \{E_1, E_2, \cdots, E_k\}$, $\mathrm{B} = \{E_{k+1}, E_{k+2}, \cdots, E_n\}$ を事象 A、事象 B とします。$p = k/n$ は事象 A が起こる確率です。そして、等確率の場合の不確実性の大きさの計算式 $f(x) = \log_2 x$ bit を使います。

次の順序で不確実性の変化を追ってみます。

[1]　最初、n 個全体が持つ不確実性の大きさは $f(n) = \log_2 n$ bit です。これが最初にある不確実性の大きさ（＝情報量）です。

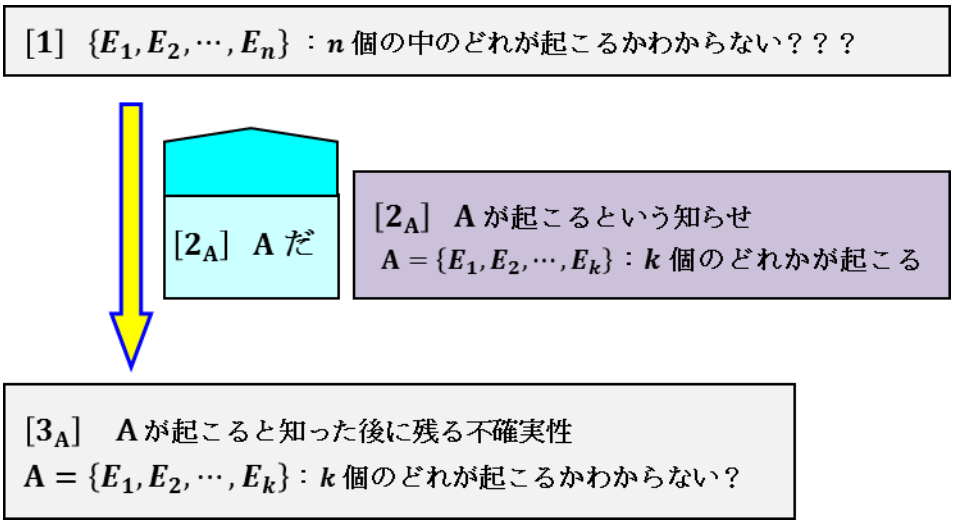
[2_A]　「A が起こる」という知らせが届きました。この知らせ「A が起こる」の情報量を I_A とすると、不確実性は $\log_2 n - I_\mathrm{A}$ bit に減ります。

[3_A]　しかし、「A が起こる」というだけでは、k 個あるどの E_i が起こるのか分かりません。したがって、残っている不確実性は $f(k) = \log_2 k$ bit です。

$\log_2 n - I_\mathrm{A} = \log_2 k$ より、

$$I_\mathrm{A} = \log_2 n - \log_2 k = \log_2 \frac{n}{k} = \log_2 \frac{1}{p} = -\log_2 p \text{ bit}$$

確率 p の「A が起こる」という知らせのもつ情報量は $I_\mathrm{A} = -\log_2 p$ bit となります。確率 p がわかるときは、$[2_\mathrm{A}] = [1] - [3_\mathrm{A}]$ で計算する必要はありません。確率 p の事象が起こるという知らせのもつ情報量は $-\log_2 p$ bit です。



この例は、確かに「知識の不確実性がどの程度減ったかをもって情報量を計る」という方法で情報量を計算しています。しかし実際には、「A が起こる」という情報を、先の例のように確定的に扱うときの情報量ではなく、次の例のように、「A が起こるか？ B が起こるか？」を書いた手紙のように、不確定的なものの情報量を評価する計算も必要です。「A が起こるか？ B が起こるか？」を、ひとまとめで扱うところがポイントです。

先ほどと同様に、順を追って不確実性の変化をみます。

[1]　最初 n 個のどれが起こるか分からない不確実性があるとします。等確率であることはわかっています。このとき、n 個全体が持つ不確実性の大きさは $f(n) = \log_2 n$ bit です。これが最初にある不確実性の大きさです。 n 個のどれが起こるかを当てるとガッポリ賞金 n 億円が手に入るとしましょう。

[2]　そこにマスクで口元を隠した怪しい女が手紙を売りに来ました。手紙には「A が起こるか？ B が起こるか？」が書いてあるとのこと。手紙を買って開けるまで何を書いてあるかはわかりません。「A が起こるか？ B が起こるか？」がわかれば、不確実性が減ります。手紙の情報量を計算し、それに応じて X 円で購入することにしました。手紙の購入価格は、手紙の情報量に見合った値段にすべきです。このとき、A が起こる確率は $p = {}^{k}/_{n}$ 、B が起こる確率は $1-p$ であることは事前に知っています。[2] の次は p の確率で $[2_{\mathrm{A}}]$ が起こり、$1-p$ の確率で $[2_{\mathrm{B}}]$ が起こります。手紙の情報量を計算するのに使えるデータは、これらの確率だけです。$[1]-[3]=[2]$ で求めることが可能かもしれません。

$[2_{\mathrm{A}}]$　手紙を開けると「A が起こる」と書いてありました。この知らせ「A が起こる」の情報量を I_{A} とすると、不確実性は $\log_2 n - I_{\mathrm{A}}$ bit に減ります。

$[3_{\mathrm{A}}]$　しかし、「A が起こる」というだけでは、k 個あるどの A_i が起こるのか分かりません。したがって、残っている不確実性は $f(k) = \log_2 k$ bit です。「A が起こる」という知らせの情報量は $I_{\mathrm{A}} = -\log_2 p$ bit 。

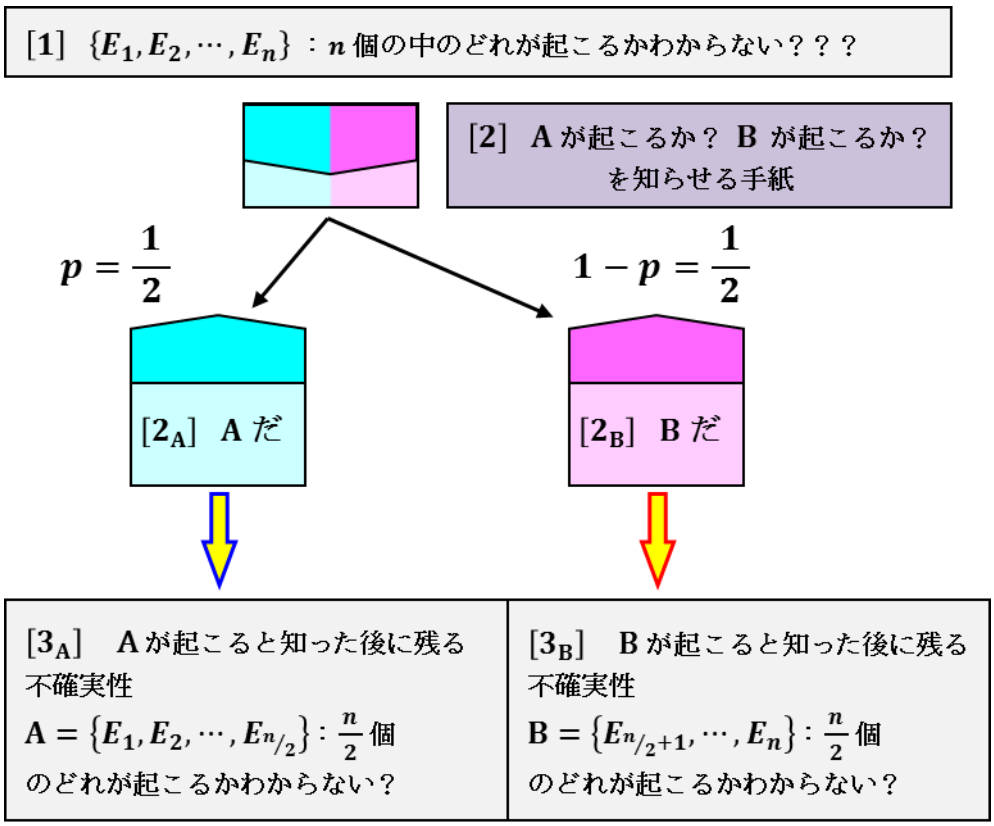
$[2_{\mathrm{B}}]$　手紙を開けると「B が起こる」と書いてありました。この知らせ「B が起こる」の情報量を I_{B} とすると、不確実性は $\log_2 n - I_{\mathrm{B}}$ bit に減ります。

$[3_{\mathrm{B}}]$　しかし、「B が起こる」というだけでは、$n-k$ 個あるどの A_i が起こるのか分かりません。したがって、$f(n-k) = \log_2(n-k)$ bit の不確実性が残っています。「B が起こる」という知らせの情報量は $I_{\mathrm{B}} = -\log_2(1-p)$ bit 。

[3]　手紙を購入し開けて「A が起こるか？ B が起こるか？」を知った後に残っている不確実性の大きさはいくらになるでしょうか。確率 p で起こる $[3_{\mathrm{A}}]$ 、確率 $1-p$ で起こる $[3_{\mathrm{B}}]$ を使って計算する必要があります。

情報量について「知識の不確実性がどの程度減ったかをもって情報量を計る」という立場から、$[1]-[3_A]=[2_A]$ 、$[1]-[3_B]=[2_B]$ 、$[1]-[3]=[2]$ という関係の成立が期待されます。

まず、A と B が等確率で起こる場合を検討してみます。



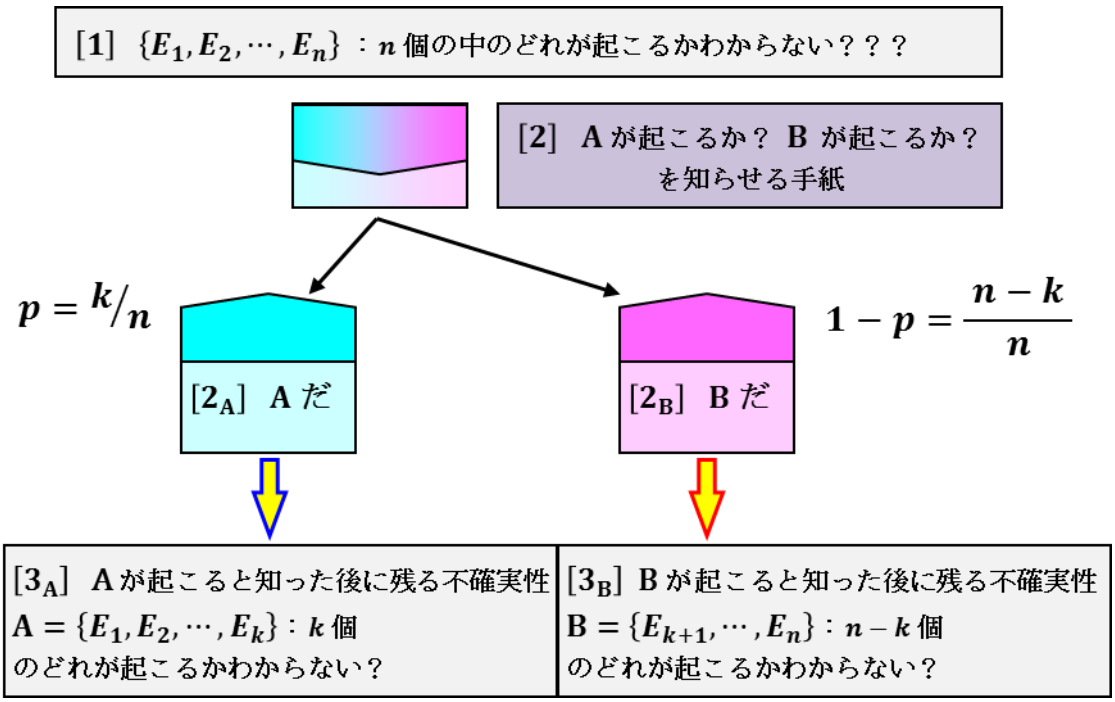
これは、A が起こる場合も、B が起こる場合も不確実性の変化は同じなので計算は簡単です。確率 $p=(1-p)=1/2$ なので $I_A=I_B=-\log_2(1/2)=1$ bit

したがって、封を開ける前の手紙の価値（情報量）も 1 bit であると考えて構わないでしょう。

では次に、A と B が等確率でない場合を検討します。

（3）「等確率でない場合」への加法性を保持した一般化の失敗 と 代替案

情報量の形式 $-\log p$ 、つまり「積⇔和の加法性」を保ったまま A と B が等確率でない場合に一般化することはできないようです。そこで、宝クジの期待値 $\sum(\text{確率}\times\text{賞金})$ を求めるのと同じ形式で情報量の確率的平均、つまり**平均情報量：** $\sum(\text{確率}\times\text{情報量})=p_1I_1+p_2I_2+p_3I_3+\cdots$ を扱うことになったようです。



[1]　n 個のどれが起こるか分からない？不確実性の大きさ＝情報量

$$I_1 = \log_2 n \text{ bit} \qquad = H_1$$

代替案 : H_1　「ある特定の E_i が起こる」確率は等しく $\frac{1}{n}$ であり、ひとつの知らせの情報量は $-\log_2 \frac{1}{n} = \log_2 n$ bit であるので、平均情報量 $\sum$(確率 × 情報量) は：

$$H_1 = \frac{1}{n}\log_2 n + \frac{1}{n}\log_2 n + \frac{1}{n}\log_2 n + \cdots = n \times \frac{1}{n} \times \log_2 n = \log_2 n \text{ bit} \qquad = I_1$$

[2]　「A が起こるか？ B が起こるか？」を知らせる手紙の情報量は幾らか？この手紙を開くことで消える不確実性の大きさは幾らか？

代替案 : H_2　確率 p の $[2_A]$ と、確率 $1-p$ の $[2_B]$ との確率的平均、つまり、「A が起こる」の情報量と「B が起こる」の情報量の確率的平均 $\sum$(確率 × 情報量) は：

$$H_2 = p \times I_{2A} + (1-p) \times I_{2B}$$

このとき $H_2 = H_1 - H_3$ となっているか？

$[2_A]$　「A が起こる」の情報量	$[2_B]$　「B が起こる」の情報量
$I_{2A} = -\log_2 p$ bit	$I_{2B} = -\log_2(1-p)$ bit

[3]　手紙を開いて「A が起こるか？ B が起こるか？」を知った後に残っている不確実性の大きさは幾らか？

代替案 : H_3　確率 p の $[3_A]$ と、確率 $1-p$ の $[3_B]$ との確率的平均は：

$$H_3 = p \times I_{3A} + (1-p) \times I_{3B}$$

$[3_A]$　「A が起こる」を知った後、k 個のどれが起こるかわからない？という不確実性の大きさが残る	$[3_B]$　「B が起こる」を知った後、$n-k$ 個のどれが起こるかわからない？という不確実性の大きさが残る
$I_{3A} = \log_2 k$ bit　　$= H_{3A}$	$I_{3B} = \log_2(n-k)$ bit　　$= H_{3B}$

「知識の不確実性（＝情報量）がどの程度減ったかをもって情報量を計る」という立場から、$[1]-[3_A]=[2_A]$ 、$[1]-[3_B]=[2_B]$ という関係が成立していますが、同じように、等確率でないときの代替案でも「知識の平均情報量がどの程度減ったかをもって平均情報量を計る」という立場から、$[1]-[3]=[2]$ という関係の成立が期待されます。

それでは、計算式を整理してから、幾つか具体例で計算してみましょう。

まず n 個の排反事象（独立事象）$\{E_1, E_2, \cdots, E_n\}$ が 等確率 で起こる場合、個々の事象の出現確率は等しく、全確率 $\sum q_i = q_1 + q_2 + \cdots q_n = 1$ であり、確率と事象の数との関係は $q \times n = 1$ 、$q = 1/n$ 、$n = 1/q$ となっています。

「n 個の中のどれが起こるか？わからない」という n 個全体が持つ不確実性の大きさは $f(n) = \log_2 n$ bit $= f\left(\frac{1}{q}\right) = \log_2 \frac{1}{q} = -\log_2 q$ bit です。

n 個の中の「特定の E_i が起こる」というどの知らせによっても、先ほどの不確実性 $\log_2 n$ bit はすべて消えますから、「特定の E_i が起こる」という知らせがもつ情報量（不確実性を消す能力量）は $\log_2 n = -\log_2 q$ bit です。ある知らせが示す確率的選択肢が、すべて等確率 q で起こる場合は、その知らせのもつ情報量は $-\log_2 q$ bit で計算できます。

次に、等確率でない場合は、平均情報量を計算します。各々の確率的選択肢の出現確率 p_i に、各選択肢がもつ情報量 I_i を掛けて、その総和をとります。

つまり、$\sum p_i I_i = p_1 I_1 + p_2 I_2 + p_3 I_3 + \cdots, (\sum p_i = 1)$ が平均情報量となります。

また、「出現確率 p_i の確率的選択肢が起こる」という知らせがもつ情報量は、$I_i = -\log_2 p_i$ bit なので、平均情報量は、

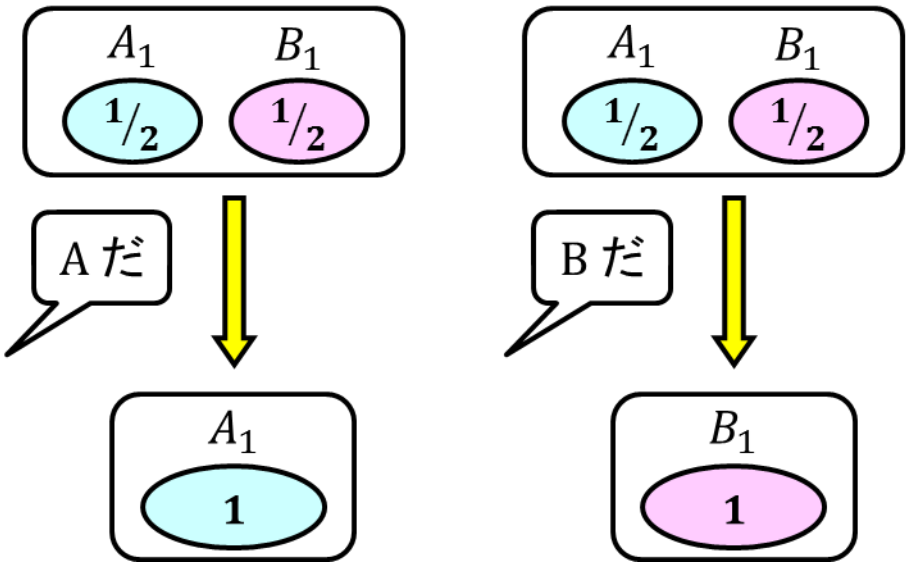
$$\sum p_i I_i = p_1 I_1 + p_2 I_2 + p_3 I_3 + \cdots = p_1(-\log p_1) + p_2(-\log p_2) + p_3(-\log p_3) + \cdots$$

$$= -p_1 \log p_1 - p_2 \log p_2 - p_3 \log p_3 - \cdots = -\sum p_i \log p_i = \sum p_i \log \frac{1}{p_i}$$

という式で計算できます。等確率 $q = 1/n$ の場合の平均情報量は、情報量と一致します。

$$-\sum q \log q = -\frac{1}{n}\log q - \frac{1}{n}\log q - \cdots - \frac{1}{n}\log q = -n \times \frac{1}{n}\log q = -\log q$$

では、具体例をいくつか紹介します。



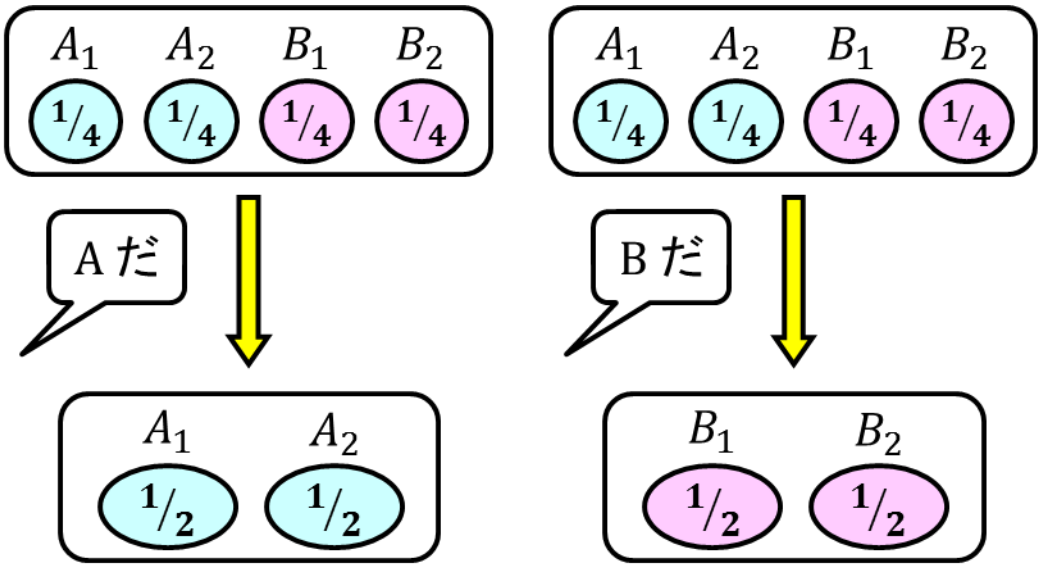
Case 1.｛A：1 個　B：1 個　等確率｝

[1]　A_1 or B_1 ??　：2 個のどれが起こるか分からない？？不確実性の大きさ＝情報量 $I_1 = \log 2 = 1$ bit 「A_1 が起こる」確率は $\frac{1}{2}$ であり、その知らせの情報量は $-\log_2 \frac{1}{2} = \log_2 2 = 1$ bit である。「B_1 が起こる」についても同じ。したがって「ある特定の E_i が起こる」という知らせがもつ情報量の確率的平均 $\sum$(確率 × 情報量) は： $H_1 = \frac{1}{2} \times \left(-\log \frac{1}{2}\right) + \frac{1}{2} \times \left(-\log \frac{1}{2}\right) = 2 \times \frac{1}{2} \times \left(-\log \frac{1}{2}\right) = 1$ bit　　$= I_1$	
[2]　A or B??　：$H_2 = \frac{1}{2} I_{2A} + \frac{1}{2} I_{2B} = 1$ bit　　$= H_1 - H_3$	
[2_A]　A：$I_{2A} = \log 2 = 1$ bit	[2_B]　B：$I_{2B} = \log 2 = 1$ bit
[3]　A or B　：　「A or B??」を知った後に残る不確実性 $H_3 = \frac{1}{2} I_{3A} + \frac{1}{2} I_{3B} = 0$ bit	
[3_A]　A_1　：　$I_{3A} = \log 1 = 0$ bit	[3_B]　B_1　：　$I_{3B} = \log 1 = 0$ bit

情報量について、$I_1 - I_{3A} = I_{2A}$，$I_1 - I_{3B} = I_{2B}$ という関係が成立し、平均情報量について、$H_1 - H_3 = H_2$ という関係が成立しています。

等確率なので、$H_2 = I_{2A} = I_{2B}$ となっています。

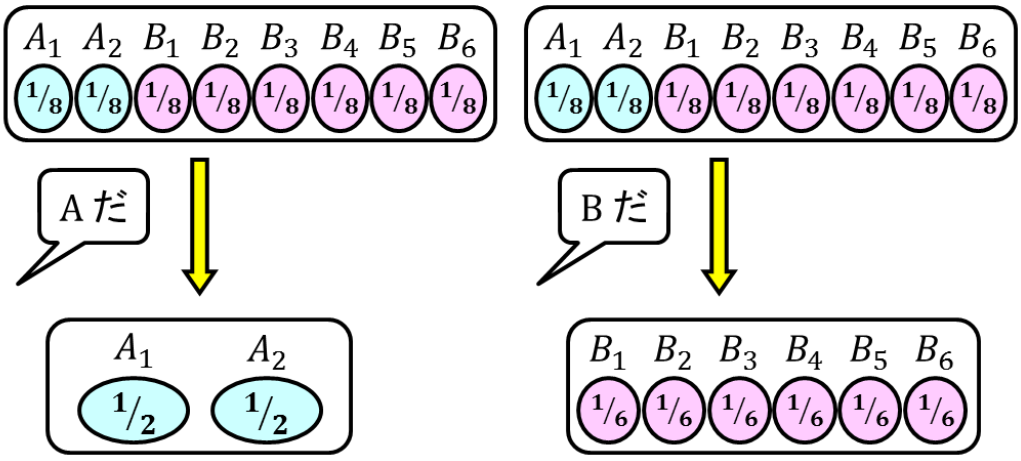
なお、誤解を生じない範囲で計算式 $\log(1/p) = -\log p$ はどちらかを適当に用いています。対数の底の 2 は省略しています。ビットの値は少数第 5 位以下切り捨てとしています。



Case 2.｛A：2 個　B：2 個　等確率｝

[1]　A_1 or A_2 or B_1 or B_2????　：4 個のどれが起こるか分からない？？？？ $I_1 = \log 4 = 2$ bit $H_1 = 4 \times \frac{1}{4} \times \left(-\log \frac{1}{4}\right) = 2$ bit	
[2]　A or B??　： $H_2 = \frac{1}{2} I_{2A} + \frac{1}{2} I_{2B} = 1$ bit　　$= H_1 - H_3$	
[2_A]　A： $I_{2A} = \log 2 = 1$ bit	[2_B]　B： $I_{2B} = \log 2 = 1$ bit
[3]　A or B　：　「A or B??」を知った後に残る不確実性 $H_3 = \frac{1}{2} I_{3A} + \frac{1}{2} I_{3B} = 1$ bit	
[3_A]　A_1 or A_2??　： $I_{3A} = \log 2 = 1$ bit	[3_B]　B_1 *or* B_2??　： $I_{3B} = \log 2 = 1$ bit

等確率なので、$H_2 = I_{2A} = I_{2B}$ となっています。



Case 3.｛A：2 個　B：6 個　等確率｝

[1]　8 個のどれが起こるか分からない？そういう時の不確実性の大きさ＝情報量

$$I_1 = \log 8 = 3 \text{ bit}$$

A_1～B_6 の中の「ある特定の E_i が起こる」という知らせのもつ情報量の確率的平均

$$H_1 = 8 \times \left(-\frac{1}{8}\log\frac{1}{8}\right) = 3 \text{ bit}$$

[2]　「A が起こるか？ B が起こるか？」、「 A が起こる」の情報量と「B が起こる」の情報量の確率的平均

$$H_2 = \frac{2}{8} I_{2\mathrm{A}} + \frac{6}{8} I_{2\mathrm{B}} = 0.8112 \text{ bit} \qquad = H_1 - H_3$$

[2_A]　「 A が起こる」の情報量

$$\mathrm{A}：I_{2\mathrm{A}} = -\log(2/8) = 2 \text{ bit}$$

[2_B]　「B が起こる」の情報量

$$\mathrm{B}：I_{2\mathrm{B}} = -\log(6/8) = 0.415 \text{ bit}$$

[3]　「A が起こるか？ B が起こるか？」を知った後に残っている不確実性の大きさの確率的平均：A_1～B_6 のどれが起こるのか？という不確実性の平均的な大きさ

$$H_3 = \frac{2}{8} I_{3\mathrm{A}} + \frac{6}{8} I_{3\mathrm{B}} = 2.1887 \text{ bit}$$

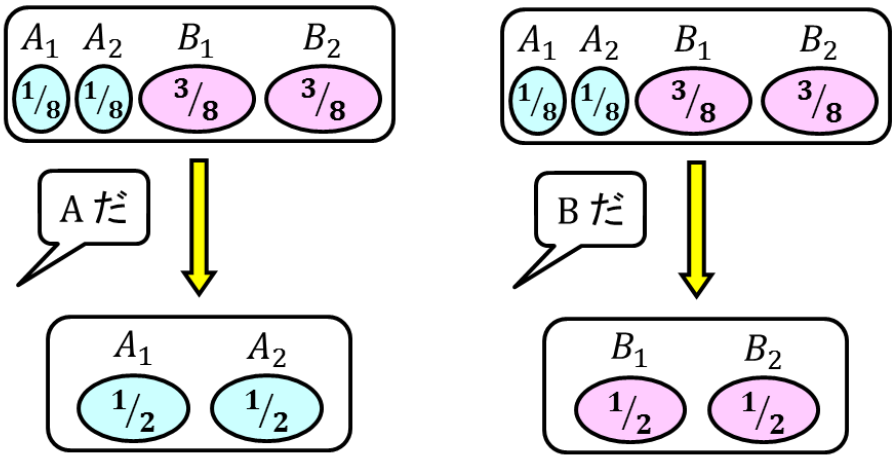
[3_A]　A が起こると分かった後に残る不確実性の大きさ：A_1～A_2 のどれが起こるのか？

$$\mathrm{A}：I_{3\mathrm{A}} = \log 2 = 1 \text{ bit}$$

[3_B]　B が起こると分かった後に残る不確実性の大きさ：B_1～B_6 のどれが起こるのか？

$$\mathrm{B}：I_{3\mathrm{B}} = \log 6 = 2.5849 \text{ bit}$$

等確率ではないので、$H_2 \neq I_{2\mathrm{A}} \neq I_{2\mathrm{B}} \neq H_2$ となっています。平均情報量について、$H_1 - H_3 = H_2$ という関係が成立しています。



Case 4.｛A：2 個　B：2 個　Non-Equiprobable｝

[1]　4 個のどれが起こるか分からない？そういう時の不確実性の大きさ＝情報量

$$I_1? = \log 4? = 2? \text{ bit}$$　等確率でないので計算対象外

A_1～B_2 の中の「ある特定の E_i が起こる」という知らせのもつ情報量の確率的平均

$$H_1 = -2 \times \frac{1}{8}\log\frac{1}{8} - 2 \times \frac{3}{8}\log\frac{3}{8} = 1.8112 \text{ bit}$$

[2]　「A が起こるか？ B が起こるか？」、「 A が起こる」の情報量と「B が起こる」の情報量の確率的平均

$$H_2 = \frac{2}{8}I_{2\mathrm{A}} + \frac{6}{8}I_{2\mathrm{B}} = 0.8112 \text{ bit} \qquad = H_1 - H_3$$

[2_A]　「 A が起こる」の情報量

$$\mathrm{A}：I_{2\mathrm{A}} = -\log(2/8) = 2 \text{ bit}$$

[2_B]　「B が起こる」の情報量

$$\mathrm{B}：I_{2\mathrm{B}} = -\log(6/8) = 0.415 \text{ bit}$$

[3]　「A が起こるか？ B が起こるか？」を知った後に残っている不確実性の大きさの確率的平均：A_1～B_2 のどれが起こるのか？という不確実性の平均的な大きさ

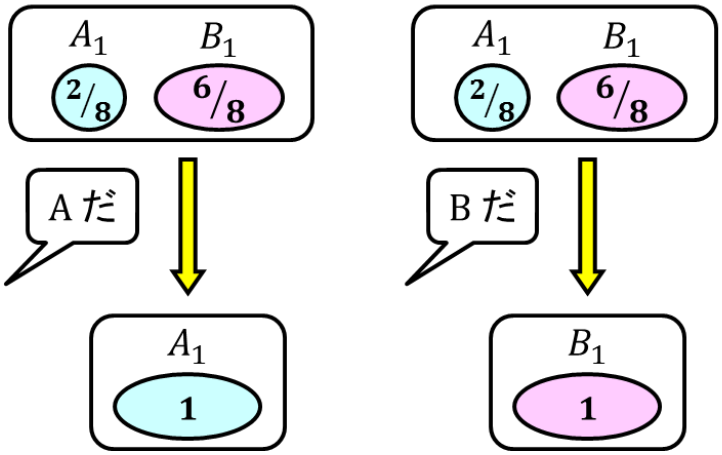
$$H_3 = \frac{2}{8}I_{3\mathrm{A}} + \frac{6}{8}I_{3\mathrm{B}} = 1 \text{ bit}$$

[3_A]　A が起こると分かった後に残る不確実性の大きさ：A_1～A_2 のどれが起こるのか？

$$\mathrm{A}：I_{3\mathrm{A}} = \log 2 = 1 \text{ bit}$$

[3_B]　B が起こると分かった後に残る不確実性の大きさ：B_1～B_2 のどれが起こるのか？

$$\mathrm{B}：I_{3\mathrm{B}} = \log 2 = 1 \text{ bit}$$



Case 5.｛A：1 個　B：1 個　Non-Equiprobable｝

[1]　2 個のどれが起こるか分からない？そういう時の不確実性の大きさ＝情報量

$$I_1? = \log 2? = 1?\ \text{bit}$$　等確率でないので計算対象外

A_1～B_1 の中の「ある特定の E_i が起こる」という知らせのもつ情報量の確率的平均

$$H_1 = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8} = 0.8112\ \text{bit} \quad S_{\text{AB}} = 0.8112\ \text{bit}$$

[2]　「A が起こるか？ B が起こるか？」、「 A が起こる」の情報量と「B が起こる」の情報量の確率的平均

$$H_2 = \frac{2}{8}I_{2\text{A}} + \frac{6}{8}I_{2\text{B}} = 0.8112\ \text{bit} \qquad = H_1 - H_3$$

[2_{A}]　「 A が起こる」の情報量

$$\text{A}：I_{2\text{A}} = -\log(2/8) = 2\ \text{bit}$$

[2_{B}]　「B が起こる」の情報量

$$\text{B}：I_{2\text{B}} = -\log(6/8) = 0.4150\ \text{bit}$$

[3]　「A が起こるか？ B が起こるか？」を知った後に残っている不確実性の大きさの確率的平均：A_1～B_1 のどれが起こるのか？という不確実性の平均的な大きさ

$$H_3 = \frac{2}{8}I_{3\text{A}} + \frac{6}{8}I_{3\text{B}} = 0\ \text{bit}$$

[3_{A}]　A が起こると分かった後に残る不確実性の大きさ：A_1 が起こる

$$\text{A}：I_{3\text{A}} = \log 1 = 0\ \text{bit}$$

$S_{\text{A}} = 0.5\ \text{bit}$

[3_{B}]　B が起こると分かった後に残る不確実性の大きさ：B_1 が起こる

$$\text{B}：I_{3\text{B}} = \log 1 = 0\ \text{bit}$$

$S_{\text{B}} = 0.3112\ \text{bit}$

参考：状況変化の各段階で確率的正規化（全確率を 1 にリセットすること）をしないで、[3_{A}] における A_1 の確率を 2/8、[3_{B}] における B_1 の確率を 6/8 のまま計算したエントロピーは、$S_{\text{AB}} = S_{\text{A}} + S_{\text{B}}$ という加法性が保たれています。シャノンは、関数がこの加法性を持つようにエントロピーの計算式を求め $-\sum p_i \log p_i$ を得ました。

Case3～5 を比較すると、平均情報量の計算式は、宝クジの期待値を求める式と形は似ていますが、性質がかなり異なっていることがわかります。

宝クジの計算式 $\sum(\text{確率} \times \text{賞金})$ には、「期待値の加法性」と呼ばれている性質があります。平均情報量には無い性質です。

たとえば{3/10 の確率で 300 円が当たる宝クジ}の期待値は 90 円ですが、{2/10 の確率で 300 円が当たる宝クジ} の期待値は 60 円、{1/10 の確率で 300 円が当たる宝クジ} の期待値は 30 円、{3/10 の確率で 200 円が当たる宝クジ} の期待値は 60 円、{3/10 の確率で 100 円が当たる宝クジ} の期待値は 30 円であり、確率や賞金を分割しても加法性が保たれています。

「期待値の加法性」があるので、賞金を分割しても期待値は変わりません。

しかし、情報量は確率の関数なので、宝クジの期待値の式で賞金に相当する部分に情報量を置いている平均情報量の式では、宝クジのときのような期待値の加法性がありません。

そのため、Case3～5 で H_1 が異なっています。

この、平均情報量のことをエントロピー（特に、シャノンの情報エントロピー）と呼びます。

平均情報量は、等確率のときは情報量と一致します。ここで次のような問題提議をします。「情報量と平均情報量とをこのまま区別して扱い続けるべきか」それとも「平均情報量をもって、それを情報量であると再定義し、これまでの対数形式の情報量を等確率の場合の特殊形として扱うべきか」という問題です。

では、より一般化された状態でエントロピーの計算式を求めてみましょう。方法はこれまでと同じです。より一般化された状態で $H_2 = H_1 - H_3$ を計算します。

次図は、n 個の等確率排反事象 $\{E_1, E_2, \cdots, E_n\}$ が、m 種類に分かれている様子を示しています。

$\mathrm{A}_{k1} = \{E_1, E_2, \cdots, E_{k1}\}$ は k_1 個の事象から成り、その出現確率は $p_1 = k_1/n$ です。

$\mathrm{B}_{k2} = \{E_{k1+1}, E_{k1+2}, \cdots, E_{k1+k2}\}$ は k_2 個の事象から成り、その出現確率は $p_2 = k_2/n$ です。

各々 k_1 個、k_2 個、k_3 個、・・・k_m 個から成る、全部で m 種類の事象を考えます。全確率は $\sum p_i = 1$ です。

[1] $\{E_1, E_2, \cdots, E_n\}$ ：n 個の中のどれが起こるかわからない？？？

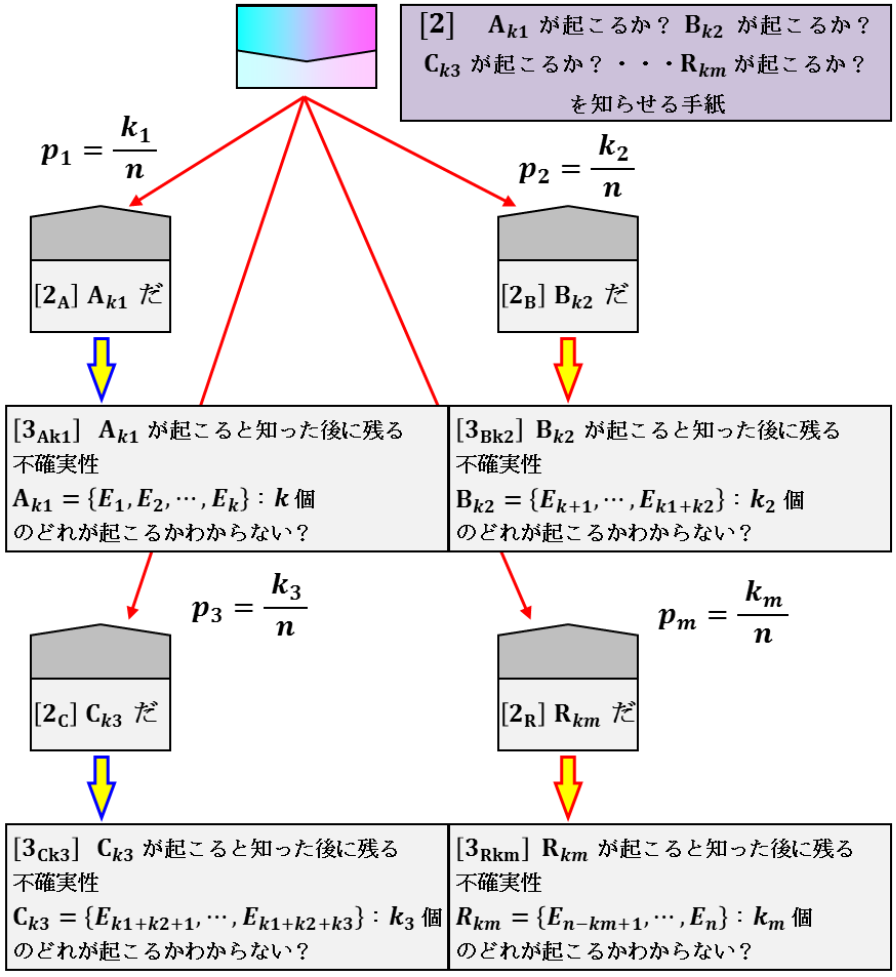
$$\mathrm{A}_{k1} = \{E_1, E_2, \cdots, E_{k1}\} \quad : \quad k_1 \text{ 個}$$

$$\mathrm{B}_{k2} = \{E_{k1+1}, E_{k1+2}, \cdots, E_{k1+k2}\} \quad : \quad k_2 \text{ 個} \qquad \sum_{i=1}^{m} k_i = n$$

$$\mathrm{C}_{k3} = \{E_{k1+k2+1}, E_{k1+k2+2}, \cdots, E_{k1+k2+k3}\} \quad : \quad k_3 \text{ 個}$$

$$\vdots$$

$$\mathrm{R}_{km} = \{E_{n-km+1}, E_{n-km+2}, \cdots, E_n\} \quad : \quad k_m \text{ 個}$$



$$H_1 = \log n$$

$$H_3 = p_1 \log k_1 + p_2 \log k_2 + \cdots + p_m \log k_m$$

$p_i = k_i/n$ 、$k_i = np_i$ を用いて上式を変形すると

$$H_3 = \frac{k_1}{n} \log np_1 + \frac{k_2}{n} \log np_2 + \cdots + \frac{k_m}{n} \log np_m$$

$$= \frac{k_1}{n}(\log n + \log p_1) + \frac{k_2}{n}(\log n + \log p_2) + \cdots + \frac{k_m}{n}(\log n + \log p_m)$$

$$= \frac{k_1 + k_2 + \cdots + k_m}{n}\log n + \frac{k_1}{n}\log p_1 + \frac{k_2}{n}\log p_2 + \cdots + \frac{k_2}{n}\log p_2$$

$$= \log n + p_1 \log p_1 + p_2 \log p_2 + \cdots + p_m \log p_m$$

$$= \log n + \sum p_i \log p_i$$

「知識の平均情報量がどの程度減ったかをもって平均情報量を計る」という立場から H_2 を求めます。

$$H_2 = H_1 - H_3 = \log n - \log n - \sum p_i \log p_i = -\sum p_i \log p_i$$

$$H_2 = -\sum p_i \log p_i = \sum p_i \log \frac{1}{p_i}$$

これは平均情報量：$\sum(\text{確率} \times \text{情報量}) = p_1 I_1 + p_2 I_2 + p_3 I_3 + \cdots$ の式そのものです。

確率 $p_1, p_2, p_3, \cdots, p_m$ がわかっているとき、情報量は $I_i = -\log p_i$ なので、「知識の平均情報量がどの程度減ったかをもって平均情報量を計る」という間接的な方法ではなく、平均情報量 $\sum(\text{確率} \times \text{情報量})$ を直接計算して求めることができます。

一般的に、全部で n 個の独立事象 $E_1, E_2, E_3, \cdots, E_n$ があって、各々の出現確率が p_i のとき、特定の事象が持つ情報量は $-\log_2 p_i$ bit ですが、この事象全体の平均情報量（エントロピー、不確実性の平均的な大きさ）を

$$-\sum p_i \log_2 p_i \text{ bit}$$

とします。

平均情報量（エントロピー）には、「積⇔和の加法性」がありません。特殊な状況、つまり n 個の事象がすべて等確率 $p = 1/n$ の時には、そのエントロピーが

$$-\sum p_i \log_2 p_i = -n \times \frac{1}{n}\log_2 p = -\log_2 p$$

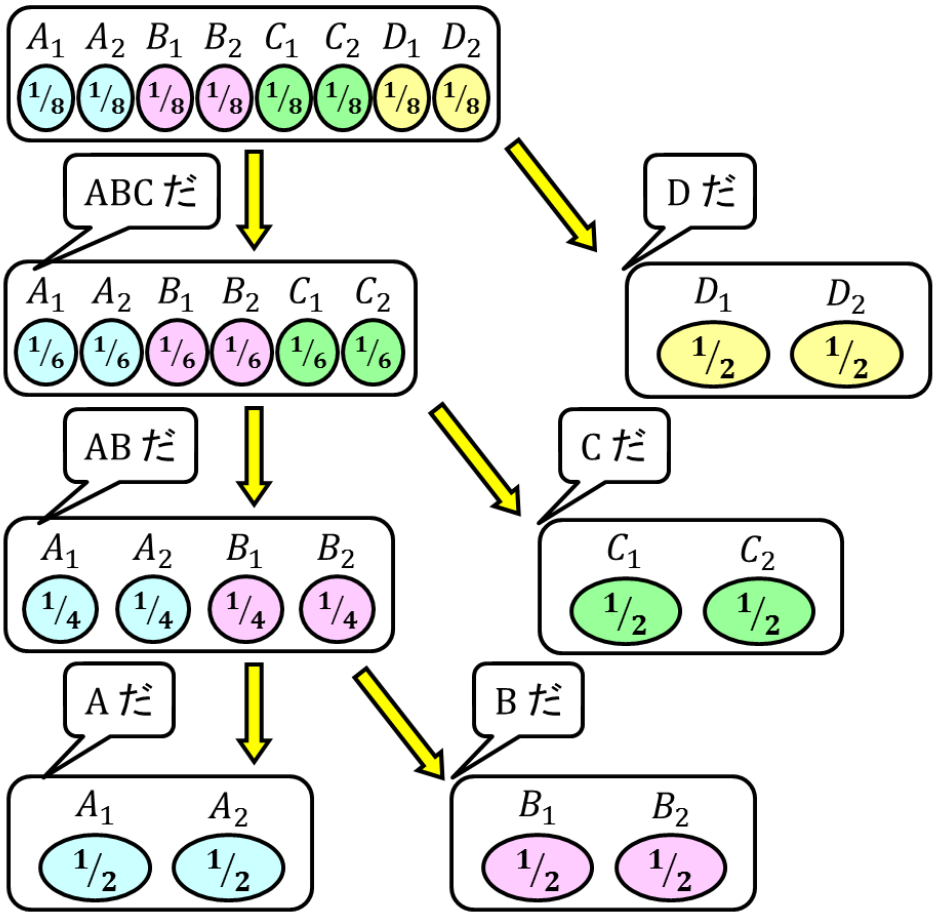
と、情報量と同じ数学的形式になるので「積⇔和の加法性」があります。

統計力学では分子運動論の立場から、熱力学的エントロピーを、$S = k_B \log_e W$ と解釈します。W は、平衡状態において系がとり得る状態の数ですが、この式においては様々な状態が出現する確率は等しいと仮定しています（等重率の原理、等確率の原理）。

熱力学のエントロピーが、本質的には情報エントロピーに由来するとしても、等確率を前提としているので情報量と同じく単純な対数の形式になっています。

では、ここで、確率的状況が段階的に明らかになっていく例をひとつ紹介します。

Case 6.｛A：2 個　B：2 個　C：2 個　D：2 個　等確率｝



参考：上図は、状況変化の各段階で確率的正規化（全確率を 1 にリセットすること）をしたものですが、最初の確率のまま（1/8 のまま）各段階のエントロピーを計算したものを S_i として示しておきます。$S_{ABCD} = S_{ABC} + S_D = S_{AB} + S_C + S_D = S_A + S_B + S_C + S_D$ という加法性が保たれています。

[1]　8 個のどれが起こるか分からない？そういう時の不確実性の大きさ＝情報量

$$I_1 = \log 8 = 3 \text{ bit}$$

$A_1 \sim D_2$ の中の「ある特定の E_i が起こる」という知らせのもつ情報量の確率的平均

$$H_1 = 8 \times \left(-\frac{1}{8}\log\frac{1}{8}\right) = 3 \text{ bit} \quad \boldsymbol{S_{ABCD} = 3 \text{ bit}}$$

[2]　「ABC が起こるか？ D が起こるか？」、「 ABC が起こる」の情報量と「D が起こる」の情報量の確率的平均　（等確率でないので情報量は計算できない）

$$H_2 = \frac{6}{8} I_{2\text{ABC}} + \frac{2}{8} I_{2\text{D}} = 0.8112 \text{ bit} \qquad = H_1 - H_3$$

[2_{ABC}]　「 ABC が起こる」の情報量

$$\text{ABC} : I_{2\text{ABC}} = -\log(6/8) = 0.4150 \text{ bit}$$

[2_{D}]　「D が起こる」の情報量

$$\text{D} : I_{2\text{D}} = -\log(2/8) = 2 \text{ bit}$$

[3]　「ABC が起こるか？ D が起こるか？」を知った後に残っている不確実性の大きさの確率的平均：$A_1 \sim D_2$ のどれが起こるのか？という不確実性の平均的な大きさ

$$H_3 = \frac{6}{8} I_{3\text{ABC}} + \frac{2}{8} I_{3\text{D}} = 2.1887 \text{ bit}$$

[3_{ABC}]　ABC が起こると分かった後に残る不確実性の大きさ：　6 個のどれが起こるか分からない？

$$\text{ABC} : I_{3\text{ABC}} = \log 6 = 2.5849 \text{ bit}$$

$H_{3\text{ABC}} = 2.5849 \text{ bit}$　$\boldsymbol{S_{\text{ABC}} = 2.25 \text{ bit}}$

[3_{D}]　D が起こると分かった後に残る不確実性の大きさ：　2 個のどれが起こるか分からない？

$$\text{B} : I_{3\text{D}} = \log 2 = 1 \text{ bit}$$

$\boldsymbol{S_{\text{D}} = 0.75 \text{ bit}}$

[4]　「AB が起こるか？ C が起こるか？」、「 AB が起こる」の情報量と「C が起こる」の情報量の確率的平均　（等確率でないので情報量は計算できない）

$$H_4 = \frac{4}{6} I_{4\text{AB}} + \frac{2}{6} I_{4\text{C}} = 0.9182 \text{ bit} \qquad = H_{3\text{ABC}} - H_5$$

[4_{AB}]　「 AB が起こる」の情報量

$$\text{AB} : I_{4\text{AB}} = -\log(4/6) = 0.5849 \text{ bit}$$

[4_{C}]　「C が起こる」の情報量

$$\text{C} : I_{4\text{C}} = -\log(2/6) = 1.5849 \text{ bit}$$

[5]　「AB が起こるか？ C が起こるか？」を知った後に残っている不確実性の大きさの確率的平均：$A_1 \sim C_2$ のどれが起こるのか？という不確実性の平均的な大きさ

$$H_3 = \frac{4}{6} I_{6\text{AB}} + \frac{2}{6} I_{6\text{C}} = 1.6666 \text{ bit}$$

[6_{AB}]　AB が起こると分かった後に残る不確実性の大きさ：　4 個のどれが起こるか分からない？

$$\text{AB} : I_{6\text{AB}} = \log 4 = 2 \text{ bit}$$

$H_{6\text{AB}} = 2 \text{ bit}$　$\boldsymbol{S_{\text{AB}} = 1.5 \text{ bit}}$

[6_{C}]　C が起こると分かった後に残る不確実性の大きさ：　2 個のどれが起こるか分からない？

$$\text{C} : I_{6\text{C}} = \log 2 = 1 \text{ bit}$$

$\boldsymbol{S_{\text{C}} = 0.75 \text{ bit}}$

[7]　「A が起こるか？ B が起こるか？」、「 A が起こる」の情報量と「B が起こる」の情報量の確率的平均　（★すべて等確率なので情報量を計算できる）

$$H_7 = \frac{2}{4} I_{7\mathrm{A}} + \frac{2}{4} I_{7\mathrm{B}} = 1 \text{ bit} \qquad = H_{6\mathrm{AB}} - H_9$$

情報量：$I_7 = H_7$

[8_A]　「 A が起こる」の情報量	[8_B]　「B が起こる」の情報量
$\mathrm{A} : I_{8\mathrm{A}} = -\log(2/4) = 1 \text{ bit}$	$\mathrm{B} : I_{8\mathrm{B}} = -\log(2/4) = 1 \text{ bit}$

[9]　「A が起こるか？ B が起こるか？」を知った後に残っている不確実性の大きさの確率的平均：$A_1 \sim B_2$ のどれが起こるのか？という不確実性の平均的な大きさ

$$H_3 = \frac{2}{4} I_{9\mathrm{A}} + \frac{2}{4} I_{9\mathrm{B}} = 1 \text{ bit}$$

[9_A]　A が起こると分かった後に残る不確実性の大きさ：　2 個のどれが起こるか分からない？	[9_B]　B が起こると分かった後に残る不確実性の大きさ：　2 個のどれが起こるか分からない？
$\mathrm{A} : I_{9\mathrm{A}} = \log 2 = 1 \text{ bit}$ $\boldsymbol{S_\mathrm{A} = 0.75 \text{ bit}}$	$\mathrm{B} : I_{9\mathrm{B}} = \log 2 = 1 \text{ bit}$ $\boldsymbol{S_\mathrm{A} = 0.75 \text{ bit}}$

●　定義 1.3 の解釈上の問題について

さて、甘利先生の文庫版 022－023 ページに、次のような定義が置かれています（注意：甘利先生の説明では、情報量（$-\log p$）の意味でも、平均情報量（$-\sum p_i \log p_i$）の意味でも、区別しないで同じ記号 I を用いているので、どちらを指すかは文脈で判断する必要があります）。

エントロピーがわかったので，これを用いると情報量の定義が一般の場合にすっきりとできる．すなわち，情報とは状況の不確定度を減少させるものであり，その量は不確定度の減少分で計ればよい．だから次のように定義できる．

定義 1.3　情報を得ることによって，状況のエントロピーが H から H' へ変わるとき，この情報のもつ情報量を $I = H - H'$ とする．

この定義によれば，さきほどまでの話はすっきりとする．たとえば，n 個の事象のエントロピーを H とし，どれが起こったかを教えてくれる情報を考える．どれが起こったかがわかれば，不確定度は 0 となるので $H' = 0$ である．だからこの情報の情報量は $I = H - H' = H$　まさしくエントロピー自身である．

定義 1.3 の「情報量：$\boldsymbol{I}$」は「$-\sum p_i \log p_i$」の形式でなければなりません。したがって、定義 1.3 は 2 種類の解釈が可能になります。

解釈 1	定義 1.3 は、情報量を定義し直しているのである。等確率を対象とした特殊なもの（$-\log p$）から、等確率でない場合も対象にできるもの（$-\sum p_i \log p_i$）へ、情報量の定義を一般化しているのである。つまり、今後はエントロピー（$-\sum p_i \log p_i$）の変化量をもって情報量（$-\sum p_i \log p_i$）の測定だと見なしますよという宣言である。したがって、定義 1.3 以降は、「エントロピー＝情報量」であり、平均情報量、情報量の平均値（期待値）といった呼び方はしない。
解釈 2	甘利先生の説明では、定義 1.3 以降も、情報量（$-\log p$）と平均情報量（$-\sum p_i \log p_i$）とが明確に区別されて用いられており、定義 1.3 にある情報量は平均情報量の間違いである。

まず、これからの説明に用いる用語についての注意です。混乱を避けるために、情報量（$-\log p$）、情報量（$-\sum p_i \log p_i$）、平均情報量（$-\sum p_i \log p_i$）、エントロピー（$-\sum p_i \log p_i$）のように、必要に応じて数式を明示します。

●　まず、「解釈 1：情報量の定義についての新たな宣言」である可能性について検討します。

最初に、等確率を仮定し、情報量の関数を求めて対数の形としましたが（初期型情報量：いわば情報量の第 1 試作品）、これは等確率でないときへの単純な拡張、つまり「$-\log p$」という同一の数学的形式を保って一般化することには失敗しました。

そこで、積⇔和の加法性は失われますが、代替として情報量（$-\log p$）の確率的平均（期待値：$-\sum p_i \log p_i$）を用いるようになりました。一般化の流れから、平均情報量と呼ばれることになりますが、これがシャノンの情報エントロピーのことです。

そして今度は、逆に、このエントロピー（$-\sum p_i \log p_i$）を用いて**情報量**を改めて再定義しよう（情報量の完成品）というのが定義 1.3 であると解釈できます。

したがって、甘利先生は、「エントロピーの変化量で情報量（の変化量）を定義します。したがって、定義 1.3 以降は、情報量という語で、まず $-\sum p_i \log p_i$ をイメージしてください。初期型情報量（$-\log p$）は情報量（$-\sum p_i \log p_i$）の特殊なタイプにすぎません」と、定義 1.3 で宣言したのです。

ところで、そのように情報量の定義を変更することで何か不都合なことはあるでしょうか。

もともと等確率の時は、情報量（$-\log p$）＝平均情報量（$-\sum p_i \log p_i$）であり、等確率でないときは平均情報量（$-\sum p_i \log p_i$）しか計算できないので、定義 1.3 によって不都合なことは何も生じないように期待できます。定義 1.3 は、「$-\log p$」という情報量の定義を「$-\sum p_i \log p_i$」に変え、等確率でない場合に一般化したに過ぎないものです。「積⇔和の加法性」など失われた性質もあるので、取扱い上の注意は必要です。

ちなみに、定義 1.3 のルーツ（最初に言い出した人、書物、論文）は不明です。

クロード・E・シャノン（Claude.E.Shannon）、ワレン・ウィーバー（Warren Weaver）「通信の数学的理論」（植村友彦訳、ちくま学芸文庫 2009）と原書「The Mathematical Theory of Communication」（1949）での英語表現を見比べながら、ザーっとチェックしました。（いつか読みたいと思って持っているだけの本です）

ウィーバーによる解説では、等確率時の $-\log p$ 形式の情報量から、平均情報量という説明なしに、一気に $-\sum p_i \log p_i$ 形式の情報量に移行して「エントロピー＝情報量」の意味で、説明が行われています。

・・・, then the actual expression for the information is $H = -[p_1 \log p_1 + p_2 \log p_2 + \cdots + p_n \log p_n]$, $H = -\sum p_i \log p_i$.

● では次に、解釈 2 を検討します。

甘利先生の説明文中の「この情報のもつ情報量を $I = H - H'$ とする」と「だからこの情報の情報量は」の部分は、「平均情報量」の間違いかもしれないと解釈可能です。

読者は、最初は解釈 1 で理解するかもしれません。しかし、その後に続く説明を読んでいて、解釈 1 ではなく解釈 2 なのかもと疑念を抱き困惑する可能性があります。

それは、定義 1.3 の後も、情報量（$-\log p$）と平均情報量（$-\sum p_i \log p_i$）を明確に区別した説明が続くからです。

文庫版 023 ページの定義 1.3 の後、文庫版 032 ページ（条件付エントロピー）には、次のように、「情報量の平均値」、「不確定度の期待値」であると明示されています。

検温によって得られる風邪についての情報量の平均値は，エントロピーの減少分

$$I = H(B) - H_A(B)$$
$$= 0.36 \text{ ビット}$$

である．$H_A(B)$ は $H_{A_1}(B)$ と $H_{A_1}(B)$ の平均値であり，検温の結果が何であるかは問わないで，とにかく検温をしたときの，検温だけでは確定できない残りの不確定度の期待値を表すことを，いま一度注意しておこう．

もし定義 1.3 が解釈 1 であるならば、定義 1.3 以降は、「エントロピー＝情報量＝不確定度」と呼ばれることになるはずです。もう「平均情報量」、「情報量の平均値」、「不確定度の期待値」などの呼び方は不要になるはずです。

それにも関わらず甘利先生が「情報量の平均値」、「不確定度の期待値」と表現したのは、「$-\log p$」型の理解を引きずっている可能性のある読者に対して「$-\sum p_i \log p_i$」型であることを重ねて注意喚起するためだったと考えられますが、逆に、文字通りの「エントロピーの平均値：$\sum p_j\left(-\sum p_{j,i} \log p_{j,i}\right)$」を思い浮かべる読者もいないわけではありません。

また、文庫版 036 ページ（相互情報量）でも「平均値」であることが強調されています。042 ページでは、「ところで，相互情報量は得られる情報量の期待値であることに注意しよう」と強く念押しされています。

解釈 1 の立場では、情報量は「$-\sum p_i \log p_i$」なので、「相互情報量」と呼んで問題ないのですが。わざわざ「情報量の期待値であることに注意しよう」と念押しされると、定義 1.3 にも関わらず甘利先生は、なお情報量を「$-\log p$」の意味で使っていることになります。

甘利先生の説明は、定義 1.3 以降も、情報量と情報量の平均値とを区別しており、情報理論初心者である読者は困惑し、解釈 2 の可能性を考えることになります。

このような混乱を防ぐために、等確率時の $-\log p$ 形式の情報量を「選択情報量」、「自己エントロピー」などと呼び変える工夫をしている方もいらっしゃるようですが、私は、そのような用語を使わない甘利先生のシンプルな説明が一番優れていると感じています。

第 1 章第 1 節は、入門者向けのとても優れた説明であり、もし改訂の機会があるならば、是非、説明の一貫性にご配慮いただきたいものです。（なお、逆説的な話ではありますが、定義 1.3 の研究によって私の情報量に対する理解は随分と深まったと感じています。）