情報量とエントロピー

アメリカのＡＴ＆Ｔ ベル研究所のシャノン(*Claude. E. Shannon*)は、1948 年に「A Mathematical Theory of Communication」という論文を発表し、それまで曖昧な形でしか把握されていなかった「情報」の概念を明確に定義しました。

シャノンは、情報を通信する立場から定量的にとらえる方法を、数学的理論として展開し、「情報理論」の基礎を築きました。このようなことから、シャノンは、「情報理論の創始者」といわれています。そこで、シャノンの理論の一部を簡単に紹介しましょう。

まず、理屈抜きに次のような話をしましょう。「情報量」ですが、これは個々の事柄について計算できる量です。一方、エントロピーは起こりうる事柄の全体から計算される量です。

いま、ある事柄が確率 pで起こるとき、情報量は log(1/p) で表わされます。これだけを、無条件に信じて次の話を聞いて下さい。

よく喩え話に出されるサハラ砂漠の天気を考えてみます。

サハラ砂漠の天気はほとんどといっていいほど「晴れ」で、ごくたまに「雨」が降ると仮定しましょう。このとき、「天気は晴れ」という情報量を考えてみます。

サハラ砂漠はたいてい晴れているので、晴れる確率 pはほぼ1です。このとき、*log*(1/p) の値はほぼゼロになります。つまり、サハラ砂漠では「天気は晴れ」という情報量はほぼゼロというわけです。つまり、天気は「晴れ」と聞いても、「いつも晴れているんだから、当たり前」という印象と一致します。

ところが逆に、サハラ砂漠で「天気は雨」という情報は、とても大きな情報量を持ちます。雨が降る確率q がほとんどゼロなので、log(1/q) の値は大きくなります。天気が雨と聞いて「えっ、サハラ砂漠で雨が降ったの？」という印象と一致します。

つまり、情報量とは、ある事柄に関する「驚きの大きさ」といってよいでしょう。

次に、エントロピーですが、これは個々の事柄について計算されるものではなく、ある系全体について計算される量です。

情報理論では

エントロピー＝情報量の平均

です。たとえば、サハラ砂漠の天気を 1 年間観測して、晴れのときは晴れの情報量、雨のときは雨の情報量を足し合わせて、年間の 365 日で割れば、「サハラ砂漠の 1 日の天気が持つエントロピー」になるわけです。

前準備はこのくらいにして先に進みましょう。

実は、学生時代に興味本位で読んだ shannon の「A Mathematical Theory of

Communication」がこんなところに顔を出したのに驚いています。読んだのももう 40 年ぐらい前なので、読み直そうと本を探しましたが既にありませんでした。どうもとっくに処分したようです。

さて、蝶の環境評価の手法の中に、Shannon—Wiener Index というのがあります。Shannon-Wiener Index というのは次式で表されます。

$$H' = -\sum_{i=1}^{S} \frac{n_i}{N} \log_2 \frac{n_i}{N}$$

ただし、N は総個体数
n_i は第 i 番目の種に属する個体数
S は種数

この式だけでは、ピンときません。そこで、基礎からこの式を追っかけてみることにしましょう。

情報量

蝶のトランセクト調査では、蝶の「種数」と「個体数」を調べますが、Shannon は情報量というものをまず考えます。すなわち、ある事象を観察したときに得られる情報の量やメッセージに含まれる受け手にとって有用な情報の量を考えるわけです。

私たちは、情報というと定性的なものとすぐ考えてしまいますが、Shannon は情報というものを定量的に評価する「物指し」を考えたわけです。すなわち、情報という漠然とした量に数学的な手法を用いて「情報量」という概念を考え出したわけです。

情報の量を、ある事象が発生する確率から考えてみましょう。ある事象の情報の量を測ろうとする場合、基準を揃える必要があります。そこで、確率 1/2 で起こる事を伝える情報の量を単位とするのです。これが、「1ビット」と呼ばれるものです。一般に、確率$(1/2)^n$ で起こる事を伝える情報量が n ビットになります。

直感的には、変数 pに対して、pの対数 $\log_m p$ はm進数での pの桁数を表します、したがって、確率 1/p で起こる事象の情報量はpの桁数を表します。
つまり、確率を p とすると

$$p = \left(\frac{1}{2}\right)^n$$

のとき、n ビットが成り立ちますから、両辺の対数（底を2）をとれば、

$$\log_2 p = \log_2 \left(\frac{1}{2}\right)^n$$

$$\therefore n = -\log_2 p \qquad \text{ビット}$$

つまり、確率　pで起こる情報の持つ情報量を　I（p）とすると、I（p）は次式で与えられます。

$$I(p) = -\log_2 p \qquad \text{ビット}$$

ちょっと練習してみましょう。

（1）サイコロを1個投げたときに出た目を知る情報量はどうなるでしょうか。

p＝1/6 より、

$$I\left(\frac{1}{6}\right) = -\log_2 \left(\frac{1}{6}\right) = \log_2 6 \cong 2.58 \qquad \text{ビット}$$

（2）A君の大学合格の可能性は 1/10 だとすれば、A君の合格、不合格を伝える情報量はどうなるでしょうか。

$$I(\text{合格}) = -\log_2 \left(\frac{1}{10}\right) = 3.322 \qquad \text{ビット}$$

$$I(\text{不合格}) = -\log_2 \left(\frac{9}{10}\right) = 0.152 \qquad \text{ビット}$$

A君の合格の可能性が 1/10 と低いため、合格を伝える情報は不合格を伝える情報量より大きいのがわかりますね。

さて、情報量の直感的な定義は何でしょうか、列記してみましょう。

〇ある事象を観察した時に、観察者が得る情報の量は、データの量とは一致しません。すなわち、必ず起きると誰もが知っていること（確率 1）を知っても情報量は少なく（情報量＝0）、滅多に起きないことを知ったときの情報量は大きい。

〇確率に対して単調減少である。つまり、発生確率が小さい事象を観察したときの情報量は大きい。

〇犬が人を噛んだという情報と、人が犬を噛んだという情報では、われわれの「驚き」は大きく異なります。犬が人を噛むことはしょっちゅうあり、我々は驚きもしません（情報量小）が、人が犬を噛んだという話になるとまずありえませんから驚いてしまいます（情報量大）。

そこで、情報量を表す関数を I とします。

（1）確率が小さいほど、情報量が大きい。すなわち、関数は単調減少である。

$$p_1 > p_2 \Rightarrow I(p_1) < I(p_2)$$

（2）独立な事象に対しては加法性が成り立つ。
どういうことかを、具体的な例で示しましょう。
サイコロを 1 回振る場合を考えます。ただし、2 つの事象a, bを次のように定めます。
aは偶数の目が出る事象。
bは3の倍数が出る事象。
とします。
情報量I(a)、(b)、I(ab)を求めます。
I(a) ＝$-\log_2 p(a) = -\log_2(1/2) = 1$ ビット
I(b) ＝$-\log_2 p(b) = -\log_2(1/3) \fallingdotseq 1.58$ ビット
I(ab) ＝$-\log_2 p(ab) = -\log_2(1/6) = 2.58$ ビット

したがって、
I (ab) ＝I(a) ＋I(b)が成立します。

（3）確率 p に対して連続である。

（4）非負の関数で確率1のとき0である。

$$I(p) \geq 0, I(1) = 0$$

したがって、これらの条件を満たす関数 I は対数しかありません。

$$I(p) = -\log_b p$$

また、底を変えても、定数倍になるだけで、単位を何にするかで決まります。

したがって、底bを底xに変換するには底の変換公式を使い、次式のように変換すれ

ばよいわけです。式中$I_b(p)$の底はb、左辺の$I_x(p)$の底がxです。

$$I_x(p) = -\log_x p = -\frac{\log_b p}{\log_b x} = \frac{1}{\log_b x} I_b(p)$$

で表されます。$I(p)$が常用対数（底が 10）で表されていて、2を底とする対数で表したい場合、すなわち、ディットからビットへの変換は、x=2、b=10 ですから

$$I_2(p) = \frac{1}{\log_{10} 2} I_{10}(p) = \frac{1}{0.3010} I_{10}(p) = 3.3223 I_{10}(p)$$

となります。

平均情報量

小学校では夏休みの宿題で毎日天気をノートに観測、記録するのが定番でしたが、インターネットの発達により、夏休みの終わりに検索、リストアップすれば全国の過去の天気さえ知ることが可能になり、学校でもあまりやらなくなりました。

A市とB市の過去の天気は
A市：晴れが 99%、雨が 1%
B市：晴れが 50%、雨が 50%であったとします。

A市の場合、ある日が雨という情報は 1%しか降らないということを考えれば、滅多に雨は降らないのに降ったわけですから、この「へえ、降ったの」という驚きをみんなが持ちますから、情報量は大ということになります。その場合、1 回の観測で得られる情報量の平均はどうなるのでしょうか。
この場合、観測していない日でも、晴れと書いておけば大体あたるわけですから、1 回の観測で得られる平均的な情報量は小さいということになります。

B市の場合はどうでしょう。ある日が雨という情報は、50%ですから、情報量としては小ということになります。

では、1 回の観測で得られる情報の平均はどうでしょうか。まず、観測していない日をどちらにすべきか迷うくらいですから、1 回の観測で得られる平均値的な情報量は大きくなります。

つまり、個々の結果の情報量よりも、平均的に得られる情報量のほうが意味があるということになります。

さて、次に以下のような問題を考えましょう。

敦賀市の 1 月のある日の天気予報が雪 50%、雨 30%、曇り 20%だとします。それぞれの情報量を求めてみます。
まず、それぞれの情報量を求めると以下のようになります。

I(雪) $=-\log_2 0.5=1.00$

I(雨) $=-\log_2 0.3=1.74$

I(曇り) $=-\log_2 0.3=2.32$

これらの情報量の平均はどうなるのでしょうか。足して、3 で割っても意味がありませんね。というのも、雪、雨、曇りの起こる確率はそれぞれ違うからです。
情報量の平均というのは、それぞれの情報量に発生する確率を掛けて足すことで得られます。すなわち、

平均情報量 $=I$（雪）$\times 50\% + I$（雨）$\times 30\% + I$（曇り）$\times 20\%$
$=1.00 \times 0.5+1.74 \times 0.3+2.32 \times 0.2$
$=1.57$ ビット

この平均値のことをエントロピーといいます。エントロピーは「不確かさの物指し」というわけです。

したがって、平均情報量は次式で表されます。

$$H = \sum_{i=1}^{S} p_i I(p_i) = -\sum_{i=1}^{S} p_i \log_2 p_i$$

情報量は本来無次元量ですが、情報量は確率の逆数の桁数を表しますから、情報量の単位として桁数の単位を使うわけです。したがって、対数の底として2、e、10 を選んだときの情報量の単位はそれぞれビット(bit)、ナット（nat）、ディット(dit)となります。

以上で、Shannon-Wiener のIndexが求まりました。

このエントロピーの概念は少し難しく、なかなか理解しにくいので、もう少し具体的に蝶を例にとって説明することにしましょう。

1． キチョウとモンシロチョウが、それぞれ1頭ずつ捕虫網の中にいるものとして、その中の1頭を取り出すときの平均情報量（エントロピー）H は

キチョウである確率＝1/2
モンシロチョウである確率＝1/2
したがって、

$$H = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1(\text{ビット})$$

2.　同じくキチョウとモンシロチョウがそれぞれ5頭ずつ入っている捕中網から、蝶を1頭取り出すときの平均情報量（エントロピー）H は
キチョウである確率＝5/10＝1/2
モンシロチョウである確率＝5/10＝1/2
したがって、

$$H = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1(\text{ビット})$$

このように、1．、2．のエントロピーはまったく同じ1ビットになりました。これは、キチョウとモンシロチョウの出現確率がまったく同じなので、曖昧さも同じというわけです。

1．や2．の場合のように、キチョウとモンシロチョウの出現確率が等しい場合は、エントロピーは最大値

$$H_{\max} = -2 \times \frac{1}{2}\log_2\frac{1}{2} = \log_2 2 = 1 \quad (\text{ビット})$$

になります。

3.　今度は、キチョウ7頭とモンシロチョウ3頭が入っている捕虫網から、蝶を1頭取り出すときの平均情報量（エントロピー）H を求めてみましょう。
取り出したものがキチョウである確率＝7/10
モンシロチョウである確率＝3/10
したがって、

$$H = -\left(\frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}\right) = 0.881(\text{ビット})$$

この場合のエントロピーは、0．881で1．や2．のような同じ確率の場合の最大値＝1よりも小さくなっています。これは、同じ確率の場合よりも情報量が大きい、すなわちキチョウの確率のほうが 7/10 と高いことがわかっているので、予測しやすくなったわけです。というのも、1．や2．の場合はどんな予測をしても 1/2 の確率でしか、当てる

ことはできませんが、3．の場合は、常に「キチョウである」と予測すれば 7/10 の確率で当てることができるからです。

4．　さらに今度は、キチョウとモンシロチョウの出現確率の違いが、更に大きい場合を考えます。

キチョウ9頭とモンシロチョウ1頭が入っている捕虫網から、蝶を1頭取り出すときのエントロピーH を求めてみましょう。

キチョウである確率＝9/10

モンシロチョウである確率＝1/10

したがって、

$$H = -\left(\frac{9}{10}\log_2\frac{9}{10} + \frac{1}{10}\log_2\frac{1}{10}\right) = 0.469(\text{ビット})$$

エントロピーがますます小さくなりました。このことは、情報量が更に大きくなり、すなわちキチョウが出てくる確率が 9/10 と圧倒的に高いことがわかっているので、何が出てくるかますます予測しやすくなったことを意味します。

ところで、捕虫網の中の蝶が全てキチョウだけのときのエントロピーは

$$H = 1 \times \log_2 1 = 0$$

になります。このことは、キチョウが出る確率が1の時は、当然何が出るかの予測できるので、エントロピーは0（ゼロ）になったというわけです。

5．　最後に、捕虫網の中の蝶が、キチョウ、モンシロチョウ、アゲハの3種類になった場合を考えてみましょう。

キチョウ3頭、モンシロチョウ5頭、アゲハ2頭が入っている捕虫網から、蝶を1頭取り出すときのエントロピーを求めてみましょう。

キチョウである確率 3/10

モンシロチョウである確率＝5/10＝1/2

アゲハである確率＝2/10＝1/5

したがって、

$$H = -\left(\frac{3}{10}\log_2\frac{3}{10} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{5}\log_2\frac{1}{5}\right) = 1.485(\text{ビット})$$

なお、キチョウ、モンシロチョウ、アゲハの3種類の蝶が出る確率が等しい時、エントロピーは最大値をとり、

$$H_{max} = -3 \times \frac{1}{3}\log_2 \frac{1}{3} = \log_2 3 = 1.585(\text{ビット})$$

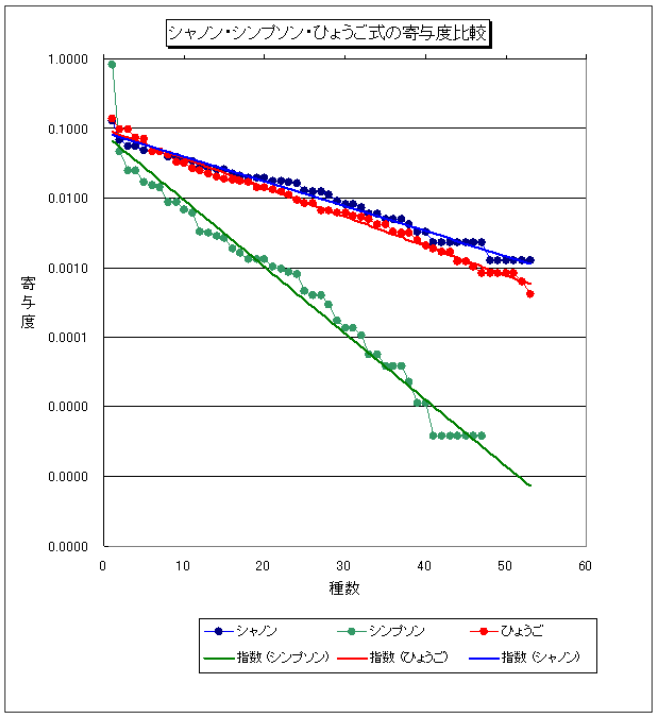
となります。

さて、下図をご覧ください。これはシャノン、シンプソン、ひょうご方式の導入の途中で出てくる 2006 年度に観察された 53 種の全体に与える寄与度を比較したものです。図では、全ての種の寄与度を均等に評価しているのはシャノンの方法で、次がひょうご方式、シンプソンの順になっています。

特にシンプソンの方式では、各種の個体数の多さを過大に評価する傾向にあり、個体数が1の種については無視するという乱暴な取り扱いをしています。生物多様性を調査するという点から考えると、希少種が考慮されないというのは問題です。

図中、各直線は最小二乗法で近似した直線（ただし縦軸は log 表示）ですが、シャノンの方法は優占種などは情報量として小さく評価し、確率で再度評価しなおすというように巧みに情報を操作しており、勾配もなだらかで全ての種について均等に考慮されていることが判ります。

ひょうご式は「平均値」で評価するのはまずいと「チョウのトランセクト調査」で指摘しましたが、このように寄与度で見るとシャノン・ウィーナーの方法に近いことが判り、平均値で考えない限り、評価できます。



この項は、「蝶のトランセクト調査」のために設けたものです。