3 情報理論

情報を数量的に取り扱う情報理論の基礎を与える。1 節では情報量の定義と基本的性質を与える。2 節では情報の損失を考える。人間の情報活動を 3 節で簡単に議論する。4 節ではエントロピーとその性質について述べる。分布の間の差を示すカルバック情報量についても議論する。分割表あるいは事象間の関連を示す結合エントロピーを 5 節で、条件付きエントロピーを 6 節で議論する。分割表のオッズ比の情報量としての解釈も与える。最後に、7 章では連続分布での情報量を考える。相対情報量としてのカルバック情報と相関係数の関連についても調べる。

3.1. 情報量

情報とは何か。日常的にも情報という言葉はよく用いられる。情報は“知る”ことに関係する文書、画像、音声、刺激などである。新聞、雑誌、テレビやラジオなどで伝えられる事柄は情報であり、また、文字や信号列などの成分も情報である。我々が情報の量を感覚的に捉えるとき、その情報の価値は別として意外な事象、すなわち起こる確率の小さい事象が発生したときに特に驚きや印象が強い。情報を数量的に考えることは、狭義の数理的議論だけでなく、生命科学や工学などの、学際的研究にも有益である。例えば、ヒトや生物の活動を情報処理の観点から数理的に議論すること、データ解析、あるいは情報通信の分野での情報伝達や処理に関する議論にも情報の数理的考察は応用されている。情報理論では偶然に発生する事象 A に対して、それが起こったときの情報量を

$$I(A) = \log(1/P(A))$$
$$= -\log P(A) \qquad (3.1)$$

で定義する。ここに P(A)は A が起こる確率である。対数 log の底は 1 より大きい実数であるが、情報理論では底を 2 にしてその単位をビット(bit)とする場合が多い。また、底を自然底 e とする場合は微積分の計算をする場合に便利であり、単位はナット(nat)である。この章では、底を省略した場合は底 2 の略する。情報量(3.1)は次の性質をもつ。事象 A,B に対して $0<P(A)<P(B)$とすれば

$$I(A) = -\log P(A)$$
$$> -\log P(B)$$
$$= I(B) \qquad (3.2)$$

が成立する。また、事象 A,B が独立のときは

$$P(A \cap B) = P(A)P(B)$$

であるから、

$$\begin{aligned} I(A\cap B) &= -\log P(A\cap B) \\ &= -\log\{P(A)P(B)\} \\ &= -\log P(A) - \log P(A) \\ &= I(A) + I(B) \end{aligned} \tag{3.3}$$

が成立する。(3.3)は独立な事象が起こったときの情報量はそれぞれの量の和であり、この性質も我々の情報に関する感覚を反映している。コンピュータでは情報量単位としてビットを用いている。定義(3.1)から分かるように、確率 1/2 で起こる事象が起こったときに我々が受ける情報量が 1 ビットである。コンピュータの内部で記憶素子は 2 個の状態(例えば電気の正と負)を、ほぼ 1/2 の確率で取り、このことから記憶素子 1 個を 1 ビットと呼ぶ。

例 3.1. サイコロを 1 個投げるとき、特定の目 i は確率 1/6 で起こる。各々の目が出た場合の情報量は log6 = 2.585 ビットである。また、6 の約数が出た場合の情報量は

$$-\log P(6\text{ の約数}) = \log(3/2) = 0.584 \text{ ビット}$$

である。　　□

問 3.1 N 個の根元事象が同じ確率で起こるとき、特定の M 個の中に含まれる事象が起こったことを知ったときの情報量はいくらか。

3.2. 情報の損失

記憶していた事象の全部または一部を忘れた場合に、我々は情報の損失を感覚する。例えば、局番を含めて 7 桁の友人電話番号の下 2 桁を忘れた場合を考える。この場合に、完全な情報に復元するには確率 1/100 に相当する情報が必要である。すなわち、100 通りの中の 1 つを偶然に当てる以外に方法は無く、この場合の情報の損失は $I_{\mathrm{Loss}} = \log(1/100) = 2\log 10$ ビットである。これはおよそ 6 ビットである。もし、末尾の桁が偶数であったことが記憶にあればそれだけ情報の損失は小さいことになり、$I_{\mathrm{Loss}} = \log 50 = 5.64$ ビットである。正確な情報の一部を隠して伝えるような場合も正確な情報に対する情報量の損失を評価することが出来る。正確な情報を A、不正確な情報を B とするとき、この 2 つの情報の間に次のことを仮定する。

$$A\subset B$$

このとき、

$$I_{\mathrm{Loss}} = I(A) - I(B) \tag{3.4}$$

である。また、完全情報 A に対して失われた情報の比率は

$$(I(A) - I(B))/I(A)$$

で定義される。

[例 3.2] 情報学の試験結果が表 3.1 のようになっている。これを、成績 A,B,C を合格、D を不合格として発表した場合（表 3.2）の情報の損失を考える。

表 3.1. 成績分布 I

成績	A	B	C	D	合計
人数	24	45	18	15	102

表 3.2. 成績分布 II

成績	合格	不合格	合計
人数	87	15	102

任意に 1 名をこの集団から選出するとき成績が A の人は確率 24/102 で抽出される。また、合格者が抽出される確率は 87/102 である。このことから、成績が A である場合の情報の損失は

$$I_{\mathrm{Loss}} = \log(102/24) - \log(102/87)$$
$$= \log(87/24) = 1.858 \text{ ビット}$$

である。成績が B および C の場合も同様に考えればよい。次に、表 3.1 を表 3.2 にして発表した場合の情報の損失を考える。合格の 87 人を 3 群に分ける場合の数は ${}_{89}C_2$ であるので、表 3.1 から表 3.2 を復元するのに $\log {}_{89}C_2$ の情報が必要である。従って、損失は

$$\log {}_{89}C_2 = \log(89\times 44)$$
$$= 11.92 \text{ ビット}$$

である。これに対して(3.4)を用いて情報の損失を計算する。合計人数を与えたとき表 3.1 と表 3.2 の情報量はそれぞれ

$$I(\text{表 3.1}) = \log {}_{105}C_3, \quad I(\text{表 3.2}) = \log({}_{105}C_3/{}_{89}C_2)$$

であるから

$$I(\text{表 3.1}) - I(\text{表 3.2}) = \log {}_{105}C_3 - \log({}_{105}C_3/{}_{89}C_2)$$
$$= \log {}_{89}C_2$$

として、上の計算と同じになる。 □

[例 3.3] ある実験で確率 p, q, r で非常に良好、良好、効果なしの結果が起こるとき、非常に良好と良好を効果ありとして記録した。このとき、結果が非常に良好であった場合の情報の損失は

$$(-\log p) - \{-\log(p+q)\} = \log(p+q)/p$$

である。ただし、$p, q, r>0$ とする。

問 3.2 サイコロを 1 個投げるとき、出た目の数を正確に教えないで、目が偶数か奇数かのみを教える場合の情報の損失を求めよ。

3.3. 人間の情報処理

コンピュータ内では記憶素子 8 個をまとめて 1 バイト（8 ビット）という情報単位を用いている。この場合 1 バイトで $2^8 = 256$ 通りの文字が表現できることになる。アルファベット、数字および汎用的特殊記号は高々100 個程度であるので、1 バイト内で表現可能である。しかし、日本語は漢字まで合わせればおよそ 2000 字であり、2 バイトの情報素子で表現することになる。日本語の場合は 1 字当たりの情報量は全ての文字が等しい確率で起こるとして

$$\log 2000 = 10.... \text{ビット}$$

である。実際には使われる文字の頻度には相当な偏りがあるので、日本語 1 文字あたりの情報量は 10 ビットより小さい。全ての文字が等しい確率で起こる場合に、1 秒間におよそ 10 字を話して聞き手に情報を伝えれば、100 ビット/秒位の情報を”話す”ことにより処理している。また、人の話を聞くときも 10 字であれば聞き取れる。従って、聞くことも 100 ビット/秒ぐらいの情報処理をしていることになる。目で見る場合の情報処理について考えてみる。デジタル写真などは写真の用紙上に縦と横罫線を引き、その格子点上に色と明暗の情報を付けることになる。仮に縦に 300 本、横に 500 本の罫線を考えれば

$$300 \times 500 = 150{,}000 \text{ 個}$$

の格子点が存在する。その上の色として 1000 色、明暗が 10 段階とすれば、写真 1 枚で $1/(1{,}000 \times 10)^{150{,}000}$ で起こる事象の我々は見ることになる。従って、写真を 1 枚見ることによって処理する情報量は

$$I = \log(1{,}000 \times 10)^{150{,}000}$$
$$= 600{,}000 \text{ ビット}$$

である。このことから、動画を見る場合の処理する情報量は相当大きいことが分かる。我々が日頃処理している情報を、このように離散的に評価することによって、人間の情報処理能力を計量的に議論出来る。

人間の情報処理では、相互に連関性のある事象の処理である場合が多く、処理される情報量は上の議論より小さくなる。人間の情報処理能力評価については、生命科学的な観点からの十分な議論が必要である。例えば、次の文章（情報）が伝えられたとする。

“私は昨日、学校と勉強しました。”

このとき、我々は

“私は昨日、学校で勉強しました。”

のように修正を加えることができる。これは、過去の経験に基づく学習(learning)、あるいは前後の文字列の連関によるものである。汚れた画像を見るときも同様で、破損した部分を識別して、ある程度の修復を加えることができる。人間の活動を情報処理と捉えて、情報理論の観点から議論することは、行動科学や生命科学上有効であると思われる。

問 3.3 手のひらで物を認識するとき、感覚として認識する点が 1,000 点あり、圧力を 10 段階、温度を 10 段階識別できるとする。この仮定の下で物に触れて、その感触を感知する場合の処理する情報量はいくらか。

3.4. 平均情報量

標本空間 $\Omega = \{\omega_1,\omega_2,\omega_3,...,\omega_n\}$で、各事象 ω_i が起こる確率を $p(\omega_i)$とする。任意に 1 つの事象を観測するとき、その平均情報量を

$$H = \sum_{i=1}^{n} p(\omega_i) \log \frac{1}{p(\omega_i)}$$
$$= -\sum_{i=1}^{n} p(\omega_i) \log p(\omega_i) \qquad (3.5)$$

で定義する。但し、$p(\omega_i) = 0$ のとき

$$0 \log 0 = 0$$

と約束する。平均情報量はエントロピー(entropy)とも呼ばれる。

[例 3.5] 標本空間 $\Omega = \{\omega_1,\omega_2\}$で、$p = p(\omega_1)$とするとき平均情報量は

$$H = -p\log p-(1-p)\log(1-p)$$

である。このエントロピーを確率 p の関数として、その増減を調べる。

$$\frac{dH}{dp} = \log e(-\log p + \log(1-p))$$
$$= \log e \times \log\{(1-p)/p\}\} = 0$$

から、$p = 1/2$ で最大値（極大値）H=log2 をとる。二つの事象の生起を予測しにくい場合に平均情報量が最大になっている。また、不確実性が無い場合、すなわち $p = 0, 1$ のとき、平均情報量は 0 である。

問 3.4. エクセルを用いて p を水平軸にして、$H = -p\log p-(1-p)\log(1-p)$のグラフを描け。

問 3.5. 血液型が AO 型の父親と BO 型の母親から子が一人生まれるとき、その子の血液型のエントロピーを求めよ。また、両親が AB 型のときの子の血液型エントロピーを求めよ

上の例に見られるように、平均情報量は標本空間の不確実性を示す。

定理 3.1 （ギブスの不等式）$\boldsymbol{p} = (p_1,p_2,...,p_n)$と $\boldsymbol{q}= (q_1,q_2,...,q_n)$を 2 つの確率分布とするとき、次の不等式が成立する。

$$\sum_{i=1}^{n} p_i \log \frac{1}{p_i} \le \sum_{i=1}^{n} p_i \log \frac{1}{q_i} \qquad (3.6)$$

等号成立は $\boldsymbol{p} = \boldsymbol{q}$ のときである。

証明 関数 $\log_e x$ に対して次の不等式が成立する。

$$\log_e x \leqq x-1 \qquad (3.7)$$

この不等式の等号成立は x=1 のときである。この不等式を利用して、

$$\text{左辺} - \text{右辺} = \log e \sum_{i=1}^{n} p_i \log_e \frac{q_i}{p_i} \le \log e \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) = 0$$

である。以上から、定理が示された。 □

(3.6)から

$$KL(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \ge 0 \qquad (3.8)$$

であり、KL($\boldsymbol{p}$,$\boldsymbol{q}$)をカルバックーライブラー情報という。これは、$\boldsymbol{p}$ が真の分布であるとき、それを $\boldsymbol{q}$ で代用する場合の情報を損失と考えられる。

問 3.6. (3.7)を示せ。

問 3.7. 条件

$$\sum_{i=1}^{n} q_i = 1$$

の下で、

$$\sum_{i=1}^{n} p_i \log \frac{1}{q_i}$$

を $\boldsymbol{q}$ に関して最小化することによって、定理 3.1 を証明せよ（ラグランジュの乗数法）。

上の定理から次の不等式が得られる。

系 3.1 $\boldsymbol{q}$ = (1/n,1/n,...,1/n)のとき、

$$H(p) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \le \log n \qquad (3.9)$$

が成立する。

この不等式はエントロピーの上限を与えている。全ての根元事象が等確率で起こるとき（一様分布）のとき、その不確実性が最大であることを示している。

[例 3.5] 箱 A に赤、青、白の石がそれぞれ、10, 20, 30 個、また、箱 B には 15, 20, 15 個入

っている。これらの箱から 1 個の石を取り出す場合に石の色を予測する。箱 A, B のエントロピーをそれぞれ H(A), H(B)とすれば

H(A)= (10/60)log(60/10) + (20/60)log(60/20) + (30/60)log(60/30) = 1.459,

H(B) = (15/50)log(50/15) + (20/50)log(50/20) + (15/50)log(50/15) = 1.571

である。以上から、箱 A の方が石の色を予測し易いことになる。

問 3.8. 分布 $\boldsymbol{p} = (p_1,p_2,...,p_n)$と $\boldsymbol{q} = (q_1,q_2,...,q_n)$に関して、0 ≦λ≦1 のとき

$$H(\lambda \boldsymbol{p} + (1-\lambda)\boldsymbol{q}) \geqq \lambda H(\boldsymbol{p}) + (1-\lambda) H(\boldsymbol{q}) \qquad (3.10)$$

が成立することを示せ。（ヒント）ギブスの不等式を用いよ。

定理 3.2　分布 $\boldsymbol{p} = (p_1,p_2,...,p_n)$に対して、$\boldsymbol{p}_1 = (p_1+p_2,p_3,...,p_n)$とすれば

$$H(\boldsymbol{p}) = H(\boldsymbol{p}_1) + (p_1+p_2)H(q_1,q_2) \qquad (3.11)$$

が成立する。ここに、$q_1 = p_1/(p_1+p_2)$, $q_2 = p_2/(p_1+p_2)$である。

証明　両辺を直接計算すれば、定理は示される。

系 3.2　　定理 3.2 と同じ条件の下で

$$H(\boldsymbol{p}) > H(\boldsymbol{p}_1) \qquad (3.12)$$

が成立する。

この定理と系は、事象を合併して観測する場合は情報が損失することを意味する。定理 3.2 を帰納的に適用して次の定理が成立する。

定理 3.3　（情報処理定理）分布 $\boldsymbol{p} = (p_1,p_2,...,p_n)$に対して、事象を適当に合併した場合の分布を $\boldsymbol{r} = (r_1,r_2,...,r_n)$とすれば

$$H(\boldsymbol{p}) > H(\boldsymbol{r}) \qquad (3.13)$$

が成立する。　□

得られた情報に処理を施す場合、事象を適当に合併し簡素な構造に加工する。この場合の平均情報量の損失は上の定理から

$$H(\boldsymbol{p}) - H(\boldsymbol{r}) \qquad (3.14)$$

と考えられる。

問 3.9. ある大きな母集団では、血液型 AA, AO, BB, BO, AB, OO をもつヒトの比率は p^2, $2pr$, q^2, $2qr$, $2pq$, r^2 となる。これを、A 型, B 型, AB 型，O 型と再分類した場合の平均情報量の損失を求めよ。ただし、$p+q+r=1$ とする。

[例 3.6] 外見で区別できない球が 12 個あり、その内の 1 つの球は他より重い。この球を特定するために天秤を用いて量る。確実に重さの異なる球を特定するために行う計量の最低回数を求める。ただし、重さの差は僅少であり、感覚では判別できないとする。1 個の球を特定するための情報は log(12) = 3.585 ビットである。1 回の計量で得られる情報量の最

大値は次のように考えられる。球全体を同じ個数の 3 群に分けられる場合が理想的で、任意の 2 群を秤に乗せた場合に、もし釣り合えば秤に乗せなかった群に重い球がある。一方、秤が傾けば、傾いた方の群に重い球が存在する。何れにしても、確率 1/3 に相当する事象の情報が得られ、log3 = 1.585 ビットが 1 回の計量で得られる情報の最大値となる。

log12/log3 = 2.262

であるから、2 回の計量では確実に重い球を特定できないことになる。1 回目に 4 個ずつの 3 群に分けて秤量すれば、4 個に絞ることができる。2 回目に 2 個ずつ分けて秤にかけると、2 個に特定され、3 回目の計量で重い球が特定される。したがって、重い球の特定のために最小の計量回数は 3 回である。　　□

問 3.10.　外見で区別できない球が 16 個あり、その内の 1 つの球は他と重さが異なる。この球を特定するために天秤を用いて量る。確実に重さの異なる球を特定するために行う計量の最低回数を求めよ。

3.5. 結合エントロピー

確率変数あるいは事象 X と Y の標本空間を、それぞれ{1,2,...,I}と{1,2,...,J}で示し、$p(i,j)$ = P(X=i,Y=j)とする。このとき、

$$H(X,Y) = -\sum_{i=1}^{I}\sum_{j=1}^{J} p(i,j)\log p(i,j) \qquad (3.15)$$

を X と Y の結合エントロピーという。さらに、３変数以上の結合エントロピーも同様に定義される。X と Y の周辺エントロピーを単に H(X)と H(Y)で示すとき次の定理が成立する。

定理 3.4　X と Y の結合エントロピーに関して次の不等式が成立する。

$$H(X,Y) \leqq H(X)+H(Y) \qquad (3.16)$$

等号成立は X と Y が独立のときのみである。

証明　記号の簡単のために、$p_X(i) = P(X = i)$, $p_Y(j) = P(Y = j)$とする。定理 3.1 のギブスの不等式を用いると

$$\begin{aligned} H(X,Y) &= -\sum_{i=1}^{I}\sum_{j=1}^{J} p(i,j)\log p(i,j) \\ &\leq -\sum_{i=1}^{I}\sum_{j=1}^{J} p(i,j)\log\left(p_X(i)p_Y(j)\right) \\ &= -\sum_{i=1}^{I}\sum_{j=1}^{J} p(i,j)\log p_X(i) - \sum_{i=1}^{I}\sum_{j=}^{J} p(i,j)\log p_Y(j) \\ &= -\sum_{i=1}^{I} p_X(i)\log p_X(i) - \sum_{j=}^{J} p_Y(j)\log p_Y(j) \\ &= H(X)+H(Y) \end{aligned}$$

□

上の定理を帰納的に用いて、次の系を得る。

系 3.3 $X_1,X_2,...,X_n$ の Y の結合エントロピー$H(X_1,X_2,...,X_n)$に関して次の不等式が成立する。

$$H(X_1,X_2,...,X_n) \leqq H(X_1)+H(X_2)+...+H(X_n) \qquad (3.17)$$

等号成立は $X_1,X_2,...,X_n$ が互いに独立のときのみである。　□

結合事象から得られる平均情報量は、それぞれの事象または確率変数の情報量和より少なく、このことは事象および確率変数間の連関による。すなわち、一方の事象の情報が得られれば、他方の事象について予測する場合の不確定性が減少するからである。例えば天気とある交差点での交通量を考える。通勤前に、天気が雨であれば交差点での渋滞が高い確率予測される。このような現象は我々が日常経験していることである。上の議論は理論上もこのことを示している。

分割表のオッズ比を情報理論で解釈する。簡単のために X と Y の標本空間を、それぞれ{1,2}と{1,2}とする。このとき、

$$OR = \frac{p(1,1)\,p(2,2)}{p(1,2)\,p(2,1)}$$

であり、対数オッズ比

$$\log OR = \log p(1,1) + \log p(2,2) + (-\log p(1,2)) + (-\log p(1,2))$$

$$= \{(-\log p(1,2))+(-\log p(2,1))\} - \{(-\log p(1,1))+(-\log p(2,2))\}$$

を得る。第 1 項は X と Y の値が異なるときの情報量、第 2 項は X と Y の値が同じときの情報量であるから、対数オッズ比は不確実性の変化量と解釈できる。

問 3.11.　X と Y の同時分布を表のように与えるとき、$H(X,Y)$, $H(X)$, $H(Y)$を求めよ。

X ＼ Y	1	2	3	4	計
1	1/8	1/8	0	0	1/4
2	1/16	1/8	1/16	0	1/4
3	0	1/16	1/8	1/16	1/4
4	0	0	1/8	1/8	1/4
計	3/16	5/16	5/16	3/16	1

3.7. 条件付きエントロピー

確率変数 X と Y に対して、$X=i$ を与えたときの Y の条件付きエントロピーを

$$H(Y \mid X = i) = -\sum_{j=1}^{J} p(j \mid i) \log p(j \mid i) \qquad (3.18)$$

で定義する。ここに、$p(j|i)$= P(Y= j|X=i)とする。上のエントロピーの X=i に関する平均

$$H(Y \mid X) = \sum_{i=1}^{I} p(i) H(Y \mid X = i) \qquad (3.19)$$

を、X を与えたときの Y の条件付きエントロピーという。ここに、$p(i)$=P(X=i)である。このエントロピーに関しては次の定理が成立する。

定理 3.5　確率変数 X と Y の結合エントロピーと X の周辺エントロピーを、それぞれ H(X,Y)と H(X)で示すとき次の関係が成立する。

$$\mathrm{H}(Y|X) = \mathrm{H}(X,Y) - \mathrm{H}(X) \qquad (3.20)$$

証明　確率変数 X と Y の同時確率を $p(i,j)$= Pr(X=i,Y=j)とするとき、条件付き確率は $p(j|i)$= $p(i,j)/p(i)$である。このとき、

$$\begin{aligned} H(Y \mid X) &= -\sum_{j=1}^{J} \sum_{i=1}^{I} p(j \mid i)\, p(i) \log \frac{p(i, j)}{p(i)} \\ &= -\sum_{j=1}^{J} \sum_{i=1}^{I} p(i, j) \log p(i, j) + \sum_{i=1}^{I} p(i) \log p(i) \\ &= H(X, Y) - H(Y) \end{aligned}$$

を得る。　□

さらに、次の定理を得る。この定理のイメージは図 3.2 に示している。

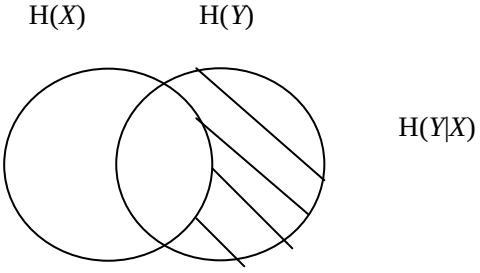

図 3.2 条件付エントロピー

定理 3.6　確率変数 X と Y の結合エントロピーと Y の周辺エントロピーに関して次の不等式が成立する。

$$\mathrm{H}(Y|X) \leqq \mathrm{H}(Y) \qquad (3.21)$$

また、等号成立は X と Y が独立のときである。

証明　定理 3.4 の(3.16)式と(3.20)を辺々加えて整理すれば、定理が得られる。ف

この定理から

$$\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) \qquad (3.22A)$$

は、Y に関する情報の内 X が持つ情報と解釈される。しかし、(3.20)から

$$\mathrm{I}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \qquad (3.22B)$$

$$= H(X) - H(X|Y) \qquad (3.22C)$$

が成立する。この情報量を相互情報量といい、X と Y の関連性を意味する。

問 3.12.　相互情報量 $I(X,Y)$をカルバックーライブラー情報量で表現せよ。

(3.22A)に関して

$$C(Y|X) = (H(Y) - H(Y|X))/H(Y) = I(X;Y)/H(Y) \qquad (3.23)$$

は、Y の不確実性で X によって説明される割合と解釈される。これを不確実性係数という。

[例 3.6] 天気予報 X とその翌日の実際の天気 Y の同時分布を表のように仮定する。このとき、不確実性係数 $C(Y|X)$を求める。

X ＼ Y	晴れ (0)	雨 (1)	計
晴れ (0)	5/8	1/16	11/16
雨 (1)	1/16	1/4	5/16
計	11/16	5/16	1

この表から、

$$H(X,Y) = 1.424, \quad H(X) = 0.896, \quad H(Y) = 0.896,$$

であり、

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = 0.368$$

を得る。以上から、不確実性係数は

$$C(Y|X) = I(X;Y)/H(Y) = 0.368/0.896 = 0.411$$

となり、Y の不確実性のおよそ 40%が X によって説明されたことになる。　□

問 3.13. 下の同時分布で不確実性係数 $C(Y|X)$を計算せよ。

X ＼ Y	1	2	3	4	計
1	1/8	1/8	0	0	1/4
2	1/8	1/8	0	0	1/4
3	0	0	1/8	1/8	1/4
4	0	0	1/8	1/8	1/4
計	1/4	1/4	1/4	1/4	1

定理 3.7. 確率変数 X,Y,Z に対して、$H(Z|Y) \geqq H(Z|X,Y)$が成立する。

証明　証明は左辺から右辺を引くことによって直接導かれる。

問 3.14. 上の定理を証明せよ。

問 3.15. 4 つの信号{1,2,3,4}を送信 X し、受信 Y する。このとき、X と Y の同時分布が問 3.8 のようになっている。不確実性係数 $C(Y|X)$を求めよ。

問 3.12. ある薬剤がある疾病に有効か調べるために、無作為に抽出した患者に確率 1/2 でプラセボと薬剤を割り付けて実験を行うとき、下の表のような確率で反応が起こるとする。この薬剤の疾病に対するする効果を条件付きエントロピーで評価せよ。

X（薬剤）＼ Y（回復）	なし (0)	あり (1)	計
なし (0)	7/20	3/20	1/2
投与 (1)	1/10	2/5	1/2
計	9/20	11/20	1

定理 3.8　確率変数 $X_1,X_2,...,X_n$ に対して、

$$H(X_1,X_2,...,X_n)= H(X_1)+H(X_2|X_1)+...+H(X_n|X_1,X_2,...,X_{n-1}) \qquad (3.24)$$

が成立する。ここに、$H(X_k|X_1,X_2,...,X_{k-1})$は X_k の $X_1,X_2,...,X_{n-1}$ を与えたときの条件付きエントロピーである。

証明　定理 3.5 から

$$H(X_1,X_2,...,X_n)= H(X_1)+H(X_2,...,X_n|X_1)$$

である。これを帰納的に適用して、定理を得る。　　ف

定理 3.7 を用いて次の定理を得る。

定理 3.9. 不確実性係数に関して次式が成立する。

$$C(Z|X) \le C(Z|X,Y) \qquad \square$$

この定理から、応答変量の不確実性は予測変量が増加するに従って、単調に減少することが分かる。

3.8. 連続確率変数の情報量

確率変数 X の密度関数を $g(x)$とするとき、微小増分 Δ に対して $g(x_i^*)\Delta$ は区間$[x_i,x_i+\Delta)$の確率である。但し、$x_i^* \in [x_i,x_i+\Delta)$とする。このとき、エントロピーは

$$H(X;\Delta) = -\Sigma_i g(x_i^*)\Delta \log_e(g(x_i^*)\Delta) \qquad (3.25)$$

の極限として、定義するのが自然のように思われる。ここに、Σ_i は全ての分割$[x_i,x_i+\Delta)$に関する和とする。しかし、

$$H(X;\Delta)\to\infty \quad (\Delta\to 0)$$

であり、常に発散する。そこで、連続確率変数のエントロピーを形式的に

$$H(X) = -\int g(x)\log g(x)dx \qquad (3.26)$$

で定義する。この量は確率変数の不確実性に関する意味は持たない。

問 3.16.　指数分布 g(*x*)=λexp(−λ*x*)のエントロピーを求めよ。

次にカルバックーライブラー情報量について考える。確率変数 *X* と *Y* の密度関数をそれぞれ *g*(*x*)、*h*(*y*)とする。このとき、分割[x_i,x_i+ Δ)に対して

$$\begin{aligned} KL(X,Y;\Delta) &= \sum_i g(x_i^*)\Delta\log\left(g(x_i^*)\Delta/(h(x_i^*)\Delta)\right) \\ &= \sum_i g(x_i^*)\Delta\log\left(g(x_i^*)/h(x_i^*)\right) \\ &\to \int g(x)\log\left(g(x)/h(x)\right)dx \quad (\Delta\to 0) \end{aligned}$$

が成立する。従って、

$$KL(X,Y) = \int g(x)\log\left(g(x)/h(x)\right)dx \qquad (3.27)$$

を、カルバックーライブラー情報量として定義する。この量は情報量としての意味をもつことになる。

[例 3.8] 指数分布 *g*(*x*)= $\lambda_1\exp(-\lambda_1 x)$と *h*(*x*)= $\lambda_2\exp(-\lambda_2 x)$の KL 情報量 KL(*g*,*h*)を求める。

$$\begin{aligned} KL(g,h) &= \int \lambda_1 e^{-\lambda_1 x}\left\{\log\frac{\lambda_1}{\lambda_2} + (\lambda_2-\lambda_1)x\right\}dx \\ &= \log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{\lambda_2-\lambda_1}{\lambda_1} \end{aligned}$$

である。上の例で λ_1 を固定して、λ_2 に関する KL(*g*,*h*)の増加減少を考える。カルバック情報を λ_2 に関して微分すると

$$\frac{d}{d\lambda_2}KL(g,h) = \frac{1}{\lambda_1} - \frac{1}{\lambda_2}$$

であるから、$\frac{d}{d\lambda_2}KL(g,h) = 0$ より

$$\lambda_1=\lambda_2$$

のとき最小値 0 となる。さらに、

$$KL(g,h)\to\infty \qquad (\lambda_2\to 0+0, +\infty)$$

である。

[例 3.9] 正規分布 *g*(*x*): $N(\mu_1,\sigma^2)$と *h*(*x*): $N(\mu_2,\sigma^2)$の KL 情報量 *KL*(*g*,*h*)を求める。

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}, \quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$$

より、

$$KL(g, f) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}$$

を得る。KL 情報量は分布に位置の平方に比例、分散に反比例することが分かる。また、関数 g と h の対称性から

$$KL(g, f) + KL(f, g) = \frac{(\mu_1-\mu_2)^2}{\sigma^2}$$

を得る。

最後に相関係数とカルバック情報量の関係について述べる。2 章で述べた 2 変量正規分布を考える。確率ベクトル(X,Y)が平均ベクトル(μ_1,μ_2)と分散共分散行列

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

をもつ 2 変量正規分布に従うとき、その密度関数は次のようになる。

$$f(x,y) = (2\pi)^{-1}|\mathbf{\Sigma}|^{-1/2}\exp\{-(x-\mu_1,y-\mu_2)\mathbf{\Sigma}^{-1}(x-\mu_1,y-\mu_2)'/2\}$$

ここに

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

$$(x-\mu_1,y-\mu_2)\mathbf{\Sigma}^{-1}(x-\mu_1,y-\mu_2)' = (1-\rho^2)^{-1}\{(x-\mu_1)^2/\sigma_1^2 - 2\rho(x-\mu_1)(y-\mu_2)/\sigma_1\sigma_2 + (x-\mu_2)^2/\sigma_2^2\}$$

である。この分布では X と Y の周辺分布はそれぞれ、$N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$であり、X と Y の密度関数をそれぞれ $f_X(x)$、$f_Y(y)$とすれば、

$$KL(f(x,y),f_X(x)f_Y(y)) = -(1/2)\log_e(1-\rho^2) \qquad (3.28)$$

$$KL(f_X(x)f_Y(y),f(x,y)) = (1/2)\log_e(1-\rho^2) + \rho^2/(1-\rho^2) \qquad (3.29)$$

を得る。上の情報量は ρ^2 の増加関数である。従って、$f(x,y)$と独立モデル $f_X(x)f_Y(y)$の乖離は相関係数の絶対値に従って増加する。上の情報を加えると

$$KL(f(x,y),f_X(x)f_Y(y))+KL(f_X(x)f_Y(y),f(x,y)) = \rho^2/(1-\rho^2)$$

である。

問 3.17. (3.28)と(3.29)を確かめよ。