・平均対数尤度　$E(logL)$

ここでは真の統計モデルからランダムで発生させたデータの対数尤度の平均

現実には推定用データと評価用データを分けておく

平均種子数から最尤推定値を求める

$\lambda_1 = \exp \beta_1$

観測データ　　真の統計モデル

$\hat{\beta}_1 = 2.04$　　$\beta_1 = 2.08$

のポアソン分布　　のポアソン分布

・最大対数尤度　$\log L^1$

一定モデルの当てはまりの良さ

真の統計モデルと観測データから推定された一定モデルには差が生じる

図4.9

・バイアス　　$b=\log L^1-E(logL)$

推定パラメータ数をk個持つモデルのバイアスはkとされている

・バイアス補正　　$E(logL)=\log L^1-b$

パラメータ1個なのでb=1を代入して

$$AIC=-2(logL^1-1)$$

これがAICとなる

AICは真の統計モデルから得た対数尤度に-2をかけているので、予測の悪さと解釈できる

$$log\lambda_i = \beta_1$$ 一定モデル

$$log\lambda_2 = \beta_1 + \beta_2 x_i$$ xモデル

$$AIC = -2\left(logL^1 - 1\right)$$

AICを用いることでパラメータ数が多いものが良いモデルとはならない

図4.13

情報

- どんなことか教えてくれるもの、不確実な知識を確実にしてくらるもの

情報の量

- その情報を得たことで知識の不確実さがどのくらい減ったかを計ったもの

例　n個の事柄 $A_1,A_2\cdots,A_n$ があり、そのうち一つが起こったとする

・サイコロを振った時の情報の量　　n = 6

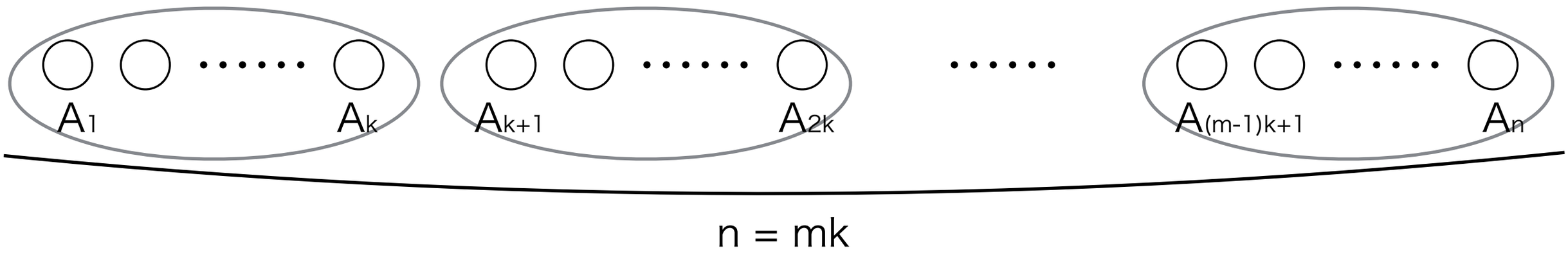
$A_1,A_2,\cdots,A_6$ はそれぞれ「1の目が出た」,「2の目が出た」…となる

・コイントスをした時　　n = 2

「表が出た」、「裏が出た」となる

どのAiが起こったかを知らせる情報の量は、事柄の数 n に関係する

情報の量をI(n)とする



n = mk

I(n) = I(mk) = I(m) + I(k)　　- 情報の加法性

小出しした情報の量を加えれば、全体の情報の量になる

情報量

加法性より

$$I(xy) = I(x) + I(y)$$

微分可能な関数と仮定すると

$$I(x + \epsilon x) = I\left((1 + \epsilon)x\right) = I(1 + \epsilon) + I(x)$$

$$\therefore I(x + \epsilon x) - I(x) = I(1 + \epsilon)$$

$$\begin{aligned}(I'(x) =)\ \lim_{\epsilon \to 0} \frac{I(x + \epsilon x) - I(x)}{\epsilon x} &= \lim_{\epsilon \to 0} \frac{I(1 + \epsilon)}{\epsilon x} \\ &= \frac{1}{x}\left(\lim_{\epsilon \to 0} \frac{I(1 + \epsilon)}{\epsilon}\right) = \frac{c}{x}\end{aligned}$$

xについて積分して

$$I(x) = c \ln x + d. \quad (\ln \text{は自然対数}.\ c, d \text{は定数})$$

xについて積分して

$$I(x) = c \ln x + d. \quad (\ln \text{は自然対数. } c, d \text{ は定数})$$

I(1) = 0より

$$I(1) = c \ln 1 + d = d = 0$$

ここで、二者択一の情報を情報単位に取り、これを1ビット(bit)とすると

$$I(2) := 1 \ (\text{bit})$$

$$I(2) = c \ln 2 = 1$$

$$\therefore c = \frac{1}{\ln 2} = \frac{\ln e}{\ln 2} = \log_2 e.$$

$$I(x) = \log_2 e \cdot \ln x = \log_2 e \cdot \frac{\log_2 x}{\log_2 e} = \log_2 x$$

x を確率p に変えると

$$\log_2 n = I + \log_2 k.$$

$$I = \log_2 n - \log_2 k = \log_2 \frac{n}{k} = \log_2 \frac{1}{p} = -\log_2 p$$

エントロピー … 情報量は-log p である。しかし-log p は事象が起こったことを知らせる情報量であり、これからもらう情報が「起こった」という答えが来るとは限らない。ここで、どの事象が起こったのか教えてもらうことにすると、事象A_1が起こればp_1の確率で-log p_1 の情報量が得られる。つまり、A_1における情報の量の期待値は$-p_1 * \log p_1$ となる。これがA_1が起こったかを聞くときに得られる情報の量と考えられる。これを情報の不確定度を表す量と捉え、エントロピーと定める。

全ての事象(それぞれの確率は $p_1,p_2,\ldots,p_n$)について考えると、

エントロピー $\mathrm{H}(\mathbf{p}) = -\sum_{i=1}^{N} p_i \log p_i$

性質1. エントロピー H は非負であり, H = 0 が成り立つのはどれか一つのp_i が1で他が全て0のときである。

束縛条件 $\sum_{i=1}^{N} p_i = 1$

性質2. n個の事象が表すエントロピーの最大値はH(n) = log n で, 全ての事象が等しい確率で起こるとき

束縛条件より $g(\mathbf{p}) := \sum_{i=1}^{n} p_i - 1 = 0$

ラグランジュの未定乗数λを用いて

$L(\mathbf{p};\lambda) := \mathrm{H}(\mathbf{p}) - \lambda g(\mathbf{p})$

$$\begin{cases} \dfrac{\partial \mathrm{H}}{\partial p_i}(\mathbf{p}) = -\log p_i - \dfrac{1}{\ln 2} \\ \dfrac{\partial g}{\partial p_i}(\mathbf{p}) = 1 \end{cases}$$

$\dfrac{\partial L}{\partial p_i}(\mathbf{p};\lambda) = 0$ より

$$\log p_i = -\frac{1}{\ln 2} - \lambda$$

右辺にi が存在しないことから, log p_i は定数であり、 $p_1 = p_2 = \cdots = p_N$

束縛条件より

$p_i = 1/N \ (i = 1, \ldots, n)$

p_i をH(p)に代入して

$$\mathrm{H}(\mathbf{p}) = -\sum_{i=1}^{N} \frac{-\log N}{N} = \log N$$

確率変数のエントロピーに一般化して

X　離散確率変数

$$p(x) = \Pr\{X = x\} \quad x \in \mathcal{X}$$

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

確率分布p(x), q(x)の間のカルバック・ライブラー情報量(相対エントロピー)

$$D(p||q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

・最尤推定

$p(x)$：真の確率分布

$q(x)$：推定した確率分布

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x)$$

一定　これが大きくなればK-L情報量は小さくなる

・最尤推定量　$\theta = (\theta_1, \cdots, \theta_k)$

$$l(x;\theta) := \sum_{i=1}^{n} \log q(x_i|\theta)$$

$$\frac{\partial l}{\partial \theta}(x;\theta) = 0$$

平均対数尤度　$< \log q(x) >_p := \sum_{x \in \mathcal{X}} p(x) \log q(x)$

平均対数尤度は求まらないので、観測された対数尤度を大きくすれば良い

$$\frac{1}{n} \sum_{i=1}^{n} \log q(x_i) \rightarrow < \log q(x) >_p \quad (n \rightarrow \infty) \qquad l(x) := \sum_{i=1}^{n} \log q(x_i)$$

連続確率変数では

エントロピー

$$h(X) := -\int p(x) \ln p(x) dx$$

K-L情報量

$$D(p||q) := \int p(x) \ln \frac{p(x)}{q(x)} dx$$

・AICの導出

最尤モデルの平均対数尤度を真のパラメータでTaylor展開したときの2次の項までを考慮し、近似する