適応的サンプリングによる標本推定法の設計

―発見科学からのアプローチ―

荒木　純道

― 目 次 ―

- 問題の背景
- 数学的準備
- 一括サンプリングとその問題点
- 適応サンプリング
- まとめ

確率事象の発生確率の標本推定

$$P \cong \frac{\sum_{i=1}^{n} X_i}{n} \triangleq \tilde{P}$$

P：真の発生確率（未知の定数）　$\tilde{P}$：推定発生確率（確率変数）

ただし，X_i：2値（0，1）の確率変数

事象が発生した時：$X_i = 1$

発生しなかった時：$X_i = 0$

問　題　点

- P が低い時にどれだけの標本数 n　を用意しなければならないか？
- 推定値 $\tilde{P}$ と真値 P との (絶対, 相対) 誤差はどのくらいか？
- 推定値 $\tilde{P}$ の信頼性は？

発見科学（機械学習，データマイニング）の立場と目標

- 確率事象の発生と確認は容易で，サンプル数は必要なだけ十分得られる．

cf. 古典的統計推定理論とは前提が少し違う．

- 推定結果が，指定された信頼度と相対誤差の規範を満足するアルゴリズムを設計提供する．

標本推定の設計規範

$$\Pr\left[\left|\tilde{P}-P\right|<\varepsilon P\right]<1-\delta \qquad (1)$$

信頼度：$1-\delta \quad (0<\delta<1)$

相対誤差：$\varepsilon \quad (0<\varepsilon<1)$

$\delta \to 0, \varepsilon \to 0$ **が望ましい**

数学的準備（Chernoff限界）

$\overline{X_i} = P$ の時，任意の $\varepsilon(0 < \varepsilon < 1)$ に対して

$$\Pr\left[\tilde{P} > (1+\varepsilon)P\right] < \exp\left(-nP\varepsilon^2/3\right) \qquad (2a)$$

$$\Pr\left[\tilde{P} < (1-\varepsilon)P\right] < \exp\left(-nP\varepsilon^2/2\right) \qquad (2b)$$

ただし，$\tilde{P} = \sum_{i=1}^{n} X_i / n$

Chernoffのアイデア：Min－Max原理

$$\Pr[X > C] = \Pr[f(X:t) > f(C:t)]$$

X：乱数　C：定数

だし，$f(x:t)$：パラメタ t を含む x に関する単調増加関

例　$f(x:t) = e^{tx}$　$(t > 0)$

次に，確率上限値をパラメタ t に関して最小化を
はかり，上界評価をより正確にする．

Markovの不等式

$$\Pr\left[X > C\overline{X}\right] \le \frac{1}{C}$$

ただし，X：正値乱数　$\overline{X}$：X の平均値　C：正の定数

$$\Pr\left[X > C\overline{X}\right] = \int_{C\overline{X}}^{\infty} P(X)\,dX \le \frac{1}{C\overline{X}} \int_{C\overline{X}}^{\infty} XP(X)\,dX$$

$$\le \frac{1}{C\overline{X}} \int_{0}^{\infty} XP(X)\,dX = \frac{1}{C}$$

一括サンプリング（固定サンプル数）とその問題点

規範（1）を満足するためには（2）から

$$\exp\left(-nP\varepsilon^2/3\right)<\frac{\delta}{2}$$

$$\therefore \quad n>\frac{3}{P\varepsilon^2}\ln\left(\frac{2}{\delta}\right) \qquad (3)$$

のサンプル数があればよい．

しかし，ε, δ は指定できるが，P は未知であるので

（3）からでは必要なサンプル数 n は決まらない．

適応サンプリングの原理

- nP →ほぼ事象の「発生回数」, $\frac{3}{\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$
- サンプルn 数を事前に決めるのではなく，事象の発生回数の上限を指定すればよい．

While文構造にする．（くり返し回数は確率変数）

cf. 一括サンプリングはFor文構造

（くり返し回数は確定数）

[定理　1]

アルゴリズムA の出力結果 $\tilde{P}$ は，(1) の規範を必ず満足

begin

$m \leftarrow 0; \quad n \leftarrow 0;$

while　$m < a$　do

$m \leftarrow m + x_i; \quad n \leftarrow n + 1;$

$\tilde{P} = m/n$

end

ただし，

$$a = \frac{3(1+\varepsilon)}{\varepsilon^2} \ln\left(\frac{2}{\delta}\right.$$

[定理　2]

アルゴリズム A のくり返し回数 n は

$$\Pr\left[n \le \frac{3(1+\varepsilon)}{(1-\varepsilon)p\varepsilon^2} \ln\left(\frac{2}{\delta}\right) \right] > 1 - \frac{\delta}{2}$$

を満足する．$(1+\varepsilon)/(1-\varepsilon)$ 倍だけ最適値より多い．

なお，P が既知の時，必要くり返し回数は

$$n = \frac{3}{p\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$$

数値例

$$\delta = 0.01 \quad \varepsilon = 0.1$$

$$a = \frac{3(1+\varepsilon)}{\varepsilon^2} \ln\left(\frac{2}{\delta}\right) = 1747.9$$

- a の計算は容易
- $\delta \to 0, \quad \varepsilon \to 0, \quad a \to \infty$
- $a \propto \varepsilon^{-2}, \quad a \propto \ln\left(\frac{1}{\delta}\right)$

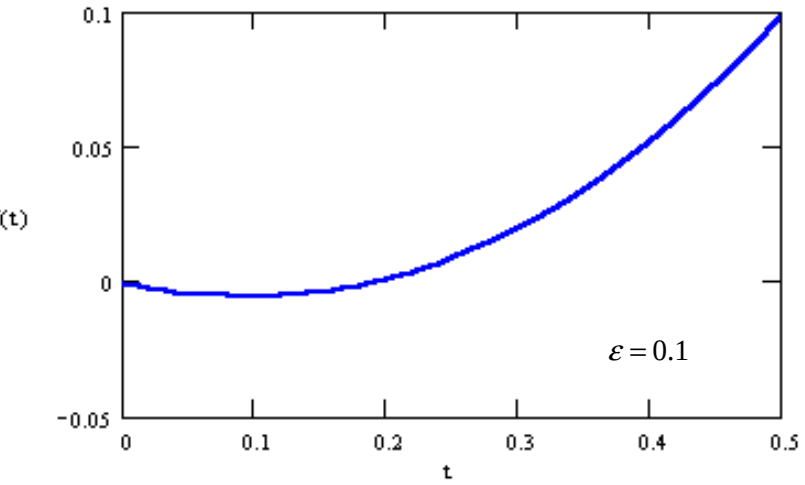
ま と め

- (δ, ε) 設計規範を満足する適応サンプリング
  アルゴリズムを示した.
- 必要な事業発生回数は, 相対誤差ε , 信頼度 δ に対して
  $\varepsilon^{-2}, \ln\left(1/\delta\right)$ に比例.
- アルゴリズムはwhile文構造で容易に記述.
- Chernoff限界ではなく, 中心極限定理を用いれば, もっと効率的なアルゴリズムが設計できる. (2～3倍早い)
- さぁ, みんなでアルゴリズム A を使ってみて下さい！

附　録

$$
\begin{aligned}
\Pr\left[\tilde{P} > (1+\varepsilon)P\right] &= \Pr\left[\sum_{i=1}^{n} X_i > (1+\varepsilon)nP\right] \\
&= \Pr\left[e^{t}\sum_{i=1}^{n} X_i > e^{(1+\varepsilon)nPt}\right] \qquad \forall t > 0 \text{に対して} \\
&< \frac{E\left[e^{t}\sum_{i=1}^{n} X_i\right]}{e^{(1+\varepsilon)nPt}} \qquad (\nearrow \text{Markov不等式}) \\
&= \frac{\left\{E\left[e^{tX_i}\right]\right\}^{n}}{e^{(1+\varepsilon)nPt}} \qquad (\nearrow \{X_i\}\text{のiid性}) \\
&= \frac{\left(e^{t}P + 1 - P\right)^{n}}{e^{(1+\varepsilon)nPt}} \qquad (\nearrow e^{x} \geq 1 + x) \\
&< \frac{e^{nP\left(e^{t}-1\right)}}{e^{nP(1+\varepsilon)t}} \\
&= e^{nP\left[e^{t}-1-(1+\varepsilon)t\right]}
\end{aligned}
$$

$$f(t)=e^{t}-1-(1+\varepsilon)t$$ のグラフ



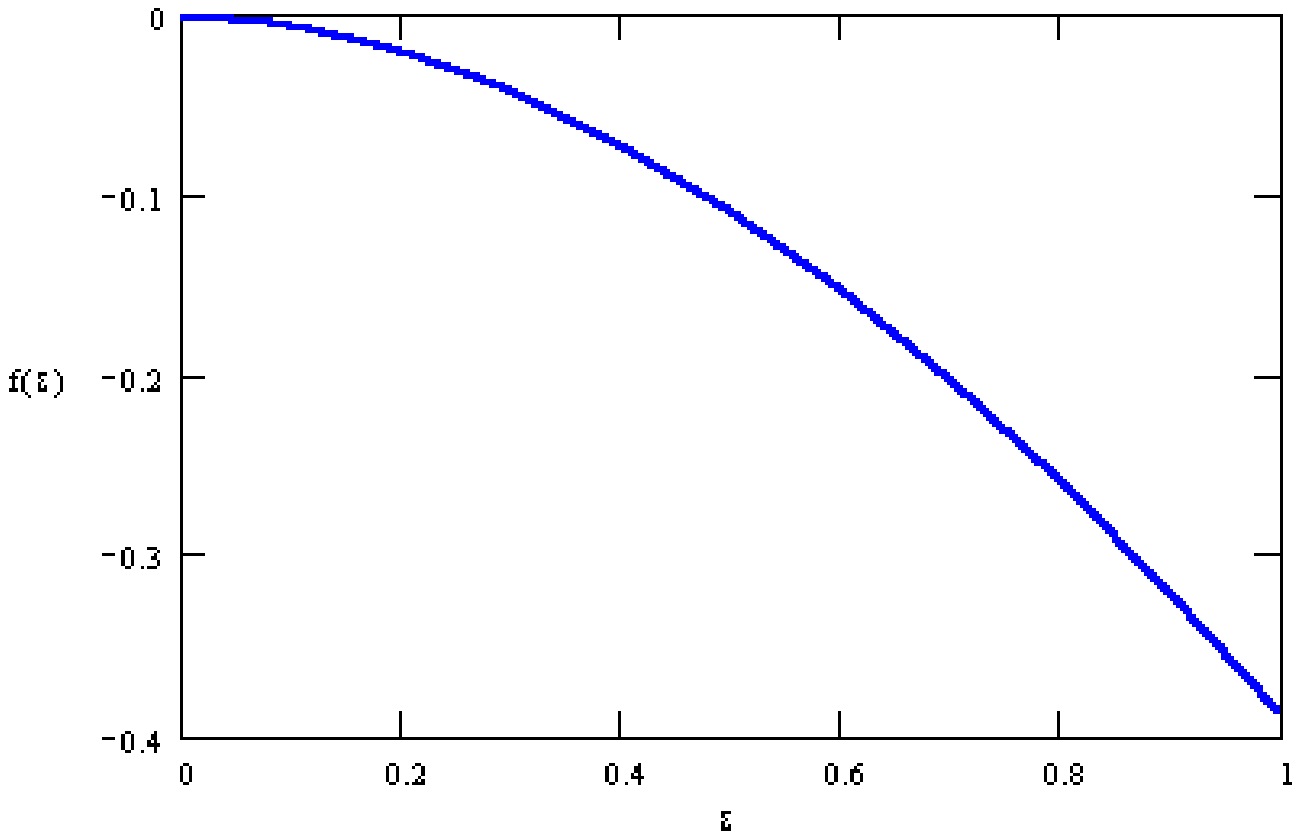
$$\text{Min } f'(t)=e^{t}-(1+\varepsilon)=$$

$$\therefore t_{\min}=\ln(1+\varepsilon)$$

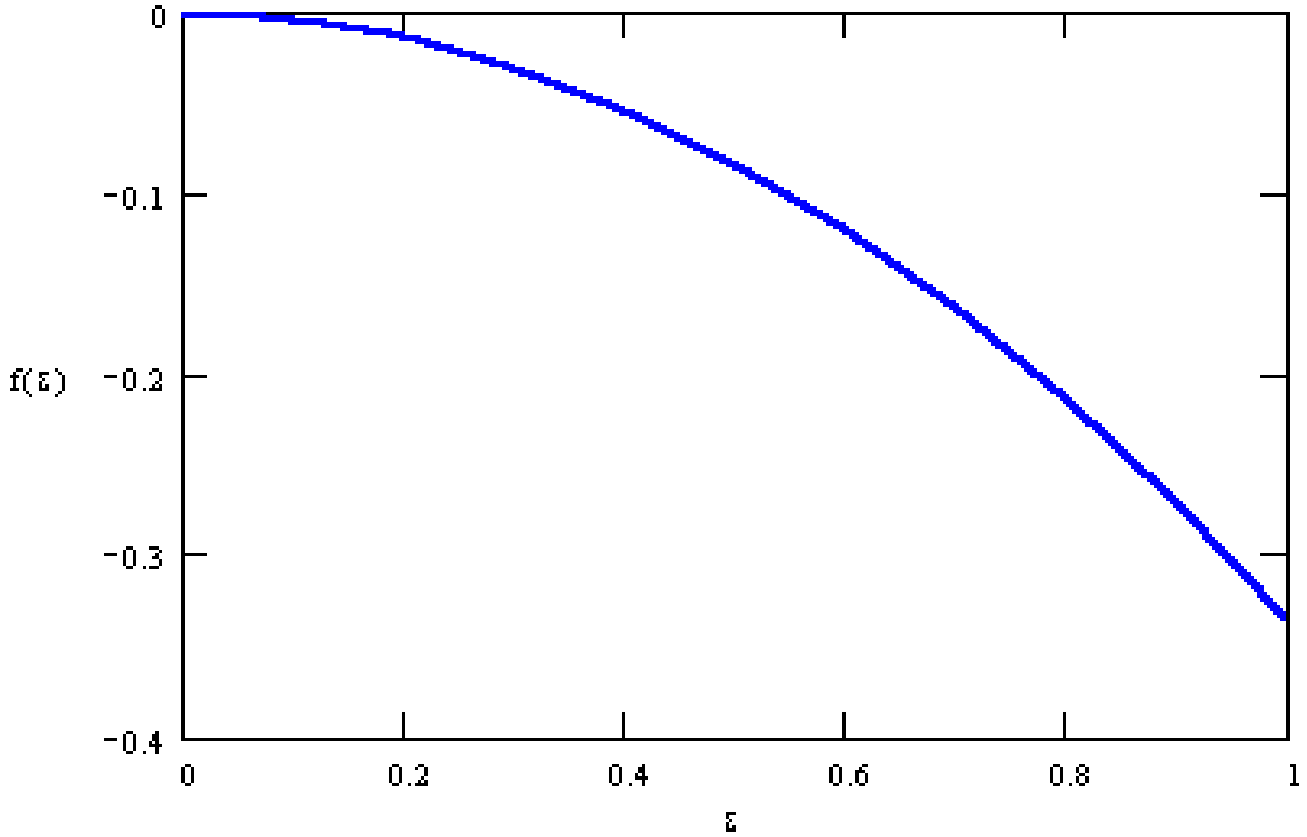
$$\text{Min } f(t_{\min})=(1+\varepsilon)-1-(1+\varepsilon)\ln(1+\varepsilon)$$

$$=\varepsilon-(1+\varepsilon)\ln(1+\varepsilon)$$

$f(\varepsilon)=\varepsilon-(1+\varepsilon)\ln(1+\varepsilon)$ のグラフ



$f(\varepsilon) = -\varepsilon^2/3$ のグラフ



$$\therefore \varepsilon - (1+\varepsilon)\ln(1+\varepsilon) < -\varepsilon^2/3$$

$$\therefore \Pr\left[\tilde{P} > (1+\varepsilon)P\right] < e^{-nP\varepsilon^2/3}$$

同様に $-\varepsilon - (1-\varepsilon)\ln(1-\varepsilon) < -\varepsilon^2/2$ **だから**

$$\Pr\left[\tilde{P} < (1-\varepsilon)P\right] < e^{-nP\varepsilon^2/2}$$