1995年度日本オペレーションズ・リサーチ学会
春季研究発表会

誤差を含む数値の表記法

米田 清 YONEDA Kiyoshi yoneda@ssel.toshiba.co.jp
(株) 東芝 研究開発センター

1 はじめに

計測装置の出力や, 確率シミュレーションの結果を表す数値には, 誤差がつきまとう.

誤差を含んだ数値を表現するには, 正規分布を仮定して信頼区間を使うのが普通である. たとえば測定値が $\mu = 123.46$ で, その誤差の標準偏差が $\sigma = 0.32$ の場合なら, 「区間 [122.50, 124.42] の間に正しい数値 X が含まれる確率は 99.7% である.」のような言い方をする. X が母数で μ が確率変数の実現値だからこの解釈は Bayes 流である. 頻度的な解釈をするのなら上より更にめんどうなことになる.

信頼区間による表現は統計の知識が必要で, 解釈に時間がかかる上に読み誤る可能性が高く, とっさの判断には使えない. 最大の難点は, 最も興味のある点推定値 123.46 が, 計算しないと出てこないことである. 点推定を中心にした 123.46 ± 0.96 という表示でも, たとえば 123.50 ± 0.96 と差があると思うべきか否か, すぐにはわからない.

現場では判断の誤りを避けるために, 数値が正確である桁数だけを表示して, 残りを抹消することが多い. たとえば上の場合, 有効数字が 4 桁であるなら, 123.5 とだけ表示する. この方式は上述の欠点がなく, 実用的である.

数値が正確である桁数だけを表示するためには有効桁数を決定する方式が必要であり, Song と Schmeiser [1] がいろいろな提案をしている. それらは基本的には最後の桁が正しい確率を計算して, その確率が予め定められた値よりも大きければその桁を表示し, 小さければそこで表示を打ち切るというものである. 信頼区間で危険率を天下りに決めて使ったのと同じように, やはり天下りの確率を使うことになってしまっている.

この報告は, そのような天下りの数値を使わずに, 最適な表示桁数を決定する方式を示す. 手計算用の簡便法も, あわせて提案する. 表現したい数値 X は μ と σ が既知の正規分布に従うものと仮定する.

2 定式化

ある桁 k で打ち切った数値 μ_k は, それ以下の桁について情報が表示されない. そこで, Bayes 流の無知の表現に習い, たとえばその数値が $\mu_{-1} = 123.5$ ならば, これは区間 $[123.5 - 0.05, 123.5 + 0.05]$ 上の一様分布を表すと解釈する. もとの分布は平均と標準偏差の与えられた正規分布であったので, 結局, 正規分布を一様分布で近似していることになる. 図 1 が典型的な例である.

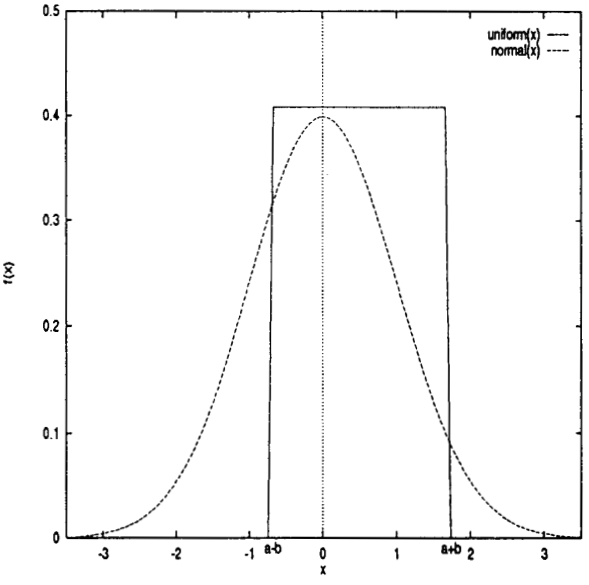

図 1: 正規分布の一様分布による近似

この近似の良さを計量し, それを最適化するような一様分布を求めれば, 最適な表示桁数が求められる. ここで近似の良さを Kullback-Leibler の情報量 [2] で計ることにする.

3 最適解

$f()$, $g()$ を各々, 区間 $[\mu_k - b_k, \mu_k + b_k)$ 上の一様分布と平均 μ, 標準偏差 σ の正規分布の密度関

数とすると, $a_k := \mu_k - \mu$ として情報量は

$$h(k) := \int_{-\infty}^{\infty} f(x)\log\{f(x)/g(x)\}dx = \frac{1}{2}(\frac{a_k}{\sigma})^2 + \frac{1}{6}(\frac{b_k}{\sigma})^2 - \log\frac{b_k}{\sigma} + \frac{1}{2}\log\frac{\pi}{2}$$

で, これを最小化するように b_k を決めれば良い. 最適解は整数

$$k = \arg\min_i h(i) . \quad (1)$$

4 近似解

今, k が $f()$ に従うと見なすと, $E[]$ を期待値として

$$E[a_k^2] = b_k^2/3 . \quad (2)$$

これを $h()$ の式に代入して $\partial h(k)/\partial b_k = 0$ を解くと $b_k = \sqrt{3/2}\sigma$. これを (2) に代入して, 典型は図 1 に示した $a_k = \sigma^2/2$ の場合であることがわかる. $b_k = (1/2)\times 10^r$ と書くと $r = \log_{10}(\sqrt{6}\sigma)$. これを図 1 と同じ条件で描くと図 2 が得られる.

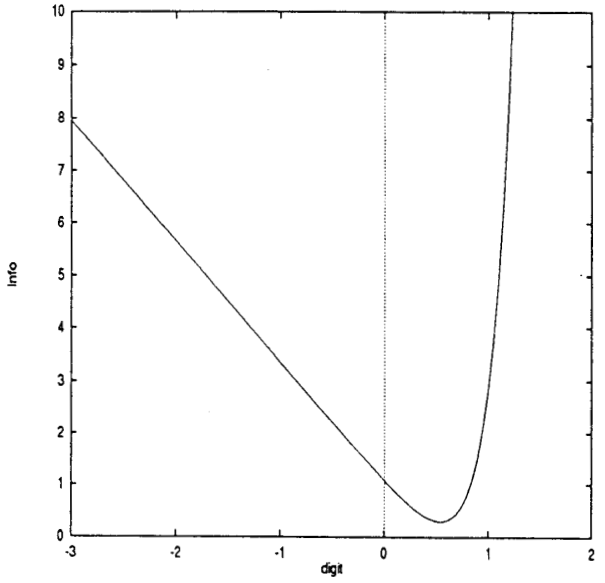

図 2: 桁数 r に対する情報量の例

この形から, 桁数を多め, すなわち k を小さめに取った方が安全なことが読みとれる. したがって, ほぼ最適な桁は, $\lfloor\ \rfloor$ を床関数として,

$$k = \lfloor\log_{10}(\sqrt{6}\sigma)\rfloor . \quad (3)$$

5 例

表示すべき数値として [1] の例 $\mu = 123.46, \sigma = 0.32$ を取る. 桁 k に対する情報量は図 3 のようになり, (1) から最適解は $k = \arg\min_i h(i) = 0$, $\mu_0 = 123$. 近似解は (3) から, $k = \lfloor\log_{10}(\sqrt{6}\times 0.32)\rfloor = -1$, $\mu_{-1} = 123.5$. 最適表示 $\mu_0 = 123$ の最後の桁が正しい確率は 0.55 で, 近似 $\mu_{-1} = 123.5$ の場合は 0.12 である.

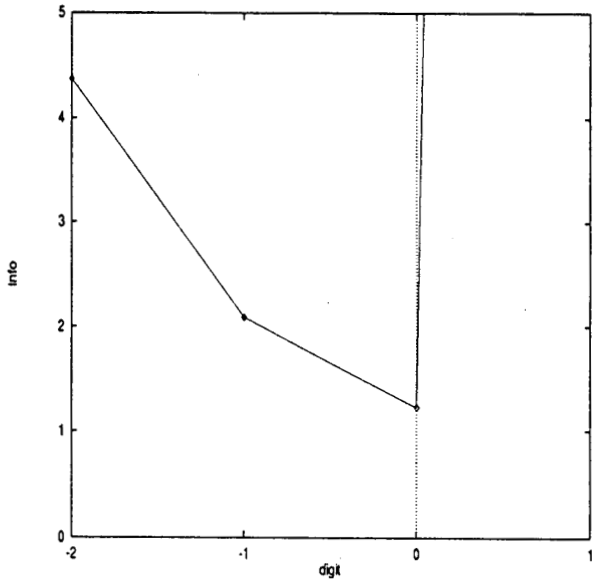

図 3: 桁数 k に対する情報量

6 逆問題

$(\mu, \sigma) \mapsto \mu_k$ の逆問題, すなわち丸めた数値から正規分布を求める問題も考えられる. その最適解は $\mu_k \mapsto (\mu_k, \sigma)$, $\sigma = 10^k/(2\sqrt{3})$.

7 おわりに

ここで提案した方式は, 任意の定数を判断に持ち込まないという意味で完結している. しかし分布間の距離の選び方には任意性がある. 他の計量, 例えば χ^2 で類似の結果が得られれば, この方法が常識に合うことの支持になる.

参考文献

[1] W. T. Song and B. Schmeiser. Reporting the precision of simulation experiments. In S. Morito *et al*, editor, *New Directions in Simulation for Manufacturing and Commmunications*, pp. 402–407. Operations Research Society of Japan, 1994.

[2] S. Kullback. *Information Theory and Statistics.* Dover, 1968.