Kolmogorov-Smirnov test

Kolmogorov-Smirnov test(コルモゴロフ-スミルノフ検定) とは任意の分布を示す 2 つの標本が同じ分布であるかどうかを評価する検定である。χ^2 検定や t 検定といった検定は適用できる分布が決まっているため、どのような分布か分からない標本を検定するには向かない。その点 Kolmogorov-Smirnov test はどのような分布にも適応できるため有用である。ただ、t 検定などのような高精度な検定結果は出せないので分布が分かっている場合はその分布に合った検定を使用すべきである。

Kolmogorov-Smirnov test を行う場合は、累積分布関数を考える。累積分布関数とは $(-\infty : a]$ に事象が発見される確率を表し、事象の密度を $f(x)$ とした場合

$$F(x) = \frac{\int_{-\infty}^{a} f(x)\mathrm{d}x}{\int_{-\infty}^{\infty} f(x)\mathrm{d}x} \tag{1}$$

と表される。例えば、次のような data が存在したとする。

$$\text{Sample A} \quad \begin{pmatrix} 1.26 & 0.34 & 0.70 & 1.75 & 50.57 & 1.55 & 0.08 & 0.42 & 0.50 & 3.20 \\ 0.15 & 0.49 & 0.95 & 0.24 & 1.37 & 0.17 & 6.98 & 0.10 & 0.94 & 0.38 \end{pmatrix}$$

$$\text{Sample B} \quad \begin{pmatrix} 2.37 & 2.16 & 14.82 & 1.73 & 41.04 & 0.23 & 1.32 & 2.91 & 39.41 & 0.11 \\ 27.44 & 4.51 & 0.51 & 4.50 & 0.18 & 14.68 & 4.66 & 1.30 & 2.06 & 1.19 \end{pmatrix}$$

この data の累積分布関数は次の様になる。

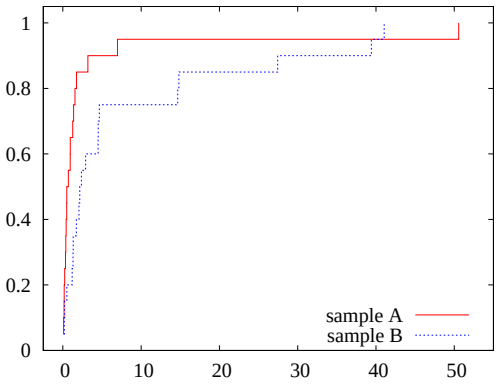

図 1　Sample A, B の累積分布関数

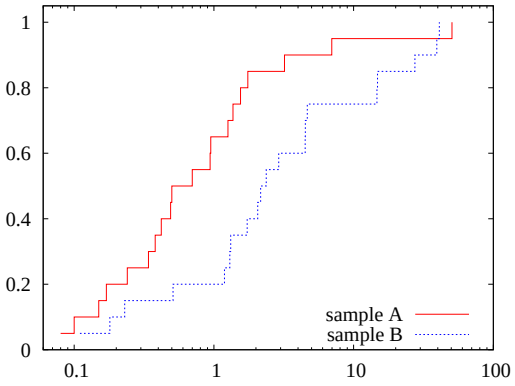

図 2　Sample A, B の累積分布関数 (x を logscale に)

この 2 つの data が同じような分布かそうでないかを考えるに当たって、この 2 つのグラフがどれだけ離れているかを考える。その前に、このカクカクのヒストグラム (もどき?) を百分率の折れ線グラフの書き換える。そのために次の式を使う

$$\text{percentile} = \frac{n}{N+1} \tag{2}$$

ここで n は data の集合を昇べきの順に並べ直したときのその data が n 番目に来る事を表す。一方 N は data 集合内の data の個数になる。これを使うと Sample A および Sample B の百分率グラフは図 3 及び図 4 になる。この変更によりヒストグラムは x が与えられたとき $(-\infty : x]$ に全体の何 % の事象が含まれるかを表す確率のグラフになる。

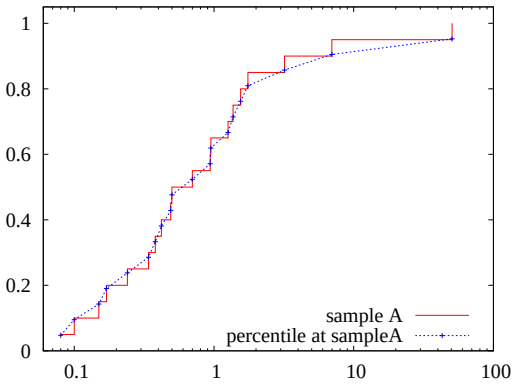

図 3　Sample A の累積分布函数 (ヒストグラムと折れ線グラフ)

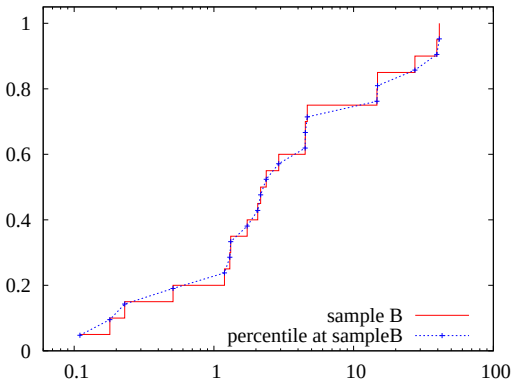

図 4　Sample B の累積分布函数 (ヒストグラムと折れ線グラフ)

最後にこの 2 つの確率のグラフを重ねて比較する。この比較こそ Kolmogorov-Smirnov test になる。2 つの累積分布函数を比較して一番値が離れたところの距離 D を求める。この D が大きければ大きいほど 2 つの

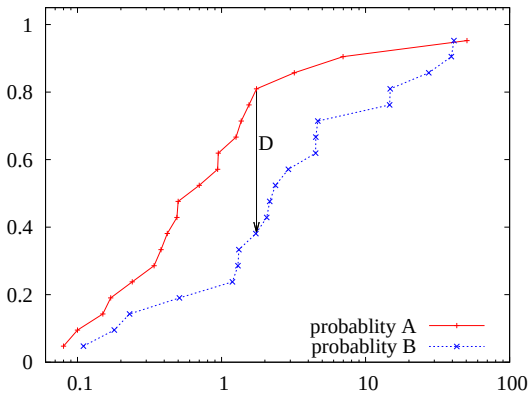

図 5　2 つの Sample の累積分布函数の比較

Sample は異なった分布を示すと考えられるのだが、それがどれくらいの確率で fit しているのかを表す有意確率は

$$P(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp\left[-2j^2\lambda^2\right] \tag{3}$$

$$\lambda = \sqrt{\frac{N_1 N_2}{N_1+N_2}}D \tag{4}$$

となる。[*1] ただし、Kolmogorov-Smirnov test を行う際に data の集合を 1 つだけ使い、比較対象を既知の分布にする場合、

$$\lambda = \sqrt{N}D \tag{5}$$

となる。

[*1] (3) の導出はどうやら N.Smirnov "*On the estimation of the discrepancy between empirical curves of distribution for two independent samples*" *Bulletin Mathématique de l'Université de Moscou*, Vol.2 (1939), fasc.2. に記載されているらしいが発見できなかった。