http://hdl.handle.net/10787/00033922

2-SAMPLE DATA analyzer

2019.7.11

石黒真木夫@統計数理研究所名誉教授

概要

統計的検定の場合を典型例として、何らかの観測値が理論の予想する値に一致するか見ることが必要になる場合がある。

観測誤差の存在を想定すると、観測値を確率変数とみなすべきであり、この観測値を生み出している確率変数と「理論値」の間にはっきりした大小差があるか否かを見ることになる。一般化すれば2つの確率変数の大小をデータにもとづいて判定する問題となる。

確率変数の形に過大な要請をすることなくこのような判定ができるソフトがあれば便利に違いない。そのようなソフトを、

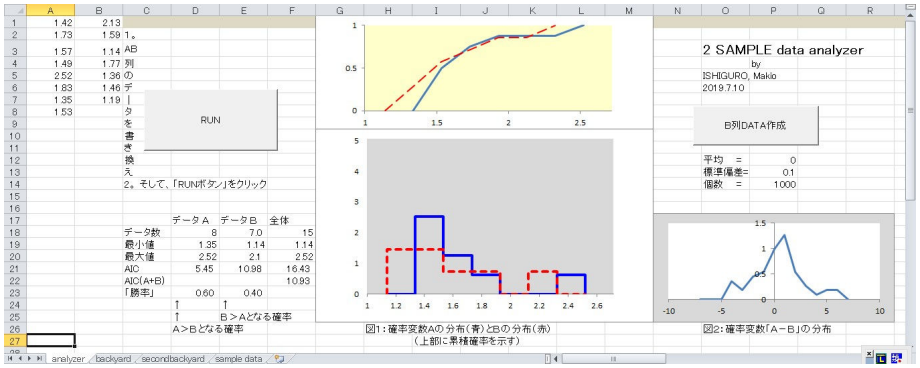
1. 確率密度関数の推定にヒストグラムを使い、
2. 確率密度関数の差の判定に2つのヒストグラムの畳みこみ積分を利用し
3. 2つのサンプルが異なる分布を持っているかを AIC で判定する

ことによって実現したのが 2 SAMPLE DATA analyzer である。

2 SAMPLE DATA analyzer はエクセルファイル、twosample.xlsm の形で提供される。

このソフトは「対応がない 2 SAMPLE 」を想定している。対応がある場合は「差」を「1 SAMPLE」と考えればいいだけのこと。

Twosample.xls 利用マニュアル



使い方は簡単。Twosample.xls をダウンロードし、画面左はしのA、B列にデータを入力して「RUN」ボタンをクリックするだけ。

　2つのヒストグラムで記述される確率変数の大小を調べることができる。データを2グループに分けることの妥当性はAICの値で調べられる。

確率変数の大小

確率変数 A とB の確率密度関数 $f_A(x)$ と $f_B(x)$ が与えられていれば、$A-B$ の確率密度関数は

$$f_{A-B}(u)=\int f_A(x)f_B(x-u)\,dx$$

で計算出来る。 $f_A(x)$ と $f_B(x)$ がそれぞれ 0-次スプライン関数

$$f_A(x)=\sum_{j=1}^{M} a_j\,h_j(x) \qquad f_B(s)=\sum_{k=1}^{M} b_k\,h_k(s)$$

つまりヒストグラム、の場合、

$$f_{A-B}(u)=\sum_{j=1}^{M}\sum_{k=1}^{M} a_j b_k \int h_j(x)\,h_k(x-u)dx$$

となる。

$$\mathrm{Prob(A<B)}=\mathrm{Prob}(A-\mathrm{B}<0)=\int_{-\infty}^{0} f_{A-B}(u)\,du$$

$$\mathrm{Prob(A>B)}=\mathrm{Prob}(\mathrm{A}-\mathrm{B}>0)=\int_{0}^{\infty} f_{A-B}(u)\,du$$

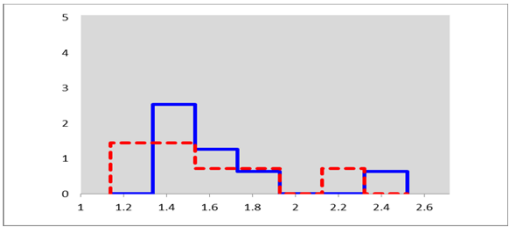
を確率変数 A とB の間の「大小関係」の指標として使える。

　なんらかの測定値の大小を論ずるにはなんらかの判定基準を決めておく必要があるのが通例であるが、確率については「1」という絶対的基準があり、直観的判断が可能である。

数式とグラフ

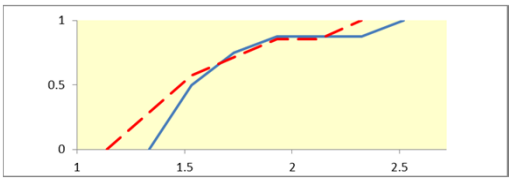
$$f_A(x) = \sum_{j=1}^{M} a_j\, h_j(x)$$

$$f_B(s) = \sum_{k=1}^{M} b_k\, h_k(s)$$

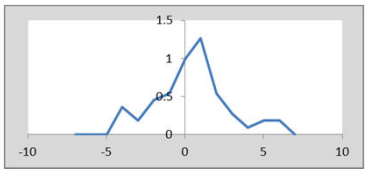


$$F_A(x) = \int_{-\infty}^{x} f_A(s)\, ds$$

$$F_B(x) = \int_{-\infty}^{x} f_B(s)\, ds$$



$$f_{A-B}(u) = \int f_A(x) f_B(x-u)\, dx$$



数式と数値

$D_A = \{1.42,1.73,1.57,1.49,2.52,1.83,1.35,1.53\}$

$D_B = \{2.13,1.59,1.14,1.77,1.36,1.46,1.19\}$

$N_A = 8$

$N_B = 7$

$Min(D_A) = 1.35$

$Max(D_A) = 2.52$

$Min(D_B) = 1.14$

$Max(D_B) = 2.13$

$Min(D_A \cup D_B) = 1.14$

$Max(D_A \cup D_B) = 2.52$

$$AIC_A = -2 \sum_{x \in D_A} \log \sum_{j=1}^{M} \hat{a}_j\, h_j(x) + 2(M-1) = 5.45$$

$$AIC_B = -2 \sum_{x \in D_B} \log \sum_{j=1}^{M} \hat{b}_j\, h_j(x) + 2(M-1) = 10.98$$

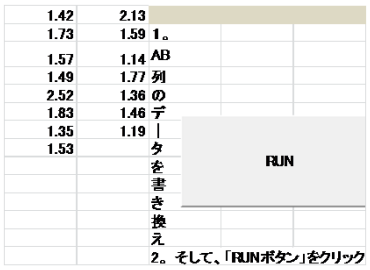
$$AIC_A + \mathrm{AIC_B} = 16.43$$

$$AIC_{A \cup B} = -2 \sum_{x \in D_A \cup D_B} \log \sum_{j=1}^{M} \hat{c}_j\, h_j(x) + 2(M-1) = 10.93$$

$$\hat{c}_j = \frac{N_A}{N_A + N_B} \hat{a}_j + \frac{N_B}{N_A + N_B} \hat{b}_j$$

$$Prob(A > B) = \int_0^{\infty} f_{A-B}(u)\, du = 0.6$$

$$Prob(B > A) = \int_{-\infty}^{0} f_{A-B}(u)\, du = 0.4$$

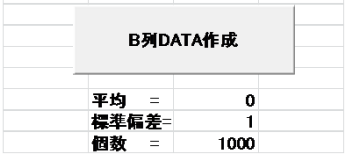


	データ A	データ B	全体
データ数	8	7	15
最小値	1.35	1.14	1.14
最大値	2.520	2.13	2.52
AIC	5.45	10.98	16.43
AIC(A+B)			10.93
「勝率」	0.60	0.40	
	↑	↑	
	↑	B＞Aとなる確率	
	A＞Bとなる確率		

「B列データ作成」ボタン、ヒストグラム区分変更、及び 「対応のある」データの分析

所与のデータを正規分布と比較したい場合がある。そのような場合のために「B 列データ作成」ボタンが用意されている。 所与のデータをA 列に置き、作成したい正規分布データの「平均」、「標準偏差」、「個数」を、P12, P13, P14 セルに設定してボタンをクリックすればよい。標準偏差を小さくすることによって、A 列データと定数値の大小較が出来る。

このボタンはシート上の情報を壊してしbまったときの再設定の機能も持っているので、そのために使用して A. B列のデータを本来比較したいものに戻せばよい。



おおざっぱな大小比較をするにはヒストグラム区分数 7 で十分と思われるが、区分数を変えたい場合には、C33 セルの7 を書き換えて「RUN」すればよい。

「対応のあるデータ」の分析には「jobtype」を 1 とする。

動作パラメータ設定 ----------							
ヒストグラム区分数			jobtype 2：2-sample データ解析；1：「A－B」データ解析				
7			2				

付録：$\sum_{j=1}^{M}\sum_{k=1}^{M} a_j b_k \int h_j(x)\,h_k(x-u)dx$ の計算

$$h_j(u)=\begin{cases} h & (j-1<uh\le j) \\ 0 & otherwise \end{cases}$$

とすると、

$$h_{jk}(u)=\int h_j(x)\,h_k(x-u)dx=\begin{cases} 0 & (uh\le j-k-1) \\ (uh-j+k+1)h & (j-k-1<uh\le j-k) \\ (j-k+1-uh)h & (j-k<uh\le j-k+1) \\ 0 & (j-k+1<uh) \end{cases}$$

となる、この式から $h_{jk}(u)=h_{(j-k),0}(u)$ であることが分かるので

$$f_{A-B}(u)=\sum_{j=1}^{M}\sum_{k=1}^{M} a_j b_k h_{jk}(u)\ =\sum_{j=1}^{M}\sum_{k=1}^{M} a_j b_k h_{(j-k),0}(u)$$

となる。$f_{A-B}(u)$ が1次スプライン関数で表されるわけである。