

観測値の距離の重みを考えた中央値推定法と検定について

A Method of Median Estimation and Statistical Test by Distance Weighting of Observation Data

北海学園大学工学部社会環境工学科

北海学園大学工学部社会環境工学科

北海学園大学工学部社会環境工学科

○学生員 中島秀弥 (Shuya Nakajima)

学生員 岡田多聞 (Tamon Okada)

フェロー 許士達広 (Tatsuhiro Kyoshi)

1. はじめに

情報処理・計算技術の発展により、ビッグデータの処理などの統計的な解析能力は飛躍的に向上している。しかし高度なモデル計算が可能になった一方で、その前提となり身近な計算に用いられる基礎的な技術はあまり変わっておらず、多くの問題が依然として残されている。中でも平均や回帰等による推定・予測の技術や検定は非常に使われる頻度が多いが、以下に示すような問題点がある。

データ分布を表す代表値として、各データからの距離の2乗を最小にする「算術平均」をとるのが一般的であるが、分布が一定以上非対称であるか外れ値が存在するとデータの中心部分から外れて代表値としての意味が低下する。

t 検定は平均値の差などに使われる代表的検定手法であるが、非対称分布では正確さが損なわれることが知られている。このためデータ分布の対称性が認められなければ、ウィルコクソンの順位と検定等の中央値の差ノンパラメトリック検定が用いられているが、実用上はこれらの使い分けの境界が明確でなく、考え方の違う複数の手法が併存している。また中央値の差のノンパラメトリック検定は順位和の検定であり、順位のみ用いて分布の50%中央値ではないため、データの値や外れの度合いは考慮されない。これらのことから分布の対象非対称にかかわらず平均値などのパラメータの検定ができることが望ましい。

2. 距離による影響度の理論

1) 距離による信頼性の違い

地理的距離や経過した時間、性質などが異なる幾つかの観測データを用いて、ある状況の値を推定する時、近い観測値から推定したほうが遠いものより信頼性が高い。これは一般的に自明のことであるが、なぜそうなるか、どの程度の信頼性の違いであるかが問題である。ごく初歩的な話であるが、図1のようにデータに回帰直線を引くとその信頼性を表す信頼区間が描かれる。信頼区間はそれが狭いほど推定値のばらつきが小さいことを示し、推定線上の点の信頼性が高くなる。信頼区間は中心から遠いほど広がるため、端部で信頼性が小さくなることがわかる。

2) 距離による信頼性からの重みの導出

推定値に対する距離に応じたデータの重みづけは以下のように考えられる。信頼区間に用いられる回帰式の誤差分散（推定誤差分散）は以下のように変形される。

$$S_h^2 = s_e^2 \left\{ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right\} \quad (2)$$

$$= \frac{s_e^2}{n} \left\{ 1 + \frac{(x_i - \bar{x})^2}{S_{xx}/n} \right\} = \frac{s_e^2}{n} \left\{ 1 + \frac{(x_i - \bar{x})^2}{\sigma_x^2} \right\} \quad (3)$$

S_h^2 : 推定誤差分散 s_e^2 : 残差分散, σ_x^2 : x の標本分散

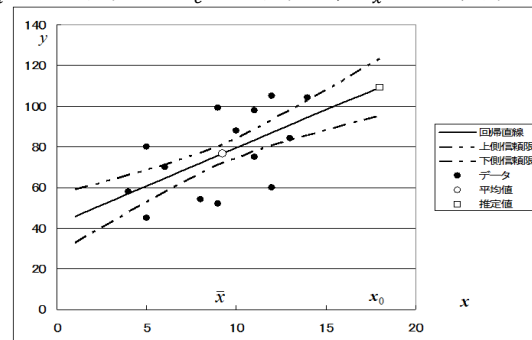


図-1 信頼区間

(3)式の{ }内はいわゆる「てこ比」の n 倍であり、 $x_0 - \bar{x}$ は中心から推定点 x_0 までの距離、 σ_x^2 はデータの x 方向の標本分散である。(3)式の逆数から(4)式のデータ点 x_i から分布の中心 $\bar{x}$ に対する重みを考えることができる。これは分布内の点 x_i の信頼度であると同時に、点 x_i の分布の中心 $\bar{x}$ に対する影響度を示す関数と考えることができる。 $(x_i - \bar{x})/\sigma_x$ は標準化距離であり、(4)式の重みは「標準化距離の2乗プラス1の逆数」である。

$$w_i = \frac{1}{1 + \frac{(x_i - \bar{x})^2}{\sigma_x^2}} \quad (4)$$

次に集団内の2点の関係を考える。(4)式の σ_x^2 は $\bar{x}$ を中心とする集団の分散であるが、 x_i と $\bar{x}$ を入れ替えて考えれば、任意の x_i を中心とする場合の分散とも考えることができ、 $\bar{x}$ を移動させて考えれば x_i を中心にした同様な信頼区間が存在する。母集団内の各点 x_i の分散は通常等しいとすると、母集団内の2点間の影響度の関係は等しいと考えられ、これによりデータ点 x_i から分布内の任意の位置 x_0 への重みは(5)式となる。

$$w_i = \frac{1}{1 + \frac{(x_i - x_0)^2}{\sigma_x^2}} \quad (5)$$

3. 距離による重みと t 分布・正規分布の関係

前節の内容は統計学的に別な観点からも説明される。スチューデントの t 分布は以下の式で表される。

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} \quad (6)$$

t 分布自体はデータ数 n が 2 個以上で定義されるが、データ点 x_i と推定点 x_0 との 2 点間の t 分布を考える場合、データ数は 1 で平均をとらないので自由度 ν が 1 であり、このとき t 値を入れると以下の(7)式のように表される。

$$f(t) = \frac{1}{\pi \left[1 + \left(\frac{x_i - x_0}{\sigma_x} \right)^2 \right]} \quad (7)$$

π は分布の全体を 1 にする定数で、各データに共通として除けば(7)式は(5)式による距離の重みと一致する。

(5)式は点 x_i を中心として x_0 を変数としたデータごとの尤度関数と考えると、これを用いてデータの分布の中心を求めることができる。このとき分布全体の尤度関数は下記の(8)式で表され、(9)式は対数尤度、(10)式はそれを x_0 で微分したもので、DL=0 となるように定めた解 x_0 は(12)式に示すように(5)式の重みを用いた重み付平均となる。

$$L = \prod_{i=1}^n \left(\frac{1}{\left[1 + \left(\frac{x_i - x_0}{\sigma_x} \right)^2 \right]} \right) \quad (8)$$

$$\begin{aligned} LL &= \sum_{i=1}^n \log \left(\frac{1}{\left[1 + \left(\frac{x_i - x_0}{\sigma_x} \right)^2 \right]} \right) \\ &= - \sum_{i=1}^n \log \left(1 + \left(\frac{x_i - x_0}{\sigma_x} \right)^2 \right) \end{aligned} \quad (9)$$

$$DLL = \sum_{i=1}^n \frac{2(x_i - x_0)}{\sigma_x^2 \left[1 + \left(\frac{x_i - x_0}{\sigma_x} \right)^2 \right]} = 0 \quad (10)$$

$$\sum_{i=1}^n w_i (x_i - x_0) = 0 \quad (11)$$

$$x_0 = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i \quad (12)$$

(12)式による最適な x_0 に対し(8)式から x の分布は(13)式で表される。()内は (14) 式のように変形され、平方根をとると(15)式になる。

$$f(x) = \left(\frac{1}{1 + \left(\frac{x - x_0}{\sigma_x} \right)^2} \right)^n \quad (13)$$

$$\begin{aligned} \frac{1}{\left[1 + \left(\frac{x - x_0}{\sigma_x} \right)^2 \right]} &= \frac{1}{\left[1 + \frac{(x - x_0)^2}{((n-1)/n) s_x^2} \right]} \\ &= \frac{1}{\left[1 + \frac{(x - x_0)^2}{(s_x^2/n)(n-1)} \right]} = \frac{1}{\left[1 + \frac{t^2}{(n-1)} \right]} \end{aligned} \quad (14)$$

$$f(t) = \left(\frac{1}{1 + \frac{t^2}{n-1}} \right)^{(n/2)} \quad (15)$$

(6)式の t 分布と比較すると(15)式は t 分布の関数の一部で、 $\alpha = \Gamma\left(\frac{\nu+1}{2}\right) / \left(\Gamma\left(\frac{\nu}{2}\right) \sqrt{\nu\pi}\right)$ 倍し分布全体を 1 にすると自由度 $n-1$ の t 分布関数となる。通常の t 分布関数との違

いは、式中の x_0 は平均値ではなく分布の 50% 中央値であることである。(13)式においてデータ数 n を十分大きくすると正規分布になることから、t 分布および正規分布の形状は元来中央値を算出する尤度関数であり、距離の重みから導かれることが分かる。すなわち(5)式の重みをもって任意のデータから合成した関数は t 分布または正規分布と同じ性質を持ち、その中心は最尤解としての 50% 中央値で同時に最頻値となる。

4. 非対称分布の検定

距離による重みを用いて非対称分布に対する中央値検定を行うことができる。(12)式における x_0 は最尤推定値としての中央値である。 x_0 が平均値となるのは分布が左右対称で平均値と中央値が一致する特別の場合であり、このため実際に t による平均値の検定はデータ分布が左右対称に近い場合に限定して使われている。本研究で中央値にすることにより、左右が非対称の分布についても t 検定を行うことができる。中央値の t 検定は通常用いられる平均が中央値に置き換わるだけで、t 値の式も自由度も全く同じであり、従来からの t 分布表が使える。

図 2 は非対称な分布の二つのデータについてヒストグラムにしたものである。この二つの分布の平均値の差の検定と中央値の差の検定、および参考までにノンパラメトリックな検定法としてウィルコクソンの順位和検定も行なって比較する。二つの分布の差の t 検定では t 値は以下ようになる。

$$t = \frac{x_1 - x_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (16)$$

x_1, x_2 : データの平均値または 50% 中央値

標準偏差は $\sigma_1 : 21.46$ $\sigma_2 : 32.68$ 自由度 : 50.5898

ウェルチ検定とする。両側有意水準 $\alpha = 0.05$ 、表中の p は超過確率の両側合計値である。平均値の t 検定では、2 つのデータ分布に差があることになるが中央値の t 検定では差がないことになり、ノンパラメトリック検定の結果も同様に差がない。

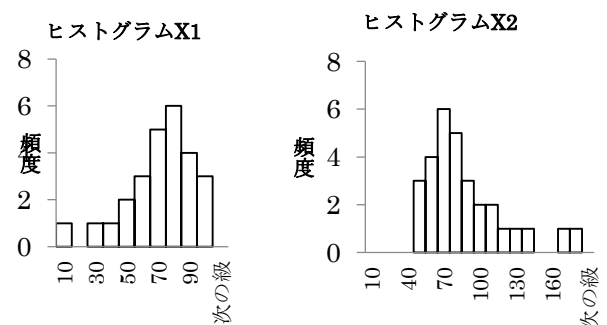


図 - 2 2つのデータ分布

表 1 検定結果

		x_1	x_2	統計量	P値
t検定	平均値	65.96	82.57	2.234	0.030
	中央値	70.01	74.12	0.538	0.593
ウィルコクソン	順位和	658.0	938.0	307.0	0.345