明治末期における小学生の理想人物調査
— キャリブレーション手法の比較 —

土屋　隆裕　データ科学研究系　准教授

【明治末期における小学生の理想人物調査】

讀賣新聞社では1908 (明治41) 年4月21日から5月3日の新聞紙上において，全国の高等尋常小学校長ならびに教員に対し，尋常小学校5，6年生および高等小学校1，2年生を対象に「理想の人物」あるいは「何某の様な人になりたい」ということを調査し，その結果を学級ごとにまとめた上で新聞社宛に送付するよう求めている．その結果は同年5月5日(火)から7月2日(木)の讀賣新聞紙上で，学校・学年・性ごとに報告されている．表1は上位10名の属性別割合である．

表1：上位10名の属性別割合 (%)

	全体	男子	女子	11歳	12歳	北海道・東北	関東	中部	近畿	中国・四国	九州
楠木正成	599 (*11.8*)	*15.0*	*5.5*	*11.3*	*12.4*	*10.2*	*10.4*	*11.2*	*12.7*	*19.7*	*8.2*
二宮尊徳	533 (*10.5*)	*11.2*	*9.2*	*11.0*	*9.9*	*15.1*	*9.8*	*9.8*	*9.7*	*10.5*	*4.8*
豊臣秀吉	492 (*9.7*)	*10.2*	*8.7*	*14.7*	*3.5*	*7.4*	*6.7*	*11.6*	*13.7*	*5.7*	*12.6*
中江藤樹	391 (*7.7*)	*8.8*	*5.5*	*2.3*	*14.3*	*7.2*	*7.4*	*11.2*	*8.3*	*2.9*	*1.4*
東郷平八郎	222 (*4.4*)	*6.2*	*0.8*	*4.7*	*3.9*	*5.5*	*4.6*	*3.6*	*3.9*	*5.9*	*1.4*
ナイチンゲール	204 (*4.0*)	*0.8*	*10.4*	*2.1*	*6.4*	*3.6*	*5.2*	*3.8*	*4.2*	*1.7*	*2.9*
紫式部	135 (*2.7*)	*0.2*	*7.6*	*1.5*	*4.0*	*1.9*	*3.5*	*2.2*	*3.2*	*1.0*	*1.9*
大山巌	122 (*2.4*)	*3.5*	*0.2*	*2.4*	*2.5*	*2.5*	*2.4*	*1.6*	*3.1*	*1.4*	*3.4*
リンカーン	96 (*1.9*)	*1.8*	*2.1*	*2.6*	*1.1*	*5.0*	*1.7*	*1.8*	*0.6*	*1.0*	*1.0*
瓜生岩子	94 (*1.9*)	*0.0*	*5.5*	*1.4*	*2.4*	*1.7*	*2.6*	*2.8*	*1.2*	*0.5*	*0.0*

【どの補助変数を用いてキャリブレーションを行えばよいか】

本研究では，上記の調査データを素材としてウェイトのキャリブレーション手法の比較を行う．一般にキャリブレーションとは，標本 s の抽出方法を反映した，第 i 児童の抽出ウェイト w_i に対し (本研究では $w_i = 400$ とした)，あるgウェイト g_i を乗じて新たなウェイト $w_i g_i$ を求めることである．ただしgウェイト g_i は，新たなウェイト $w_i g_i$ を用いたとき，補助変数ベクトル $\boldsymbol{x}$ に関する母集団総計の推定値のベクトル $\hat{\boldsymbol{\tau}}_{x,\mathrm{C}}$ が母集団 U における真の値のベクトル $\boldsymbol{\tau}_x$ に一致するよう定める．

$$\hat{\boldsymbol{\tau}}_{x,\mathrm{C}} = \sum_{i \in s} w_i g_i \boldsymbol{x}_i = \sum_{i \in U} \boldsymbol{x}_i = \boldsymbol{\tau}_x \tag{1}$$

(1) 式を満たす g_i は何通りも存在するので，g_i に関する距離関数 $G(g_i)$ を定め，加重距離 $\sum_s w_i G(g_i)$ を最小とする g_i を求める．図1は，横軸に示す変数を補助変数として線形関数 $\sum_{i \in s} w_i G(g_i) = \sum_{i \in s} w_i (g_i - 1)^2/2$ を用いたときの，表1の上位10名の推定値である．

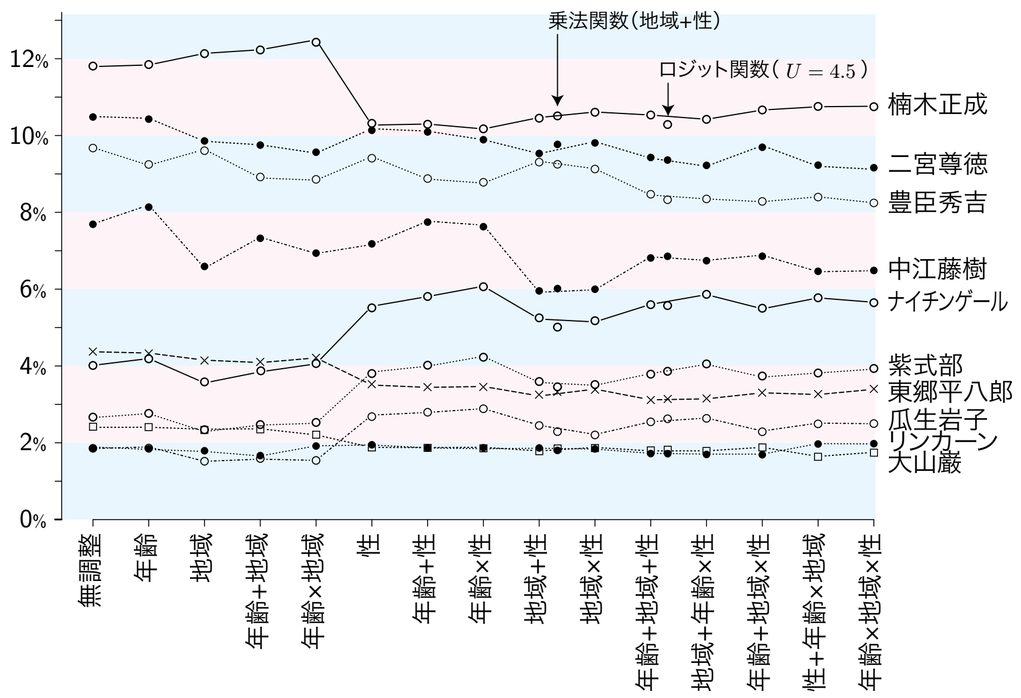

図1：線形関数を用いたキャリブレーション推定値

キャリブレーションに用いる補助変数の選択には二つの基準を用いる．まず，目的とする変数と補助変数との関連の強さに関しては，表1の20名のキャリブレーション前の推定値 $\hat{p}_j$ とキャリブレーション推定値 $\hat{p}_{j,\mathrm{C}}$ との差の大きさで測る．

$$\mathrm{Diff} = 100 \times \sum_{j=1}^{20} |\hat{p}_{j,\mathrm{C}} - \hat{p}_j| \tag{2}$$

次に標準誤差に関連した指標としては不等加重効果を用いる．

$$\mathrm{UWE} = n \frac{\sum_{i \in s} (w_i g_i)^2}{\left(\sum_{i \in s} w_i g_i\right)^2} = 1 + \frac{n-1}{n} CV(w_i g_i)^2 \approx 1 + CV(w_i g_i)^2 \tag{3}$$

図2によれば，“年齢+性” や “年齢x性” を用いればよさそうである．

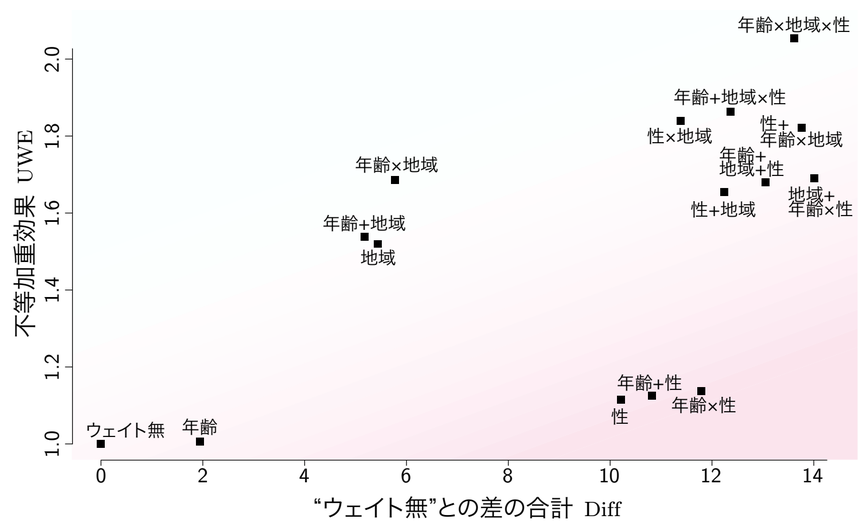

図2：各キャリブレーションウェイト (線形関数) の不等加重効果

【範囲制約をどう課せばよいか】

極端な g_i が得られることを避けるため，乗法関数 $\sum_{i \in s} w_i G(g_i) = \sum_{i \in s} w_i \{g_i \log(g_i) - g_i + 1\}$ によるウェイトを $g_i^* = c_i g_i$ とトリミングする方法，

$$c_i = \begin{cases} L/g_i & \text{if } g_i < L \\ a & \text{if } L \le g_i \le U \\ U/g_i & \text{if } g_i > U \end{cases} \tag{4}$$

切断型の乗法関数

$$G(g_i) = \begin{cases} g_i \log(g_i) - g_i + 1 & \text{if } L \le g_i \le U \\ \infty & \text{otherwise} \end{cases} \tag{5}$$

を用いる方法，ロジット関数

$$\sum_{i \in s} w_i G(g_i) = \sum_{i \in s} w_i \left[\frac{1}{A} \left\{ (g_i - L) \log \frac{g_i - L}{1 - L} + (U - g_i) \log \frac{U - g_i}{U - 1} \right\} \right] \tag{6}$$

を用いる方法という，g_i に範囲制約を課す三つの方法を比較する．用いる補助変数は “年齢＋地域＋性” とする．乗法関数による不等加重効果は1.783であり，最大の g_i は $g_i = 8.138$ であった．そこで下限は $L = 0$ とし，上限 U を8.0から4.0まで0.5ずつ小さくしながら，上記三つの方法で g_i を求めた．図3には g_i の箱ヒゲ図を，最大値を黒丸として示す．ロジット関数は他の二つに比べて上限 U の影響を受けにくい．

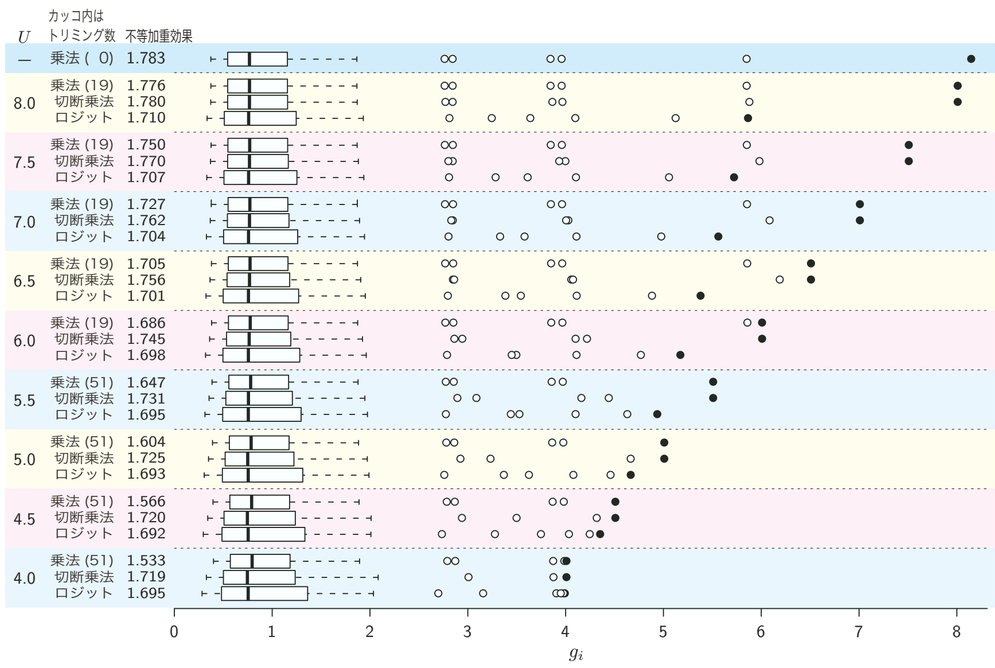

図3：各 U に応じた g_i の箱ヒゲ図