

OMR の読み取りエラー解析の ための確率モデル

太 田 拓 男

1. は じ め に

光学式マーク読み取り式装置 (Optical Mark Reader, OMR) が、試験の解答入力に広く利用されている。しかし、この OMR は僅かな汚れをマークとして読みとったり、逆に薄いマークを読み落とすことがあるため、通常、以下のような方策がとられる。まず、一連のマークポジション (マーク欄) 中の一つにマークがなされるようにし、複数のマークを禁止する。この条件下で複数のマークが検出された、あるいは一つもマークが検出されなかったマーク欄を持つシートは読み取りが正しく行なわれなかった可能性があるとして抽出され、これについて直接人の目によるマーク位置の判断が行われる。

機械読み取り作業は、このように人の目による判断が部分的に必要なため、その高速性を十分に発揮できない場合が生じる。この作業を軽減し機械読み取りの効果を高めるために汚れの主たる原因であるマークの消し残しを減らす工夫がなされる。しかし、その効果の測定は、次の理由によって簡単ではない。解答者が書き直しを行なう頻度がマークの消し残しの出現に影響を与えるためである。

ここでは、この複数マーク発生の要因である書き直しと消し残しを分離し、消し残し発生抑制の効果を測定できるようにするための確率モデルを

提示する.

2. 基本モデル

個人 n は, 確率分布 $\mathbf{p}^{(n)} = (p_1^{(n)}, p_2^{(n)}, \dots, p_k^{(n)})$ に従って対象 $o_1, o_2, \dots, o_k$ よりの選択を独立に 2 回行なう. 1 回目と 2 回目の選択結果が異なる場合, 確率 $1 - \beta$ で 1 回目の選択を取り消す. つまり, 確率 β で 1 回目の選択の取り消しに失敗し, 2 ケ選択 (複数マーク) を行なうことになる.

まず, いくつかの基本的事象の確率と確率変数の期待値を求める. 「個人 n が o_i を選択」のインディケータを $\delta_i^{(n)}$ とおけば,

$$\begin{aligned} (1) \quad E(\delta_i^{(n)}) &= p_i^{(n)^2} + p_i^{(n)}(1 - p_i^{(n)})\beta + (1 - p_i^{(n)})p_i^{(n)} \\ &= p_i^{(n)}(1 + (1 - p_i^{(n)})\beta) \end{aligned}$$

である. 個人の総数を N として, この中での o_i の選択者の総数 $\sum_n \delta_i^{(n)}$ について

$$\begin{aligned} (2) \quad E(\sum_n \delta_i^{(n)}) &= \sum_n p_i^{(n)}(1 + (1 - p_i^{(n)})\beta) \\ &= \sum_n p_i^{(n)} + \beta(\sum_n p_i^{(n)} - \sum_n p_i^{(n)^2}) \end{aligned}$$

が成立する. 個人 n が 2 ケ選択を行なう確率は,

$$\begin{aligned} (3) \quad P(\sum_i \delta_i^{(n)} = 2) &= \sum_i p_i^{(n)}(1 - p_i^{(n)})\beta \\ &= \beta(1 - \sum_i p_i^{(n)^2}) \end{aligned}$$

であり, N 人中 2 ケ選択を行なう個人の割合の期待値は,

$$\begin{aligned} (4) \quad \sum_n P(\sum_i \delta_i^{(n)} = 2) / N &= \sum_n \beta(1 - \sum_i p_i^{(n)^2}) / N \\ &= \beta(1 - \sum_n \sum_i p_i^{(n)^2} / N) \end{aligned}$$

$$=\beta(1-\sum_i \bar{p}_i^2 - \sum_i 1/N \sum_j (p_i^{(n)} - \bar{p}_i)^2)$$

となる。ここで $\bar{p}_i = 1/N \sum_j p_i^{(n)}$ である。

第3項は個人間の異質性を表わしている。すなわち同質な集団であればこの項は0となり、

$$(5) \quad \sum_j P(\sum_i \delta_i^{(n)}=2)/N=\beta(1-\sum_i \bar{p}_i^2)=\beta(1-\sum_i p_i^{(n)2})$$

となる。これは個人の選択確率のエントロピー $-\sum_i p_i^{(n)} \log p_i^{(n)}$ との近似的な比例関係にあり、個人が何を選択するかについて曖昧な状況にあれば、それだけ複数マークが発生しやすいことを示している。

一方、 $p_i^{(n)}$ の値が1または0、つまり選択が deterministic である場合は、 $p_i^{(n)2}=p_i^{(n)}$ であり、(4)式は、 $\beta(1-\sum_j \sum_i p_i^{(n)2}/N)=\beta(1-\sum_j 1/N)=0$ となり、2ヶ選択者の割合が0となることが確かめられる。この場合、個人間の異質性は最大となっている。

つぎに N 人中よりランダムに選ばれた個人が2ヶ選択を行なっているとき、 o_i が選ばれている条件付確率を求める。抽出された個人が対象 o_i を選んでいることを示すインディケータを δ_i 、抽出される個人を X とすれば、

$$\begin{aligned} (6) \quad P(\delta_i=1 \mid \sum_j \delta_j=2) &= P(\delta_i=1 \cap \sum_j \delta_j=2) / P(\sum_j \delta_j=2) \\ &= \frac{\sum_n P(\delta_i^{(n)}=1 \cap \sum_j \delta_j^{(n)}=2 \mid X=n) \cdot P(X=n)}{\sum_n P(\sum_j \delta_j^{(n)}=2 \mid X=n) \cdot P(X=n)} \\ &= (\sum_j 2p_i^{(n)}(1-p_i^{(n)})\beta/N) / (\sum_j (1-\sum_i p_j^{(n)2})\beta/N) \\ &= 2(\sum_j p_i^{(n)} - \sum_j p_i^{(n)2}) / (N - \sum_j \sum_i p_j^{(n)2}) \end{aligned}$$

である。2ヶ選択を行なった個人に限定することによって取り消し失敗の確率 β に無関係な値が導出されたことになる。

なお、取り消し失敗の確率 β が個人によって異なるとするモデルを想

定することもできる. この場合, 個人 n の取り消し失敗の確率 $\beta^{(n)}$ の分布 b が選択分布の分布と独立であれば, $\beta = E_b(\beta^{(n)})$ とすることでこのモデルを基本モデルと同じものとみなすことができる.

3. 選択分布をパラメタ化したモデル

2 ケ選択者の発生頻度の分析が目的であるから選択確率の $(k-1) \cdot N$ ケの要素を推定する必要はなく, その特性値

$$(7) \quad \phi_i = 1 / N \sum_{\pi} p_i^{(n)} = \bar{p}_i,$$

$$(8) \quad \phi_i = 1 / N \sum_{\pi} p_i^{(n)^2},$$

および取り消し失敗の確率 β に関する推定を試みる. 観測できるデータを以下に示す.

N_i : o_i を選択した個人の総数, $\sum_{\pi} \delta_i^{(n)}$

N_i^0 : $\beta = 0$ であるときの $\sum_{\pi} \delta_i^{(n)}$, この値は, 複数マーク検出シートに対する目視によって得られる.

D : 2 ケ選択をした個人の総数

R_i : 2 ケ選択をした個人の中での o_i 選択者の割合

N_i^0 / N は選択確率 $\beta = 0$ としたときの (2) 式によって ϕ_i の不偏推定量 $\hat{\phi}_i$ である. また, D は, (4) によって $\beta(1 - \sum_j \phi_j)$ の不偏推定量であるが, β 単独の, あるいは $(1 - \sum_j \phi_j)$ 単独の推定量をこれによって構成することはできない. 一方, R_i は (6) 式によって $2(\phi_i - \phi_i) / (1 - \sum_j \phi_j)$ を確率とする事象の生起の割合であるからこの量の不偏推定量である. しかし, これより

$$(9) \quad R_i = 2(\hat{\phi}_i - \phi_i) / (1 - \sum_j \phi_j)$$

を満たす ϕ_j の値を一意に求めることはできない。なぜなら $\phi = \sum_j \phi_j$ のいかなる値に対しても (9) を満たす ϕ_j が存在するからである。

そこで選択確率 $p^{(n)}$ をパラメタ化して、 ϕ_i, ϕ_i に条件を設けることにする。

ケース 1

一人の個人の選択分布は、異質な集団 ($p_i = 0$ または 1 の選択分布の族 f) か、完全に同質な集団 g のいずれかに、それぞれ確率 $(1-\lambda)$, λ で属する。異質な集団からは、確率 ϕ_i で $p_i = 1, p_j = 0 (j \neq i)$ なる選択分布がとられる。完全に同質な集団は、 $p_i = \phi_i$ なる選択分布である。つまり、選択分布の分布は、混合パラメタ λ の分布 f と一つの選択分布 g の混合 h である。このとき、

$$\begin{aligned} E_f(p_i) &= E_f(p_i^2) = \phi_i, \\ E_g(p_i) &= \phi_i, \quad E_g(p_i^2) = \phi_i^2, \end{aligned}$$

であるから

$$\begin{aligned} E_h(p_i) &= (1-\lambda)\phi_i + \lambda\phi_i = \phi_i, \\ E_h(p_i^2) &= (1-\lambda)\phi_i + \lambda\phi_i^2 = \phi_i \end{aligned}$$

である。個人が 2 ケ選択を行なう確率は (3) 式をこの p の分布 h で混合化して、

$$\begin{aligned} (10) \quad P(\sum_i \delta_i = 2) &= E_h \beta (1 - \sum_i p_i^2) = \beta (1 - (1-\lambda) - \lambda \sum_i \phi_i^2) \\ &= \beta \lambda (1 - \sum_i \phi_i^2) \end{aligned}$$

となる。

ケース 2

個人の選択確率 $\mathbf{p}$ は, Dirichlet 分布 (パラメタ $\alpha_1, \alpha_2, \dots, \alpha_k, \alpha_i > 0$) に従っている, つまり

$$(p_1, p_2, \dots, p_k) \sim \Gamma(\sum_j \alpha_j) / \prod_j \Gamma(\alpha_j) \cdot \prod_j p_j^{\alpha_j - 1}$$

であり, $A = \sum_i \alpha_i, \phi_i = \alpha_i / A$ とおけば,

$$E(p_i) = \phi_i,$$

$$E(p_i^2) = \phi_i(A\phi_i + 1) / (A + 1) = \phi_i$$

となる. 個人の 2 ケ選択の確率は (3)式をこの $\mathbf{p}$ の分布で混合化して,

$$\begin{aligned} (11) \quad P(\sum_i \delta_i = 2) &= E\beta(1 - \sum_i p_i^2) = \beta(1 - (A \sum_i \phi_i^2 + 1) / (A + 1)) \\ &= \beta \cdot A / (A + 1) \cdot (1 - \sum_i \phi_i^2) \end{aligned}$$

となる. $\lambda = A / (A + 1)$ とおけば, ケース 1 の (10)式と等しくなる.

ケース 2 における $A \rightarrow \infty$, つまり $A / (A + 1) \rightarrow 1$ とケース 1 における $\lambda \rightarrow 1$ は, とともに選択確率を同質な分布に近づけ, $E(p_i) = \phi_i$ の条件下での p_i の分散を最小値 0 に, 従って 2 ケ選択の確率を最大値に近づける.

一方, ケース 2 における $A \rightarrow 0$, つまり $A / (A + 1) \rightarrow 0$ とケース 1 における $\lambda \rightarrow 0$ は, とともに $E(p_i) = \phi_i$ の条件下での p_i の分散を最大値に, 2 ケ選択の確率を 0 に近づける.

いずれのケースも 2 ケ選択の確率は, 以下の 3 つの要因,

- ・ 選択取り消し失敗の確率
- ・ 集団の同質性
- ・ 集団で観測される選択分布の曖昧性

に対応する値の積であることが示されている.

4. 推定の問題

集団で観測される選択分布の曖昧性に対応する $(1 - \sum_i \phi_i^2)$ は, $(1 - \sum_i \hat{\phi}_i^2)$ によって推定できる. しかし, 残りの 2 ケの要因は, 積の形でしか観測されず, 一組のデータからこれらを分離することは不可能である. そこで選択取り消し失敗の確率が変わらないとされる条件下で行なわれた独立な試行による 2 組のデータを用いることができる場合を考える. これより, それぞれの組のパラメタ, 統計量にインデックス I, II を加えて区別することにする. それぞれの試行について, 集団で観測される選択分布の曖昧性の推定値 $(1 - \sum_i \hat{\phi}_{Ii}^2)$, $(1 - \sum_i \hat{\phi}_{IIi}^2)$ が得られる.

(10), (11) より D_I, D_{II} はそれぞれパラメタ $(N, \beta\lambda_I(1 - \sum_i \phi_{Ii}^2))$ と $(N, \beta\lambda_{II}(1 - \sum_i \phi_{IIi}^2))$ の 2 項確率変数である. N が大となれば, その比 $D_I / D_{II} = (D_I / N) / (D_{II} / N)$ の分子, 分母の分布が共に一点へ集中してくるため, その期待値 $E(D_I / D_{II})$ は,

$$(12) \quad E(D_I / N) / E(D_{II} / N) = \lambda_I \cdot (1 - \sum_i \phi_{Ii}^2) / (\lambda_{II} \cdot (1 - \sum_i \phi_{IIi}^2))$$

に近づく. これより (12) の一致推定量として D_I / D_{II} を用いることができる.

さらに, ϕ_{Ii}, ϕ_{IIi} をその推定値 $\hat{\phi}_{Ii}, \hat{\phi}_{IIi}$ で代用して, λ_I / λ_{II} の推定量として

$$(13) \quad D_I / D_{II} \cdot (1 - \sum_i \hat{\phi}_{IIi}^2) / (1 - \sum_i \hat{\phi}_{Ii}^2)$$

が得られる.

一方, 同一データの異なる 2 種の読み感度での読み取りより得られた 2 組のデータについては, 2 ケ選択者数比, D_I / D_{II} が, $\lambda, 1 - \sum_i \phi_i^2$ の共通性より β_I / β_{II} の一致推定量となる. D_I, D_{II} は, 1, 2 回目に異なる選択をした個人について確率 β_I, β_{II} で事象の生起の観測を行なった結果

であるから、この比 D_I / D_{II} は、パラメタ (N, β_I) と (N, β_{II}) の 2 項確率変数の比である。

このように λ, β ともに特定の試行における値との比としてしか推定できない。しかし、これらのパラメタは、2 個選択者数に積の形でのみ影響を与えるため、2 個選択者数の変化を分析するのに支障ない。

5. お わ り に

複数選択のモデルとして、2 回以上の多段選択、各段階での異なる選択確率の設定が考えられる。しかし、観測可能なデータの種類と件数からみて、ここで述べた単純な 2 回選択のモデルが実用上適当と考えられる。

なお、本学の統計研究会のメンバより多くの示唆を頂いた。とくに丘本教授からは、選択分布のパラメタ化に関してアイデアを頂いた。これらの先生方に感謝の意を表します。

(平成 2 年 10 月 15 日受理)