数理解析研究所講究録
第345巻 1979年 36-42

累積されたデータ・ファイルからの推測について、

九大 理 基礎情報研
浅野長一郎

1. はじめに

近年、種々の社会調査などの資料がコンピュータのデータ・ファイルに格納され、必要に応じ、その集積によって対象母集団の特性が論じられる場合が多くなってきている。すなわち、コンピュータによる情報化の今日、このような調査結果の蓄積がデータ・ベースとかデータ・バンクと呼ばれ構成されいるのを各所で見かける。そして、このような資料の集積や情報処理が意外に機械的に無雑作になされ、統計学的推測の立場から正しい情報処理になっているのか否か疑問になることがある。

本報告者らは、かねて新・統計プログラム・パッケージ「NISANシステム」の研究開発を進めているが、この仕事にも関連してデータの集積法やデータ・ファイルの利用法には種々の工夫や注意の必要性を痛感している。すなわち、一般に

調査は何らかの制約とか枠内で行われる。このような毎回の調査成績は、数種類の条件つきデータであり、併用に際して整合性を忘れてはならない。

本報文は、上記のような立場から、問題を多次元情報の場合よりも一方端的に単純な場合として、報告者の嘗ての研究にもとづき、集収情報の処理方式に関する基本的な推測手順について論じる。

2. 集積情報による推測

上記のように、それぞれ時間的・空間的また金貨的な規模のまちまちな調査資料が逐次的に集収される場合、その累積による包括的な母集団（対象集団）の規定とそれが事後的に規定された後の所謂"調査データの集積後の跡始末"の方法論という局面に遭遇する。

母集団規定については、何らかの制約条件下で集積された資料の内容と対象の実態に関して弁別・傾向・相関・回帰など統計学に独自な諸概念が導入される筈である。とともに、対象集団を観測した種々のサンプリング技法にも依存して調査対象が論じられる。ここでは、この母集団規定が充分になされたものとしておく。

この段階では、規定された母集団の特性を推測するために、

出来るだけ多くの、または幾つかの調査資料を併用する処理方式の問題となる。一つのファイルの上に累積併合した結果をそのまま無雑作に利用すると誤った結論に導いてしまうことがあるので、この処理方式はデータ・ファイルの構成にも関係することになる。

3. 簡単な例証

いま、幾つかの属性をもつ対象集団を考え、この属性百分率、すなわち母集団の多項確率を推定する問題を考えよう。ただし、この際の観測調査は部分的に制約され、不完全資料も相当に含まれているとする。

これらの調査資料の事情を母集団属性に関してみると、一般に Nested case と Chained case の2つの場合の混合でできていると云える。次図で、これを典型で示しておく。

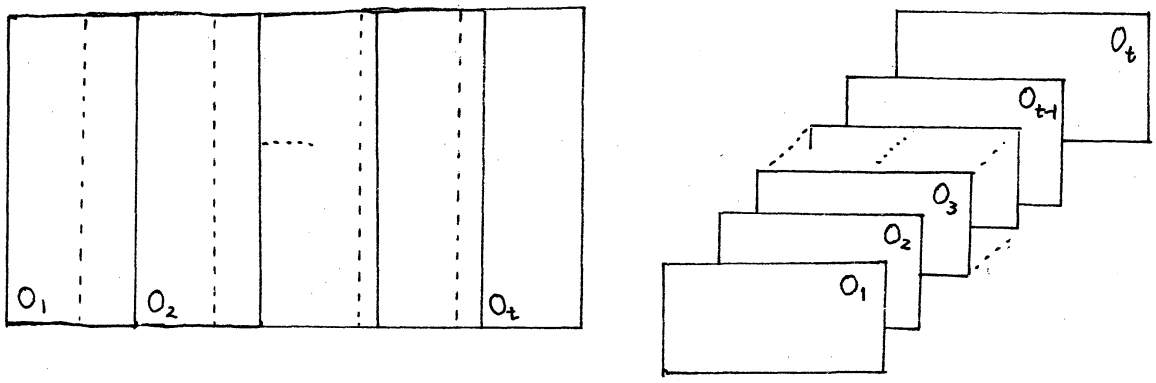

Fig.1 Chained case

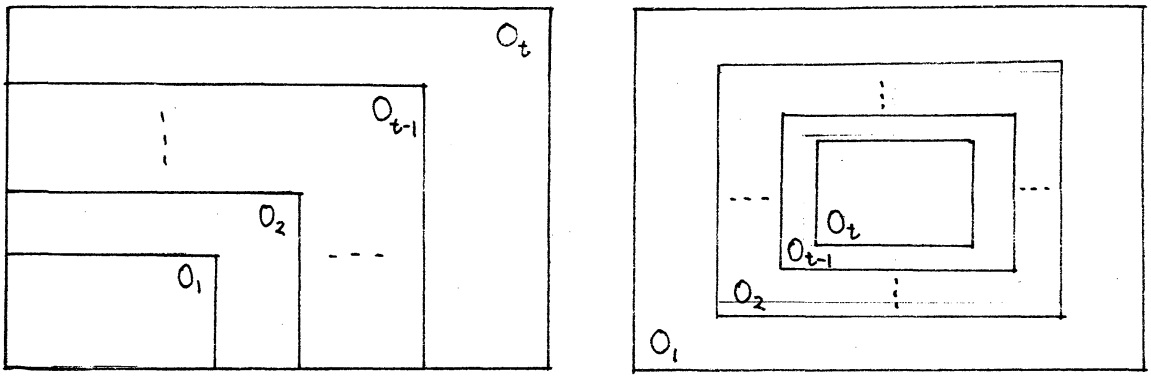

Fig.2 Nested case

次に、簡単に記号の説明と最尤推定の結果のみを示しておく。

3.1) Chained case

いま、各調査における母数空間を Ω_i, $i=1,2,\cdots,t$, とし、それらの部分空間 ω_{ij} を次のようにおく。

$$\omega_{i1} \equiv \Omega_{i-1} \cap \Omega_i \quad , \quad \omega_{i3} \equiv \Omega_i \cap \Omega_{i+1} \ ,$$

$$\omega_{i2} \equiv \Omega_i - \omega_{i1} - \omega_{i3} \ , \qquad i=1,2,\cdots,t.$$

ここに、 $\Omega_0 \equiv \Omega_{t+1} \equiv \phi$ (空集合)

このような ω_{ij} 中の多項確率の和を P_{ij}、また $\{\omega_{ij}\}$ を更めて通し番号 u により $\{\omega_u\}$、その ω_u 中の個々の多項確率を p_{uv}, $v=1,2,\cdots,l_u$、ω_u 中の p_{uv} の和を P_u と記すれば、すべての $v \in \omega_u$, $u=1,2,\cdots,2t-1$, について p_{uv} の最尤推定量は次のように得られる[1]。

$$\hat{p}_{uv} = \left[\sum_{\{i \mid \Omega_i \supset \omega_u\}} n_{uv}^{(i)} \Big/ \sum_{\{i \mid \Omega_i \supset \omega_u\}} n_{u\cdot}^{(i)} \right] \cdot \hat{P}_u$$

ここに $\hat{P}_1 = \dfrac{n^*_{11}}{N_1} \Big/ \left[1 + \sum_{g=2}^{t} \dfrac{n^*_{12}}{N_1} \dfrac{\prod_{k=2}^{g-1} n^*_{k3} (n^*_{g2} + n^*_{g3})}{\prod_{k=2}^{g} n^*_{k1}} \right]$,

$$\hat{P}_u = \hat{P}_{2i-3+j} = \left[\prod_{k=2}^{i-1} n^*_{k3} n^*_{ij} \Big/ \prod_{k=2}^{i} n^*_{k1} \right] \Big/ \left[\frac{N_1}{n^*_{12}} + \sum_{g=2}^{t} \frac{\prod_{k=2}^{g-1} n^*_{k3} (n^*_{g2} + n^*_{g3})}{\prod_{k=2}^{g} n^*_{k1}} \right],$$

for $u = 2, 3, \cdots, 2t-1$, $t \geq i \geq 2$, $j = 1, 2, 3$

また、これらは 不偏性・一致性・十分性を有している。

3.2) Nested case

この場合は $\Omega \equiv \Omega_1 \supset \Omega_2 \supset \cdots \supset \Omega_t$ で示される。また $\bigcup_{i=1}^{t} \Omega_i$ の分割 $\{w_u\}$ は、

$$\sum_{u=1}^{t} w_u = \bigcup_{i=1}^{t} (\Omega_i \cap \Omega^c_{i+1}) = \Omega$$

ここに $\Omega^c_{t+1} \equiv \Omega$ とする。

さて、このようにして、上記 3.1) と同じ意味の p_{uv} の最尤推定量は次式で示される[1]。

$$\hat{p}_{uv} = n_{uv} \Big/ \sum_{j=1}^{t} \frac{n^{(j)}_v}{1 - \sum_{r=0}^{j-1} \sum_{w_r} \hat{p}_{rv}} ,$$

ここに、$w_0 = \phi$, $p_{0v} = 0$ 。

[註] より詳細な記号の説明は原著参照のこと。

この推定量の形は、$t=3$ とすると次のように判り易い。

$$\hat{p}_{1v_1} = n_{1v_1} / N_1 \qquad \text{for } v_1 \in w_1 ,$$

$$\hat{p}_{2v_2} = n_{2v_2} \Big/ N_1 \left\{ 1 + n^{(2)}_{v_2} \Big/ N_1 \left(1 - \sum_{w_1} n^{(1)}_{v_1} / N_1 \right) \right\} \qquad \text{for } v_2 \in w_2 ,$$

$$\hat{p}_{3v_3} = n_{3v_3} \Big/ N_1 \left\{ 1 + n^{(2)}_{v_3} \Big/ N_1 \left(1 - \sum_{w_1} n^{(1)}_{v_1} / N_1\right) + n^{(3)}_{v_3} \Big/ N_1 \left[1 - \sum_{w_1} n^{(1)}_{v_1} / N_1 - \sum_{w_2} n_{2v_2} / N_1 \left(1 + n^{(2)}_{v_2} / N_1 (1 - \Sigma n^{(1)}_{v_1} / N_1) \right) \right] \right\}$$

3.3) General case

調査資料に関する一般的な集積の形態は、上記の2つの典型の混合でなりたっている。しかも、それぞれの式の形はある一定の規則があり、そのアルゴリズムは比較的容易である。この general case の解はむしろ複雑なのでここでは省略するが、この場合の推定計算のプログラミングはむしろ楽である。

4. むすび

種々の調査資料を累積して、データ・ファイルが構成されている昨今、これを利用しての推測には、なほ解決してゆかねばならない幾つかの重要問題がある。本報では、最も簡単な場合について論じたが、情報容量の圧縮の上からは分割表の解析に関しても多くの見直しが必要であらう。また、これらが多次元情報で与えられる場合の情報処理方式はより複雑であるが、日常的な必要性も多いと思われる。

このように、コンピューターによる情報処理として、大量高次元情報ファイルの活用法には統計学的研究対象が多く外

在している。たとえば、具体的に

a) 増加するデータ・容量の圧縮法(上述)

b) データ群の廃棄論

c) 次元の低減法

d) データの併用・併合の方法論

e) 数学モデル構成法 ── その推測システム

などがあげられる。

[参考文献]

[1] Asano, Ch., "On estimating multinomial probabilities by pooling incomplete samples", Ann. Instit. Stat. Math., v.17, No.1, 1965.

[2] Batschelet, E., "Spurious correlation of the age of onset, with special reference to atopic diseases", Biometr. Z., v.3, 1962.

[3] Geppert, M.P., "Erwartungstreue plausibelste Schätzer aus dreieckig gestutzten Kontingenztafeln", Biometr. Z., v.3, 1961

[4] Li, C.C., "Human Genetics, Principles and Methods", McGraw Hill Book Co., 1961

[5] Watson, G.S., "Missing and mixed-up frequencies in contingency tables", Biometrics, v.12, 1956.