2010年7月9日　統計数理研究所　オープンハウス

間接質問法の新展開

土屋　隆裕　　データ科学研究系　准教授

統計調査の現代的課題と間接質問法

調査環境は依然として深刻な状況が続いている。例えば「日本人の国民性調査」の回収率は、平成20年には52%と過去最低となった。原因の一つとしては個人情報・プライバシーに対する意識の高まりの影響があろう。個人情報に敏感な人々の信頼感をいかに獲得し、調査への協力率をいかに上げるかが、統計調査の最重要課題の一つである。

しかし見方を変えれば、似たような問題は昔から存在している。例えば社会的に望ましくない行為の経験の有無や、性行動などの非常に私的な事柄、違法な薬物使用経験などの犯罪行為については、回答者によっては率直に回答したくないのが普通である。このような、回答者によっては率直に回答したくない内容を調査する手法の開発は、統計調査における長年の課題であり、その対処法の一つが 間接質問法である。

Item Count法

例えば調査の目的は、サイズ N の母集団 U において「成人前に喫煙したことがある」(キー項目と呼ぶ) 人の割合 π を知ることであるとする。

$$\pi=\frac{1}{N}\sum_{i\in U}Y_i,\quad Y_i=\begin{cases}1 & \text{：個人 } i \text{ が未成年喫煙の経験があるとき}\\ 0 & \text{：それ以外のとき}\end{cases} \tag{1}$$

間接質問法の一つである **Item Count法**では、キー項目の他に複数の非キー項目を用意する。非キー項目だけを並べたリストAを短リストと呼び、短リストにキー項目を加えたリストBを長リストと呼ぶ。

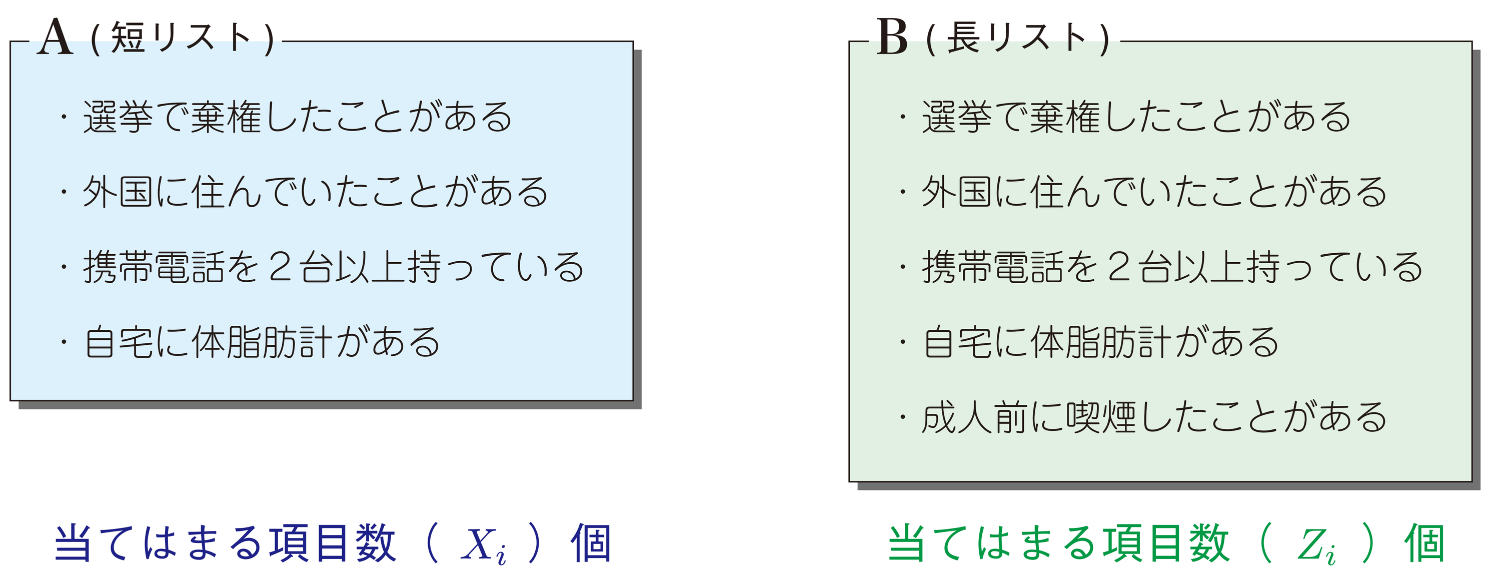

図1. Item Count法

次に大きさ n のサンプルを、大きさ n_{A} と n_{B} の二つの等質な下位サンプルA群とB群に分割する。A群には短リストを提示し、その中で「当てはまる項目の数」X_i だけを答えてもらう。B群には長リストを提示し、同様に「当てはまる項目の数」Z_i だけを答えてもらう。

ところで $Z=X+Y$ であるので次式が成り立つ。

$$\pi=\frac{1}{N}\sum_{i\in U}Y_i=\frac{1}{N}\sum_{i\in U}(X_i+Y_i)-\frac{1}{N}\sum_{i\in U}X_i=\bar{Z}-\bar{X} \tag{2}$$

$\bar{Z}$ と $\bar{X}$ はそれぞれ Z と X の母集団平均である。$\bar{Z}$ と $\bar{X}$ をそれぞれの推定量 $\hat{\bar{Z}}$ と $\hat{\bar{X}}$ で置き換えれば π の推定量となる。

$$\hat{\pi}=\hat{\bar{Z}}-\hat{\bar{X}} \tag{3}$$

二重リスト法

上述の方法ではキー項目はB群だけに提示される。 二重リスト法では A群とB群の役割を入れ替え、A群にも長リストを提示することでA群からも情報を得る。ただし非キー項目は先ほどとは別のものを用いる。

二重リスト法

問	A群	B群	推定量
1	短リスト X	長リスト Z	$\hat{\bar{Z}}-\hat{\bar{X}}$
2	長リスト W	短リスト V	$\hat{\bar{W}}-\hat{\bar{V}}$

推定量は、A群に長リストを割り当てたときの $\hat{\bar{W}}-\hat{\bar{V}}$ とB群に割り当てたときの $\hat{\bar{Z}}-\hat{\bar{X}}$ を (加重) 平均すればよい。

$$\hat{\pi}=\frac{n_{\mathrm{A}}(\hat{\bar{W}}-\hat{\bar{V}})+n_{\mathrm{B}}(\hat{\bar{Z}}-\hat{\bar{X}})}{n} \tag{4}$$

Aggregated Response法

キー項目 Y が量的な変数のときのItem Count法の変形が **Aggregated Response法**である。例えば年収 Y の母集団平均を調べたいものとする。まず2つの等質なサンプルA群とB群を用意する。A群には年収 Y (万円) に例えば電話番号の下4桁 X を加えた数値 $V=X+Y$ を回答してもらう。B群には電話番号の下4桁 X から年収 Y を引いた数値 $W=X-Y$ を回答してもらう。電話番号が秘匿されていれば、回答の数値を見ても回答者の年収は分からない。

ところで

$$E(Y)=\frac{E(X+Y)-E(X-Y)}{2}=\frac{E(V)-E(W)}{2} \tag{5}$$

が成り立つ。つまりA群とB群の回答の平均値の“差”を2で割れば平均年収 $\bar{Y}$ を推定できる。また平均値の“和”を2で割れば、電話番号の下4桁 X の母集団平均を推定することも可能である。

Three-Card法

無免許で自動車を運転している人の割合を推定したいものとしよう。話を簡単にするため、運転免許の区分は普通自動車、中型自動車、大型自動車の3区分のみとする。まず自動車を運転している人から成るサンプルを3つの等質な下位サンプルA群・B群・C群に分割する。

次に **Three-Card法**では図2のような3枚のカードを用意する。A群にはカードAだけを提示し、どの自動車まで運転できる免許を持っているかを「1」または「2」というグループ名だけで答えてもらう。無免許の人はグループ「2」に属すが、「2」と回答しても無免許であることは秘匿される。同様にB群にはカードBだけ、C群にはカードCだけを提示し、属しているグループ名だけを答えてもらう。

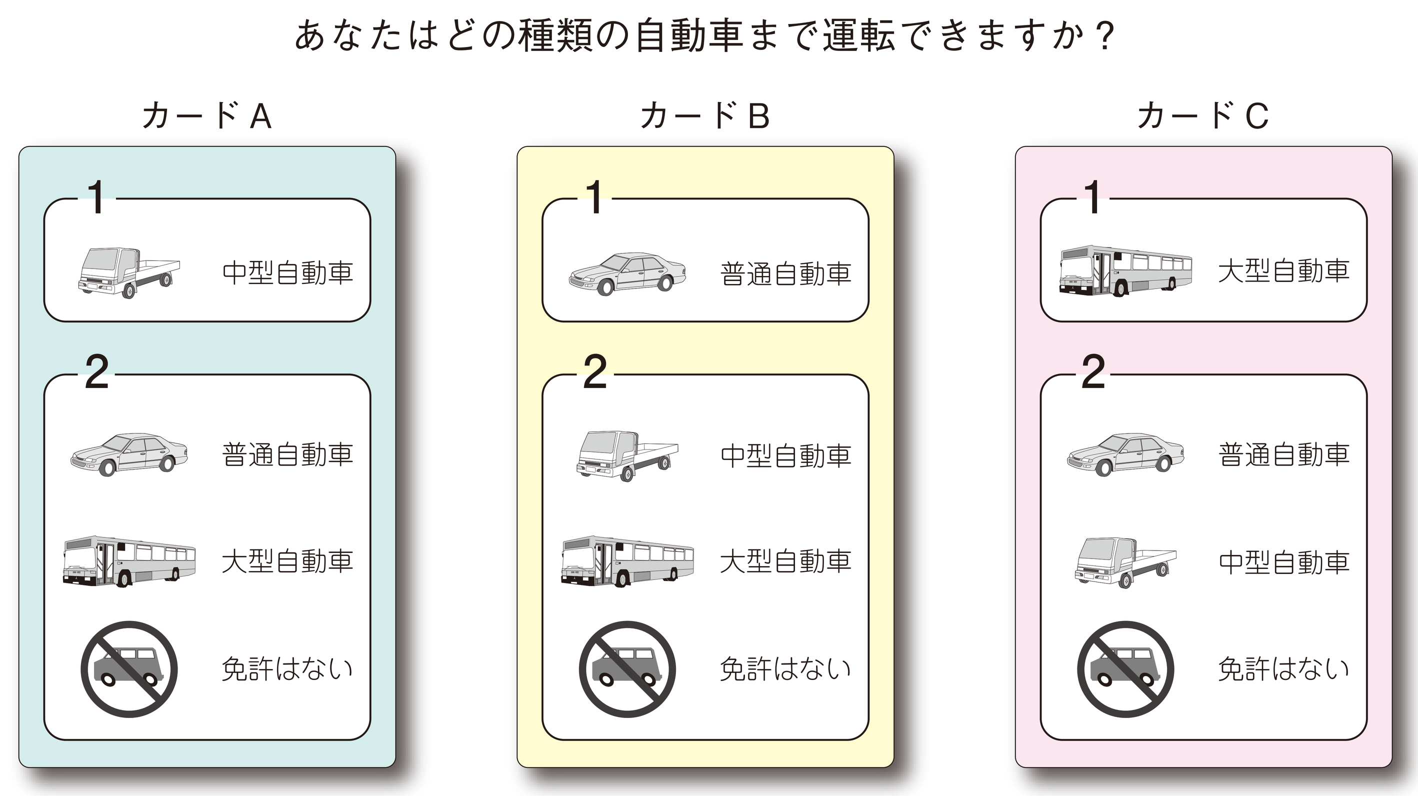

図2. Three-Card 法

カードAを提示したA群の中で、グループ「2」と回答する人の割合は、調査目的である無免許の割合に普通自動車と大型自動車の割合を足したものである。

$$\text{グループ「2」（カードA）}=\text{普通自動車}+\text{大型自動車}+\text{無免許} \tag{6}$$

ところで普通自動車の割合はカードBを提示したB群で「1」と回答する人の割合、大型自動車の割合はカードCを提示したC群で「1」と回答する人の割合として推定できる。これらをカードAの「2」の割合から引けば、A群における無免許の人の割合を推定することができる。

同様の考え方で、B群・C群それぞれにおける無免許の人の割合を推定できる。サンプル全体を用いて無免許の人の割合を推定するには、それらを (加重) 平均すればよい。