統計十話

第7話 測定誤差・測定データの変動の評価なくして統計的分析の意味はない——その3

林 知己夫
文部省統計数理研究所

今度は計測されたものか質的データ，つまりカテゴリカルな表現をとった場合を考えよう．これに誤差がある場合はどうなるか．

実例として尿検査，糖，蛋白の＋，±，－という表現，レントゲン写真の読みである項目の「あり，疑，なし」，心電図の読みの表現など思い浮かべていただければよい．これは，ミス・クラシフィケーション（誤分類）という表現で医学に登場してきたのであるが，データ解析の領域では，はるか昔から論議のあった問題である．

被測定者の一人一人を突き合わせてみれば，確かに測定結果の不一致といった現象はあるが，全体としての結果は一致している場合が多い．つまり個人の測定結果は一致しないが，測定結果を積み上げたマージナルは一致するという典型的なランダム・レスポンスである．こうしたことはいかに注意しても出てくるのであるから，これにたじろがず，これを突き抜けてそのなかにある本質的ともいえるものを導き出すデータの解析方法を考えなければならないのである．

一つの例

このため，簡単な例をあげておこう（**表1**）．

測定結果は，＋(ある)，－(ない) の二つであるとしよう．本来は＋であるが，＋と正しく測定される確率が0.7，－と測定される確率が0.3，本来は－であるが，－と正しく測定される確率が0.8，＋と測定される確率が0.2としよう．これを**表1**で示すと上のようになる．本来＋のものが5,500人（55%），－のものが4,500人（45%）としよう．＋が10%多いのである．これはわからないのである．ところが，データで＋のものは平均的に，

表1 誤差の出方

	データ / 本来	＋	－
5500人	＋	0.7	0.3
4500人	－	0.2	0.8
データ		4750人	5250人

$$5500\times0.7+4500\times0.2=4750$$

－のものは平均的に，

$$5500\times0.3+4500\times0.8=5250$$

となり，－のほうが5%多く出てしまう．データからそのまま結論すると，－のほうが多くなって本来の結果を誤ってしまうことになる．回答に確率的な変動があるにもかかわらず，これを無視したために結果が逆転し，誤りを犯したことになる．正しい結論を出すためにはデータに変換を施さねばならないが，このためには，確率的な反応の様相を基礎研究ではっきりつかんでおかねばならないので

表 2 回答確率行列

本当＼見かけ(測定)	＋	±	−	計
＋	P_{++}	$P_{+\pm}$	P_{+-}	1
±	$P_{\pm+}$	$P_{\pm\pm}$	$P_{\pm-}$	1
−	P_{-+}	$P_{-\pm}$	P_{--}	1

表 3 回答確率が１のとき

本当＼見かけ(測定)	＋	±	−
＋	1	0	0
±	0	1	0
−	0	0	1

ある．同表にある0.7だの0.8だのという数字である．これがしっかりつかまえられていれば解決を得るわけで，こうした土台が無視されてはならない．

正しい結果の導きかた

（1） 誤差の表現

測定値の揺れ（誤差）は**表 2** のような確率の表で与えられるものとしよう．これは，本来（本当）のところ＋のものが，測定で正しい値＋を示す確率が P_{++}，誤った値±となって現れる確率を $P_{+\pm}$，－となって大いに誤ったものとして測定される確率を P_{+-} とする．これが測定誤差の表現である．$P_{+-}=1$，$P_{+\pm}=0$，$P_{+-}=0$，$P_{\pm+}=0$，$P_{\pm\pm}=1$，$P_{\pm-}=0$，$P_{-+}=0$，$P_{-\pm}=0$，$P_{--}=1$ であれば，つまり**表 3** の通りであれば，測定誤差がないことになる．

しかしここでも，測定誤差があるときこれを考慮に入れないと，結論で誤りを犯すことになる．これからは，誤差が確率的に独立に起こるものとして話を進めてみよう．

（2） 確率的な誤差の例と偏りのない推定値

簡単にするため，＋，－の二つとして数値例で説明しよう．本当は＋のものは，測定誤差があるため本当の値＋を示す確率は0.7，－と誤った値を示す確率が0.3としておく（**表 4**）．本当は－であるものは測定誤差により，－と正しい値を示す確率は0.8，＋と誤った測定を出す確率を0.2としておこう．いま，仮に本当は＋である人数は5500人，本当は－である人数は4500人（それぞれ55%，45%）であるとするが，これは未知なのである．測定結果をまとめてみると，平均的な意味で＋，－と現れる人数は前に示したが，繰り返して書くと，

表 4 誤差の出現（表１再出）

本当＼見かけ(測定)	＋	−
5500 ＋	0.7	0.3
4500 −	0.2	0.8
	4750	5250

$$5500\times0.7+4500\times0.2=4750\text{人}$$
$$5500\times0.3+4500\times0.8=5250\text{人}$$

となる．この4750人，5250人が観測された値となる．観測された値が4750人，5250人であるから，＋が47.5%，－が52.5%と結論しがちであるが，これは誤りである．本当は＋が55%，－が45%なのである．これはいわゆるサンプリングによる誤差ではない．測定結果が確率的に起こるための誤差なのである．このようなとき，測定誤差の確率がわかっていれば観則から正しい値を求めることができる．本当の＋の人数を n_+，－の人数を n_- とすれば，見かけの＋，－の人数，すなわち観測によって得られる人数は平均的に

$$\left.\begin{aligned}0.7n_+ + 0.2n_- &= 4750\\ 0.3n_+ + 0.8n_- &= 5250\end{aligned}\right\}$$

となるから，この連立方程式を解けば

$$n_+=5500,\quad n_-=4500$$

と正しい値を得ることができる．これは一般化して表現し，

$$\begin{pmatrix}0.7 & 0.3\\ 0.2 & 0.8\end{pmatrix}' \begin{pmatrix}n_+\\ n_-\end{pmatrix}=\begin{pmatrix}300\\ 700\end{pmatrix}$$

（ただし $\begin{pmatrix}\ \\ \ \end{pmatrix}'$ は転置行列 $\begin{pmatrix}0.7 & 0.2\\ 0.3 & 0.8\end{pmatrix}$ を表す）

となることから，

$$\begin{pmatrix}n_+\\ n_-\end{pmatrix}=\begin{pmatrix}0.7 & 0.3\\ 0.2 & 0.8\end{pmatrix}'^{-1} \begin{pmatrix}4750\\ 5250\end{pmatrix}$$

として求められる．

以上は，平均的な関係式を使って求めたのである．しかし，必ずしも平均値が出てくるわけではないので，そのときはどうするかとなる．このときは上の関係式を形式的に使って求めたものが，n_+，n_-の偏りのない推定値となることが証明でき，その分散も容易に計算できる．

関連分析の場合のミスリード

実際の例からみよう．前章の(1)で述べたP_{++}，…，P_{--}の表としてつぎのようなものを考えてみよう（**表5**）．

誤差のない場合，本当の＋，±，－の数は100，200，700としておこう．こうした測定誤差があるとき，2回の測定結果のクロス表は平均的に**表6**のようになる．こうなる理由を少し説明してみよう．

1回目の測定で＋を示すものは

$$100\times0.7+200\times0.2+700\times0.1$$
$$=70+40+70=180\text{人}$$

である．100人が180人と出るところからして誤っている．このうち2回目の測定で＋を示すものは平均的に，

表5 回答確率行列

本当＼見かけ	＋	±	－
＋	0.7	0.2	0.1
±	0.2	0.6	0.2
－	0.1	0.1	0.8

表6 データ

1回目＼2回目	＋	±	－	計
＋	64	45	71	180
±	45	83	82	210
－	71	82	457	610
計	180	210	610	1000

表7 正しい場合

1回目＼2回目	＋	±	－
＋	100	0	0
±	0	200	0
－	0	0	700

$$70\times0.7+40\times0.2+70\times0.1=64\text{人}$$

である．1回目＋で2回目±を示すものは，

$$70\times0.2+40\times0.6+70\times0.1=45\text{人}$$

となる．－を示すものは，

$$70\times0.1+40\times0.2+70\times0.8=71\text{人}$$

となる．こうして**表6**ができ上がる．

測定誤差が1回目，2回目とも同様であるならば，周辺分布は同一となっている．ここに注意すべきは，周辺分布が真の分布0.10，0.20，0.70と大きく異なっていることである．もし，測定誤差がなければ，2回測定をやったとき**表7**のようになるべきである．それが測定誤差のため**表6**のようにばらついた表が

得られる．ばらついているだけならばそれまでであるが，**表6**をもとにし，測定誤差がないものとして，実体的な結論を引き出そうとすれば大きな誤りを犯す．1回目に＋であったものが180人である．2回目の測定で＋にとどまるものは64人で約1/3にしかすぎない．そこで＋であったものは±あるいは－へ，特に－へ変化したものが多いと結論できる．ここに，測定誤差を無視して，いわゆる検定論を用いたところで結論は同様に誤りである．

ところがこれは，測定誤差のいたずらにすぎないのである．1回目±のところも2回目の測定で±にとどまるものは（＋，－）になるものにくらべ少なくなっているのである．

こうしたことも，測定誤差のことを心得ておけば，妥当性ある結論を導くことができるが，これを無視して分析していたとすれば，非常に誤りを犯すことになる．こうしたことを無視した統計学は空理空論である．