論　壇

誤差と残差と偏差と精度を中心に

鈴　木　栄　一*

「気象集誌」,「天気」に発表される統計学的研究論文を読んだり，レフェリーを依頼されたりした時の用語表現をみると，時に混乱が見られるので上記用語を中心に私見を述べておきたい．

一般に2つの確率変数 $\boldsymbol{X}$, $\boldsymbol{Y}$（ベクトル，スカラーのどちらでもよい），パラメータ（行列かベクトル）を $\boldsymbol{\beta}$ とするとき想定される母集団モデルを $F(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\beta}) = \varepsilon$ とかくとき，ε は誤差(error)といわれその期待値が0，有限な分散をもち，大きさ n の標本変量 $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ $(i=1, \cdots, n)$（確率変数）から $\boldsymbol{\beta}$ の推定量（estimator）$\tilde{\boldsymbol{\beta}}$（確率変数）をもとめ，上記モデルに代入した $F(\boldsymbol{X}_i, \boldsymbol{Y}_i, \tilde{\boldsymbol{\beta}}) = e_i$ $(i=1,\cdots, n)$ での e_i は残差（residual）といわれるのが普通である．

また，何らかの目的に合う基準または推定値 $\boldsymbol{a}$（たとえば期待値，平均値，中央値など）をもとに作られた確率変数 $(\boldsymbol{X}_i - \boldsymbol{a})$ $(i=1,\cdots, n)$ は偏差（deviation）といわれる．

この偏差の関数に対し，○○偏差といった名がつけられてきた．たとえば $\boldsymbol{a}$ をスカラー期待値 μ としてもとめられる $\sigma = \sqrt{\sum_{i=1}^{n} (X_i - \mu)^2 / n}$ は標準偏差で，特殊な順序標本から得られる4分偏差などがそれである．

ところで，気象関係では観測誤差（observation error）と観測精度（observation accuracy または precision）が，しばしば同じような意味の用語として論文にかかれたこともあったが，それでよいだろうか？　英文通りとすると，前者は観測のエラー（過誤），つまりまちがった観測と受けとられかねない．後者はたとえば真値が20/3のとき，小数1位までしか読み取れない計器の目盛から6.7となったとき，この実測値と真値との小さい差と受けとられ，エラーではない．したがって観測にエラーがない以上，観測精度とかかれる方が妥当である．

Eiichi Suzuki, 青山学院大学.

1982年4月

最近インドの A.K. Bansal (1980) が分散未知のことを unknown precision と書いているが，unknown variance の方が分り易い．気象関係でよく用いられる重回帰では

母集団：$Y = \alpha + \boldsymbol{\beta}' \boldsymbol{X} + \varepsilon$

標　本：$y_i = a + \boldsymbol{b}' \boldsymbol{X}_i + e_i \quad i=1,\cdots, n$

となり，ε が誤差，e_i が残差とみなされるのがよい．エラー ε を導入する理由は現時点での知識では Y と $\boldsymbol{X}$ の関係が線形か非線形か分らないし，$\boldsymbol{X}$ の中に入れるべき要因に見落しがあるかも知れず，真の完全な関係(未知)からみてエラーがあると考えるからで，e_i を残差とする理由は，母集団から大きさ n の標本をとる仕方が数多くあり（母集団が有限個で大きさ N なら，標本のとり方は ${}_N C_n$ 通り，母集団が無限個なら，標本のとり方は無限通りある！），その「とり方」に依存し，ε とはちがう性格をもつので区別したいからである．実際，最小2乗法で a, $\boldsymbol{b}$ の推定量（確率変数）をもとめるとき，残差2乗和 (Residual Sum of squares)の最小化といい，誤差2乗和最小化とはあまりいわない．

この他，独立 (independency)，無関係 (no relation)，無相関 (no correlation) もよく混同されて用いられる．

独立とは2つの確率変数の同時確率分布が，それぞれの確率分布の積となる場合で，独立な標本変量とは正にこの場合である．無関係といったときは変数が確率変数でなくてもよいであろうが，無相関というときは，確率変数でなくてはならない．というのは相関という考え方が統計学であらわれたとき確率変数間で定義されたもので，たとえば，時間 t とともに物理量 X が増加または減少しているとき，t と X に相関があるとするのは正しくない．つまり t は確率変数でなく，指定変数 (fixed variate) とされているからである．

また人によっては「$P(X=1)=1$ つまり $X=1$ なので」という方がいる．これは正しくない．$P(X=1)$ とは正確にかくと $P[\{X(\omega) \mid X(\omega)=1\}]$ で事象の元 ω に割当てられた実確率変数 $X(\omega)$ が1となるような事

象（集合）の確率だが，単に $X=1$ とするときは X が確率変数である必要がない．また確率変数 x の確率密度関数 $f(x)$ とかくのも正確でない．確率変数 $X(\omega)$ の確率分布 $P[\{X(\omega)|\ X(\omega)\leqq x\}]\equiv F(x)$ が連続で微分可能のとき確率密度関数 $f(x)=dF(x)/dx$ が存在し，一般に $X(\omega)$ がいろいろな値をとり得るので，一般化するため実変数 x を用いる．この小文字 x は確率変数というより実変数と考えられるべきものである．この他，事象（集合）の非負測度（non-negative measure）である確率（Probability）と単なる可能性（Possibility）ともよく混用されるし，標本抽出誤差と標本誤差（この定義と意味内容は人により不明確である）との区別もはっきりかかれていない．

偶然誤差（random error），系統的誤差（systematic error），偏り（bias），上下の偏差（anomaly），…といった各種用語には，その専門書なり研究文献でも多少混乱が散見されるので，実際応用家がその被害（？）をうけて誤解した表現をするのも無理ないであろう．そこで，私の提案は，

(1) 一応，身近にいる数理統計学専門家にきいてみる．
(2) 用語を最初に用いるとき，その各用語について自分なりの定義，意味，内容を明確に示す（村上多喜雄の集誌論文はこの良い例）．
(3) 同じ意味内容を示すのに2つ以上の相異なる表現をせず，ちがう意味内容なのに同一の用語を用いない（この例をあげるのは遠慮したい）．

の3つで，とくに (1) が実際に不可能なら，(2)，(3) に留意すればよく，それほど神経質にならなくてもよかろう．たとえば (3) についていえば「推論」，「推測」，「推定」，「推察」，「予測」，「予知」，「予報」，「予察」といった表現をあまり適当に（悪くいえば便宜的に）統計処理過程やその結果に用いられると，これを読む人毎にちがった受け取り方をされる懸念がある．とくに推測，この対象は母集団確率分布モデルか，あるいはその中のパラメータ）と予測（この対象は一応定義された確率変数）の混用は避けた方が賢明である．

要は，論文を読まれる人々の間で「異なった理解や受け取り方をされない」用語表現を上記 (2) によって実行されれば十分であろう．つまり定義や概念説明をしない用語をつかわなければよいだけである．

〔付記〕

気象関係での「予報」，「予知」を水産関係では「予察」といわれる由．これも上記 (1) の実行で混乱はさけられよう．

高校数学Ⅰにも，最近確率が登場した．そこで，根元事象，事象，全事象（標本空間）と確率が説明されているが，多少の混乱があるようだ．混乱をさけるため，次のようにしたらよいと私は考えている．

1つの観測，試行の結果を「事象の元（要素）」という（これを根元事象というと，「事象（集合）の一種」と間違えられる懸念があり，標本点とすると，母集団からの標本を示す点と誤解され易い．ここではまだ標本の概念はいらない．）．空事象を含めて，「事象の元」の「集合」を事象という（集合としての事象に対して非負測度の確率を定義するため，事象概念を明確にする訳である．）．

可能なすべての「事象の元」の集合を全事象 Ω とする（これを標本空間とはなるべくいわない．ここでもまだ母集団に対する標本は登場してこないから．）．

つまり「事象の元」$\omega\rightarrow$事象 $E\rightarrow$事象族 F という集合の論理に合う形にし，確率変数 $X(\omega)$，事象 E の確率 $P(E)$ という定義をもとに混乱をおこさない表現が望ましいだろう．