統計十話

第5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2

林 知己夫
文部省統計数理研究所

第5話で，測定データ変動を無視するこわさを実例をもって示した．これをどう処理して正しい結論を得るかをここで示してみよう．

誤差の出方の説明

なお，これまで，あるいはこれからも変動といったり誤差といったりしているが，分析のうえでは同じものとみなしていただきたい．誤差とは，われわれの情報を覆い隠す雑音という意味であるから，生体変動による変動であっても，平均値としての情報を覆い隠す誤差とみなすわけである．いずれにしても真の値（あるいは平均値）のまわりの変動を誤差という表現で述べてみる．したがって，測定者の読みの誤差もあるし，測定条件や生体の変動によるものも含まれているわけである．同じ時期に2回以上測定して標識づけたとき同一の値が示されない．すなわちデータに変動があるとき，これを“誤差があるため”と表現して取り扱う．誤差をこのように値の変動を含めた広い意味に用いることにする．

（1） 誤差の出方

前述のように全体の集団の分散は441程度であるが，これには測定誤差が含まれている．測定誤差は，変動の分散を前に述べた議論からよくコントロールされた条件下で一応50とみなしておき，かつ見通しをよくするため各人の変動の分散も一応同一とみなしておけば，測定誤差のないときの分散 $\equiv_o^2$ は，

$$\equiv_o^2 = 391$$

となる．前に述べた誤差分散と $\equiv_o^2$ との比は，0.13程度でかなり大きい．さて，こうした変動は一応ガウス分布とみなしておく．つまり平均0，分散 s^2 のガウス分布とみなしておく．$s^2=50$ としておくわけである．この確率論的仮定も，データからみて，妥当な仮定と考えられる．さて，こうなると2回測定を行えば図のように，測定の誤差のないとき (x_0, y_0)，ただし $x_0=y_0$ であるものが測定されたとき，それらの値はある確率をもって円形の範囲にばらつくことが考えられる．

このような (x, y) が観測されるはずである．このような人々が多くあれば，$x_0=y_0$ であっても45°の上にのらずに散らばってくることになる．T 年および $T+1$ 年において，測定誤差がなければ，等しい値をもつ，すなわち上の表現に従えば，各人について $x_0=y_0$ であるとしておこう．これは，前に示した T 年，$T+1$ 年の平均，分散，分布の形の恒常性から一応認められよう．

（2） 測定誤差のない場合の測定値

さて，測定誤差のないときの血圧値の分布も簡単のため，ガウス分布としておく．これは，さきのデータからみるとやや無理な仮定であるようにみえる——実際は分布が少し流れ，ピアソンIII型の分布である——が，計算が容易で，見通しをよくするためひとまずこうしておく．ピアソンIII型にしても以下の計

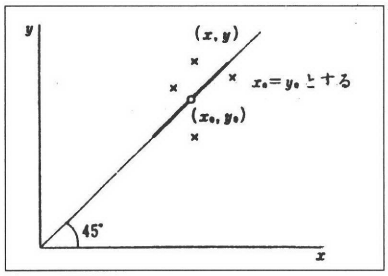

図1　誤差の出方

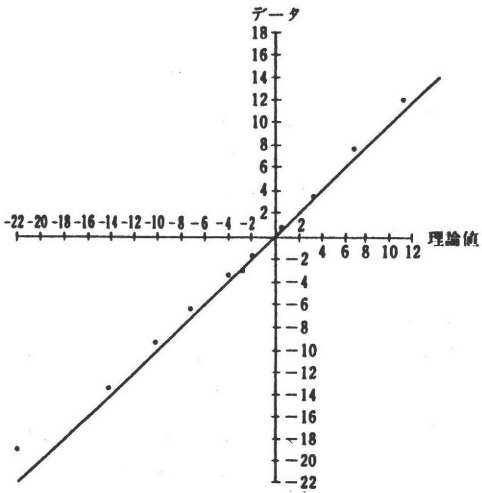

図2　理論値とデータ

算は可能であるが，見通しのよい形で書けないだけである．これを

$$\frac{1}{\sqrt{2\fallingdotseq}\equiv}e^{-(x-M)^2/2\sigma^2}dx=(x|M,\equiv^2)dx$$

としておく．M は平均値，$\equiv^2$ は分散である．これまでの表現に従えば，本来 $x, \boldsymbol{\sigma}^2$ でなく $x_0, \boldsymbol{\sigma}_o^2$ を用いるべきであるが，煩雑を避けるため添字をおとして x としておく．T 年の測定誤差のない血圧値の分布も，$T+1$ 年のそれも同一であるとしておく．したがって，T年の M も $\equiv^2$ も $T+1$ の M も $\boldsymbol{\sigma}^2$ も同一であり，かつ T 年であるxを示すものは，$T+1$ 年でもその x をもつとしておく．さて，測定誤差 の分布は平均0，分散 s^2 のガウス分布は，

$$\frac{1}{\sqrt{2\fallingdotseq}s}e^{-\varepsilon^2/2s^2}d$$

であるとしておくので，測定誤差のないとき血圧値 x を示すものの測定値 z の分布は，

$$\frac{1}{\sqrt{2\fallingdotseq}s}e^{-(z-x)^2/2s^2}dz$$

となる．$s^2=g(x)$ と，x の関数であることも予想されるが，いまは，一応 s^2 は常数としておく．$s^2=g(x)$ でも計算はめんどうになるが可能である．計算は，面倒なので省略するが(詳しくは，林：行動計量学序説．朝倉書店，1993，p79～81)，こうして一応の説明は可能になったが，まだ不十分であった．これは，s^2 が一定であるとして計算したからである．また，もとの血圧分布がガウス分布をしていないこともその原因でもある．

より現実に近づける……

血圧値の大きいところに理論とデータに誤差が高いので，さらに精密な手続きを用いて検討を重ねてみた．つまり，血圧値の分布がガウス分布でないとして，データから求める．血圧値の $s^2=g(x)$ についてデータから適当とみなせる形を想定する，ということを行い，誤差を含む現実の血圧分布に近づけた．

こうして計算した結果とデータを目盛ってみると図2のようにきわめてよい一致を得た．

以上の議論からみて，測定誤差のいたずらによって思わぬ歪んだ様相があらわれてきたことが示された．つまり，測定誤差によってわれわれの得た“データの相関関係”の様相がよく説明されたことがわかった．T 年も $T+1$ 年の測定も，もし測定誤差がなければまったく相等しい値を示すものとしても，T 年からみた $T+1$ 年の平均値は45°の線上にのらないで，偏った直線上にのることになるわけである．つまり，回帰直線は45°とならないわけである．これが45°にのったとしたら，むしろ T 年と $T+1$ 年とはまったく相等しい値を示すものではない——測定誤差がないとき——といえることになる．