統計的予測の形式と方法

1.

統計的予測の主な形式について述べよう．

$X_1, X_2, \cdots, X_n$をデータとし，Yを予測すべき量とする．$X_1, \cdots, X_n$, Yは未知母数θをふくむある分布に従うものとする．

いまYが実数であるとするとき，Yの値そのものを直接予測しようとする方式は，点予測 point prediction とよばれる．$X_1, \cdots, X_n$から計算される予測量

$$\hat{Y}=\hat{Y}(X_1, \cdots, X_n)$$

は，すべてのθについて，予測誤差の期待値が0になる，すなわち，

$$E_\theta(Y-\hat{Y})=0 \quad \forall\theta$$

であるとき，不偏予測量 unbiased predictor であるとよばれる．不偏予測量の中で誤差分散，

$$V_\theta(Y-\hat{Y})=E_\theta(Y-\hat{Y})^2$$

を最小にするものがあれば，それは(一様)最小分散不偏予測量とよばれる．

多くの具体的な例では$X_1, \cdots, X_n$とYが(確率的に)独立になる．このときには，

$$E_\theta(Y)=g(\theta)$$

とおけば，$\hat{Y}$が不偏予測量であるとき，

$$E_\theta(\hat{Y})=g(\theta)$$

$$E_\theta(Y-\hat{Y})^2=E_\theta(Y-g(\theta))^2+E_\theta(\hat{Y}-g(\theta))^2$$
$$=V_\theta(Y)+V_\theta(\hat{Y})$$

となる．$V_\theta(Y)$は$\hat{Y}$の定め方と独立であるから，$\hat{Y}$は$g(\theta)$の不偏推定量になり，かつそれが最小分散不偏予測量になることは，それが$g(\theta)$の最小分散不偏推定量になることを意味する．したがって不偏予測論は不偏推定論に帰着する．

$X_1, \cdots, X_n$とYが互いに独立でないときには点予測の問題はやや複雑になる．その場合には，$X_1, \cdots, X_n$が与えられたときの条件付期待値を，

$$E_\theta(Y|X)=g(\theta, X_1, \cdots, X_n)$$

と表わせば，不偏性の条件は，

$$E_\theta(\hat{Y}-g(\theta, X_1, \cdots, X_n))=0$$

また分散最小の条件は，

$$V_\theta(Y-\hat{Y})=V_\theta(Y|X)$$
$$+E_\theta(\hat{Y}-g(\theta, X_1, \cdots, X_n))^2$$

であるから，

$$E_\theta(\hat{Y}-g(\theta, X_1, \cdots, X_n))^2：最小$$

となる．

この問題は一般には簡単な解をもたないが，

$$g(\theta, X_1, \cdots, X_n)=g(\theta)+h(X_1, \cdots, X_n)$$

という形に分解される場合には，$g(\theta)$の最小分散不偏推定量$\hat{g}(\theta)$を求めて，

$$\hat{Y}=\hat{g}(\theta)+h(X_1, \cdots, X_n)$$

とすればよい．

［例1］$X_1, \cdots, X_n, Y$が多変量正規分布に従い，その平均，分散がすべてμ，およびσ^2，$X_1, \cdots, X_n$は互いに独立，X_iとYとの相関がρ_iであるとき（ρ_iは既知とする），

$$E(Y|X)=\mu+\rho_1(X_1-\mu)+\cdots+\rho_n(X_n-\mu)$$
$$=(1-\rho_1-\cdots-\rho_n)\mu$$
$$+\rho_1X_1+\cdots+\rho_nX_n$$

であるから，

$$\hat{Y}=(1-\rho_1-\cdots-\rho_n)\bar{X}+\rho_1X_1+\cdots\cdots+\rho_nX_n$$
$$=\bar{X}+\rho_1(X_1-\bar{X})+\cdots\cdots+\rho_n(X_n-\bar{X})$$

とすればよい．

［例2］ もう一つの例として．$Z_1, \cdots, Z_N$ が互いに独立につぎのような密度関数をもつ指数分布に従うとし，

$$f(z)=\frac{1}{\theta}e^{-z/\theta} \qquad z>0$$

それを大きさの順に並べたものを $Z_{(1)}<\cdots<Z_{(N)}$ とし，$X_1=Z_{(1)}, \cdots, X_n=Z_{(n)}$, $Y=Z_{(N)}$ $(N>n)$とする．すなわち $Z_1, \cdots, Z_N$ のうち小さいほうから n 個を観測して，その中の最大のものを予測する問題である．$Z_1, \cdots, Z_N$ が寿命を表わすものとすれば，寿命試験の観測を途中で打ち切って，最大値を予測する問題と考えられる．このとき，

$$Y_i=(N-i+1)(Z_{(i)}-Z_{(i-1)}) \quad i=1, \cdots, N$$

ただし $Z_{(0)}=0$ とする．

とおけば $Y_1, \cdots, Y_N$ は互いに独立に同じ指数分布に従う．そうして $X_1, \cdots, X_n$ を与えることは $Y_1, \cdots, Y_n$ を与えることと同じであるから，

$$\begin{aligned}Z_{(N)}&=Z_{(N)}-Z_{(N-1)}+(Z_{(N-1)}-Z_{(N-2)})+\cdots\\&=Y_N+\frac{1}{2}Y_{N-1}+\cdots+\frac{1}{N}Y_1\end{aligned}$$

となることを用いれば，

$$\begin{aligned}E(Z_{(N)}|X)&=\left(1+\frac{1}{2}+\cdots\cdots+\frac{1}{N-n}\right)\theta\\&\quad+\frac{1}{N-n+1}Y_n+\cdots\cdots+\frac{1}{N}Y_1\\&=\left(1+\frac{1}{2}+\cdots\cdots+\frac{1}{N-n}\right)\theta+X_n\end{aligned}$$

また θ の最小分散不偏推定量は，

$$\begin{aligned}\hat{\theta}&=\frac{1}{n}(Y_1+\cdots+Y_n)\\&=\frac{1}{n}((N-n+1)X_n+X_1+\cdots+X_{n-1})\\&=\frac{1}{n}\{(N-n)X_n+n\bar{X}\}\end{aligned}$$

となるから，結局 $Z_{(N)}$ の最小分散不偏予測量は，

$$\hat{Z}_{(N)}=\left(1+\frac{1}{2}+\cdots\cdots+\frac{1}{N-n}\right)\hat{\theta}+X_n$$

となる．

2.

つぎに区間予測について考えよう．予測される量 Y に対して $X_1, \cdots, X_n$ から2つの量

$$\underline{Y}=\underline{Y}(X_1, \cdots, X_n) \qquad \bar{Y}=\bar{Y}(X_1, \cdots, X_n)$$

を計算して，Y がほぼ $\underline{Y}$ と $\bar{Y}$ の間に入るであろうという形で予測するのが区間予測である．ここで，

$$P_\theta\{\underline{Y}<Y<\bar{Y}\}\geqq 1-\alpha \qquad \forall\theta$$

となるとき，$[\underline{Y}, \bar{Y}]$ を信頼係数 $1-\alpha$ の予測区間 prediction interval という．

予測区間を求める直観的な方法は，分布が θ をふくまないような適当な統計量

$$T=t(X_1, \cdots, X_n, Y)$$

を計算して，

$$P_T\{\underline{t}<T<\bar{t}\}=1-\alpha$$

となるような $\underline{t}$, $\bar{t}$ を求め，つぎに $\underline{t}<T<\bar{t}$ を Y に関して解くことである．このときもし Y についての区間 $[\underline{Y}<Y<\bar{Y}]$ が得られるならば，

$$P_\theta\{\underline{Y}<Y<\bar{Y}\}=1-\alpha \qquad \forall\theta$$

となる．このような区間は相似 similar であるといわれる．

［例3］ $X_1, \cdots, X_n, Y$ が互いに独立に正規分布 $N(\mu, \sigma^2)$ に従うとき，

$$T=\sqrt{\frac{n}{n+1}}\left(\frac{Y-\bar{X}}{S}\right) \quad S^2=\frac{1}{n-1}\sum(X_1-\bar{X})^2$$

とおけば，T は自由度 $n-2$ の t 分布に従うからその両側 α 点を t_α と表わせば，

$$P\{|T|<t_\alpha\}=1-\alpha$$

より，

$$\begin{aligned}&P\left\{\bar{X}-t_\alpha S\sqrt{\frac{n+1}{n}}<Y<\bar{X}+t_\alpha S\sqrt{\frac{n+1}{n}}\right\}\\&\quad=1-\alpha\end{aligned}$$

となるから，{ }内が予測区間を与える．

もう少し複雑な場合として上記の例2の場合を取り上げよう．このとき，

$$Z_{(N)}-X_n=Y_N+\frac{1}{2}Y_{N-1}+\cdots+\frac{1}{N-n}Y_{n+1}$$

は $X_1, \cdots, X_n$ と独立で，かつ，

$$T=\frac{Z_{(N)}-X_n}{\sum_{i=1}^{n}Y_i/n}=\frac{Z_{(N)}-X_n}{\hat{\theta}}$$

の分布は θ をふくまない．したがって T の分布を(簡単な形にはならないが)，計算して，

$$P\{\underline{T}<T<\bar{T}\}=1-\alpha$$

となるような $\underline{T}$, $\bar{T}$ を求めれば，

$$X_n+\underline{T}\hat{\theta}<Z_{(N)}<X_n+\bar{T}\hat{\theta}$$

より $Z_{(N)}$ の予測区間を求めることができる．

もっと複雑な問題についてはつぎのように考えればよい．予測区間 $\underline{Y}(X_1,\cdots,X_n)<Y<\bar{Y}(X_1,\cdots,X_n)$ に対応して，区間予測関数 ϕ を，

$$\phi(X_1,\cdots,X_n,\ Y)=1 \qquad \underline{Y}<Y<\bar{Y} \text{ のとき}$$
$$=0 \qquad Y\leqq\bar{Y} \quad \text{または} \quad Y\geqq\bar{Y}$$

と定義すれば，

$$E_\theta(\phi)\geqq 1-\alpha \qquad \forall\theta$$

となる．このことは $1-\phi$ が $X_1,\cdots,X_n,Y$ が仮定された同時分布に従うという仮説を検定する問題に対する，水準 α の検定関数になることを表わしている．したがってこのことから ϕ を求めることができ，それから逆に Y の予測区間を求めることができる．

このような考え方からノンパラメトリック予測区間が求められる．

いま $X_1,\cdots,X_n,Y$ が互いに独立に，同じ連続分布に従うとする．いま $X_1,\cdots,X_n,Y$ を一緒に考えて，その $n+1$ 個の値の集合を，

$$O_{n+1}=\{Z_1,\ Z_2,\ \cdots,\ Z_{n+1}\}$$

と表わすと，O_{n+1} が与えられたとき，

$$P\{Y=Z_i|O_{n+1}\}=1/(n+1) \quad i=1,\ 2,\ \cdots,n+1$$

となる．そこでいま f を任意の関数とするとき，$f(z_1),\cdots,f(z_{n+1})$ の中での $f(Y)$ の順位を R とすると，

$$P\{R=i\}=1/(n+1) \qquad i=1,\ 2,\ \cdots,\ n+1$$

となる．それゆえ ϕ を，

$$\phi=1 \qquad R\leqq j \qquad \text{のとき}$$

とすれば $E(\phi)=j/(n+1)$ となるから，j を $1-\alpha=j/(n+1)$ となるように定めれば，ϕ から信頼係数 $1-\alpha$ の予測区間が得られる．すなわち，$f(X_i)$ $i=1,\ \cdots,\ n$ の中での i 番目の値を $f(X)_{(i)}$ と表わすことにすれば，$R\leqq j$ は，

$$f(Y)<f(X)_{(j)}$$

と同値になる．そこで上から定めうる Y の範囲が区間になるならば，予測区間が得られることになる．たとえば，$f(X)=X^2$ とすれば，予測区間は，

$$-|X|_{(j)}<Y<|X|_{(j)}$$

となる．

上記の議論においてさらに f は O_{n+1} に依存してもよいことがわかる．すなわち W を $X_1,\ \cdots,\ X_n$ および Y の対称関数として $f(X_i,W)$ の中で j 番目の値を $f(X,W)_{(j)}$ と表わすとき，

$$P\{f(Y,W)<f(X,W)_{(j)}\}=j/(n+1)$$

となるから，右辺が $1-\alpha$ に等しければ，

$$f(Y,W)<f(X,W)_{(j)}$$

を Y について解いたものが，もし区間になれば信頼係数 $1-\alpha$ の予測区間が得られることになる．

[例4] いま $W=(\sum X_i+Y)/(n+1)=(n\bar{X}+Y)/(n+1)$ とし，$f(Y,W)=|Y-W|$ とする．そうすると上記の条件は，不等式

$$\frac{n}{n+1}|Y-\bar{X}|<\frac{1}{n+1}|n(X_i-\bar{X})+(X_i-Y)|$$

か，少なくとも $n-j+1$ 個の X_i について成り立つことを意味する．また上記の不等式が成り立つことは，

$$X_i\gtrless Y\gtrless\bar{X}+\frac{n+1}{n-1}(\bar{X}-X_i)$$

となることに等しいから，予測区間はこのような区間のうち少なくとも $n-j+1$ 個にふくまれるような部分として与えられる．

3.

予測区間とよく似た概念として予測限界 prediction limit がある．それは Y の値について，上側あるいは下側の限界 $\bar{Y}$，あるいは $\underline{Y}$ を $X_1,\cdots,X_n$ から計算して，

$$P_\theta\{Y>\bar{Y}\}\leqq\alpha \quad \text{あるいは} \quad P_\theta\{Y<\underline{Y}\}\leqq\alpha \qquad \forall\theta$$

となるようにするものである．このとき $\bar{Y}$，および $\underline{Y}$ をそれぞれ信頼係数 $1-\alpha$ の上側予測限界あるいは下限予測限界という．このような(片側)予

測限界を求める方法は，予測区間を求める方法とほとんど同じである．

［例5］ 1月の事故件数は，λを母数とするポアソン分布に従うことがわかっているとする．このとき過去のデータ $X_1, \cdots, X_n$ にもとづいて，ある日の事故件数 Y の上側予測限界を求める．$X_1, \cdots, X_n, Y$, の同時分布は，

$$P\{X_1=x_1, \cdots, X_n=x_n,\ Y=y\} = \frac{\lambda^{x_1+\cdots+x_n+y}}{x_1!\cdots\cdots x_n!\, y!} e^{-\lambda}$$

と表わされるから，$T=\sum X_i+Y$ が与えられたときの Y の条件付分布は，

$$P\{Y=y \mid T=t\} = \frac{t!}{y!(t-y)!}\left(\frac{1}{n+1}\right)^{y}\left(\frac{n}{n+1}\right)^{t-y}$$

という形の2項分布になることがわかる．

そこで与えられた t に対して，

$$\sum_{y'\leqq y} P\{Y=y' \mid T=t\} \leqq \alpha < \alpha - P\{Y=y+1 \mid T=t\}$$

となるような y の値を $y_\alpha(t)$ と表わせば $y_\alpha(t)$ は t の増加関数になる．そうして，

$$P\{Y\leqq y_\alpha(T)\} \leqq \alpha$$

となる．そこで与えられた X_i に対して，

$$Y\leqq y_\alpha(\sum X_i+Y)$$

をみたすような Y の値の最大値を $\bar{Y}$ とすれば，それが Y の上側予測限界を与える．

4.

予測されるべき値がベクトル値，すなわち2つ以上の実数値 $Y_1, \cdots, Y_k$ である場合，その実数値関数 $Y=g(Y_1, \cdots, Y_k)$ の点予測，あるいは区間予測については，これまでの議論をそのままあてはめることができる．また $Y_1, \cdots, Y_k$ の同時予測についても，点予測の場合にはそれぞれの成分についての不偏予測量を考えればよいから，あまり問題はない．これに対して同時区間予測については，やや新しい問題が生ずる．すなわち今度は一般に $X_1, \cdots, X_n$ に対応して k 次元ユークリッド空間内の集合 C を対応させ，

$$P_\theta\{(Y_1, \cdots, Y_k) \subset C\} \geqq 1-\alpha \qquad \forall\theta$$

となるようにする．このような C を信頼係数 $1-\alpha$ の予測域 prediction region という．C は連続な集合であることが要求され，また凸集合であることが一般には望ましいと考えられよう．

［例6］ $X_1, \cdots, X_n, Y_1, \cdots, Y_k$ が互いに独立に正規分布 $N(\mu, \sigma^2)$ に従うとする．このとき $Y_1-\bar{X}, \cdots, Y_k-\bar{X}$ はすべて平均0，分散 $(1+1/n)\sigma^2$ の正規分布に従い，かつ，これらの値の共分散は σ^2/n となるから，分散行列の逆行列を求めることにより，

$$\frac{1}{\sigma^2}\left\{\sum(Y_i-\bar{X})^2 - \frac{k^2}{n+k}(\bar{Y}-\bar{X})^2\right\}$$

が χ^2 分布に従うことがわかる．σ^2 をその推定量 S^2 でおきかえれば，F 分布に従う統計量が得られるから，

$$P\left\{\frac{1}{S^2}\left[\sum(Y_i-\bar{X})^2 - \frac{k^2}{n+k}(\bar{Y}-\bar{X})^2\right] < F_\alpha\right\} = 1-\alpha$$

となる．ただし F_α は自由度 $(k,\ n-1)$ の F 分布の上側の点である．ゆえに同時予測域は，

$$\sum(Y_i-\bar{X})^2 - \frac{k^2}{n+k}(\bar{Y}-\bar{X})^2 < F_\alpha S^2$$

となる．これは $(\bar{X}, \cdots, \bar{X})$ を中心とする楕円体となっている．もっと複雑な問題については，区間予測の場合と同じく地域予測関数 ϕ を，

$(Y_1, \cdots, Y_k) \in C$ のとき

$$\phi(X_1, \cdots, X_n,\ Y_1, \cdots, Y_k) = 1$$

$(Y_1, \cdots, Y_k) \notin C$ のとき

$$\phi(X_1, \cdots, X_n,\ Y_1, \cdots, Y_k) = 0$$

と定義すれば，

$$E_\theta(\phi) \geqq 1-\alpha \qquad \forall\theta$$

となるような ϕ を求めればよいことになる．

ここに述べたような問題についてのその他の多くの例題については，前の章でも述べた私著「統計的予測論」および石井吾郎氏がＢＡＳＩＣ数学に連載中の「統計的予測論」を参照していただきたい．