分類について II

[7] の結果の改良と適用例

藤本熙

(1964年12月受付)

On the Classification II

Hirosi Hudimoto

We shall improve on a result of [7] and illustlate a use of the mann-whitney statistic to classification. The case that ties are present is treated.

Institute of Statistical Mathematics

§1. [7] の結果の改良について

$(X_1, X_2, \cdots, X_n)$ が二つの母集団 Π_1 あるいは Π_2 のいずれかからの大きさ n のランダム標本であるとき，これをその集団のいずれかへ判別する問題は，Π_1 および Π_2 の確率分布が完全に指定されれば，尤度比を使う方法がよく知られる（例えば，T. W. Anderson [1]）．だが，確率分布の型が指定されず，ただ Π_1 および Π_2 からのランダム標本 $(X_1^{(1)}, \cdots, X_{n_1}^{(1)})$, $(X_1^{(2)}, \cdots, X_{n_2}^{(2)})$ のみが利用可能であるときには，この尤度比による方法は適用できない．その故に [7] でこのための予備的な考察を，次のように試みた．

Π_1 および Π_2 の分布関数 $F_1(x^{(1)})$, $F_2(x^{(2)})$（連続を仮定）が既知であるとき，次の2つの統計量，

$$\hat{p}_1=\frac{1}{n}\sum_{k=1}^{n}F_1(X_k) \quad \text{および} \quad \hat{p}_2=1-\frac{1}{n}\sum_{k=1}^{n}F_2(X_k) \tag{1.1}$$

を考え，

$$p=\int_{-\infty}^{\infty}F_1(t)dF_2(t) \tag{1.2}$$

が 1/2 より大であれば，$\hat{p}_1<\hat{p}_2$ ならば，$(X_1, \cdots, X_n)$ を Π_1 へ，$\hat{p}_1>\hat{p}_2$ ならば Π_2 へ属すと判別した．

その理由は，$u=1,2$ および $v=1,2$ に対して，

$$\mathrm{E}\{\hat{p}_u|(X_1, \cdots, X_n)\in\Pi_v\}=\begin{cases}1/2 & (u=v),\\ p & (u\neq v),\end{cases} \tag{1.3}$$

ただし $\mathrm{E}\{\ \ |c\}$ は条件 c のもとでの期待値を示す――であり，かつ $p>1/2$ が仮定されるからである．従って $p<1/2$ のときは，前述の判別方式は，不等号の向きを変えるだけでよいことはいうまでもない．また，このときこの判別方式による判別成功の確率は $1-e^{-n\Delta^2/2}$ より大である．ただし，$\Delta=p-(1/2)$．これは [7] での結果の一つの改良である（証明は [8] 参照）．

更にこの方式の non-parametric な場合，すなわち，$F_1(x^{(1)})$, $F_2(x^{(2)})$ が未知で，それぞれそれからの標本 $(X_1^{(1)}, \cdots, X_{n_1}^{(1)})$, $(X_1^{(2)}, \cdots, X_{n_2}^{(2)})$ のみが利用可能な場合への拡張は，

(1.4) $$\hat{p}_1=\frac{1}{n}\sum_{k=1}^{n}\hat{F}_1(X_k) \quad \text{および} \quad \hat{p}_2=1-\frac{1}{n}\sum_{k=1}^{n}\hat{F}_2(X_k)$$

を $\hat{p}_1$ および $\hat{p}_2$ の代りに使うことである．ただし，$\hat{F}_1, \hat{F}_2$ はそれぞれ F_1 および F_2 の経験分布関数である．

ここでもし $p>(1/2)$ を仮定する根拠が与えられるならば，判別方式は再び $\hat{p}_1<\hat{p}_2$ ならば，$(X_1, \cdots, X_n)$ を Π_1 へ，$\hat{p}_1>\hat{p}_2$ ならば Π_2 へ判別することになり，そのときの判別成功の確率はやはり $1-e^{-n\Delta^2/2}$ より大となる．

勿論ここで $p\geqq 1/2$ は，実際に標本 $(X_1^{(1)}, \cdots, X_{n_1}^{(1)})$ および $(X_1^{(2)}, \cdots, X_{n_2}^{(2)})$ から確かめられるのが望ましいわけだが，

(1.5) $$U_{n_1n_2}=\{X_i^{(1)}<X_j^{(2)}\text{なる組 } (X_i^{(1)}X_j^{(2)}) \text{ の数}\}$$

とし，

(1.6) $$\hat{\hat{p}}=\frac{U_{n_1n_2}}{n_1n_2}$$

を考えると，$\hat{\hat{p}}$ は p の一様最小分散不偏推定量であることが知られる（例えば E. L. Lehman [10] 参照）．また，その極限分布は正規分布になる（E. L. Lehman [11] 参照）．そこで n_1, n_2 が同程度の大きさで $(n_1/(n_1+n_2)\underset{n_1+n_2\to\infty}{\longrightarrow}\text{常数}, (\neq 0))$，その大きさがかなり大であれば，$\hat{\hat{p}}$ の評価に正規分布による近似が可能である．ただし，有意性検定のような手段で，$F_1\neq F_2$ を確かめるだけなら，$F_1\equiv F_2$ のときには期待値 $\mathrm{E}(\hat{\hat{p}})=1/2$，分散 $\mathrm{Var}(\hat{\hat{p}})=(n_1+n_2+1)/12n_1n_2$ であるから，取扱いは容易であるが，この場合には $\Delta=p-(1/2)$ の大きさも問題であるので，$F_1\neq F_2$ のときの分散が必要になる．だが，これには $\int F_u^2(t)dF_v(t)$ $(u\neq v,\ u, v=1, 2)$ の形の積分が入るから，一般の分布についてその値は求まらないが，$\mathrm{Var}(\hat{\hat{p}})\leqq p(1-p)/\min\{n_1, n_2\}$ が知られるから，それを利用する手段がある．

だが，なお一層好都合な方法は，Z. W. Birnbaum と R. C. McCarty [3] の p についての，distribution free upper confidence bound，すなわち，$P(p<\hat{\hat{p}}+\varepsilon)\geqq\gamma$ $(P(\hat{\hat{p}}-\varepsilon<p)$ も同様であるから——）を利用することであろう．ここに $\varepsilon=\delta/\sqrt{n_1+n_2}$ で，δ は，

(1.7) $$\gamma=1-\lambda e^{-2(1-\lambda)\delta^2}-(1-\lambda)e^{-2\lambda^2}$$
$$-2\sqrt{2\pi}\lambda(1-\lambda)\delta e^{-2\lambda(1-\lambda)\delta^2}\frac{1}{\sqrt{2\pi}}\left[\int_{-2(1-\lambda)\delta}^{2\lambda\delta}e^{-t^2/2}dt\right]$$

から定まる．ただし，$\lambda=n_2/(n_1+n_2)$．更にこれは D. R. Owen, K. J. Craswell と D. L. Hanson [12] によって数表にされているから，便利である．だが一層簡単な評価は W. Hoeffding [9] の結果を使うと，次の評価が得られる．

(1.8) $$P(p<\hat{\hat{p}}+\varepsilon)=P(\hat{\hat{p}}-\varepsilon<p)\leqq 1-e^{-2n\varepsilon^2}$$

ただし，$n=\min\{n_1, n_2\}$．またこれは Z. W. Birnbaum [2] の結果よりはよくなっている．

§2. p の適用例と tie について

§1 での議論は適当な一次元化の基準が与えられれば，多変量の場合にも適用可能である．

例えば，期待値 $\boldsymbol{\mu}^{(u)}(u=1, 2)$，共分散行列 $\boldsymbol{\Sigma}$ の p 変量正規分布に従う独立な確率変数 $\boldsymbol{X}^{(u)}$（p 成分よりなる縦ベクトルとする）については

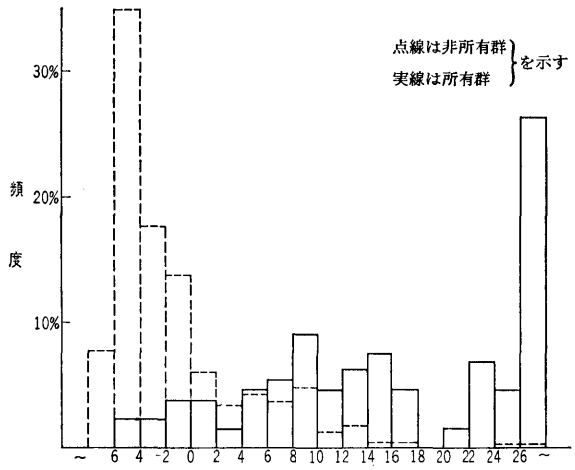
(2.1) $$p=P(\boldsymbol{\alpha}'\boldsymbol{X}^{(1)}<\boldsymbol{\alpha}'\boldsymbol{X}^{(2)})=\Phi\left(\frac{\boldsymbol{\alpha}'\boldsymbol{\mu}^{(2)}-\boldsymbol{\alpha}'\boldsymbol{\mu}^{(1)}}{\sqrt{2\boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha}}}\right)$$ （$\boldsymbol{\alpha}'$ は縦ベクトル $\boldsymbol{\alpha}$ の転置）

ただし $$\Phi(c)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{c}e^{-(1/2)t^2}dt$$

であるから，$P(\boldsymbol{a}'\boldsymbol{X}^{(1)}<\boldsymbol{a}'\boldsymbol{X}^{(2)})$（$\boldsymbol{a}'\boldsymbol{\mu}^{(1)}<\boldsymbol{a}'\boldsymbol{\mu}^{(2)}$ なら）あるいは $P(\boldsymbol{a}'\boldsymbol{X}^{(1)}>\boldsymbol{a}'\boldsymbol{X}^{(2)})$（$\boldsymbol{a}'\boldsymbol{\mu}^{(1)}>\boldsymbol{a}'\boldsymbol{\mu}^{(2)}$ なら）のいずれかを最大にする $\boldsymbol{a}$ を係数とする $\boldsymbol{a}'\boldsymbol{X}^{(u)}{}_{(u=1,2)}$ は一次判別関数である．

次の例は AOR グループの購買層分析に現われた一例で，電気冷蔵庫の所有，非所有の有様の特性分析である．特性（要因と考える）としては生活水準（5 分類，範疇化の方法は，このグループでの研究がある．例えば [5] の 141 頁），電気冷蔵庫以外の電気器具所有の有様（9 分類），年令（7 分類），学歴（3 分類），世帯主職業（10 分類），購読新聞の種類（7 分類），新聞閲読の有様（4 分類），消費および貯蓄性向（5 分類），生活合理化の態度（4 分類）の 9 種類である．また各特性の各分類（category）に与えられる数量は，級間（電気冷蔵庫所有，非所有）分散 σ_b^2 と全分散 σ^2 の比，すなわち相間比 $\eta=\sigma_b^2/\sigma^2$ を最大にするようなものとして与えられた（この方法については，例えば [4], [5] 参照）．

試みに数量化ののちの所有，非所有群の値の頻度分布を図表で示すと，第 A 図のようになって，両者の分布の分離は甚だよいことが期待できる．$\eta=0.71$ であるが，分布の分離の測度



第 A 図

として，例えば，この経験分布に対して，minimax 基準を適用して求めた分割点により，それより値の高いものを所有，そうでないものを非所有と判別したときの，判別成功率の推定値は 85% となり，確かに分離のよいことがわかる．だが，この場合には，新たな標本を所有，非所有者の群に判別することよりは，分布の分離に寄与する特性の検出が主要関心事となろうから，むしろ $p=P(Y^{(1)}<Y^{(2)})$ のような量が分布の分離の測度として (2.1) のような意味で適当ではないかと思われる．ただし，$Y^{(1)}$ は数量が与えられたのちの非所有者の値を示す確率変数，$Y^{(2)}$ は所有者のそれである．

しかしながら，この場合には，各特性のとり得る値は，そのなかの category への反応であるから，tie を考慮しないわけにはゆかない．そこで次のような修正が必要となろう．記述を簡単にするために，$Y^{(1)}$, $Y^{(2)}$ のとり得る値を $i=1,2,\cdots,r$ とし，そこでの相対頻度をそれぞれ $\hat{f}_i^{(1)}=\mathrm{est}\,P(Y^{(1)}=i)$, $\hat{f}_i^{(2)}=\mathrm{est}\,P(Y^{(2)}=i)$, その経験分布関数を $\hat{F}_i^{(1)}=\mathrm{est}\,P(Y^{(1)}\leqq i)$, $\hat{F}_i^{(2)}=\mathrm{est}\,P(Y^{(2)}\leqq i)$ としよう．ただし，$\mathrm{est}\,P(\quad)$ は $P(\quad)$ の推定量を示す．このとき，$P(Y^{(1)}<Y^{(2)})$ の推定量として，

(2.2)
$$\hat{\hat{p}}=\sum_{i=1}^{r}\hat{F}_{i-1}^{(1)}\hat{f}_i^{(2)}+\frac{1}{2}\sum_{i=1}^{r}\hat{f}_i^{(1)}\hat{f}_i^{(2)}$$
$$=\frac{1}{2}+\frac{1}{2}\sum_{i=1}^{r}[\hat{F}_i^{(1)}\hat{f}_i^{(2)}-\hat{F}_i^{(2)}\hat{f}_i^{(1)}]$$

を採用する．これについては [6] でも触れたが，類似の議論は [13] にもある．これを第 A 図の場合へ適用すると，$\hat{\hat{p}}=0.9217$ となる．また推定量の安定性という点からみても，(2.1) の $\hat{\hat{p}}$ の形からみて，$\hat{F}^{(1)}=\hat{F}^{(2)}$ を分割点として，その判別成功率を見るより安定することが期待できる．

また η を最大にする数量化では，例えば特性間の高次の association が無視できて，漸近的に正規分布が仮定できるとしても，その分散の等しいことは一般に期待できない．例えば各特性が A か non-A の二様の反応を示す dichotomous type の場合を考えると，これは容易に理解される．従って Neyman-Pearson の Lemma をよりどころになな尤度比を考える立場からは，線型のものが最適にはならないわけであるが，この例のように分布の分離がよければ，線型の便利さがむしろ優先する場合が多いであろう．またその故に何らかの方法で分布の分離の度合の評価の必要がおころう．もしこのために前述の $\hat{\hat{p}}$ を使うとすれば，§1 で述べた評価の方法が適用できる．[12] は分布関数の連続性の仮定なしに，$p=P(Y^{(1)}\leqq Y^{(2)})$ の推定量として，$\hat{\hat{p}}=\{Y_i^{(1)}\leqq Y_j^{(2)}$ なる組 $(Y_i^{(1)}, Y_j^{(2)})$ の数$\}/n_1 n_2$ とすれば，(1.7) を使った評価の可能なことを示したが，しかし経験分布が完全に一致したとき，$\hat{\hat{p}}=1/2$ となるような推定量がのぞましければ，(2.2) のような量を考えねばならない．

統計数理研究所

考 参 文 献

[1] T. W. Anderson, An Introduction to Multivariate Analysis, Wiley and Sons, New York, 1958.

[2] Z. W. Birnbaum, On a use of the Mann-Whitney statistic, Proceedings of the Third Berkeley Symposium, Vol. I (1956).

[3] Z. W. Birnbaum and R. C. McCarty, A distribution-free upper confidence bound for $Pr\{Y<X\}$ based on independent samples of X and Y, Ann. Math. Statist., Vol. 29 (1958).

[4] 林 知己夫，数量化理論とその応用例，統計数理研究所彙報，第2巻，第1号 (1953).

[5] 林 知己夫，村山孝喜，市場調査の計画と実際，日刊工業新聞社 (1964).

[6] 藤本 熙，その他，「週刊誌の買われ方」と「ニュースは新聞かテレビか」，教育統計 No. 59 (1959).

[7] 藤本 熙，分類について，I，統計数理研究所彙報，第11巻，第1号 (1963).

[8] 藤本 熙，On a distribution-free two-way classification, Ann. Inst. Statist. Math., Vol. 16 (1964).

[9] W. Hoeffding, Probability inequalities for sum of bounded random variables, Amer. Statist. Assoc., Vol. (1963).

[10] E. L. Lehman, Notes on the theory of estimation, Lectures note, (1950).

[11] E. L. Lehman, Consistency and unbiasedness of certain nonparametric tests, Ann. Math. Statist., Vol. 22 (1951).

[12] D. R. Owen, K. J. Craswell and D. L. Hanson, Nonparametric upper confidenec bound for $Pr\{Y<X\}$ and confidence limits for $Pr\{Y<X\}$ when X and Y are normal, Amer. Statist. Assoc., Vol. 59 (1964).

[13] J. Putter, The treatment of ties in some nonparametric tests, Ann. Math. Statist., Vol. 26 (1955).