入学者数予測と合格者数決定について

小林みどり, 高野 加代子

A method to estimate the percentage of university enrollment of successful candidates

Midori KOBAYASHI and Kayoko TAKANO

Abstract

We report a method to estimate what percentage of students who passed the entrance examination actually enter the school using the central limit theorem.

1 はじめに

募集人員が N 人のとき合格者数を何人にしたらよいかは, 毎年, 入試委員が頭を悩ませる問題である. 入学者 (入学手続きをする者) がほぼ N 人となるように合格者数を決めるにはどのようにしたらよいだろうか. 合格者数を決定するために, 様々な方法が考えられている. たとえば, 過去の歩留まり (入学者数/合格者) とその年の社会状況から, その年の歩留まりを予測して合格者数を決める方法や, オープンキャンパス参加者数などのいくつかの変数を組み合わせて式をつくり合格者数を決める方法などが工夫されている [3]. ここでは, それらの方法とは異なり, 個々の受験者のデータに基づき, 各受験者の入学確率を推定し, それにより入学者数を予測して合格者数を決める方法について報告する. 本学部では, 様々な方法によりいくつかの案をつくり, それらの案を比較検討して最終的な合格者数を決定している.

2 合格者数の決定方法

全受験者を得点の高い順に並べて A_1, A_2, A_3, ... とする. 受験者 A_i が, 本学部に合格したときに入学する確率を p_i とする. p_i は次のように推定する. 受験者 A_i が受験した他大学を U_1, U_2, ..., U_t とする. ここで, U_1, U_2, ..., U_t は A_i が受験したすべての他大学ではなく, 本学部とその大学の両方に合格したときに, その大学への入学を選択すると判断される大学のみを取り上げる. この判断は, 入試センターからの情報 (受験者の得点, 住所, 性別, 志望大学, 志望学部等) を基に行う. 大学 U_1, U_2, ..., U_t に合格する確率を受験者 A_i の得点から判断し, それぞれ $r_1, r_2, \ldots, r_t$ とする. これは, 予備校が算出している合格確率を利用する. このとき, $(1-r_1)(1-r_2)\cdots(1-r_t)$ が, 本学部に合格したときに本学部へ入学する確率である. この値を p_i とおく: すなわち $p_i = (1-r_1)(1-r_2)\cdots(1-r_t)$ である. そこで, 募集人員が N 人のとき, $\Sigma_{i=1}^{n} p_i \geq N$ を満たす最小の n をもって合格者数と決める.

3　中心極限定理

確率変数 $x_i\ (i=1,2,\cdots)$ は独立で，それぞれ有限な平均 $E(x_i)$ と有限な分散 $Var(x_i)$ をもつものとする．$S_n=\sum_{i=1}^{n}x_i\ (n=1,2,\cdots)$ とおく．x_i の独立性から $Var(S_n)=\sum_{i=1}^{n}Var(x_i)$ となる．

Lindeberg の中心極限定理 [2]　任意の $\varepsilon>0$ に対して

$$\frac{1}{Var(S_n)}\sum_{i=1}^{n}E[(x_i-E(x_i))^2;|x_i-E(x_i)|\geq\varepsilon\sqrt{Var(S_n)}]\to 0\ (n\to\infty)\qquad(*)$$

ならば，$(S_n-E(S_n))/\sqrt{Var(S_n)}$ の分布は，$n\to\infty$ のとき，平均 0，分散 1 の正規分布 $N(0,1)$ に収束する．

条件 $(*)$ を Lindeberg の条件という．

特別な場合として，$x_i\ (i=1,2,\cdots)$ は独立で，平均 μ，分散 $\sigma^2>0$ の同分布に従うものとする．このとき，$Var(S_n)=n\sigma^2$ であるから，任意の $\varepsilon>0$ に対して

$$\frac{1}{Var(S_n)}\sum_{i=1}^{n}E[(x_i-E(x_i))^2;|x_i-E(x_i)|\geq\varepsilon\sqrt{Var(S_n)}]$$
$$=\frac{1}{\sigma^2}E\left[(x_1-\mu)^2;|x_1-\mu|\geq\varepsilon\sqrt{n}\sigma\right]$$

となり，$P(|x_1-\mu|/\sigma\varepsilon\geq\sqrt{n})\to 0\ (n\to\infty)$ により，右辺は $n\to\infty$ のとき 0 に収束する．したがって，常に Lindeberg の条件が満たされる．この場合の例として，$\{X_i\,|\,i=1,2,\cdots,n\}$ が，平均 μ，分散 $\sigma^2>0$ の分布をもつ母集団からの十分大きな標本であるとする．その標本平均 $\bar{X}(n)=(X_1+X_2+\cdots+X_n)/n$ は中心極限定理により，近似的に平均 μ，分散 σ^2/n の正規分布 $N(\mu,\sigma^2/n)$ に従う．

また，確率変数 $x_i\ (i=1,2,\cdots)$ は独立で，それぞれ確率分布 $P(x_i=1)=p_i, P(x_i=0)=1-p_i$ に従うものとする．$E(x_i)=p_i, Var(x_i)=p_i(1-p_i)$ であるから，$S_n=\sum_{i=1}^{n}x_i$ とおくと $E(S_n)=\sum_{i=1}^{n}p_i, Var(S_n)=\sum_{i=1}^{n}p_i(1-p_i)$ となる．$n\to\infty$ のとき $Var(S_n)\to\infty$ となることを仮定する．このとき任意の $\varepsilon>0$ に対して，

$$\frac{1}{Var(S_n)}\sum_{i=1}^{n}E[(x_i-E(x_i))^2;|x_i-E(x_i)|\geq\varepsilon\sqrt{Var(S_n)}]$$
$$=\frac{1}{Var(S_n)}\sum_{i=1}^{n}\{(1-p_i)^2P(x_i=1,1-p_i\geq\varepsilon\sqrt{Var(S_n)})$$
$$+p_i^2P(x_i=0,p_i\geq\varepsilon\sqrt{Var(S_n)})\}$$

である．$Var(S_n)\to\infty\ (n\to\infty)$ により，十分大きな n に対して $\varepsilon\sqrt{Var(S_n)}>1$，したがって $P(x_i=1,1-p_i\geq\varepsilon\sqrt{Var(S_n)})=P(x_i=0,p_i\geq\varepsilon\sqrt{Var(S_n)})=0\ (i=1,2,\cdots,n)$ とな

る．ゆえに，十分大きな n に対して $\sum_{i=1}^{n} E[(x_i - E(x_i))^2 ; |x_i - E(x_i)| \geq \varepsilon\sqrt{Var(S_n)}] = 0$ となり，Lindeberg の条件が満たされる．すなわち $\sum_{i=1}^{n} p_i(1-p_i) \to \infty$ $(n \to \infty)$ と仮定すると $(S_n - \sum_{i=1}^{n} p_i)/\sqrt{\sum_{i=1}^{n} p_i(1-p_i)}$ の分布は $n \to \infty$ のとき $N(0,1)$ に収束する．

4 入学者数 S_n の分布

受験者 A_i が本学部に合格したと仮定する．そのとき，確率変数 x_i を，$x_i = 1$ (A_i が本学部に入学するとき)，$x_i = 0$ (A_i が本学部に入学しないとき) と定義する．そのとき，x_i は二項分布 $B(1, p_i)$ に従う．

$S_n = \sum_{i=1}^{n} x_i$ とおくと，$E(x_i) = p_i$，$Var(x_i) = p_i(1-p_i)$ であり，また，x_i $(i = 1, 2, \ldots)$ は互いに独立な確率変数であるから，S_n は，平均 $\sum_{i=1}^{n} p_i$，分散 $\sum_{i=1}^{n} p_i(1-p_i)$ の確率変数となる．

ある番号 i_0 が存在して，すべての i $(\geq i_0)$ について $p_i = 0$ or 1 であると仮定する．そのときは，i $(\geq i_0)$ について x_i は 0 か 1 かが確定するため，予測から除くことができる．したがって，この仮定が成り立たない場合を考える．p_i の決め方から，$p_i \neq 0, 1$ である p_i に対しては，ある一定区間 $[a, b]$ $(0 < a < b < 1)$ に入っているとして差し支えない．したがって，$n \to \infty$ のとき $\sum_{i=1}^{n} p_i(1-p_i) \to \infty$ が成り立つ．よって中心極限定理より，S_n は正規分布 $N(\sum_{i=1}^{n} p_i, \sum_{i=1}^{n} p_i(1-p_i))$ に近似的に従うと考えられる．

合格者数を n と決めたとき，入学者数 S_n は，平均 $\sum_{i=1}^{n} p_i$，分散 $\sum_{i=1}^{n} p_i(1-p_i)$ の正規分布に近似的に従い，確率 0.95 で $\sum_{i=1}^{n} p_i \pm 1.96\sqrt{\sum_{i=1}^{n} p_i(1-p_i)}$ の範囲に入ることが分かる．

5 例

この方法をある入試に適用したところ，表 1 のような結果となった．

表 1

n	S_n			
	平均	分散	標準偏差	95% 区間
85	44.0	14.52	3.81	36.5 ～ 51.5
90	48.6	14.86	3.85	41.0 ～ 56.2
95	53.6	14.86	3.85	46.0 ～ 61.2
100	57.1	15.23	3.90	49.5 ～ 64.7

この入試の合格者数は $n = 91$ であった．そのときの S_n の平均は 49.6，分散は 14.86，標準偏差は 3.85 であった．したがって入学者数は近似的に確率 0.95 で区間 [42.1, 57.1] に入る．我々は入学者を 50 名と予測したところ，実際の入学者は 53 名であった．

また別の入試に適用したところ，合格者数は $n = 131$ であり，S_n の平均は 101.0，分散は 13.31，標準偏差は 3.65 であった．したがって入学者数は近似的に確率 0.95 で区間 [93.8, 108.2]

に入る. 我々は入学者を 101 名と予測したところ, 実際の入学者は 103 名であった.

6　S_n の分散

$p_1, p_2, \ldots, p_n$ の平均を p とおく: すなわち $p = (p_1 + p_2 \cdots + p_n)/n$. そのとき, S_n の分散は $Var(S_n) = np - \sum_{i=1}^{n} p_i^2$ である. 平均が p となる $p_1, p_2, \ldots, p_n$ の組合せの中で $\sum_{i=1}^{n} p_i^2$ が最小となるのは, すべての p_i が等しいときである. したがって, $Var(S_n)$ は $p_1 = p_2 = \cdots = p_n$ のとき最大の値をとる. つまり p_i がすべて等しいとき, S_n の分散は最大となるのである. Feller はこのことに関していくつかの例を挙げている [1]. たとえば, ある地域で 1 年間におこる火事の件数 S を確率変数として扱い, S の平均は一定であるとする. 火事を起こす確率がすべての家で等しいとき, 火事の件数 S の分散は最大となる. また, たとえば, n 個の機械の質の平均を一定とすると, すべての機械が同質のときに output が一番ばらつくことになる. 我々の例では, すべての受験者の入学確率が等しいときに入学者数のばらつきが最も大きくなるのである. Feller はこの現象を “striking result” であると述べている.

参考文献

[1] W. Feller, Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed., John Wiley & Sons, 1950, pp229-231.

[2] 伊藤 清,「確率論 II」岩波講座基礎数学, 岩波書店, 1997, p203.

[3] 福田 宏, 経営情報学部合格者数と入学者数の関係, 経営と情報, Vol. 16, No. 1, 2003.