S. TRYBUŁA (Wrocław)

OCENA FREKWENCJI W POPULACJI ELEMENTÓW NALEŻĄCYCH DO KLAS NIEREPREZENTOWANYCH W PRÓBCE(1)

Załóżmy, że populacja generalna jest podzielona na skończoną lub przeliczalną ilość klas. Załóżmy dalej, że istnieje przepis pozwalający rozstrzygnąć, czy dwa elementy dowolnie wybrane z populacji należą do różnych klas, czy nie. Z populacji pobieramy, zgodnie ze schematem Bernoulliego, próbkę o liczności n. Na ogół nie wszystkie klasy będą reprezentowane w próbce. Jak ocenić frekwencję w populacji elementów należących do klas w próbce niereprezentowanych? Podamy klasę statystyk, które będą nieobciążonymi estymatorami tej frekwencji.

Oznaczmy przez $\nu_{k,n}^{(i)}$ zmienną losową zdefiniowaną w następujący sposób: $\nu_{k,n}^{(i)} = 1$ wtedy i tylko wtedy, gdy w danej próbce o liczności n i-ta klasa jest reprezentowana k razy. W pozostałych przypadkach $\nu_{k,n}^{(i)} = 0$. Jeżeli przez p_i oznaczymy nieznaną frekwencję i-tej klasy w populacji, to wartość oczekiwana $E(\nu_{k,n}^{(i)})$ będzie równa

$$E(\nu_{k,n}^{(i)}) = P(\nu_{k,n}^{(i)} = 1) = \binom{n}{k} p_i^k (1-p_i)^{n-k}. \tag{1}$$

Oznaczmy przez $\nu_{k,n}$ sumę $\sum_i \nu_{k,n}^{(i)}$; wtedy

$$E(\nu_{k,n}) = \sum_i E(\nu_{k,n}^{(i)}) = \binom{n}{k} \sum_i p_i^k (1-p_i)^{n-k}. \tag{2}$$

Zdefiniujmy zmienną losową $\mu_{k,n}^{(i)}$; $\mu_{k,n}^{(i)} = p_i$, gdy w danej próbce o liczności n i-ta klasa jest reprezentowana k razy, i $\mu_{k,n}^{(i)} = 0$ w pozostałych przypadkach. Zmienna $\mu_{k,n} = \sum_i \mu_{k,n}^{(i)}$ jest frekwencją elementów

---

(1) Frekwencja elementów w populacji należących do klas niereprezentowanych w próbce jest zmienną losową. Dlatego też termin *estymator* oznaczający statystykę pozwalającą ocenić wartość tej frekwencji jest niepoprawny. Anglicy nazywają taką statystykę *predict function*. Może należałoby wymyślić nowy termin, np. *predyktor*. Termin *prognoza*, spotykany w polskich publikacjach, uważam w danym przypadku za niezręczny.

w populacji należących do klas k razy reprezentowanych w próbce. Jej wartość oczekiwana jest równa

$$(3)\quad E(\mu_{k,n}) = \sum_i E(\mu_{k,n}^{(i)}) = \sum_i p_i P(\mu_{k,n}^{(i)} = p_i) = \binom{n}{k} \sum_i p_i^{k+1}(1-p_i)^{n-k}.$$

Z wzoru (2) wynika

$$(4)\quad \frac{\binom{m}{s}}{\binom{N+m+1}{r+s+1}} E(\nu_{r+s+1,N+m+1}) = \sum_i p_i^{r+1}(1-p_i)^{N-r}\left[\binom{m}{s} p_i^s (1-p_i)^{m-s}\right],$$

gdzie $m \geqslant s$. Połóżmy

$$(5)\quad \varkappa_{r,N}^{(m)} = \sum_{s=0}^{m} \frac{\binom{N}{r}\binom{m}{s}}{\binom{N+m+1}{r+s+1}} \nu_{r+s+1,N+m+1}.$$

Z wzoru (4) otrzymamy

$$E(\varkappa_{r,N}^{(m)}) = \binom{N}{r} \sum_{s=0}^{m} \sum_i p_i^{r+1}(1-p_i)^{N-r}\left[\binom{m}{s} p_i^s (1-p_i)^{m-s}\right] =$$

$$= \binom{N}{r} \sum_i p_i^{r+1}(1-p_i)^{N-r}\left[\sum_{s=0}^{m} \binom{m}{s} p_i^s (1-p_i)^{m-s}\right] =$$

$$= \binom{N}{r} \sum_i p_i^{r+1}(1-p_i)^{N-r},$$

a więc ostatecznie

$$(6)\quad E(\varkappa_{r,N}^{(m)}) = E(\mu_{r,N}) \qquad (m = 0, 1, 2, \ldots).$$

Zmienna losowa $\varkappa_{r,N}^{(m)}$ jest kombinacją liniową zmiennych $\nu_{r+s+1,N+m+1}$, to jest ilości klas mających w próbce o liczności $N+m+1$, $r+s+1$ reprezentantów. Zmienna $\mu_{r,N}$ jest łączną frekwencją elementów należących do klas mających r reprezentantów w próbce o liczności N. Gdy $m \ll N$, zmienna losowa może służyć do oceny frekwencji elementów, należących do klas r razy reprezentowanych w danej próbce o liczności N. W szczególności statystyka

$$\varkappa_{0,N}^{(m)} = \sum_{s=0}^{m} \frac{\binom{m}{s}}{\binom{N+m+1}{s+1}} \nu_{s+1,N+m+1} \qquad (m = 0, 1, 2, \ldots)$$

może służyć do oceny frekwencji elementów niereprezentowanych w próbce. Przypuśćmy np., że z populacji słów zawartych w polskich dziennikach z 1957 roku pobrano, zgodnie ze schematem Bernoulliego, próbkę o liczności 10000. Aby ocenić frekwencję słów, których w próbce nie spotkano, losuje się np. 5 dodatkowych słów z populacji ($m = 4$) i oblicza się wartość statystyki $\varkappa^{(4)}_{0,10000}$. Wzór (6) powiada nam, że statystyka $\varkappa^{(4)}_{0,10000}$ jest nieobciążonym estymatorem wartości oczekiwanej tej frekwencji.

Dla $r = 0$ i $m = 0, 1$ mamy

$$\varkappa^{(0)}_{0,N} = \frac{1}{N+1}\nu_{1,N+1},$$

$$\varkappa^{(1)}_{0,N} = \frac{1}{N+2}\nu_{1,N+2} + \frac{2}{(N+1)(N+2)}\nu_{2,N+2}.$$

Ogólnie biorąc, przy małych r i $N \gg m$ współczynniki przy $\nu_{r+s+1,N+m+1}$ we wzorze (5) szybko maleją, gdy $s \to m$.

Można obliczyć wariancję zmiennej $\varkappa^{(m)}_{r,N}$. Wystarczy skorzystać z wzorów

$$E(\nu^{(i)2}_{k,n}) = \binom{n}{k} p_i^k (1-p_i)^{n-k},$$

$$E(\nu^{(i)}_{k,n}\nu^{(i)}_{l,n}) = 0, \quad \text{gdy} \quad k \neq l,$$

$$E(\nu^{(i)}_{k,n}\nu^{(j)}_{l,n}) = \frac{n!}{k!l!(n-k-l)!} p_i^k p_j^l (1-p_i-p_j)^{n-k-l} \qquad (i \neq j)$$

i rozbić $E[\varkappa^{(m)}_{r,N} - E(\varkappa^{(m)}_{r,N})]^2$ na kombinację liniową wartości oczekiwanych wyrażeń $\nu^{(i)}_{k,n}\nu^{(j)}_{l,n}$. Otrzymany rezultat jest jednak bardzo skomplikowany i nie będę go tutaj podawał.

Statystyki $\varkappa^{(m)}_{r,N}$ mają tę zaletę, że przy ich użyciu zbędna jest informacja o ilości klas w populacji. Często, jak np. w podanym wyżej przykładzie, informacji takiej brak.

Zmienne losowe $\nu_{k,n}$ były badane przez różnych autorów, np. Gooda [1]. Posłużyły one, między innymi, do oceny frekwencji p_i oraz do oceny ilości klas w próbce o liczności αN ($\alpha > 1$), gdy dana jest próbka o liczności N.

Praca cytowana

[1] I. J. Good, *The population frequencies of species and the estimation of population parameters*, Biometrika 40 (1953).

INSTYTUT MATEMATYCZNY POLSKIEJ AKADEMII NAUK

Praca wpłynęła 11. 4. 1958

С. ТРЫБУЛА (Вроцлав)

ОЦЕНКА ЧАСТОТЫ В СОВОКУПНОСТИ ЭЛЕМЕНТОВ, ПРИНАДЛЕЖАЩИХ КЛАССАМ, НЕ ИМЕЮЩИМ ПРЕДСТАВИТЕЛЕЙ В ВЫБОРКЕ

РЕЗЮМЕ

Если из совокупности, разделенной на конечное или счетное число классов, выберем согласно схеме Бернулли выборку, состоящую из N элементов, то ожидаемое значение частоты элементов, принадлежащих классам имеющим по r представителей в выборке, будет равно

$$m_{r,N} = \binom{N}{r} \sum_i p_i^{r+1}(1-p_i)^{N-r} \qquad (r = 0, 1, 2, \ldots),$$

где p_i есть (неизвестная) частота в совокупности элементов, принадлежащих i-му классу. Обозначим через $\nu_{r,N}$ число классов с r представителями в выборке, состоящей из N элементов. Доказывается, что

$$\varkappa_{r,N}^{(m)} = \binom{N}{r} \sum_{\delta=0}^{m} \frac{\binom{m}{\delta}}{\binom{N+m+1}{r+\delta+1}} \nu_{r+\delta+1,N+m+1} \qquad (m = 0, 1, 2, \ldots)$$

являются несмещенными оценками параметра $m_{r,N}$. Приводятся возможные приложения этого результата, в частности к оценке частоты элементов, принадлежащих к классам, которые не имеют представителей в выборке. Заметим, что для составления статистик $\varkappa_{r,N}^{(m)}$ не нужны сведения о числе классов в совокупности.

S. TRYBUŁA (Wrocław)

THE ESTIMATION OF FREQUENCY IN A POPULATION OF ELEMENTS BELONGING TO CLASSES NOT REPRESENTED IN THE SAMPLE

SUMMARY

If we take a sample of size N according to the Bernoulli scheme from a population divided into a finite or denumerable number of classes, then the expected value of the frequency, in the population, of elements belonging to classes represented r times in the sample will be

$$m_{r\,N} = \binom{N}{r} \sum_i p_i^{r+1}(1-p_i)^{N-r} \qquad (r = 0, 1, 2, \ldots),$$

where p_i is the (unknown) frequency of the elements belonging to the i-th class in the population. Denote by $\nu_{r,N}$ the number of classes represented r times in a given sample of size N. It can be shown that

$$\varkappa_{r,N}^{(m)} = \binom{N}{r} \sum_{\delta=0}^{m} \frac{\binom{m}{\delta}}{\binom{N+m+1}{r+\delta+1}} \nu_{r+\delta+1,N+m+1} \qquad (m = 0, 1, 2, \ldots)$$

are unbiased estimators of the parameter $m_{r,N}$.

The possible applications of this result are given, in particular to the estimation of frequency of elements belonging to classes which are not represented in the sample.

It will be observed that in order to form the statistics $\varkappa_{r,N}^{(m)}$ no information concerning the number of classes in the population is needed.