JANOS JANOSITZ
Akademia Górnicza w Miszkolcu

OCENA REPREZENTATYWNOŚCI POBIERANIA PRÓBEK

UKD 543.05:519.26/.27:552.1:553.1

Zbiór wartości ξ zmiennej losowej niezależnych od siebie określeń charakteryzujących badaną cechę danego ośrodka nazywamy próbą. Proces określania tej próby nazywamy pobieraniem próbki. Od pojedynczych elementów próby wymagamy, aby były od siebie niezależne i aby w sposób reprezentatywny przedstawiały badany zbiór. Można to osiągnąć w warunkach, gdy badana cecha posiada ciągły rozkład gęstości prawdopodobieństwa. Rozkład ten jest monotonicznie rosnący w obrębie odcinka (a, b). Gęstość prawdopodobieństwa $f(x)$ badanej cechy określa dystrybuanta:

$$F(x) = \int_{a}^{x} f(x)\,dx$$

Gdy badana cecha zależy od funkcji położenia $F(x)$, wówczas miejsca pobrania próbek powinny być rozłożone równomiernie w całej przestrzeni badanego obszaru. Gdy warunek równomiernego rozmieszczenia próbek nie jest spełniony, np. gdybyśmy próbki pobierali tylko z części danego rejonu możemy nie otrzymać próby reprezentatywnej. Funkcję położenia można jednak rozłożyć na sumy rozkładów $F_i(x)$ odcinkowych o s_i wagach proporcjonalnych do długości a_i odcinków [1]:

$$F(x) = \sum_{i=1}^{n} s_i F_i(x)$$

gdzie:

$$F_i(x) = \int_{a_i}^{a_i + \Delta a_i} f_i(x)\,dx \qquad [1]$$

$$s_i = \frac{a_i}{A}$$

i gdzie na ogół $f_i(x) \neq f(x)$
przy czym:

$F(x)$ — jest funkcją rozkładu badanej wartości x zbioru A,

a_i — jest „wielkością" elementarnego zbioru położonego wewnątrz badanego zbioru,

$F_i(x)$ — jest funkcją rozkładu x w danym cząstkowym zbiorze.

Ogólnie $F_i(x)$ różni się od $F(x)$, więc nawet przy próbkach pobranych z rejonu zawężonego założony wzór empiryczny funkcji rozkładu nie będzie zbliżony do $F(x)$. Próbki należy pobierać tak, aby „odległości", zmierzone między sąsiednimi miejscami poboru próbek w obrębie danego rejonu były możliwie jednakie, tzn. aby ich zagęszczenie w całym rejonie było jednakowe.

W ten sposób można najskuteczniej zapewnić reprezentatywność pobranych próbek, gdy ilość tych próbek jest dostatecznie duża, ponieważ każda elementarna próbka ma tę samą wagę, mimo że pochodzi z różnych miejsc badanego rejonu. Gdy więc wszystkie próbki jednakowe mają wspólne wagi i pochodzą z różnych miejsc badanego rejonu, a ze względu na ciągłość tych zmian próbki będą charakterystyczne dla strefy o coraz to innych właściwościach, wówczas taką próbkę możemy uznać za reprezentatywną.

Takie określenie obszarów próbek eliminuje przypadek, gdy odległości między próbkami są dowolnie małe, a zatem gdy zmienna ich wartość ma charakter ciągły. Wówczas oba punkty pobrania próbek leżą w bezpośrednim sąsiedztwie. Próbki dublujące się charakteryzują się tym samym obszarem. Obszar ten byłby w tym przypadku bez uzasadnienia w próbce liczony podwójnie, co nie byłoby prawidłowe. Dlatego też jako wartości takich dublujących się próbek należy traktować łącznie jako wartość średnią z obu próbek dublujących się.

Rozważmy dla jakiego obszaru odnosi się to ważne stwierdzenie. Wyjdźmy z założenia, iż każdy obszar próbki ma tę samą wagę, zatem w naszych rozważaniach nie ma takiej próbki, której należałoby przypisać większą wagę. Gdy liczba próbek $= N$, a każda próbka jest dla swego otoczenia reprezentatywna, tzn. czy spełnia warunek reprezentatywności dla obszaru o promieniu równym r. Obszar ten przedstawia n-tą część całości badanego obszaru. Wychodząc z tego założenia, gdyby miejsca pobrania dwóch sąsiednich próbek oddalone były od siebie o odległość mniejszą jak r, to nie mogą być one uznane za samoistne jednostki próbki ogólnej. Na podstawie powyższego, zajmijmy się teraz problemem związanym z określeniem wagi kilku różnych próbek. Zagadnienie to można przeprowadzić jedną z metod:

a) Przez kilkakrotne całkowanie stochastycznej zmiennej metodą Monte-Carlo, rozwiązując całkę:

$$I = \int_{G_n} f_n(P_n)\, dP_n \qquad [2]$$

przy czym:

G_n — jest danym obszarem w przestrzeni n-wymiarowej, określonym zbiorem,

P_n — $P_n(x_1, x_2, \ldots, x_n)$ punktów przynależnych do obszaru G_n,

dP_n — $dx_1 \ldots dx_n$.

Jeżeli w obszarze G_n posiadamy ilość N punktów P_i rozłożonych regularnie, stosując stochastyczne całkowanie metodą Monte-Carlo możemy z [2] napisać:

$$I = |G_n|\, E[f_n(P_n)] \cong |G_n| \frac{1}{N} \sum_{i=1}^{N} f_n(P_{ni}) \qquad [3]$$

gdzie:

E — operator wartości oczekiwanej (nadzieja matematyczna),

$|G_n|$ — oznacza obszar G_n w przestrzeni n wymiarowej.

Przyjęto, że wielowymiarowy rozkład $f_n(P_n) = \eta_n$ w obrębie ograniczonego obszaru (a, b) ma przebieg monotoniczny. Jeżeli całkowanie przeprowadzimy dla zmiennej x_n, *otrzymamy:*

$$I = \int_{G_{n-1}} f_{n-1}(P_{n-1})\, dP_{n-1} \qquad [4]$$

gdzie

$$f_{n-1}(P_{n-1}) = \int_{x_n = \nu_1(P_{n-1})}^{\mu(P_{n-1})} f_n(P_n)\, dx_n \qquad [5]$$

Znaczenie wyrażeń G_{n-1}, P_{n-1}, dP_{n-1} oraz η_{n-1} jest analogiczne jak poprzednio. Ogólnie, jeżeli się całkuje względem zmiennej „k", otrzyma się w wyniku całkowania zmienną $\eta_{(n-k)}$. W przypadku pierwszym $\eta_{(n-k)}$ posiada wartość oczekiwaną w przybliżeniu:

$$\overline{\eta}_{(n-k)} = \frac{\sum_{i=1}^{N} (n-k)\, i}{\sum_{i=1}^{N} |G_{(n-k)\,i}|} \qquad [6]$$

przy czym:

$|G_{(n-k)i}|$ — jest powierzchnią takich obszarów o ilości wymiarów niższym od tego, dla którego określona jest całka [2].

Próbkę złożoną z $\eta_{(n-k)i}$ elementów uważać należy za reprezentatywną wtedy, jeżeli dwa sąsiednie miejsca pobrania leżą w odległości mniejszej od r. Wielkość r można oznaczyć wychodząc z założenia, że suma poszczególnych płaszczyzn w rejonie o promieniu r da wartość zgodną z wartością $|G_n|$.

Jeżeli próbka pobrana została w punkcie $(k = 0)$, wówczas wartość przeciętna może być obliczona na podstawie [3] ze stopniem prawdopodobieństwa przy odchyleniu granicznym ε. Przy pobraniu próbek o wymiarze niepunktowym, jeżeli z [6] otrzymamy wartość $\overline{\eta}_{(n-k)}$, wtedy ε—t określić możemy na podstawie następującego wzoru na N'

$$N' \cong \frac{|G_n|}{|G_{er}|} \qquad [7]$$

Przy czym:

$|G_{er}|$ — jest wielkością obszaru o promieniu r miejsca poboru próbki o wymiarze punktu.

Wartość r należy określić na podstawie metody podanej poprzednio. Jeżeli zaś można podać wartość rozkładu $f_n(P_n)$ w obszarze G_n, wtedy można również określić standard σ. Pobraną próbkę z $\eta_{(n-k)i}$ jednostek możemy przyrównać do próbki punktowej. Jeżeli zaś w wyrażeniu $\eta_{(n-k)i}$ także i k jest zmienną, wtedy średnią η_n należy obliczyć według

$$\overline{\eta}_{ni} = \frac{\eta_{(n-k)\,i}}{|G_{(n-k)\,i}|} \qquad [8]$$

Próbki także będą reprezentatywne, jeśli dane części $|G_{(n-k)i}|$ hiperpłaszczyzn nie są oddalone od

siebie mniej, jak o r. Tak jak poprzednio r określić można z warunku, gdy części hiperpłaszczyzn $|G_{(n-k)t}|$ wraz z otoczeniem o promieniu r dadzą w sumie $|G_n|$.

Gdy wielkość pojedynczych obszarów leżących w płaszczyznach wyższego rzędu z przestrzenią o promieniu r będzie $|G_{(n-k)tr}|$, wtedy:

$$\bar{\eta}_n = \frac{\sum_{i=1}^{N} |G_{(n-k)tr}| \bar{\eta}_{ni}}{|G_n|} \qquad [9]$$

czyli pojedyncze próbki $\bar{\eta}_{ni}$ należy wstawić do rachunku z takimi wagami, którym odpowiada wielkość $|G_n|$ reprezentowana w przestrzeni próby częściowej.

Wartość r wyznaczamy podobnie, jak to już uczyniono uprzednio według formuły [7], wówczas określimy N', a więc można również obliczyć σ_n'.

b) Pobieranie próbki wieloelementarnej.

Przyjąwszy, że x oznacza pewną cechę, której dystrybuanta oznaczona jest przez $F(x)$. Jeżeli $F(x)$ w obrębie obszaru jest ściśle monotoniczne, wówczas szacowanie oczekiwanej wartości może nastąpić przy użyciu metod poprzednio opisywanych na podstawie x próbek w ilości N.

W praktyce pobieranie próbki x_i w punkcie jest niemożliwe. Z materiału zbioru o objętości V przy pobieraniu próbki zwykle wyodrębnia się ilość ΔV_i, a w laboratorium oznacza się dla tej ilości przynależną wartość średnią x_i. Jeżeli ze zbioru o rozkładzie $F(x)$ pobiera się próbki V_i o wielkości stałej, wówczas średni rozkład wartości x_i w objętości V_i będzie $F_i(x)$. Jeżeli do niego przynależna wartość σ jest σ_i, wtedy:

$$\sigma_i \leqq \sigma \qquad [10]$$

gdzie:

σ — jest rozrzutem $F(x)$.

Przypadek ten ważny jest dla każdej próbki o wielkości $\Delta V_i < V$. Do oznaczenia odchylenia ważne są zależności podane [2]. W dalszym ciągu należy zbadać, jak można oszacować granice błędu ε dla różnych poziomów prawdopodobieństwa, gdy jest różna wielkość próbek. Przy analizie tego zagadnienia należy rozróżnić dwa przypadki:

1) gdy:

$$\sum_{i=1}^{N} \Delta V_i = \Delta V \qquad [11]$$

czyli próbka całkowita jest wobec całości badanego zbioru nikła (gdy ΔV wynosi tylko kilka % V),

2) gdy ΔV w stosunku do V nie da się pominąć.

Dla uproszczenia przyjmiemy próbki o objętości V_i jako kule o promieniu r_i, co daje:

$$\frac{4r_i^3\pi}{3} = \Delta V_i \quad (i = 1, 2, \ldots, N) \qquad [12]$$

więc jeżeli wartość r jest rozumiana jak podano poprzednio, to do jej określania posłużyć może zależność:

$$\sum_{i=1}^{N} \frac{4(r_i + r)^3\pi}{3} = V \qquad [13]$$

Z tego wynika, że w pierwszym przypadku:

$$r \cong \sqrt[3]{\frac{3V}{4N\pi}}; \qquad [14]$$

gdyż wartość $\frac{\Delta V}{V}$ jest nikła. Dlatego też wobec r pominąć możemy także V_i. Tym samym próbki o różnej objętości V_i są wyrazem objętości prawie że równych, więc przy określaniu na ich podstawie wartości x_i należy przyjąć jednakowe uwagi. W drugim przypadku wartość r należy określić według [13]. Znając już tę wartość oblicza się $\bar{X}$, jak następuje:

$$\bar{X} = \frac{4\pi}{3} \cdot \frac{\sum_{i=1}^{N} (r_i + r)^3 \bar{x}_i}{V} \qquad [15]$$

Znaczy to, iż każda próbka częściowa do próbki ogólnej wstawiona będzie z taką wagą, jaka odpowiada jej udziałowi w całej objętości V.

Naturalnie, obliczenia te słuszne są tylko wówczas, gdy próbki są od siebie niezależne oraz reprezentatywne. Wartości średnie można oznaczyć na podstawie znanych z literatury (2, 3) metod odnośnie do odchylenia odpowiadającego danemu stopniowi prawdopodobieństwa. Oczywiście, że wartość σ należy wypośrodkować z empirycznego wzoru na rozkład. (Także w przypadku różnych wag próbek). Wzór [4] udowadnia, że przy ważonych wartościach przeciętnych, lepsze wyniki daje zwykła średnia jednakowych wag, gdyż wykazuje ona najmniejszy rozrzut. To samo można udowodnić przy założeniu, iż próbka reprezentuje zawartość danej przestrzeni o promieniu r, gdy pobrano próbki łącznego zbioru. Przy warunku [16] zależność [15] określi nam wartość r,

$$\sum_{i=1}^{N} \Delta V_i = \Delta V = \text{constans} \qquad [16]$$

która wtedy osiągnie swoją wartość najniższą przy określonych wartościach N i V, jeżeli próbki posiadają jednakową wielkość.

$$\left(\Delta V_i = \Delta V_j = \frac{\Delta V}{N} \; i \neq j\right).$$

LITERATURA

1. Edwin F., Backenbach — Modern matematika mérnököknek. Müszaki Könyvkiadó, Budapest, 1960.
2. Janositz J. — Adott intervallumon változó folytonos eloszlású valószinüségi változó várható értékének vizsgálata. Nehezipari Müszaki Egyetem Közleményi (w druku).
3. Janositz J. — A várhato értélk analizise. Bányászati és Kohászati Lapek. Bányászat (w druku).
4. Jánossy L. — A valószinüségélmélet alapjai és nehány alkalmazása különös tekintetel mérési eredmenyek kiértékélese. Tankönyvkiadó, 1965.
5. Prékopa A. — Valószinüségelmét. Müszaki Könyvkiadó, Budapest, 1962.
6. Rényi A. — Valószinüségszámitás. Tankönyvkiado, Budapest, 1954, 1966.
7. Srejgyer I. U. — Monte Carló módszerek. Müszaki Könyvkiadó, Budapest, 1965.
8. Smirnov N. W., Dunin-Borkowski I. W. — Mathematische Statistik in der Technik. VEB Deutscher Verlag der Wissenschaften, Berlin, 1963.

SUMMARY

The paper deals with the problem of taking representative samples in the case when any feature of a given series analysed changes continuously. To obtain a representative picture of the whole series we must take samples within the series analysed, according to a uniform density net, i.e. we must take care to obtain samples that represent similar volume. Under such conditions, when the amount of sampling points are determined, any sample must representatively determine the features of a given series within an area (having radius r). On this basis a possibility exists to determine average weight of a sample. The weight should always be proportional to the area with the radius r determined previously.

Average value of an error ε, calculated on the basis of samples having known probability level, may also be determined. The amount of samples characterized by various weight is a quotient of the entire area and of the elementary areas with the radius r.

РЕЗЮМЕ

В статье обсуждается методика отбора представительных проб, в том случае когда один из признаков исследуемого объекта изменяется последовательно. Для того, чтобы пробы были представительные для всего объекта, они должны удовлетворять условию, чтобы места опробования располагались по всему объему с одинаковой густотой, следовательно, чтобы каждая проба представляла одинаковый объем. В условиях определенного количества мест опробования проба определяет свойства объекта на площади с разиусом r. Таким образом, при пробах величиной k ($<$n) в — размерном пространстве можно определить средний вес пробы, когда k ($<$n) является постоянной величиной, или же когда взятые пробы различны (k переменное). Вес всегда должен быть прямопропорционален ранее определенной площади с радиусом r.

Можно также определить величину средней ошибки E на основании проб с известным уровнем вероятности. Количество проб разного веса является кратностью всей площади и элементарных участков с радиусом r.