ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
VIII (1965)

A. PLUCIŃSKA (Warszawa)

O PEWNYCH ZAGADNIENIACH ZWIĄZANYCH Z PODZIAŁEM POPULACJI NORMALNEJ NA CZĘŚCI

Przy poprawnie przebiegającej produkcji masowej przyjmuje się na ogół, że cechy produkowanych elementów mają rozkłady normalne. Przeprowadzane testy statystyczne dają empiryczne potwierdzenie słuszności tej hipotezy.

W pewnych praktycznych zagadnieniach powstaje konieczność wyboru spośród całej populacji jej „lepszej części". Problem taki powstaje na przykład wtedy, gdy produkowane elementy mają być wmontowywane w bardzo precyzyjne urządzenia. Wówczas dozwolone są tylko bardzo nieznaczne różnice między rzeczywistymi wartościami ich cech i wartościami znamionowymi. Zakładamy tutaj, że produkcja jest dobrze uregulowana i wartość znamionowa pokrywa się z wartością przeciętną.

Całą populację normalną dzielimy wtedy na dwie populacje, z których jedna charakteryzuje się bardzo dużym skupieniem wartości cech wokół wartości znamionowej, a druga stosunkowo dużym rozproszeniem wartości cech. Segregacja elementów realizowana jest przez automatyczne urządzenie, które może popełniać błędy.

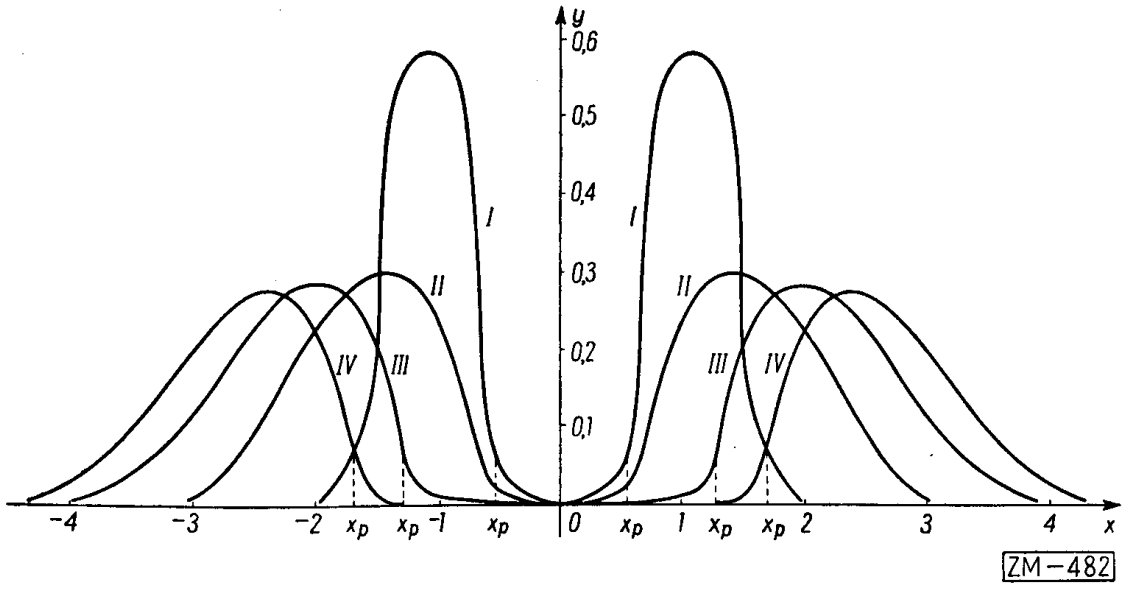
Oznaczając przez X określoną cechę rozważanych elementów i dobierając układ współrzędnych tak, by wartość przeciętna cechy była zerem, możemy w następujący sposób określić optymalny podział populacji na dwie części: dla pierwszej z nich powinien być spełniony warunek $|X| \leqslant a$, dla drugiej $|X| > a$. Liczbowa wartość stałej a wynika z konkretnego zagadnienia technicznego.

Efekt optymalnego podziału populacji można więc opisać za pomocą rozkładów uciętych. Jednakże taki opis wyniku segregacji jest nie zawsze możliwy, bo postuluje bezbłędną pracę automatu segregującego. Zagadnienie błędów szeregowania zostało szczegółowo omówione przez Borodaczewa [1] (§§ 4 i 6). Niezależnie od tego operowanie rozkładami uciętymi jest niewygodne ze względów analitycznych. W wielu zagadnieniach praktycznych trzeba badać własności różnych funkcji zmiennych losowych, co prowadzi do skomplikowanych rachunków w przypadku, gdy argumenty tych funkcji mają rozkłady ucięte.

Funkcją, za pomocą której wygodnie jest opisywać resztową populację powstałą z populacji normalnej po symetrycznej selekcji elementów centralnych, jest funkcja (rys. 1)

$$(1) \qquad f(x) = \frac{1}{\Gamma(k+\frac{1}{2})(2\sigma^2)^{k+1/2}} x^{2k} e^{-x^2/2\sigma^2}, \qquad -\infty < x < +\infty, \; k \geqslant 0.$$

Metoda rozumowania, które doprowadziło do uzyskania takiej właśnie postaci funkcji, podana jest w oddzielnym komentarzu na końcu pracy.



Rys. 1. Wykresy funkcji $y = \frac{1}{\Gamma(k+\frac{1}{2})(2\sigma^2)^{k+1/2}} x^{2k} e^{-x^2/2\sigma^2}$.

I: $k = 2$, $\sigma = \frac{1}{2}$; II: $k = 1$, $\sigma = 1$; III: $k = 2$, $\sigma = 1$; IV: $k = 3$, $\sigma = 1$

Funkcja (1) ma, między innymi, następujące własności dogodne dla opisu rozważanej populacji: jest ona ciągła dla $-\infty < x < +\infty$; szybko zbieżna do prostej $y = 0$ przy $|x| \to \infty$, symetryczna względem prostej $x = 0$; ma dwie wartości maksymalne położone symetrycznie względem początku układu współrzędnych; pozwala na taki dobór parametrów, by

$$(2) \qquad \int_{-a}^{a} f(x)\,dx \leqslant \varepsilon,$$

gdzie stała ε charakteryzuje w pewien sposób dokładność selekcji.

Zauważmy, że dla $k = 0$ funkcja (1) jest gęstością rozkładu normalnego.

Niech X będzie zmienną losową o gęstości (1). Moment rzędu $2r$ ($r = 1, 2, \ldots$) zmiennej losowej X określony jest wzorem

$$(3)\quad E(X^{2r}) = 2\int_0^\infty \frac{1}{\Gamma(k+\frac{1}{2})(2\sigma^2)^{k+1/2}} x^{2k+2r} e^{-x^2/2\sigma^2} dx =$$

$$= \frac{\Gamma(k+r+\frac{1}{2})(2\sigma^2)^{k+r+1/2}}{\Gamma(k+\frac{1}{2})(2\sigma^2)^{k+1/2}} 2\int_0^\infty \frac{x^{2(k+r)}}{\Gamma(k+r+\frac{1}{2})(2\sigma^2)^{k+r+1/2}} e^{-x^2/2\sigma^2} dx$$

$$= \frac{\Gamma(k+r+\frac{1}{2})}{\Gamma(k+\frac{1}{2})} (2\sigma^2)^r.$$

Z symetrii rozkładu wynika, że

$$(4)\quad E(X^{2r+1}) = 0, \quad r = 0, 1, 2, \ldots$$

Wyprowadzimy teraz wzory na funkcję charakterystyczną, a następnie wzory dokładne i przybliżone na gęstość sumy s niezależnych zmiennych losowych o jednakowych rozkładach (1).

Niech $X_1, X_2, \ldots, X_s$ będą niezależnymi zmiennymi losowymi o jednakowych rozkładach (1). Funkcję charakterystyczną tych zmiennych znajdziemy jedynie dla przypadku, gdy parametr k jest liczbą naturalną. Przyjmijmy ponadto, że $\sigma = 1$. Nie zmniejsza to oczywiście ogólności rozważań, oznacza jedynie przyjęcie na osi układu współrzędnych skali, której jednostką jest σ.

W wyniku $2k$-krotnego różniczkowania stronami względem t tożsamości

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx} e^{-x^2/2} dx = e^{-t^2/2},$$

otrzymujemy

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (ix)^{2k} e^{itx} e^{-x^2/2} dx = W_{2k}(t) e^{-t^2/2},$$

gdzie $W_{2k}(t)$ jest wielomianem Hermite'a (patrz np. [2], str. 131) stopnia $2k$.

Funkcję charakterystyczną zmiennej losowej o rozkładzie (1) można teraz napisać w następującej postaci

$$\varphi(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx = \frac{1}{\sqrt{2\pi}\,(2k-1)!!} \int_{-\infty}^{+\infty} e^{itx} x^{2k} e^{-x^2/2} dx =$$

$$= \frac{(-1)^k}{(2k-1)!!} W_{2k}(t) e^{-t^2/2}.$$

Funkcja charakterystyczna zmiennej losowej

$$Y_s = \sum_{i=1}^{s} X_i$$

określona jest wzorem

$$\varphi_s(t) = \frac{(-1)^{ks}}{[(2k-1)!!]^s}[W_{2k}(t)]^s e^{-st^2/2}. \tag{5}$$

Znając funkcję charakterystyczną można wyznaczyć gęstość i obliczyć wszystkie interesujące prawdopodobieństwa.

W szczególności dla $k = 1$ wzór (5) przyjmuje postać

$$\varphi_s(t) = (1-t^2)^s e^{-st^2/2}, \tag{6}$$

a gęstość można wyrazić za pomocą wzoru

$$\begin{aligned} f_s(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\varphi_s(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty}(1-t^2)^s e^{-st^2/2-itx}\,dt = \\ &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left[1-\frac{1}{s}\left(u-\frac{ix}{\sqrt{s}}\right)^2\right]^s e^{-u^2/2}\,du = \\ &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\sum_{r=0}^{s}\frac{(-1)^r}{s^r}\left(u-\frac{ix}{\sqrt{s}}\right)^{2r} e^{-u^2/2}\,du = \\ &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\sum_{r=0}^{s}\frac{(-1)^r}{s^r}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left(u-\frac{ix}{\sqrt{s}}\right)^{2r} e^{-u^2/2}du = \\ &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\sum_{r=0}^{s}\frac{(-1)^r}{s^r}\mu'_{2r}, \end{aligned} \tag{7}$$

gdzie przez μ'_{2r} oznaczony jest moment rzędu $2r$ unormowanej zmiennej losowej normalnej względem wartości $ix/\sqrt{s}$. Wyrażając moment μ'_{2r} przez momenty centralne ([3]) możemy napisać wzór (7) w następującej postaci

$$\begin{aligned} f_s(x) &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\sum_{r=0}^{s}\frac{(-1)^r}{s^r}\binom{s}{r}\sum_{j=0}^{2r}\binom{2r}{j}\left(-\frac{xi}{\sqrt{s}}\right)^j\mu_{2r-j} = \\ &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\sum_{r=0}^{s}\frac{(-1)^r}{s^r}\binom{s}{r}\sum_{j=0}^{s}\binom{2r}{2j}\left(-\frac{x^2}{s}\right)^j\mu_{2(r-j)} = \\ &= \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s}\sum_{l=0}^{s}(-1)^l x^{2l}\sum_{j=l}^{s}\binom{2j}{2l}\mu_{2(j-l)}\frac{(-1)^j}{s^{j+l}}\binom{s}{j}. \end{aligned} \tag{8}$$

Ze wzoru (8) dla $s = 2$, $s = 3$, $s = 4$ kolejno otrzymujemy gęstości

$$f_2(x) = \frac{1}{\sqrt{4\pi}}\left(\frac{3}{2^2} - \frac{1}{2^2}x^2 + \frac{1}{2^4}x^4\right)e^{-x^2/4}, \tag{9}$$

$$f_3(x) = \frac{1}{\sqrt{6\pi}}\left(\frac{4}{3^2} + \frac{2}{3^2}x^2 - \frac{2}{3^4}x^4 + \frac{1}{3^6}x^6\right)e^{-x^2/6}, \tag{10}$$

$$f_4(x) = \frac{1}{\sqrt{8\pi}}\left(\frac{153}{4^4} - \frac{5}{4^4}x^2 + \frac{66}{4^6}x^4 - \frac{12}{4^7}x^6 + \frac{1}{4^8}x^8\right)e^{-x^2/8}. \tag{11}$$

Przy obliczaniu prawdopodobieństwa tego, że zmienna losowa Y_s przyjmuje wartości z przedziału (α, β), czyli całki

$$\int_\alpha^\beta f_s(x)\,dx$$

można skorzystać z tablicy wartości całek

$$\frac{1}{\sqrt{2\pi}}\int_0^\infty x^k e^{-x^2/2}\,dx$$

umieszczonych w [4] i [5].

Wzór (8) daje dokładną gęstość. Został on wyprowadzony dla $k = 1$. W podobny sposób można wyprowadzić wzór na gęstość zmiennej losowej Y_s w przypadku, gdy parametr k jest liczbą naturalną. Korzystanie z dokładnych wzorów dla dużych wartości s jest jednak bardzo skomplikowane.

Znajdziemy teraz wzór przybliżony, prawdziwy dla dowolnego $k \geqslant 0$, korzystając z rozwinięcia funkcji w szereg Edgewortha (patrz np. [2], str. 223).

Oznaczmy

$$\psi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \tag{12}$$

$$Z_i = \frac{X_i}{\mu_2} = \frac{X_i}{2k+1}, \quad i = 1, 2, \ldots, s, \tag{13}$$

zaś przez $\varkappa_r$ półniezmiennik rzędu r zmiennej losowej o rozkładzie (1).

Ponieważ $\mu_{2r+1} = 0$, zatem

$$\varkappa_{2r+1} = 0 \quad \text{dla} \quad r = 1, 2, \ldots$$

Gęstości $g_s(x)$ zmiennej losowej $\frac{1}{\sqrt{s}}\sum_{i=1}^{s} Z_i$ możemy w przybliżeniu wyrazić jako

$$g_s(x) = \psi(x) + \frac{1}{s}\left[\frac{\varkappa_4}{4!(2k+1)^2}\psi^{(4)}(x)\right] + \tag{14}$$

$$+ \frac{1}{s^2}\left[\frac{\varkappa_6}{6!(2k+1)^3}\psi^{(6)}(x) + \frac{\varkappa_4^2}{2(4!)^2(2k+1)^4}\psi^{(8)}(x)\right],$$

z błędem rzędu $1/s^3$.

Jeśli przyjmiemy, że

$$g_s(x) = \psi(x) + \frac{1}{s} \cdot \frac{\varkappa_4}{4!(2k+1)^2} \psi^{(4)}(x), \tag{15}$$

to popełnimy błąd rzędu $1/s^2$. A więc w zależności od żądanej dokładności i wartości s możemy stosować wzór (14) lub (15).

Wartości półniezmienników $\varkappa_4$ i pochodnych funkcji $\psi(x)$ występujących we wzorach (14), (15) są następujące

$$\begin{aligned}
\varkappa_2 &= \mu_2 = 2k+1, \\
\varkappa_4 &= \mu_4 - 3\mu_2^2 = -4k(2k+1), \\
\varkappa_6 &= \mu_6 - 30\mu_4\mu_2 + 15\mu_2^2 = 16k(2k+1)(4k+1), \\
\psi^{(4)}(x) &= (3 - 6x^2 + x^4)\psi(x), \\
\psi^{(6)}(x) &= (-15 + 45x^2 - 15x^4 + x^6)\psi(x), \\
\psi^{(8)}(x) &= (105 - 420x^2 + 210x^4 - 28x^6 + x^8)\psi(x).
\end{aligned}$$

Dla $k = 1$ wzór (15) przyjmuje postać

$$g_s(x) = \psi(x)\left[1 - \frac{1}{18s}(3 - 6x^2 + x^4)\right].$$

Przechodząc do zastosowań przeprowadzonych rozważań teoretycznych wskażemy metody ułatwiające dobór takich wartości parametrów k, σ funkcji (1), by funkcja ta była wygodną aproksymacją rzeczywistego rozkładu.

Na rysunku 1, na którym przykładowo podanych jest kilka wykresów funkcji rodziny (1), widać, iż krzywa stromo wznosi się ku górze na prawo od punktu x_p i analogicznie na lewo od punktu $-x_p$, gdzie x_p, $-x_p$ są odciętymi punktów przegięcia krzywej.

Przyjmijmy, że punkty podziału $-a$ oraz a pokrywają się z odciętymi punktów przegięcia krzywej, czyli

$$a = x_p = \sqrt{\frac{4k+1-\sqrt{16k+1}}{2}}\,\sigma, \quad \text{dla} \quad k > \tfrac{1}{2}. \tag{16}$$

Ponadto nałóżmy warunek

$$\int_{-a}^{a} f(x)\,dx = \varepsilon. \tag{17}$$

Stała ε występująca we wzorze (17) charakteryzuje w pewien sposób błędy selekcji. Możemy powiedzieć, iż ε jest tą frakcją populacji wyjściowej, która powinna być zakwalifikowana do pierwszej grupy, a została zakwalifikowana do drugiej.

Napiszmy wzór (17) w postaci

$$\frac{1}{\Gamma(k+\frac{1}{2})(2\sigma^2)^{k+1/2}}\int_0^a x^{2k}e^{-x^2/2\sigma^2}dx = \frac{\varepsilon}{2}.$$

Podstawiając $x^2/2\sigma^2 = y$ i oznaczając $k-\frac{1}{2} = m$ mamy

$$\frac{1}{\Gamma(m)}\int_0^{m-1/4-\sqrt{m-7/16}} e^{-y}y^{m-1}dy = \varepsilon. \tag{18}$$

Na podstawie tablic niekompletnej funkcji gamma ([4]) w przybliżeniu wyznaczamy z równania (18) niewiadomą m.

Przykładowo kilka wartości m, ε związanych wzorem (18) podanych jest w tablicy 1.

TABLICA 1

k	$\frac{3}{4}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3	$\frac{7}{2}$
ε	0,05	0,07	0,09	0,095	0,1	0,11	0,12

Zamiast warunku (17) możemy do wyznaczania parametrów k, σ posłużyć się również następującą metodą:

Przyjmujemy równość (16), obliczamy na podstawie próbki wariancję empiryczną s^2 i przyjmujemy, że spełniona jest równość

$$s^2 = (2k+1)\sigma^2. \tag{19}$$

Z układu równań (16), (19) wyznaczamy k oraz σ.

Oczywiście można proponować wiele różnych metod wyznaczania parametrów k, σ. Ostatecznym etapem powinno być sprawdzenie, za pomocą odpowiedniego testu statystycznego, zgodności rozkładu empirycznego z proponowanym rozkładem teoretycznym.

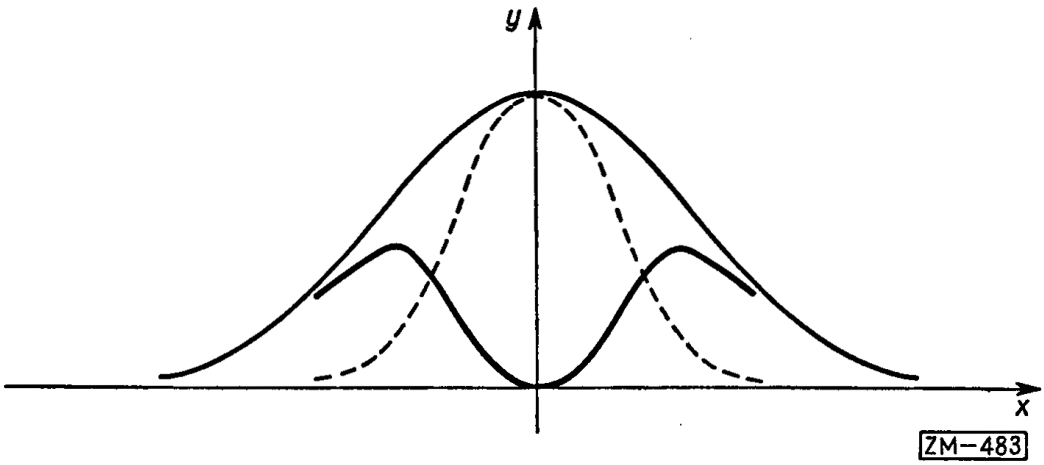
Zagadnienie sum zmiennych losowych jest, jak to było wspomniane we wstępie pracy, zagadnieniem często występującym w zastosowaniach. Na przykład w elektrotechnice, przy analizie obwodów elektrycznych, spotykamy się często z taką sytuacją, w której pewna cecha dotycząca układu elementów równa jest sumie odpowiednich cech każdego z elementów: może to być oporność lub indukcyjność w przypadku szeregowo połączonych oporników lub cewek albo pojemność w przypadku równolegle połączonych kondensatorów. Wyprowadzone w niniejszej pracy wzory dla rozkładu sumy można więc na przykład zastosować do badania łącznego oporu s połączonych szeregowo oporników, przy czym przyjmujemy, że opór opornika jest zmienną losową o funkcji gęstości (1).

Komentarz do wzoru (1). Metoda rozumowania, która doprowadziła do uzyskania wzoru (1) w przypadku, gdy parametr k jest liczbą naturalną, była następująca. Dla uproszczenia zapisu przyjmijmy $\sigma = 1$. Populacja normalna ma gęstość

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{20}$$

Z populacji tej automat segregujący wybiera część środkową. Matematycznie proces selekcji można scharakteryzować jako odejmowanie (rys. 2) od funkcji (20) pewnej funkcji $r(x)$, spełniającej warunek

$$0 < r(x) \leqslant \varphi(x).$$



Rys. 2. Krzywe charakteryzujące podział populacji normalnej na dwie części
$y = \varphi(x)$ ———, $y = r(x)$ — — — —, $y = \varphi(x) - r(x)$ ▬▬

Przyjmijmy

$$r(x) = \varphi(x) e^{-x^k},$$

gdzie k jest liczbą naturalną.

Populacja powstała z populacji normalnej przez symetryczne wycięcie środka jest więc scharakteryzowana funkcją

$$\varphi(x) - r(x) = \varphi(x)\{1 - e^{-x^k}\} = \varphi(x)\{1 - [1 - Cx^k + \ldots]\} \approx Cx^k \varphi(x).$$

Ponieważ otrzymana funkcja ma być gęstością, należy ją pomnożyć przez taką stałą, by całka niewłaściwa w przedziale $(-\infty, +\infty)$ z tej funkcji była równa jedności. Stąd otrzymujemy wzór (1).

Prace cytowane

[1] Н. А. Бородачев, *Основные вопросы теории точности производства*, Москва 1950.

[2] H. Cramér, *Metody matematyczne w statystyce*, Warszawa 1958.

[3] M. G. Kendall, *The advanced theory of statistics*, London 1952.

[4] В. И. Пагурова, *Таблицы неполной гамма-функции*, Москва 1963.

[5] K. Pearson, *Tables for statisticians and biometricians*, Cambridge, vol. 1, 1924, vol. 2, 1931.

[6] G. Rouzet, *Etude des moments de la loi normale tronquée*, Révue de Statistique Appliquée 10 (1962), str. 49-61.

Praca wpłynęła 11. 4. 1964

A. ПЛЮЦИНЬСКА (Варшава)

О НЕКОТОРЫХ ВОПРОСАХ, СВЯЗАННЫХ С ДЕЛЕНИЕМ НОРМАЛЬНОЙ СОВОКУПНОСТИ НА ЧАСТИ

РЕЗЮМЕ

Элементы массового производства делят часто на определенные классы с различной точностью характеристик этих элементов. В настоящей работе указана функция, с помощью которой удобно схарактеризовать совокупность, возникшую с нормальной совокупности путем симметрического выреза центра. Этой функцией является (1).

Введены следующие формулы для суммы s независимых случайных величин плотности (1):

а) формула (5) для характеристической функции в случае, когда k является натуральным числом,

б) формула (8) для плотности при $k = 1$,

в) приближенные формулы (14), (15) для плотности действительные при любом $k \geqslant 0$; ошибка при их применении имеет порядок $1/s^3$ и $1/s^2$ соответственно.

A. PLUCIŃSKA (Warsaw)

ON CERTAIN PROBLEMS CONNECTED WITH A DIVISION OF A NORMAL POPULATION INTO PARTS

SUMMARY

Elements of mass production are often divided into classes according to the different accuracy of the characteristic features of those elements. This paper contains a formula by means of which it is convenient to characterize a population arising from a normal population by cutting out the centre symmetrically. This is function (1).

The following formulas for the sum of s independent random variables with density function (1) have been derived:

a) formula (5) for the characteristic function in the case where k is a natural number,

b) formula (8) for the density function where $k = 1$,

c) approximate formulas (14), (15) for the density function, true for any $k \geqslant 0$; the errors arising in their application are quantities of order $1/s^3$ and $1/s^2$ respectively.
