ZESZYTY NAUKOWE Instytutu Zarządzania i Marketingu Akademii im. Jana Długosza w Częstochowie
Seria: Pragmata tes Oikonomias 2009, z. III

Witold Miszczak
Walenty Ostasiewicz

Mieszanki rozkładów

1. Elementarne wprowadzenie

Rozpatrzmy proste zdanie.

Mamy 10 monet, w tym jest 8 prawidłowych, a 2 są zniekształcone. Przy losowym podrzucaniu jedną monetą mamy następujące rozkłady:

$$\mathrm{P}\,(\text{orzeł} \mid \text{prawidłowa}) = \frac{1}{2}$$

$$\mathrm{P}\,(\text{reszka} \mid \text{prawidłowa}) = \frac{1}{2}$$

$$\mathrm{P}\,(\text{orzeł} \mid \text{zniekształcona}) = \frac{2}{3}$$

$$\mathrm{P}\,(\text{reszka} \mid \text{zniekształcona}) = \frac{2}{3}.$$

Na przestrzeni zdarzeń elementarnych{orzeł, reszka} określmy dwie zmienne losowe:

$$Y, X\colon \{\text{orzeł, reszka}\} \to \{0,1\}$$

w następujący sposób:

$$X\,(\text{orzeł}) = 0, \quad X\,(\text{reszka}) = 1,$$

$$Y\,(\text{orzeł}) = 0, \quad Y\,(\text{reszka}) = 1.$$

Rozkłady prawdopodobieństwa tych zmiennych są następujące:

$$f_X(x)=\begin{cases}\frac{1}{2}, \text{gdy } x=0\\ \frac{1}{2}, \text{gdy } x=1\end{cases} \qquad f_Y(y)=\begin{cases}\frac{2}{3}, \text{gdy } y=0\\ \frac{1}{3}, \text{gdy } y=1\end{cases}$$

Zmieszajmy wszystkie monety i określmy zmienną losową Z: {orzeł, reszka}$\rightarrow$ {0,1} której rozkład jest następującą mieszanką dwóch rozkładów:

$$f_z(x)=\frac{8}{10}f_X(x)+\frac{2}{10}f_Y(x)$$

$$f_z(x)=\begin{cases}\frac{8}{10}\cdot\frac{1}{2}+\frac{2}{10}\cdot\frac{2}{3}, & \text{gdy } x=0\\ \frac{8}{10}\cdot\frac{1}{2}+\frac{2}{10}\cdot\frac{1}{3}, & \text{gdy } x=1\end{cases}$$

Dla większej przejrzystości przedstawmy tę mieszankę następująco:

$$\frac{8}{10}\cdot\begin{array}{|cc|}\hline 0 & 1\\ \hline \frac{1}{2} & \frac{1}{2}\\ \hline\end{array}+\frac{2}{10}\cdot\begin{array}{|cc|}\hline 0 & 1\\ \hline \frac{2}{3} & \frac{1}{3}\\ \hline\end{array}=\begin{array}{|cc|}\hline 0 & 1\\ \hline \frac{8}{15} & \frac{7}{15}\\ \hline\end{array}$$

Widzimy, że „zmieszać” dwa rozkłady jest dość łatwo i sens mieszanki dwóch rozkładów łatwo jest zrozumieć na podstawie prostego eksperymentu losowego.

O wiele trudniejszy jest problem odwrotny, polegający na tym, że dany jest jakiś rozkład, o którym wiadomo, że jest mieszanką i trzeba wyodrębnić składowe tej mieszanki.

Zauważmy przede wszystkim, że zadanie takie nie ma jednoznacznego rozwiązania. Istotę tego problemu rozpatrzmy na prostym przykładzie rozkładu zero-jedynkowego.

Dana jest pewna populacja polityków i wiadomo, że nie jest ona jednorodna. Na przykład wiadomo, że są to jastrzębie i gołębie lub że są to zieloni i czerwoni, wierni i niewierni itp. Przyjmijmy, że udział pierwszej podpopulacji oznaczony będzie symbolem α, zaś szansa wylosowania osobnika należącego do tej podpopulacji oznaczona będzie jako θ_1. Funkcję rozkładu prawdopodobieństwa w całej populacji można więc przedstawić następująco:

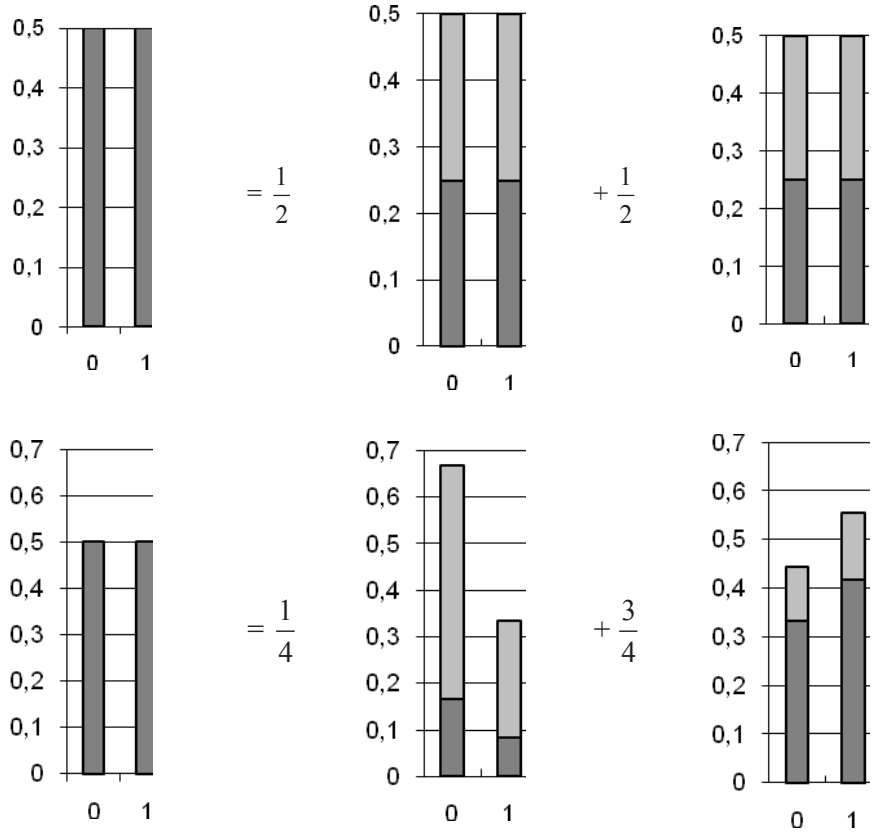
$$f_Z(z,\theta)=\alpha f_X(x,\theta_1)+(1-\alpha)f_Y(y,\theta_2)$$

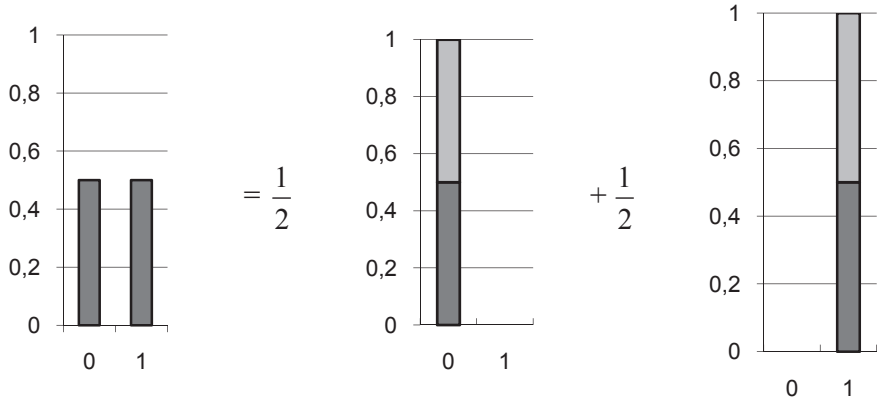
gdzie

$$f_X(x;\theta_1)=\begin{cases}\theta_1, & \text{jeśli } x=0\\ 1-\theta_1 & \text{jeśli } x=1\end{cases} \qquad f_Y(y;\theta_2)=\begin{cases}\theta_2, & \text{jeśli } y=0\\ 1-\theta_2 & \text{jeśli } y=1\end{cases}$$

Problem polega na tym, aby określić prawdopodobieństwo θ_1 i θ_2 oraz strukturę populacji, tzn. parametr α.

Istnieje wiele możliwych rozwiązań. Trzy rozwiązania przedstawiają poniższe rysunki:





Na rysunkach po lewej stronie mamy ten sam rozkład, a po prawej różne rozkłady przedstawione za pomocą wykresu kolumnowego. Cała kolumna oznacza prawdopodobieństwo dla danego x w rozkładzie będącym składnikiem mieszanki. Zawarta w niej ciemniejsza kolumna oznacza tę część prawdopodobieństwa w danym rozkładzie αP($X = x$), która wchodzi do mieszanki rozkładów. Z powyższego widzimy, że nie ma rozwiązania jednoznacznego.

Rozpatrzmy teraz przykład związany z rozkładem dwumianowym.

Załóżmy, że grupa n osób spotyka się regularnie w każdy czwartek, 15 razy w ciągu jednego semestru. Korzystając z prawa gwarantowanego przez demokrację, grupa ta przed każdym spotkaniem decyduje poprzez głosowanie, czy jakaś osoba ma wygłosić referat, czy też nie. Liczbę głosów za tym, aby wykład się odbył w i-tym tygodniu, oznaczmy symbolem k_i, $i = 1, 2, ..., 15$.

Po zakończeniu semestru znane są wyniki w postaci wektora $\mathbf{k} = (k_1, k_2, ..., k_{15})$, gdzie $k_i \in \{0,1,2,...,n\}$. Ponieważ na wynik głosowania mogą mieć wpływ różne czynniki losowe, takie jak: niewyspana noc, nieudana randka, przejedzenie itp., wektor **k** potraktujemy jako 15 realizacji zmiennej losowej K.

Przyjmijmy, że prawdopodobieństwo głosowania „za” oznaczone jest symbolem θ. Wówczas możemy traktować zmienna losowa K jako zmienną o rozkładzie dwumianowym $K \sim \mathrm{B}(n,q)$. Oznaczmy funkcję rozkładu prawdopodobieństwa tej zmiennej w postaci $B(k;n,q)$, tzn.:

$$B(k;n,\theta) = C_n^k \theta^k (1-\theta)^{n-k}$$

Niezależnie od wymienionych zakłóceń losowych cała grupa może stanowić jednorodną i zwartą społeczność, może też być podzielona na pewne klasy lub kliki.

Oznaczmy liczbę klas symbolem *c*, czyli w przypadku grupy jednorodnej mamy $c = 1$, zaś w drugim przypadku ekstremalnym, gdy grupa składa się tylko z samolubów, mamy $c = n$.

Zadanie, jakie tu jest rozwiązywane, polega na tym, aby na podstawie obserwacji zmiennej losowej *K*:

$$k_1 k_2,...,k_{15}$$

zidentyfikować klasy, z jakich się składa cała grupa.

Rozpatrzmy zadanie możliwie najprostsze, przyjmując, że w grupie są dwie frakcje, np. gołębi i jastrzębi lub doświadczonych i nieopierzonych itp. Przyjmijmy ponadto, że osoby pierwszej klasy z prawdopodobieństwem q_1 głosują „za", osoby z drugiej klasy głosują „za" z prawdopodobieństwem θ_2.

Oznacza to, że osoby należące do klasy pierwszej postępują zgodnie z rozkładem $\mathrm{B}(n,q_1)$, zaś osoby z klasy drugiej głosują zgodnie z rozkładem $\mathrm{B}(n,q_2)$.

Przyjmijmy na koniec, że do pierwszej klasy należy 100 α_1% osób, zaś do klasy drugiej 100 α_2%, przy czym $\alpha_1 + \alpha_2 = 1$ oraz $\alpha_1 > 0$, $\alpha_2 > 0$.

Przyjęte wyżej założenia oznaczają, że cała grupa jest „mieszanką" dwóch klas. To z kolei oznacza, że obserwowana funkcja rozkładu $B(k;n,\theta)$ jest mieszanką dwóch rozkładów:

$$B(k;n,\theta) = \alpha_1 \cdot B(k;n,\theta_1) + \alpha_2 B(k;n,\theta_2) \quad k = 0,1,2,...,n.$$

Sformułowane wyżej zadanie identyfikacji klas oznacza teraz, w ujęciu probabilistycznym, estymację parametrów $\alpha_1, \alpha_2, \theta_1$ i θ_2. Ponieważ $\alpha_1 + \alpha_2 = 1$, to zadanie polega ma estymacji trzech parametrów : $\alpha_1, \theta_1, \theta_2$.

W celu uproszczenia zapisów, zamiast α_1 stosowany będzie symbol α.

Zadanie polega na estymacji parametrów funkcji rozkładu prawdopodobieństwa: $B(k;n,q)$.

Załóżmy, że w grupie są tylko dwie osoby, tzn. $n = 2$.

W celu identyfikacji klas należy wówczas rozwiązać następujący układ równań:

$$B(0;2,\theta) = \alpha B(0;2,\theta_1) + (1-\alpha)B(0;2,\theta_2)$$

$$B(1;2,\theta) = \alpha B(1;2,\theta_1) + (1-\alpha)B(1;2,\theta_2)$$

$$B(2;2,\theta) = \alpha B(2;2,\theta_1) + (1-\alpha)B(2;2,\theta_2)$$

względem niewiadomych α, θ_1 i θ_2.

Korzystając z definicji funkcji rozkładu dwumianowego, układ ten zapisać można następująco:

$$B(0;2,\theta) = \alpha(1-\theta_1)^2 + (1-\alpha)(1-\theta_2)^2$$

$$B(1;2,\theta) = \alpha 2\theta_1(1-\theta_1) + (1-\alpha)2(1-\theta_2)$$

$$B(2;2,\theta) = \alpha\theta_1^{\,2} + (1-\alpha)\theta_2^{\,2}$$

Zauważmy, że

$$\sum_{k=0}^{n} B(k;n,\theta) = \sum_{k=0}^{n} C_n^k \theta^k (1-\theta)^{n-k} = 1$$

czyli w powyższym układzie trzech równań z trzema niewiadomymi jedno równanie zależy liniowo od dwóch pozostałych. Oznacza to, że układ powyższy nie ma jednoznacznego rozwiązania. Klasy są nieidentyfikowalne.

Rozpatrzmy więc przypadek, gdy $n > 2$.

Załóżmy, iż wiadomo, że liczba podpopulacji jest znana, oznaczmy ją symbolem c.

Rozkład analizowanej cechy w każdej podpopulacji jest taki sam z dokładnością do parametrów.

Oznaczmy go symbolem $f(x;\theta)$, gdzie θ jest tym parametrem. W dalszej części opracowania przyjmuje się, że jest on skalarem.

Rozkład w całej populacji jest następującą mieszanką:

$$f(x;\eta) = \sum_{i=1}^{c} \alpha_i \cdot f(x;\theta_i)$$

gdzie $\eta = (\alpha_1,...,\alpha_c,\ \theta_1,...,\theta_c)$.

Ponieważ $\alpha_i \geq 0$ oraz $\sum_{i=1}^{c} \alpha_i = 1$, to wektor $\alpha = (\alpha_1,...,\alpha_c)$, można interpretować jako rozkład prawdopodobieństwa na przestrzeni Θ, tak że

$$\alpha_i = P(\tilde{\theta} = \theta_i).$$

Do estymacji składowych wektora η stosowane są różne metody.

W kolejnych paragrafach dokładnie omówione są dwie z nich:

1) metoda największej wiarygodności,
2) metoda momentów.

2. Metoda największej wiarygodności

Załóżmy, że dane są obserwacje

$$x_1, x_2, ..., x_n$$

zmiennej losowej X, której funkcja rozkładu prawdopodobieństwa jest $f(x)$.

Funkcja wiarygodności tych obserwacji jest następująca:

$$L(x_1,...,x_n;\alpha_1,...,\alpha_c,\theta_1,...,\theta_c) = \prod_{j=1}^{n} f(x_i) = \prod_{j=1}^{n} \sum_{i=1}^{c} \alpha_i f(x_j;\theta_i) .$$

Oznaczmy jej logarytm jako l, tzn.:

$$l = \log L(x_1,...,x;\alpha_1,...,\alpha_c,\theta_1,...,\theta_c) .$$

Ponieważ nieznane parametry α_i muszą spełniać warunek

$$\sum_{i=1}^{n} \alpha_i - 1 = 0 ,$$

to maksymalizacji funkcji l dokonujemy za pomocą metody mnożników Lagrange'a. Funkcję, którą trzeba maksymalizować, oznaczmy tym samym symbolem. Ma więc ona następującą postać:

$$l = \log \prod_{j=1}^{n} \sum_{i=1}^{c} \alpha_i f(x_j;\theta_i) - \lambda \left(\sum_{i=1}^{c} \alpha_i - 1 \right)$$

gdzie λ jest to mnożnik Lagrange'a.

W celu określania nieznanych parametrów trzeba więc rozwiązać następujący układ równań:

$$\frac{\partial l}{\partial \alpha_i} = 0$$

$$\frac{\partial l}{\partial \theta_i} = 0$$

Rozwińmy lewe strony pierwszego zestawu równań:

$$\frac{\partial l}{\partial \alpha_i} = \frac{\partial\left[\log\prod_{j=1}^{n} f(x_j) - \lambda\left(\sum_{j=1}^{c}\alpha_j - 1\right)\right]}{\partial \alpha_i} = \frac{\partial \log\prod_{j=1}^{n} f(x_j)}{\partial \alpha_i} - \lambda =$$

$$= \frac{\partial\sum_{j=1}^{n}\log f(x_j)}{\partial \alpha_i} - \lambda = \sum_{j=1}^{n}\frac{\partial \log f(x_j)}{\partial \alpha_i} - \lambda =$$

$$= \sum_{\mathrm{j=1}}^{\mathrm{n}}\frac{1}{f(x_j)}\cdot\frac{\partial f(x_j)}{\partial \alpha_i} - \lambda = \sum_{j=1}^{n}\frac{1}{f(x_j)}\cdot\frac{\partial\sum_{k=1}^{c}\alpha_k \cdot f_k(x_j;\theta_k)}{\partial \alpha_i} - \lambda =$$

$$= \sum_{j=1}^{\mathrm{n}}\frac{1}{f(x_j)}\cdot f_i(x_j;\theta_i) - \lambda = \sum_{j=1}^{n}\frac{f_i(x_j;\theta_i)}{f(x_j)} - \lambda$$

Podobnie przedstawione są rozwinięcie lewej strony drugiego zestawu równań

$$\frac{\partial l}{\partial \theta_i} = \frac{\partial\left[\log\prod_{j=1}^{n} f(x_j) - \lambda\left(\sum_{k=1}^{c}\alpha_k - 1\right)\right]}{\partial \theta_i} = \frac{\partial \log\prod_{j=1}^{n} f(x_j)}{\partial \theta_i} =$$

$$= \sum_{j=1}^{n}\frac{\partial \log f(x_j)}{\partial \theta_i} = \sum_{j=1}^{n}\frac{1}{f(x_j)}\cdot\frac{\partial f(x_j)}{\partial \theta_i} =$$

$$= \sum_{j=1}^{n}\frac{1}{f(x_j)}\frac{\partial\sum_{k=1}^{c}\alpha_k \cdot f(x_j;\theta_k)}{\partial \theta_i} = \sum_{j=1}^{n}\frac{1}{f(x_j)}\cdot\alpha_i\cdot\frac{\partial f_i(x_j;\theta_i)}{\partial \theta_i}$$

W celu określenia parametrów α_i oraz θ_i należy rozwiązać następujący układ dwóch zestawów równań:

$$\sum_{j=1}^{n}\frac{f_i(x_j;\theta_i)}{f(x_j)} = \lambda \ , \quad i = 1, 2, ..., c$$

$$\sum_{j=1}^{n}\frac{\alpha_i}{f(x_j)}\frac{\partial f_i(x_j;\theta_i)}{\partial \theta_i} = 0 \ , \ i = 1, 2, ..., c$$

Mnożąc stronami równania pierwszego zestawu przez α_i i sumując po wszystkich i, uzyskujemy:

$$\sum_{i=1}^{c}\alpha_i \sum_{j=1}^{n}\frac{f_i(x_j;\theta_i)}{f(x_j)} = \sum_{i=1}^{c}\alpha_i \cdot \lambda$$

$$\sum_{j=1}^{n}\frac{\sum_{i=1}^{c}\alpha_i f_i\left(x_j;\theta_i\right)}{f\left(x_j\right)} = \sum_{j=1}^{n}\frac{f\left(x_j\right)}{f\left(x_j\right)} = \sum_{j=1}^{n}1 = n = \lambda$$

W ten sposób został wyznaczony nieokreślony mnożnik Lagrange'a.

Korzystając z tego, że $\lambda = n$, pierwsze równanie można zapisać następująco:

$$\sum_{j=1}^{n}\frac{f_i(x_j;\theta_i)}{f(x_j)} = n \quad , \quad i = 1, 2, ..., c$$

Pomnóżmy je stronami przez α_i i uzyskamy

$$\sum_{j=1}^{n}\frac{\alpha_i f_i(x_j;\theta_i)}{f(x_j)} = \alpha_i n$$

Ponieważ udziały α_i mieszanki rozkładów traktujemy jako następujące prawdopodobieństwo:

$$\alpha_i = P(\tilde{\theta} = \theta_i),$$

to uzyskujemy następujące równanie:

$$\sum_{j=1}^{n}\frac{P(\tilde{\theta} = \theta_i)\cdot f_i(x_j;\theta_i)}{f(x_j)} = \alpha_i \cdot n$$

Zamiast oznaczenia $P\left(\tilde{\theta} = \theta_i\right)$ wprowadźmy proste oznaczenie p_i oznaczające w dalszym ciągu prawdopodobieństwo przynależności dowolnej obserwacji do *i*-tej podpopulacji (klasy). Przedstawmy teraz powyższe równanie w postaci:

$$\sum_{j=1}^{n}\frac{f(x_j \mid i)\cdot p_i}{f(x_j)} = \alpha_i n$$

Bez większego trudu rozpoznajemy pod znakiem sumy wzór Bayesa:

$$P(i \mid x_j) = \frac{f_i(x_j \mid i)p_i}{f(x_j)}$$

Korzystając z twierdzenia Bayesa, ostatecznie otrzymujemy następujące równanie:

$$\sum_{j=1}^{n} P(i \mid x_j) = \alpha_i n$$

gdzie $P(i|x_j)$ oznacza prawdopodobieństwo, że obserwacja x_j należy do *i*-tej podpopulacji (*i*-tej klasy). Dzieląc stronami to równanie przez *n*, wyznaczamy nieznaną wielkość α_i, którą traktujemy jako estymator parametru α_i.

Jako estymator parametru α_i przyjmiemy następujące wyrażenie:

$$\hat{\alpha}_i = \frac{1}{n} \sum_{j=1}^{n} P(i \mid x_j)$$

W celu uzyskania wyrażenia określającego estymatory parametrów θ_i przedstawmy drugi zestaw równań w postaci:

$$\sum_{j=1}^{n} \frac{\alpha_i}{f(x_j)} \cdot f_i(x_j;\theta_i) \cdot \frac{\partial \log f_i(x_j;\theta_i)}{\partial \theta_i} = 0$$

Korzystając ponownie w podobny sposób z twierdzenia Bayesa, uzyskujemy:

$$\sum_{j=1}^{n} P(i \mid x_j) \frac{\partial \log f_i(x_j;\theta_i)}{\partial \theta_i} = 0 \quad , \; i = 1, 2, ..., c.$$

Estymatory $\hat{\theta}_i$ parametrów θ_i uzyskujemy jako rozwiązania układu równań, do którego należy podstawić konkretne postaci $f_i(x;\theta_i)$.

Rozpatrzmy konkretny przykład liczbowy w przypadku mieszanki dwóch rozkładów dwumianowych. Funkcja rozkładu prawdopodobieństwa jest również następująca:

$$f(k;\alpha,\theta_1,\theta_2) = \alpha C_m^k \theta_1^k (1-\theta_1)^{m-k} + (1-\alpha) C_m^k \theta_2^k (1-\theta_2)^{m-k}$$

Funkcja wiarygodności próby

$$k_1, k_2, ..., k_n$$

jest następująca:

$$L(k_1 k_2, ..., k_n; \alpha, \theta_1, \theta_2) = \prod_{j=1}^{n} f(k_i, \alpha, \theta_1, \theta_2).$$

Estymator parametru α określany jest następująco:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(1 \mid k_i)$$

gdzie $\hat{P}\left(1\middle|k_j\right)$ oznacza estymator prawdopodobieństwa tego, że obserwacja k_i pochodzi z pierwszej klasy. Zgodnie z twierdzeniem Bayesa mamy:

$$\hat{P}\left(1\middle|k_i\right) = \frac{\hat{\alpha} C_m^{k_i} \hat{\theta}_1^{k_i} \left(1-\hat{\theta}_1\right)^{m-k_i}}{\hat{\alpha} C_m^{k_i} \hat{\theta}_1^{k_i} \left(1-\hat{\theta}_1\right)^{m-k_i} + \left(1-\hat{\alpha}\right) C_m^{k_i} \hat{\theta}_2^{k_i} \left(1-\hat{\theta}_2\right)^{m-k_i}}$$

Estymatory $\hat{\theta}_1$ i $\hat{\theta}_2$ określane są następująco:

$$\sum_{j=1}^{n} P(1 \mid k_j) \cdot \frac{\partial \log C_m^{k_j} \hat{\theta}_1^{k_j} (1-\hat{\theta}_1)^{m-k_j}}{\partial \hat{\theta}_1} = 0$$

$$\sum_{j=1}^{n} P(2 \mid k_j) \cdot \frac{\partial \log C_m^{k_j} \hat{\theta}_2^{k_j} (1-\hat{\theta}_2)^{m-k_j}}{\partial \hat{\theta}_2} = 0$$

Zauważmy, że

$$\frac{\partial \log C_m^k \theta^k (1-\theta)^{m-k}}{\partial \theta} = \frac{\partial \log C_m^k + k \log \theta + (m-k)\log(1-\theta)}{\partial \theta} =$$
$$= k\frac{1}{\theta} - \frac{m-k}{1-\theta} = \frac{k(1-\theta)-(m-k)\theta}{\theta(1-\theta)} = \frac{k-m\theta}{\theta(1-\theta)}$$

Stąd też równania powyższe przyjmują następującą postać:

$$\sum_{j=1}^{n} \left[P(1 \mid k_j) \frac{k_j - m\hat{\theta}_1}{\hat{\theta}_1(1-\hat{\theta}_1)} \right] = 0$$

$$\sum_{j=1}^{n} \left[P(2 \mid k_j) \frac{k_j - m\hat{\theta}_2}{\hat{\theta}_2(1-\hat{\theta}_2)} \right] = 0$$

Pomnóżmy pierwsze równanie przez $\hat{\theta}_1\left(1-\hat{\theta}_1\right)$, a drugie przez $\hat{\theta}_2\left(1-\hat{\theta}_2\right)$ i zapiszmy te równania w następującej postaci:

$$\sum_{j=1}^{n} P\left(1\middle|k_j\right) k_j = m\hat{\theta}_1 \sum_{j=1}^{n} P\left(1\middle|k_j\right)$$

$$\sum_{j=1}^{n} P\left(2\middle|k_j\right)k_j = m\hat{\theta}_2 \sum_{j=1}^{n} P\left(2\middle|k_j\right)$$

Rozwiązanie tego układu równań uzyskujemy bardzo łatwo, a mianowicie dzieląc stronami oba równania przez $m \cdot n$, ostatecznie otrzymujemy:

$$\hat{\theta}_1 = \frac{1}{m}\frac{\sum_{j=1}^{n} P\left(1\middle|k_j\right)k_j}{\sum_{j=1}^{n} P\left(1\middle|k_j\right)} = \frac{1}{m}\bar{k}_1$$

$$\hat{\theta}_2 = \frac{1}{m}\frac{\sum_{j=1}^{n} P\left(2\middle|k_j\right)k_j}{\sum_{j=1}^{n} P\left(2\middle|k_j\right)} = \frac{1}{m}\bar{k}_2$$

gdzie średnie $\bar{k}_1$ są średnimi ważonymi liczby sukcesów dla pierwszego składnika mieszanki, a $\bar{k}_2$ dla drugiego składnika. Bezpośrednie korzystanie z tych wzorów w celu obliczenia wartości estymatorów $\hat{\alpha}, \hat{\theta}_1$ i $\hat{\theta}_2$ jest niemożliwe, gdyż do obliczenia na przykład $\hat{\alpha}$ potrzebna jest znajomość wartości $\hat{P}\left(1\middle|k_j\right)$, z kolei do obliczania $\hat{P}\left(1\middle|k_j\right)$ potrzebna jest znajomość $\hat{\alpha}$. Trudność obliczeniową omija się, stosując iteracyjnie kolejne przybliżenia potrzebnych wartości.

Rozpatrzmy sposób takiego postępowania na fragmentarycznym przykładzie. Załóżmy, że rozkład zmiennej losowej K jest mieszanką dwóch rozkładów dwumianowych:

$$f(k;\alpha_1,\alpha_2,\theta_1,\theta_2) = \alpha_1 C_m^k \theta_1^k (1-\theta_1)^{m-k} + \alpha_2 C_m^k \theta_2^k (1-\theta_2)^{m-k}$$

przy czym

$$m = 20$$

$$\alpha_1 = \frac{1}{4},\ \theta_1 = \frac{4}{5},\ \theta_2 = \frac{3}{5}$$

Załóżmy, że parametry rozkładu nie są znane, znamy zaś następujące obserwacje:

2, 3, 8, 12, 15, 1, 9, 3, 19.

W celu obliczenia wartości estymatorów parametrów $\alpha_1, \alpha_2, \theta_1$ i θ_2, przyjmijmy następujące wartości początkowe:

$$\alpha_1^{(0)} = 0{,}2\,,\ \theta_1^{(0)} = 0{,}5\,,\ \theta_2^{(0)} = 0{,}6$$

Na ich podstawie obliczymy wartości

$$\alpha_1^{(1)},\ \theta_1^{(1)},\ \theta_2^{(1)}$$

w sposób następujący:

$$\hat{P}^{(1)}\left(1\middle|k_1\right) = \frac{0{,}2C_{20}^2 0{,}5^2 0{,}5^{18}}{0{,}2C_{20}^2 0{,}5^2 0{,}5^{18} + 0{,}8C_{20}^2 0{,}6^2 0{,}4^{18}} = \frac{0{,}000036}{0{,}00004} = 0{,}906$$

$$\hat{P}^{(1)}\left(1\middle|k_2\right) = \frac{0{,}2C_{20}^3 0{,}5^3 0{,}5^{17}}{0{,}2C_{20}^3 0{,}5^3 0{,}5^{17} + 0{,}8C_{20}^3 0{,}6^3 0{,}4^{17}} = \frac{0{,}00022}{0{,}00025} = 0{,}8653$$

$$\hat{P}^{(1)}\left(1\middle|k_3\right) = \frac{0{,}2C_{20}^8 0{,}5^8 0{,}5^{12}}{0{,}2C_{20}^8 0{,}5^8 0{,}5^{12} + 0{,}8C_{20}^8 0{,}6^8 0{,}4^{12}} = \frac{0{,}024}{0{,}0524} = 0{,}4532$$

Pozostałe wartości $\hat{P}\left(1\middle|k_j\right)$ są równe odpowiednio: 0,1432, 0,0472, 0,9353, 0,3606, 0,8653 i 0,0097.

Po obliczeniu tych wartości, obliczamy wartości $\hat{\alpha}^{(1)}$ i $\hat{\alpha}^{(2)}$:

$$\hat{\alpha}_1^{(1)} = \frac{1}{9}(\hat{P}^{(1)}(1\,|\,k_1) + \ldots + \hat{P}^{(1)}(1\,|\,k_9) = \frac{4{,}5909}{9} = 0{,}5101$$

oraz wartości $\hat{\theta}_1^{(1)}$ i $\hat{\theta}_2^1$

$$\hat{\theta}_1^{(1)} = \frac{1}{20}\frac{2\hat{P}^{(1)}\left(1\middle|k_1\right) + \ldots + 19\hat{P}^{(1)}\left(1\middle|k_9\right)}{\hat{P}^{(1)}\left(1\middle|k_1\right) + \ldots + \hat{P}^{(1)}\left(1\middle|k_9\right)} = \frac{17{,}4615}{20 \cdot 4{,}5909} = 0{,}1902$$

$$\hat{\theta}_2^{(1)} = \frac{1}{20}\frac{2\hat{P}^{(1)}\left(2\middle|k_1\right) + \ldots + 19\hat{P}^{(1)}\left(2\middle|k_9\right)}{\hat{P}^{(1)}\left(2\middle|k_1\right) + \ldots + \hat{P}^{(1)}\left(2\middle|k_9\right)} = \frac{1}{20}\frac{2\hat{P}^{(1)}\left(2\middle|k_1\right) + \ldots + 19\hat{P}^{(1)}\left(2\middle|k_9\right)}{9 - \hat{P}^{(1)}\left(1\middle|k_1\right) - \ldots - \hat{P}^{(1)}\left(1\middle|k_9\right)}$$

gdzie $P(2\,|\,k_j) = 1 - P(1\,|\,k_j)$. Stąd $\hat{\theta}_2^{(1)} = \frac{54{,}5385}{20 \cdot 4{,}4091} = 0{,}6185$.

Obliczone wartości $\hat{\alpha}_1^{(1)}$, $\hat{\alpha}_2^{(1)}$, $\hat{\theta}_1^{(1)}$, i $\hat{\theta}_2^{(1)}$ wykorzystujemy do obliczenia wartości $\hat{\alpha}_1^{(2)}$, $\hat{\alpha}_2^{(2)}$, $\hat{\theta}_1^{(2)}$, $\hat{\theta}_2^{(2)}$.

Procedurę powtarzamy tak długo aż kolejne przybliżenia $\left|P^{(r)}\left(i\middle|k_j\right)-P^{(r-1)}\left(i\middle|k_j\right)\right|$ będą się różniły nie więcej niż zadana wcześniej mała liczba $\varepsilon > 0$, tzn.

$$\max_{i,j}\left|P^{(r)}\left(i\middle|k_j\right)-P^{(r-1)}\left(i\middle|k_j\right)\right| < \varepsilon ,$$

gdzie *r* oznacza numer iteracji obliczeń.

3. Idea metody momentów

Istota tej metody polega na tym, że trzeba określić pewną funkcję „teoretyczną", której argumentami są oceniane parametry. Oznaczmy ją jako $\mathbf{\Phi}(\eta)$. Oprócz tego trzeba określić jej odpowiednik empiryczny, którego argumentami są zaobserwowane wartości zmiennej losowej. Oznaczmy tę statystykę jako $T(X_1, X_2, ..., X_n)$. Estymacji parametru η dokonujemy poprzez „rozwiązanie" następującego równania:

$$\mathbf{\Phi}(\eta) = T(X_1, X_2, ..., X_n)$$

względem niewiadomej η.
Symbolicznie zapiszmy więc w postaci:

$$\hat{\eta} = \mathbf{\Phi}^{-1}\left(T(X_1, X_2, ..., X_n)\right)$$

Zamiast rozwiązywania równania, można minimalizować pewną funkcję błędu:

$$\mathbf{\Delta} = \mathbf{\Delta}\left(\mathbf{\Phi}(\eta) - T(X_1, X_2, ..., X_n)\right)$$

Jeden z prostszych sposobów takiego postępowania polega na tym, że funkcje $T(X_1, X_2, ..., X_n)$ definiowane są jako momenty empiryczne, w ten sposób aby

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = \mathrm{E}\left(X^k\right).$$

Przedstawiony wyżej ogólny schemat postępowania rozpatrzmy na konkretnym przykładzie mieszanki dwóch rozkładów dwumianowych.

W przypadku zmiennych losowych typu dyskretnego, szczególnie tych o charakterze dwumianowym, zamiast momentów „addytywnych" postaci

$$\mu_k = \mathrm{E}\left(X^k\right) = \sum x_i^k p_i$$

wygodniejsze są tzw. momenty multiplikatywne.
Momentem multiplikatywnym rzędu *k* nazywa się wyrażenie:

$$\mu_k = \sum_i x_i^{[k]} p_i \,, \quad k = 1, 2, \ldots$$

gdzie

$$x_i^{[k]} = x_i (x_i - 1) \ldots (x_i - k + 1).$$

Trzy kolejne momenty tego typu względem zera są więc następujące:

$$\mu_1 = \sum x_i p_i$$

$$\mu_2 = \sum x_i (x_i - 1) p_i$$

$$\mu_3 = \sum x_i (x_i - 1)(x_i - 2) p_i$$

Na podstawie próby losowej $(X_1, X_2, \ldots, X_n)$ momenty empiryczne określa się następująco:

$$M_k = \frac{1}{n} \sum_i X_i^{[k]}.$$

Rozpatrzmy mieszankę dwóch rozkładów dwumianowych

$$f(x, \eta) = \alpha_1 \mathrm{B}(x; m, \theta_1) + \alpha_2 \mathrm{B}(x; m, \theta_2)$$

gdzie

$$\mathrm{B}(x; m, \theta) = C_m^x \theta^x (1 - \theta)^{m-x} \,, \quad x = 0, 1, \ldots, m$$

Dokonamy estymacji parametrów $\eta = (\alpha_1, \alpha_2, \theta_1, \theta_2)$ metodą momentów, korzystając z próby prostej $(X_1, X_2, \ldots, X_n)$.
Określmy trzy pierwsze momenty empiryczne:

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^{[k]} \,, \quad k = 1, 2, 3$$

Można sprawdzić, że

$$\mu_1 = m(\alpha_1 \theta_1 + \alpha_2 \theta_2)$$

$$\mu_2 = m(m-1)(\alpha_1 \theta_1^2 + \alpha_2 \theta_2^2)$$

$$\mu_3 = m(m-1)(m-2)(\alpha_1 \theta_1^3 + \alpha_2 \theta_2^3)$$

Porównując momenty empiryczne z teoretycznymi, otrzymujemy następujący układ równań:

$$\frac{1}{n\,m}\sum_{i=1}^{n} x_i = \alpha_1\theta_1 + \alpha_2\theta_2$$

$$\frac{1}{n\,m(m-1)}\sum_{i=1}^{n} x_i(x_i-1) = \alpha_1\theta_1^2 + \alpha_2\theta_2^2$$

$$\frac{1}{n\,m(m-1)(m-2)}\sum_{i=1}^{n} x_i(x_i-1)(x_i-2) = \alpha_1\theta_1^3 + \alpha_2\theta_2^3$$

Ten układ trzech równań trzeba rozwiązać względem trzech niewiadomych α_1, θ_1 i θ_2 gdyż $\alpha_2 = 1-\alpha_1$.

W celu uproszczenia zapisów przyjmijmy, że lewe strony oznaczone będą symbolami l_1, l_2 i l_3, zaś zamiast α_1 stosowany będzie symbol α i w konsekwencji $\alpha_2 = 1 - \alpha$. Powyższy układ równań przyjmuje następującą postać:

$$l_1 = \alpha\theta_1 + (1-\alpha)\theta_2$$

$$l_2 = \alpha\theta_1^2 + (1-\alpha)\theta_2^2$$

$$l_3 = \alpha\theta_1^3 + (1-\alpha)\theta_2^3$$

Z pierwszego równania wyznaczymy α:

$$\alpha = \frac{l_1 - \theta_2}{\theta_1 - \theta_2}$$

Podstawmy to wyrażenie do drugiego równania:

$$l_2 = \alpha\theta_1^2 + \theta_2^2 - \alpha\theta_2^2 = \alpha\left(\theta_1^2 - \theta_2^2\right) + \theta_2^2 = \frac{l_1 - \theta_2}{\theta_1 - \theta_2}\left(\theta_1^2 - \theta_2^2\right) + \theta_2^2 = (l_1 - \theta_2)(\theta_1 + \theta_2) + \theta_2^2$$

Uzyskamy stąd równanie:

$$(l_1 - \theta_2)(\theta_1 + \theta_2) + \theta_2^2 - l_2 = 0,$$

w którym są dwie niewiadome θ_1 i θ_2.

W celu wyeliminowania jednej z nich obliczmy różnicę

$$l_3 - l_1 l_2$$

i podzielmy ją przez $l_2 - l_1^2$.

$$l_3 - l_1 l_2 = \alpha(1-\alpha)(\theta_1 + \theta_2)(\theta_1 - \theta_2)^2$$

$$l_2 - l_1^2 = \alpha(1-\alpha)(\theta_1 - \theta_2)^2$$

po podzieleniu uzyskujemy

$$\frac{l_3 - l_1 l_2}{l_2 - l_1^2} = \theta_1 + \theta_2$$

Po podstawieniu do uzyskanego równania otrzymujemy:

$$(l_1 - \theta_2)\frac{l_3 - l_1 l_2}{l_2 - l_1^2} + \theta_2^2 - l_2 = 0$$

Wprowadźmy oznaczenie

$$B = \frac{l_3 - l_1 l_2}{l_2 - l_1^2}$$

Powyższe równanie przyjmie wówczas postać:

$$(l_1 - \theta_2)B + \theta_2^2 - l_2 = 0\,.$$

Zapiszmy je następująco:

$$\theta_2^2 - B\theta_2 + C = 0$$

gdzie

$$C = l_1 B - l_2$$

Uzyskaliśmy równie drugiego stopnia. Ma ono dwa rozwiązania, które traktujemy jako estymatory $\hat{\theta}_1, \hat{\theta}_2$, gdy $B^2 - 4C > 0$.
Czyli

$$\hat{\theta}_1 = \frac{B - \sqrt{B^2 - 4C}}{2}$$

$$\hat{\theta}_2 = \frac{B + \sqrt{B^2 - 4C}}{2}$$

Na ich podstawie uzyskujemy estymator $\hat{\alpha}$

$$\hat{\alpha} = \frac{l_1 - \hat{\theta}_2}{\hat{\theta}_1 - \hat{\theta}_2}\,.$$

Rozpatrując poprzedni przykład, uzyskujemy następujące wyniki:

$$l_1 = \frac{1}{nm}\sum_{i=1}^{n} k_i = \frac{1}{9\cdot 20}(2+3+8+12+15+1+9+3+19) = 0{,}4$$

$$l_2 = \frac{1}{nm(m-1)}\sum_{i=1}^{n} k_i(k_i-1) =$$

$$= \frac{1}{9\cdot 20\cdot 19}(2\cdot 1+3\cdot 2+8\cdot 7+12\cdot 11+15\cdot 14+1\cdot 0+9\cdot 8+3\cdot 2+19\cdot 18) = 0{,}2415$$

$$l_3 = \frac{1}{nm(m-1)(m-2)}\sum_{i=1}^{n} k_i(k_i-1)(k_i-2) =$$

$$= \frac{1}{9\cdot 20\cdot 19\cdot 18}(2\cdot 1\cdot 0+3\cdot 2\cdot 1+8\cdot 7\cdot 6+12\cdot 11\cdot 10+$$
$$+15\cdot 14\cdot 13+1\cdot 0(1)+9\cdot 8\cdot 7+3\cdot 2\cdot 1+19\cdot 18\cdot 17) =$$
$$= \frac{10716}{180\cdot 19\cdot 18} = 0{,}1741$$

Stąd obliczamy

$$B = \frac{l_3 - l_2 l_1}{l_2 - l_1^2} = 0{,}9503$$

$$C = l_1 \cdot B - l_2 = 0{,}1386$$

Na ich podstawie obliczamy potrzebne parametry:

$$\hat{\theta}_1 = \frac{B - \sqrt{B^2 - 4C}}{2} = 0{,}1799$$

$$\hat{\theta}_2 = \frac{B + \sqrt{B^2 - 4C}}{2} = 0{,}7704\,.$$

Rozkłady złożone

Ogólnie mieszankę rozkładów można symbolicznie zapisać jako $X \sim \sum_{i=1}^{r} \alpha_i X_i$, gdzie znak tyldy oznacza rozkład, a symbol po prawej stronie nie oznacza sumy zmiennych losowych, tylko rozkład tej sumy przy założeniu, że

zmienne losowe X_i mają różne rozkłady. Rozkład tej mieszanki w przypadku dyskretnym ma postać

$$P(X = x) = \sum_{i=1}^{r} \alpha_i P(X_i = x) \tag{1}$$

lub gdy X jest zmienną losową ciągłą

$$f(x) = \sum_{i=1}^{r} \alpha_i f_i(x), \tag{2}$$

przy czym

$$\alpha_i > 0 \text{ dla } i = \overline{1,r}, \text{ oraz } \sum_{i=1}^{r} \alpha_i = 1$$

Symbol $i = \overline{1,r}$ oznacza to samo, co $i = 1, 2, \ldots, r$.

Przyglądając się wagom w powyższych wzorach, nietrudno zauważyć, że mogą być one interpretowane jako wartości prawdopodobieństw z pewnego rozkładu. Zdanie to na razie nic nie znaczy, póki nie powiemy, czego te prawdopodobieństwa mają dotyczyć.

Załóżmy, że gęstości $f(x)$ oraz prawdopodobieństwa $P(X = x)$ zależą od pewnego parametru θ, którego wartość jest nieokreślona. Zapiszemy tę zależność jako $f(x|\theta)$ oraz $P(X = x|\Theta = \theta)$. Jako miarę tej nieokreśloności możemy przyjąć pewien rozkład o gęstości $\pi(\theta)$ lub $P(\Theta = \theta_j) = p_j$, dla $j = \overline{1,k}$. Wówczas wzory (1) i (2) można zapisać jako:

$$\left.\begin{aligned} P(X = x) &= \sum_{j\geq 0} P(\Theta = \theta_j) P(X = x|\Theta = \theta_j) \\ f(x) &= \sum_{j\geq 0} P(\Theta = \theta_j) f(x|\theta_j) \end{aligned}\right\} \text{ dla } \theta \text{ dyskretnego,} \tag{3}$$

$$\left.\begin{aligned} P(X = x) &= \int_{\Omega} \pi(\theta) P(X = x|\Theta = \theta) d\theta \\ f(x) &= \int_{\Omega} \pi(\theta) f(x|\theta) d\theta \end{aligned}\right\} \text{ dla } \theta \text{ ciągłego,}$$

gdzie Ω jest przestrzenią parametrów. Uważny czytelnik zwróci uwagę, że po prawej stronie wzorów (3) mamy wartości oczekiwane pewnych funkcji, który-

mi są prawdopodobieństwo lub gęstość. Powyższe wzory można więc zapisać w równoważnej postaci

$$P(X = x) = E_{\Theta}\left(P(X = x|\theta)\right),$$
$$f(x) = E_{\Theta}\left(f(x|\theta)\right).$$

Dalszym uogólnieniem tych wzorów może być mieszanka rozkładów ze względu na pewien podzbiór parametrów w przypadku, gdy parametr $\boldsymbol{\theta}$ jest wektorem parametrów jak np. w rozkładzie normalnym $\boldsymbol{\theta} = (\mu, \sigma)$. Mieszanki rozkładów tego typu często są nazywane rozkładami złożonymi i oznaczane są specjalnym symbolem

$$f(x|\Theta = \theta) \underset{\theta}{\wedge} \pi(\theta)$$

$$P(X = x|\Theta = \theta) \underset{\theta}{\wedge} \pi(\theta)$$

który odczytujemy w następujący sposób: rozkład złożony jest złożeniem dwóch rozkładów (może ich być więcej) – rozkładu zmiennej losowej X, który jest określony z dokładnością do nieznanego parametru θ, którego rozkład $\pi(\theta)$ jest znany.
Ogólniejszy zapis ma postać

$$F_A \underset{\Theta}{\wedge} F_B,$$

gdzie po lewej stronie mamy rozkład X ze zmiennym parametrem Θ, a po prawej stronie rozkład zmiennej Θ.

Przykład 1.

Rozważmy rozkład typu:

$$\mathrm{B}(N, p) \underset{N}{\wedge} \mathrm{P}(\lambda).$$

Odczytując ten symbol, widzimy, że jest to mieszanka rozkładu dwumianowego o nieznanej liczbie doświadczeń x_1, x_2, …, x_N wykonanych wg schematu Bernoulliego. Niepewność co do wartości N określamy za pomocą rozkładu Poissona z parametrem λ. Oczywiście parametry p i λ nie muszą być znane i wtedy są estymowane na podstawie wyników próby losowej. Zauważmy, że skoro w próbie o nieokreślonej liczbie N obserwacji odnotowano x sukcesów, to liczba obserwacji musi być co najmniej x.

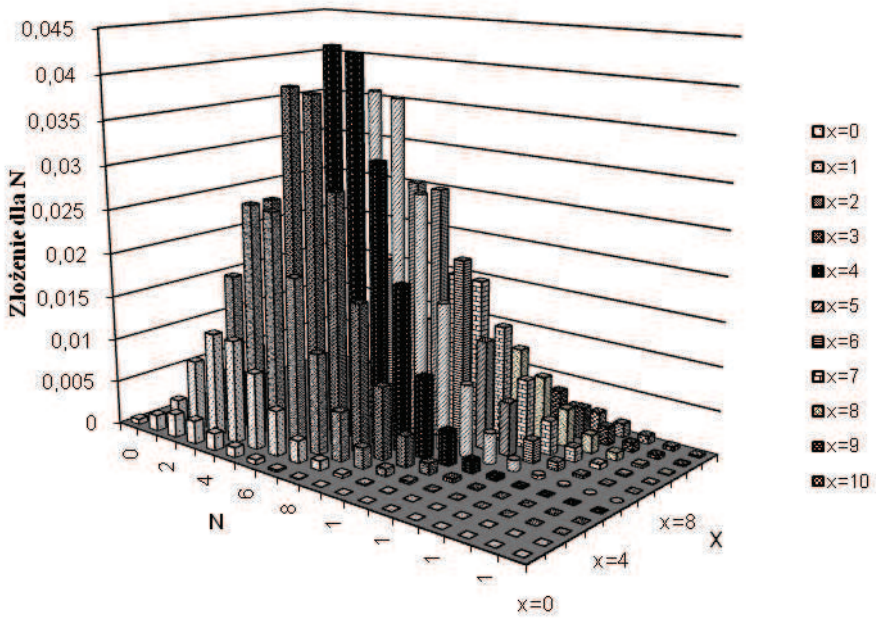
$$P(X=x)=\mathrm{E}_N\left(P\left(X=x\middle|N=n\right)\right)=\sum_{n=x}^{\infty}\left(\underbrace{\binom{n}{x}p^x(1-p)^{n-x}}_{\substack{\text{Prawdopodobieństwo}\\ X=x \text{ z rozkładu } \mathrm{B}(n,p)}}\underbrace{\frac{\lambda^n}{n!}e^{-\lambda}}_{\substack{\text{Prawdopodobieństwo}\\ N=n \text{ z rozkładu } P(\lambda)}}\right)=$$

$$=\frac{e^{-\lambda}}{x!}\sum_{n=x}^{\infty}\frac{(\lambda p)^x\left(\lambda(1-p)\right)^{n-x}}{(n-x)!}=\frac{e^{-\lambda}}{x!}(\lambda p)^x\sum_{n=x}^{\infty}\frac{\left(\lambda(1-p)\right)^{n-x}}{(n-x)!}=$$

$$\overset{k=n-x}{=}\frac{e^{-\lambda}}{x!}(\lambda p)^x\sum_{k=0}^{\infty}\frac{\left(\lambda(1-p)\right)^k}{k!}=\frac{e^{-\lambda}}{x!}(\lambda p)^x e^{\lambda(1-p)}=\frac{(\lambda p)^x}{x!}e^{-\lambda p}$$

Jak widać, mieszanka ma rozkład Poissona o parametrze λp, co symbolicznie można zapisać

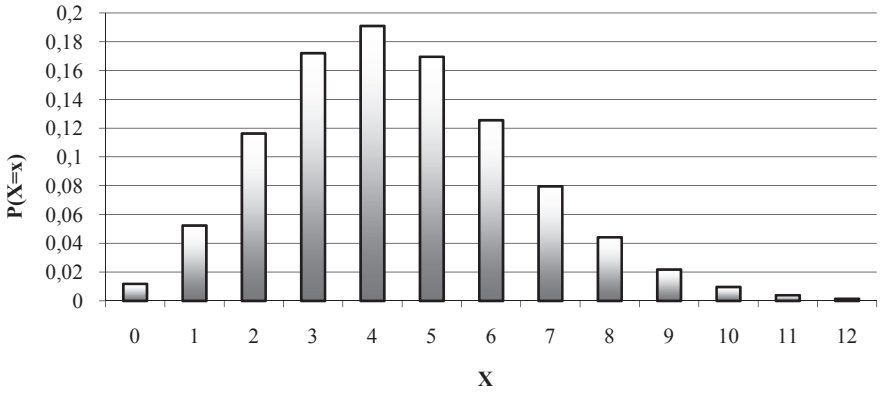
$$\mathrm{P}(\lambda p)=\mathrm{B}(N,p)\underset{N}{\wedge}\mathrm{P}(\lambda)$$



Rysunek 1. Składowe rozkładów B(N, 0,6) i P(7,4).

Opracowanie własne

Rysunek 1 przedstawia wykres kolumnowy poszczególnych składników sumy $\binom{n}{x}0{,}6^x 0{,}4^{n-x}\frac{7{,}4^n}{n!}e^{-7{,}4} = \frac{1{,}5^x 2{,}96^n}{x!(n-x)!}e^{-7{,}4}$, za pomocą których wyznaczany jest rozkład $\mathrm{P}(X = x)$ mieszanki $\mathrm{B}(N,0{,}6)\underset{N}{\wedge}\mathrm{P}(7{,}4)$. Dokonujemy tego, sumując dla wybranego x na osi X długości wszystkich słupków histogramu wzdłuż osi N. Ostatecznie otrzymujemy w tym przykładzie dla wybranych parametrów rozkład Poissona o parametrze równym 4,44.



Rysunek 2. Rozkład $\mathrm{P}(4{,}44) = \mathrm{B}(N;0{,}6)\underset{N}{\wedge}\mathrm{P}(7{,}4)$

Opracowanie własne.

Przykład 2.

Zauważmy, że jeżeli mamy praktycznie pewność, że nieznany parametr przyjmuje wartość θ_0, rozkład złożony sprowadza się do zwykłego rozkładu. Niech

$$P(\Theta = \theta_0) = 1$$

to

$$P(X = x) = \mathrm{E}_\Theta\left(P(X = x|\theta)\right) = \sum_{j\geq 0} P(X = x|\theta_j)P(\Theta = \theta_j) =$$
$$= P(X = x|\theta_0),$$
$$f(x) = \mathrm{E}_\Theta\left(f(x|\theta)\right) = f(x|\theta_0).$$

Przykład 3.

Znaleźć rozkład będący złożeniem rozkładu dwumianowego o nieznanym parametrze P, którego rozkład jest rozkładem beta o parametrach a i b.

$$\mathrm{B}(N,P)\underset{P}{\wedge}Beta(a,b)$$

$$f(x)=\int_0^1 \underbrace{\binom{n}{x}p^x(1-p)^{n-x}}_{\text{rozkład dwumianowy}}\underbrace{\frac{p^{a-1}(1-p)^{b-1}}{Beta[a,b]}}_{\text{rozkład beta}}dp=\frac{\binom{n}{x}}{Beta[a,b]}\underbrace{\int_0^1 p^{x+a-1}(1-p)^{n-x+b-1}dp}_{Beta(x+a,n-x+b)}=$$

$$=\frac{\binom{n}{x}Beta[x+a,n-x+b]}{Beta[a,b]},$$

gdzie $Beta[a,b]=\dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Jest to rozkład beta-dwumianowy, który oznacza się symbolem BB(n, a, b).

Przykład 4.

Znaleźć rozkład będący złożeniem dwóch rozkładów dwumianowych

$$\mathrm{B}(N,p)\underset{N}{\wedge}\mathrm{B}(m,q)$$

Wtedy

$$P(X=x)=\sum_{n=x}^{m}\binom{n}{x}p^x(1-p)^{n-x}\binom{m}{n}q^n(1-q)^{m-n}=$$

$$=\frac{m!}{x!}\left(\frac{p}{1-p}\right)^x(1-q)^m\sum_{n=x}^{m}\frac{1}{(n-x)!(m-n)!}\left(\frac{(1-p)q}{1-q}\right)^n=$$

$$=\frac{m!}{x!(m-x)!}\left(\frac{pq}{1-q}\right)^x(1-q)^m\sum_{n=x}^{m}\frac{(m-x)!}{(n-x)!(m-n)!}\left(\frac{(1-p)q}{1-q}\right)^{n-x}=$$

$$=\binom{m}{x}\left(\frac{pq}{1-q}\right)^x(1-q)^m\left(\frac{(1-p)q}{1-q}+1\right)^{m-x}=\binom{m}{x}(pq)^x(1-pq)^{m-x}$$

W wyniku otrzymuje się rozkład dwumianowy B(m, pq).

Przykład 5.

Niech będzie dana gęstość rozkładu gamma G(α, β):

$$f(x)=\frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)},\quad \alpha>0,\beta>0,x>0.$$

Znaleźć mieszankę rozkładu Poissona P(λ) z powyższym rozkładem:

$$\mathrm{P}(\Lambda)\underset{\Lambda}{\wedge}\mathrm{G}(\alpha,\beta).$$

W wyniku obliczeń otrzymuje się

$$P(X=x)=\int_0^{\infty}\frac{\lambda^x}{x!}e^{-\lambda}\frac{\lambda^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}e^{-\frac{\lambda}{\beta}}d\lambda=\frac{1}{x!\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty}\lambda^{x+\alpha-1}e^{-\lambda\left(\frac{\beta+1}{\beta}\right)}d\lambda=$$

$$=\frac{\left(\frac{\beta}{\beta+1}\right)^{x+\alpha}\Gamma(x+\alpha)}{x!\beta^{\alpha}\Gamma(\alpha)}\underbrace{\frac{1}{\left(\frac{\beta}{\beta+1}\right)^{x+\alpha}\Gamma(x+\alpha)}\int_0^{\infty}\lambda^{x+\alpha-1}e^{-\lambda\left(\frac{\beta+1}{\beta}\right)}d\lambda}_{=1}=\frac{\beta^x(x+\alpha-1)\cdots\alpha}{x!(\beta+1)^{x+\alpha}}$$

$$=\frac{(x+\alpha-1)!}{(\alpha-1)!x!}\left(\frac{\beta}{\beta+1}\right)^x\left(\frac{1}{\beta+1}\right)^{\alpha}=\binom{x+\alpha-1}{\alpha-1}\left(\frac{\beta}{\beta+1}\right)^x\left(\frac{1}{\beta+1}\right)^{\alpha}$$

Złożenie jest rozkładem ujemnym dwumianowym oznaczanym symbolicznie przez NB(α, β).

Przykład 6.

Znajdźmy rozkład $\mathrm{N}\left(0,\frac{1}{\sqrt{\lambda}}\right)\underset{\lambda}{\wedge}\mathrm{G}(\alpha,\beta)$, gdzie $\lambda > 0$. Na podstawie wzorów (3) dla obu rozkładów ciągłych mamy

$$f(x)=\int_{\Omega}\pi(\lambda)f(x|\lambda)d\lambda=\int_0^{\infty}\frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}\sqrt{\lambda}}{\beta^{\alpha}\Gamma(\alpha)\sqrt{2\pi}}e^{-\frac{\lambda x^2}{2}}d\lambda=\frac{1}{\sqrt{2\pi}\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty}\lambda^{\alpha-\frac{1}{2}}e^{-\lambda\left(\frac{x^2}{2}+\frac{1}{\beta}\right)}d\lambda=$$

$$=\frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\sqrt{2\pi}\beta^{\alpha}\Gamma(\alpha)}\left(\frac{1}{\beta}+\frac{x^2}{2}\right)^{-\left(\alpha+\frac{1}{2}\right)}$$

Jest to rozkład Studenta o parametrach α i β.

Interesująca jest interpretacja mieszanki rozkładów poprzez twierdzenie Bayesa. Z (3) wynika, że

$$\left.\begin{aligned}&\sum_{j\geq 0}\frac{P(\Theta=\theta_j)P(X=x|\Theta=\theta_j)}{P(X=x)}=1\\&\sum_{j\geq 0}\frac{P(\Theta=\theta_j)f(x|\theta_j)}{f(x)}=1\end{aligned}\right\}\text{ dla }\theta\text{ dyskretnego}$$

$$\left.\begin{aligned}&\int_{\Omega}\frac{\pi(\theta)P(X=x|\Theta=\theta)}{P(X=x)}d\theta=1\\&\int_{\Omega}\frac{\pi(\theta)f(x|\theta)}{f(x)}d\theta=1\end{aligned}\right\}\text{ dla }\theta\text{ ciągłego}$$

Wstawiając za mianowniki wyrażenia z (3), otrzymujemy pod znakami sum lub całek wzory analogiczne do wzorów Bayesa, skąd wynika, że funkcje podcałkowe są rozkładami aposteriori dla rozkładów apriori $\pi(\theta)$ dla przypadku ciągłego i $P(\Theta = \theta)$ w przypadku dyskretnym

$$\left.\begin{aligned}&P(\Theta=\theta_j|X=x)=\frac{P(\Theta=\theta_j)P(X=x|\Theta=\theta_j)}{\sum_{j\geq 0}P(\Theta=\theta_j)P(X=x|\Theta=\theta_j)}\\&P(\Theta=\theta_j|x)=\frac{P(\Theta=\theta_j)f(x|\theta_j)}{\sum_{j\geq 0}P(\Theta=\theta_j)f(x|\theta_j)}\end{aligned}\right\}\text{ dla }\theta\text{ dyskretnego,}$$

$$\left.\begin{aligned}&\pi(\theta|x)=\frac{\pi(\theta)P(X=x|\Theta=\theta)}{\int_{\Omega}\pi(\theta)P(X=x|\Theta=\theta)d\theta}\\&\pi(\theta|x)=\frac{\pi(\theta)f(x|\theta)}{\int_{\Omega}\pi(\theta)f(x|\theta)d\theta}\end{aligned}\right\}\text{ dla }\theta\text{ ciągłego.}$$

Stąd

$$\left.\begin{aligned}&P(X=x)=\sum_{j\geq 0}P(\Theta=\theta_j|X=x)\\&f(x)=\sum_{j\geq 0}P(\Theta=\theta_j|x)\end{aligned}\right\}\text{ dla }\theta\text{ dyskretnego,}$$

$$\left.\begin{aligned} P(X=x) &= \int_{\Omega} \pi(\theta|x)d\theta \\ f(x) &= \int_{\Omega} \pi(\theta|x)d\theta \end{aligned}\right\} \text{ dla } \theta \text{ ciągłego.}$$

Na zakończenie przytoczone zostaną twierdzenia mające zastosowanie w analizie rozkładów złożonych. Dla mieszanek rozkładów o niezależnych rozkładach F_1, F_2 i F_3 zachodzą :

1. $(F_1 \wedge F_2) \wedge F_3 \sim F_1 \wedge (F_2 \wedge F_3)$
2. $\mathrm{E}(g(X)) = \mathrm{E}_{\theta}\left(\mathrm{E}_{x|\theta}(g(X))\right)$
3. $\mathrm{V}(g(X)) = \mathrm{V}_{\theta}\left(\mathrm{E}_{x|\theta}(g(X))\right) + \mathrm{E}_{\theta}\left(\mathrm{V}_{x|\theta}(g(X))\right)$.

D o w ó d 1 dla rozkładów ciągłych:

$X \sim F_1(x|\theta_1)$, $\theta_1 \subset \Omega_1$, $Y \sim F_2(y|\theta_2)$, $\theta_2 \subset \Omega_2$, $Z \sim F_3(x|\theta_3)$. Wtedy lewa strona jest równa

$$\left(F_1(x|\Theta_1) \underset{\Theta_1}{\wedge} F_2(\theta_1|\Theta_2)\right) \underset{\Theta_2}{\wedge} F_3(\theta_2) = \int_{\Omega_2} \left(F_1(x|\Theta_1) \underset{\Theta_1}{\wedge} F_2(\theta_1|\Theta_2)\right) f_3(\theta_2)d\theta_2 =$$
$$= \int_{\Omega_2}\int_{\Omega_1} f_1(x|\theta_1) f_2(\theta_1|\theta_2) f_3(\theta_2)d\theta_1 d\theta_2,$$

a prawa

$$F_1(x|\Theta_1) \underset{\Theta_1}{\wedge} \left(F_2(\theta_1|\Theta_2) \underset{\Theta_2}{\wedge} F_3(\theta_2)\right) = \int_{\Omega_1} f_1(x|\theta_1) \left(F_2(\theta_1|\Theta_2) \underset{\Theta_2}{\wedge} F_3(\theta_2)\right) d\theta_1 =$$
$$= \int_{\Omega_1}\int_{\Omega_2} f_1(x|\theta_1) f_2(\theta_1|\theta_2) f_3(\theta_2)d\theta_2 d\theta_1,$$

co kończy dowód. Dla pozostałych przypadków przebiega on analogicznie.

D o w ó d 2 :

$$\mathrm{E}(g(X)) = \int_{\Omega}\int_{X} g(x) f(x,\theta)dx d\theta = \int_{\Omega}\int_{X} g(x) f(x|\theta)dx \pi(\theta)d\theta =$$
$$= \int_{\Omega} \mathrm{E}_{X|\vartheta}(g(X))\pi(\theta)d\theta = \mathrm{E}_{\theta}\left(\mathrm{E}_{x|\theta}(g(X))\right)$$

Dowód 3:

$$\begin{aligned}
&\mathrm{V}(g(X)) = \mathrm{E}(g(X) - \mathrm{E}(g(X)))^2 = \mathrm{E}(g^2(X)) - \mathrm{E}^2(g(X)) = \\
&= \mathrm{E}_\theta\left(\mathrm{E}_{X|\theta}(g^2(X))\right) - \left(\mathrm{E}_\theta\left(\mathrm{E}_{X|\theta}(g(X))\right)\right)^2 = \mathrm{E}_\theta\left(\mathrm{E}_{X|\theta}(g^2(X))\right) - \mathrm{E}_\theta\left(\mathrm{E}^2_{X|\theta}(g(X))\right) + \\
&+ \mathrm{E}_\theta\left(\mathrm{E}^2_{X|\theta}(g(X))\right) - \left(\mathrm{E}_\theta\left(\mathrm{E}_{X|\theta}(g(X))\right)\right)^2 = \mathrm{E}_\theta\left(\mathrm{E}_{X|\theta}(g^2(X)) - \mathrm{E}^2_{X|\theta}(g(X))\right) + \\
&+ \mathrm{E}_\theta\left(\mathrm{E}^2_{X|\theta}(g(X))\right) - \mathrm{E}^2_\theta\left(\mathrm{E}_{X|\theta}(g(X))\right) = \mathrm{E}_\theta\left(\mathrm{V}_{x|\theta}(g(X))\right) + \mathrm{V}_\theta\left(\mathrm{E}_{x|\theta}(g(X))\right).
\end{aligned}$$

Przykład 7.

Obliczyć parametry rozkładu mieszanki $\mathrm{P}(\Lambda)\underset{\Lambda}{\wedge}\mathrm{G}(\alpha,\beta)$. Wartość oczekiwana jest równa

$$p = \omega^2\,\mathrm{E}(X) = \mathrm{E}_\Lambda\left(\underbrace{\mathrm{E}_{x|\Lambda}(X)}_{\text{Wartość oczekiwana P}(\Lambda)}\right) = \underbrace{\mathrm{E}_\Lambda(\Lambda)}_{\text{wartość oczekiwana G}(\alpha,\beta)} = \alpha\beta$$

a wariancja

$$\mathrm{V}(X) = \mathrm{V}_\Lambda\left(\mathrm{E}_{x|\Lambda}(X)\right) + \mathrm{E}_\Lambda\left(\mathrm{V}_{x|\Lambda}(X)\right) = \mathrm{V}_\Lambda(\Lambda) + \mathrm{E}_\Lambda(\Lambda) = \alpha\beta^2 + \alpha\beta = \alpha\beta(1+\beta)$$

Przedstawione w artykule rozkłady złożone w tym ujęciu służą przede wszystkim do konstruowania nowych rozkładów. Nie zawsze udaje się uzyskać ich postać analityczną. Przykładem tego jest rozkład $\mathrm{N}\left(0,\sqrt{T}\omega\right)\underset{T}{\wedge}\mathrm{Geometryczny}(p)$, gdzie nie jest znana analityczna postać następującej sumy

$$f(x) = \sum_{t=1}^{\infty} \frac{p(1-p)^{t-1}}{\sqrt{2\pi t}\,\omega} e^{-\frac{x^2}{2t\omega^2}}.$$

W takim przypadku do obliczeń należy użyć metod numerycznych. Za pomocą wyżej podanych wzorów można wyliczyć wartość oczekiwaną tego rozkładu, która jest równa zeru i wariancję w tym przypadku postaci ω^2 / p. Zadając wariancję rozkładu złożonego, np. równą 1, mamy warunek $p = \omega^2$. Znając p, można wyznaczyć ω, co upraszcza obliczenie powyższej sumy.

Literatura

[1] Gerstenkorn T., Śródka T., *Kombinatoryka i rachunek prawdopodobieństwa*, PWN, Warszawa 1972.
[2] Fisz M., *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa 1968.
[3] Johnson N.L., Kemp A.W., Kotz S., *Univariate Discrete Distributions*, Wiley, New Jersey 2005.
[4] Rose C., Smith M.D., *Mathematical Statistics with Matematica*, Springer, New York 2002.

Summary

Mixture Distributions

Apart from elementary introduction into the problem of mixture distributions, in a great detail there are discussed the basic methods of estimation: maximum likelihood and method of moments. The paper ends with presentation of compound distributions applied in insurance.