Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 37
Ekonometria 23 2009

Albert Gardoń
Uniwersytet Ekonomiczny we Wrocławiu

METODY TESTOWANIA HIPOTEZ O LICZBIE SKŁADNIKÓW MIESZANKI ROZKŁADÓW

Streszczenie: W badaniach statystycznych często wydają się uzasadnione podejrzenia, że próba nie jest jednorodna, tzn. obserwacje można podzielić na kilka podgrup, z których każda pochodzi z innego rozkładu. W takim wypadku mówi się, że rozkład, z którego została wylosowana próba, jest skończoną mieszanką rozkładów tych podgrup. W niniejszym artykule omówione zostaną istniejące metody testowania hipotez dotyczących liczby składników, z których złożona jest mieszanka, z głównym uwzględnieniem przypadku dwuskładnikowego.

Słowa kluczowe: mieszanki rozkładów, testowanie hipotez statystycznych, liczba składników mieszanki, symulacje komputerowe.

1. Wstęp

Niech $\mathbf{X} = \left[X_i \right]_{i=1}^{n}$ będzie wektorem złożonym z n niezależnych zmiennych losowych o wartościach w przestrzeni IR^d (choć najczęściej będzie to prosta rzeczywista) i jednakowym rozkładzie, zadanym gęstością f (względem odpowiedniej miary na IR^d) i niech

$$f(x) = \sum_{i=1}^{K} \pi_i f_i(x),$$

gdzie $x \in IR^d$ oraz $\sum_{i=1}^{K} \pi_i = 1$, $\forall 1 \le i \le K \; \pi_i \ge 0$, natomiast $f_i(x)_{i=1}^{K}$ będą gęstościami pewnych rozkładów prawdopodobieństwa. W takim wypadku mówi się, że składowe wektora $\mathbf{X}$ mają skończony rozkład mieszany, a ich gęstość f nazywana jest gęstością skończonej mieszanki rozkładów. Parametry $\left(\pi_i \right)_{i=1}^{K}$ nazywa się wagami mieszanki, a gęstości $\left(f_i \right)_{i=1}^{K}$ – składowymi gęstościami mieszanki.

Oczywiście, poszczególne składowe nie muszą mieć ze sobą nic wspólnego, jednak w praktyce najczęściej problem sprowadzany jest do sytuacji, w której wszystkie gęstości należą do tej samej rodziny parametrycznej, czyli:

$$f(x) = \sum_{i=1}^{K} \pi_i f_{\theta_i}(x).$$

Właśnie taki przypadek będzie głównie omawiany w niniejszym artykule, którego celem jest omówienie sposobów testowania hipotez dotyczących liczby składników mieszanki oznaczanej literą K, ponieważ ten parametr musi być znany w większości problemów estymacyjnych związanych z mieszankami i jest w nich punktem wyjścia. Oczywiście, wyjściową wartość K można sztucznie zawyżyć. Numeryczne skutki będą co najwyżej takie, że wartości nadmiarowo wprowadzonych wag nie będą równe 0, a jedynie bliskie tej wartości. Faktycznie jednak istotną wadą takiego podejścia jest znaczne wydłużenie czasu obliczeń i z tego punktu widzenia znajomość K jest szczególnie cenna.

Najczęściej testuje się po prostu, czy rozkład składowych wektora $\mathbf{X}$ w ogóle jest mieszanką, czyli sprawdza się następującą hipotezę:

$$H_0: f = f_{\theta_1};\ H_1: f = \pi f_{\theta_1} + (1-\pi) f_{\theta_2}. \tag{1}$$

Jednak nawet w tym przypadku pojawiają się pewne problemy. Wystarczy wyobrazić sobie sytuację, w której testuje się, czy próba jest jednorodna o jednowymiarowym rozkładzie normalnym, czy jest mieszanką dwóch różnych (jednowymiarowych) rozkładów normalnych. „Naturalnym" podejściem jest zastosowanie testu opartego na ilorazie wiarogodności, który asymptotycznie powinien mieć rozkład $\sim \chi^2$ o liczbie stopni swobody równej liczbie nałożonych na H_1 restrykcji doprowadzających ją do H_0. Jednak w omawianym przypadku brzmienie hipotezy alternatywnej można doprowadzić do hipotezy zerowej na dwa sposoby: wprowadzając 1 restrykcję

$$\pi = 1$$

lub 2 restrykcje

$$\theta_1 = \theta_2,$$

gdyż jednowymiarowy rozkład normalny określany jest przez dwuwymiarowy wektor parametrów, składający się z wartości oczekiwanej i dyspersji. Niestety, nie ma jasnej odpowiedzi, co jest w takim przypadku właściwą liczbą stopni swobody, dlatego tego typu testy są najczęściej bezużyteczne w odniesieniu do mieszanek. Pojawiające się trudności na ścieżce formalnej sprawiły, że dużą popularnością przy określaniu liczby parametrów mieszanki cieszą się metody nieformalne, głównie graficzne. Polegają one np. na szacowaniu liczby modalnych i punktów prze-

gięcia na krzywej estymującej gęstość czy też na sile odchylenia od prostej wykresu kwantyl-kwantyl (zob. [Everitt, Hand 1981]). W tym artykule nie będą one jednak analizowane.

2. Mieszanki dwuskładnikowe

W tym punkcie zostaną omówione sposoby testowania hipotez dotyczących liczby składowych mieszanki dla wybranych, szczególnych, jednowymiarowych przypadków. Jak wspomniano w poprzednim punkcie, najczęściej dotyczą one zagadnienia, czy dane pochodzą z jednorodnego rozkładu, czy też z mieszanki (przynajmniej dwóch składowych). Stanowią one odpowiedzi na konkretne wyzwania, jakie pojawiły się w trakcie badań naukowych w różnych dziedzinach, jak np. w medycynie, geologii czy ichtiologii (zob. [Makov i in. 1985]).

2.1. Dowolne sprecyzowane rozkłady

Pierwszy przypadek dotyczy sytuacji, gdy testowana hipoteza przybiera postać:

$$H_0: f \equiv f_1;\ H_1: f \equiv \pi f_1 + (1-\pi) f_2, \tag{2}$$

a gęstości składowe są znane, choć mogą być dowolne. Wtedy, jeśli tylko istnieje statystyka g, dla której możliwe są do obliczenia

$$m = E\left(g\left(X_1\right)\middle|H_0\right) \text{ oraz } s = \frac{1}{\sqrt{n}} D\left(g\left(X_1\right)\middle|H_0\right),$$

i jeśli H_0 jest prawdziwa, to statystyka

$$T(\mathbf{X}) = \frac{g(\mathbf{X}) - m}{s} \overset{A}{\sim} N(0,1).$$

Ponieważ znany jest przybliżony rozkład statystyki T, może ona być statystyką testową. W tym przypadku zbiór krytyczny rozmiaru α jest postaci:

$$W = (-\infty, -u_{1-\frac{\alpha}{2}}] \cup [u_{1-\frac{\alpha}{2}}, \infty),$$

przy czym u_γ będzie oznaczać kwantyl rzędu γ standardowego rozkładu normalnego. Powyższy test może mieć jednak szersze zastosowanie, jeśli dodatkowo będzie można policzyć momenty statystyki g przy założeniu prawdziwości hipotezy alternatywnej i znajomości wagi π. Wtedy za jego pomocą można testować również wielkość udziału poszczególnych składowych, czyli wartość π. Ponadto, mając takie dane, łatwo również obliczyć moc powyższego testu (zob. [Tiago de Oliveira 1965]). Warto też zauważyć, że podany zbiór krytyczny mógłby być jednostronny, postaci:

$$W = (-\infty, -u_{1-\alpha}] \text{ lub } W = [u_{1-\alpha}, \infty),$$

co powinno zwiększyć tę moc. Można sobie czasem pozwolić na taką modyfikację, ponieważ test wymaga wyspecyfikowania obu gęstości składowych. Wystarczy więc rozważyć, jaki wpływ na $E\left(g\left(X_1\right)\right)$ ma prawdziwość hipotezy alternatywnej w porównaniu z sytuacją, gdy prawdziwa jest hipoteza testowana. Jeśli powoduje jej wzrost, to (statystycznie) zawyżona powinna zostać również wartość statystyki $g(\mathbf{X})$, co sugeruje prawostronny zbiór krytyczny, natomiast jeśli powoduje jej spadek, to odwrotnie – wartość statystyki $g(\mathbf{X})$ powinna zostać zaniżona, co sugerowałoby lewostronny zbiór odrzucenia. Niestety, wpływ prawdziwości H_1 na $E\left(g\left(X_1\right)\right)$ nie zawsze jest tak jednoznaczny, a wtedy zbiór krytyczny musi pozostać obustronny. Powyższe rozważania zostaną zilustrowane przykładem numerycznym w ostatnim punkcie.

Następny przypadek obejmuje sytuacje, w których gęstości składowe są znane, a dodatkowo składowe rozkłady są symetryczne i mają jednakowe wariancje. Tym razem testowana hipoteza wyjątkowo przybiera postać (zob. [Johnson 1973]):

$$H_0: f \equiv \pi f_1 + (1-\pi) f_2;\ H_1: f \equiv f_1.$$

Statystyka testowa oparta jest na różnicy dwóch nieobciążonych estymatorów parametru π i przybiera postać:

$$T(\mathbf{X}) = \frac{\overline{X} - \mu_2}{\mu_1 - \mu_2} - \frac{\hat{F}(c) - F_2(c)}{F_1(c) - F_2(c)},$$

gdzie μ_1 i μ_2 są wartościami oczekiwanymi w rozkładach składowych, F_1 i F_2 składowymi dystrybuantami, a $\hat{F}$ dystrybuantą empiryczną wyznaczoną na podstawie wartości wektora $\mathbf{X}$. Punkt c może być zasadniczo wybrany dowolnie, ale jeśli przyjmie się, że $c = \frac{1}{2}\left(\mu_1 + \mu_2\right)$, to dyspersja $s = S_{T(\mathbf{X})}$ nie zależy od parametru π i statystyka testowa ma asymptotycznie rozkład $\sim N(0, s)$. Powoduje to, że zbiór krytyczny przybiera postać:

$$W = (-\infty, -su_{1-\frac{\alpha}{2}}] \cup [su_{1-\frac{\alpha}{2}}, \infty).$$

W pracy [Johnson 1973] obliczone zostały moce niektórych tego typu testów o rozmiarze 5% przy alternatywie, że próba jest jednorodna i pochodzi z rozkładu normalnego.

2.2. Rozkłady normalne

Kolejne dwa testy przeznaczone są dla mieszanek rozkładów normalnych. Podobnie jak na początku testowana hipoteza będzie miała postać:

$$H_0: f = f_{\theta_1};\ H_1: f = \pi f_{\theta_1} + (1-\pi) f_{\theta_2}, \tag{3}$$

przy czym, jak zaznaczono, obie składowe są gęstościami rozkładów normalnych, które nie muszą być znane, natomiast θ_1 i θ_2 są tu dwuwymiarowe. Jeśli jednak oba składniki mieszanki mają jednakową wariancję, a testowaną hipotezę przedstawi się w nieco zmodyfikowany sposób:

$$H_0 : \{X_i\}_{i=1}^{n}\ iid \sim N(\mu_1, \sigma);$$
$$H_1 : \{X_{\tau(i)}\}_{i=1}^{n_0}\ iid \sim N(\mu_1, \sigma),\ H_1 : \{X_{\tau(i)}\}_{i=n_0+1}^{n}\ iid \sim N(\mu_2, \sigma),$$

gdzie τ oznacza pewną permutację zbioru $IN \cap [1, n]$, to jej sprawdzianem może być następująca statystyka testowa:

$$T(\mathbf{X}) = \max_{n_0 \in IN \cap (0,n)} \frac{n_0 (n - n_0)\left(\overline{X}_{(1:n_0)} - \overline{X}_{(n_0+1:n)}\right)}{n\left((n_0 - 1) S^2_{\overline{X}_{(1:n_0)}} + (n - n_0 - 1) S^2_{\overline{X}_{(n_0+1:n)}}\right)},$$

przy czym $\overline{X}_{(1:n_0)}$ i $S^2_{\overline{X}_{(1:n_0)}}$ są statystykami opartymi na pierwszych n_0 elementach z próby uporządkowanej, natomiast $\overline{X}_{(n_0+1:n)}$ i $S^2_{\overline{X}_{(n_0+1:n)}}$ na pozostałych elementach. Dzięki temu, że asymptotycznie ma ona rozkład normalny (zob. [Engleman, Hartigan 1969]), możliwe jest wyznaczenie odpowiedniego zbioru krytycznego rozmiaru α:

$$W = \left[\left(1 - \frac{2}{\pi}\right)\exp\left(\frac{\mu_\alpha \sqrt{n-2} + 2{,}4}{n-2}\right) - 1, \infty\right).$$

Niestety, powyższy test nie nadaje się do stosowania w sytuacjach wielowymiarowych, gdyż wtedy asymptotyczny rozkład może nie być normalny. Pewnym ograniczeniem może być również założenie o równości wariancji. Można je jednak obejść, używając do hipotezy (3) statystyki testowej:

$$T_j(\mathbf{X}) = \overline{\left|\hat{F}\left(X_{(\cdot)}\right) - \Phi\left(\frac{X_{(\cdot)} - \overline{X}}{S_{\overline{X}}}\right)\right|^j} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{i}{n} - \Phi\left(\frac{X_{(i)} - \overline{X}}{S_{\overline{X}}}\right)\right|^j,$$

w której Φ oznacza dystrybuantę standardowego rozkładu normalnego, a $X_{(i)}$ i-tą statystykę porządkową. W tym przypadku próba nie musi pochodzić z mieszanki rozkładów normalnych o równych wariancjach, a zbiór krytyczny budowany jest prawostronnie (zob. [White 1984]).

2.3. Rozkłady Poissona

Następny test odnosi się do przypadku, w którym gęstości składowe są poissonowskie, mogą być nieznane, a testowana hipoteza ma typową postać:

$$H_0: f = f_{\theta_1};\ H_1: f = \pi f_{\theta_1} + (1-\pi) f_{\theta_2}. \tag{4}$$

Jeśli H_0 jest prawdziwa, to znany jest asymptotyczny rozkład następującej statystyki testowej (zob. [Tiago de Oliveira 1965]):

$$T(\mathbf{X}) = \frac{\sqrt{n}\left(S^2_{\overline{X}} - \overline{X}\right)}{\sqrt{1 - 2\sqrt{\overline{X}} + 3\overline{X}}} \overset{A}{\sim} N(0,1).$$

Natomiast gdy prawdziwa jest H_1, to, jak łatwo udowodnić (zob. [Makov i in. 1985]), wariancja w takim mieszanym rozkładzie jest większa od wartości oczekiwanej, a więc i różnica $S^2_{\overline{X}} - \overline{X}$ powinna przyjmować większe wartości, co sugeruje prawostronny zbiór krytyczny rozmiaru α:

$$W = [u_{1-\alpha}, \infty).$$

2.4. Rozkłady wykładnicze

Ostatni przypadek obejmuje mieszanki rozkładów wykładniczych i został zaproponowany przez polskiego matematyka J. Spławę-Neymana (zob. [White 1984]). Bez szkody dla ogólności, ze względu na możliwość skalowania rozkładów wykładniczych, można przyjąć hipotezę:

$$H_0: f \equiv f_{\theta_1};\ H_1: f \equiv \pi f_{\theta_1} + (1-\pi) f_{\theta_2},\ \theta > 1,$$

w której gęstość f_1 jest znaną gęstością rozkładu ~Exp(1), natomiast druga gęstość rozkładu ~Exp(θ) może być nieznana. Do testowania można w tym przypadku użyć statystyki:

$$T(\mathbf{X}) = \sqrt{n}(1 - \overline{X}) \overset{A}{\sim} N(0,1).$$

Ze względu na postać hipotezy alternatywnej zbiór krytyczny powinien zostać zbudowany prawostronnie, ponieważ im większy parametr θ, tym mniejsza śred-

nia, a więc tym większe wartości powinna przyjmować statystyka testowa. Ostatecznie zbiór krytyczny rozmiaru α przybiera następującą postać:

$$W = [u_{1-\alpha}, \infty).$$

3. Mieszanki wieloskładnikowe

Pomimo wielu metod podanych w poprzednim punkcie, jakie można stosować w szczególnych zagadnieniach, oczywista wydaje się chęć posiadania metody uniwersalnej, która mogłaby być stosowana dla szerokiej klasy problemów. Konkretnie chodziłoby o skonstruowanie sprawdzianu dla hipotezy dotyczącej dowolnej liczby składników, w której nie będzie istotny typ rozkładu, mającej następującą postać:

$$H_0: K = K_0;\ H_1: K = K_1 > K_0. \tag{5}$$

Oczywiście „naturalnym" sprawdzianem dla tej hipotezy byłby test oparty na ilorazie wiarogodności postaci:

$$T(\mathbf{X}) = -2 \ln \frac{L_0(X)}{L_1(X)}, \tag{6}$$

który, przy założeniu prawdziwości hipotezy zerowej i regularności f względem wszystkich swoich parametrów (a więc i wag $(\pi_i)_{i=1}^{K_1}$), miałby asymptotycznie rozkład $\sim \chi^2_{K_1-K_0}$ (o $K_1 - K_0$ stopniach swobody), co dawałoby zbiór krytyczny rozmiaru α postaci:

$$W = \left[\chi^2_{K_1-K_0,1-\alpha}, \infty\right),$$

w którym $\chi^2_{r,\gamma}$ oznacza kwantyl rzędu γ z rozkładu $\sim \chi^2_r$. Niestety, jak już wspomniano w punkcie 1, pojawiają się tu pewne problemy z liczbą stopni swobody statystyki testowej. Przeprowadzane jeszcze pod koniec lat sześćdziesiątych poprzedniego stulecia symulacje wskazywały początkowo, że dla d-wymiarowych danych normalnych i gdy $K_1 - K_0 = 1$ statystyka (6) powinna mieć $d+1$ stopni swobody, gdy macierze kowariancji poszczególnych składników mieszanki są sobie równe, lub $(d+1)\left(\frac{1}{2}d+1\right)$ stopni swobody w przeciwnym przypadku, lecz wkrótce, po przeprowadzeniu wnikliwszych symulacji, wycofano się z tego. Pod koniec lat 70. podejrzewano, że w przypadku hipotezy (1), gdy parametr θ jest q-wymiarowy, liczba stopni swobody powinna zawierać się między q a $q+1$. Ostatecznie jednak jedyny teoretycznie uzasadniony wynik otrzymano w połowie lat 80. Orzeka on, że

w tym ostatnio wyszczególnionym przypadku liczba stopni swobody jest równa q, jednak pod warunkiem że dokładna wartość π jest wyspecyfikowana w alternatywie.

Oprócz powyższego problemu dochodzą jeszcze kłopoty z regularnością gęstości względem wag mieszanki (jest ona jednym z głównych założeń w rozważaniach nad asymptotyką testów opartych na ilorazie wiarogodności). Przede wszystkim, gdy część z nich się zeruje, a tak się dzieje w wypadku prawdziwości testowanej hipotezy, wtedy leżą one na brzegu przestrzeni parametrów. Mało tego, jeśli jedna z wag znika, dajmy na to π_i, wtedy wiarogodność alternatywy jest stała dla wszystkich wartości odpowiedniego parametru θ_i. Podobnie dzieje się w przypadku, gdy założona zostanie równość dwóch parametrów, np. $\theta_i = \theta_{i+1}$. Wtedy wiarygodność alternatywy pozostanie stała, gdy tylko suma odpowiednich wag $\pi_i = \pi_{i+1}$ będzie stała bez względu na ich wartości. To powoduje, że teoretyczne założenia dotyczące asymptotyki w pewnych przypadkach mogą nie być spełnione (zob. [Makov i in. 1985]). Co prawda, dla rozkładów ~Exp( ), ~Poi( ) i ~Bin( ) symulacje dają zadowalające wyniki (choć uzasadnienie teoretyczne budzi wątpliwości – zob. [White 1984]), ale w przypadku rozkładu normalnego zbieżność do podanego rozkładu asymptotycznego niestety nie zachodzi. Można sobie z tym poradzić co najwyżej przy hipotezie postaci (1). Rozkład graniczny jest dla niej poprawny wówczas, gdy przy prawdziwości hipotezy alternatywnej estymator parametru π, uzyskany metodą największej wiarogodności, będzie mniejszy niż 1, co zdarza się dokładnie z prawdopodobieństwem $\frac{1}{2}$. Daje to następujący zbiór krytyczny rozmiaru α:

$$\left(\hat{\pi}_L, T(\mathbf{X})\right) \in W^* = [0,1) \times \left[\chi^2_{1,1-2\alpha}, \infty\right),$$

gdzie $\hat{\pi}_L$ jest wspomnianym estymatorem największej wiarogodności, a statystyka T zadana jest równaniem (6). Natomiast dla hipotezy (5), przy d-wymiarowych danych o rozkładzie normalnym i jednakowych macierzach kowariancji, skonstruowano w latach 70. następujący test (zob. [Makov i in. 1985]) będący pewną modyfikacją ilorazu wiarogodności:

$$T(\mathbf{X}) = -\frac{1}{2}(2n - 2 - 2d - K_1) \ln \frac{L_0(\mathbf{X})}{L_1(\mathbf{X})} \overset{A}{\sim} \chi^2_{2d(k_1 - K_0)}.$$

Oczywiście, powyższy rozkład graniczny jest osiągany pod warunkiem prawdziwości testowanej hipotezy, co daje zbiór krytyczny rozmiaru α postaci:

$$W = \left[\chi^2_{2d(K_1 - K_0),1-\alpha}, \infty\right).$$

Niestety, również powyższy wynik został uzyskany głównie na podstawie symulacji, trudno więc się dziwić, że w późniejszych latach wytknięto mu wiele nie-

dociągnięć. Przede wszystkim na początku lat 80., po przeprowadzeniu wnikliwszych symulacji (!), stwierdzono, że podany rozkład przybliżony jest bliski prawdziwemu dopiero dla $n > 10d$. Dodatkowo wykazano, że moc testu jest bardzo mała, gdy $K_1 - K_0 \leq 2$.

Podsumowując rozważania w tym punkcie, trudno nie zauważyć, że poważną wadą istniejących metod są ich mizerne podstawy teoretyczne. Symulacje są oczywiście ważnym narzędziem we współczesnej matematyce i pozwalają na uzyskanie cennych podejrzeń na temat natury badanego zjawiska. Jednak ze względu na to, że mogą się odnosić tylko do konkretnych przypadków, należy być bardzo ostrożnym przy wysuwaniu ogólnych wniosków na ich podstawie, gdyż później może pojawić się osoba, która przeprowadzi „wnikliwsze" symulacje.

4. Przykład numeryczny

Zaprezentowane w tym punkcie testy będą dotyczyły mieszanki dwóch rozkładów Poissona. Do 100-elementowej próby wylosowanych zostało niezależnie, za pomocą kongruentnego generatora liczb pseudolosowych, 60 obserwacji z rozkładu ~Poi(3) i 40 obserwacji z rozkładu ~Poi (4) o gęstościach, które będą dalej oznaczane odpowiednio jako f_3 i f_4. Można więc uważać, że wygenerowana w ten sposób próba pochodzi z rozkładu będącego następującą mieszanką:

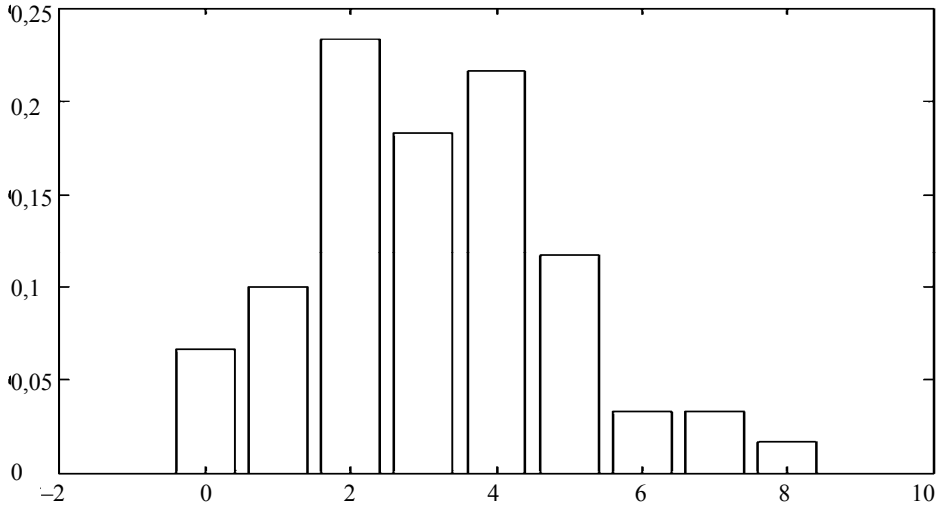
$$f(x) = \frac{3}{5} f_3(x) + \frac{2}{5} f_4(x) = \frac{3}{5} \cdot e^{-3} \frac{3^x}{x!} II_{IN_0}(x) + \frac{2}{5} \cdot e^{-4} \frac{4^x}{x!} II_{IN_0}(x), \qquad (7)$$

gdzie *II* oznacza funkcję charakterystyczną (indykator) zbioru (ma wartość 1, gdy argument należy do zbioru podanego jako indeks, i 0 w przeciwnym przypadku). Oczywiście, jest to gęstość rozkładu prawdopodobieństwa względem miary zliczającej na zbiorze IN_0. Wyniki losowania zostały przedstawione na histogramach gęstości zawartych na rys. 1-3. Do obliczeń będą jednak potrzebne tylko realizacje dwóch statystyk: średniej arytmetycznej i wariancji, które w wygenerowanej próbie miały wartości: $\overline{X} = 3{,}66$ i $S_X^2 = 3{,}8$.

4.1. Udział składników w mieszance

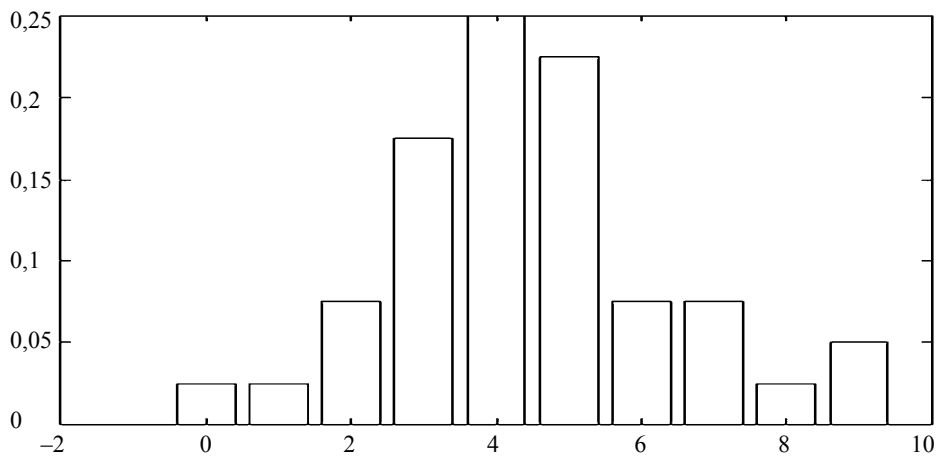
Na początek, na podstawie metody przedstawionej przy hipotezie (2), zostanie przeprowadzony test dotyczący wartości parametru π. W tym celu postawiona zostanie następująca hipoteza:

$$H_0: f \equiv \pi f_3 + (1-\pi) f_4,\ \pi = \frac{1}{2};\ H_1: f \equiv \pi f_3 + (1-\pi) f_4,\ \pi < \frac{1}{2}.$$



Rys. 1. Histogram gęstości dla 60 obserwacji wylosowanych z rozkładu ~Poi(3)

Źródło: opracowanie własne.

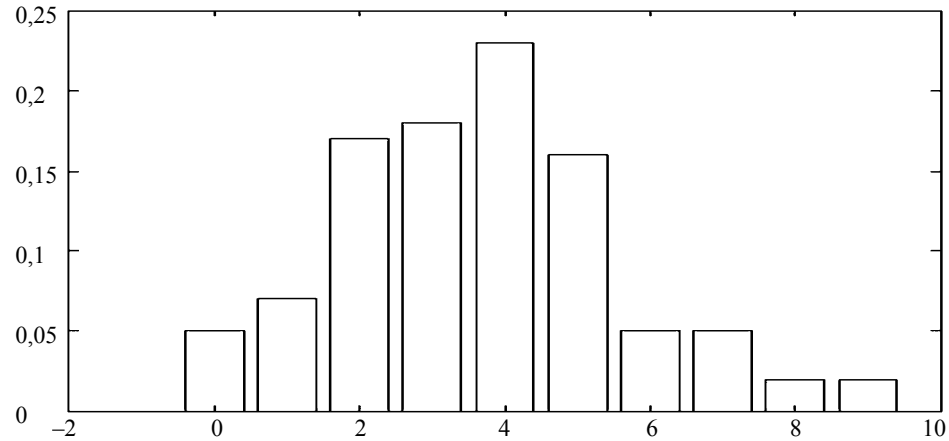


Rys. 2. Histogram gęstości dla 40 obserwacji wylosowanych z rozkładu ~Poi(4)

Źródło: opracowanie własne.

Za konieczną tu statystykę g przyjęte zostanie odwzorowanie identycznościowe, czyli $g(x) = x$. Należy zwrócić uwagę, że gęstości składowe musiały zostać w tym przypadku dokładnie wyspecyfikowane, co pozwala łatwo obliczyć:

$$m = E\left(g\left(X_1 \middle| H_0\right)\right) = E\left(X_1 \middle| H_0\right) = 3{,}5\,,$$



Rys. 3. Histogram gęstości dla całej 100-elementowej próby (rozkład mieszany)

Źródło: opracowanie własne.

$$s = \frac{1}{\sqrt{n}} D\left(g\left(X_1\right)\middle|H_0\right) = \frac{1}{\sqrt{n}} D\left(X_1\middle|H_0\right) = \sqrt{0{,}0375} = 0{,}194,$$

czyli jeśli H_0 jest prawdziwa, to statystyka

$$T(\mathbf{X}) = \frac{\overline{X} - 3{,}5}{0{,}194} \overset{A}{\sim} N(0,1).$$

Ponieważ znana jest realizacja statystyki $\overline{X}$ i jest ona większa od 3,5, sugeruje to, że większy udział w mieszance ma rozkład o większej wartości oczekiwanej, czyli ~Poi(4). Dlatego sensowne jest postawienie alternatywy $\pi < \frac{1}{2}$. Jej prawdziwość, jak właśnie wspomniano, oznaczałaby większy udział rozkładu ~Poi(4) w mieszance, co powinno zawyżać wartość statystyki testowej. Z tego względu generowany będzie przez nią prawostronny zbiór krytyczny:

$$W = \left[u_{1-\alpha}, \infty\right).$$

Można też zignorować wiedzę o $\overline{X}$ i sformułować alternatywę jako proste zaprzeczenie ($\pi \neq \frac{1}{2}$), co również będzie poprawne, ale będzie wymagało skonstruowania obustronnego zbioru krytycznego i zmniejszy moc testu. Oczywiście, zaobserwowana wartość $\overline{X}$, a co za tym idzie – również zaproponowana do hipotezy zerowej alternatywa, nie zgadza się z rozkładem zdefiniowanym gęstością (7), z którego faktycznie próba została wygenerowana. Należy jednak pamiętać, że takie sytuacje są bardzo częste w teorii prawdopodobieństwa, gdzie w konkretnych

losowaniach mogą się realizować zdarzenia mniej prawdopodobne. Trzeba również być świadomym, że w rzeczywistych sytuacjach statystyk dysponuje tylko próbą, nie mając wiedzy o rozkładzie, z którego ona pochodzi. Ponadto postawienie przeciwnej alternatywy ($\pi > \frac{1}{2}$) spowodowałoby, że bez wykonywania obliczeń można by orzec o braku podstaw do odrzucenia testowanej hipotezy przy każdym rozsądnym poziomie istotności, gdyż dane wskazywałyby, że taka alternatywa jest znacznie gorsza od hipotezy zerowej.

Zaobserwowana w tym przypadku (zdarzenie ω_0) wartość statystyki testowej to:

$$T\left(X\left(\omega_0\right)\right) = 0{,}826\,,\ p = 0{,}205\,,$$

gdzie p to *p-value*, czyli taki poziom istotności, przy którym zaobserwowana wartość statystyki testowej leży na brzegu zbioru krytycznego, a więc przy każdym większym poziomie istotności H_0 zostałaby odrzucona, a przy mniejszym nie byłoby do tego podstaw. W przypadku zbioru obustronnego *p-value* byłoby dwukrotnie większe. Dane wskazują więc, że przy odrzuceniu hipotezy zerowej, z prawdopodobieństwem nie mniejszym niż 0,205 popełniony zostałby błąd, co w praktyce oznacza, że nie ma podstaw do stwierdzenia, że π jest istotnie mniejsze niż $\frac{1}{2}$. Ostatecznie więc test dał wynik zgodny z oczekiwaniami wynikającymi z postaci gęstości (7). Można teraz wyznaczyć moc tego testu. Będą do tego potrzebne wartość oczekiwana i odchylenie standardowe elementów wektora X w sytuacji, gdy prawdziwa jest hipoteza alternatywna. Proste obliczenia prowadzą do wyników:

$$m_1(\pi) = E\left(g\left(X_1\right)\middle|H_1\right) = E\left(X_1\middle|H_1\right) = 4 - \pi\,,$$

$$s_1(\pi) = \frac{1}{\sqrt{n}} D\left(g\left(X_1\right)\middle|H_1\right) = \frac{1}{\sqrt{n}} D\left(X_1\middle|H_1\right) = \frac{\sqrt{4-\pi^2}}{10}\,,$$

z czego wynika, że w razie prawdziwości alternatywy statystyka testowa ma postać:

$$T(\mathbf{X}) = \frac{\overline{X} - 3{,}5}{0{,}194} \overset{A}{\sim} N\left(\frac{0{,}5-\pi}{0{,}194}, \frac{\sqrt{4-\pi^2}}{0{,}194}\right),$$

zatem moc testu w zależności od poziomu istotności i alternatywnej wartości wagi π wynosi:

$$M(\pi,\alpha) = P\left(Y(\mathbf{X}) \in W\middle|H_1\right) = \Phi\left(\frac{5 - 1{,}94u_{1-\alpha} - 10\pi}{\sqrt{4-\pi^2}}\right).$$

Stąd dla dwóch najczęściej przyjmowanych poziomów istotności $\alpha = 0{,}01$ oraz $\alpha = 0{,}05$ moc wyraża się wzorami:

$$M(\pi,0,01)=\Phi\left(\frac{0,49-10\pi}{\sqrt{4-\pi^2}}\right) \text{ oraz } M(\pi,0,05)=\Phi\left(\frac{1,81-10\pi}{\sqrt{4-\pi^2}}\right),$$

co dla wybranych wartości π wynosi:

α \ π	0,4	0,3	0,2	0,1	0
0,01	0,038	0,102	0,224	0,399	0,597
0,05	0,131	0,274	0,463	0,657	0,817

Warto zauważyć następującą rzecz: ostatnia kolumna zawiera moce testu hipotezy o równym udziale rozkładów ~Poi(3) i ~Poi(4) w mieszance, przeciwko alternatywie, że próba pochodzi tylko z rozkładu ~Poi(4).

4.2. Jednorodność próby

Kolejny przykład będzie dotyczył tej samej metody, czyli tej stosowanej w przypadku problemu (2), ale przy inaczej postawionej hipotezie, stawiającej pytanie, czy próba jest jednorodna i pochodzi z rozkładu ~Poi(4), czy też jest mieszanką rozkładów ~Poi(3) i ~Poi(4):

$$H_0: f \equiv f_4,\ H_1: f \equiv \pi f_3 + (1-\pi) f_4,\ \pi \in (0,1).$$

Ponownie za statystykę g zostanie przyjęte odwzorowanie identycznościowe, zmienią się jednak odpowiednie momenty składowych wektora **X**, ponieważ zmianie uległa hipoteza zerowa:

$$m = E\left(g\left(X_1 \middle| H_0\right)\right) = E\left(X_1 \middle| H_0\right) = 4,$$

$$s = \frac{1}{\sqrt{n}} D\left(g\left(X_1\right) \middle| H_0\right) = \frac{1}{\sqrt{n}} D\left(X_1 \middle| H_0\right) = 0,2,$$

Stąd zaobserwowana wartość statystyki testowej wyniesie

$$T\left(X\left(\omega_0\right)\right) = 5 \cdot (3,66 - 4) = -1,7.$$

Prawdziwość hipotezy alternatywnej spowodowałaby zmniejszenie wartości m w stosunku do sytuacji, gdy prawdziwa jest hipoteza testowana, więc jedynie zbyt małe wartości statystyki $T(\mathbf{X})$ powinny skłaniać do odrzucenia H_0, a co za tym idzie – zbiór krytyczny powinien być lewostronny, postaci:

$$W = \left(-\infty, u_{1-\alpha}\right].$$

Powoduje to, że *p-value* wynosi w tym przypadku $p = 0,045$, co pozwala odrzucić hipotezę zerową przy rozsądnym poziomie istotności 5%, ale gdyby wyma-

gany był większy stopień pewności przy przyjęciu alternatywy, np. 99%, próba odrzucenia testowanej hipotezy nie powiodłaby się. Można jednak uznać, że test spisał się w tym przypadku całkiem poprawnie. Jeśli chodzi o jego moc, to w związku z tym, że przy założeniu prawdziwości hipotezy alternatywnej statystyka testowa

$$T(\mathbf{X}) = 5(\bar{X} - 4) \overset{A}{\sim} N\left(-5\pi, \frac{\sqrt{4-\pi^2}}{2}\right),$$

wyraża się ona wzorem:

$$M(\pi, \alpha) = \Phi\left(\frac{10\pi - 2u_{1-\alpha}}{\sqrt{4-\pi^2}}\right),$$

co dla $\alpha = 0{,}01$ i $\alpha = 0{,}05$ daje:

$$M(\pi, 0{,}01) = \Phi\left(\frac{10\pi - 4{,}65}{\sqrt{4-\pi^2}}\right) \text{ oraz } M(\pi, 0{,}05) = \Phi\left(\frac{10\pi - 3{,}29}{\sqrt{4-\pi^2}}\right).$$

Oto kilka wybranych wartości mocy:

α \ π	0,1	0,3	0,5	0,75	1
0,01	0,034	0,202	0,572	0,938	0,999
0,05	0,126	0,442	0,811	0,988	0,9999

Ze względu na wartość zaobserwowanej średniej arytmetycznej, która jest bliższa wartości 4 niż 3, wyniki byłyby bardziej wyraziste, gdyby próbowano testować hipotezę o pochodzeniu próby z rozkładu ~Poi(3), postaci:

$$H_0: f \equiv f_3;\ H_1: f \equiv \pi f_3 + (1-\pi) f_4,\ \pi \in (0,1).$$

Wtedy

$$m = E\left(g\left(X_1 \middle| H_0\right)\right) = E\left(X_1 \middle| H_0\right) = 3,$$

$$s = \frac{1}{\sqrt{n}} D\left(g\left(X_1\right) \middle| H_0\right) = \frac{1}{\sqrt{n}} D\left(X_1 \middle| H_0\right) = \sqrt{0{,}03} = 0{,}173,$$

a zbiór krytyczny, ze względu na powiększenie średniej przy prawdziwości alternatywy, jest prawostronny, postaci:

$$W = \left[u_{1-\alpha}, \infty\right).$$

To daje:

$$T\left(X\left(\omega_0\right)\right) = \frac{3{,}66 - 3}{0{,}173} = 3{,}81,\ p = 7 \cdot 10^{-5},$$

czyli hipoteza zerowa zostanie odrzucona przy każdym rozsądnym poziomie istotności, a prawdopodobieństwo, że został przy tym popełniony błąd, będzie zerowe z dokładnością nawet do czterech miejsc po przecinku! Moc testu, wyrażona równaniem:

$$M(\pi,\alpha)=\Phi\left(\frac{10-\sqrt{3}u_{1-\alpha}-10\pi}{\sqrt{4-\pi^2}}\right),$$

będzie miała następujące przykładowe wartości:

α \ π	0,9	0,7	0,5	0,25	0
0,01	0,045	0,292	0,699	0,96	0,999
0,05	0,151	0,532	0,867	0,99	0,9998

Jest to zatem najmocniejszy test spośród tu weryfikowanych, służący do rozróżnienia, czy próba pochodzi z równomiernej mieszanki rozkładów ~Poi(3) i ~Poi(4) ($\pi=\frac{1}{2}$), czy też jest jednorodna.

Na koniec zostanie sprawdzone, jak z identyfikacją wygenerowanej mieszanki radzi sobie metoda, która stosowana jest, gdy hipoteza statystyczna wyrażona jest formułą (4). Należy zauważyć, że tym razem zadanie polega na udzieleniu odpowiedzi na pytanie, czy dane pochodzą z jednego rozkładu Poissona, czy też są mieszanką dwóch rozkładów tego typu, bez specyfikowania ich parametrów. Zaobserwowana na podstawie próby wartość statystyki testowej oraz *p-value* w tym przypadku to:

$$T\left(X\left(\omega_0\right)\right)=\frac{10\cdot(3,8-3,66)}{\sqrt{1-2\sqrt{3,66}+3\cdot 3,66}}=0,49\ ;\ p=0,31,$$

co praktycznie nie daje żadnych podstaw do uważania, że wygenerowana próba pochodzi z rozkładu mieszanego. Innymi słowy różnica między S_X^2 i $\overline{X}$ w tej próbie nie była statystycznie istotna. Przyczyną znacznie mniejszej czułości tego testu na istnienie mieszanki, w porównaniu z poprzednimi, jest oczywiście to, że nie wymaga on dokładnej specyfikacji ani składowych gęstości, ani wag. Jego zaletą jest jednak to, że jest znacznie ogólniejszy i może być stosowany dla szerszej klasy problemów. Niestety, w związku z nieznajomością rozkładu statystyki testowej, gdy prawdziwa jest hipoteza alternatywna, nie da się wyznaczyć mocy tego testu.

Literatura

Engleman L., Hartigan J.A., *Percentage points for a test for clusters*, „Journal of American Statistical Association" 1969, vol. 64, s. 1647-1648.

Everitt B.S., Hand D.J., *Finite mixture distributions*, Chapman and Hall, London-New York 1981.

Johnson N.L., *Some simple tests of mixtures with simmetrical components*, „Communication Statistics" 1973, vol. 1, s. 17-25.

Makov U.E., Smith A.F.M., Titterington D.M., *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons, Chichester-New York-Brisbane-Toronto-Singapore 1985.

Tiago de Oliveira J., *Classical and Contagious Discrete Distributions*, Pergamon, New York 1965.

White H., *Asymptotic Theory for Econometricians*, Academic Press, San Diego 1984.

STATISTICAL TESTS FOR THE NUMBER OF COMPONENTS IN MIXED DISTRIBUTIONS

Summary: The main aim of this work is to demonstrate methods which allow to recognize, whether the sample is homogeneous or comes from a convex combination of two distributions. However, in the case of normal distributions a test for any fixed number of components is also shown. In the last section results of a numerical example are presented, in which the sample has been generated from two different Poisson distributions.