O pewnym zadaniu egzaminacyjnym

Antoni Kościelski

1 Treść zadania

Zmienne losowe $X_1, X_2, \ldots, X_n$ przyjmują wartości dodatnie, są niezależne i mają ten sam rozkład. Dla $k = 1, 2, \ldots n$ oblicz wartości oczekiwane

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

2 Rozwiązanie

Jak zwykle, najtrudniej rozwiązać najprostsze zadania. Zadanie to można rozwiązywać w następujący sposób:

Krok 1. Pokazujemy dla $j = 1, 2, \ldots, n$, że zmienne

$$\frac{X_j}{\sum_{i=1}^{n} X_i}$$

mają ten sam rozkład.

Krok 2. Jeżeli uda nam się wykonać krok 1, to jako wniosek otrzymyjemy, że zmienne te mają te same wartości oczekiwane, a więc dla pewnej liczby c i dla $j = 1, 2, \ldots, n$ mamy

$$E\left(\frac{X_j}{\sum_{i=1}^{n} X_i}\right) = c.$$

Krok 3. Liczymy stałą c. Oczywiście zachodzą następujące równości

$$n \cdot c = \sum_{j=1}^{n} E\left(\frac{X_j}{\sum_{i=1}^{n} X_i}\right) = E\left(\sum_{j=1}^{n} \frac{X_j}{\sum_{i=1}^{n} X_i}\right) = E\left(\frac{\sum_{j=1}^{n} X_j}{\sum_{i=1}^{n} X_i}\right) = E(1) = 1.$$

Stąd otrzymujemy, że $c = \frac{1}{n}$.

Krok 4. Teraz wystarczy zauważyć, że

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

3 Komentarz

Nietrudno zauważyć, że jeżeli coś budzi wątpliwości, to jest to krok 1. Przeprowadzenie pozostałych rachunków wymaga jedynie znajomości wzoru $E(X+Y) = E(X) + E(Y)$, a więc zasady, że wartość oczekiwana sumy zmiennych losowych jest sumą wartości oczekiwanych tych zmiennych. Może trzeba jeszcze zauważyć, że rozkład zmiennej wyznacza jej wartość oczekiwaną.

Stwierdzenie z kroku pierwszego wydaje się dość oczywiste. Często jest tak, że jeżeli weźmiemy rzeczy podobne, zrobimy z nimi to samo, to otrzymamy coś podobnego. Mogłoby się więc wydawać, że jeżeli weźmiemy na przykład dwie pary zmiennych losowych X_1 i X_2 oraz Y_1 i Y_2, pary, których elementy mają ten sam rozkład, a następnie na pierwszych i drugich z tych elementów wykonamy to samo przekształcenie F, to zmienne $F(X_1, Y_1)$ oraz $F(X_2, Y_2)$ otrzymane w wyniku tego przekształcenia też powinny mieć ten sam rozkład. Takim przekształceniem mogłoby być na przykład dodawanie zmiennych $F(X,Y) = X+Y$ lub dzielenie przez sumę $F(X,Y) = X/(X+Y)$. Niestety, tego typu stwierdzenia nie zawsze są prawdziwe. W związku z tym, jeżeli są prawdziwe, to wymagają uzasadnienia.

4 Przykład

Przypuśćmy, że losujemy punkt z jednostkowego kwadratu $(0,1) \times (0,1)$, a więc o współrzędnych x i y spełniających warunek $0 < x, y < 1$, a sposób losowania podlega rozkładowi jednostajnemu. Możemy rozważać dwie zmienne losowe: X, która wylosowanemu punktowi przyporządkowuje jego pierwszą współrzędną, oraz Y określoną jako druga współrzędna wylosowanego punktu.

Nietrudno zauważyć, że dla $a \in (0,1)$

$$P(X < a) = \text{ pole kwadratu } (0,a) \times (0,1) = a$$

oraz

$$P(Y < a) = \text{ pole kwadratu } (0,1) \times (0,a) = a.$$

Tak więc dystrybuanty obu zmiennych są funkcjami identycznościowymi na odcinku (0,1), a poza nim są stałe. Stąd otrzymujemy, że obie zmienne mają ten sam rozkład i jest to rozkład jednostajny na odcinku $(0,1)$.

Ze zmiennych X i Y możemy utworzyć dwie pary X_1, X_2 oraz Y_1, Y_2 przyjmując $X_1 = X_2 = Y_2 = X$ oraz $Y_1 = Y$. Jest oczywiste, że zmienne w obu parach mają ten sam rozkład. Teraz dodajmy pierwsze i drugie współrzędne z obu par tworząc zmienne

$$S_1 = X_1 + Y_1 = X + Y \ \text{ oraz } \ S_2 = X_2 + Y_2 = X + X = 2X.$$

Mogłoby się wydawać, że zmienne S_1 i S_2 mają ten sam rozkład. Można jednak przekonać się, że tak nie jest.

W przypadku zmiennych X i Y znamy ich łączny rozkład. To pozwala znaną metodą znaleźć łączny rozkład innej, przekształconej pary zmiennych

$$S = S_1 = X + Y \ \text{ oraz } \ Y.$$

Przekształcenie odwrotne jest dane wzorami

$$X = S - Y \ \text{ oraz } \ Y = Y$$

i jest określone na zbiorze

$$A = \{(s, y) \in R^2 : (0 < S \leqslant 1 \wedge 0 < Y < S) \vee (1 < S < 2 \wedge (S - 1 < Y < 1)\},$$

czyli na rombie o wierzchołkach $(0, 0)$, $(1, 0)$, $(1, 1)$ i $(2, 1)$. Gęstość $f_{X,Y}$ pary (X, Y) jest stale równa 1, jakobian przekształcenia odwrotnego też stale równa się 1, więc gęstość $f_{S,Y}$ pary (S, Y) jest również stale równa 1 na wyżej podanym rombie A. Całkując znajdujemy gęstość sumy $S = S_1$

$$f_{S_1}(s) = f_S(s) = \int_{\{y \in R : (s,y) \in A\}} 1\, dy = \begin{cases} s & \text{jeżeli } 0 < s \leqslant 1, \\ 2 - s & \text{jeżeli } 1 < s < 2. \end{cases}$$

Analogicznie (ale dla przekształcenia pojedyńczej zmiennej) lub posługując się dystrybuantą znajdujemy gęstość zmiennej $S_2 = 2X$. Łatwo przekonać się, że

$$f_{S_2}(s) = f_{2X}(s) = \frac{s}{2}$$

dla $s \in (0, 2)$. Z udowodnionych wzorów łatwo wynika, że rozkłady zmiennych S_1 i S_2 są różne. Świadczą o tym różne (ciągłe) gęstości, można też wyliczyć prawdopodobieństwo jakiegoś zdarzenia dla każdego rozkładu. Prawie zawsze powinniśmy otrzymać różne wartości.

5 Pierwszy krok

Wiele zadań z ćwiczeń, a także przedstawione wyżej rozumowanie pozwalają wyobrazić sobie plan uzasadnienia pierwszego kroku. Znajomość rozkładu łącznego kilku zmiennych często pozwala na znalezienie rozkładów nowych, dziwnych zmiennych losowych, utworzonych z danych. Skąd wziąć rozkład łączny układu zmiennych? W przypadku zmiennych niezależnych jest to bardzo proste: znamy odpowiedni wzór. Dalsze rozumowanie zostanie przedstwione przy dodatkowym założeniu, że $n = 3$. Zamiast zmiennych X_1, X_2 i X_3 będziemy zajmować zmiennymi X, Y i Z (bez indeksów).

Zgodnie z warunkami zadania zmienne X, Y i Z przyjmują dodatnie wartości, są niezależne i mają ten sam rozkład. Przyjmijmy, że $f : (0, \infty) \to [0, \infty)$ jest gęstością rozkładu tych zmiennych. Ponieważ zmienne są niezależne, gęstość rozkładu łącznego tych zmiennych, określona na zbiorze $A = \{(x, y, z) \in R^3 : x, y, z > 0\}$, jest dana wzorem

$$f_{X,Y,Z}(x, y, z) = f(x)f(y)f(z).$$

Interesują nas dwie zmienne

$$U = \frac{X}{X + Y + Z} \quad \text{oraz} \quad W = \frac{Y}{X + Y + Z}$$

i chcemy porównać ich rozkłady. W tym celu najpierw znajdziemy łączny rozkład trzech zmiennych U, W oraz $S = X + Y + Z$. Rozważamy więc przekształcenie h określone na zbiorze A, zdefiniowane wzorem

$$h(x, y, z) = (\frac{x}{x + y + z}, \frac{y}{x + y + z}, x + y + z).$$

Najpierw musimy ustalić, jakie wartości przyjmuje przekształcenie h na zbiorze A. Zauważmy, że obraz zbioru A wyznaczony przez h jest dany wzorem

$$h[A] = \{(u, w, s) \in R^3 : u, w, s > 0 \wedge u + w < 1\} = B.$$

Można się o tym przekonać biorąc funkcję g określoną na zbiorze B, zdefiniowaną wzorem

$$g(u, w, s) = (us, ws, s - us - ws) = (x(u, w, s), y(u, w, s), z(u, w, s))$$

i dowodząc, że jest to funkcja odwrotna do h. Przy okazji mamy przekształcenie odwrotne do h.

Teraz liczymy jakobian przekształcenia g:

$$J = \begin{vmatrix} s & 0 & u \\ 0 & s & w \\ 0 & 0 & 1 \end{vmatrix} = s^2.$$

W tej sytuacji, zgodnie ze znanym wzorem, łączna gęstość zmiennych U, W i S wyraża się równaniem

$$f_{U,W,S}(u, w, s) = f(us)f(ws)f(s - us - ws)s^2.$$

Gęstości poszczególnych zmiennych znajdujemy całkując gęstość łączną. Tak więc

$$f_U(u) = \int_{\{(w,s):(u,w,s)\in B\}} f_{U,W,S}\, dw\, ds = \int_0^\infty \int_0^{1-u} f(us)f(ws)f(s-us-ws)s^2\, dw\, ds.$$

Analogicznie,

$$f_W(w) = \int_{\{(u,s):(u,w,s)\in B\}} f_{U,W,S}\, dw\, ds = \int_0^\infty \int_0^{1-w} f(us)f(ws)f(s-us-ws)s^2\, du\, ds.$$

Całek tych już raczej nie uprościmy, ale jeżeli w pierwszym wzorze literkę w zastąpimy przez v, a u przez x, zaś w drugim wzorze u zamienimy na v i w na x, to okaże się, że gęstości zmiennych U i W są opisane tymi samymi wzorami

$$f_U(x) = f_W(x) = \int_0^\infty \int_0^{1-x} f(xs)f(vs)f(s - xs - vs)s^2\, dv\, ds,$$

a więc są takie same.

Pokazaliśmy już równość rozkładów zmiennych U i W. Mamy jednak pokazać równość rozkładów trzech zmiennych: U, W oraz

$$W' = \frac{Z}{X + Y + Z}.$$

Można ten fakt uznać już za oczywisty. Dokładnie to samo rozumowanie powinno prowadzić do wniosku, że zmienne U i W', a w konsekwencji także wszystkie trzy rozważane zmienne, mają ten sam rozkład.

6 Końcowy komentarz

Przedstawione wyżej rozumowanie wymaga kilku założeń, których nie ma w treści zadania. Na przykład milcząco zostało założone, że rozważamy zmienne o ciągłych rozkładach. Umożliwiło to skorzystanie z twierdzenia o podstawianiu dla całek z funkcji wielu zmiennych, bardzo skomplikowanego, ale znanego i wielokrotnie wykorzytywanego.

Rozwiązywane zadanie ma raczej proste, elementarne rozwiązanie, nie wymagające zaawansowanych twierdzeń. Chyba wymaga za to posłużenia się mniej znanym (mniej przećwiczonym) aparatem pojęciowym, na przykład rozkładami zmiennych, czyli pewnymi przestrzeniami probabilistycznymi. W rozwiązaniu istotne są trzy fakty. Po pierwsze, dzięki niezależności danych zmiennych losowych, łatwo możemy odtworzyć przestrzeń probabilistyczną opisującą te zmienne łącznie. Po drugie, fakt, że dane zmienne mają ten sam rozkład pociąga za sobą, że ta łaczna przestrzeń nie zależy od kolejności, w jakiej zmienne zostaną wymienione. W końcu, wyliczenie rozkładu zmiennej $X_1/(X_1 + X_2 + \ldots + X_n)$ jest możliwe dzięki przekształceniu takiemu, jak wyżej zdefiniowane h, a zwłaszcza jego różnowartościowości.