Podstawy statystyki opisowej

Szymon Wąsowicz

2014 – · · ·

Przedmowa

Niniejszy skrypt zawiera wykład podstaw statystyki opisowej. Ta część statystyki stanowi pierwszy etap w analizie danych. Nie stosuje się tu metod rachunku prawdopodobieństwa (który jest językiem statystyki matematycznej). Prezentowane pojęcia ilustrowane są szeroką gamą przykładów. Zamieszczono też zestaw ćwiczeń do samodzielnego rozwiązania, a wszystkie ćwiczenia opatrzono odpowiedziami.

Skrypt wyrósł na gruncie zajęć z podstaw statystyki prowadzonych przez szereg lat i przeznaczonych dla studentów różnych kierunków: inżynierskich, humanistycznych czy medycznych. Wyrażam nadzieję, że będzie on pomocny w opanowaniu najprostszych metod statystyki opisowej.

Niektóre zamieszczone w tym opracowaniu przykłady wzbogacono o obliczenia wykonane w środowisku **R**. Jest to całkowicie bezpłatny pakiet komputerowy przeznaczony głównie do obliczeń statystycznych. Można go pobrać ze strony `www.r-project.org`. Skrypt nie stanowi jednak podręcznika środowiska **R**. Czytelnicy zainteresowani jego użyciem powinny samodzielnie zapoznać się z jego podstawami np. z licznych podręczników dostępnych w Internecie. Inne osoby mogą całkowicie pominąć zaprezentowane obliczenia bez szkody dla zrozumienia treści statystycznych.

Składam gorące podziękowanie Panu Sylwestrowi Błaszczukowi za cenne sugestie merytoryczne i redakcyjne.

Szymon Wąsowicz

Spis treści

1

Uwagi o przedmiocie statystyki

Populacja generalna

Populacją generalną nazywamy zbiorowość, której dotyczy badanie statystyczne. Może to być zbiorowość ludzka (jak np. w badaniu wzrostu czy masy ciała), ale niekoniecznie. W badaniu średniego przebiegu opon samochodowych danego typu populacją generalną jest zbiór wszystkich opon tego typu.

Badanie statystyczne

Cechy

Badaniu statystycznemu podlegają *cechy*. W populacji wszystkich Polaków badanymi cechami mogą być np. wzrost, masa ciała, wysokość zarobków, wykształcenie, wyznanie, preferencje wyborcze itp. Bada się *cechy zmienne*, czyli takie, których wartości mogą być różne dla różnych elementów populacji.Wymienione wyżej cechy są oczywiście zmienne. *Cechy stałe* mają identyczne wartości w całej populacji i jako takie nie są przedmiotem badań statystycznych. Przykładem cechy stałej w populacji wszystkich osób narodowości polskiej jest właśnie narodowość.

Cechy mierzalne

Cechy mierzalne (inaczej *ilościowe*) wyrażają się liczbami. Są nimi np. wzrost, masa ciała, liczba posiadanych dzieci itp.

Cechy niemierzalne

Cechy niemierzalne (inaczej *jakościowe*) nie wyrażają się wartościami liczbowymi. Są nimi np. wykształcenie, wyznanie, kolor oczu, preferencje wyborcze itp.

Badanie pełne

Badaniu pełnemu podlega cała populacja generalna. Prostymi przykładami badania pełnego są Narodowy Spis Powszechny, a także jakiekolwiek wybory.

Badanie częściowe

Badaniu częściowemu podlega podzbiór populacji generalnej zwany *próbą*. Próbę wybiera się z populacji generalnej, najczęściej w sposób losowy. Losowość zapewnia podobną strukturę próby i populacji generalnej, np. to, że w badanej próbie udział osób z wyższym wykształceniem będzie podobny jak w populacji generalnej, analogicznie ze strukturą wieku, zamieszkiwaniem w mieście czy we wsi itp. Często w mediach słyszymy stwierdzenia typu „badanie przeprowadzono w grupie reprezentatywnej 967 osób dorosłych".

Badanie częściowe przeprowadza się co najmniej z kilku powodów. Przede wszystkim dlatego, że wykonanie badania pełnego często jest zbyt kosztowne (spójrzmy na koszty przeprowadzenia wyborów) lub czasochłonne. Sama populacja generalna może być bardzo liczna, co uczyni przeprowadzenie badania pełnego wręcz niemożliwym. Inną przyczyną jest niszczący charakter badania (testy zderzeniowe samochodów, przydatność do spożycia puszki z konserwą rybną itp.).

Rozkład cechy

Celem badania statystycznego jest poznanie *rozkładu* badanej cechy w populacji generalnej. Jeśli badanie jest pełne, poznajemy *rozkład dokładny*. Jeśli badanie jest częściowe, poznajemy rozkład cechy w badanej próbie, czyli *rozkład przybliżony*. Metody statystyki matematycznej (oparte na rachunku prawdopodobieństwa) pozwalają ocenić, w jakim stopniu rozkład przybliżony zgodny jest z rozkładem dokładnym, który można by było uzyskać w badaniu pełnym. Metody te nie są przedmiotem niniejszego skryptu.

Statystyka opisowa

Statystyka opisowa zajmuje się wstępnym opracowaniem danych pochodzących z badania statystycznego bez stosowania metod rachunku prawdopodobieństwa. To opracowanie danych obejmuje ich odpowiednią prezentację, a także obliczenie różnych parametrów. Jeśli badanie jest pełne, to poprzestaje się na etapie statystyki opisowej. Jeśli badanie jest częściowe, to uzyskany w wyniku badania próby *rozkład empiryczny* może posłużyć do wnioskowania statystycznego, które należy już do statystyki matematycznej.

2

Statystyka opisowa jednej zmiennej

W niniejszym rozdziale omawia się badania statystyczne ze względu na jedną cechę, tzn. w próbie bądź w populacji generalnej bada się tylko jedną cechę. Jednoczesne badania dwóch cech omówione zostaną w następnym rozdziale.

2.1 Gromadzenie danych statystycznych

Szeregi statystyczne

Dane uzyskane w badaniu statystycznym mogą mieć postać ciągu uporządkowanego niemalejąco, który nazywa się *szeregiem szczegółowym*. Najczęściej dane pochodzące z szeregu szczegółowego zapisujemy w tabeli zwanej *szeregiem rozdzielczym*. Wśród szeregów rozdzielczych wyróżniamy *szeregi punktowe* i *szeregi przedziałowe*.

Rodzaje cech mierzalnych

Istnieją dwa rodzaje cech mierzalnych: *skokowe* i *ciągłe*. Dla każdego z nich szereg rozdzielczy sporządza się inaczej.

Cechy skokowe

Cechy te przyjmują wartości należące do pewnego zbioru skończonego lub przeliczalnego. Najczęściej jest to zbiór liczb całkowitych nieujemnych. Nie dopuszczają one stanów pośrednich. Np. rodzina nie posiada dzieci, albo posiada jedno dziecko, dwójkę dzieci itd.

Dla cech skokowych konstruuje się *szeregi punktowe*. W najprostszej wersji zawierają one wszystkie zaobserwowane wartości cechy oraz odpowiadające im liczebności, tj. liczby elementów badanej próby, dla których cecha przyjmuje konkretne wartości.

Cechy ciągłe

Cechy te mogą przyjmują (przynajmniej w teorii) wszystkie wartości należące do pewnego przedziału liczbowego. W praktyce jesteśmy ograniczeni dokładnością przyrządów pomiarowych. Otóż w badaniu wzrostu można oczywiście zaobserwować wartość 171,2387 cm. Jednak podawanie wzrostu z taką dokładnością nie ma sensu, wystarczy ograniczyć się do pełnych centymetrów.

Dla cech ciągłych buduje się *szeregi przedziałowe* dzieląc zbiór zaobserwowanych wartości cechy na pewną liczbę klas. Liczba ta zależy od liczności badanej próby. Każdej klasie przyporządkowuje się liczebność, która w tym przypadku informuje, dla ilu elementów badanej próby wartość cechy leży w danej klasie. Dla przykładu można podać płace pracowników pewnego zakładu uszeregowane następująco: poniżej 1000 zł, 1000 – 1200 zł, 1200 – 1400 zł itd.

W praktyce cechy skokowe od ciągłych odróżniamy przez powtarzalność wartości cechy. Jeśli jest duża, jak np. przy ocenach z danego przedmiotu czy liczbie dzieci w rodzinie, cechę traktuje się jak skokową. Jeśli w szeregu szczegółowym wartości cechy powtarzają się rzadko lub w ogóle się nie powtarzają, a badana próba jest wystarczająco liczna (w praktyce już ok. 25–elementowa), to cechę traktujemy jako ciągłą.

Konstrukcja szeregu punktowego

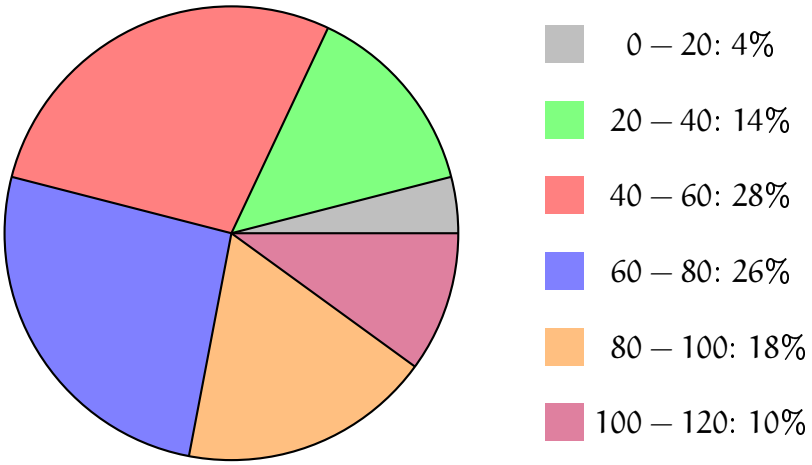
Przykład 2.1. Zapytano 25 rodzin o liczbę posiadanych dzieci otrzymując dane:

$$1, 3, 1, 2, 2, 1, 0, 3, 4, 1, 2, 3, 2, 5, 2, 3, 1, 0, 1, 2, 2, 4, 2, 6, 2.$$

Zbudować na ich podstawie szereg punktowy.

Badaną cechą jest liczba dzieci w rodzinie. Badaną próbą jest 25 rodzin. Oczywiście dane tego rodzaju cechuje duża powtarzalność, cecha jest skokowa. Widzimy, że wśród 25 badanych rodzin liczba dzieci waha się od 0 do 6. Zatem w powyższym ciągu danych jest 7 różnych wartości, tj. 0, 1, 2, 3, 4, 5 i 6. Liczbę różnych wartości cechy oznaczymy przez k, więc $k = 7$. Oznaczmy różne wartości cechy przez x_1, x_2 itp. Zatem mamy $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$, $x_7 = 6$. Dalej, niech n oznacza liczbę elementów badanej próby. W naszym przykładzie mamy więc $n = 25$. Każdej z danych $x_1, x_2, \ldots, x_k$ przypiszemy teraz odpowiednią *liczebność*, tj. liczbę elementów badanej próby, dla której cecha ma taką, a nie inną wartość. Skoro zatem 6 rodzin miało po jednym dziecku, to $n_2 = 6$ (danej $x_2 = 1$ odpowiada liczebność $n_2 = 6$). Pozostałe liczebności wyznaczamy podobnie.

Omawianie przykładu zakończymy sporządzeniem wykresu kołowego. Wykonuje się go identycznie jak dla szeregu punktowego, a wartości cechy zastępuje się klasami.



2.2 Miary tendencji centralnej

Miary te wskazują na środkowe wartości badanej cechy. Omówimy *średnią arytmetyczną*, *dominantę* i *medianę*. Oprócz tych miar zostaną także przedstawione *kwantyle*, wśród których szczególną rolę odgrywają *kwartyle* i *centyle*.

Średnia arytmetyczna

Z nauki szkolnej, a także z życia codziennego wiemy, że *średnia arytmetyczna* iluś liczb to suma tych liczb podzielona przez ich ilość. Podobnie oblicza się średnią w statystyce.

W szeregu szczegółowym o wartościach cechy $x_1, x_2, \ldots, x_n$ średnia arytmetyczna wyraża się wzorem

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{1}$$

W szeregu punktowym o wartościach cechy $x_1, x_2, \ldots, x_k$ z liczebnościami odpowiednio $n_1, n_2, \ldots, n_k$, średnia arytmetyczna wyraża się wzorem

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \,, \tag{2}$$

gdzie $n = n_1 + n_2 + \cdots + n_k$ jest liczbą elementów badanej próby. Równoważny wzór ma postać

$$\bar{x} = \sum_{i=1}^{k} x_i w_i \,, \tag{3}$$

gdzie $w_1, w_2, \ldots, w_k$ są częstościami (wagami) wartości cechy odpowiednio $x_1, x_2, \ldots, x_k$. Dlatego mówi się, że średnią arytmetyczną w szeregu punktowym oblicza się według *formuły ważonej*. Im większa częstość (waga) w_i, tym większy wpływ na średnią $\bar{x}$ wywiera odpowiednia wartość x_i.

W szeregu przedziałowym średnią arytmetyczną oblicza się analogicznie, zastępując w powyższych wzorach dane $x_1, x_2, \ldots, x_k$ środkami klas $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$. Popełniamy przy tym pewien błąd wynikający z zastąpienia rzeczywistych danych środkami klas. Jak zobaczymy, błąd ten nie jest na ogół zbyt duży i zazwyczaj mieści się w granicach dokładności przyrządu pomiarowego.

Przykład 2.3. Dla danych z Przykładu 2.1 obliczymy według wzoru (2) średnią liczbę dzieci w rodzinie:

$$\bar{x} = \frac{0 \cdot 2 + 1 \cdot 6 + 2 \cdot 9 + 3 \cdot 4 + 4 \cdot 2 + 5 \cdot 1 + 6 \cdot 1}{25} = \frac{55}{25} = 2{,}2\,.$$

Wykonamy teraz obliczenia w środowisku **R**. W tym celu wprowadzamy ciąg danych, który nazwiemy `dzieci`. Następnie wywołujemy funkcję `mean`, która wyznaczy średnią arytmetyczną naszego ciągu danych. W konsoli środowiska **R** wydajemy następujące komendy, zatwierdzając każdą z nich klawiszem **Enter**.

```
> dzieci=c(1,3,1,2,2,1,0,3,4,1,2,3,2,5,2,3,1,0,1,2,2,4,2,6,2)
> mean(dzieci)
[1] 2.2
```

Zadanie zostało wykonane. Poniżej zaprezentujemy dodatkowo kilka możliwości środowiska **R**. Poniżej wyświetla się szereg punktowy dla cechy `dzieci`.

```
> table(dzieci)
dzieci
0 1 2 3 4 5 6
2 6 9 4 2 1 1
```

Spróbujmy teraz dowiedzieć się jedynie tego, ile rodzin ma dwójkę dzieci.

```
> table(dzieci)[names(table(dzieci))==2]
2
9
```

Ciąg danych można uporządkować rosnąco.

```
> sort(dzieci)
 [1] 0 0 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 4 4 5 6
```

```
> kwantyl=function(dane,rzad){
+   n=sum(dane[,2]) # liczebność próby
+   k=1             # początkowy numer klasy kwantyla
+   sk=0            # początkowa poprzednia liczebność skumulowana
+   while(sk+dane[k,2]<n*rzad) {sk=sk+dane[k,2];k=k+1} # klasa kwantyla
+   return(dane[k,1]+(n*rzad-sk)*(dane[2,1]-dane[1,1])/dane[k,2])}
> Q=c(kwantyl(czasy,0.25),kwantyl(czasy,0.5),kwantyl(czasy,0.75))
> names(Q)=c("Q1","Me","Q3")
> Q
      Q1       Me       Q3
45.00000 63.07692 83.33333
```

Obliczymy jeszcze powyższe kwantyle w oparciu o szereg szczegółowy, na podstawie którego zbudowano szereg przedziałowy.

```
> czasy=c(12,17,23,25,27,28,28,30,36,41,42,48,49,50,52,52,53,54,55,57,
+         58,59,60,61,63,64,64,65,67,69,70,71,73,74,76,80,81,82,86,89,
+         90,90,93,95,96,101,108,110,112,116)
> Q=quantile(czasy,c(0.25,0.5,0.75))
> names(Q)=c("Q1","Me","Q3")
> Q
   Q1    Me    Q3
49.25 63.50 81.75
```

Widzimy, że zastosowanie szeregu przedziałowego zamiast szczegółowego rodzi pewne błędy.

2.3 Miary zróżnicowania cechy

Podstawową miarą zróżnicowania cechy jest *odchylenie standardowe*. Istnieją też pozycyjne miary zróżnicowania cechy, a wśród nich *odchylenie ćwiartkowe*.

Wariancja

Będziemy zakładać, że szeregi statystyczne występujące w tym skrypcie pochodzą z badania próby, a nie z badania pełnego. *Wariancją* w szeregu szczegółowym o wartościach cechy $x_1, x_2, \ldots, x_n$ (*wariancją z próby*) nazywamy liczbę obliczaną według wzoru

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Wariancją w szeregu punktowym o wartościach cechy $x_1, x_2, \ldots, x_k$ z liczebnościami odpowiednio n_1, $n_2, \ldots, n_k$ takimi, że $n_1 + n_2 \cdots + n_k = n$ i częstościami $w_1, w_2, \ldots, w_k$ nazywamy liczbę obliczaną według jednego dwóch równoważnych wzorów

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} (x_i - \bar{x})^2 n_i \,, \tag{6}$$

$$s^2 = \frac{n}{n-1} \sum_{i=1}^{k} (x_i - \bar{x})^2 w_i \,. \tag{7}$$

Aby wyznaczyć wariancję w szeregu przedziałowym, należy wartości cechy $x_1, x_2, \ldots, x_k$ zastąpić środkami klas $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$.

Wariancję można też obliczać według wzorów

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x})^2 n_i \quad \text{lub} \quad s^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 w_i \,. \tag{8}$$

które stosujemy, gdy badanie jest pełne, tj. obejmuje całą populację generalną. Mówimy wtedy o *wariancji z populacji*. Wielkości obliczane według wzorów (8) są nieco mniejsze niż te wyznaczone według wzorów (6), (7). Dla dużych wartości n różnice są niewielkie. Oprogramowanie statystyczne (m. in. środowisko **R**) posługuje się jednak wzorem (6). Również w tym skrypcie będziemy w ten sposób wyznaczać wariancję, jako że prezentowane przykłady omawiają badania prób, a nie badania pełne.

W środowisku **R** wariancję z próby obliczamy za pomocą funkcji `var` (od nazwy angielskiej *variance*). Wariancję z populacji wylicza się nieco inaczej.

```
> dzieci=c(1,3,1,2,2,1,0,3,4,1,2,3,2,5,2,3,1,0,1,2,2,4,2,6,2)
> # Wariancja z próby
> var(dzieci)
[1] 2.083333
> # Wariancja z populacji
> mean((dzieci-mean(dzieci))^2)
[1] 2
```

Odchylenie standardowe

Odchyleniem standardowym nazywamy pierwiastek kwadratowy z wariancji:

$$s = \sqrt{s^2} \,.$$

Informuje ono o tym, jak średnio różnią się wartości badanej cechy od średniej arytmetycznej. Im mniejsze odchylenie standardowe, tym bardziej wartości cechy skupiają

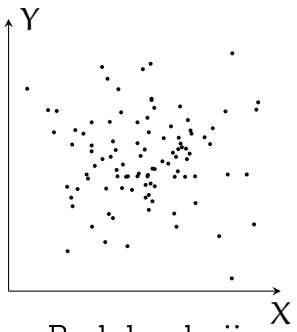
3

Statystyka opisowa dwóch zmiennych

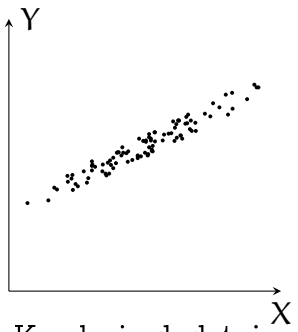
Dotychczas zajmowaliśmy się badaniem statystycznym dotyczącym tylko jednej cechy. Często wykonuje się też badania pod względem dwóch lub większej liczby cech. W niniejszym skrypcie ograniczymy się do ważnego przypadku dwóch cech. W takiej sytuacji jedna z badanych cech może zależeć w jakiś sposób od drugiej lub cechy są od siebie niezależne. Np. wydatki gospodarstw domowych zapewne zależą od ich dochodów, w pewnej mierze masa ciała może zależeć od wzrostu, kolor oczu dziecka (który jest cechą dziedziczną) zależy od koloru oczu rodzica itp. Natomiast masa ciała danej osoby oraz liczba jej dzieci z pewnością nie będą wykazywały żadnej zależności. Badanie współzależności cech rozdzielimy na dwa przypadki: w pierwszym z nich obie cechy będą mierzalne, a w drugim przynajmniej jedna cecha będzie niemierzalna.

3.1 Badanie współzależności cech mierzalnych

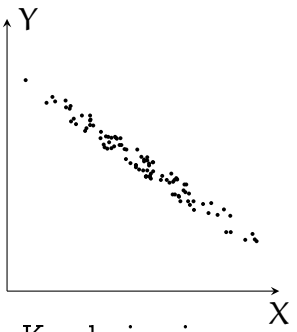
Załóżmy, że dana jest n–elementowa próba, w której bada się dwie cechy mierzalne umownie oznaczone przez X, Y. Niech $x_1, x_2, \ldots, x_n$ będą wartościami cechy X, a $y_1, y_2, \ldots, y_n$ wartościami cechy Y. Aby dostrzec możliwą współzależność obu cech, zaznaczamy w układzie współrzędnych punkty (x_i, y_i), gdzie $i = 1, 2, \ldots, n$.



Brak korelacji



Korelacja dodatnia



Korelacja ujemna

Na powyższym rysunku widać po lewej stronie cechy nieskorelowane — układ punktów nie wykazuje jakiejś regularności. Natomiast pośrodku i po prawej stronie można zauważyć wyraźną regularność układu punktów: skupiają się one wokół pewnych linii prostych. Cechy zobrazowane na tych rysunkach są niewątpliwie skorelowane: cecha Y zależy od cechy X. Na rysunku środkowym zależność ma kierunek dodatni (wraz ze wzrostem wartości cechy X następuje wzrost wartości cechy Y). Zależność na prawym rysunku ma charakter ujemny (wraz ze wzrostem wartości cechy X następuje spadek wartości cechy Y).

Przyjmijmy następujące oznaczenia:

$\bar{x}$ — średnia arytmetyczna cechy X,
$\bar{y}$ — średnia arytmetyczna cechy Y,
s_X — odchylenie standardowe cechy X,
s_Y — odchylenie standardowe cechy Y.

Kowariancją cech X i Y nazywamy wielkość

$$\operatorname{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\,. \tag{12}$$

Współczynnik korelacji

Współczynnikiem korelacji cech X i Y nazywamy liczbę

$$r = \frac{\operatorname{cov}(X,Y)}{s_X \cdot s_Y}\,.$$

W praktyce używa się wzoru równoważnego

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}\,. \tag{13}$$

Jeśli $r > 0$, to ewentualna zależność cech X i Y ma *kierunek dodatni* (wzrostowi wartości cechy X towarzyszy wzrost wartości cechy Y), a jeśli $r < 0$, to *ujemny*. (wzrostowi wartości cechy X towarzyszy spadek wartości cechy Y).

Własności współczynnika korelacji

1. Liczba r spełnia nierówność $-1 \leqslant r \leqslant 1$.

Obliczenia przeprowadzimy w tabeli.

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	1	0,3	−0,5	0,09	0,25	−0,15
2	1	2	−0,7	0,5	0,49	0,25	−0,35
3	1	2	−0,7	0,5	0,49	0,25	−0,35
4	0	2	−1,7	0,5	2,89	0,25	−0,85
5	4	1	2,3	−0,5	5,29	0,25	−1,15
6	2	2	0,3	0,5	0,09	0,25	0,15
7	3	1	1,3	−0,5	1,69	0,25	−0,65
8	1	0	−0,7	−1,5	0,49	2,25	1,05
9	0	1	−1,7	−0,5	2,89	0,25	0,85
10	3	3	1,3	1,5	1,69	2,25	1,95
Razem					16,10	6,50	0,50

Obliczamy współczynnik korelacji według wzoru (13) (str. 47):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{0,50}{\sqrt{16,10 \cdot 6,50}} = 0,049\,.$$

Wartość współczynnika r jest bardzo bliska zeru. Dlatego badane cechy są praktycznie nieskorelowane.

Rozwiązanie w środowisku **R**.

```
> kawa=c(2,1,1,0,4,2,3,1,0,3)
> dzieci=c(1,2,2,2,1,2,1,0,1,3)
> r=cor(kawa,dzieci)
> # Współczynnik korelacji liniowej Pearsona
> r
[1] 0.04887653
```

Przykład 3.2. Poniższe dane dotyczą miesięcznych dochodów i wydatków dziesięciu wybranych gospodarstw domowych.

Dochód w tys. zł	2,0	2,2	2,3	2,6	2,9	3,0	3,2	3,5	4,0	5,0
Wydatki w tys. zł	1,8	1,7	2,0	2,3	2,4	2,3	3,0	3,0	3,5	4,0

Zbadać czy cechy te są skorelowane.

zmiennych będzie objaśniana, a która objaśniająca, decydują intuicja i zdrowy rozsądek. Mówiąc np. o dochodach i wydatkach, to raczej wydatki zależą od dochodów, a nie na odwrót. Dysponujemy jakimś dochodem i w zależności od niego projektujemy nasze wydatki.

Jeśli X jest zmienną czasową, to zamiast o regresji mówimy o *trendzie liniowym*. Pisząc t zamiast X otrzymujemy równanie trendu liniowego o postaci

$$Y = at + b\,,$$

gdzie t oznacza czas.

Parametry strukturalne modelu regresji liniowej

Parametry strukturalne a, b prostej regresji (15) wyznaczamy według wzorów

$$a = \frac{\operatorname{cov}(X,Y)}{s_X^2}\,, \qquad b = \bar{y} - a\bar{x}\,. \tag{16}$$

W praktyce używa się wzoru równoważnego

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.$$

Przykład 3.3. Jak widzieliśmy w Przykładzie 3.2 (str. 49), współczynnik korelacji r pomiędzy dochodami X a wydatkami Y był bardzo bliski 1 ($r = 0{,}975$). Można więc powiedzieć, że wydatki badanych gospodarstw domowych zależą od ich dochodów (zob. też rysunek na str. 50). Wyznaczymy parametry strukturalne a, b modelu regresji.

Wszystkie potrzebne dane zawiera tabela w Przykładzie 3.2 (str. 50).

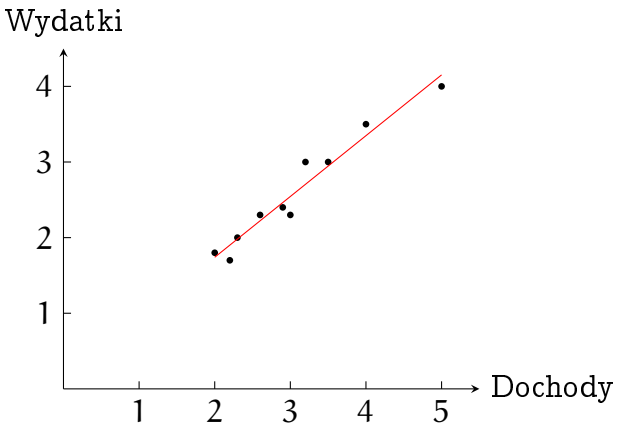
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{6{,}060}{7{,}5410} = 0{,}804\,,$$

$$b = \bar{y} - a\bar{x} = 2{,}60 - 0{,}804 \cdot 3{,}07 = 0{,}132\,.$$

Dlatego równanie prostej regresji ma postać

$$Y = 0{,}804X + 0{,}132\,.$$

Wzór ten opisuje w przybliżony sposób zależność wydatków od dochodów w badanych gospodarstwach domowych. Prostą regresji przedstawia poniższy rysunek.



Na podstawie równania regresji można przewidywać wartości zmiennej objaśnianej na podstawie wartości zmiennej objaśniającej. Mowa tu o *prognozowaniu*. Gdyby np. gospodarstwo domowe z badanej populacji osiągnęło dochód 2,1 tys. zł, to jego wydatki można określić na poziomie

$$Y = 0{,}804X + 0{,}132 = 0{,}804 \cdot 2{,}1 + 0{,}132 = 1{,}8204 \approx 1{,}8 \text{ tys. zł.}$$

Na zakończenie wyznaczymy równanie prostej regresji liniowej w środowisku **R**.

```
> dochody=c(2.0,2.2,2.3,2.6,2.9,3.0,3.2,3.5,4.0,5.0)
> wydatki=c(1.8,1.7,2.0,2.3,2.4,2.3,3.0,3.0,3.5,4.0)
> a=cov(dochody,wydatki)/var(dochody)
> b=mean(wydatki)-a*mean(dochody)
> # Równanie prostej regresji
> #  Współczynnik a
> a
[1] 0.8036069
> #  Współczynnik b
> b
[1] 0.1329267
```

Do wyznaczania równania regresji można też użyć specjalnej funkcji.

```
> lm(wydatki~dochody)

Call:
lm(formula = wydatki ~ dochody)

Coefficients:
```

4

Zbiór zadań

Zadania

Zadanie 1. W pewnym mieście przez 50 kolejnych dni notowano liczbę kolizji drogowych otrzymując dane:

$$2,1,1,4,2,3,2,0,5,0,0,0,0,1,2,0,2,1,1,2,1,2,2,1,1,$$
$$3,0,0,2,2,0,3,0,1,1,1,3,2,1,4,2,0,1,1,1,0,1,0,1,0.$$

a) Zbudować na podstawie powyższych danych szereg punktowy.

b) Obliczyć średnią arytmetyczną, medianę i dominantę.

c) Obliczyć kwartyle oraz decyle.

d) Obliczyć odchylenie standardowe.

e) Wyznaczyć przedział typowych wartości cechy.

f) Obliczyć klasyczny współczynnik zmienności, współczynnik skośności oraz klasyczny współczynnik asymetrii.

Zadanie 2. Zmierzono wzrost 80 studentek pewnego kierunku i otrzymano dane:

$$169,153,148,154,170,157,175,159,160,165,156,160,151,174,169,186,$$
$$170,159,161,173,163,164,159,158,169,155,159,148,163,161,171,171,$$
$$167,180,183,162,163,177,166,159,172,162,172,161,145,162,168,155,$$
$$145,154,153,167,166,153,179,162,163,162,166,168,165,163,161,177,$$
$$170,149,160,164,148,154,161,172,163,149,159,155,171,156,170,154.$$

Zbudować na ich podstawie szereg przedziałowy oraz wykonać polecenia b)–f) z Zadania 1.