Rozdział 1

Statystyka opisowa

1.1 Dane statystyczne

Statystyka zajmuje się zbieraniem, przetwarzaniem, przedstawianiem danych oraz wnioskowaniem na podstawie danych. Dane statystyczne dotyczą pewnej zbiorowości, zwanej *populacją*. Obserwuje się lub bada elementy tej zbiorowości, czyli *jednostki badania*. Interesują nas zwykle pewne *cechy* jednostek. Mogą to być cechy *ilościowe*, czyli wielkości liczbowe lub cechy *jakościowe*, czyli opisowo zdefiniowane własności jednostek.

Przykład. W populacji studentów WMiI UMK mogą nas interesować takie cechy jak: ocena z matematyki, wiek studenta, płeć itd. Pierwsze z dwóch cech są ilościowe, trzecia jest jakościowa. Dane mają postać tabelki, podającej cechy poszczególnych studentów:

student	ocena	wiek	płeć
1	3	25	M
2	5	41	K
3	4	28	M
...	...	...	...

□

Ogólnie, zapis wartości cech dla poszczególnych elementów populacji może mieć postać

jednostka	cecha $\mathcal{X}$	cecha $\mathcal{Y}$	cecha $\mathcal{Z}$	...
1	x_1	y_1	z_1	...
2	x_2	y_2	z_2	...
3	x_3	y_3	z_3	...
...	...	...	...	...

Jeśli, powiedzmy, cechy $\mathcal{X}$ i $\mathcal{Y}$ są ilościowe, to $x_1, x_2, \ldots$ i $y_1, y_2, \ldots$ są liczbami. Jeśli cecha $\mathcal{Z}$ jest jakościowa, to jej wartości $z_1, z_2, \ldots$ traktujemy po prostu jak *nazwy* lub umowne symbole. Oczywiście, można tu też użyć symboli liczbowych. W ostatnim przykładzie, moglibyśmy zakodować płeć pisząc, czysto umownie, $M = 1$, $K = 2$.

Badanie statystyczne może obejmować całą populację. Mówimy wtedy o badaniu *pełnym*. W tym przypadku, zebrane dane opisują wartości cech dla wszystkich jednostek populacji. Bywa tak, że dysponujemy tylko danymi dla pewnej części populacji. Mówimy w takiej sytuacji o badaniu *reprezentacyjnym*, a poddaną badaniu część populacji nazywamy *próbką*. Często (choć nie zawsze) próbkę wybieramy *losowo*. Jest jasne, że wnioskowanie o całej populacji na podstawie losowej próbki wymaga metod rachunku prawdopodobieństwa.

Statystyka opisowa zajmuje się opracowaniem zgromadzonych danych bez użycia rachunku prawdopodobieństwa. Po prostu przetwarzamy posiadane dane nie wnikając w to, czy dotyczą one całej populacji, czy też tylko próbki z populacji.

1.2 Pojedyncza cecha ilościowa

Zajmiemy się teraz najprostszą sytuacją. Przypuśćmy, że interesuje nas tylko jedna cecha ilościowa $\mathcal{X}$. Dane mają postać ciągu liczb:

$$x_1,\ x_2,\ \ldots,\ x_n,$$

gdzie n jest liczbą zbadanych (zaobserwowanych) jednostek, zaś x_i oznacza wartość cechy $\mathcal{X}$ dla i-tej spośród tych jednostek.

ŚREDNIA

Najprostszym sposobem „streszczenia” danych jest obliczenie średniej.

DEFINICJA. Średnią (lub wartością przeciętną) danych $x_1, x_2, \ldots, x_n$ nazywamy liczbę

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

□

Zauważmy, że $n\bar{x} = \sum x_i$ oraz $\sum (x_i - \bar{x}) = 0$.

Przykład. Wartość sprzedaży (w tys. zł) w pewnym sklepie w 10 kolejnych dniach wyniosła

$$12.0,\ 10.5,\ 17.3,\ 21.1,\ 14.7,\ 18.0,\ 11.5,\ 12.7,\ 10.9,\ 9.3.$$

Sumaryczna sprzedaż w ciągu 10 dni ma wartość 138. Średnia dzienna wartość sprzedaży jest równa 138/10=13.8. □

Średnia w Definicji 1.2 jest to tak zwana *średnia arytmetyczna*. Od interpretacji danych zależy, czy obliczanie średniej arytmetycznej jest uzasadnione, czy nie. Rozpatruje sie także inne rodzaje średnich. W naszych rozważaniach ważną rolę odgrywać będzie *średnia ważona*.

DEFINICJA. Średnią ważoną liczb $x_1, x_2, \ldots, x_n$ z odpowiadającymi im wagami $w_1, w_2, \ldots, w_n$ nazywamy liczbę

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}.$$

□

Jeśli wagi są jednakowe: $w_1 = w_2 = \cdots = w_n$, to średnia ważona jest po prostu średnią arytmetyczną. Pojęcie średniej ważonej wyjaśnimy na przykładzie.

Przykład. 100 kg paszy zawiera trzy składniki:

składnik	A	B	C
ilość (kg)	50	30	20
cena (zł/kg)	15	20	30

Za 100 kg paszy musimy zapłacić $50 \cdot 15 + 30 \cdot 20 + 20 \cdot 30 = 1950$ zł. Zatem cena 1 kg paszy jest równa $1950/100 = 19.5$ zł. Jest to średnia ważona cen 15, 20 i 30 z wagami, odpowiednio, 50, 30 i 20. Widać, że obliczenie zwykłej średniej arytmetycznej cen nie ma w tym przykładzie większego sensu. □

MEDIANA i KWANTYLE

Rozważamy ciąg n liczb

$$x_1,\ x_2,\ \ldots,\ x_n$$

Dla dowolnej liczby ξ, oznaczmy przez $N^{<}(\xi)$ i $N^{\leqslant}(\xi)$ liczbę wyrazów naszego ciągu, które są, odpowiednio, miejsze (niewiększe) od ξ. (niektóre wyrazy ciągu mogą się powtarzać; liczymy każdy wyraz tyle razy, ile razy się powtarza):

$$N^{<}(\xi) = \#\{1 \leqslant i \leqslant n : x_i < \xi\} = \text{liczba tych wyrazów } x_i\text{, które są } < \xi;$$

$$N^{\leqslant}(\xi) = \#\{1 \leqslant i \leqslant n : x_i > \xi\} = \text{liczba tych wyrazów } x_i\text{, które są } \leqslant \xi;$$

Sens definicji kwantyla rzędu $p = 100p\%$ jest taki: jest to taka liczba ξ_p, że *około* $100p\%$ danych leży poniżej ξ_p, czyli

$$\frac{N^{<}(\xi_p)}{n} \simeq \frac{N^{\leqslant}(\xi_p)}{n} \simeq p.$$

Nie możemy pominąć słówka ,,około", bo np nie zawsze jest liczbą całkowitą. Dlatego w formalnej definicji kwantyla, którą podamy poniżej, wymagamy spełnienia podwójnej nierówności.

DEFINICJA. Niech $0 < p < 1$. **Kwantylem** rzędu p nazywamy taką liczbę ξ_p, że

$$\frac{N^{<}(\xi_p)}{n} \leqslant p \leqslant \frac{N^{\leqslant}(\xi_p)}{n}.$$

Medianą ciągu $x_1, x_2, \ldots, x_n$ nazywamy kwantyl rzędu 0.5. Używamy oznaczenia

$$\xi_{0.5} = \mathrm{med}(x_1, x_2, \ldots, x_n).$$

□

Kwantyle $\xi_{0.25} = Q_1$ i $\xi_{0.75} = Q_3$ nazywamy **kwartylami** (pierwszym i trzecim).

Kwantyle $\xi_{0.1}$, $\xi_{0.2}$, $\ldots \xi_{0.9}$ nazywamy **decylami**.

Kwantyle $\xi_{0.01}$, $\xi_{0.02}$, $\ldots \xi_{0.99}$ nazywamy **centylami**.

Przykład. Rozważmy dane z Przykładu 1.2. Uporządkujmy nasze dane w kolejności rosnącej:

$$9.3,\ 10.5,\ \underline{10.9},\ 11.5,\ \underline{12.0,\ 12.7},\ 14.7,\ \underline{17.3},\ 18.0,\ 21.1.$$

Pierwszy kwartyl jest równy

$$\xi_{0.25} = 10.9,$$

bo $N^{<}(10.9) = 2$, $N^{\leqslant}(10.9) = 3$ i mamy $\frac{2}{10} \leqslant 0.25 \leqslant \frac{3}{10}$.

Podobnie, trzeci kwartyl jest równy

$$\xi_{0.75} = 17.3,$$

bo $N^{<}(17.3) = 7$, $N^{\leqslant}(17.3) = 8$ i mamy $\frac{7}{10} \leqslant 0.75 \leqslant \frac{8}{10}$.

Mediana nie jest tu wyznaczona jednoznacznie: dowolna liczba $\xi \in [12.0, 12, 7]$ jest medianą. Niekiedy wybiera się *środek* tego przedziału:

$$\mathrm{med} = 12.35$$

□

Uwaga. Czasami, dla uniknięcia niejednoznaczności, przyjmuje się nieco inne określenie kwantyli, ale sens jest podobny, jak w naszej definicji.

Średnia (arytmetyczna) i mediana są *miarami położenia*. Każda z nich w inny sposób precyzuje ,,wokół jakiej liczby dane się koncentrują". Najczęściej używana jest średnia.

Przykład. Dla rozpatrywanego wyżej ciągu 10 liczb

$$12.0,\ 10.5,\ 17.3,\ 21.1,\ 14.7,\ 18.0,\ 11.5,\ 12.7,\ 10.9,\ 9.3.$$

$$\text{średnia } = 12.8; \quad \text{mediana } = 12.35.$$

Rozpatrzymy ciąg 11 liczb, powstały z poprzedniego przez dołączenie liczby 62:

$$12.0,\ 10.5,\ 17.3,\ 21.1,\ 14.7,\ 18.0,\ 11.5,\ 12.7,\ 10.9,\ 9.3,\ \mathbf{62}.$$

Teraz

$$\text{średnia } = 20, \quad \text{mediana } = 12.7$$

Widać, że średnia jest *bardzo wrażliwa na ekstremalne wartości danych* (wyjątkowo duże lub małe). Mediana jest ,,bardziej odporna" na ekstremalne wartości.

Możemy sobie wyobrazić, że jedenastego dnia sprzedaż była, z jakiejś przyczyny, wyjątkowo wysoka. Z drugiej strony, możemy sobie wyobrazić, że dane zawierają błąd: jedenastego dnia sprzedaż była naprawdę równa 6.2. ,,Wrażliwość" średniej może być czasem zaletą, a czasem wadą! □

GRAFICZNE PRZEDSTAWIENIE DANYCH

DEFINICJA. Dystrybuantą empiryczną nazywamy funkcję daną wzorem

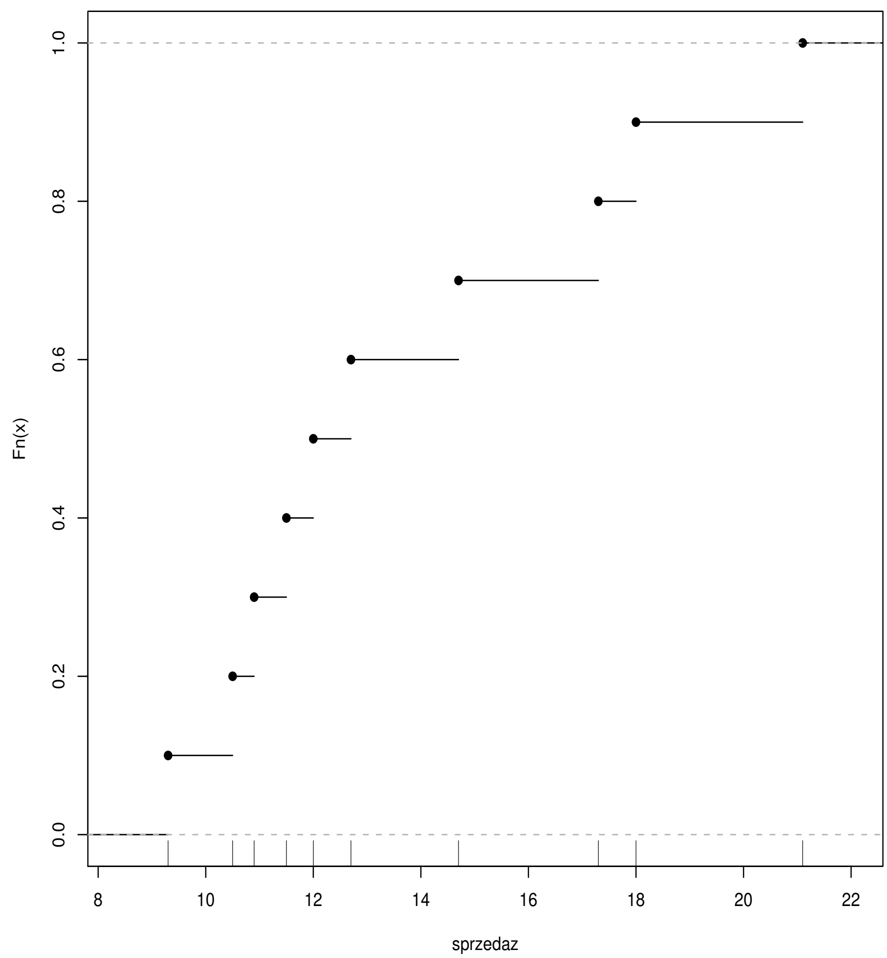
$$F_n(x) = \frac{N^{\leqslant}(x)}{n}.$$

Przypomnijmy, że $N^{\leqslant}(x)$ jest liczba tych danych spośród $x_1, x_2, \ldots, x_n$, które są $\leqslant x$, zaś n to liczba wszystkich danych. Dystrybuanta empiryczna jest funkcją ,,schodkową". Jeśli dane nie zawierają powtarzających się liczb, to wszystkie ,,skoki" mają wysokość $1/n$. Liczby $x_1, x_2, \ldots, x_n$ są poziomymi współrzędnymi punktów skoku.

Niemal wszystkie informacje o danych można odczytać z dystrybuanty empirycznej (z wyjątkiem kolejności w której dane zostały zapisane). Dla przykładu zauważmy, że mediana jest miejscem, w którym dystrybuanta empiryczna przekracza wysokość 0.5. Łatwo sformułować w podobny sposób definicję innych kwantyli.

Dla danych z Przykładu 1.2 wykres dystrybuanty empirycznej jest pokazany na Rys. 1.1. Zauważmy, że pionowe kreski na dole odpowiadają danym, czyli liczbom

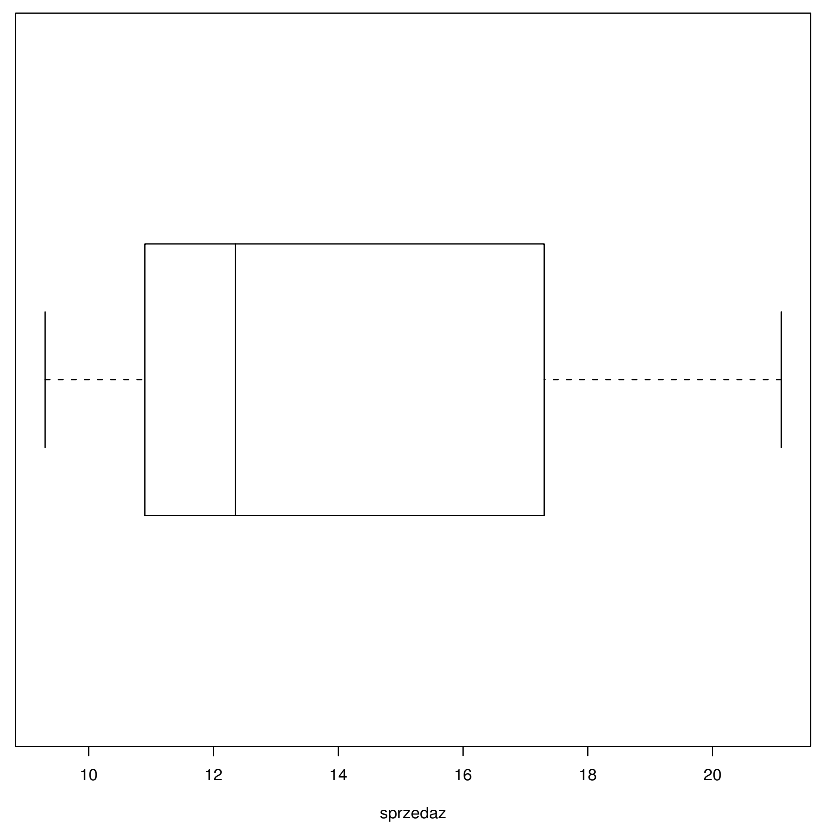
$$9.3,\ 10.5,\ 10.9,\ 11.5,\ 12.0,\ 12.7,\ 14.7,\ 17.3,\ 18.0,\ 21.1.$$



Rysunek 1.1: Dystrybuanta empiryczna.

Tak zwany wykres ,,pudełkowy” przedstawia dane w sposób streszczony, ale bardzo sugestywny. Pokazuje on najmniejszą i największą wartość danych, medianę i kwartyle. Zobaczmy to na danych z Przykładu 1.2. Przypomnijmy, że

$$\min = 9.3,\ Q_1 = 10.9,\ \mathrm{med} = 12.35,\ Q_3 = 17.3,\ \max = 21.1$$



Rysunek 1.2: Wykres ,,pudełkowy".

WARIANCJA i ODCHYLENIE STANDARDOWE

Zajmiemy się teraz liczbami, które informują o stopniu ,,rozproszenia" lub ,,rozrzutu" danych.

DEFINICJA. Wariancją danych $x_1, x_2, \ldots, x_n$ nazywamy liczbę

$$\tilde{S}^2 = \frac{1}{n}\big[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\big],$$

gdzie $\bar{x}$ jest średnią. Innymi słowy: wariancja jest *średnią kwadratów odchyleń od średniej.*

Odchyleniem standardowym nazywamy pierwiastek z wariancji:

$$\tilde{S} = \sqrt{\tilde{S}^2}.$$

$\square$

Inny wzór na obliczanie wariancji jest następujący:

$$\tilde{S}^2 = \frac{1}{n}\big[x_1^2 + x_2^2 + \cdots + x_n^2\big] - \bar{x}^2.$$

Wariancja jest równa *średniej kwadratów minus kwadrat średniej.*

Przykład. Wróćmy do rozpatrywanego wcześniej ciągu danych z Przykładu 1.2, opisujących wysokość sprzedaży:

$$12.0,\ 10.5,\ 17.3,\ 21.1,\ 14.7,\ 18.0,\ 11.5,\ 12.7,\ 10.9,\ 9.3.$$

Obliczmy wariancję:

$$\begin{aligned}\tilde{S}^2 = \frac{1}{10}\big[&(12.0-13.8)^2 + (10.5-13.8)^2 + (17.3-13.8)^2 + (21.1-13.8)^2 \\ &+ (14.7-13.8)^2 + (18.0-13.8)^2 + (11.5-13.8)^2 + (12.7-13.8)^2 \\ &+ (10.9-13.8)^2 + (9.3-13.8)^2\big] = 13.328\end{aligned}$$

Inaczej:

$$\begin{aligned}\tilde{S}^2 = \frac{1}{10}\big[&12.0^2 + 10.5^2 + 17.3^2 + 21.1^2 + 14.7^2 + 18.0^2 + 11.5^2 + 12.7^2 \\ &+ 10.9^2 + 9.3^2\big] - 13.8^2 = \frac{2037.68}{10} - 13.8^2 = 13.328\end{aligned}$$

Odchylenie standardowe:

$$\tilde{S} = \sqrt{13.328} = 3.65.$$

Zauważmy, wariancja jest wyrażona w „jednostkach kwadratowych”. W naszym przykładzie sprzedaż podana jest w tys. zł, a więc wariancja jest równa 13328000 zł^2. Odchylenie standardowe jest wyrażone w tys. zł: czyli jest równe 3650 zł. Oczywiście, łatwiej jest zinterpretować odchylenie standardowe. Jest to, mówiąc bardzo nieprecyzyjnie, „typowa” wartość rozrzutu danych wokoł średniej. $\square$

Uwaga. Rozpatruje się także inne miary rozrzutu. Na przykład można zdefiniować *odchylenie przeciętne* ciągu danych wzorem

$$\frac{1}{n}\big[|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|\big].$$

Ta wielkość wydaje się nawet prostsza, niż odchylenie standardowe, ale okazuje się mniej wygodna i jest rzadko używana.

Dla naszych danych dotyczących sprzedaży, odchylenie przeciętne jest równe

$$
\begin{aligned}
&= \frac{1}{10}\big[|12.0 - 13.8| + |10.5 - 13.8| + |17.3 - 13.8| + |21.1 - 13.8| \\
&\quad + |14.7 - 13.8| + |18.0 - 13.8| + |11.5 - 13.8| + |12.7 - 13.8| \\
&\quad + |10.9 - 13.8| + |9.3 - 13.8|\big] = 3.18
\end{aligned}
$$

Dość często używa się jeszcze innej miary rozrzutu. *Rozstępem międzykwartylowym* nazywamy liczbę

$$Q_3 - Q_1,$$

gdzie $Q_1 = \xi_{0.25}$ i $Q_3 = \xi_{0.75}$ oznaczją kwartyle. Na wykresie ,,pudełkowym" rozstęp międzykwartylowy jest długością pudełka.

W naszm przykładzie, rozstęp międzykwartylowy wielkości sprzedaży jest równy

$$17.7 - 10.9 = 6.8.$$

Najczęściej używaną miarą rozproszenia jest wariancja (lub, co jest równoważne, odchylenie standardowe). Jest to jednak miara wrażliwa na ekstremalne wartości danych. Odchylenie przeciętne jest mniej wrażliwe, a rozstęp międzykwartylowy jest najbardziej ,,odporny" na ekstremalne wartości.

Ogólnie, wybór miary rozproszenia (podopbnie jak wybór miary położenia) zależy od tego, *jaką informację* o danych chcemy przekazać.

TABLICA KONTYNGENCJI

Bardzo często dane wygodnie jest zapisać w postaci ,,tablicy kontyngencji", czyli ,,tablicy powtórzeń". Ogólnie, taka tablica ma postać:

wartość cechy	x_1	x_2	$\cdots$	x_k	razem
liczba jednostek	n_1	n_2	$\cdots$	n_k	n

Zauważmy, że, k oznacza tu *liczbę możliwych wartości cechy*, zaś n – liczbę jednostek. Oczywiście $n = n_1 + n_2 + \cdots + n_k$. Sposób budowania takiej tablicy wyjaśnimy na przykładzie.

Przykład. W grupie, składającej się z 20 studentów, oceny z egzaminu ze statystyki były następujące:

$$
\begin{aligned}
&2.0,\ 3.0,\ 3.5,\ 4.0,\ 4.5,\ 4.0,\ 5.0,\ 3.0,\ 3.0,\ 3.0,\\
&3.0,\ 4.0,\ 3.0,\ 3.5,\ 3.5,\ 2.0,\ 4.0,\ 3.5,\ 3.5,\ 5.0
\end{aligned}
$$

Możemy te dane zapisać w skróconej postaci, notując *ile razy* powtórzyły się poszczególne wartości:

ocena	2.0	3.0	3.5	4.0	4.5	5.0	razem
liczba studentów	2	6	5	4	1	2	20

Równoważnie, możemy podać w podobnej tabelce odpowiednie *ułamki* (procenty) całkowitej liczby studentów:

ocena	2.0	3.0	3.5	4.0	4.5	5.0	razem
ułamek studentów	0.10	0.30	0.25	0.20	0.05	0.10	1

Przejrzystym sposobem przedstawienia tablicy kontyngencji jest wykres ,,słupkowy", pokazany na Rys. 1.3.

Ponieważ statystyka interesują własności *zbiorowości*, a nie poszczególne jednostki, tablica kontyngencji zawiera pelną informację o danych. Pokażemy jak obliczyć wielkości, którymi się wcześniej zajmowaliśmy, na podstawie tablicy kontyngencji.

Dla danych z naszego przykładu średnią obliczymy tak:

$$\bar{x} = \frac{2 \cdot 2.0 + 6 \cdot 3.0 + 5 \cdot 3.5 + 4 \cdot 4.0 + 1 \cdot 4.5 + 2 \cdot 5.0}{20} = 3.5$$

Zauważmy, że *średnia arytmetyczna* wyjściowych 20 ocen jest tym samym, co *średnia ważona* 6 rórnych możliwych ocen, z wagami odpowiadającymi *liczbie powtórzeń*. Ten oczywisty fakt wyjaśnia, dlaczego w statystyce często posługujemy się średnią ważoną (Definicja 1.2).

Podobnie, wariancją jest *ważoną* średnią kwadratów odchyleń od średniej:

$$\begin{aligned}\tilde{S}^2 &= \frac{2}{20} \cdot (2.0 - 3.5)^2 + \frac{6}{20} \cdot (3.0 - 3.5)^2 + \frac{5}{20} \cdot (3.5 - 3.5)^2 \\ &+ \frac{4}{20} \cdot (4.0 - 3.5)^2 + \frac{1}{20} \cdot (4.5 - 3.5)^2 + \frac{2}{20} \cdot (5.0 - 3.5)^2 = 0.625,\end{aligned}$$

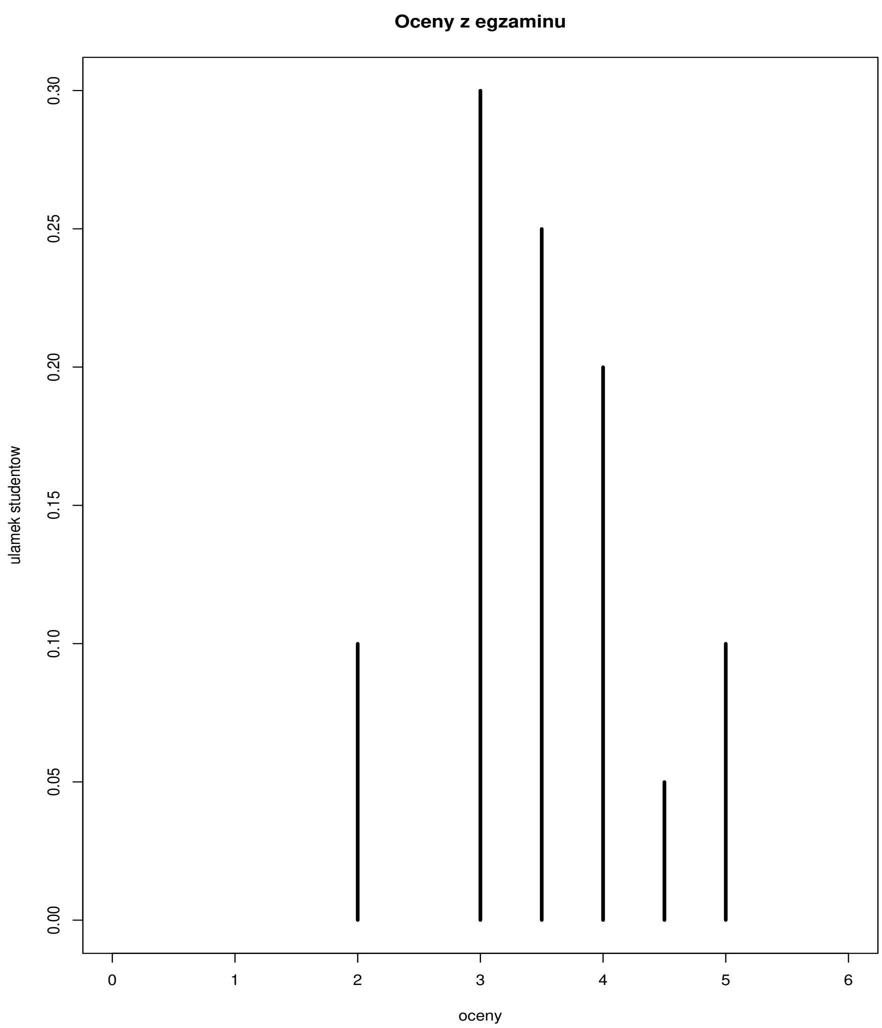
zatem

$$\tilde{S} = \sqrt{0.625} = 0.79.$$

Wreszcie, *medianą* ocen jest 3.5, bo liczba studentów o ocenie mniejszej niż 3.5, czyli 8, nie przekracza połowy, zaś liczba studentów o ocenie mniejszej lub równej 3.5, czyli 13, przekracza połowę. Innymi słowy, przy oznaczeniach wprowadzonych w Definicji 1.2 mamy

$$\frac{N^{<}(3.5)}{20} = \frac{8}{20} \leqslant 0.5 \leqslant \frac{N^{\leqslant}(3.5)}{20} = \frac{13}{20}.$$

□



Rysunek 1.3: Wykres ,,słupkowy".

Podsumujmy: dla tablicy kontyngencji wartość średnią i wariancję obliczamy według wzorów:

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n_1 + n_2 + \cdots + n_k} = \frac{n_1}{n} x_1 + \cdots + \frac{n_k}{n} x_k,$$

$$\tilde{S}^2 = \frac{n_1}{n}(x_1 - \bar{x})^2 + \cdots + \frac{n_k}{n}(x_k - \bar{x})^2.$$

Są to *te same wielkości*, które określiliśmy w Definicjach 1.2 i 1.2. Wzory wyglądają inaczej, bo symbole x_i w tablicy kontyngencji oznaczają coś trochę innego, niż poprzednio. Dla danych indywidualnych x_i oznaczało wartość cechy dla i-tej jednostki. W tablicy kontyngencji x_i jest i-tą spośród różnych wartości cechy, ktora to wartość występuje u n_i jednostek.

SZEREG PRZEDZIAŁOWY

Często dane nie zawierają wartości cechy $\mathcal{X}$ dla pojedynczych jednostek, tylko informację o tym, ile jednostek ma cechę w pewnych przedziałach wielkości. Jest to tak zwany ,,szereg rozdzielczy przedziałowy”

Przykład. Wielkości mieszkań w pewnym osiedlu (w m^2), (po uporządkowaniu) podaje następująca tabelka:

33.46	34.71	34.96	35.29	35.51	37.82	38.25	38.37	38.81	39.67
40.26	40.87	41.49	41.81	42.33	42.35	43.41	43.90	44.00	44.53
44.71	45.64	46.94	47.33	47.37	47.65	48.32	48.44	48.72	49.40
49.95	50.10	50.16	50.85	50.94	51.43	51.50	51.55	51.73	51.83
51.89	51.91	51.96	51.97	52.00	52.46	53.99	54.64	55.05	55.35
55.62	56.01	56.16	56.30	56.34	56.45	57.03	57.22	57.92	58.72
58.80	59.07	59.34	59.49	59.90	60.95	61.17	63.04	64.11	66.90
68.51	68.73	71.20	71.74	72.06	72.52	73.17	73.73	75.33	75.69
77.53	78.08	78.21	79.59	80.90	81.32	84.00	85.38	86.82	88.97
90.07	93.64	95.55	96.98	100.35	104.22	107.00	112.70	115.70	118.11

Dla przypomnienia obliczmy średnią, medianę i kwantyle:

$$\text{średnia} = 60.66$$

$$Q_1 = 47.58, \text{ med} = 55.49, \ Q_3 = 72.18$$

Zauważmy, że mediana i kwantyle nie są tu jednoznacznie wyznaczone. Program komputerowy wybrał konkretne liczby mieszczące się pomiędzy podkreślonymi danymi (w pewien sposób, który nie jest dla nas zbyt ważny).

Pogrupujmy te dane w następujący sposób: zanotujmy, ile mieszkań mieści się w przedziałach wielkości po 10 m^2. Przedstawia to tabelka:

przedział wielkości	liczba mieszkań
30–40	10
40–50	21
50–60	34
60–70	7
70–80	12
80–90	6
90–100	4
100–110	3
110–120	3
razem	100

Oczywiście, należałoby się umówić, do jakiego przedziału zaliczmy mieszkania o metrażu 40, 50, ... lub 110. Powiedzmy, że zawsze zaliczamy je do przedziału ,,niższego".

Postać i interpretacja szeregu przedziałowego są, oczywiście, podobne jak dla omawianej wcześniej tablicy kontyngencji. Zwróćmy jednak uwage na istotną różnicę. Podsumowując dane w postaci szeregu przedziałowego *tracimy część informacji*. Z ostatniej naszej tabelki *nie możemy* się dowiedzieć na przykład, ile jest mieszkań o metrażu 30–35 (na podstawie pełnych danych wiemy, że jest ich 3). Na podstawie tejże tabelki nie możemy *dokładnie* obliczyć średniej oryginalnych danych. □

Ogólnie, przedziałowy szereg rozdzielczy ma postać

cecha $\mathcal{X}$	liczba jednostek
x_0–x_1	n_1
x_1–x_2	n_2
...	...
x_{n-1}–x_k	n_k
razem	n

Tutaj n_i oznacza liczbę jednostek, dla których cecha $\mathcal{X}$ ma wartość w przedziale $(x_{i-1}, x_i]$. Liczba przedziałów jest oznaczona przez k. Oczywiście. $n = n_1 + \cdots + n_k$. Taką tabelkę możemy wyprodukować sami na podstawie danych indywidualnych. Czasem po prostu jest to jedyny opis danych, jakim dysponujemy.

Histogram jest wygodnym graficznym przedstawieniem przedziałowego szeregu rozdzielczego.

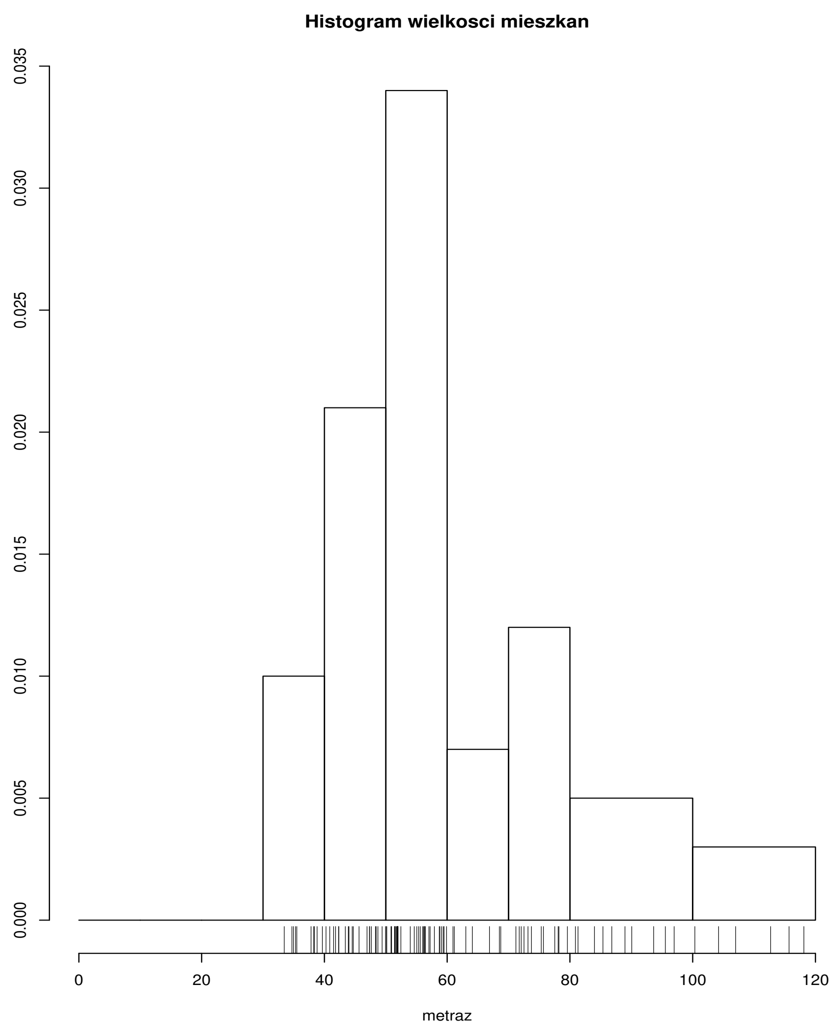
Nad każdym odcinkiem $[x_{i-1}, x_i]$ rysujemy prostokąt o polu równym n_i/n (w odpowiednio dobranych jednostkach. Wysokość tego prostokąta jest więc równa $n_i/n(x_i - x_{i-1})$. Jeśli przedziały $[x_{i-1}, x_i]$ nie są równej długości, wtedy istotne jest, żeby prostokąty histogramu miały *pole* (a nie *wysokość*) równą n_i/n.

Przykład. Przedstawmy dane mieszkaniowe z poprzedniego przykładu w postaci nieco ,,bardziej zwięzłego" szeregu, z mniejszą liczbą przedziałów (nierównej długości):

przedział wielkości	liczba mieszkań
30–40	10
40–50	21
50–60	34
60–70	7
70–80	12
80–100	10
100–120	6
razem	100

Tej tabelce odpowiada histogram przedstawiony na Rys 1.3. Kreski u dołu pokazują wielkości poszczególnych mieszkań (dane indywidualne).

Zauważmy, że pole ,,szerokiego" prostokąta prostokąta nad przedziałem 100–120 jest rowne $6/100$ i jest równe sumie pól dwóch prostokątów o polu po $3/100$ każdy. Jak zmieniłby się obrazek, gdybyśmy zbudowali histogram na podstawie poprzedniej, bardziej szczegółowej tabelki? $\square$



Rysunek 1.4: Histogram.