STATYSTYKA STOSOWANA

Tadeusz Inglot

STATYSTYKA STOSOWANA

Krótki kurs

Oficyna Wydawnicza GiS
Wrocław 2020

Tadeusz Inglot
Wydział Matematyki
Politechnika Wrocławska
tadeusz.inglot@pwr.edu.pl

Projekt okładki:
IMPRESJA Studio Grafiki Reklamowej

Skład książki w systemie LaTeX wykonał autor.

ISBN 978-83–62780–76–1

Wydanie I, Wrocław 2020
Oficyna Wydawnicza GiS, s.c., www.gis.wroc.pl
Druk i oprawa: Drukarnia I-BiS sp. z o.o., sp. kom.

Spis treści

Przedmowa

W latach 2001-2019 prowadziłem w Politechnice Wrocławskiej przedmiot Statystyka Stosowana w wymiarze 15 godzin wykładu i 15 godzin ćwiczeń dla studentów kierunków technicznych, którzy nie mieli żadnego przygotowania z rachunku prawdopodobieństwa. Ze względu na skromny wymiar czasowy wykładu, zdecydowałem się na niestandardowe wprowadzenie podstawowych pojęć. Zrezygnowałem z aksjomatycznej definicji prawdopodobieństwa i oparłem się na intuicyjnej, choć nieścisłej matematycznie, definicji częstościowej. To pozwoliło względnie szybko przejść do pojęć statystycznych i omówić najważniejsze klasyczne modele oraz metody wnioskowania statystycznego. Ponadto ułatwiło słuchaczom zrozumienie i przyswojenie materiału oraz wytworzenie prawidłowych intuicji w krótkim czasie.

Wykład w dużej mierze oparłem na podręczniku J. Koronackiego i J. Mielniczuka *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Dlatego w niniejszych notatkach przedstawiłem tylko tyle materiału, ile rzeczywiście da się wyłożyć w czasie 15–20 godz., ograniczając się do minimalnej liczby pojęć i faktów. Z tego samego powodu w wielu dowodach i wyprowadzeniach świadomie pominąłem szczegóły, a nawet zrezygnowałem ze ścisłości, skupiając uwagę słuchacza i czytelnika na istocie rozumowania. Do tekstu wykładu dołączyłem zestaw zadań na ćwiczenia, podzielony na kolejne rozdziały. Zawiera tyle zadań, aby zdecydowaną większość z nich dało się rozwiązać ze studentami na ćwiczeniach audytoryjnych.

Czytelnik, zainteresowany bardziej dogłębnym studium, ma możliwość uściślenia, uzupełnienia i rozszerzenia wiadomości w oparciu o wspomniany podręcznik, czy o wiele innych opracowań za statystyki matematycznej. Materiał z rozdziałów oznaczonych $*$ nie zawsze udało mi się wyłożyć.

Wrocław, kwiecień 2020 r. Tadeusz Inglot

1

Dane. Zmienne losowe

Podstawową metodą opisu zjawisk otaczającego nas świata jest ich wielokrotna obserwacja. Spotykamy się ze zjawiskami, dla których kolejne wyniki obserwacji (wykonanych w tych samych warunkach) są takie same. Mówimy wtedy o zjawisku *deterministycznym*. Jednak znacznie częściej mamy do czynienia z sytuacją, w której każda kolejna obserwacja daje „nieco" inny wynik, mimo iż staramy się przeprowadzić ją w tych samych warunkach i nie widać przyczyn tych różnic. Mówimy wtedy, że opisywane zjawisko jest *losowe*. Zmienność obserwacji może być niewielka, ale też bardzo duża, znacząca. Jeśli z naszego punktu widzenia zmienność jest pomijalnie mała, możemy ją zaniedbać i mówić o powtarzalności obserwacji wykonanych w tych samych warunkach. Podejście statystyczne stosujemy, gdy obserwowane zjawisko wykazuje zmienność, której nie możemy lub nie chcemy zaniedbać. Zatem ciąg obserwacji wielkości wykazującej zmienność tworzy *dane*.

Powyżej przyjęliśmy, że obserwacje dotyczą pewnej interesującej nas wielkości liczbowej (parametru, cechy) danego zjawiska, którą będziemy nazywać zmienną losową. Zmienność (losowość) lub jej brak wynikają z jednej strony z natury obserwowanego zjawiska, a z drugiej strony z przyjętej lub wymaganej dokładności obserwacji danej wielkości. I te dwa aspekty decydują, z którą sytuacją mamy do czynienia. Zilustrujemy to na kilku prostych przykładach.

1. Pomiar „metrówką" szerokości:
a) stołu z dokładnością do 1 cm – brak zmienności i za każdym razem wynik 80 cm;
b) stołu z dokładnością do 1 mm – pojawia się niewielka zmienność ze względu na konstrukcję „metrówki" i wynikające stąd błędy odczytu;
c) budynku z dokładnością do 1 cm – zmienność wydaje się oczywista i jest dosyć duża, gdyż „metrówkę" trzeba wielokrotnie przekładać i raczej nie dziwimy się wynikom (w cm), takim jak: 903, 899, 896, itp. W

tym przypadku zmienność powodują błędy pomiaru i przyjęta dokładność nieodpowiednia do metody pomiaru.

2. Bieżący kurs € (w zł) w kantorze:
a) na wyświetlaczu podającym co sekundę średni kurs na giełdzie. Może to wyglądać następująco: 4.2764, 4.2761, 4.2758 itd. Zmienność wynika z podawania wyników z dokładnością do 0.01 gr.;
b) przy zakupie 100 € zapłacimy 429 zł po stałym kursie przez cały dzień i zmienność widoczna w pkt. a) nie ma tu praktycznego znaczenia.

3. Chwilowe zużycie paliwa przez silnik (w l/100 km) na wyświetlaczu w samochodzie może wyglądać następująco: 12.3, 21.6, 15.5, 7.3, 5.4, 5.1, 5.9, 0.0, itd. Widać ogromną zmienność, wynikającą oczywiście z natury pracy silnika przy zmiennym obciążeniu, bardzo zmiennych warunkach i wielu innych istotnych czynnikach. Natomiast dokładność rejestracji wyników jest adekwatna do ich ewentualnego wykorzystania.

4. Rzuty kostką do gry mogą dać następujące wyniki: 3, 1, 1, 4, 6, 4, itd. Również tutaj mamy dużą zmienność, wynikającą z natury zjawiska, przy naturalnej dokładności zapisu wyników. Gdyby jednak interpretować te wyniki jako wypłatę w dziesiątkach groszy, przy wcześniej wnoszonej opłacie 35 gr za rzut, i wygraną podawać w zaokrągleniu do pełnych złotych, zmienność zniknie.

Powyższe przykłady pokazują, że określenia „wahanie losowe", „przypadkowa zmienność", czy „losowy błąd" są tylko różnymi sposobami nazywania zmienności obserwowanej wielkości.

Podsumujmy nasze rozważania.

• Obserwowaną wielkość liczbową wykazującą zmienność, nazywamy *zmienną losową* (krócej zmienną) lub *cechą* (interesującego nas „obiektu") i oznaczamy dużą literą, zwykle z końca alfabetu, np. X, Y, itp.

• Ciąg kolejnych obserwacji wartości zmiennej losowej X nazywamy *próbą* i oznaczamy $x_1, ..., x_n$. Używamy małych liter, odpowiadających dużej literze będącej nazwą zmiennej losowej. W statystyce rozważa się także równoczesne obserwowanie kilku zmiennych losowych, np. dwóch zmiennych losowych X oraz Y. Wtedy próba jest ciągiem par $(x_1, y_1), ...,$

(x_n, y_n). Termin statystyczny *próba* jest zatem tożsamy z potocznym określeniem *dane*.

• *Liczebnością próby* nazywamy liczbę obserwacji występujących w próbie. W postępowaniu statystycznym liczebność próby ma istotne znaczenie dla ilości informacji o zmiennej losowej, którą ta próba zawiera, i w konsekwencji wpływa na dalsze wnioskowanie.

Jeśli zbiór możliwych wartości zmiennej losowej X wypełnia przedział (ograniczony lub nie), to X nazywamy *zmienną ciągłą* (lub typu ciągłego, gdyż nie chodzi tu o ciągłość rozumianą jak w analizie funkcji jednej zmiennej). Jeśli zmienna losowa może przyjmować co najwyżej skończenie wiele lub przeliczalnie wiele „izolowanych" wartości, to mówimy o *zmiennej dyskretnej*. Jeśli obserwacje nie są liczbami i określamy je w postaci opisowej, to mówimy o zmiennej jakościowej. Numerując poszczególne kategorie, łatwo przekształcamy ją na zmienną dyskretną.

2

Wstępne opracowanie danych

Niech $x_1, \ldots, x_n$ będzie próbą o liczebności n obserwacji zmiennej losowej X.

Z powodu zmienności X spodziewamy się, że liczby x_i będą różne, choć przecież niektóre mogą się powtarzać (nawet dla zmiennej ciągłej, bo wyniki notujemy z określoną liczbą miejsc dziesiętnych). Zastanówmy się, jak z takiego ciągu liczb wydobyć pewne podstawowe informacje? Na pewno interesuje nas w jakiej części prostej koncentrują się obserwacje i jak duża jest ich zmienność.

Najpierw jednak uporządkujemy obserwacje $x_1, ..., x_n$ od najmniejszej do największej, wypisując powtarzające się obserwacje tyle razy, ile razy wystąpiły w próbie. Tak otrzymany ciąg nazywamy *szeregiem pozycyjnym* i oznaczamy $x_{(1)}, x_{(2)}, ..., x_{(n)}$. W szczególności mamy więc $x_{(1)} = \min\{x_1, ..., x_n\}$ oraz $x_{(n)} = \max\{x_1, ..., x_n\}$. Wyrazy tego ciągu nazywamy *statystykami pozycyjnymi*.

a) Wskaźniki położenia. Najbardziej naturalnym wskaźnikiem położenia jest średnia z próby.

Definicja 1. *Średnią z próby* nazywamy liczbę $\overline{x} = \dfrac{x_1 + ... + x_n}{n}$, czyli średnią arytmetyczną wszystkich obserwacji.

Przykład 1. Zanotowano czas T dojazdu samochodem do pracy (w min.) w kolejnych 15 dniach i otrzymano: 25, 21, 28, 25, 52, 27, 28, 28, 24, 48, 21, 28, 27, 30, 22. Mamy zatem $\overline{t} = (25+21+...+22)/15 = 28.93$. Widać, że tylko trzy razy czas dojazdu przekroczył 28.93 min., a w 12 przypadkach był mniejszy. Więc liczba 28.93 nie bardzo jest „środkiem koncentracji" naszych obserwacji. Przyczyną są dwie *obserwacje odstające* 48 i 52, zapewne związane z nieprzewidzianymi utrudnieniami w ruchu.

Definicja 2. *Medianą z próby*, oznaczaną x_{med}, nazywamy „środkową" obserwację w szeregu pozycyjnym, a dokładnie

$$x_{med} = \begin{cases} x_{((n+1)/2)}, & \text{gdy } n \text{ jest nieparzyste,} \\ (x_{(n/2)} + x_{((n+2)/2)})/2, & \text{gdy } n \text{ jest parzyste.} \end{cases}$$

k-tą średnią obciętą z próby, gdzie $k = 1, 2, ..., \lfloor (n-1)/2 \rfloor$, nazywamy liczbę

$$\overline{x}_k = \frac{x_{(k+1)} + x_{(k+2)} + ... + x_{(n-k)}}{n - 2k},$$

czyli średnią arytmetyczną po usunięciu k początkowych i k końcowych wyrazów szeregu pozycyjnego.

Mediana z próby i k-ta średnia obcięta z próby należą do wskaźników odpornych na dane odstające. Zobaczmy, jak to wygląda w naszym przykładzie.

Przykład 1 cd. Szereg pozycyjny ma postać: 21, 21, 22, 24, 25, 25, 27, 27, 28, 28, 28, 28, 30, 48, 52. Ponieważ n jest liczbą nieparzystą, to mediana z próby jest ósmym wyrazem szeregu pozycyjnego, czyli mamy $x_{med} = 27$. Natomiast $\overline{x}_1 = 27.77$ oraz $\overline{x}_2 = 26.55$. Widać, że wszystkie trzy wskaźniki mają zbliżone wartości, a zatem zawierają praktycznie tę samą informację o obserwowanej zmiennej. Ale są istotnie mniejsze od średniej z próby.

Mimo wskazanej wyżej wady, średnia z próby jest najczęściej używanym wskaźnikiem położenia.

b) Wskaźniki rozproszenia. Jest oczywiste, że wskaźniki położenia nic nie mówią o skali zmienności obserwowanej zmiennej losowej. Potwierdza to prosty przykład.

Przykład 2. Pięć obserwacji zmiennej losowej X dało wyniki: 1.2, 1.5, 0.9, 1.3, 1.1, a pięć obserwacji innej zmiennej Y wyniki: 1.8, 0.5, 1.0, 0.7, 2.0. Łatwo sprawdzić, że $\overline{x} = \overline{y} = 1.2$, ale wahania Y wokół tej średniej są zdecydowanie większe niż X.

Najczęściej używanym wskaźnikiem rozproszenia jest wariancja z próby.

Definicja 3. *Wariancją z próby* nazywamy liczbę

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

Liczbę $s = \sqrt{s^2}$ nazywamy *odchyleniem standardowym z próby.*

Odchylenie standardowe z próby wyraża się w tych samych jednostkach co obserwowana zmienna i to uzasadnia jego wprowadzenie.

Twierdzenie 1. Dla dowolnej liczby rzeczywistej c

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

Dowód twierdzenia 1 pozostawiamy czytelnikowi jako ćwiczenie (zad. 1). Twierdzenie 1 pozwala ułatwić obliczanie s^2 przez wybór dogodnej liczby c. Jednym z możliwych wyborów jest $c = 0$. Wtedy

$$s^2 = \frac{x_1^2 + ... + x_n^2}{n} - \overline{x}^2.$$

Mimo prostej postaci, ostatni wzór zwykle nie ułatwia obliczeń.

Przykład 2 cd. Wariancje z obu prób obliczymy bezpośrednio z definicji $s_X^2 = (0 + 0.3^2 + 0.3^2 + 0.1^2 + 0.1^2)/5 = 0.04$, $s_X = 0.2$ oraz $s_Y^2 = (0.6^2 + 0.7^2 + 0.2^2 + 0.5^2 + 0.8^2)/5 = 0.356$, $s_Y \approx 0.6$. Zatem możemy powiedzieć, że zmienność Y jest ok. 3 razy większa niż X.

Przykład 1 cd. Dla wygody obliczeń weźmiemy $c = 29$ i zastosujemy twierdzenie 1. Wtedy $s^2 = \frac{1}{15}(4^2 + 8^2 + 1^2 + 4^2 + 23^2 + 2^2 + 1^2 + 1^2 + 5^2 + 19^2 + 8^2 + 1^2 + 2^2 + 1^2 + 7^2) - 0.07^2 = 75.800 - 0.0049 \approx 75.80$, $s \approx 8.71$. Wariancja z próby jest „wrażliwa" na dane odstające. W naszym przykładzie, pomijając dwie największe obserwacje, otrzymalibyśmy odpowiednio wartości 8.06 i 2.84.

Definicja 4. *Kwartylem dolnym* Q_1 z próby nazywamy medianę z pierwszej części szeregu pozycyjnego, tj. na lewo od x_{med}, ale bez niej. Podobnie określamy *kwartyl górny* Q_3. *Rozstępem międzykwartylowym* nazywamy liczbę $IQR = Q_3 - Q_1$. Rozstęp międzykwartylowy jest odporny na dane odstające.

Przykład 1 cd. Mamy $Q_1 = 24$ (czwarta statystyka pozycyjna, bo na lewo od x_{med} jest 7 obserwacji) i $Q_3 = 28$ (dwunasta statystyka pozycyjna), czyli $IQR = 4$ i niewiele różni się od odchylenia standardowego z próby po usunięciu danych odstających, ale jest ponad dwukrotnie mniejszy od s.

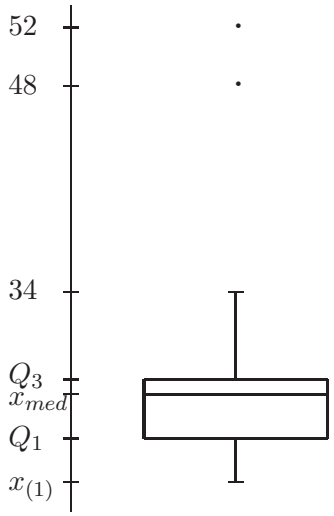
Do mierzenia rozproszenia używa się także *rozstępu z próby* $r = x_{(n)} - x_{(1)}$ oraz *odchylenia średniego z próby* $d_1 = (|x_1 - \overline{x}| + ... + |x_n - \overline{x}|)/n$.

c) Wykres ramkowy. W celu szybkiej oceny „rozkładu" jednej zmiennej losowej lub orientacyjnego porównania „rozkładów" dwóch lub więcej zmiennych losowych wyniki dotychczasowej analizy są przedstawiane graficznie. Jednym ze sposobów takiej wizualizacji jest *wykres ramkowy.*

Oś liczbową przedstawiamy jako prostą pionową i na niej zaznaczamy $x_{(1)}, Q_1, x_{med}, Q_3, x_{(n)}$. Następnie rysujemy prostokąt o dowolnej szerokości, dolnej podstawie na wysokości Q_1 i górnej podstawie na wysokości Q_3 i przekreślamy otrzymany prostokąt linią poziomą na wysokości x_{med}. Następnie w środkach obu podstaw prostokąta wyciągamy na zewnątrz prostokąta „wąsy" do $x_{(1)}$ w dół i do $x_{(n)}$ w górę, o ile długość każdego „wąsa"nie przekracza półtora IQR. W przeciwnym razie rysujemy „wąs" („wąsy") długości półtora IQR. Natomiast pozostałe dane traktujemy jako odstające i każdą z nich zaznaczamy kropką odpowiednio poniżej lub powyżej danego „wąsa".

Wykres ramkowy dla przykładu 1 jest przedstawiony na rys. 1.

Przykład 3. ([GR], str. 144) Pobrano próbki betonu tej samej klasy z dwóch betoniarni i zbadano ich wytrzymałość na ściskanie. Wyniki (po uporządkowaniu rosnąco) w daN/cm^2 były następujące: betoniarnia A



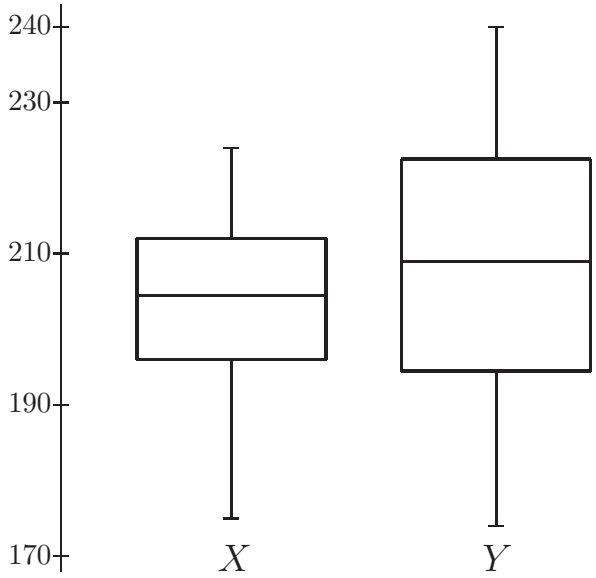
Rys. 1

(zmienna losowa X, $n_1 = 26$) 175, 176, 180, 189, 190, 190, 196, 196, 198, 199, 200, 200, 204, 205, 205, 206, 206, 208, 210, 212, 212, 213, 216, 219, 219, 224; betoniarnia B (zmienna losowa Y, $n_2 = 29$) 174, 181, 185, 186, 188, 193, 194, 195, 195, 195, 201, 201, 202, 203, 209, 209, 211, 215, 217, 217, 217, 220, 225, 225, 229, 229, 231, 233, 240. Dla betoniarni A mamy $Q_{1X} = 196$, $x_{med} = 204.5$, $Q_{3X} = 212$, $IQR_X = 16$, a dla betoniarni B odpowiednio $Q_{1Y} = 194.5$, $y_{med} = 209$, $Q_{3Y} = 222.5$, $IQR_Y = 28$. W obu próbach nie występują dane odstające. Wykresy ramkowe są przedstawione na rys. 2.

Z porównania wykresów widać, że rozkłady wytrzymałości betonu z obu wytwórni nie są takie same i beton z drugiej wytwórni ma nieco większą wytrzymałość, jednak z wyraźnie większym rozproszeniem w kierunku większych wartości.

d) Histogram. Wskaźniki położenia i rozproszenia i ich wizualizacja w postaci wykresu ramkowego mają charakter orientacyjny i nie pozwalają dokładniej ocenić stopnia koncentracji obserwacji w różnych częściach przedziału $[x_{(1)}, x_{(n)}]$. Dotyczy to zwłaszcza sytuacji, gdy próba jest duża (jej liczebność wynosi co najmniej kilkadziesiąt). Na dokładniejszą analizę pozwala dopiero histogram.

Najpierw rozważymy przypadek zmiennej losowej ciągłej i dużej próby, tj. $n \geqslant 50$. Wszystkie obserwacje grupujemy w przyległych do siebie, jednostronnie domkniętych, przedziałach jednakowej długości h_n określo-



Rys.2

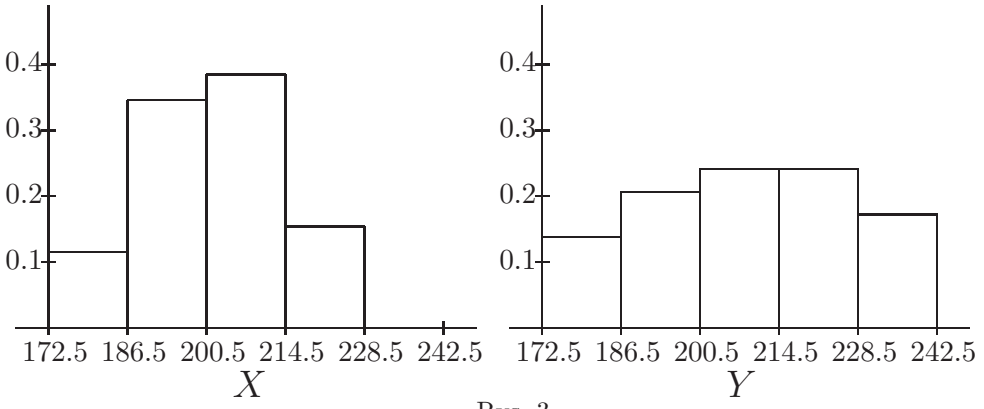
nej przybliżonym wzorem

$$h_n = 2.64\frac{IQR}{\sqrt[3]{n}}.$$

Lewy koniec pierwszego (od lewej) przedziału przyjmujemy mniej więcej równy $x_{(1)} - h_n/2$ i tworzymy tyle przedziałów, aby prawy koniec ostatniego przedziału wypadł powyżej $x_{(n)}$. Oznaczmy liczbę tak określonych przedziałów przez k_n, a same przedziały przez $I_1, ..., I_{k_n}$. Niech n_j oznacza liczbę obserwacji wpadających do przedziału I_j. *Histogramem* nazywamy funkcję $f_n(x)$ określoną na $\mathbb{R}$ wzorem $f_n(x) = n_j/n$ dla $x \in I_j,\ j = 1, ..., k_n$, oraz 0 dla pozostałych $x \in \mathbb{R}$. Frakcję $\nu_j = n_j/n$ nazywamy *częstością* wpadania do I_j. Oczywiście, liczby n_j sumują się do n, a częstości do 1. Wzór, określający h_n, jest orientacyjny i zwykle wybieramy pewne zaokrąglenie w górę lub w dół, także w celu uzyskania bardziej regularnego wyglądu wykresu histogramu.

W przypadku mniejszej próby, tj. n pomiędzy 20 i 50, postępujemy podobnie, ale wybieramy $k_n = 4$ lub 5 i dobieramy odpowiednio do tego h_n. Dla $n < 20$ histogram przestaje być przydatny do analizy rozkładu badanej zmiennej losowej.

Przykład 3 cd. Skonstruujemy histogramy dla obu zmiennych, przyjmując $k = 5$ oraz przedziały $[172.5, 186.5)$, $[186.5, 200.5)$, $[200.5, 214.5)$, $[214.5, 228.5)$, $[228.5, 242.5]$. Liczby obserwacji w kolejnych przedziałach wynoszą: 3, 9, 10, 4, 0 dla X oraz 4, 6, 7, 7, 5 dla Y. Wykresy histogramów są przedstawione na rys. 3.



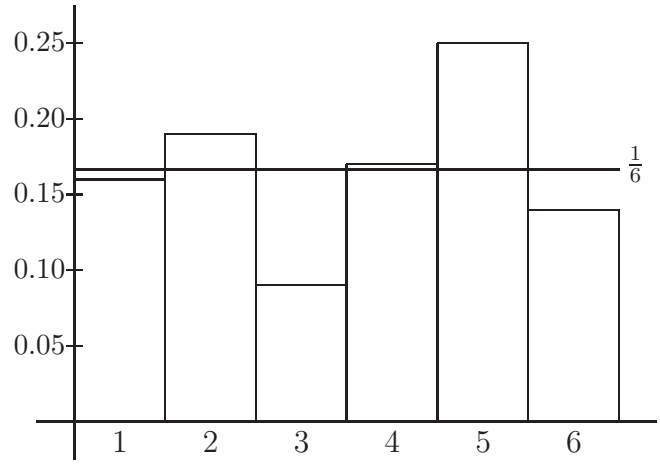
Rys. 3

Widać, że rozkład wytrzymałości betonu z pierwszej wytwórni jest dość mocno skupiony wokół mediany z próby, a dla drugiej wytwórni jest bardzo spłaszczony i przez to bardziej rozproszony. Zatem histogramy mają wyraźnie inny kształt.

Jeśli zmienna losowa jest dyskretna i przyjmuje niezbyt dużo wartości $w_1, ..., w_k$, to dla każdej z nich budujemy prostokąt o wysokości równej częstości $\nu_j = n_j/n$, gdzie n_j oznacza liczbę powtórzeń w próbie wartości w_j. Podstawy prostokątów są jednakowe i dobrane tak, aby histogram wyglądał regularnie. Przy dużej liczbie wartości zmiennej i niedużej liczebności próby sąsiadujące wartości (w uporządkowaniu rosnącym) grupujemy, tak aby liczba prostokątów w histogramie nie była zbyt duża.

Przykład 4. ([KM], Przykład 1.2) Rzucono 100 razy kostką do gry i otrzymano: 16 jedynek, 19 dwójek, 9 trójek, 17 czwórek, 25 piątek i 14 szóstek. Histogram dla dyskretnej zmiennej losowej X, oznaczającej liczbę wyrzuconych oczek, jest przedstawiony na rys. 4.

Spodziewamy się, że częstości wszystkich wyników powinny być bliskie 1/6, tzn. wszystkie prostokąty powinny mieć wysokość bliską 1/6. Dla lepszego porównania poprowadzono na rysunku cienką linię poziomą



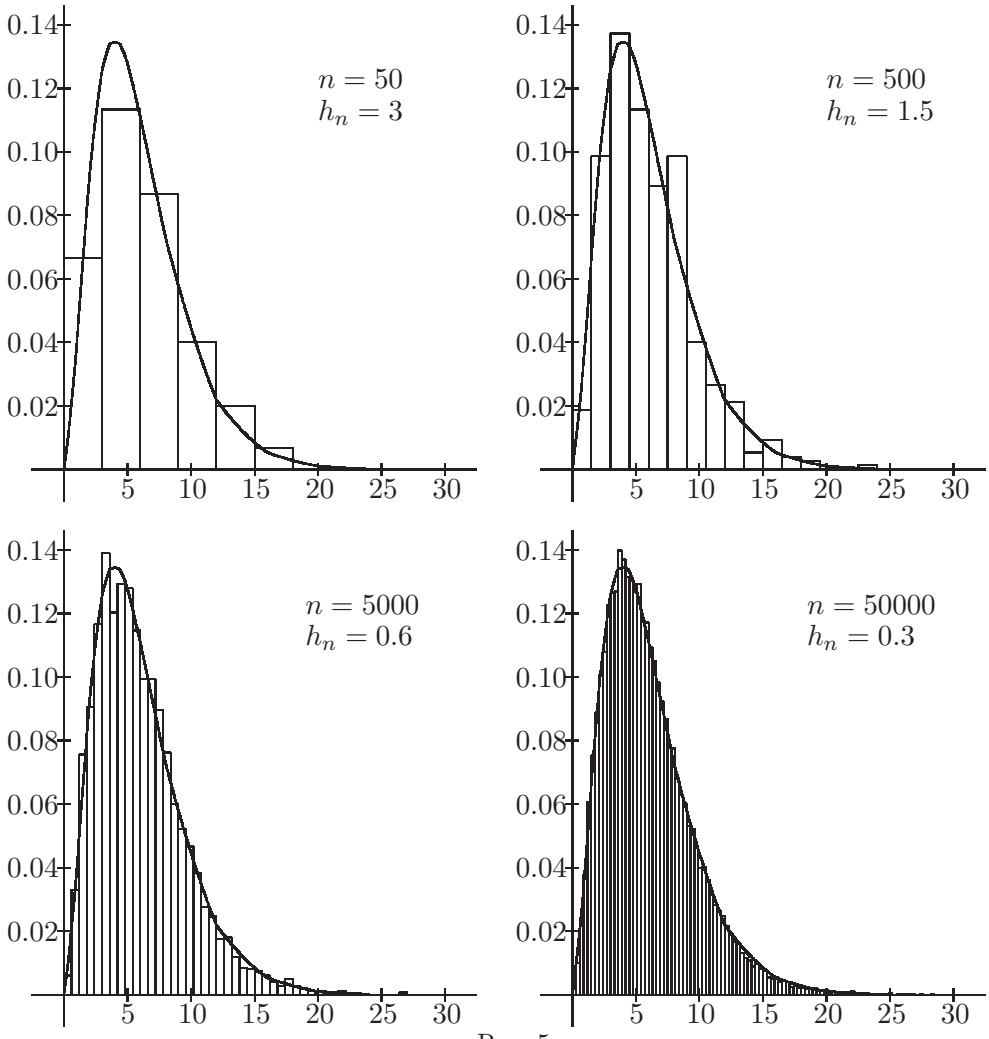
Rys. 4

na wysokości 1/6. Widać, że, histogram znacznie odbiega od naszych oczekiwań. To skutek dużej zmienności, której nie „niweluje" próba o niezbyt dużej liczebności $n = 100$.

Rozważmy ponownie zmienną losową ciągłą i próbę o liczebności co najmniej 50. Pole pod wykresem histogramu $f_n(x)$ wynosi $\sum_{j=1}^{k_n} \nu_j h_n = h_n$. Ponieważ h_n maleje ze wzrostem n, to pole pod wykresem histogramu staje się coraz mniejsze i histogram staje się bezużyteczny. Dlatego dla dużych n rozważamy *skalowany histogram* $\widetilde{f}_n(x)$, który ma stałe pole pod wykresem równe 1. To oznacza, że $\widetilde{f}_n(x) = f_n(x)/h_n$.

e) Granica histogramu. Załóżmy, że zmienna losowa jest ciągła i możemy dowolnie zwiększać liczebność próby n i budować dla każdej z nich skalowany histogram $\widetilde{f}_n(x)$. Ze względu na malejącą długość przedziałów h_n ich liczba k_n rośnie, wykres staje się coraz bardziej „wygładzony", ale pole pod nim jest stałe. Dla „bardzo dużej" próby wykres przestanie być schodkowy i „praktycznie" stanie się wykresem pewnej funkcji „granicznej" $f(x)$. Wydaje się dość intuicyjne, że wykonując nowy ciąg obserwacji rozważanej zmiennej losowej i powtarzając poprzedni proces „nieograniczonego zwiększania n" i budowania skalowanych histogramów, otrzymamy tę samą funkcję „graniczną" $f(x)$, która wobec tego opisuje „naturę" rozkładu badanej zmiennej na osi liczbowej. Można powiedzieć, że funkcja „graniczna" $f(x)$ zawiera „pełną informację" o naszej zmiennej losowej z punktu widzenia częstości jej

wpadania w otoczenia różnych punktów. Dla przykładu, gdy dla pewnych x_1, x_2 mamy $f(x_1) = cf(x_2)$, to częstość wpadania w „małe" otoczenie punktu x_1 jest mniej więcej c razy większa niż częstość wpadania w otoczenie punktu x_2 o tej samej „małej" długości. Tak więc $f(x)$ możemy uważać za „idealny histogram" w sytuacji posiadania „nieskończenie wielu" obserwacji. „Graniczną" funkcję $f(x)$ nazywamy *gęstością rozkładu* rozważanej zmiennej losowej lub krótko gęstością tej zmiennej.



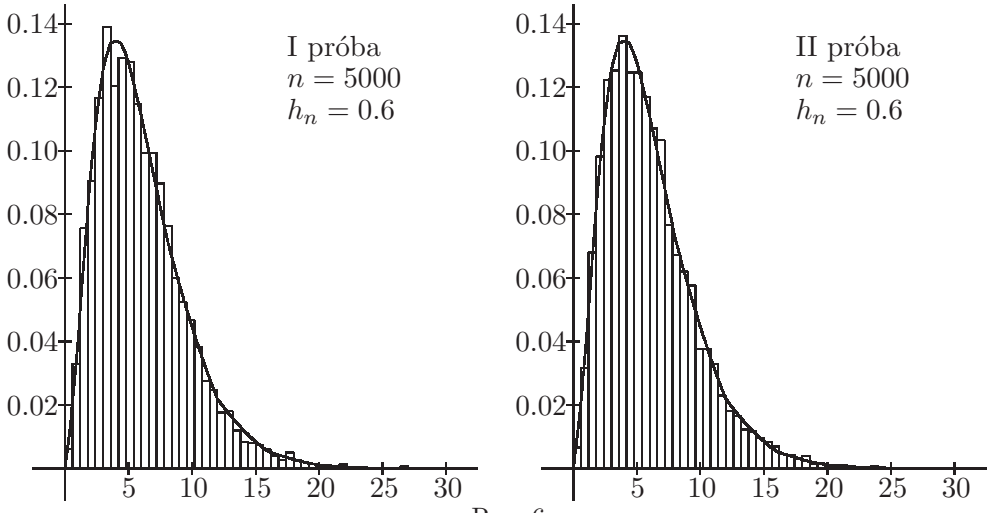
Rys. 5

Przykład 5. Zmierzono czasy życia T (w pewnych jednostkach) masowo produkowanych elementów dla czterech liczebności prób i sporzą-

dzono skalowane histogramy, gdzie długości przedziałów h_n zostały wyznaczone według wzoru podanego w punkcie d) i następnie zaokrąglone do wygodnej wartości. Otrzymane histogramy pokazane są na rys. 5. Dla lepszej oceny zgodności $f_n(x)$ z $f(x)$, na wszystkich wykresach naniesiono gęstość $f(x)$ zmiennej T. Dla $n = 5000$ pobrano także drugą próbę i liczebności obserwacji $l_I,\ l_{II}$ w poszczególnych 50 przedziałach długości $h_{5000} = 0.6$, tj. $I_1 = [0, 0.6),\ I_2 = [0.6, 1.2), ..., I_{50} = [29.4, 30)$, dla obu

nr przedziału	1	2	3	4	5	6	7	8	9	10	11	12	13
l_I	18	99	227	272	350	417	361	388	384	344	298	298	269
l_{II}	20	95	204	295	367	376	409	374	374	351	321	308	230
nr przedziału	14	15	16	17	18	19	20	21	22	23	24	25	26
l_I	229	180	157	140	115	83	74	53	54	36	25	24	22
l_{II}	202	186	173	113	113	99	69	54	50	37	35	29	25
nr przedziału	27	28	29	30	31	32	33	34	35	36	37	38	39
l_I	18	12	8	15	8	6	4	3	2	0	4	1	0
l_{II}	21	14	13	10	12	5	5	2	2	3	0	0	1
nr przedziału	40	41	42	43	44	45	46	47	48	49	50		
l_I	0	0	0	0	0	2	0	0	0	0	0		
l_{II}	2	1	0	0	0	0	0	0	0	0	0		

Tabela 1

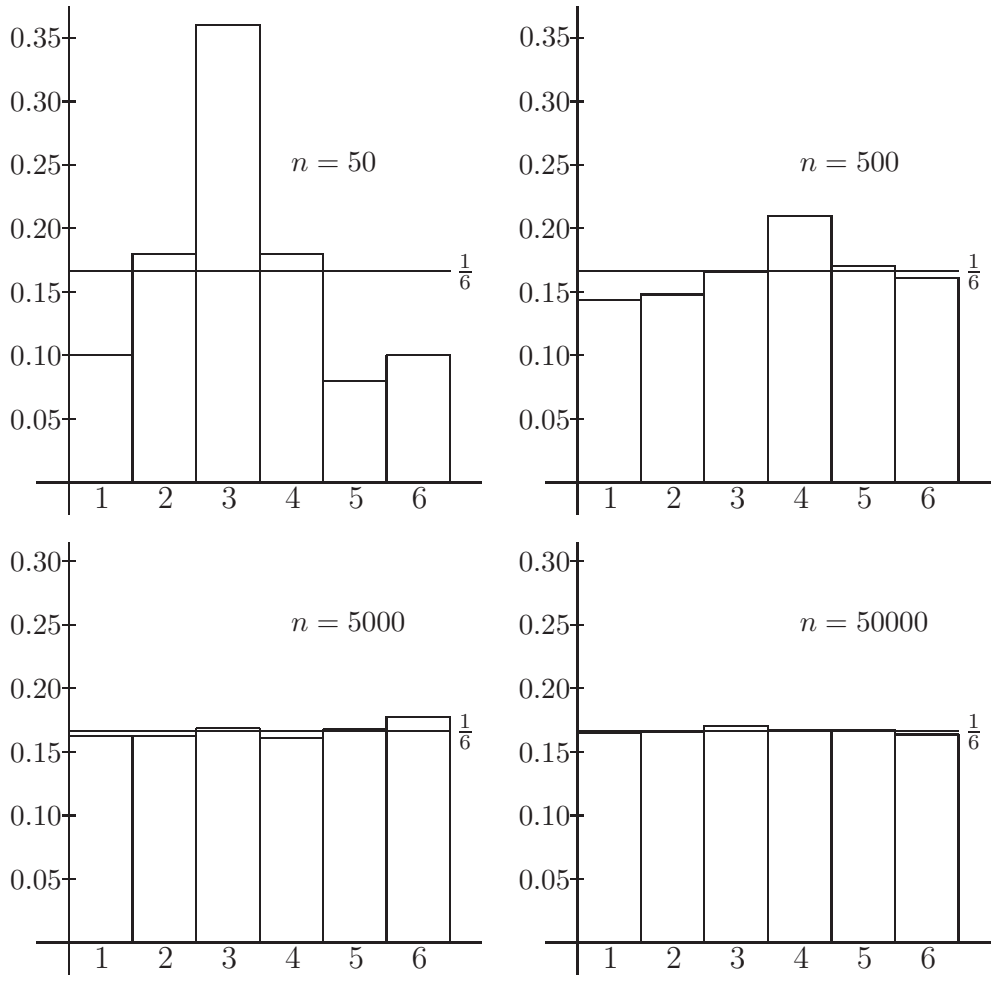


Rys. 6

prób są podane w tabeli 1. Jak widać, liczebności różnią się, w siód-

mym przedziale nawet o blisko 50. Tak właśnie wyraża się w praktyce zmienność obserwowanego czasu życia. Jednak histogramy nie różnią się istotnie. Dla porównania przedstawiamy je na rys. 6.

Przedstawiony powyżej fenomen dla ciągłych zmiennych losowych ma odpowiednik dla zmiennych dyskretnych. I jest on nawet bardziej zgodny z intuicją. W tym przypadku wartości zmiennej i szerokości prostokątów histogramu nie zależą od n. Wahaniom ulegają jedynie ich wysokości, które są równe częstościom. Naturalna intuicja podpowiada nam, że wraz z rosnącym n, częstości $\nu_j = n_j/n$ stabilizują się, czyli „zbliżają się" do



Rys.7

pewnych wartości „granicznych" p_j, które nazywamy *prawdopodobień-*

stwami przyjmowania wartości w_j przez obserwowaną zmienną. Dla wyników rzutu kostką do gry, rozważanym w przykładzie 4, wydaje się oczywiste, że wszystkie częstości są zbieżne do granicy 1/6. Zobaczmy, jak to wygląda w rzeczywistości.

Przykład 4 cd. Wykonano 50 rzutów kostką i otrzymano: 5 jedynek, 9 dwójek, 18 trójek, 9 czwórek, 4 piątki oraz 5 szóstek. Dla 500 rzutów wyniki były odpowiednio: 72, 74, 83, 105, 85 oraz 81; dla 5000 rzutów odpowiednio: 813, 812, 843, 805, 839 oraz 888; a dla 50000 rzutów odpowiednio: 8266, 8300, 8528, 8352, 8362 oraz 8192. Na rys. 7 powyższe wyniki zostały przedstawione w postaci 4 histogramów.
Rysunki potwierdzają, że istotnie ze wzrostem n częstości każdej liczby oczek są coraz bliższe prawdopodobieństwom $p_j = 1/6,\ j = 1, ..., 6$, zaznaczonym cienką linią na wszystkich histogramach. Nawet dla $n = 50000$ zdarzają się odchylenia od idealnej liczby 8333 powtórzeń każdej liczby oczek bliskie 200. Jednak częstości różnią się od 1/6 nie więcej niż 0.004.

Uwaga. W powyższych rozważaniach wprowadziliśmy pojęcie gęstości zmiennej losowej ciągłej oraz prawdopodobieństw dla zmiennej losowej dyskretnej w oparciu o intuicję i potwierdzenie jej eksperymentalnie. Takie „częstościowe" rozumienie prawdopodobieństwa funkcjonowało w rachunku prawdopodobieństwa i statystyce przez ponad 200 lat, aż do początków XX wieku. Niestety, nie może być ono przyjęte jako poprawna definicja. W 1933 roku matematyk rosyjski A. N. Kołmogorow wprowadził definicję aksjomatyczną prawdopodobieństwa na gruncie teorii miary. Od tego momentu rozwinęły się teoria prawdopodobieństwa oraz statystyka matematyczna jako działy matematyki. Ich rozwój trwa do dzisiaj i spowodował, że metody statystyki matematycznej stosuje się praktycznie w każdej nauce przyrodniczej i eksperymentalnej. Dzięki metodom statystycznym nastąpił ogromny postęp w tych naukach. Dla przykładu, opracowanie coraz skuteczniejszych terapii lekowych czy metod diagnostycznych w medycynie nie byłoby możliwe bez zastosowania wnioskowania statystycznego. Warto dodać, że postulowana wyżej zbieżność częstości do prawdopodobieństwa oraz histogramu do gęstości są treścią ważnych twierdzeń teorii prawdopodobieństwa i nasze „pójście na skróty" nie prowadzi do błędu.

Ponieważ podejście aksjomatyczne nie jest intuicyjne i jego wprowa-

dzenie wymaga pewnego czasu, w tak krótkim wykładzie wprowadzającym nie da się tego uczynić bez szkody dla prezentacji podstawowych idei i metod istotnie statystycznych.

Zadania

1. Niech $\overline{x}$ i s^2 będą średnią i wariancją z próby. Wykazać, że dla dowolnego $c \in \mathbb{R}$ zachodzi wzór $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 - (\overline{x} - c)^2$.

2. ([KM] zad. 1.2). Katalogowe zużycie X paliwa (w l na 100 km) 24 modeli samochodów wynosi: 6.3, 8.0, 8.5, 9.3, 5.5, 5.9, 5.9, 6.5, 6.4, 6.6, 8.2, 10.1, 6.3, 6.8, 7.6, 6.7, 7.3, 7.1, 9.2, 6.9, 5.9, 7.5, 8.6, 6.0. Wyznaczyć wskaźniki położenia i rozproszenia zmiennej X, sporządzić wykres ramkowy i histogram.

3. Wynagrodzenia (w tys. zł) 41 pracowników pewnej firmy w uporządkowaniu rosnącym wynoszą: 2.6, 2.6, 2.6, 2.6, 2.8, 2.8, 2.8, 2.8, 3, 3, 3, 3, 3, 3, 3, 3.2, 3.2, 3.2, 3.2, 3.2, 3.2, 3.2, 3.2, 3.2, 3.5, 3.5, 3.5, 3.5, 3.5, 3.5, 4.4, 4.4, 4.4, 4.4, 5.2, 5.2, 5.2, 6.0, 6.0, 6.5. Wyznaczyć wskaźniki położenia i rozproszenia, sporządzić wykres ramkowy oraz histogram.

4. ([KM] zad. 1.3). Suma opadów (w mm) w Warszawie w lipcu w kolejnych latach, poczynając od roku 1811 do roku 1960, wynosiła: 35, 82, 48, 75, 77, 123, 117, 75, 92, 101, 116, 113, 42, 44, 36, 71, 9, 74, 114, 49, 83, 94, 223, 28, 57, 46, 33, 86, 85, 74, 72, 104, 37, 229, 41, 50, 73, 40, 76, 100, 171, 41, 160, 120, 144, 46, 143, 105, 29, 92, 138, 44, 26, 80, 50, 84, 78, 74, 53, 51, 76, 30, 48, 6, 54, 63, 20, 74, 81, 45, 50, 174, 82, 18, 139, 31, 47, 78, 173, 71, 72, 20, 85, 19, 35, 39, 120, 92, 172, 98, 37, 77, 143, 26, 96, 13, 132, 109, 116, 132, 37, 32, 91, 101, 77, 87, 99, 181, 166, 68, 5, 122, 33, 84, 66, 64, 149, 23, 20, 115, 71, 108, 55, 166, 124, 115, 53, 71, 49, 73, 93, 76, 113, 53, 77, 37, 78, 124, 84, 44, 68, 26, 65, 136, 154, 82, 88, 38, 80, 159. Obliczyć średnią, wariancję, medianę z próby i rozstęp międzykwartylowy. Sporządzić histogram i wykres ramkowy. Wyznaczyć średnią obciętą, odrzucając po 15% skrajnych wyników. Ocenić własności rozkładu sumy opadów (jednomodalność, skośność, wyostrzenie).

5. Dla danych z poprzedniego zadania rozważyć oddzielnie sumy opadów z lat 1811 - 1860 oraz z lat 1911 - 1960. Sporządzić wykresy ramkowe

i histogramy dla tych danych i ocenić zgrubnie, czy po 100 latach zmienił się rozkład sumy opadów w lipcu.

6. Przeciętna długość życia mężczyzn w 29 państwach świata wynosi: 74, 76, 77, 72, 69, 65, 72, 68, 70, 72, 72, 73, 74, 75, 72, 74, 68, 72, 75, 72, 69, 71, 75, 76, 81, 73, 69, 78, 74. Wyznaczyć wskaźniki położenia i rozproszenia, sporządzić wykres ramkowy i histogram. Ocenić, czy rozkład jest skośny, czy symetryczny. Wyznaczyć średnią obciętą, odrzucając po 10% skrajnych wyników.

7. Przeciętna długość życia kobiet w 29 państwach świata, tych samych co w zadaniu poprzednim, wynosi: 80, 74, 76, 77, 80, 82, 81, 80, 84, 79, 81, 75, 71, 73, 76, 78, 83, 81, 73, 74, 75, 79, 81, 75, 80, 79, 75, 77, 81. Wyznaczyć wskaźniki położenia i rozproszenia, sporządzić wykres ramkowy i histogram. Porównać jakościowo rozkłady obu zmiennych. Czy z powyższych danych można wnioskować, że rozkład długości życia kobiety jest przesuniętym rozkładem długości życia mężczyzny? O ile lat?

8. Z dowolnego tekstu w języku polskim wybrać 6 wierszy i wypisać z nich po kolei samogłoski *a, e, i, o, u, y*. Sporządzić histogram dla otrzymanej próby. Przyporządkować kolejnym samogłoskom liczby od 1 do 6 i wyznaczyć wskaźniki położenia i rozproszenia oraz sporządzić wykres ramkowy.

9. Podobnie jak w poprzednim zadaniu, wybrać 30 wierszy z dowolnego tekstu w języku polskim i wypisać kolejno samogłoski *ą, ę, ó*. Wykonać te same czynności dla otrzymanej próby. Samogłoskom przyporządkować odpowiednio liczby 1, 2 i 3.