RACHUNEK PRAWDOPDODOBIEŃSTWA I STATYSTYKA

OKNO - Ośrodek Kształcenia na Odległość
Politechnika Warszawska

Krystyna Lipińska Dominik Jagiełło Rafał Maj

2010

Spis treści

Słowo wstępne

Celem przedmiotu Rachunek Prawdopodobieństwa i statystyka jest dostarczenie studentom aparatu pojęciowego niezbędnego w toku studiowania przedmiotów kierunkowych.

Materiał wykładów i ćwiczeń zawartych w podręczniku OKNA zawiera podstawowe elementy tych działów rachunku prawdopodobieństwa i statystyki, które mogą być użyteczne w przedmiotach specjalistycznych.

Student powinien opanować umiejętność odnajdywania w podręczniku odpowiednich metod i wzorów ułatwiających rozwiązanie problemów opisanych modelem matematycznym. Przystępując do samodzielnego opanowania materiału należy starać się zrozumieć rolę podanych definicji i wzorów ułatwiających rozwiązywanie zadań i ustalić relacje między nimi. Jest to bardzo przyjemny proces w wyniku którego można samodzielnie rozwiązać umieszczone na końcu rozdziału zadania uzyskując wynik zgodny z podaną odpowiedzią.

Zaliczenie przedmiotu polega na rozwiązaniu dwóch zestawów projektowych oraz zdaniu egzaminu. Egzamin polega na sprawdzeniu czy student opanował materiał objęty przedmiotem.

Studiując samodzielnie można korzystać z literatury uzupełniającej, pamiętając jednak że mogą występować różne metody i oznaczenia rozwiązywania zadań a nawet mogą występować różnice w definicjach.

Pomocą w opanowaniu systematycznym obowiązującego do egzaminu materiału są zajęcia stacjonarne na których wykładowca omawia trudniejsze zadania i wyjaśnia wątpliwości w postaci indywidualnych konsultacji.

Przedmiot jest realizowany w jednym półsemestrze.

Szczegóły dotyczące prowadzenia przedmiotu w danym semestrze będą podawane w witrynie przedmiotu.

Życzymy wytrwałości i satysfakcji z trudnych ale ciekawych studiów.

Zespół prowadzących przedmiot RPiS

Wstęp

Podręcznik zawiera podstawowe elementy rachunku prawdopodobieństwa i statystyki matematycznej.

Rachunek prawdopodobieństwa zajmuje się badaniem i wykrywaniem prawidłowości w zjawiskach, na które działają czynniki losowe oraz budowaniem modeli matematycznych tych zjawisk.

Statystyka matematyczna zajmuje się natomiast metodami wnioskowania o całej zbiorowości danych na podstawie zbadania pewnej jej części zwanej próbką.

Czynniki losowe występują w wielu dziedzinach jak: teorii sterowania, miernictwie, kontroli jakości a także w organizacji i zarządzaniu w ekonomii.

W podręczniku umieszczone są definicje i twierdzenia bez dowodów. Wykorzystanie teorii ilustrowane jest przykładami. W zadaniach wymagających żmudnych obliczeń podawany jest jedynie algorytm ułatwiający uzyskanie wyniku oraz wynik końcowy. Ważny jest bowiem sposób uzyskania rozwiązania i interpretacja otrzymanego rezultatu.

Uwaga *Przy czytaniu podręcznika proszę zwórcić uwagę na fakt iż w większości rysunków osie* Ox, Oy *zostały w rożny sposób skalibrowane, tzn. jedna jednostka na jednej osi może być innej długości od 1 jednostki na drugiej osi.*

Wykład 1

Zdarzenia elementarne

W tym wykładzie omówione są pojęcia z kombinatoryki, które są wykorzystywane w najprostyszych przykładach prezentujących rozważany materiał. Następnie omówione są podstawowe pojęcia rachunku prawdopodobieństwa.

1.1 Elementy kombinatoryki

Definicja 1.1. Symbol $n!$ dla $n \in \mathbb{N}$ nazywamy *silnią*. Wyraża się on wzorem

$$n! = \begin{cases} 1 & \text{dla } n = 0 \vee n = 1 \\ 1 \cdot 2 \cdot \ldots \cdot n & \text{dla } n \geqslant 2 \end{cases} \tag{1.1}$$

Definicja 1.2. *Symbolem Newtona* nazywamy wyrażenie

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \qquad \text{dla } n, k \in \mathbb{N}, k \leqslant n. \tag{1.2}$$

Reguła możenia. Jeżeli pewien wybór zależy od skończenie wielu decyzji, powiedzmy k, przy czym podejmując pierwszą decyzję mamy n_1 możliwości, drugą n_2 możliwości, ..., k-tą n_k możliwości, bo wybór ten może być zrobiony na

$$n = n_1 \cdot n_2 \cdot \ldots \cdot n_k \tag{1.3}$$

możliwości.

Przykład 1.3. Na ile sposobów można podzielić 3 role męskie i 2 kobiece pomiędzy trzech aktorów i dwie aktorki?

Rozwiazanie. Role męskie możemy przydzielić na 3! sposobów, role kobiece na 2! sposobów, więc korzystając z reguły mnożenia wynika, że role te możemy przydzielić na

$$3! \cdot 2! = 6 \cdot 2 = 12$$

sposobów. □

Przykład 1.4. Ile nastąpi powitań, gdy jednocześnie spotka się 6 znajomych?

Rozwiązanie. Mamy $n = 6, k = 2$, czyli $\binom{6}{2} = 15$ powitań. □

Definicja 1.5. *Permutacją bez powtórzeń* zbioru n-elementowego $A = \{a_1, a_2, \ldots, a_n\}$, dla $n \in \mathbb{N}$ nazywamy każdy n-wyrazowy `ciąg` utworzony ze wszystkich n-elementów zbioru A, czyli każde uporządkowanie elementów zbioru A.

Stwierdzenie 1.6. *Liczba wszystkich różnych permutacji bez powtórzeń zbioru n-elementowego jest równa*

$$P_n = n! \tag{1.4}$$

Permutacje wykorzystujemy, gdy:

- występują wszystkie elementy zbioru,
- kolejność elementów jest istotna.

Przykład 1.7. Na ile sposobów można ułożyć na półce 4 tomową encyklopedię?

Rozwiazanie. Rozmieszczamy wszystkie elementy i kolejność elementów ma znaczenie, zatem można to zrobić na 4! sposobów. □

Definicja 1.8. *Permutacją n-wyrazową z powtórzeniami* zbioru k-elementowego $A = \{a_1, a_2, \ldots, a_k\}$, w której element a_1 występuje n_1 razy, element a_2 występuje n_2 razy, ..., element a_k występuje n_k razy, przy czym $n_1 + n_2 + \ldots + n_k = n$, nazywamy każdy n-wyrazowy `ciąg`, w ktrórym element a_i występuje n_i razy, $i = 1, 2, \ldots, k$,

Stwierdzenie 1.9. *Liczba wszystkich różnych n-wyrazowych permutacji z powtórzeniami ze zbioru k-elementowego jest równa*

$$P_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!},$$

gdzie $n_i \in \mathbb{N}$, $i = 1, 2, \ldots, k$, n_i – liczba powtórzeń elementu $a_i \in A$, $n_1 + n_2 + \ldots + n_k = n$.

Przykład 1.10. Ile różnych liczb 6-cyfrowych można utworzyć z cyfr: $1, 1, 3, 3, 3, 5$?

Rozwiązanie. Są to permutacje z powtórzeniami. Zatem korzystając ze wzoru otrzymujemy, że możemy utworzyć

$$P_6(2, 3, 1) = \frac{6!}{2! \cdot 3! \cdot 1!}$$

liczb. □

Definicja 1.11. *Waracją k-wyrazową z powtórzeniami* zbioru A, n-elementowego, gdzie $k \in \mathbb{N}$, nazywamy każdy k-wyrazowy `ciąg`, którego wyrazami są elementy danego zbioru A.

Stwierdzenie 1.12. *Liczba wszystkich różnych k-wyrazowych wariacji z powtórzeniami zbioru n-elementowego jest równa*

$$W_n^k = n^k. \tag{1.5}$$

Waracje z powtórzeniami wykorzystujemy, gdy:

- kolejność elementów jest istotna,
- elementy mogą się powtarzać (losowanie ze zwracaniem),
- niekoniecznie wszystkie elementy zbioru są wykorzystane.

Przykład 1.13. Na ile sposobów można umieścić 11 piłeczek w czterech szufladach?

Rozwiązanie. Można to zrobić na $W_4^{11} = 4^{11}$ sposobów.

Definicja 1.14. *Wariacją k-wyrazową bez powtórzeń* zbioru A, n-elementowego, gdzie $k \in \mathbb{N}$, nazywamy każdy k-wyrazowy `ciąg` różnowartościowy, którego wyrazami są elementy danego zbioru A.

Stwierdzenie 1.15. *Liczba wszystkich różnych k-wyrazowych wariacji bez powtórzeń zbioru n-elementowego jest równa*

$$V_n^k = \frac{n!}{(n-k)!} = n(n-1)(n-2)\cdot \ldots \cdot (n-k+1). \tag{1.6}$$

Wariacje bez powtórzeń wykorzystujemy, gdy:

- kolejność elementów jest istotna,
- elementy nie mogą się powtarzać (losowanie bez zwracania),
- niekoniecznie wszystkie elementy zbioru są wykorzystane.

Przykład 1.16. Ile jest liczb czterocyfrowych utworzonych tylko z cyfr nieparzystych?

Rozwiązanie. Cyfr nieparzystych jest 5. Możemy je rozmieszczać na 4 pozycjach. Mamy zatem, że tych liczb jest $V_5^4 = \frac{5!}{(5-4)!} = 120$. □

Definicja 1.17. *Kombinacją k-elementową bez powtórzeń* zbioru A, n-elementowego, gdzie $k, n \in \mathbb{N}$, nazywamy każdy `podzbiór` k-elementowy zbioru A, przy czym elementy nie mogą się powtarzać.

Stwierdzenie 1.18. *Liczba wszystkich różnych kombinacji k-elementowych bez powtórzeń jest równa*

$$C_n^k = \binom{n}{k}. \tag{1.7}$$

Kombinacje stosujemy wtedy, gdy kolejność elementów nie ma znaczenia.

Przykład 1.19. Na ile sposobów można wypełnić kupon Dużego Lotka?

Rozwiązanie. Są to kombinacje 6 elementowe ze zbioru 49 elementowego, czyli kupon można wypełnić na $C_{49}^6 = \binom{49}{6}$ sposobów. □

1.2 Definicja prawdopodobieństwa

Niech Ω będzie dowolnym zbiorem, którego elementy oznaczamy przez ω. Zbiór ten będziemy nazywali *przestrzenią zdarzeń elementarnych*, a jego elementy zdarzeniami elementarnymi. Przestrzeń zdarzeń elementarnych jest pojęciem pierwotnym (nie definiowanym) w rachunku prawdopodobieństwa. W konkretnych przykładach będziemy Ω utożsamiać ze zbiorem wszystkich możliwych wyników doświadczenia losowego. Ponieważ zdarzenia losowe będzimy rozumieli jako podzbiory zbioru wszystkich możliwych zdarzeń losowych w danym doświadczeniu, więc wygodnie jest wprowadzić następującą definicję.

Definicja 1.20. Rodzinę podzbiorów $\mathcal{A}$ zbioru Ω nazywamy algebrą zbiorów, jeżeli:

(i) $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$,

(ii) $A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$,

(iii) $A \in \mathcal{A} \Rightarrow A' = (\Omega \setminus A) \in \mathcal{A}$,

(iv) $\Omega \in \mathcal{A}, \varnothing \in \mathcal{A}$.

Jako *zdarzenia losowe* będziemy rozumieć podzbiory z pewnej algebry podzbiorów zbioru Ω.

Definicja 1.21. Rzeczywistą fukcję $P(A)$ określoną na algebrze $\mathcal{A}$ podzbiorów zbioru Ω nazywamy *prawdopodobieństwem*, jeżeli spełnia warunki

(i) $A \in \mathcal{A} \Rightarrow P(A) \geqslant 0$,

(ii) $A, B \in \mathcal{A}, A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$,

(iii) $P(\Omega) = 1, \quad P(\varnothing) = 0$.

Podstawowe własności prawdopodobieństwa są zamieszczone poniżej.

1. Jeżeli $A \cap B = \varnothing, A, B \in \mathcal{A}$, to $P(A \cup B) = P(A) + P(B)$.
2. Jeżeli $A \subset B, A, B \in \mathcal{A}$, to $P(B \setminus A) = P(B) - P(A)$.
3. Jeżeli $A \subset B, A, B \in \mathcal{A}$, to $P(A) \leqslant P(B)$.
4. Dla każdego $A \in \mathcal{A}, 0 \leqslant P(A) \leqslant 1$.
5. Jeżeli $A, B \in \mathcal{A}$, to $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Trójkę $(\Omega, \mathcal{A}, P)$ nazywamy *przestrzenią probabilistyczną*.

Często użyteczna jest też klasyczna definicja prawdopodobieństwa.

Definicja 1.22. Jeżeli zbiór zdarzeń elementarnych Ω jest skończony i każde zdarzenie elementarne ma tą samą szansę zaistnienia to prawdopodobieństwo zdarzenia A wyraża się wzorem

$$P(A) = \frac{|A|}{|\Omega|}, \tag{1.8}$$

gdzie $|A|, |\Omega|$ oznacza liczność zbioru.

Tak określone prawdopodobieństwo spełnia wszystkie aksjomaty z definicji 1.21.

Przykład 1.23. Partia odbiorników telewizyjnych składa się z 10 sztuk, z których 3 są wadliwe. Jakie jest prawdopodobieństwo zdarzenia polegającego na tym, że dwa wybrane losowo odbiorniki są wadliwe?

Rozwiązanie. Przez A oznaczmy zbiór odbiorników wadliwych. Zatem zdarzeń sprzyjających jest $|A| = \binom{3}{2} = 3$ – wybieramy dwie sztuki wadliwe spośród trzech. Natomiast wszystkich zdarzeń $|\Omega| = \binom{10}{2} = 45$ – wybieramy dwie sztuki spośród 10. Stąd

$$P(A) = \frac{3}{45} = \frac{1}{15} \approx 0,06.$$

□

Przykład 1.24. Egzaminator przygotował 30 pytań, wypisując na każdej kartce 4 pytania. Zdający umie odopowiedzieć poprawnie na połowę pytań. Jakie jest prawdopodobieństwo że zdający odpowie poprawnie na 4 pytania?

Rozwiązanie. Oznaczmy przez A zdarzenie polegające na wylosowaniu zestawu z pytaniami, na które zdający zna odpowiedź. Mamy $|A| = \binom{15}{4}$ – losujemy 4 pytania z 15 „dobrych", oraz $|\Omega| = \binom{30}{4}$ – losujemy 4 pytania spośród wszystkich. Zatem

$$P(A) = \frac{\binom{15}{4}}{\binom{30}{4}} = \frac{13}{261} \approx 0,05,$$

czyli 5%. □

Przykład 1.25. Spośród 100 studentów, 25 wybrało język angielski, 40 niemiecki, 20 rosyjski a 20 angielski i niemiecki. Jakie jest prawdopodobieństwo, że losowo wybrany student uczy się języka angielskiego lub niemieckiego?

Rozwiązanie. Oznaczmy przez A student uczy się języka angielskiego, przez B - niemieckiego. Mamy $P(A) = \frac{25}{100}, P(B) = \frac{40}{100}, P(A \cap B) = \frac{20}{100}$. Stąd

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,25 + 0,4 - 0,2 = 0,45.$$

□

Przykład 1.26. W 30 osobowej grupie studentów jest 8 kobiet. Grupa otrzymała 6 biletów bezpłatnych do teatru, które losowano w grupie. Jakie jest prawdopodobieństwo, że wśród posiadaczy bezpłatnych biletów są dokładnie 3 kobiety?

Rozwiazanie.

$$P(A) = \frac{\binom{8}{3}\binom{22}{3}}{\binom{30}{6}} = 0,02.$$

□

1.3 Zadania do samodzielnego rozwiązania

1.1. Komisja złożona z 3 kobiet i 5 mężczyzn wybiera spośród siebie przewodniczącego. Jakie jest prawdopodobieństwo, że zostanie nim kobieta, przy założeniu, że wszyscy członkowie komisji mają takie same szanse? Odp. $\frac{3}{8}$.

1.2. Spośród pięciu piłek o różnej wielkości wybieramy dwie. Jakie jest prawdopodobieństwo, że że wśród wylosowanych będzie najmniejsza, przy założeniu, że wszystkie piłki mają równe szanse być wybrane? Odp. $\frac{2}{5}$.

1.3. Znaleźć prawdopodobieństwo, że przy 5 krotnym rzucie kostką otrzymamy 5 różnych wyników? Odp. $\frac{6!}{6^5}$.

1.4. Jakie jest prawdopodobieństwo, że przy jednoczesnym rzucie trzema kostkami do gry wypadnie

a) 11,

b) 12.

Odp. a) $\frac{27}{216}$, b) $\frac{25}{216}$.

1.5. W urnie mamy 14 kul czarnych, 16 kul białych i dwie kule niebieskie. Obliczyć prawodopodobieństwo, że przy losowaniu jednej kuli

a) wylosowana kula będzie niebieska,

b) wylosowana kula będzie niebieska lub czarna.

Odp. a) $\frac{1}{16}$, b) $\frac{1}{2}$.

1.6. W skrzyni znajduje się 6 dobrych i 4 wadliwe elementy. Obliczyć prawdopodobieństow, że wśród 4 wybranych losowo elementów, nie będzie ani jednego wadliwego. Odp. $\frac{1}{14}$.

1.7. Autobus zatrzymuje się na 10 przystankach. Jakie jest prawdopodobieństwo, że spośród 6 osób znajdujących się w autobusie, każda wysiądzie na innym przystanku? Odp. $0,1512$.

1.8. W pudełku znajduje się 25 długopisów, z czego 5 jest zepsutych. Wybieramy losowo 3 długopisy. Obliczyć prawdopodobieństwo, że co najmniej dwa są dobre. Odp. $\frac{209}{230}$.

1.9. Student umie odpowiedzieć na 20 spośród 25 pytań egzaminacyjnych. Na egzaminie losuje 3 pytania. Jakie jest prawdopodobieństwo, że uczeń odpowie na 2 pytania? Odp. $\frac{209}{230}$.

1.10. W sześciu szufladach umieszczamy sześć krawatów. Zakładając, że każde rozmieszczenie krawatów jest jednakowo prawdopodobne obliczyć prawdopodobieństwo zdarzenia, że co najmniej dwie szuflady będą puste. Odp. $\frac{61}{81}$.

Wykład 2

Prawdopodobieństwo warunkowe i wzór Bayesa

W tym wykładzie omówione są jedne z najważniejszych pojęć rachunku prawdopodobieństwa, mianowicie prawdopodobieństwo warunkowe oraz wzór Bayesa.

2.1 Prawdopodobieństwo warunkowe

Rozważmy dwa zdarzenia w tym samym skończonym zbiorze zdarzeń elementarnych.

Definicja 2.1. Prawdopodobieństwo warunkowe zdarzenia A pod warunkiem, że zaszło zdarzenie B wyliczamy ze wzoru

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{jeżeli } P(B) > 0. \tag{2.1}$$

Analogicznie obliczamy prawdopodobieństwo warunkowe zdarzenie B pod warunkiem, że zaszło zdarzenie A

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad \text{jeżeli } P(A) > 0. \tag{2.2}$$

Przekształcając powyższy wzór otrzymujemy wzór na *prawdopodobieństwo ilorazowe*

$$P(A \cap B) = P(A)P(B|A). \tag{2.3}$$

Definicja 2.2. Zdarzenia A i B nazywamy niezależnymi, jeżeli

$$P(A \cap B) = P(A) \cdot P(B).$$

Twierdzenie 2.3. *Jeżeli niezależne są zdarzenia A oraz B, z tej samej rodziny zdarzeń, to niezależne są także zdarzenia*

a) A *i* B'*;*

b) A' *i* B*;*

c) A' *i* B'*.*

Przykład 2.4. Do windy na parterze sześciopiętrowego bloku wsiadło 4 pasażerów. Jakie jest prawdopodobieństwo, że każda z osób wysiadła na innym piętrze?

Rozwiazanie. Pradopodobieństwo zdarzenia $A_i, i = 1, 2, 3, 4$ wynoszą $P(A_1) = \frac{6}{6}, P(A_2) = \frac{5}{6}, P(A_3) = \frac{4}{6}, P(A_4) = \frac{3}{6}$. Zdarzenia te są niezależne, więc

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot P(A_4) \approx 0,28,$$

czyli 28%. □

Przykład 2.5. W pewnym przedsiębiorstwie 96% wyrobów jest dobrych. Na 100 dobrych wyrobów średnio 75 jest piewszego gatunku. Znaleźć prawdopodobieństwo tego, że dobra sztuka wyprodukowana w tym przedsiębiorstwie jest piewszego gatunku.

Rozwiązanie. Oznaczmy przez A zdarzenie polegające na tym, że sztuka jest dobra, a przez B - że sztuka jest piewszego gatunku. Mamy

$$P(A \cap B) = P(A) \cdot P(B|A) = 0,96 \cdot 0,75 = 0,72.$$

□

Przykład 2.6. Obliczyć prawdopodobieństwo, że losowo wybrany los loteryjny wygrywa największą stawkę, jeżeli wiadomo, że 25% losów przegrywa a 20% to losy wygrywające największą stawkę.

Rozwiązanie. Oznaczmy przez A zdarzenie, że los jest wygrywający (cokolwiek). Stąd $P(A) = 1 - P(A') = 1 - \frac{25}{100} = \frac{3}{4}$, natomiast przez B zdarzenie, że los wygrywa najwyższą stawkę. Stąd $P(B|A) = \frac{20}{100}$. Zatem

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{3}{4} \cdot \frac{1}{5} = 0,15.$$

□

Przykład 2.7. W pudełku zawierającym 15 rezystorów 3 są wybrakowane. Rezystory z pudełka wyjmujemy w sposób losowy. Obliczyć prawdopodobieństwo

a) wyjęcia kolejno dwóch rezystorów dobrych,

b) wyjęcia dwóch rezystorów, z których jeden jest dobry a drugi wybrakowany.

Rozwiązanie. Oznaczmy przez A_1 zdarzenie, że pierwszy z wyjętych rezystorów jest dobry, zaś przez A_2, że drugi z wyjętych rezystorów jest dobry.

W przypadku a) $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{12}{15} \cdot \frac{11}{14} = \frac{22}{35} \approx 0,63$.

Oznaczmy przez B_1 zdarzenie, że piewszy z wyjętych rezystorów jest wybrakowany, zaś przez B_2, że drugi z wyjętych rezystorów jest wybrakowany.

W przypadku b) $P(A_1 \cap A_2) = P(A_1) \cdot P(B_2|A_1) = \frac{12}{15} \cdot \frac{3}{14} = \frac{6}{35} \approx 0,17$. □

Przykład 2.8. Policja uzyskała informację, że terroryści podłożyli bomby w dwóch spośród 8 odlatujących samolotów. Jakie jest prawdopodobieństwo, że zostaną one znalezione już po przeszukaniu dwóch pierwszych samolotów.

Rozwiązanie. Oznaczmy przez A_1 zdarzenie, że w piewszym samolocie jest bomba, zaś przez A_2 zdarzenie, że w drugim samolocie jest bomba. Stąd

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{2}{8} \cdot \frac{1}{7} = \frac{1}{28} \approx 0,04.$$

□

2.2 Prawdopodobieństwo całkowite i wzór Bayesa

Definicja 2.9. Układ zdarzeń $A_1, A_2, \ldots, A_n$ nazywamy *zupełnym*, jeżeli zdarzenia te są parami niezależne (wykluczają się), tzn. $A_i \cap A_j = \varnothing$ dla $i \neq j$ oraz

$$A_1 \cup A_2 \cup \ldots \cup A_n = \Omega. \tag{2.4}$$

Twierdzenie 2.10. *Jeżeli zdarzenia $A_1, A_2, \ldots, A_n$ tworzą układ zupełny oraz $P(A_i) > 0$, dla $i = 1, 2, \ldots, n$, to dla dowolnego zdarzenia B*

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i). \tag{2.5}$$

Przykład 2.11. W jednej urnie mamy 3 kule białe i 4 czarne, a w drugiej 5 kul białych i 4 czarne. Rzucamy kostką do gry. Gdy wypadnie liczba podzielna przez 3, to losujemy z piewszej urny, w przeciwnym przypadku z drugiej. Jakie jest prawdopodobieństwo wylosowania kuli białej?

Rozwiązanie. Oznaczmy przez B_1 zdarzenie, że wypadła liczba podzielna przez 3, przez B_2 zdarzenie, że wypadła liczba niepodzielna przez 3, przez A zdarzenie, że wylosowaliśmy kulę białą. Mamy

$$P(A) = P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) = \frac{1}{3} \cdot \frac{3}{7} + \frac{2}{3} \cdot \frac{5}{9} = \frac{97}{189} \approx 0,51.$$

□

Przykład 2.12. Trzy fabryki wytwarzają pewien towar, dla którego określona jest norma. Przy czym

Fabryka 1 dostarcza na rynek 30% towaru w którym normę spełnia 80% towaru,

Fabryka 2 dostarcza na rynek 40% towaru w którym normę spełnia 70% towaru,

Fabryka 3 dostarcza na rynek 30% towaru w którym normę spełnia 60% towaru.

Obliczyć prawdopodobieństwo, że towar dostarczony na rynek spełnia normę.

Rozwiązanie. Oznaczmy przez B zdarzenie, że towar dostarczony na rynek spełnia normę. Mamy $P(A_1) = 0,3; P(B|A_1) = 0,8; P(A_2) = 0,4; P(B|A_2) = 0,7; P(A_3) = 0,3; P(B|A_3) = 0,6$. Stąd prawdopodobieństwo speplnienia normy przez towar na rynku wynosi

$$P(B) = 0,3 \cdot 0,8 + 0,4 \cdot 0,7 + 0,3 \cdot 0,6 = 0,7.$$

□

Twierdzenie 2.13. *Jeżeli $P(B) > 0$ i spełnione są założenia Twierdzenia 2.10, to zachodzi wzór zwany* wzorem Bayesa

$$P(A_j|B) = \frac{P(A_j) \cdot P(B|A_j)}{P(B)}.$$

Przykład 2.14. Korzystając z danych z Przykładu 2.12 obliczyć prawdopodobieństwo spełnienia normy przez towar wyprodukowany w fabryce 3.

Rozwiązanie. Mamy

$$P(A_3|B) = \frac{0,3 \cdot 0,6}{0,7} = 0,25.$$

□

2.3 Zadania do samodzielnego rozwiązania

2.1. Rzucamy dwukrotnie kostką do gry. Obliczyć prawdopodobieństwo, że suma oczek będzie większa od 9, jeżeli za piewszym razem wypadło 6 oczek? Odp. $\frac{1}{2}$.

2.2. W skrzyni znajduje się 12 elementów, z czego 6 jest dobrych a 6 wadliwych. W sposób losowy, bez zwracania wybieramy dwa elementy. Oblicz prawdopodobieństwo wylosowania za drugim elementu wadliwego, pod warunkiem, że za wpiewszym razem wybrano element dobry. Odp. $\frac{6}{11}$.

2.3. Na dworcu kolejowym znajdują się dwoje schodów ruchomych. Pierwsze są sprawne z prawdopodobieństwem $\frac{1}{2}$, natomiast drugie $\frac{1}{3}$. Prawdopodobieństwo, że działają piewsze schody, gdy zepsute są drugie wynosi $\frac{1}{2}$.

a) Jakie jest prawdopodobieństwo, że działają drugie schody, pod warunkiem, że nie działają pierwsze?

b) Jakie jest prawdopodobieństwo, że działają przynajmniej jedne schody?

Odp. a) $\frac{1}{3}$, b) $\frac{2}{3}$.

2.4. Rzucamy dwiema symetrycznymi kostkami do gry. Oblicz prawdopodobieństwo otrzymania sumy oczek równej 3, jeżeli na piewszej kostce wypadła 1. Odp. $\frac{1}{6}$.

2.5. Rozważmy rodziny z dwojgiem dzieci. Niech d oznacza dziewczynkę, c - chłopca. Zdarzeniami elementarnymi będą pary: (d,d), (d,c), (c,d), (c,c), gdzie pierwsza litera w parze oznacza płeć starszego dziecka, druga zaś młodszego. Zakładając, że wszystkie zdarzenia są jednakowo prawdopodobne obliczyć prawdopodobieństwo, że w losowo wybranej rodzinie z dwojgiem dzieci, jest dwóch chłopców, pod warunkiem, że w tej rodzinie jest co najmniej jeden chłopiec. Odp. $\frac{1}{3}$.

2.6. Student dojeżdza na uczelnię rowerem średnio co drugi dzień, autobusem co trzeci dzień, a tramwajem co szósty. Jadąc rowerem, spóźnia się z prawdopodobieństwem raz na sześciesiąt razy jadąc autobusem - raz na dwadzieścia razy, a tramwajem raz na dziesięć razy. Oblicz prawdopodobieństwo, że student spóźni się na uczelnię. Odp. $\frac{1}{24}$.

2.7. Potrzeby świerkowych sadzonek dla nadleśnictwa pokrywa produkcja dwóch szkółek leśnych. Pierwsza szkółka pokrywa 75% zapotrzebowania, przy czym na 100 sadzonek z tej szkółki 80 jest piewszej jakości. Druga szkółka pokrywa 25% zapotrzebowania, przy czym na 100 zadzonek z tej szkółki 60 jest pierwszej jakości. Obliczyć prawdopodobieństwo, że losowo wybrana sadzonka jest piewszej jakości. Odp. $0,75$.

Wykład 3

Zmienna losowa jednowymiarowa

W tym wykładzie omówione jest pojęcie zmiennej losowej, typy zmiennych losowych, parametry zmiennych losowych oraz przykłady rozładów prawdopodobieństwa.

3.1 Zmienna losowa

Definicja 3.1. Zmienną losową nazywamy funkcję X przyporządkowującą zdarzeniu elementarnemu dokładnie jedną liczbę rzeczywistą, tj. $X : \Omega \to \mathbb{R}$, spełniającą warunek:
dla dowolnego $a \in \mathbb{R}$ zbiór $\{\omega \in \Omega : X(\omega) < a\}$ należy do zbioru zdarzeń losowych.

Notacja. Symbolu $A := B$ będziemy używali do oznaczenia, że pewne oznaczenie jest równe z definicji. Należy go rozumieć następująco: Symbol (wyrażenie) A jest równy z definicji obiektowi B.

Zatem poniższe oznaczenia należy rozumieć następująco: zbiór stojący z prawej storny symbolu $:=$ dla skrócenia zapisu będzie oznaczany przez symbol stojący z lewej strony tego znaku.

Będziemy korzystali z następującego zapisu
$(X < a) := \{\omega \in \Omega : X(\omega) < a\}$,
$(X \leqslant a) := \{\omega \in \Omega : X(\omega) \leqslant a\}$,
$(X > a) := \{\omega \in \Omega : X(\omega) > a\}$,
$(X \geqslant a) := \{\omega \in \Omega : X(\omega) \geqslant a\}$.

Zmienne losowe pozwalają przedstawić wyniki doświadczeń losowych za pomocą liczb. Prawdopodobieństwo przyjęcia przez zmienną losową danych wartości można wyznaczyć za pomocą dystrybuanty.

Definicja 3.2. *Dystrybuantą* zmiennej losowej nazywamy funkcję F określoną następująco

$$F(x) = P(X \leqslant x). \tag{3.1}$$

Dystrybuanta zmiennej losowej ma następujące własności

1. Jest funkcją niemalejącą, prawostronnie ciągłą.
2. $\lim\limits_{x\to-\infty} F(x) = 0$ oraz $\lim\limits_{x\to\infty} F(x) = 1$.
3. $0 \leqslant F(x) \leqslant 1$ dla każdego x.
4. Dla każdego przedziału $\langle a, b\rangle$ mamy $P(a \leqslant x \leqslant b) = F(b) - F(a-)$, gdzie $F(a-) = \lim\limits_{x\to a^-} F(x)$ oznacza granicę lewostroną funkcji F w punkcie a.
5. $P(xb) = 1 - F(b)$.

Dalej zostaną podane typy zmiennych losowych: zmienną losową skokową (dyskretną) i zmienną losową ciągłą.

Zmienna losowa skokowa (dyskretna)

Zmienna losowa skokowa X, jest to zmienna losowa, której zbiór wartości jest zbiorem skończonym lub przeliczalnym (tnz. jest równoliczny ze zbiorem liczb natrualnych $\mathbb{N}$), czyli przyjmuje wartości pewnego ciągu x skończonego lub nieskończonego z prawdopodobieństwem p, czyli jest określona funkcja prawdopodobieństwa

$$p_k = P(X = x_k), \quad p_k > 0, \quad \sum_k p_k = 1. \tag{3.2}$$

Zależność tę można przedstawić za pomocą tabeli

x_k	x_1	x_2	$\cdots$	x_n
p_k	p_1	p_2	$\cdots$	p_n

Przykład 3.3. Zmienna losowa X dyskretna ma funkcję prawdopodobieństwa określoną tabelą

x_i	-1	0	2
p_i	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$

Wyznaczyć dystrybuantę zmiennej losowej X.

Rozwiazanie. Dystrybuanta zmiennej losowej X jest funkcją $F(x) = P(X \leqslant x)$. Wtedy

$$\begin{aligned}
F(-2) &= P(X \leqslant -2) = 0 \\
F(-1) &= P(X \leqslant -1) = \frac{1}{2} \\
F(-0,5) &= P(X \leqslant -0,5) = \frac{1}{2} \\
F(0) &= P(X \leqslant 0) = P(-1) + P(0) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} \\
F(1) &= P(X \leqslant 1) = P(-1) + P(0) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} \\
F(3) &= P(X \leqslant 3) = P(-1) + P(0) + P(2) = \frac{1}{2} + \frac{1}{6} + \frac{1}{3} = 1
\end{aligned}$$

Zatem

$$F(x) = \begin{cases} 0 & \text{dla } x < -1 \\ \frac{1}{2} & \text{dla } -1 \leqslant x < 0 \\ \frac{2}{3} & \text{dla } 0 \leqslant x < 2 \\ 1 & \text{dla } x > 2 \end{cases}$$

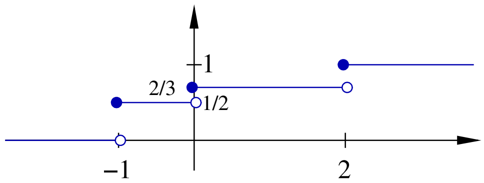
□

Zmienna losowa ciągła

Zmienna losowa X ciągła, jest to zmienna losowa zdefiniowana za pomocą funkcji $f(x)$ zwanej *gęstością prawdopodobieństwa zmiennej* X w postaci

$$P(a < X < b) = \int_a^b f(x)dx. \tag{3.3}$$

Gęstość prawdopodobieństwa spełnia warunki



Rysunek 3.1: Dystrybuanta zmiennej losowej z Przykładu 3.3

1. $f(x) > 0$, czyli jest funkcją nieujemną,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Dystrybuanta zmiennej losowej ciągłej może być przedstawiona w postaci

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{3.4}$$

więc jest funkcją górnej granicy cakowania.

	zmienna losowa skokowa	zmienna losowa ciągła
wartość oczekiwana	$E(X) = \sum_{k=1}^{n} x_k \cdot p_k$	$E(X) = \int_{-\infty}^{\infty} x f(x) dx$
wariancja	$\sigma^2(X) = \sum_{k=1}^{n} (x_k - E(X))^2 \cdot p_k$	$\sigma^2(X) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) dx$
odchylenie standardowe	$\sigma = \sqrt{\sigma^2(X)}$	$\sigma = \sqrt{\sigma^2(X)}$

Tablica 3.1: Parametry zmiennych losowych.

3.2 Rozkład prawdopodobieństwa zmiennej losowej

Mówimy, że znamy rozkład prawdopodobieństwa zmiennej losowej, jeżeli jest znana

- dystrybuanta zmiennej losowej,
- funkcja prawdopodobieństwa, gdy zmienna jest skokowa, lub
- funkcja gęstości, jeżeli zmienna losowa jest ciągła.

Parametry rozkładów

Zmienne losowe i ich rozkłady jednowymiarowe nie zawsze są wystarczające lub wygodne do opisania bardziej złożonych problemów. Przy analizie danych uzyskanych w wyniku przeprowadzonych obserwacji istotną rolę odgrywają parametry rozkładów zmiennych losowych takie jak wartość oczekiwana $E(X)$, wariancja $\sigma^2(X)$, odchylenie standardowe $\sigma(X)$. Wariancja jest miarą rozproszenia zmiennej losowej dookoła jej wartości oczekiwanej, także miarą rozproszenia jest odchylenie standardowe σ.

Definicje parametrów umieścimy w zestawieniu

Obliczanie wariancji upraszcza wzór

$$\sigma^2(X) = E(X^2) - [E(X)]^2. \tag{3.5}$$

Teraz podamy podstawowe własności parametrów.

1. $E(aX) = aE(X)$, gdzie a jest stałą.
2. Jeżeli istnieją wartości oczekiwane zmiennych losowych X i Y, to
$$E(X + Y) = E(X) + E(Y).$$
3. $E(XY) = E(X)E(Y)$, jeżeli zmienne losowe X i Y są niezależne.
4. $E(a) = a$, gdzie a jest stałą.
5. $\sigma^2(aX) = a\sigma^2(X)$.
6. $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$.

7. $\sigma^2(a) = 0$, gdzie a jest stałą.

Przykład 3.4. Zmienna losowa ma rozkład orkreślony tabelą

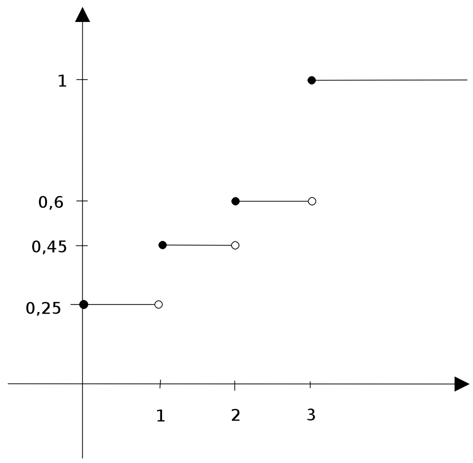
x_k	0	1	2	3
p_k	0,25	0,2	0,15	0,4

Wyznaczyć: dystrybuantę rozkładu, $P(X < 3), E(X), D^2(X), \sigma$.

Rozwiązanie. Mamy $P(X \leqslant 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0,25+0,2+0,15 = 0,6$. $E(X) = 0\cdot 0,25+1\cdot 0,2+2\cdot 0,15+3\cdot 0,4 = 1,70$, $[E(X)]^2 = 2,89$,
$E(X^2) = 0^2 \cdot 0,25 + 1^2 \cdot 0,2 + 2^2 \cdot 0,15 + 3^2 \cdot 0,4 = 4,4$.
$\sigma^2(X) = E(X^2) - [E(X)]^2 = 4,4 - 2,89 = 1,51$. $\sigma = \sqrt{1,51} \approx 1,23$.

$$F(x) = \begin{cases} 0 & x \in (-\infty, 0) \\ 0,25 & x \in \langle 0, 1) \\ 0,45 & x \in \langle 1, 2) \\ 0,60 & x \in \langle 2, 3) \\ 1 & x \in \langle 3, \infty) \end{cases}$$

Wykres dystrybuanty natomiast histogram □



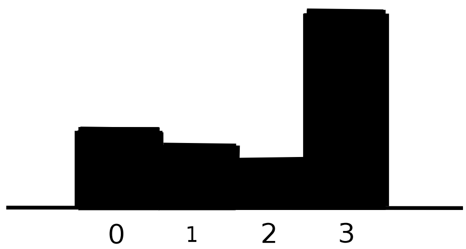
Rysunek 3.2: Dystrubunata zmiennej losowej z Przykładu 3.4

Rozkłady zmiennej skokowej

Rozkład zero-jednynkowy

Zmienna losowa X ma rozkład zerojedynkowy z parametrem p, jeżeli jej funkcja prawdopodobieństwa wyraża się równościami

$$P(X = 1) = p, P(X = 0) = 1 - p = q. \tag{3.6}$$



Rysunek 3.3: Histogram zmiennej losowej z Przykładu 3.4

Wartość oczekiwana tej zmiennej $E(X) = p$, wariancja $\sigma^2(X) = pq$.

Rozkład dwumianowy (Bernoullego)

Jeżeli pewne doświadczenie losowe składa się z serii n prób, przy czym kolejne próby są niezależne oraz w każdej próbie możliwe sa dwa wyniki: `sukces` z prawdopodobieństwem p, oraz `porażka` z prawdopodobieństwem $q = 1 - p$, to takie doświadczenie nazywamy *schematem Beronoullego* z parametrami n i p. Prawdopodobieństwo tego, że w serii n-prób uzyskamy k sukcesów i $n - k$ porażek wyraża się wzorem

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n. \tag{3.7}$$

Wartość oczekiwana tej zmiennej wynosi $E(X) = p$, natomiast wariancja $\sigma^2(X) = pq$.

Przy stosowaniu rozkładu dwumianowego należy zwracać uwagę na rodzaj warunków wynikających ze zdarzenia. Są to sformułowania „dokładnie ...", „co najmniej ...", „co najwyżej ...".

Przykład 3.5. W hali fabrycznej pracuje 5 maszyn. Każda z nich psuje się z prawdopodobieństwem $p = \frac{1}{3}$ niezależnie od siebie. Wyznaczyć prawdopodobieństwa

a) zepsuła się jedna maszyna, tj. $P(X = 1)$,

b) żadna maszyna się niepopsuła, tj. $P(X = 0)$,

c) zepsuły się trzy maszyny, tj. $P(X = 3)$,

d) zepsuła się co najmniej jendna maszyna, tj. $P(X \geqslant 1)$,

e) zepsuła się co najwyżej jedna maszyna, tj. $P(X \leqslant 1)$,

f) zepsuło się więcej niż jedna maszyna, tj. $P(X > 1)$.

Rozwiązanie. a) $P(X = 1) = \binom{5}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^4 = \frac{80}{243} \approx 0,320$

b) $P(X = 0) = \binom{5}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^5 = \frac{32}{243} \approx 0,132$

c) $P(X=3)=\binom{5}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2=\frac{40}{243}\approx 0,167$

d) $P(X\geqslant 1)=1-P(X<1)=P(X=0)=0,868$

e) $P(X\leqslant 1)=P(X=0)+P(X=1)=0,461$

f) $P(X>1)=1-P(X\leqslant 1)=0,539$

□

Przykład 3.6. Środek owadobójczy zabija przeciętnie 90% owadów. Środek ten zastosowano na 10 owadach. Obliczyć prawdopodobieństwo, że co najwyżej dwa osobniki przeżyją.

Rozwiązanie. Oznaczmy przez A zdarzenie, że owad przeżyje. Mamy więc $p=0,1, q=0,9, n=10$.

$$\begin{aligned}P(X\leqslant 2)&=P(X=0)+P(X=1)+P(X=2)=\\&=\binom{10}{0}(0,1)^0(0,9)^{10}+\binom{10}{1}(0,1)^1(0,9)^9+\binom{10}{2}(0,1)^2(0,9)^8=\\&=0,929.\end{aligned}$$

□

Rozkład Poissona

Zmienna skokowa X ma *rozkład Poissona*, jeżeli

$$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!},\qquad \text{dla } k=0,1,2,\ldots \tag{3.8}$$

Wiele występujących w praktyce rozkładów może być aproksymowane rozkładem Poissona.

Jeżeli liczba doświadczeń n jest duża a prawdopodobieństwo p małe, to obliczenie k sukcesów jest bardzo utrudnione przy zastosowaniu rozkładu dwumianowego Bernullego. Można wówczas przyjąć $\lambda=np$ i zastosować wzór graniczny przy $n\to\infty$. Otrzymujemy wówczas wzór przybliżony

$$P(X=k)\approx\frac{(np)^k e^{-np}}{k!}. \tag{3.9}$$

Przykład 3.7. Daltonizm stwierdza się o 1% mężczyzn. Obliczyć prawdopodobieństwo, że w próbie liczącej $n=100$ mężczyzn

a) nie będzie ani jednego daltonisty,

b) będzie co najmniej trzech.

Rozwiązanie. Mamy $n=100, p=0,01$.

a) $P(X=0)=\frac{1}{0!}e^{-1}=0,37$,

b) $P(X\geqslant 3)=1-P(X<3)=1-[P(X=0)+P(X=1)+P(X=2)]$.

$P(X=1)=\frac{1}{1!}e^{-1}=0,37$. $P(X=2)=\frac{1}{2!}e^{-1}=0,18$. Stąd

$$P(X\geqslant 3)=1-[0,37+0,37+0,18]=0,08.$$

□

Rozkłady zmiennej losowej ciągłej

Rozkład normalny

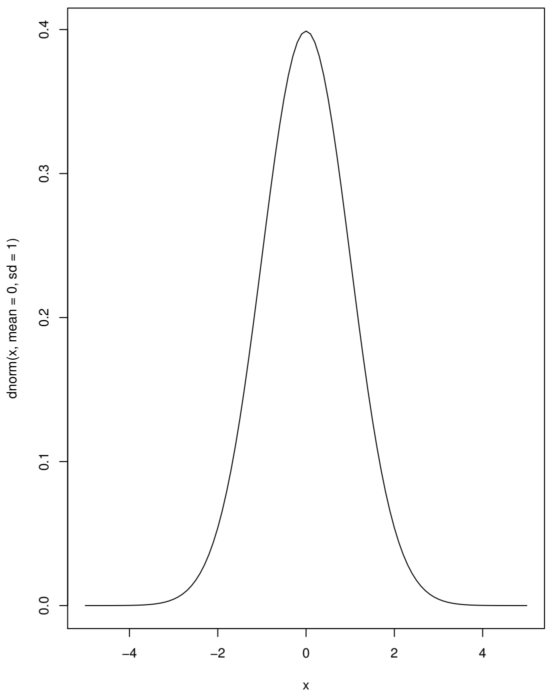
Rozkład normalny jest rozkładem o funkcji gęstości prawdopodobieństwa

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.10}$$

gdzie $\mu = E(X)$ jest wartością oczekiwaną a σ odchyleniem standardowym. Symbolicznie zapisujemy ten rozkład jako $N(\mu, \sigma)$. Szczególnym przypadkiem jest rozkład $N(0, 1)$ o gęstości

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

której wykres przedstawia poniższy rysunek



Rysunek 3.4: Gęstość rozkładu normalnego $N(0, 1)$.

Dystrybuanta tego rozkładu jest równa

$$F(x) = P(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \Phi(x). \tag{3.11}$$

Przykład 3.8. Dla rozkładu $N(0,1)$ obliczyć

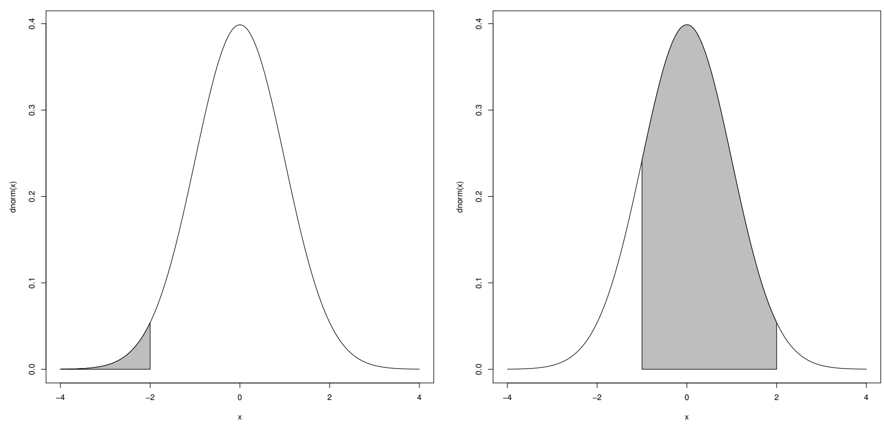
a) $P(X < -2)$,

b) $P(-1 \leqslant X \leqslant 2)$,

c) $P(X > 6)$.

Rozwiązanie. Mamy a)

$$P(X < -2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2} e^{-\frac{t^2}{2}} dt = \Phi(-2) =$$
$$= 1 - \Phi(2) = 1 - 0,97725 = 0,02275.$$



Rysunek 3.5: Przykład 3.8 a) i b)

b)

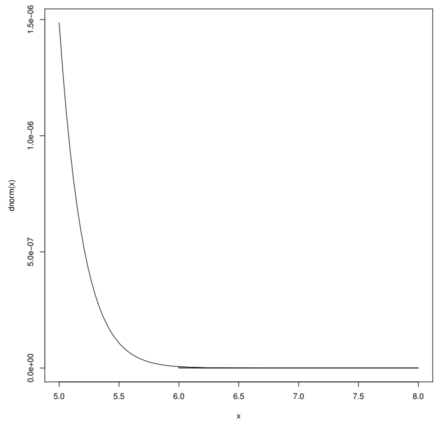
$$\begin{aligned} P(-1 \leqslant X \leqslant 3) &= F(3) - F(1) = \Phi(3) - (1 - \Phi(-1)) = \\ &= \Phi(3) + \Phi(1) - 1 = 0,9987 + 0,8413 - 1 \approx \\ &\approx 0,84. \end{aligned}$$

c)

$$P(X > 6) = 1 - P(X \leqslant 6) = 1 - F(6) = 1 - \Phi(6) \approx 1 - 1 \approx 0.$$

□

Uwaga 3.9. W podpunkcie c) powyższego zadania przyjeliśmy, że $\Phi(6) \approx 1$, co jak spojrzymy do tablic rozkładu normalnego popełniamy mały błąd, gdyż już dla wartości 5 ta różnica między prawdziwą wartością a 1 jest bardzo niewielka (w praktyce zaniedbywalna). Widoczne jest to też na ostatnim rysunku.



Rysunek 3.6: Przykład 3.8 c)

W praktyce występują jednak najczęściej rozkłady $N(m,\sigma)$, gdzie $m \neq 0$ i $\sigma \neq 1$. Wówczas wprowadzamy *zmienną standaryzowaną*

$$Y = \frac{X-m}{\sigma}, \tag{3.12}$$

która ma już rozkład $N(0,1)$, co umożliwia nam skorzystanie z funkcji $\Phi(x)$.

Przykład 3.10. Wydajność pracy jest mierzona liczbą detali wykonanych przez pracownika na danym stanowisku. Liczba detali dana jest zmienną losową X dla $N(8,2)$. Obliczyć $P(X<5)$.

Rozwiązanie. Mamy $Y = \frac{X-8}{2}$, czyli

$$\begin{aligned} P(X<5) &= P\left(\frac{X-8}{2} < \frac{5-8}{2}\right) = P\left(Y < -1,5\right) = F(-1,5) = \Phi(-1,5) = \\ &= 1 - 0,93319 = 0,6681 \end{aligned}$$

□

Pewne własności rozkładu normalnego

1. Jeżeli zmienna X ma rozkład normalny $N(m,\sigma)$, to zmienna losowa $Y = aX+b$ ma rozkład $N(am+b, |a|\sigma)$.

2. Jeżeli zmienne losowe X i Y mają niezależne rozkłady $N(m_1,\sigma_1), N(m_2,\sigma_2)$, to zmienna losowa $Z = X+Y$ ma rozkład $N(m_1+m_2, \sqrt{\sigma_1^2+\sigma_2^2})$.

Przykład 3.11. Urządzenie złożone z dwóch bloków pracuje w ten sposób, że najpierw włączony jest pierwszy blok, a w chwili awarii tego bloku włącza się drugi blok. Czas bezawaryjnej pracy bloków są zmiennymi losowymi o rozkładach $N(60; 4)$ i $N(80; 3)$ odpowiednio. Obliczyć prawdopodobieństwo, że urządzenie będzie pracować co najmniej 150 godzin.

Rozwiązanie. $Z = X + Y$, zatem Z ma rozkład $N(60 + 80, \sqrt{4^2 + 3^2}) = N(140, 5)$.

$$P(Z \geqslant 150) = P\left(\frac{Z - 140}{5} \geqslant \frac{150 - 140}{5}\right) = P\left(R \geqslant 2\right) = 1 - P(R < 2) =$$
$$= 1 - \Phi(2) = 1 - 0,97725 \approx 0,022.$$

Czyli około $2,2\%$. □

3.3 Zadania do samodzielnego rozwiązania

3.1. Zmienna losowa skokowa X ma funkcję prawdopodobieństwa:

x_i	-1	1	4
p_i	0,5	0,4	0,1

Wyznaczyć dystrybuantę zmiennej losowej X i narysować jej wykres.

$$\text{Odp. } F(x) = \begin{cases} 0 & x \leqslant -1 \\ 0,5 & -1 < x \leqslant 1 \\ 0,9 & 1 < x \leqslant 4 \\ 1 & x > 4 \end{cases}$$

3.2. Zmienna losowa ciągła X ma gęstość

$$f(x) = \begin{cases} 0 & x \leqslant -1 \\ \frac{1}{2} & -1 < x \leqslant 0 \\ x & 0 < x \leqslant 1 \\ 0 & x > 0 \end{cases}$$

Wyznaczyć dystrybuantę zmiennej X.

$$\text{Odp. } F(x) = \begin{cases} 0 & x \leqslant -1 \\ \frac{1}{2}(x+1) & -1 < x \leqslant 0 \\ \frac{1}{2}(x^2+1) & 0 < x \leqslant 1 \\ 1 & x > 1 \end{cases}$$

3.3. Dystrybuanta zmiennej losowej ciągłej dana jest wzorem

$$F(x) = \begin{cases} 0 & x \leqslant 0 \\ \frac{1}{16}x^2 & 0 < x \leqslant 4 \\ 1 & x > 4 \end{cases}$$

Wyznaczyć funkcję gęstości zmiennej X oraz wyznaczyć wartość oczekiwaną i wariancję zmiennej X.

$$\text{Odp. } f(x) = \begin{cases} 0 & x \leqslant 0 \\ \frac{1}{8}x & 0 < x \leqslant 4 \\ 0 & x > 4 \end{cases}, E(X) = \frac{8}{3}, \sigma^2(X) = \frac{8}{9}.$$

3.4. Gęstością zmiennej losowej X jest funkcja

$$f(x) = \begin{cases} 0 & x \leqslant 1 \\ \frac{1}{2} & 1 < x \leqslant 3 \\ 0 & x > 3 \end{cases}$$

Wyznaczyć dystrybuantę zmiennej losowej X oraz wyznaczyć $P(X > 2)$.

$$\text{Odp. } F(x) = \begin{cases} 0 & x \leqslant 1 \\ \frac{1}{2}x - \frac{1}{2} & 1 < x \leqslant 3 \\ 1 & x > 3 \end{cases}, P(X > 2) = \frac{1}{2}.$$

3.5. Zmienna losowa X przyjmuje wartości $0, 1, 2, 3$ z prawdopodobieństwami odpowiednio równymi $0,1; 0,1; 0,2; 0,6$. Obliczyć wartość oczekiwaną i wariancję zmiennej X.

Odp. $E(X) = 2,3, \sigma^2(X) = 1,01$.

3.6. Zmienna losowa X ma rozkład normalny $N(1,5; 2)$. Obliczyć prawdopodobieństwo:

a) $P(X < -2,5)$

b) $P(X > -0,5)$,

c) $P(0,5 < X < 2)$.

Odp. a) 0,02275, b) 0,8413, c) 0,2902.

3.7. Masa gruszek odmiany klops ma rozkład normalny $N(160, 30)$. Oblicz prawdopodobieństwo, że gruszka tego gatunku waży od 130 do 160 gramów.

Odp. 0,3413.

3.8. W populacji studentów uczęszczających na zajęcia ze statystyki dokonano pomiaru wzrostu mężczyzn. W wyniku badania stwierdzono, że zmienna losowa X wyrażająca wzrost studenta ma rozkład normalny $N(178, 10)$. Oblicz prawdopodobieństwo, że

a) wzrost studenta jest mniejszy niż 188 cm,

b) wzrost studenta jest większy niż 172,

c) wzrost studenta jest większy niż 200 cm,

d) wzrost studenta należy do przedziału (166 cm, 186 cm).

Odp. a) 0,8413, b) 0,7257, c) 0,0139, d) 0,673.

Wykład 4

Zmienne losowe dwuwymiarowe

Omówione są zmienne losowe dwuwymiarowe, ich parametry. Następnie funkcje zmiennych losowych oraz twierdzenia graniczne.

4.1 Zmienna losowa dwuwymiarowa

Rozkład dwuwymiarowej zmiennej losowej $Z = (X, Y)$ jest rozkładem dwóch zmiennych lowowych X i Y.

W przypadku, gdy zmienne losowe X i Y są dyskretne, możemy ich rozkład opisać łączną funkcją prawdopodobieństwa

$$P(X = x, Y = y) = P(x, y), \tag{4.1}$$

która daje prawdopodobieństwo tego, że zmienna losowa X osiąga wartość x i jednocześnie zmienna losowa Y osiąga wartość y. Łączne prawdopodobieństwo $P(x, y)$ możemy podać w postaci tzw. tablicy korelacyjnej. Przy założeniu, że różnych wartości zmiennej X jest r a różnych wartości zmiennej Y jest s, ta tablica obejmuje $r \cdot s$ łącznych prawdopodobności możliwych kombinacji wartości x i y.

x		y			suma
	y_1	y_2	$\cdots$	y_s	
x_1	$P(x_1, y_1)$	$P(x_1, y_2)$	$\cdots$	$P(x_1, y_s)$	$P_1(x_1)$
x_2	$P(x_2, y_1)$	$P(x_2, y_2)$	$\cdots$	$P(x_1, y_s)$	$P_1(x_2)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_r	$P(x_r, y_1)$	$P(x_r, y_2)$	$\cdots$	$P(x_r, y_s)$	$P_1(x_r)$
suma	$P_2(y_1)$	$P_2(y_2)$	$\cdots$	$P_2(y_s)$	1

Poziome sumy tych prawdopodobieństw w tej tablicy są wartościami brzegowej funkcji prawdopodobieństwa $P_1(x)$, która podaje prawdopodobieństwa, że zmienna losowa X osiąga watość x bez względu na wartości zmiennej Y. Podobnie pionowe sumy tych prawdopodobieństw dają wartości brzegowej funkcji prawdopodobieństwa $P_2(y)$. Mamy zatem

$$\sum_y P(x, y) = P_1(x), \qquad \sum_x P(x, y) = P_2(y),$$

(4.2)

oraz

$$\sum_x \sum_y P(x, y) = \sum_x P_1(x) = \sum_y P_2(y) = 1.$$

(4.3)

Rozkład dwóch dysktretnych lub ciągłych zmiennych losowych można opisać łączną dystrybuantą

$$F(x, y) = P(X \leqslant x, Y \leqslant y), \tag{4.4}$$

która podaje prawdopodobieństwo, że zmienna X osiągnie wartość mniejsze niż x a jednocześnie zmienna Y osiągnie watość mniejszą od y. Łączna dystrybuanta spełnia warunki

$$F(-\infty, y) = F(-\infty, x) = F(-\infty, -\infty) = 0, \quad F(\infty, \infty) = 1. \tag{4.5}$$

Prawdopodobieństwo, że ciągła zmienna losowa X osiągnie wartość z przedziału $\langle x_1, x_2 \rangle$ i jednocześnie ciągła zmienna losowa Y osiągnie wartość z przedziału $\langle y_1, y_2 \rangle$ jest równa

$$P(x_1 \leqslant X \leqslant x_2, y_1 \leqslant Y \leqslant y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1). \tag{4.6}$$

Możemy także otrzymać dystrybuanty zmiennych losowowych brzegowych kładąc

$$F_1(x) = F(x, \infty), \qquad F_2(x) = F(\infty, y). \tag{4.7}$$

Rozkład dwuwymiarowej zmiennej losowej ciągłej możemy także opisać za pomocą łącznej gęstości prawdopodobieństwa $f(x, y)$. Dwuwymiarowa gęstość prawdopodobieństwa jest tak samo jak jednowymiarowa funkcją nieujemną i spełnia warunek

$$\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(x, y) dx \right] dy = 1. \tag{4.8}$$

Brzegowe gęstości prawdopodobieństwa

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy,$$
$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Łączna dystrybuanta dwuwymiarowej zmiennej losowej możemy otrzymać z łącznej gęstości i na odwrót

$$F(x, y) = \int_{-\infty}^{y} \left[\int_{-\infty}^{x} f(t, u) dt \right] du, \tag{4.9}$$
$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}. \tag{4.10}$$

Kolejnym typem rozkładów (oprócz łącznego i brzegowego) są rozkłady warunkowe. Rozkładem warunkowym zmiennej losowej X względem y rozumiemy rozkład tej zmiennej przy założeniu, że zmienna Y przyjmuje wartość y i analogicznie rozkładem warunkowym zmiennej X względem x rozumiemy rozkład tej zmiennej przy założeniu, że zmienna X przyjmuje wartość x. Rozkład warunkowy jest zdefiniowany jako iloraz łącznego i brzegowego rozkładu.

Dla dwóch zmiennych losowych dyskretnych X i Y funkcje prawdopodobności są dane

$$P(x|y) = \frac{P(x, y)}{P_2(y)}, \quad P_2(y) \neq 0, \tag{4.11}$$

$$P(y|x) = \frac{P(x,y)}{P_1(x)}, \quad P_1(x) \neq 0, \tag{4.12}$$

dystrybuanty warunkowe

$$F(x|y) = \frac{\sum\limits_{t<x} P(t,y)}{P_2(y)}, \qquad P_2(y) \neq 0, \tag{4.13}$$

$$F(y|x) = \frac{\sum\limits_{u<y} P(x,u)}{P_1(x)}, \qquad P_1(x) \neq 0. \tag{4.14}$$

Dla zmiennych losowych ciągłych X i Y gęstości warunkowe są określone wzorami

$$f(x|y) = \frac{f(x,y)}{f_2(y)}, \qquad f_2(y) \neq 0, \tag{4.15}$$

$$f(y|x) = \frac{f(x,y)}{f_1(x)}, \qquad f_1(x) \neq 0 \tag{4.16}$$

i warunkowe dystrybuanty

$$F(x,y) = \frac{\int_{-\infty}^{x} f(t,y)dt}{f_2(y)}, \qquad f_2(y) \neq 0, \tag{4.17}$$

$$F(y|x) = \frac{\int_{-\infty}^{y} f(x,t)dt}{f_1(x)}, \qquad f_1(x) \neq 0. \tag{4.18}$$

Definicja 4.1. Mówimy, że dwie zmienne losowe X oraz Y są niezależne, jeżeli dla wszyskich i, j zachodzi równść

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j). \tag{4.19}$$

Czyli zmienne losowe X i Y są niezależne, jeżeli rozkład jednej zmiennej nie zależy od wartości drugiej zmiennej. Ponadto prawdziwe są twierdznia, że zmienne losowe X oraz Y są niezależne wtedy i tylko wtedy, gdy zachodzi jedna z poniższych równości

$$P(x,y) = P_1(x) \cdot P_2(y), \tag{4.20}$$

$$F(x,y) = F_1(x) \cdot F_2(y), \tag{4.21}$$

$$f(x,y) = f_1(x) \cdot f_2(y). \tag{4.22}$$

Przykład 4.2. Dana jest dwuwymiarowa zmienna losowa $Z = (X, Y)$ dyskretna, której wartości prawdopodobieństw podane są w poniższej tablicy

		y	
		1	2
x	-1	$\frac{3}{16}$	$\frac{5}{16}$
	0	$\frac{3}{32}$	$\frac{5}{32}$
	1	$\frac{3}{32}$	$\frac{5}{32}$

a) Wyznaczyć rozkłady brzegowe zmiennych losowych X i Y.

b) Wyznaczyć prawdopodobieństwa warunkowe tych zmiennych.

c) Wykazać, że zmienne są niezależne.

Rozwiązanie. Mamy a)

$$P_2(1) = \frac{3}{16} + \frac{3}{32} + \frac{3}{32} = \frac{3}{8}, \quad P_2(2) = \frac{5}{8}, \quad P_1(-1) = \frac{3}{16} + \frac{5}{16} = \frac{1}{2},$$

$$P_1(0) = \frac{1}{4}, \quad P_1(1) = \frac{1}{4}.$$

b) Korzystając ze wzorów (4.11) oraz (4.12) otrzymujemy

$$\begin{aligned}
P(X=-1|Y=1) &= \frac{P(X=-1, Y=1)}{P_2(1)} = \frac{\frac{3}{16}}{\frac{3}{8}} = \frac{1}{2},\\
P(X=-1|Y=2) &= \frac{\frac{5}{16}}{\frac{5}{8}} = \frac{1}{2},\\
P(X=0|Y=1) &= \frac{\frac{3}{32}}{\frac{3}{8}} = \frac{1}{4},\\
P(X=0|Y=2) &= \frac{\frac{5}{32}}{\frac{5}{8}} = \frac{1}{4},\\
P(X=1|Y=1) &= \frac{\frac{3}{32}}{\frac{3}{8}} = \frac{1}{4},\\
P(X=1|Y=2) &= \frac{\frac{5}{32}}{\frac{5}{8}} = \frac{1}{4}.\\
P(Y=1|X=-1) &= \frac{P(X=-1, Y=1)}{P_1(-1)} = \frac{\frac{3}{16}}{\frac{1}{2}} = \frac{3}{8},\\
P(Y=2|X=0) &= \frac{\frac{3}{32}}{\frac{1}{4}} = \frac{3}{8},\\
P(Y=1|X=1) &= \frac{\frac{3}{32}}{\frac{1}{4}} = \frac{3}{8},\\
P(Y=2|X=-1) &= \frac{\frac{5}{16}}{\frac{1}{2}} = \frac{5}{8},\\
P(Y=2|X=0) &= \frac{\frac{5}{32}}{\frac{1}{4}} = \frac{5}{8},\\
P(Y=2|X=1) &= \frac{\frac{5}{32}}{\frac{1}{4}} = \frac{5}{8}.
\end{aligned}$$

c) Mamy

$$\begin{aligned} P_1(-1)\cdot P_2(1) &= \frac{1}{2}\cdot\frac{3}{8} = \frac{3}{16} = P(X=-1, Y=1), \\ P_1(-1)\cdot P_2(2) &= \frac{1}{2}\cdot\frac{5}{8} = \frac{5}{16}, \\ &\text{itd.} \end{aligned}$$

co dowodzi, że zmienne te są niezależne. □

4.2 Charakterystyki zmiennych losowych dwuwymiarowych

Brzegowe charakterystyki, które informują nas o własnościach zmiennych brzegowych X i Y dane są wzorami dla dysktrenej zmiennej losowej

$$E(X) = \sum_x x P_1(x), \tag{4.23}$$

$$\sigma^2(X) = \sum_x (x - E(X))^2 \cdot P_1(X), \tag{4.24}$$

a dla zmiennej ciągłej

$$E(X) = \int_{-\infty}^{\infty} x f_1(x) dx, \tag{4.25}$$

$$\sigma^2(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f_1(x) dx. \tag{4.26}$$

Analogicznie definiuje się te charakterystyki dla zmiennej Y.

Charakterystyki zmiennych warunkowych definiujemy wzorami

$$E(X|y) = \begin{cases} \sum_x x P(x|y), \\ \int_{-\infty}^{\infty} x f(x|y) dx, \end{cases} \tag{4.27}$$

$$\sigma^2(X|y) = \begin{cases} \sum_x (x - E(X|y))^2 \cdot P(x|y), \\ \int_{-\infty}^{\infty} (x - E(X|y))^2 \cdot f(x|y) dx. \end{cases} \tag{4.28}$$

Analogicznie definiujemy $E(Y|x), \sigma^2(Y|x)$.

Charakterystyki, które dostarczają nam informację o zależnościach między zmiennymi X i Y. Do tych charakterystyk należy kowariancja $C(X,Y)$ oraz współczynnik korelacji $\varrho(X,Y)$.

Kowariancja jest zdefiniowana jako wartość oczekiwana iloczynu odchyleń zmiennych X i Y od ich wartości oczekiwanych

$$C(X,Y) = E\left[(X - E(X)) \cdot (Y - E(Y))\right]. \tag{4.29}$$

Przy obliczeniach wygodnie jest skorzystać ze wzoru

$$C(X,Y) = E(XY) - E(X)E(Y). \tag{4.30}$$

Kowariancja może osiągać wartości ze zbioru $(-\infty, \infty)$ i pomaga nam stwierdzić o istnieniu lub jego braku między zmiennymi.

Użyteczniejszą charakterystyką jest współczynnik korelacji liniowej $\varrho(X,Y)$ dany wzorem

$$\varrho(X,Y) = \frac{C(X,Y)}{\sigma(X) \cdot \sigma(Y)}. \tag{4.31}$$

Przyjmuje wartości z przedziału $\langle -1, 1 \rangle$. Jeżeli jego wartość jest równa ± 1, to wtedy między zmiennymi X i Y mamy zależność liniową, natomiast gdy jest równa 0, to nie ma zależności liniowej między tymi zmiennymi losowymi.

Przykład 4.3. Zmienna losowa $Z = (X, Y)$ ma gęstość prawdopodobieństwa zadaną wzorem

$$f(x,y) = \begin{cases} 2 & \text{dla } 0 < x < y < 1, \\ 1 & \text{poza.} \end{cases}$$

Wyznaczyć

a) brzegowe wartości oczekiwane i wariancje;

b) $E(X|y), \sigma^2(X|y)$;

c) $C(X,Y), \varrho(X,Y)$.

Rozwiązanie. Najpierw wyznaczymy gęstości brzegowe

oraz warunkową gęstość prawdopodobieństwa zmiennej losowej X, jeżeli zmienna Y przyjmuje wartość y

$$f(x|y) = \frac{2}{2y} = \frac{1}{y} \qquad \text{dla } 0 < x < y, 0 < y < 1.$$

Poza wartościami wyróżnionymi te gęstości są zerowe.

a) Mamy

$$E(X) = \int_0^1 2x(1-x)dx = \frac{1}{3}, \qquad E(Y) = \int_0^1 2y^2 dy = \frac{2}{3},$$

$$\sigma^2(X) = \int_0^1 2x^2(1-x)dx - \left(\frac{1}{3}\right)^2 = \frac{1}{18}, \qquad \sigma^2(Y) = \int_0^1 2y^3 dy - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

b) Korzystając ze wzrorów wcześniej podanych otrzymujemy

$$E(X|y) = \int_0^y \frac{x}{y} dx = \frac{y}{2}, \quad \sigma^2(X|y) = \int_0^y \left(x - \frac{y}{2}\right)^2 \frac{1}{y} dx = \frac{y^2}{12}.$$

c) Korzystając ze wzorów na kowariancję i współczynnik korelacji otrzymujemy

$$C(X,Y) = \int_0^1 \left[\int_0^y 2xydx\right] dy - \left(\frac{2}{3}\right)^2 = \frac{1}{36}$$

$$\varrho(X,Y) = \frac{\frac{1}{36}}{\sqrt{\frac{1}{18} \cdot \frac{1}{18}}} = \frac{1}{2}.$$

$\square$

4.3 Funkcje zmiennych losowych

Funkcje jednej zmiennej losowej

W niektórych zagadnieniach spotykamy się z sytuacją, że znamy rozkład prawdopodobieństwa zmiennej losowej X a interesuje nas rozkład zmiennej losowej Y, która jest funkcją zmiennej losowej X

$$Y = y(X). \tag{4.32}$$

Jeżeli funkcja $y(x)$ w zbiorze możliwych wartości zmiennej X jest ściśle monotoniczna, tzn. jeżeli ma funkcję odwrotną $x = y^{-1}(y) = x(y)$, to istnieje między zmiennymi X i Y wzajemnie jednoznaczna odpowiedniość i łatwo wyznaczyć rozkład prawdopodobieństwa zmiennej Y.

Jeżeli funkcja $y(x)$ jest rosnąca, to dystrybuanta zmiennej losowej Y zadana jest wzorem

$$G(y) = P(Y < y) = P(X \leqslant x(y)) = F(x(y)), \tag{4.33}$$

jeżeli natomiast funkcja $y(x)$ jest malejąca, to dystrybuanta zmiennej Y dana jest wzorem

$$G(y) = P(Y \leqslant y) = P(X \geqslant x(y)) = 1 - F(x(y)). \tag{4.34}$$

Jeżeli zmienna losowa X jest zmienną ciągłą o gęstości $f(x)$ oraz jeżeli funkcja $x(y)$ ma we wszystkich punktach wewnętrznych przedziału możliwych wartości ciągłą pochodną, to wtedy gęstość prawdopodobieństwa $g(y)$ zmiennej losowej Y dla rosnącej $y(x)$ dana jest wzorem

$$g(y) = \frac{dG(y)}{dy} = f(x(y)) \cdot x'(y) \tag{4.35}$$

a dla malejącej $y(x)$

$$g(y)\frac{dG(y)}{dy} = -f(x(y)) \cdot x'(y). \tag{4.36}$$

Jeżeli funkcja $y(x)$ w obszarze możliwych wartości zmiennej losowej X nie jest ściśle monotoniczna, to wtedy nie istnieje związek pomiędzy zmiennymi losowymi X i Y wzajemnie jednoznaczny związek.

Funkcje dwóch ciągłych zmiennych losowych

Jeżeli znamy łączną gęstość prawdopodobieństwa $f(x_1, x_2)$ zmiennych losowych X_1, X_2 a interesuje nas rozkład zmiennej losowej Y, która jest funkcją tych dwóch zmiennych losowych

$$Y = y(X_1, X_2). \tag{4.37}$$

Dystrybuanta zmiennej losowej Y

$$G(y) = P(Y \leqslant y) = P(y(x_1, x_2) \leqslant y) \tag{4.38}$$

uzyskujemy całkując gęstość prawdopodobieństwa $f(x_1, x_2)$ po zbiorze S takim, że $y(x_1, x_2) < y$

$$G(y) = \iint\limits_S f(x_1, x_2) dx_1 dx_2. \tag{4.39}$$

Różniczkując dystrybuantę $G(y)$ otrzymujemy gęstość prawdopodobieństwa zmiennej Y

$$g(y) = \frac{dG(y)}{dy}. \tag{4.40}$$

Przykład 4.4. Zmienna losowa dyskretna X dana jest

x_i	-1	0	1
p_i	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Wyznaczyć zmienną losową $Y = X^2$.

Rozwiazanie. Mamy

y_i	$(-1)^2$	0^2	1^2
p_i	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

tj.

y_i	0	1
p_i	$\frac{1}{3}$	$\frac{1}{6} + \frac{1}{2} = \frac{2}{3}$

.

□

Przykład 4.5. Pomiędzy zmiennymi losowymi X i Y jest zależność

$$Y = 2X + 3.$$

X jest ciągłą zmienną losową o dystrybuancie $F(x)$. Wyznaczyć gęstość prawdopodobieństwa $g(y)$?

Rozwiązanie. Dystrybuanta zmiennej losowej Y dana jest

$$G(y) = P(Y < y) = P(2X + 3 < y) = P\left(X < \frac{y-3}{2}\right) = F\left(\frac{y-3}{2}\right).$$

Różniczkójąc dystrybuantę $G(y)$ otrzymujemy gęstość prawdopodobieństwa zmiennej losowej Y

$$g(y) = \frac{dG(y)}{dy} = \frac{1}{2} f\left(\frac{y-3}{2}\right),$$

gdzie $f = \frac{d}{dx} F(x)$. □

Przykład 4.6. Wyznaczyć funkcję prawdopodobieństwa zmiennej losowej Y, jeżeli

$$P(X) = \begin{cases} \frac{1}{3} & \text{dla } x = 1, 2, 3, \\ 0 & \text{poza.} \end{cases}$$

oraz $Y = 2X + 1$.

Rozwiązanie. Mamy $y(x) = 2x + 1$, więc

$$P(y) = \begin{cases} \frac{1}{3} & \text{dla } y = 3, 5, 7, \\ 0 & \text{poza.} \end{cases}$$

□

Przykład 4.7. Zmienna losowa Y jest funkcją ciągłej zmiennej losowej X. Wyznaczyć gęstość prawdopodobieństwa $g(y)$, jeżeli gęstość zmiennej losowej X dana jest

$$f(x) = \begin{cases} 2x & \text{dla } 0 < x < 1, \\ 0 & \text{poza,} \end{cases}$$

oraz $Y = X^3$.

Rozwiazanie. Mamy $y(x) = x^3$ dla $0 < x < 1$. W tym przedziale funkcja $y(x)$ jest rosnąca a funkcją do niej odwrotną jest $x(y) = \frac{1}{2}y^{\frac{1}{3}}$ dla $0 < y < 8$. Na mocy wzoru (4.35) otrzymujemy

$$g(y) = \begin{cases} 2 \cdot \frac{1}{2}y^{\frac{1}{3}} \cdot \frac{1}{6}y^{-\frac{2}{3}} = \frac{1}{6}y^{-\frac{1}{3}} & \text{dla } 0 < y < 8, \\ 0 & \text{poza.} \end{cases}$$

□

Przykład 4.8. Niech X będzie zmienną losową o wartości oczekiwanej $E(X) = -1$ i wariancji $\sigma^2(X) = 4$. Rozważmy zmienną losową

$$Y = 2 - 3X.$$

Wyznaczyć wartość średnią, wariancję zmiennej Y oraz kowariancję i współczynnik korelacji zmiennych X, Y.

Rozwiązanie. Z własności wartości oczekiwanej i wariancji otrzymujemy

$$E(Y) = E(2 - 3X) = E(2) - 3E(X) = 2 - 3 \cdot (-1) = 5,$$
$$\sigma^2(Y) = \sigma^2(2 - 3X) = \sigma^2(2) + \sigma^2(-3X) = 0 + (-3)^2 D(X) = 9 \cdot 4 = 36$$

korzystając ponadto z definicji kowariancji otrzymujemy

$$\begin{aligned} C(X, Y) &= E[XY] - E(X)E(Y) = E(X(2 - 3X)) - E(X)E(Y) = \\ &= 2E(X) - 3E(X^2) + 5 \end{aligned}$$

a ponieważ

$$E(X^2) = \sigma^2(X) + (E(X))^2 = 4 + (-1)^2 = 5,$$

więc

$$C(X, Y) = 2 \cdot (-1) - 3 \cdot 5 + 5 = -12$$

oraz współczynnik korelacji
$$\varrho(X,Y) = \frac{-12}{\sqrt{4 \cdot 36}} = -1,$$
co oznacza, że między zmiennymi losowymi X i Y mamy zależność liniową (tak przecież została określona zmienna Y), czego się należało spodziewać.

□

4.4 Twierdzenia graniczne

Do tej pory zajmowaliśmy się zmienną losową o rozkładzie teoretycznym, któremu przypisawaliśmy teoretyczne charakterystyki. Jeżeli jednak powtórzymy niezależnie pewne doświadczenie losowe, możemy z obserwowanych wartości rozkład względnych częstości i informacje o tym rozkładzie sprowadzić znowu do charakterystyk. Ten rozkład, ewentualnie jego charakterystyki nazwiemy dla odróżnienia od poprzednich empirycznym rozkładem, ewentualnie empirycznymi charakterystykami.

Przy zachowaniu pewnych warunków możemy oczekiwać, że rozkład empiryczny (ewntualnie jeto charakterystyki) będzie się zbliżało do rozkładu teoretycznego (ewentualnie teoretycznych charakterystyk), tym bardziej im więcej będzie realizowanych doświadczeń. Musimy jednak uświadomić sobie, że zbieżność wartości empirycznych do wartości teoremtycznych nie ma charakteru zbieżności matematycznej ale zbieżności w sensie prawdopodobieństwa.

Definicja 4.9. Mówimy, że ciąg zmiennych losowych $X_1, X_2, \ldots, X_n, \ldots$ jest zbieżny do zmiennej X według prawdopodobieństwa 1, jeżeli

$$\bigwedge_{\varepsilon>0} \lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

Podamy teraz kilka twierdzeń dotyczących własności granicznych sum zmiennych losowych. Prawa wielkich liczb

Twierdzenie 4.10 (Bernoullego). *Niech X_n, $n = 1, 2, \ldots$ będzie ciągiem zmiennych losowych o rozkładzie Berunullego z parametrami n, p, gdzie $0 < p < 1$. Dla dowolnego $\varepsilon > 0$ zachodzi*

$$\lim_{n\to\infty} P\left[\left|\frac{X_n}{n} - p\right| < \varepsilon\right] = 1.$$

Twierdzenie 4.11. *Niech $X_1, X_2, \ldots$ będą parami niezależnymi zmiennymi losowymi takimi, że*

$$E(X_i) = a, \quad D(X_i) < c \quad i = 1, 2, \ldots$$

gdzie $|a| < \infty, c < \infty$. Wtedy dla dowolnego $\varepsilon > 0$

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \varepsilon\right] = 1.$$

oraz centralne twierdznie graniczne Lindeberga-Levy'ego

Twierdzenie 4.12 (Lindeberga-Levy'ego). *Niech $X_1, X_2, \ldots$ będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie z wartością oczekiwaną $E(X_k) = \mu$ oraz odchyleniem standardowym $\sigma(X_k) = \sigma$, $k = 1, 2, \ldots$. Wtedy*

$$\lim_{n\to\infty} P\left(a < \frac{S_n - E(X_n)}{\sigma} \le b\right) = \Phi(b) - \Phi(a), \tag{4.41}$$

gdzie $S_n = \sum\limits_{k=1}^{n} X_k$ oraz Φ oznacza dystrybuantę rozkładu $N(0, 1)$.

Szczególnym przypadkiem powyższego twierdzenia jest

Twierdzenie 4.13 (Moiver'a-Laplace'a)**.** *Niech* X_k*,* $k = 1, 2, 3, \ldots$ *będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie postaci*

$$P(X_k = 1) = 1 - P(X_k = 0) = p$$

, $k = 1, 2, 3, \ldots$. *Wtedy*

$$\lim_{n\to\infty} P\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \Phi(b) - \Phi(a). \tag{4.42}$$

4.5 Zadania do samodzielnego rozwiązania

4.1. Zmienna losowa X dana jest jak w Przykładzie 4.4. Wyznaczyć zmienną losową $Y = X^3$.

Odp. Zmienna Y ma taki sam rozkład jak zmienna X.

4.2. Zmienna losowa dyskretna X ma rozkład prawdopodobieństwa

$$\begin{array}{c|c|c|c} x_i & -2 & 0 & 2 \\ \hline p_i & \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \end{array}.$$

Wyznaczyć zmienną losową $Y = X^3 - 2$.

Odp. $\begin{array}{c|c|c|c} x_i & -10 & -2 & 6 \\ \hline p_i & \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \end{array}$

Wykład 5

Elementy statystyki opisowej

Statystyka opisowa zajmuje się opracowaniem danych statystycznych bez posługiwania się rachunkiem prawdopodobieństwa. Pozwala przedstawić dane w sposób uporządkowany, dający możliwość ich analizy.

Populacją statystyczną nazywamy zbiór wszystkich możliwych elementów (jednostek), które podlegają badaniu. Przykładem populacji są np. wszystkie elementy wyprodukowane przez daną maszynę, mieszkańcy Polski, mieszkańcy Warszawy itp.

Ponieważ często populacja jest zbyt duża aby można było przeprowadzić badanie całej populacji (np. ze względu na koszty, lub czas potrzebny do realizacji), więc wybiera się podzbiór (próbę) z populacji, która powinna być reprezentatywna dla całej populacji, tzn. aby badanie przeprowadzone na części populacji można było odnieść do wszystkich elementów populacji.

Próba (próba losowa) jest podzbiorem elementów (jednostek) populacji.

Jednostki statystyczne charakteryzują się pewnymi właściwościami, które określa się mianem *cech statystycznych*. Cechy statystyczne ogólnie dzieli się na

1. *Cechy niemierzalne* (jakościowe). Są to na ogół określane słownie np. płeć, rozmieszczenie przestrzene czy geograficzne.

2. *Cechy mierzalne* (ilościowe). Są to właściwości, które można zmierzyć i wyrazić za pomocą jednostek fizycznych, np. waga, wysokość, długość, ilość itp. Ze względu na przyjmowane wartości cechy mierzalne dzielimy na:

   (a) *dyskretne* (skokowe), to takie, które przyjmują skończony lub przeliczalny zbiór wartości na danej skali liczbowej, przy czym jest to na ogół zbiór liczb naturalnych (np. liczba dzieci w rodzinie, ilość wyprodukowanych elementów przez fabrykę itp.).

   (b) *ciągłe*, to takie, które mogą przyjąć każdą wartość z określonego przedziału liczbowego $\langle a, b\rangle$.

5.1 Dane statystyczne

Materiał otrzymany w wyniku przeprowadzonej obserwacji statystycznej należy odpowiednio usystematyzować i pogrupować w postaci tzw. szeregów statystycznych.

Szeregiem statystycznym nazywamy ciąg wielkości statystycznych uporządkowany według określonych kryteriów.

Ze względu na kryteria uporządkowania szeregi statystyczne dzielimy na

Szeregi szczegółowe są to uporządkowane ciągi wartości badanej cechy statystycznej. Taki sposób prezentacji danych statystycznych jest stosowany na ogół w przypadku, gdy przedmiotem badania jest niewielka liczba jednostek. Załóżmy, że zmienna X przyjmuje wartości $x_1, x_2, \ldots, x_n$. Wartości tej cechy możemy uporządkować rosnąco

$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n \tag{5.1}$$

lub malejąco

$$x_1 \geqslant x_2 \geqslant \ldots \geqslant x_n. \tag{5.2}$$

Szereg rozdzielczy stanowi zbiorowość statystyczną podzieloną na części (klasy) według określonej cechy jakościowej lub ilościowej z podaniem liczebności każdej z wyodrębnionych klas. W przypadku szeregów rozdzielczych cechy ilościowej jej warianty można określić

- *punktowo*, wtedy szereg ma postać

i	x_i	n_i	n_i^{sk}
1	x_i	n_1	n_1^{sk}
2	x_i	n_2	n_2^{sk}
3	x_i	n_3	n_3^{sk}
$\ldots$	$\ldots$	$\ldots$	$\ldots$
k	x_k	n_k	n_k^{sk}
$\sum$		n	

gdzie x_i jest i-tą wartościa badanej cechy oraz n_i jest licznością cechy x_i w badanej próbce a n_i^{sk} są licznościami skumulowanymi, tj. $n_s^{sk} = n_1+n_2+\ldots+n_s$.

- *przedziałowo*, wtedy szerego ma postać

i	$x_{0,i}$	$x_{1,i}$	$\dot{x_i}$	n_i	n_i^{sk}
1	$x_{0,1}$	$x_{1,1}$	$\dot{x_1}$	n_1	n_1^{sk}
2	$x_{0,2}$	$x_{1,2}$	$\dot{x_2}$	n_2	n_2^{sk}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
k	$x_{0,k}$	$x_{1,k}$	$\dot{x_k}$	n_k	n_k^{sk}
$\sum$				n	

gdzie $x_{0,i}$ jest początkiem i-tego przedziału, $x_{1,i}$ jest końcem i-tego przedziału, $\dot{x_i}$ - jest reprezentantem i-tego przedziału oraz n_i licznością i-tego przedziału a n_i^{sk} jak poprzednio licznościami skumulowanymi.

5.2 Miary położenia, zróżnicowania, asymetrii

Miary położenia (tendencji centralnej)

Miary tendencji centralnej służą do wyznaczenia wartości cechy, wokół której skupiają się dane. Czyli można taką wartość cechy mierzalnej uważać za „typowego reprezentanta" naszych danych. Do najczęściej używanych miar tendencji centralnej należą: średnia arytmetyczna, mediana i dominanta. Przyjmujemy oznaczenia n jest to liczba elemntów w próbie, k jest to liczba klas (przedziałów) w szeregu rozdzielczym.

Średnia arytmetyczna dla poszczególnych szeregów wyraża się wzorami

- szczegółowego

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{5.3}$$

- rozdzielczego punktowego

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i, \tag{5.4}$$

- rozdzielczego przedziałowego

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i \cdot n_i. \tag{5.5}$$

Mediana jest to element środkowy w uporządkowanej próbie (zbiorowości) cechy X. Obliczamy ją ze wzorów

- dla szeregu szczegółowego

$$\mathrm{Me} = \begin{cases} x_{\frac{n+1}{2}} & \text{dla } n \text{ nieparzystego,} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{dla } n \text{ parzystego,} \end{cases} \tag{5.6}$$

- dla szeregu rodzielczego punktowego, określamy pozycję mediany tak jak dla szeregu szczegółowego i odczytujemy w którym przedziale dana pozycja się znajduje. Wartość tego przedziału przyjmujemy za medianę.
- dla szeregu rozdzielczego przedziałowego, wyznaczamy pozycję mediany ze wzoru $\frac{n}{2}$ i patrzymy w którym przedziale znajduje się mediana. Następnie wartość mediany liczymy ze wzoru

$$M = x_{0m} + \frac{\frac{1}{2}n - \sum\limits_{i=1}^{m-1} n_i}{n_m} h_m, \tag{5.7}$$

gdzie x_{0m} - dolna granica przedziału mediany; $\frac{n}{2}$ - pozycja mediany; $\sum\limits_{i=1}^{m-1} n_i$ - liczność wszystkich przedziałów poprzedzających przedział mediany (bez liczebności klasy mediany); n_m - liczność przedziału mediany; h_m - długość przedziału mediany; m - numer przedziału mediany.

Dominanta. Oznaczamy ją przez D. Definiuje się ją dla

- szereg rozdzielczy punktowy - jest to jest to ta cecha, która występuje najczęściej. Jeżeli najczęściej występującą cechą jest x_d, to $D = x_d$.

- szereg rozdzielczy przedziałowy - najpierw wyznaczamy klasę najliczniejszą a następnie dominantę wyliczamy ze wzoru

$$D = x_{0d} + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} h_d, \tag{5.8}$$

gdzie x_{0d} - dolna granica przedziału dominanty; n_d - liczebność przedziału dominanty; n_{d-1} - liczebność przedziału poprzedzającego przedział dominanty; n_{d+1} - liczebność przedziału następnego po przedziale dominanty; h_d - rozpiętość przedziału dominanty.

Kwartyle. Definiujemy je jako wartości cechy badanej zbiorowości, przestawionej w postaci szeregu statystycznego, które dzielą zbiorowość na określone części pod względem liczby jednostek. *Kwartyl pierwszy* Q_1 - dzieli zbiorowość na dwie różne części w ten sposób, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu Q_1, 75% równe bądź wyższe od tego kwartyla. *Kwartyl drugi* – to jest modalna. *Kwartyl trzeci* Q_3 - dzieli zbiorowśc na dwie części w ten sposób, że 75% jednostek ma wartości cechy nieższe bądź równe Q_3, a 25% równe bądź wyższe od tego kwartyla.

W przypadku szeregów szczegółowych kwartyle pierwszy i trzeci wyznacza się analogicznie jak medianę. Można bowiem przyjąć, że zbiorowość podzielimy na dwie części: pierwszą, której jednostki przyjmują wartości mniejsze od mediany oraz drugą w której przyjmują wartości większe od mediany.

W szeregach rozdzielczych wyznaczenie kwartyli poprzedza ustalenie ich pozycji według wzorów

$$N_{Q_1} = \frac{n}{4}, \tag{5.9}$$

$$N_{Q_3} = \frac{3n}{4}. \tag{5.10}$$

Do szeregów rozdzielczych przedziałowych stosujemy wzory

$$Q_1 = x_{0m} + \frac{N_{Q_1} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m, \tag{5.11}$$

$$Q_3 = x_{0m} + \frac{N_{Q_3} - \sum_{i=1}^{m-1} n_i}{n_m} \cdot h_m, \tag{5.12}$$

gdzie m - numer przedziału (klasy), w którym występuje odpowiadający mu kwartyl, x_{0m} - dolna granica tego przedziału, n_m- liczność przedziału, w którym występuje odpowiedni kwartyl, $\sum_{i=1}^{m-1} n_i$ - liczność skumulowana przedziału poprzedzającego przedział odpowiedniego kwartyla, h_m - długość przedziału klasowego, w którym jest odpowiedni kwartyl.

Przykład 5.1. Dwóch pracowników wykonuje detale tego samego typu. Przeprowadzono obserwację czasu wykonywania pięciu detali przez robotnika pierwszego R_1 oraz sześciu dla pracownika drugiego R_2. Otrzymano wyniki (w min)

- R_1 - 13, 16, 16, 19, 21,
- R_2 - 11, 11, 13, 13, 15, 15.

Średnie dla obu robotników

$$
\begin{aligned}
\overline{x}_{R_1} &= \frac{13+16+16+19+21}{5} = 17 \text{ min},\\
\overline{x}_{R_2} &= \frac{11+11+13+13+15+15}{6} = 13 \text{ min}.
\end{aligned}
$$

Dominanty $D_{R_1} = 16$ min, ponieważ najczęściej występującą wartościa jest 16, natomiast w przypadku robotnika drugiego dominanty nie jesteśmy w stanie wyznaczyć.

Mediana w przypadku robotnika piewszego. Mamy pięć elementów, jest to liczba nieparzysta, więc za wartość mediany przyjamujemy wartość elementu stojącego na miejscu $\frac{5+1}{2} = 3$, czyli $Me_{R_1} = x_3 = 16$ min. Natomiast w przypadku drugim mamy parzystą liczbę elementów, zatem z faktu iż $\frac{6}{2} = 3$ wynika, że $Me_{R_2} = \frac{1}{2}(x_{\frac{6}{2}} + x_{\frac{6}{2}+1}) = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(13 + 13) = 13$ min.

Kwartyle dla pierwszego robotnika. Dzielimy nasze dane na dwie cześci 13, 16, 16 oraz 16, 19, 21. Stąd otrzymujemy, że $Q_{1,R_1} = 16$ oraz $Q_{3,R_1} = 19$. Natomiast dla drugiego robotnika 11, 11, 13 oraz 13, 15, 15. Stąd otrzymujemy, że $Q_{1,R_2} = 11$ oraz $Q_{3,R_2} = 15$.

Charakterystyki zróżnicowania (rozproszenia)

Miary zróżnicowania zwane także miarami rozproszenia lub dyspresji pozwalają nam stwierdzić, czy dane są bardzo rozproszone czy też bardziej skoncentrowane, tj. mierzą jak się zachowują wokół miary centralnej (np. średniej).

Wariancja

- szereg szczegółowy

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{5.13}$$

- szereg rozdzielczy punktowy

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_i, \tag{5.14}$$

- szereg rozdzielczy przedziałowy

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 \cdot n_i. \tag{5.15}$$

Odchylenie standardowe

$$s = \sqrt{s^2}. \tag{5.16}$$

Odchylenie ćwiartkowe

$$Q = \frac{Q_3 - Q_1}{2}. \tag{5.17}$$

Współczynnik zmienności

$$V = \frac{s}{|\overline{x}|}100\%, \tag{5.18}$$

przy założeniu, że $\overline{x} \neq 0$.

Roztęp

$$R = x_{\max} - x_{\min}, \tag{5.19}$$

gdzie $x_{\max}$ - największa dana statystyczna, $x_{\min}$ - namniejsza dana statystyczna.

Przykład 5.2. Dla danych z Przykładu 5.1 obliczymy miary rozproszenia. Mamy dla pracownika pierwszego

$$s^2_{R_1} = \frac{1}{5}\sum_{i=1}^{5}(x_i - \overline{x}_{R_1})^2 = \frac{1}{5}\left[(13-17)^2 + (16-17)^2 + (16-17)^2 + (19-17)^2 + (21-17)^2\right] = 7,6,$$

stąd $s_{R_1} = \sqrt{s^2_{R_1}} = \sqrt{7,6} = 2,76$. Natomiast dla robotnika drugiego

$$s^2_{R_2} = \frac{1}{6}\sum_{i=1}^{6}(x_i - \overline{x})^2 = 2,7,$$

stąd $s_{R_2} = 1,63$.

Współczynnik zmienności

$$V_{s,R_1} = \frac{2,76}{17}100\% = 16,2\%, \quad V_{s,R_2} = \frac{1,63}{13}100\% = 12,5\%.$$

Rozstęp $R_{R_1} = 21 - 13 = 8$, $R_{R_2} = 15 - 11 = 4$.

Odchylenie ćwiartkowe

$$Q_{R_1} = \frac{19-16}{2} = 1,5, \quad Q_{R_2} = \frac{15-11}{2} = 2.$$

Zatem możemy stwierdzić, że drugi pracownik wykonuje dany detal szybciej oraz różnice w czasie wykonywania tego detalu dla drugiego pracownika są mniejsze. Chociaż z odchylenia ćwiartkowego wynikałoby by coś odwrotnego.

Charakterystyki asymetrii

Asymtria mówi nam z której strony wartości centralnej (np. średniej) bardziej skupiają się wartości badanej cechy.

Współczynnik asymetrii obliczamy ze wzoru

$$A = \frac{\mu_3}{(s)^3}, \tag{5.20}$$

gdzie s - oznacza odchylenie standardowe, a μ_3 - trzeci moment centralny, który obliczamy ze wzorów

- szeregu szczegółowego

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3, \tag{5.21}$$

- szeregu rozdzielczego punktowego

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 \cdot n_i, \tag{5.22}$$

- szeregu rozdzielczego przedziałowego

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(\dot{x}_i - \overline{x})^3 \cdot n_i. \tag{5.23}$$

Gdy współczynnik asymetrii równa się 0 to mówimy, że rozkład jest symetryczny. Gdy jest ujemny to mówimy o asymetrii lewostronnej, w przeciwnym przypadku o prawostronnej.

Przykład 5.3. Policzymy asymetrię dla danych z Przykładu 5.1.

Dla robotnika pierwszego mamy

$$\mu_3 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \overline{x})^3 = \frac{1}{5}\left[(13-17)^3 + (16-17)^3 + (16-17)^3 + (19-17)^3 + (21-17)^3\right] = 1,2$$

więc $A_{R_1} = \frac{1,2}{(2,76)^3} = 0,057$. Analogicznie dla robotnika drugiego

$$\mu_3 = \frac{1}{6}\sum_{i=1}^{6}(x_i - \overline{x})^3 = 0,$$

więc $A_{R_2} = \frac{0}{(1,63)^3} = 0$.

Przykład 5.4. W grupie 100 studentów przeprowadzono badanie liczby wypalanych dziennie papierosów. Oznaczając przez x_i liczbę wypalanych dziennie papierosów a przez n_i liczbę studentów (wypalających taką liczbę papierosów) otrzymanow wyniki

x_i	0	5	10	15	20	25	30
n_i	5	10	20	30	20	10	5

Wyznaczyć średnią, wariancję, odchylenie standardowe, współczynnik zmienności, asymetrię, medianę, dominantę, kwartyle (pierwszy i trzeci), odchylnie ćwiatkowe.

Rozwiązanie. Dla wygody obliczenia wykonujemy w tabeli

i	x_i	n_i	$x_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2 \cdot n_i$	$(x_i - \overline{x})^3 \cdot n_i$	n_i^{sk}	
1	0	5	0	-15	1125	-16875	5	
2	5	10	50	-10	1000	-10000	15	
3	10	20	200	-5	500	-2500	35	Q_1
4	15	30	450	0	0	0	65	Me
5	20	20	400	5	500	2500	85	Q_3
6	25	10	250	10	1000	10000	95	
7	30	5	150	15	1125	16875	100	
$\sum$	$\times$	100	1500	$\times$	5250	0	$\times$	

Zatem otrzymujemy

$$\begin{aligned}
\overline{x} &= \frac{1}{100}\sum_{i=1}^{7} x_i \cdot n_i = \frac{1}{100}1500 = 15(sztuk),\\
s^2 &= \frac{1}{100}\sum_{i=1}^{7}(x_i - \overline{x})^2 \cdot n_i = \frac{1}{100}5250 = 52,5,\\
s &= \sqrt{52,5} \approx 7,2,\\
\mu_3 &= \frac{1}{100}\sum_{i=1}^{7}(x_i - \overline{x})^3 \cdot n_i = \frac{1}{100}0 = 0,\\
A &= \frac{0}{(7,2)^3} = 0,\\
V_s &= \frac{7,2}{15}100\% = 48\%.
\end{aligned}$$

Widzimy stąd, że średnia ilość wypalanych papierosów w tej grupie studentów wynosi 15 sztuk, z ochyleniem standardowym na plus lub minus 7,2 sztuki. Co więcej wiemy, że rozkład jest symetryczny. Współczynnik zmienności wynosi 48%, czyli jest duże zróżnicowanie w tej grupie pod względem wypalanych papierosów. Teraz przejdziemy do wyznaczenia pozostałych charakterystyk. Dominanta jest najprostsza do wyznacznia, wynosi

$$D = 15,$$

ponieważ największa liczba studentów (30) wypala taką ilość papierosów. Mamy 100 wyników, $\frac{100}{2} = 50$, więc szukamy do której klasy wpada 50 element. Korzystając z liczności skumulowanych widzimy, że dla klasy czwartej ($i = 4$). Zatem

$$Me = 15,$$

Liczymy miejsce kwartyli $N_{Q_1} = \frac{100}{4} = 25, N_{Q_3} = \frac{3 \cdot 100}{4} = 75$. Szukamy do których klas należą elementy o tych numerach. Są to odpowiednio klasy trzecia i piąta. Stąd

$$Q_1 = 10, \quad Q_3 = 20.$$

Zatem

$$Q = \frac{20 - 10}{2} = 5.$$

□

Przykład 5.5. Analizując liczbę wyprodukowanych elementów pewnej brygady otrzymano wyniki, które zanotowano w poniższej tabeli, gdzie x_i - liczbę detali, n_i - ilość pracowników wyrabiających daną ilość elementów

x_i	$12-14$	$14-16$	$16-18$	$18-20$
n_i	6	7	11	6

.

Wyznaczyć średnią, wariancję, odchylenie standardowe, współczynnik zmienności, asymetrię, medianę, dominantę, kwartyle (pierwszy i trzeci), odchylnie ćwiatkowe.

Rozwiązanie. Tak jak w zadaniu poprzednim wygodnie będzie wykonywać rachunki w tabeli

i	x_i	n_i	$\dot{x}_i$	$\dot{x}_i \cdot n_i$	$\dot{x}_i - \overline{x}$	$(\dot{x}_i - \overline{x})^2 \cdot n_i$	$(\dot{x}_i - \overline{x})^3 \cdot n_i$	n_i^{sk}	
1	$12-14$	6	13	78	$-3,1$	$57,66$	$-178,75$	6	
2	$14-16$	7	15	105	$-1,1$	$8,47$	$-9,32$	13	Q_1
3	$16-18$	11	17	187	$0,9$	$8,91$	$8,02$	24	Me, D, Q_3
4	$18-20$	6	19	114	$2,9$	$50,46$	$146,33$	30	
$\sum$	$\times$	30	$\times$	484	$\times$	$125,5$	$-33,72$	$\times$	

Zatem

$$\begin{aligned}
\overline{x} &= \frac{1}{30}\sum_{i=1}^{4} \dot{x}_i \cdot n_i = \frac{1}{30} 484 \approx 16,1, \\
s^2 &= \frac{1}{30}\sum_{i=1}^{4} (\dot{x}_i - \overline{x})^2 \cdot n_i = \frac{1}{30} 125,5 \approx 4,2, \\
s &= \sqrt{4,2} \approx 2, \\
\mu_3 &= \frac{1}{30}\sum_{i=1}^{4} (\dot{x}_i - \overline{x})^3 \cdot n_i = \frac{1}{30}(-33,72) \approx -1,1, \\
A &= \frac{-1,1}{2^3} = -0,14, \\
V_s &= \frac{2,0}{16,1} 100\% = 12,4\%.
\end{aligned}$$

Otrzymujemy stąd, że średnio robotnik wykonuje 16,1 sztuki elementu, z odchyleniem standardowym plus, minus 2 elementy, przy dosyść małym zróżnicowaniu (12,4%) oraz niewielkim większym skupieniu poniżej średniej. Pozostałe charakterystyki. Miejsce mediany, to $\frac{30}{2} = 15$. Jest to klasa 3, zatem korzystając ze wzoru (5.7) otrzymujemy

$$Me = 16 + \frac{\frac{1}{2}30 - (6+7)}{11} \cdot 2 = 16 + 0,4 = 16,4.$$

Przedziałem dominanty jest w tym przypadku ten sam przedział, co przedział mediany, zatem

$$D = 16 + \frac{11-7}{(11-7)+(11-6)} \cdot 2 = 16 + 0,9 = 16,9.$$

Miejsce kwartyli $N_{Q_1} = \frac{30}{4} \approx 8, N_{Q_3} = \frac{3\cdot 30}{4} \approx 23$. Zatem

$$\begin{aligned} Q_1 &= 14 + \frac{8-6}{7} \cdot 2 = 14,6, \\ Q_3 &= 16 + \frac{23-13}{11} \cdot 2 = 17,8. \end{aligned}$$

Stąd też

$$Q = \frac{17,8 - 14,6}{2} = 1,6.$$

□

Wszystkie charakterystyki możemy podzielić na miary klasyczne - to te przy wyliczaniu wykorzystujemy wartości wszystkich elementów w próbce oraz miary pozycyjne - to te gdzie wyznaczamy miejsce (pozycję) elementów w uporządkowanej próbce. Pojawia się naturalne pytanie dlaczego nie ograniczyć sie do jednego typów miar? Wykorzystanie jednych czy drugich zależy od kontestu zadnia oraz postaci samych danych. Czasami w próbce pojawia się wartość, która wyraźnie odstaje (jest dużo mniejsza lub dużo większa od naszych danych). Przy liczeniu średniej arytmetycznej zostanie jej wartość uwzględniona przy liczeniu i może w znaczny sposób zawyżyć lub zaniżyć liczoną daną, natomiast przy miarach pozycyjnych nie zostanie to uwzględnione. Z drugiej strony nasze dane mogą być danymi wziętymi np. z Rocznika Statystycznego i dotyczyć dochodów gospodarstw rolnych w zależności od powierzchni. Na ogół w takich tabelach klasy skrajne są podawane jako: poniżej 1ha, powyżej 50ha. I przy liczeniu miar klasycznych mamy problem jakiego wybrać reprezentanta dla tych klas (tj. jakie wybrać $\dot{x}_i$). Zatem użycie konkretnej miary zależy od konkretnego zagadnienia którym się zajmujemy. W podręczniku zostały policzone zarówno jedne jak i drugie mimo, że w konkretnych zadaniach powinno korzystać z albo z jednych miar albo z drugich.

5.3 Zadania do samodzielnego rozwiązania

5.1. W fabryce w ciągu pięciu dni roboczych wyprodukowano pięć wyrobów o wadze: 12, 14, 16, 18, 20. Obliczyć średnią i odchylenie standardowe.

5.2. W pewnej szkole badano wzrost dziewcząt klas czwartych. Otrzymano wyniki: 140, 148 148, 148, 150, 150, 156, 156, 160, 160, 160, 160, 162, 163, 164, 166, 168, 169, 170, 175, 175, 180. Obliczyć: medianę, dominantę, kwartyle pierwszy i drugi oraz odchylenie ćwiartkowe.

5.3. Oceny studentów z przedmiotu statystyka przedstawia tabela

Ocena	3	3,5	4	4,5	5
Liczba studentów	25	30	10	15	20

Obliczyć: średnią, odchylenie standardowe, współczynnik zmienności, asymetrię.

5.4. Poniższa tabela przedstawia dane dotyczące wydajności pracy (w szt/h) pewnego wydziału zakładu produkcyjnego. Wyznaczyć medianę, dominantę, kwartyle pierwszy i trzeci oraz odchylnie ćwiartkowe.

Wydajność	40	50	60	30	25	20
ilość pracowników	15	10	30	10	15	10

5.5. Dla poniższego szeregu rozdzielczego przedziałowego, przedstawiającego staż pracy pracowników pewnego przedsiębiorstwa, obliczyć: średnią, wariancję, odchylenie standardowe, współczynnik zmienności, asymetrię.

przedział	1 − 5	5 − 10	10 − 15	15 − 20	20 − 25	25 − 30	30 − 35
ilość elementów	10	20	25	30	35	10	20

5.6. Dla poniższego szeregu rozdzielczego przedziałowego obliczyć: medianę, dominantę, kwartyl pierwszy i drugi oraz odchylenie ćwiartkowe.

0 − 4	4 − 8	8 − 12	12 − 16	16 − 20	20 − 24	24 − 28
10	20	30	40	10	5	10

Wykład 6

Elementy stytystyki matematycznej

W tym wykładzie omówione są podstawy estymacji oraz testowania hipotez statystycznych.

6.1 Pewne rozkłady stosowane w statystyce

Rozkład chi-kwardrat (χ^2). Jeżeli zmienne losowe $X_1, X_2, \ldots, X_k$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(0,1)$, to zmienną losową χ^2 określamy następująco

$$\chi^2 = \sum_{i=1}^{k} X_i^2,$$

i mówimy, że ma rozkład χ^2 o k „stopniach swobody". Zmienna losowa o rozkładzie chi-kwadrat przyjmuje wartości dodatnie, a jej rozkład zależy od liczby stopni swodoby k. Dla małych wartości k jest to rozkład silnie asymetryczny, natomiast w miarę wzrostu k asymetria jest mniejsza. k wyznaczamy najczęściej jako $k = n-1$, gdzie n jest liczebnością próby. Paremetry tego rozkładu, to

$$E(\chi^2) = k, \quad \sigma(\chi^2) = \sqrt{2k}.$$

Piszemy wtedy $\chi^2 \sim \chi^2(k, \sqrt{2k})$.

Jeżeli k wzrasta, to rozkład chi-kwadrat zbliża się do rozkłady normalnego o tych samych parametrach. Przyjmuje się, że przy $k = 30$ przyliżenie wartości rozkładu chi-kwadrat wartościami rozkładu normalnego jest wystarczająco dokładne.

Rozkład t-Studenta. Jeżeli zmienna losowa Z ma rozkład $N(0,1)$ i χ^2 ma rozkład $\chi^2 \sim \chi^2(k, \sqrt{2k})$, oraz powyższe zmienne losowe są niezależne, to mówimy, że zmienna $T = \frac{Z}{\chi^2}\sqrt{k}$ ma rozkład t-Studenta o k stopniach swobody. Parametry rozkłady t-Studenta

$$E(T) = 0, \text{dla} k \geqslant 2, \quad \sigma(T) = \sqrt{\frac{k}{k-2}}, \text{dla} k \geqslant 3.$$

Dla $k > 30$ zmienna o rozkładzie t-Studenta ma rozkład zbliżony do rozkładu normalnego standaryzowanego $N(0,1)$.

6.2 Estymacja

Mając do dyspozycji jedynie próbkę pobraną z całej populacji losową możemy oszacować wartość interesujących nas parametrów na podstawie tej próbki. Takie szacowanie na podstawie próbki nazywa sie *estymacją*.

Wskaźniki, które możemy obliczyć z próby, będziemy nazywali statystykami, a odpowiadające im wskaźniki dotyczące populacji parametrami populacji.

Dobry estytmator powinien posiadać trzy podstawowe cechy:

- Powinien być *nieobciązony*, co oznacza, że powinien być wolny od błędów systematycznych. Błędy systematyczne to takie, które są popełniane „stale”, np. robiąc pomiary zawsze zawyżamy lub zaniżamy wartość parametru.
- Powinien być *efektywny*, tzn. minimalizuje błąd oszacownia. Inaczej mówiąc powinien mieć jak najmniejszą wariancję.
- Powinien być zgodny, tzn. wraz ze wzrostem liczebności próbki zwiększa się prawdopoodbieństwo, że jego wartość zbliża się do wartości szacowanego paremetru.

Wyróżniamy dwa sposoby szacowania nieznanego parametru: estymacja punktowa i estymacja przedziałowa.

*Estymacja punktowa*polega na wybraniu statystyki na podstawie której będziemy szacowali wartość interesującego nas parametru. Istnieją różne metody wyznaczania estymatorów. My ograniczymy się do padania kilku gotowych estymatorów.

- Estymacja wartości oczekiwanej dla rozkładu normalnego. Jeżeli cecha X z populacji ma rozkład normalny $N(\mu, \sigma)$, przy czym znane jest σ. Wtedy estymatorem wartości oczekiwanej jest średnia z próby. Jest to estymator nieobciążony, efektywny i zgodny.
- Estymacja wariancji dla rozkładu normalnego. Jeżeli cecha X z populacji ma rozkład normalny $N(\mu, \sigma)$, przy czym znane jest μ, to wtedy za estymator wariancji możemy przyjąć $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, który jest estymatorem nieobciążonym, zgodnym i efektywdnym. Jeżeli μ jest nieznane, to za estymator przyjmujemy $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, który jest estymatorem efektywnym, zgodnym i niebociążonym. Zwykła wariancja z próby, którą rozważaliśmy w porzednim rozdziale, jest estymatorem obciążonym.

Estymacja przedziałowa polega na szacowaniu wartości nieznanego parametru za pomocą tzw. przedziału ufności.

Przedziałem ufności nazywamy taki przedział, który z zadanym z góry prawdopodobieństwem $(1 - \alpha)$, zwanym poziomem ufności (lub współczynnikiem ufności), pokrywa nieznaną wartość szacowanego parametru. Interpretacja poziomu ufności: przy wielokrotnym pobieraniu prób n-elementowych i wyznaczaniu na ich podstawie granic przedziałów ufności, średnio w $(1-\alpha)\cdot 100\%$ przypadków otrzymujemy przedziały pokrywające nieznaną wartość. Sposób konstrukcji przedziału ufności związany jest z rozkładem odpowiedniego estymatora. Teraz podamy przedziały ufności dla podstawowych parametrów rozkładu cechy w zbiorowości generalnej.

Przedział ufności dla przeciętne μ. Zakładając, że cecha X w zbiorowości generalnej ma rozkład $N(\mu, \sigma)$ oraz znane jest σ lub $n > 30$. Wtedy przedział ufności dla parametru μ (wartości oczekiwanej) ma postać

$$\overline{x} - t_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + t_\alpha \frac{\sigma}{\sqrt{n}}, \tag{6.1}$$

gdzie t_α odczytuje się z tablic rozkładu normalnego, korzystając z relacji

$$\Phi(t_\alpha) = 1 - \frac{\alpha}{2}.$$

Jeżeli natomiast $n < 30$ i σ jest nieznane, to wtedy przedział przyjmuje postać

$$\overline{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n-1}} < \mu < \overline{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n-1}}, \tag{6.2}$$

gdzie $t_{\alpha,n-1}$ odczytuje się z tablic rozkładu Studenta dla $n-1$ stopni swobody.

Przykład 6.1. Zakładając, że roczne wydatki na paliwo można uznać za cechę o rozkładzie $N(\mu, \sigma)$, pobrano próbę losową liczącą 100 małych zakładów. Uzyskano $\overline{x} = 12$ oraz $s = 4,72$ (w tys. zł). Wyznaczyć przedział ufności dla wartości oczekiwanej na poziomie ufności $1 - \alpha = 0,96$.

Rozwiązanie. Ponieważ $n > 100$ oraz σ jest nieznane, więc korzystamy z przedziału postaci (6.1). Otrzymujemy

$$12 - t_\alpha \frac{4,72}{\sqrt{100}} < \mu < 12 + t_\alpha \frac{4,72}{\sqrt{100}},$$

gdzie $\Phi(t_\alpha) = 1 - \frac{0.04}{2} = 0.98$. Z tablic rozkładu normalnego otrzymujemy $t_\alpha = 2,05$. Zatem

$$\mu \in (11,2; 12,8).$$

□

Przykład 6.2. Poddano analizie wydatki na odzież w wiejskich rodzinach 5-osobowych. Z populacji tych rodzin wylosowano próbę 289-elementów. Na podstawie przeprowadzonych obserwacji ustalono przeciętną skalę wydatków na odzież $\overline{x} = 100$ zł. Badania z lat ubiegłych wykazały, że rozkład wydatków na odzież jest rozkładem normalnym o stałej wariancji $\sigma^2 = 576$. Wyznaczyć przedział ufności średnich miesięcznych wydatków na odzież w wiejskich rodzinach 5-osobowych przyjmując poziom ufności $1 - \alpha = 0,98$.

Rozwiązanie. Korzystając ze wzoru (6.1) dla $\Phi(t_\alpha) = 0.99, t_\alpha = 2,35$. Zatem $96,682 < \mu < 103,318$. W rodzinach 5-osobowych miesięczne wydatki na odzież zawierają się w przedziale $\mu \in (96,68\text{zł}; 103,32\text{zł})$. □

6.3 Testowanie hipotez statystycznych

Hipotezą statystyczną nazywamy każdy sąd o całej populacji, wydany bez przeprowadzenia badania całej populacji. Prawdziwość przypuszczenia (hipotezy) sprawdza się na podstawie próby losowej.

Hipoteza H_0. Jest to hipoteza, której prawdziwość sprawdzamy.

Hipoteza H_1. Jest to hipotez, którą jesteśmy skłonni przyjąć w przypadku odrzucenia hipotezy H_0.

Test statystyczny. Są to reguły postępowania na podstawie których przyjmujemy lub odrzucamy hipotezę $H_{0.}$

Przy testowaniu hipotez statystycznych możemy popełnić dwa błędy. Odrzucić hipotezę H_0 pomimo, że jest ona prawdziwa. Błąd tego rozdzaju nazywamy *błędem I rodzaju.* Lub też możemy przyjąć hipotezę mimo, że jest ona fałszywa. Błąd tego rodzaju nazywamy *błędem II rodzaju.*

Poziom istotności. Oznaczmy przez α i jest to prawdopodobieżstwo popełnienia błędu I rodzaju.

Prawdopodobieństwo popełnienia błędu drugiego rodzaju oznaczamy przez β. Dobry test statystyczny powinien charakteryzować się tym, że β powinno być bliskie zeru.

Sprawdzianem hipotezy nazywamy taką statystykę, której wartość obliczona na podstawie pobranej próby losowej, pozwalana na podjęcie decyzji o orzuceniu (lub nie) hipotezy H_0. Zbiorem krytycznym nazywamy zbiór tych wartości sprawdzianu hipotezy, które przemawiają za odrzuceniem hipotezy H_0.

Zbiór krytyczny jest to zbiór tych wartości sprawdzianu hipotezy, które przemawiają za odrzuceniem hipotezy H_0.

Testy dla wartości oczekiwanej dla jednej próby

Rozważamy hipotezę zerową:

$$H_0: \quad \mu = \mu_0 \tag{6.3}$$

wobec jednej z trzech hipotez alternatywnych:

$$H_1: \quad \mu \neq \mu_0 \tag{6.4}$$
$$H_1': \quad \mu < \mu_0 \tag{6.5}$$
$$H_1'': \quad \mu > \mu_0 \tag{6.6}$$

Model 1 Zakładamy, że badana cecha x ma rozkład normalny $N(\mu, \sigma)$ o znanym odchyleniu standardowym σ. Statystyka testowa:

$$T = \frac{\overline{\overline{x}} - \mu_0}{\sigma}\sqrt{n} \tag{6.7}$$

która przy założeniu prawdziwości hipotezy H_0 ma rozkład normalny $N(0,1)$, w związku z czym obszar krytyczny – w zależności od przyjętej hipotezy alternatywnej (H_1, H_1' albo

H_1'') – ma postać:

$$W_\alpha = \left(-\infty, -t_{1-\frac{\alpha}{2}}\right\rangle \cup \left\langle t_{1-\frac{\alpha}{2}}, \infty\right) \tag{6.8}$$

$$W'_\alpha = (-\infty, -t_{1\alpha}\rangle \tag{6.9}$$

$$W''_\alpha = \langle t_{1-\alpha}, \infty) \tag{6.10}$$

gdzie $t_{1-\frac{\alpha}{2}}$ i $t_{1-\alpha}$ są kwantylami rozkładu normalnego $N(0,1)$ rzędów $1-\frac{\alpha}{2}$ i $1-\alpha$.

Model 2 Jeżeli cecha x ma rozkład normalny $N(\mu,\sigma)$ o nieznanym odchyleniu standardowym σ, to weryfikacja hipotezy H_0 dokonujemy za pomocą statystyki testowej

$$T = \frac{\overline{x} - \mu_0}{s}\sqrt{n} \tag{6.11}$$

która ma rozkład t-Studenta o $n-1$ stopniach swobody (przy założeniu prawdziwości hipotezy H_0). W zależności od przyjętej hipotezy alternatywnej obszar krytyczny przybiera postać:

$$W_\alpha = \left(-\infty, -t_{n-1;1-\frac{\alpha}{2}}\right\rangle \cup \left\langle t_{n-1;1-\frac{\alpha}{2}}, \infty\right) \tag{6.12}$$

$$W'_\alpha = (-\infty, -t_{n-1;1\alpha}\rangle \tag{6.13}$$

$$W''_\alpha = \langle t_{n-1;1-\alpha}, \infty) \tag{6.14}$$

gdzie $t_{n-1;1-\frac{\alpha}{2}}$ oraz $t_{n-1;1-\alpha}$ są kwantylami rozkładu t-Studenta o $n-1$ stopniach swobody.

Model 3 Jeżeli próba pochodzi z dowolnego rozkładu (posiadającego jednakże skończoną wariancję), ale jest wystarczająco duża ($n \geq 100$), wówczas statystyka testowa przyjmuje postać:

$$T = \frac{\overline{x} - \mu_0}{s}\sqrt{n} \tag{6.15}$$

Przy założeniu prawdziwości hipotezy H_0 i dla dostatecznie dużej próby statystyka powyższa ma rozkład (w przybliżeniu) normalny $N(0,1)$, w związku z czym obszar krytyczny w zależności od hipotezy alternatywnej ma postać:

$$W_\alpha = \left(-\infty, -t_{1-\frac{\alpha}{2}}\right\rangle \cup \left\langle t_{1-\frac{\alpha}{2}}, \infty\right) \tag{6.16}$$

$$W'_\alpha = (-\infty, -t_{1\alpha}\rangle \tag{6.17}$$

$$W''_\alpha = \langle t_{1-\alpha}, \infty) \tag{6.18}$$

Przykład 6.3. Załóżmy, że długość „życia opon” samochodowych ma rozkład normalny $N(\mu,\sigma)$. Producent twierdzi, że wartość przeciętna tej charakterystyki jest równa 50 tys. km. Na podstawie 100 losowo wybranych opon otrzymano $\overline{x} = 45$ tys. km oraz $s = 8$ tys. km. Czy na poziomie istotności $\alpha = 0,05$ można uważać, że producent ma rację?

Rozwiązanie. Będziemy korzystali z modelu trzeciego. Mamy

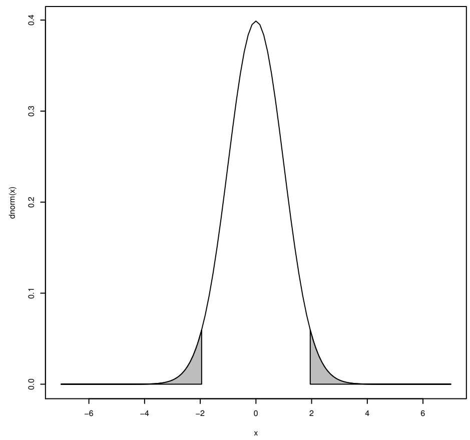
$$H_0 : \mu = 50,$$

$$H_1 : \mu \neq 50.$$

Obliczamy teraz wartość statystyki testowej

$$T = \frac{\overline{\overline{x}} - \mu_0}{s}\sqrt{n} = \frac{45 - 50}{8}\sqrt{100} = -6,25.$$

Teraz t_α. Mamy $\Phi(t_\alpha) = 1 - \frac{0,05}{2} = 0.975$, stąd $t_\alpha = 1,95$. Zatem wartość statystyki testowej wpada do zbioru krytycznego, czyli należy odrzucić hipotezę H_0 na rzecz hipotezy alternatywnej H_1. Innymi słowy producent nie ma racji twierdząc, że przeciętna długość życia opon wynosi 50 tys. km. Na poniższym obrazku na szaro został zaznaczony zbiór krytyczny.



Rysunek 6.1: Interpretacja zbioru krytycznego z Przykładu 6.3

□

Przykład 6.4. W pewnym rejonie morza dokonano 5 niezależnych pomiarów głębokości morza. Otrzymano średnią głębokość morza $\overline{x} = 770$m oraz odchylenie statndardowe $s = 6,2$. Na poziomie istotności $\alpha = 0,02$ zweryfikować hipotezę, że średnia głębokość morza w tym rejnie wynosi $\mu = 775$m, przyjmując że rozkład pomiarów głębokości w tym rejonie morza ma rozkład normalny.

Rozwiązanie. Korzystamy z modelu drugiego. Testujemy hipotezę

$$\begin{aligned} H_0 &: \quad \mu = 775m \\ H_1 &: \quad \mu \neq 775m \end{aligned}$$

Statystyka testowa przyjmuje wartość

$$T = \frac{770 - 775}{6,2}\sqrt{4} = -1,6.$$

Wartość $t_{4;0,02} = 3.747$ odczytujemy z tablic kwantyli rozkładu t-Studenta. Widzimy stąd, że wartość statystyki testowej nie należy do zbioru krytycznego, zetem nie ma podstaw do odrzucenia hipotezy zerowej. □

6.4 Zadania do samodzielnego rozwiązania

6.1. Poddano analizie wydatki na opłaty za telefon TP S.A w 100 gospodarstwach domowych w pewnym mieście. Na podstawie przeprowadzonych obserwacji ustalono średnią miesięczną opłatę za telefon $\overline{x} = 95$zł i odchylenie standardowe $\sigma = 15$zł. Zakładamy że wydatki mają rozkład normalny. Na poziomie ufności $1 - \alpha = 0,98$ wyznaczyć przedział ufności dla wartości przeciętnej miesięcznych opłat za telefon.

Odp $91,55 < \mu < 98,45$.

6.2. W zakładzie „Alfa" zbadano staż pracowników fizycznych. Z populacji tych pracowników wylosowano próbę 169-elementową, z której obliczono $\overline{x} = 7,2$ lat. Rozkład stażu pracowników fizycznych jest rozkładem normalnym z odchyleniem standardowym $\sigma = 3,2$ lat. Przyjmując współczynnik ufności zbudować przedział ufności dla nieznanego średniego stażu pracy w populacji pracowników fizycznych w tym zakładzie.

Odp. $6,634 < \mu < 7,766$.

6.3. Cecha X ma rozkład $N(\mu, \sigma)$, gdzie μ, σ są nieznane. Na podstawie próby 17 elementowej obliczono $\overline{x} = 60, s = 0,5$. Zweryfikować hipotezę $H_0 : \mu = 61,5$, wobec hipotezy alternatywnej $H_1 : \mu \neq 61,5$ na poziomie istotności $\alpha = 0,05$.

Odp. Odrzucamy H_0.

6.4. Z dużej partii słópów betonowych wybrano próbkę losową 64 słupów. rednia wytrzymałość na ściskanie w tej próbie wynosiła $\overline{x} = 245$ kG/cm . Odchylenie standardowe $s = 5$kG/cm . Zweryfikować hipotezę $H_0 : \mu = 240$kG/cm , wobec hipotezy alternatywnej $H_1 : \mu \neq 240$kG/cm , na poziomie istotności $\alpha = 0,01$, przy założeniu, że wytrzymałość na ściskanie jest zmienną losową o rozkładzie normalnym.

Odp. Odrzucamy H_0.

6.5. W pewnym zakładzie wybrano losowo 10 pracowników. Otrzymano średni wiek $\overline{x} = 32$ lata oraz odchylenie standardowe $s = 4$ lata. Zakładając, że wiek pracowników ma rozkład normalny zweryfikować hipotezę, na poziomie istotności $\alpha = 0,05$, że średni wiek pracowników jest istotnie wyższy niż 30 lat. ($Wsk.H_0 : \mu = 30, H_1 : \mu > 30$). Odp. Nie ma podstaw do odrzucenia H_0.

tzn. nie możemy twierdzić, że średni wiek w przedsiębiorstwie jest istonie większy od 30 lat.

Wykład 7

Wybrane zagadnienia procesów stochastycznych

W tym wykładzie omówione jest pojęcie procesu stochastycznego.

7.1 Podstawowe definicje

Definicja 7.1. Niech $T \subset \mathbb{R}$. Rodzinę zmiennych losowych $\{X_t : t \in T\}$ określonych na tej samej przestrzeni probablistycznej nazywamy *procesem stochastycznym.*

W przypadku $T = \mathbb{Z}$ albo $T = \mathbb{Z}$ mówimy o *procesie z czasem dyskretnym* albo też o *szeregu czasowym.* Jeżeli natomiast $T = \langle a, b\rangle$, gdzie $-\infty \leq a < b \leq \infty$, to mówimy o procesie z czasem ciągłym.

Proces stochastyczny $\{X_t : t \in T\}$ możemy rozumieć jako funkcję dwóch zmiennych ω, t. Dla ustalonego t, $X_t(\cdot)$ jest zmienną losową. Natomiast dla ustalonego $\omega \in \Omega$, $X_{(\cdot)} = X_{(\cdot)}(\omega)$ jest rzeczywistą funkcją zmiennej t. Tę funkcję nazywamy *trajektorią procesu* $\{X_t : t \in T\}$.

Dla każdego skończonego zbioru $\{t_1, \ldots, t_n\} \subset T$ określamy układ zmiennych losowych $X_{t_1}, \ldots, X_{t_n}$, które mają mają rozkład zadany przez skończenie wymiarowe dystrybuanty

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n). \tag{7.1}$$

Definicja 7.2. Niech $\{X_t : t \in T\}$ będzie procesem stochastycznym takim, że dla każdego $t \in T$ istnieje wartość oczekiwana $E(X_t)$. Wtedy funkjcę $m_X(t) = E(X_t)$ określoną na zbiorze T nazywamy *wartością oczekiwaną* procesu stochastycznego $\{X_t\}$. Jeżeli ponadto $E(|X_t|^2) < \infty$ dla wszystkich $t \in T$, to wtedy funkcję dwóch zmiennych określoną na zbiorze $T \times T$ wzorem $K(s, t) = E[(X_s - m_X(s))(X_t - m_X(t))]$ nazywamy *funkcją korelacyjną* procesu stochastycznego.

Definicja 7.3. Mówimy, że proces stochastyczny $\{X_t : t \in T\}$ jest *ściśle stacjonarny*, jeżeli dla dowolnego $n \in \mathbb{N}$, dla dowolnych liczb rzeczywistych $x_1, \ldots, x_n$ i dowolnych $t_1, \ldots, t_n$ oraz h takich, że $t_i \in T, t_i + h \in T, i = 1, 2, \ldots, n$ jest

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = F_{t_1+h,\ldots,t_n+h}(x_1, \ldots, x_n). \tag{7.2}$$

Wprost z definicji procesu stochastycznego wynika, że wszystkie zmienne losowe mają takie same rozkłady oraz, że wartość oczekiwana procesu i funcja kowariancyjna procesu nie zmieniają się przy przesunięciach.

Przykład 7.4 (Biały szum)**.** Biały szum jest to proces $\{X_t : t \in \mathbb{Z}\}$ nieskorelowanych zmiennych losowych o zerowej wartości oczekiwanej i stałą skończoną funkcją kowariancyjną. Nazwa procesu pochodzi wywodzi się z jego podobieństwa do własności fizycznych białego światła.

Przykład 7.5. Niech Y_1, Y_2, Y_3 będą zmiennymi losowymi. Utwórzmy proces

$$X_t = Y_1 + Y_2 t + \frac{1}{2} Y_3 t^2, \qquad t \geqslant 0.$$

Proces X_t może być użyty do opisu położenia punktu materialnego poruszającego się ruchem jednostajnie przyspieszonym z przyspieszeniem Y_3, jezeli w chwili początkowej $t = 0$ punkt ma położenie Y_1 i prędkość Y_2.

Rozważmy ciąg zmiennych losowych $\{X_n : n \in \mathbb{N}\}$, które przyjmują tylko wartości całkowitoliczbowe. Niech S będzie zbiorem liczb całkowitych i takich, że $i \in S$ wtedy i tylko wtedy, gdy istnieje $n \in \mathbb{N}$ takie, że $P(X_n = i) > 0$. Zbiór S może być skończony lub przeliczalny (tzn. równoliczny ze zbiorem liczb naturalnych). Będziemy go nazywać *zbiorem stanów* procesu stochastycznego $\{X_n : n \in \mathbb{N}\}$ a jego punkty będziemy nazywali *stanami*. Bez straty ogólności zakładamy, że $S = \{0, 1, \ldots, N\}$ albo też $S = \{1, 2, \ldots\}$.

Definicja 7.6. Ciąg całkowitoliczbowych zmiennych losowych $\{X_n : n \in \mathbb{N}\}$ nazywamy *łańcuchem Markowa* z czasem dyskretnym i zbiorem stanów S, jeżeli

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i) \tag{7.3}$$

dla wszystkich $n = 0, 1, 2, \ldots$ i wszystkich $i, j, i_{n-1}, \ldots, i_0 \in S$ takich, że $P(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) > 0$.

Warunek (7.3) nazywany *warunkiem Markowa* oznacza, że prawdopodobieństwo znalezienia się procesu w czasie $n+1$ w stanie j zależy jedynie od tego w jakim stanie znajdował się proces w czasie n.

Prawdopodobieństwa warunkowe

$$P(X_{n+1} = j | X_n = i) = p_{ij}(n, n+1) \tag{7.4}$$

(jeżeli są określone) nazywamy *prawdopodobieństwami przejścia* ze stanu i w czasie n do stanu j w czasie $n + 1$, czasami też *prawdopodobieństwami przejścia pierwszego rzędu.* Analogicznie prawdopodobieństwa warunowe

$$P(X_{n+m} = j | X_n = i) = p_{ij}(n, n+m) \tag{7.5}$$

dla naturalnego $m \geqslant$ nazywamy *prawdopodobieństwami przejścia* ze stanu i w czasie n do stanu j w czasie $n + m$, lub też *prawdopodobieństwami przejścia m-tego rzędu.* Jeżeli prawdopodobieństwa przejścia $p_{ij}(n, n+m)$ niezależą od czasów n i $n + m$, ale tylko do ich odległości m, to mówimy, że proces Markowa jest *jednorodny.*

Rozważmy jednorodny łańcuch Markowa $\{X_n\}$. Prawdopodobieństwa przejścia pierwszego rzędu $P(X_{n+1} = j | X_n = i)$ są w tym przypadku niezależne od n, będziemy je oznaczać p_{ij}. Ponieważ dla każdego $i \in S$ istnieje $n \in \mathbb{N}$ takie, że $P(X_n = i) > 0$, więc prawdopodobieństow warunkowe $P(X_{n+1} = j | X_n = i) = p_{ij}$ są zdefiniowane dla wszystkich $j \in S$. Wszystkie te prawdopodobieństwa możemy wstawić do kwadratowej macierzy $\mathbf{P} = [p_{ij}]_{:i,j \in S}$. Dla każdego $n \in \mathbb{N}$

$$p_{ij} > 0, \quad i, j \in S; \qquad \sum_{j \in S} p_{ij} = 1, \quad i \in S. \tag{7.6}$$

Kwadratową macierz o powyższych własnościach nazywamy *macierzą stochastyczną.*

Oznaczmy dalej

$$p_i = P(X_0 = i), \qquad i \in S. \tag{7.7}$$

Oczywiście

$$p_i \geqslant 0, i \in S, \sum_{i \in S} p_i = 1. \tag{7.8}$$

Rozkład prawdopodobieństwa $\mathbf{p} = \{p_i : i \in S\}$ nazywamy *prawdopodobieństwami początkowymi.*

Twierdzenie 7.7. *Niech $\{X_n : n \in \mathbb{N}\}$ będzie procesem stochastycznym o zbiorze stanów $S = \{0, 1, \ldots\}$. Niech $\mathbf{p} = \{p_i : i \in S\}$ jest wektorem spełniającym (7.8) oraz $\mathbf{P} = [p_{ij}]_{i,j \in S}$ macierzą spełniającą (7.6). Wtedy proces X_t jest jednorodnym łańcuchem Markowa z rozkładem początkowym $\mathbf{p}$ i macierzą przejścia $\mathbf{P}$, wtedy i tylko wtedy, gdy wszystkie skończeniewymiarowe rozkłady tego procesu są postaci*

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k) = p_{i_0} p_{i_0 i_1} \ldots p_{i_{k-1} i_k} \tag{7.9}$$

dla wszystkich $i_0, i_1, \ldots, i_k \in S$ i wszystkich $k \in \mathbb{N}$.

Rozważmy teraz jednorodny łańcuch Markowa o macierzy prawdopodobieństw przejścia P. Połóżmy $p_{ij}^{(0)} = \delta_{ij}$, gdzie δ_{ij} jest symbolem Kroneckera

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Dalej $p_{ij}^{(1)} = p_{ij}$ i dla wszystkich $n \geqslant 1$ definiujemy indukcyjnie

$$p_{ij}^{(n+1)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}. \tag{7.10}$$

Można dowieść, że powyższe szeregi są zbieżne dla każdego $n \geqslant 1$, oraz że macierze $\mathbf{P}^{(n)}$ z elementami $p_{ij}^{(n)}$ są macierzami stochastycznymi. Z warunku (7.10) wynika, że

$$\mathbf{P}^{(2)} = \mathbf{P} \cdot \mathbf{P} = \mathbf{P}^2, \qquad \mathbf{P}^{(n)} = \mathbf{P}^{(n-1)} \cdot \mathbf{P} = \mathbf{P} \cdot \mathbf{P}^{(n-1)} = \mathbf{P} \cdot \mathbf{P}^n.$$

Przykład 7.8 (Zadanie o ruinie gracza)**.** Gracz A oraz jego przeciwnik B grają w pewną powtarzającą się grę, która może skończyć się tylko wygraną jednego z nich. W grze jest kapitał a jednostek, przy czym na początku gracz A ma z jednostek kapitału, natomiast jego przeciwnik $a - z$ jednostek. Jeżeli wygra gracz A zyskuje od swego przeciwnika 1 jednostkę kapitału, jeżlie przegra 1 jednostkę traci. Gracze grają tak długo aż jeden z nich straci cały swój kapitał. Zakładamy, że prawdopodobieństwa wygrania graczy A i B są p i $q = 1 - p$ odpowiednio, oraz że wszystkie partie gry są niezależne. Jeżeli X_n oznacza kapitał, który po n-tej partii posiada gracz A, to wtedy $\{X_n\}$ jest jednorodnym łańcuchem Markowa ze stanami $S = 0, 1, \ldots, a\}$ i wektorem prawdopodobieństw początkowych $p_z =$

$1, p_j = 0, j \neq z$ oraz prawdopodobieństwami przejścia $p_{00} = p_{aa} = 1, p_{i,i+1} = p, p_{i,i-1} = q, 1 \leqslant i \leqslant a-1$. Macierz prawdopodobieństw przejścia ma wtedy postać

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ q & 0 & p & 0 & \dots & 0 & 0 & 0 \\ 0 & q & 0 & p & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & q & 0 & p \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}$$

Bibliografia

[1] L.Gajek, M. Kałuszka, *Wnioskowanie statystyczne dla studnetów,* WNT, Warszawa 1998.

[2] J. Jóźwiak, J. Podgórski, *Statystyka od podstaw,* PWE, Warszawa 2006.

[3] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, M. Wasilewski, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach cz. I i II*, PWN, Warszawa 2004.

[4] J. Ombach, *Rachunek prawdopodobieństwa wspomagany komputerowo - Maple,* Wydawnictwo UJ, Kraków 2000.