Rachunek prawdopodobieństwa i statystyka

Anna Serwatka

Pod redakcją:
Anna Serwatka

Kraków 2025

Tytuł:Rachunek prawdopodobieństwa i statystyka
Autorzy: Anna Serwatka (Wydział Matematyki Stosowanej, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie)
Recenzenci:Prof. dr hab. Marcin Salamaga, Uniwersytet Ekonomiczny w Krakowie; dr hab. Andrzej Nowik, prof. Uniwersytetu Gdańskiego
Data publikacji:05.08.2025
ISBN: 978-83-973330-4-8

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Al. Mickiewicza 30
30-059 Kraków

Spis treści

1 Rachunek prawdopodobieństwa - rys historyczny **3**
1.1. Wprowadzenie . . . 3

2 Kombinatoryka **5**
2.1. Reguła mnożenia i reguła dodawania . . . 5
2.2. Wariacje . . . 9
2.3. Permutacje . . . 12
2.4. Kombinacje . . . 15

3 Zdarzenia losowe i prawdopodobieństwo **19**
3.1. Doświadczenie losowe i działania na zdarzeniach . . . 19
3.2. Prawdopodobieństwo klasyczne . . . 22
3.3. Doświadczenie losowe wieloetapowe . . . 24
3.4. Prawdopodobieństwo warunkowe, zupełne i niezależność zdarzeń . . . 26
3.5. Schemat Bernoulliego . . . 29

4 Jednowymiarowe zmienne losowe **31**
4.1. Zmienna losowa i jej dystrybuanta . . . 31
4.2. Wartość oczekiwana . . . 35
4.3. Momenty . . . 37
4.4. Nierówność Czebyszewa . . . 39

5 Wielowymiarowe zmienne losowe **41**
5.1. Dwuwymiarowa zmienna losowa . . . 41
5.2. Rozkład brzegowy i warunkowy . . . 44
5.3. Niezależność zmiennych losowych . . . 46
5.4. Parametry rozkładu wektorów losowych . . . 47
5.5. Dwuwymiarowy rozkład normalny . . . 48

6 Elementy statystyki opisowej **51**
6.1. Średnie, mediana i moda . . . 51
6.2. Miary zmienności: odchylenie standardowe, wariancja . . . 54

7 Przedziały ufności **57**
7.1. Pojęcia wstępne: populacja, próba . . . 57
7.2. Estymacja punktowa i przedziałowa . . . 58
7.3. Przedział ufności dla średniej . . . 59
7.4. Przedział ufności dla wariancji . . . 63
7.5. Przedział ufności dla wskaźnika struktury . . . 65
7.6. Niezbędna liczba pomiarów dla próby . . . 66

8 Parametryczne testy istotności **69**
8.1. Test dla wartości średniej populacji . . . 69
8.2. Test dla dwóch średnich . . . 72
8.3. Test dla wskaźnika struktury . . . 75
8.4. Test dla dwóch wskaźników struktury . . . 77
8.5. Test dla wariancji . . . 78

9 Związki korelacyjne między zmiennymi różnych typów **81**
9.1. Korelacja i kowariancja . . . 81
9.2. Prosta regresja liniowa . . . 83
9.3. Wieloraka regresja liniowa . . . 86
9.4. Regresja nieliniowa . . . 87

10 Przykłady paradoksów statystycznych i probabilistycznych **89**
10.1. Paradoks Simpsona . . . 89
10.2. Paradoks Monty'ego Halla . . . 91
10.3. Paradoks hazardzisty . . . 93
10.4. Paradoks urodzinowy . . . 94
10.5. Paradoks więźnia . . . 95
10.6. Paradoks Bertrand'a . . . 97
10.7. Błędy i pułapki w interpretacji danych statystycznych . . . 98

Rozdział 1

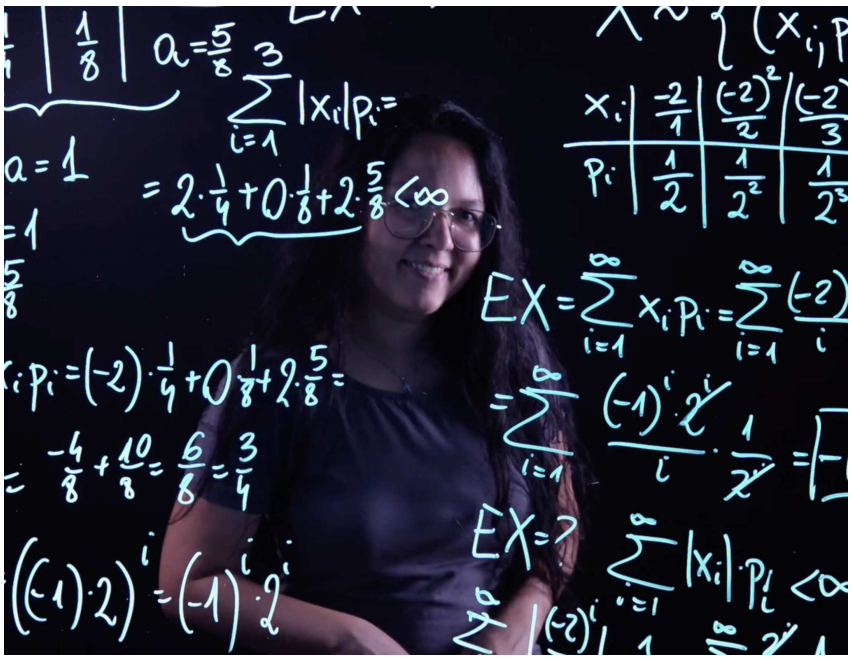
Rachunek prawdopodobieństwa - rys historyczny

1.1. Wprowadzenie

Autorzy/Autorki: Anna Serwatka

Rachunek prawdopodobieństwa jest dziedziną matematyki, która zajmuje się badaniem jak często występują zdarzenia losowe. Jego historia sięga starożytności, natomiast w XVII wieku nastąpił ogromny skok w tej dziedzinie. Pierwsze wzmianki o rachunku prawdopodobieństwa znajdziemy już w starożytnym Rzymie i Grecji, gdzie badano różne aspekty gier losowych. Możemy zatem powiedzieć, że rachunek prawdopodobieństwa narodził się z ... hazardu. Korespondencja wybitnych matematyków Blaise Pascala oraz Pierre de Fermata, zrewolucjonizowała pojęcie losowości. Pascal i Fermat rozważali problemy związane z grami i podziałem stawek, co doprowadziło ich do rozważań, które dziś są fundamentalne w rachunku prawdopodobieństwa.

W XVIII wieku ta dziedzina została rozwinięta przez matematyków takich jak Jakob Bernoulli oraz Abrahama de Moivre'a. Bernoulli w swoim dziele Ars Conjectandi wprowadził słabe prawo wielkich liczb Bernoulliego, natomiast Abraham de Moivre (którego najbardziej kojarzymy ze wzoru na potęgę liczby zespolonej w postaci biegunowej) był autorem klasycznej pozycji The Doctrine of Chances. Na początku XIX wieku kolejny matematyk zajmujący się rachunkiem prawdopodobieństwa Pierre-Simon Laplace opracował bardziej zaawansowane narzędzia matematyczne, takie jak analiza statystyczna i teoria błędów pomiarowych, które umożliwiały zastosowanie rachunku prawdopodobieństwa w naukach przyrodniczych i ekonomii. W XX wieku dziedzina ta rozwijała się intensywnie dzięki wkładowi Andrieja Kołmogorowa, który w 1933 roku sformalizował podstawy rachunku prawdopodobieństwa w ramach teorii miary.



https://vimeo.com/1029563256/1fb492d497

Wstęp

Rozdział 2

Kombinatoryka

2.1. Reguła mnożenia i reguła dodawania

Autorzy/Autorki: Anna Serwatka

Kombinatoryka jest dziedziną matematyki, zajmującą się badaniem różnych sposobów wyboru, rozmieszczenia i układania elementów w zbiorach. Jest to dziedzina, która daje odpowiedź na pytania związane z liczbą elementów, które można wybrać z danego zbioru, gdyż zajmuje się badaniem kombinacji, permutacji i wariacji, uwzględniając różne zasady dotyczące kolejności, powtarzalności i innych ograniczeń.

W praktyce oznacza to, że kombinatoryka dostarcza narzędzi do liczenia możliwych konfiguracji w różnych sytuacjach, co jest kluczowe dla zrozumienia bardziej zaawansowanych zagadnień matematycznych.

Oto cztery powody, dla których warto poznać kombinatorykę zanim przejdzie się do liczenia zadań z rachunku prawdopodobieństwa:

1. **Zliczanie możliwych zdarzeń** - rachunek prawdopodobieństwa opiera się na liczeniu prawdopodobieństwa zdarzeń, które często wymagają wcześniejszego zrozumienia, ile różnych konfiguracji jest możliwych w danej sytuacji. Kombinatoryka dostarcza podstawowych technik, które umożliwiają liczenie takich konfiguracji. Bez zrozumienia kombinatoryki, trudno byłoby zrozumieć, na czym opiera się wiele problemów probabilistycznych.
2. **Rozumienie struktury problemu** - kombinatoryka pomaga w rozkładaniu złożonych problemów na prostsze części, co jest nieocenione przy rozwiązywaniu zadań z rachunku prawdopodobieństwa.
3. **Budowanie intuicji matematycznej** - kombinatoryka rozwija intuicję matematyczną i umiejętność logicznego myślenia. Pozwala zrozumieć, jak różne elementy mogą się ze sobą łączyć i jakie są konsekwencje różnych wyborów.
4. **Zastosowanie w realnych problemach** - rachunek prawdopodobieństwa często stosuje się do analizy rzeczywistych zjawisk, takich jak gry losowe, statystyka czy procesy decyzyjne. Kombinatoryka dostarcza metod, które pomagają w modelowaniu tych zjawisk, na przykład znając kombinacje możemy wyliczyć na ile sposobów możemy wylosować 6 z 49 liczb, a dzięki temu policzymy prawdopodobieństwo trafienia "szóstki" w Dużym Lotku.

Podsumowując, znajomość kombinatoryki jest fundamentem dla zrozumienia rachunku prawdopodobieństwa. Umożliwia poprawne liczenie przestrzeni zdarzeń, co jest kluczowe dla obliczania prawdopodobieństw i rozwiązywania problemów probabilistycznych.

DEFINICJA

Definicja 1: Reguła mnożenia

Reguła mnożenia to zasada kombinatoryczna, która mówi, że jeśli pewne doświadczenie składa się z k etapów, gdzie pierwszy etap można wykonać na n_1 sposobów, drugi etap na n_2 sposobów,..., k-ty etap na n_k sposobów, to liczba możliwych sposobów wykonania całego doświadczenia wynosi $n_1 n_2 \dots n_k$.

PRZYKŁAD

Przykład 1:

Aby obliczyć ile jest wszystkich liczb trzycyfrowych musimy podjąć trzy decyzje: wybrać cyfrę setek, cyfrę dziesiątek i cyfrę jedności. Doświadczenie składa się z trzech etapów. Zauważmy, że cyfrę setek możemy wybrać na 9 sposobów, będzie to cyfra ze zbioru $\{1,2,3,4,5,6,7,8,9\}$, ponieważ 0 nie może być cyfrą setek. Cyfrę dziesiątek możemy wybrać na 10 sposobów (zero może być cyfrą dziesiątek) i analogicznie na 10 sposobów możemy wybrać cyfrę jedności. Korzystając z reguły mnożenia otrzymujemy wynik $9 \cdot 10 \cdot 10 = 900$, czyli mamy 900 liczb trzycyfrowych.

ZADANIE

Zadanie 1:

Treść zadania:

Ile jest czterocyfrowych kodów PIN [1], w których cyfry

1. mogą się powtarzać
2. nie mogą się powtarzać?

Rozwiązanie:

1. Zauważmy, że każdą z czterech cyfr możemy wybrać na 10 sposobów, stąd możliwych kodów PIN, w których cyfry mogą się powtarzać jest 10^4.
2. Pierwszą cyfrę w kodzie PIN wybieramy na 10 sposobów (możemy wybrać każdą z cyfr ze zbioru {0,1,2,3,4,5,6,7,8,9}), drugą cyfrę możemy już wybrać na 9 sposobów (ponieważ nie możemy powtórzyć cyfry, która stoi na pierwszym miejscu). Skoro zabraliśmy już 2 cyfry to na trzecim miejscu pozostaje nam 8 możliwości, a na czwartym 7 opcji wyboru (wszak 3 cyfry z 10 już zarezerwowaliśmy dla poprzednich miejsc). Stąd wynik to $10 \cdot 9 \cdot 8 \cdot 7 = 5040$.

ZADANIE

Zadanie 2:

Treść zadania:

Numer karty płatniczej składa się z 16 cyfr. Pierwszą cyfrą jest 5, drugą jest jedna z cyfr: 1,2,3,4,5. Zakładamy, że pozostałe cyfry mogą być dowolne. Ile jest numerów kart płatniczych?

Rozwiązanie:

Pierwsza cyfra jest ustalona i jest tylko jeden możliwy wybór cyfry 5, drugą cyfrę możemy wybrać na 5 sposobów. Te dwie cyfry mamy już wybrane i potrzebujemy wybrać jeszcze 14 cyfr, a każdą z nich możemy wybrać na 10 sposobów. Otrzymujemy, że przy takich założeniach jest $1 \cdot 5 \cdot 10^{14}$ numerów kart płatniczych.

Podsumowując, reguła mnożenia mówi, że jeśli pewne zadanie można rozłożyć na kilka etapów i każdy z tych etapów można wykonać na określoną liczbę sposobów, to łączna liczba wszystkich możliwych sposobów wykonania całego zadania jest równa iloczynowi liczby sposobów wykonania każdego z tych etapów [2].

DEFINICJA

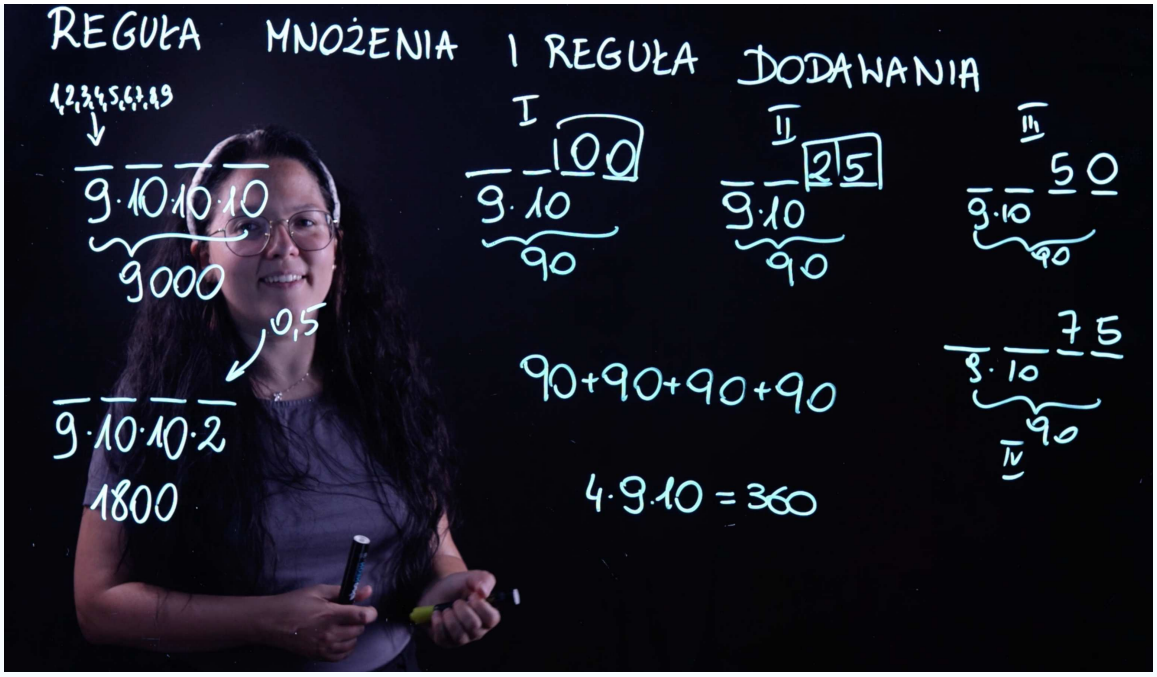
Definicja 2: Reguła dodawania

Jeśli pewne doświadczenie może być wykonane na n_1 sposobów albo na n_2 sposobów, ... albo n_k sposobów, przy czym te sposoby wzajemnie się wykluczają (nie mogą wystąpić jednocześnie), to łączna liczba możliwych sposobów wykonania tego zadania wynosi $n_1 + n_2 \ldots n_k$.

Reguła dodawania mówi zatem, że jeśli pewne doświadczenie można wykonać na kilka różnych, wzajemnie wykluczających się sposobów, to łączna liczba możliwych sposobów wykonania tego doświadczenia jest równa sumie liczby sposobów wykonania każdego z tych sposobów. Reguła mnożenia oraz reguła dodawania są często stosowane intuicyjnie, a ich działanie dobrze obrazuje następujący przykład.

PRZYKŁAD

Przykład 2: Reguła mnożenia i reguła dodawania na kreskach



https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2370/reader#openaghmod:2370:openaghhfivep:1

Reguła mnożenia i reguła dodawania

ZADANIE

Zadanie 3:

Treść zadania:

Restauracja "AGH Bistro" serwuje 12 dań kuchni polskiej, 11 dań kuchni francuskiej i 5 dań kuchni węgierskiej. Na ile sposobów można w tej restauracji wybrać:

1. jedno danie, które nie należy do kuchni polskiej,
2. dwa różne dania, z których jedno należy do kuchni polskiej a drugie nie należy do kuchni polskiej?

Rozwiązanie:

1. Jeżeli chcemy wybrać danie, które nie należy do kuchni polskiej, mamy do wyboru kuchnię francuska (11 dań) oraz kuchnię węgierską (5 dań). W sumie możemy wybrać 11+5=16 dań z kuchni niepolskiej.

2. Wybieramy danie z kuchni polskiej na 12 sposobów i danie z kuchni niepolskiej na 16 sposobów (przy wyborze dania z kuchni niepolskiej stosujemy regułę dodawania z poprzedniego przykładu). Ponieważ to doświadczenie składa się z dwóch etapów (pierwszy to wybór kuchni polskiej, a drugi kuchni niepolskiej) musimy zastosować teraz regułę mnożenia i otrzymujemy $12 \cdot 16 = 192$ takie zestawy dań.

2.2. Wariacje

Autorzy/Autorki: Anna Serwatka

W dziedzinie kombinatoryki, zajmującej się zliczaniem obiektów i ich układów, fundamentalne znaczenie ma pojęcie **wariacji**. Wariacje stanowią narzędzie do określania liczby sposobów wyboru i uporządkowania elementów z danego zbioru. Kolejność wybranych elementów ma znaczenie, a w zależności od tego, czy elementy mogą być powtarzane w tworzonych układach, rozróżniamy dwa główne typy wariacji: **wariacje z powtórzeniami** oraz **wariacje bez powtórzeń** [1]. Zrozumienie tych rozróżnień i odpowiadających im zasad zliczania jest niezbędne do rozwiązywania szerokiej gamy problemów probabilistycznych, algorytmicznych oraz w informatyce, gdzie analiza złożoności obliczeniowej często opiera się na liczbie możliwych konfiguracji. Wprowadźmy definicję wariacji z powtórzeniami i bez powtórzeń oraz przedstawimy twierdzenia dotyczące zliczania liczby wariacji [3].

DEFINICJA

Definicja 3: Wariacja bez powtórzeń

Wariacja bez powtórzeń to ciąg elementów wybranych z danego niepustego zbioru, w którym każdy element może wystąpić tylko raz.

TWIERDZENIE

Twierdzenie 1: Liczba wariacji bez powtórzeń

ZAŁOŻENIE:

Niech dany będzie n-elementowy zbiór elementów. Załóżmy, że wybieramy k różnych elementów z tego zbioru, gdzie $k \leq n$ elementów i ustalamy je w ciąg.

TEZA:

Liczba możliwych k-wyrazowych wariacji bez powtórzeń elementów zbioru n-elementowego wynosi

$$V_{n,k} = \frac{n!}{(n-k)!}, \tag{2.1}$$

gdzie $n!$ to iloczyn wszystkich liczb naturalnych od 1 do n (silnia liczby n).

DOWÓD:

Zauważmy, że mamy tutaj do podjęcia k decyzji (są to kolejne elementy ciągu jaki tworzymy). Pierwszą decyzję możemy podjąć na n sposobów, drugą na $n - 1$, ponieważ nie możemy powtórzyć elementów w ciągu, trzecią na $n - 2$ i k-tą na $n - (k - 1)$. Z reguły mnożenia otrzymujemy liczbę wszystkich wariacji bez powtórzeń $n \cdot (n-1) \ldots (n-(k-1)) = \frac{n!}{(n-k)!}$

DEFINICJA

Definicja 4: Wariacja z powtórzeniami

Wariacja z powtórzeniami to ciąg elementów wybranych z danego niepustego zbioru, w którym dopuszczamy powtarzanie się elementów.

TWIERDZENIE

Twierdzenie 2: Liczba wariacji z powtórzeniami

ZAŁOŻENIE:

Niech dany będzie n-elementowy zbiór elementów. Załóżmy, że wybieramy k elementów z tego zbioru (elementy mogą się powtarzać), gdzie $k \in \mathbb{N}_+$ i ustalamy je w ciąg.

TEZA:

Liczba możliwych k-wyrazowych wariacji z powtórzeniami elementów zbioru n-elementowego wynosi

$$W_n^k = n^k, \tag{2.2}$$

gdzie: n to liczba elementów w zbiorze, k to liczba wybieranych elementów.

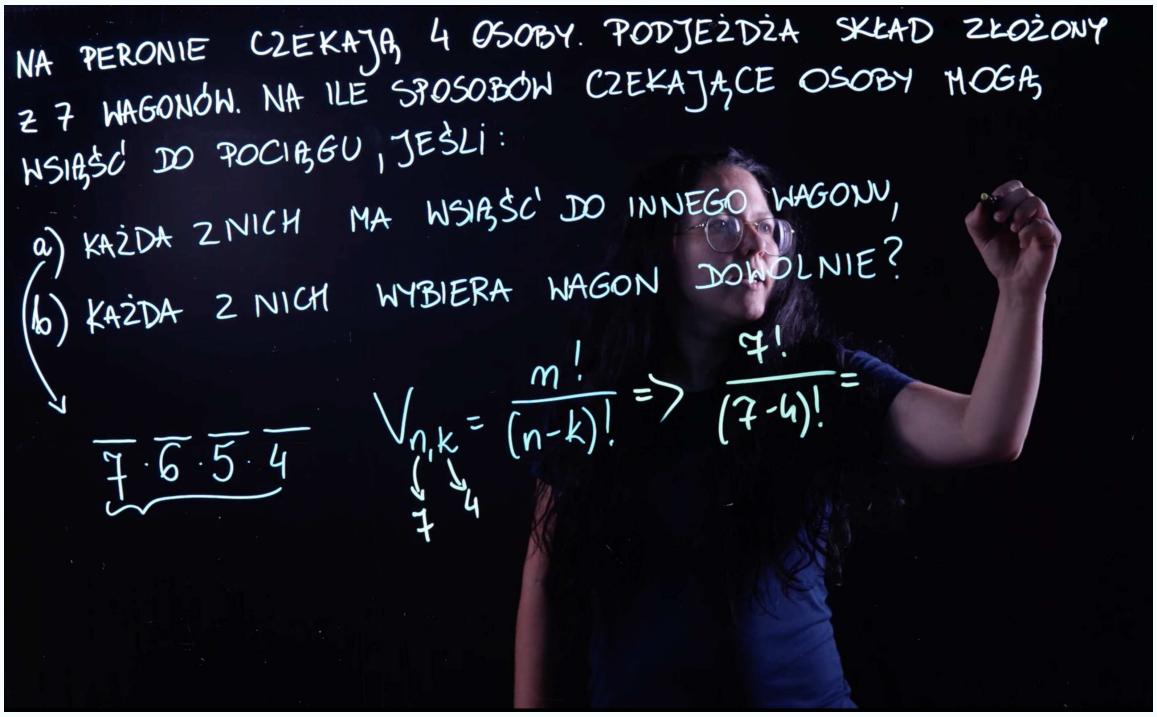
DOWÓD:

Zauważmy, że mamy tutaj do podjęcia k decyzji (są to kolejne elementy ciągu jaki tworzymy). Pierwszą decyzję możemy podjąć na n sposobów, podobnie drugą także na n sposobów (ponieważ możemy powtórzyć elementy w ciągu), itd. i k-tą również na n sposobów. Z reguły mnożenia otrzymujemy liczbę wszystkich wariacji z powtórzeniami n^k

W kombinatoryce wariacje są stosowane wtedy, gdy istotna jest kolejność wybieranych elementów (są to ciągi). Innymi słowy, wariacje odnoszą się do sytuacji, gdy wybieramy pewną liczbę elementów ze zbioru i ważne jest, w jakiej kolejności te elementy są ułożone. Aby poprawnie zastosować wzory na liczbę wariacji należy po pierwsze upewnić się, że w doświadczeniu liczy się kolejność, a po drugie ustalić czy jest to wariacja z powtórzeniami czy bez powtórzeń. Zastosowanie wariacji dobrze zobrazuje następujący przykład i zadania.

PRZYKŁAD

Przykład 3: Zadanie z wagonami



https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2373/reader#openaghmod:2373:openaghhfivep:1

Wariacje

ZADANIE

Zadanie 4:

Treść zadania:

Ile pięcioliterowych kodów można utworzyć z liter A,B,C?

Rozwiązanie:

Po pierwsze zauważmy, że w tym zadaniu liczy się kolejność. Ze zbioru, w którym są trzy litery A,B,C wybieramy 5 liter i ustawiamy je w ciąg. Jest to wariancja z powtórzeniami, ponieważ liczy się kolejność, a w treści zadania nie mamy informacji, że litery nie mogą się powtarzać (wręcz jest to niemożliwe, aby wybrać 5 liter 3 bez powtórzeń). Możemy rozwiązać to zadanie na dwa sposoby.

Pierwszy sposób to zastosowanie wzoru na liczbę wariacji z powtórzeniami, mamy $n = 3$ i $k = 5$, stąd możemy utworzyć $W_3^5 = 3^5 =$ pięcioliterowych kodów.

To zadanie możemy również rozwiązać nie stosując wzoru na liczbę wariacji z powtórzeniami, a zastosować regułę mnożenia. Pierwszą literę wybieramy na 3 sposoby, drugą także na 3 sposoby (litery mogą się powtarzać), analogicznie wybieramy trzecią, czwartą i piątą literę i otrzymujemy $3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 243$ możliwe kody.

ZADANIE

Zadanie 5:

Treść zadania:

Ile jest liczb trzycyfrowych, zbudowanych z cyfr należących do zbioru $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, jeśli cyfry w liczbie nie mogą się powtarzać?

Rozwiązanie:

Po pierwsze zauważmy, że w tym zadaniu liczy się kolejność, ponieważ tworzymy liczbę (patrzymy na liczbę trzycyfrową jak na ciąg trzech cyfr). Ze zbioru, w którym jest 9 cyfr (zauważmy, że w zbiorze nie ma cyfry 0) wybieramy 3 cyfry i ustawiamy je w ciąg (pierwsza cyfra to cyfra setek, druga to cyfra dziesiątek i trzecia to cyfra jedności. Jest to wariancja bez powtórzeń, ponieważ cyfry w liczbie nie mogą się powtarzać. To zadanie także możemy rozwiązać na dwa sposoby.

Pierwszy sposób to zastosowanie wzoru na liczbę wariacji bez powtórzeń, mamy $n = 9$ i $k = 3$, stąd możemy utworzyć $V_{9,3} = \frac{9!}{(9-3)!} = \frac{9!}{(6)!} = 7 \cdot 8 \cdot 9 = 504$ liczby.

Rozwiążmy teraz to zadanie stosując regułę mnożenia. Pierwszą cyfrę wybieramy na 9 sposobów, drugą na 8 sposobów (nie możemy powtórzyć cyfry, która już stoi na pierwszym miejscu) i trzecią na 7 sposobów. Czyli mamy $7 \cdot 8 \cdot 9 = 504$ liczby.

2.3. Permutacje

Autorzy/Autorki: Anna Serwatka

Permutacje pozwalają nam odpowiedzieć na pytanie, ile jest sposobów na uporządkowanie elementów pewnego zbioru. Niezależnie od tego, czy układamy książki na półce, losujemy kolejność mówców na konferencji, czy analizujemy możliwe stany algorytmu sortującego, permutacje dostarczają nam narzędzi do precyzyjnego określania tych unikalnych sekwencji. Zrozumienie, czym jest permutacja i jakie są jej rodzaje – zwłaszcza rozróżnienie na permutacje z powtórzeniami i bez powtórzeń – jest fundamentalne dla dalszych zagadnień kombinatorycznych i probabilistycznych, stanowiąc podstawę do rozwiązywania wielu problemów zarówno teoretycznych, jak i praktycznych.

DEFINICJA

Definicja 5: Permutacja bez powtórzeń

Permutacją bez powtórzeń nazywamy n wyrazową wariację bez powtórzeń ze zbioru n-elementowego

TWIERDZENIE

Twierdzenie 3: Liczba permutacji bez powtórzeń

ZAŁOŻENIE:

Rozważamy n-elementowy zbiór elementów, gdzie $n \in \mathbb{N}_+$.

TEZA:

Liczba możliwych permutacji bez powtórzeń elementów zbioru n-elementowego wynosi [4]

$$P_n = n!. \tag{2.3}$$

DOWÓD:

Zauważmy, że permutacja bez powtórzeń to z definicji n wyrazowa wariacja bez powtórzeń ze zbioru n-elementowego. Możemy zastosować wobec tego wzór na liczbę wariacji bez powtórzeń, stąd otrzymujemy: $V_{n,n} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$

DEFINICJA

Definicja 6: Permutacja z powtórzeniami

Permutacją z powtórzeniami nazywamy uporządkowany ciąg k-elementów wybranych ze zbioru n- elementowego, gdzie kolejne elementy zbioru powtarzają się odpowiednio $k_1, k_2, \ldots, k_n$ razy, a $k = k_1 + k_2 + \ldots + k_n$.

TWIERDZENIE

Twierdzenie 4: Liczba permutacji z powtórzeniami

ZAŁOŻENIE:

Rozważmy n- elementowy zbiór z powtórzeniami odpowiednio $k_1, k_2, \ldots, k_n$ krotnymi kolejnych elementów (gdzie $k = k_1 + k_2 + \ldots + k_n$).

TEZA:

Liczba możliwych k-wyrazowych permutacji z powtórzeniami elementów zbioru wynosi

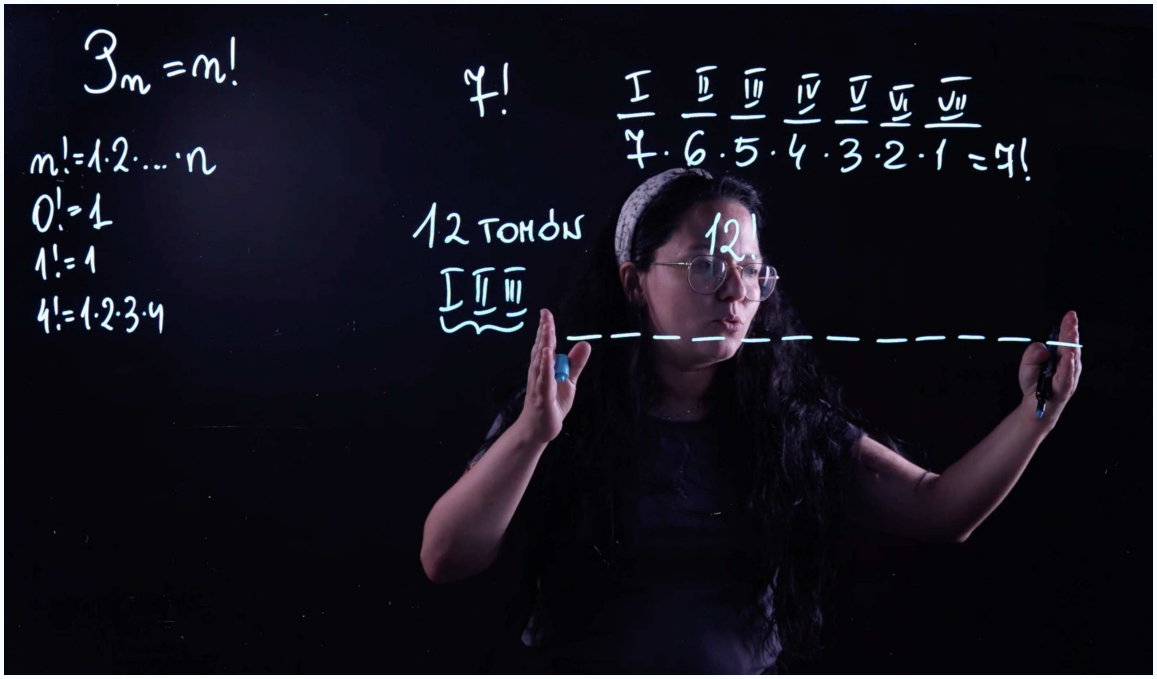
$$P_k^{k_1,k_2,\ldots,k_n} = \frac{k!}{k_1! \cdot k_2! \ldots k_n!}$$

.

Istotne jest, że w permutacjach liczy się kolejność, a ponadto należy zwrócić uwagę, czy mamy do czynienia z permutacjami z powtórzeniami, czy bez. Zadania można rozwiązywać używając gotowych wzorów, lub wykorzystać regułę mnożenia.

PRZYKŁAD

Przykład 4: Na ile sposobów można ułożyć tomy encyklopedii na półce?



https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2374/reader#openaghmod:2374:openaghhfivep:1

Permutacje

ZADANIE

Zadanie 6:

Treść zadania:

W biegu finałowym uczestniczy ośmiu sprinterów. Na ile sposobów mogą oni zająć kolejne miejsca, jeśli wszyscy ukończą bieg i nie będzie remisów?

Rozwiązanie:

Jest to permutacja 8-elementowa, stąd mamy $8! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 = 40320$ możliwości ukończenia biegu.

ZADANIE

Zadanie 7:

Treść zadania:

Ile wyrazów, mających sens lub nie, można utworzyć przestawiając słowa w wyrazie MATEMATYKA?

Rozwiązanie:

Rozważmy zbiór liter użytych w słowie MATEMATYKA, czyli $\{M, A, T, E, Y, K\}$ oraz ich krotności: $M - 2, A - 3, T - 2, E - 1, Y - 1, K - 1$. Zauważmy, że tworzymy permutację 10 elementową z powtórzeniami, z zadanymi wcześniej krotnościami dla liter. Korzystając ze wzoru otrzymujemy $\frac{10!}{2!3!2!}$ możliwych słów. [1]

2.4. Kombinacje

Autorzy/Autorki: Anna Serwatka

Po szczegółowym omówieniu permutacji i wariacji, gdzie kolejność wyboru elementów odgrywała kluczową rolę, nadszedł czas na zagadnienie kombinacji. W wielu problemach zliczania, istotne jest jedynie, jakie elementy zostały wybrane, a nie w jakiej kolejności się pojawiły. Na przykład, gdy wybieramy trzy osoby z grupy do zespołu, nie ma znaczenia, czy Janek został wybrany pierwszy, czy ostatni; liczy się jedynie to, że Janek, obok pozostałych dwóch osób, znalazł się w składzie. Kombinacje pozwalają nam odpowiedzieć na pytanie, ile jest sposobów na wybranie podzbioru elementów z większego zbioru, bez uwzględniania ich wewnętrznego uporządkowania.

DEFINICJA

Definicja 7: Kombinacja bez powtórzeń

Niech dany będzie n-elementowy zbiór elementów. Kombinacją k-elementową bez powtórzeń nazywamy każdy podzbiór tego zbioru , wybrany w dowolnej kolejności bez zwracania. Zakładamy, że $n, k \in \mathbb{N}_+$ i $k \leq n$.

TWIERDZENIE

Twierdzenie 5: Liczba kombinacji bez powtórzeń

ZAŁOŻENIE:

Niech dany będzie n-elementowy zbiór elementów i $k \in \mathbb{N}_+$ i $k \leq n$.

TEZA:

Liczba możliwych k-wyrazowych kombinacji bez powtórzeń elementów zbioru n-elementowego wynosi

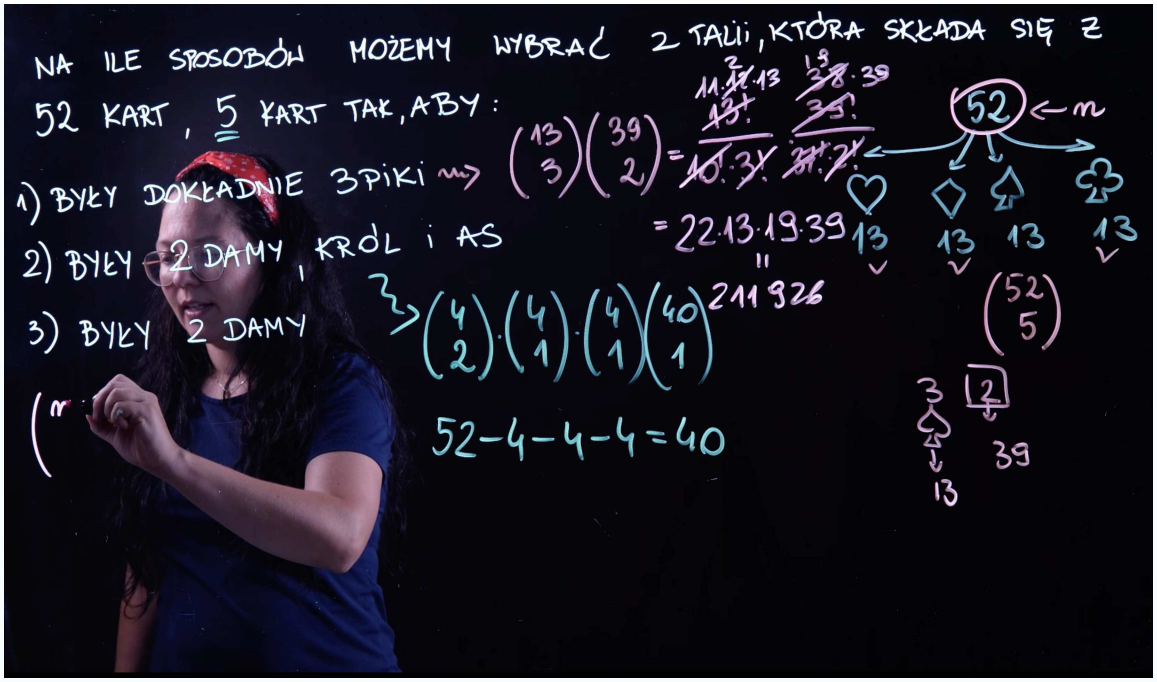
$$C_n^k = \binom{n}{k} = \frac{n!}{(n-k)!k!}. \tag{2.4}$$

DOWÓD:

Zauważmy, że mamy tutaj do podjęcia k decyzji (są to kolejne elementy ciągu jaki tworzymy). Pierwszą decyzję możemy podjąć na n sposobów, drugą na $n - 1$, ponieważ nie możemy powtórzyć elementów w ciągu, trzecią na $n - 2$ i k -tą na $n - (k - 1)$. Z reguły mnożenia otrzymujemy liczbę wszystkich wariacji bez powtórzeń $n \cdot (n-1) \ldots (n-(k-1)) = \frac{n!}{(n-k)!}$

PRZYKŁAD

Przykład 5: Losowanie z talii kart



https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2375/reader#openaghmod:2375:openaghhfivep:1

Losowanie z talii kart

ZADANIE

Zadanie 8:

Treść zadania:

Z talii 52 kart losujemy jednocześnie 4 karty. Na ile sposobów możemy to zrobić? [1]

Rozwiązanie:

Zauważmy, że nie liczy się tutaj kolejność wylosowania kart. Wybór 4 kart spośród 52 można obliczyć za pomocą kombinacji, ponieważ kolejność wyboru nie ma znaczenia. Kombinacje obliczamy za pomocą wzoru $C_n^k = \binom{n}{k} = \frac{n!}{(n-k)!k!}$, gdzie n to liczba wszystkich kart, czyli $n = 52$, a k to liczba kart wylosowanych, stąd $k = 4$. Podstawiamy te wartości do zbioru i otrzymujemy: $\binom{52}{4} = \frac{52!}{(52-4)!4!} = \frac{52!}{48!4!} = \frac{48!\cdot 49\cdot 50\cdot 51\cdot 52}{48!4!} = \frac{49\cdot 50\cdot 51\cdot 52}{24} = 270725$

Rozdział 3

Zdarzenia losowe i prawdopodobieństwo

3.1. Doświadczenie losowe i działania na zdarzeniach

Autorzy/Autorki: Anna Serwatka

Aby przejść do dalszej części podręcznika potrzebujemy podstawowych definicji.

DEFINICJA

Definicja 8: Przestrzeń zdarzeń elementarnych

Przestrzeń zdarzeń elementarnych jest to zbiór wszystkich możliwych wyników danego eksperymentu losowego, oznaczany często przez Ω.

PRZYKŁAD

Przykład 6:

Gdy chcemy wypisać przestrzeń zdarzeń elementarnych zadajmy sobie pytanie jakie zdarzenia mogły zajść w naszym doświadczeniu. Na przykład, gdy rzucamy raz symetryczną sześcienną kostką do gry. Wyniki, które określają liczbę oczek jakie uzyskaliśmy w takim pojedynczym rzucie mogą być następujące: 1,2,3,4,5 lub 6. Dlatego przestrzenią zdarzeń elementarnych w tym zadaniu będzie zbiór $\Omega = \{1, 2, 3, 4, 5, 6\}$.

PRZYKŁAD

Przykład 7:

Rozważmy doświadczenie, które polega na trzykrotnym rzucie monetą. Podczas jednego rzutu możemy otrzymać orła (co będziemy oznaczać literą o) lub reszkę (oznaczamy przez r). Używając tych oznaczeń możemy opisać przestrzeń zdarzeń elementarnych w tym doświadczeniu: $\Omega = \{(o,o,o),(o,r,r),(r,o,r),(r,r,o),(o,o,r),(o,r,o),(r,o,o),(r,r,r)\}$.

DEFINICJA

Definicja 9: Zdarzenie losowe

Dowolny podzbiór przestrzeni zdarzeń elementarnych nazywamy zdarzeniem losowym.

PRZYKŁAD

Przykład 8:

Wróćmy do doświadczenia, w którym rzucamy raz symetryczną sześcienną kostką do gry. Zdarzeniem losowym będzie na przykład wyrzucenie nieparzystej liczby oczek $A = \{1,3,5\}$, a także wyrzucenie liczby oczek, która jest liczbą pierwszą $B = \{2,3,5\}$.

PRZYKŁAD

Przykład 9:

W doświadczeniu, które polegało na trzykrotnym rzucie monetą, zdarzeniem losowym będzie na przykład wyrzucenie dokładnie dwóch orłów $A = \{(o,o,r),(o,r,o),(r,o,o)\}$ czy wyrzucenie co najmniej jednej reszki $B = \{(o,r,r),(r,o,r),(r,r,o),(o,o,r),(o,r,o),(r,o,o),(r,r,r)\}$.

PRZYKŁAD

Przykład 10: Zdarzenie niemożliwe

Zdarzenie niemożliwe to zdarzenie, które nigdy nie zachodzi, np. wyrzucenie liczby 7 w rzucie sześcienną kostką. Prawdopodobieństwo zajścia takiego zdarzenia jest równe 0.

PRZYKŁAD

Przykład 11: Zdarzenie pewne

Zdarzenie pewne to zdarzenie, które zawsze zachodzi, np. wyrzucenie liczby mniejszej niż 8 w rzucie sześcienną kostką. Prawdopodobieństwo zajścia zdarzenia pewnego jest równe 1.

PRZYKŁAD

Przykład 12: Zdarzenie przeciwne

Zdarzenie przeciwne do zdarzenia A jest to zdarzenie, które zachodzi wtedy, gdy nie zachodzi zdarzenie A , oznaczane jako A'. Zdarzeniem przeciwnym do zdarzenia polegającego na wyrzuceniu nieparzystej liczby oczek w jednokrotnym rzucie symetryczną sześcienną kostką do gry, będzie wyrzucenie parzystej liczby oczek.

DEFINICJA

Definicja 10: Prawdopodobieństwo zdarzenia

Funkcja przypisująca każdemu zdarzeniu losowemu liczbę z przedziału $[0, 1]$, oznaczająca stopień pewności jego wystąpienia. W przypadku zdarzeń elementarnych równie prawdopodobnych, prawdopodobieństwo zdarzenia A możemy obliczyć następująco:

$$P(A) = \frac{\text{liczba zdarzeń sprzyjających}}{\text{liczba wszystkich zdarzeń elementarnych}}$$

Działania na zdarzeniach pozwalają opisywać złożone sytuacje losowe, dlatego wprowadza się następujące działania:

DEFINICJA

Definicja 11: Suma zdarzeń

Niech dane będą dwa zdarzenia losowe $A, B \subset \Omega$. Sumą zdarzeń A i B nazywamy zdarzenie polegające na zajściu przynajmniej jednego z dwóch zdarzeń A lub B i oznaczamy $A \cup B$. Prawdopodobieństwo sumy zdarzeń wyraża wzór:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

DEFINICJA

Definicja 12: Iloczyn zdarzeń

Niech dane będą dwa zdarzenia losowe $A, B \subset \Omega$. Iloczynem zdarzeń A i B nazywamy zdarzenie polegające na jednoczesnym zajściu obu zdarzeń A i B i oznaczamy $A \cap B$.

DEFINICJA

Definicja 13: Różnica zdarzeń

Niech dane będą dwa zdarzenia losowe $A, B \subset \Omega$. Różnicą zdarzeń $A \backslash B$ nazywamy zdarzenie, które zachodzi, gdy zajdzie zdarzenie A, ale nie zajdzie B.

3.2. Prawdopodobieństwo klasyczne

Autorzy/Autorki: Anna Serwatka

Prawdopodobieństwo klasyczne to jedna z najstarszych definicji prawdopodobieństwa[2] [4], opracowana przez Pierre'a-Simona Laplace'a. Opiera się na założeniu, że wszystkie zdarzenia elementarne w danym doświadczeniu losowym są jednakowo prawdopodobne.

DEFINICJA

Definicja 14: Prawdopodobieństwo klasyczne

Jeśli przestrzeń zdarzeń elementarnych Ω jest skończona, niepusta i wszystkie zdarzenia elementarne są jednakowo prawdopodobne, natomiast A jest dowolnym zdarzeniem w tej przestrzeni, to prawdopodobieństwem zdarzenia A nazywamy liczbę

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

gdzie $|A|$ to moc zbioru A, tzn. liczba zdarzeń elementarnych tej przestrzeni sprzyjających zdarzeniu A, a $|\Omega|$ to moc zbioru Ω, czyli liczba wszystkich zdarzeń elementarnych przestrzeni Ω.

ZADANIE

Zadanie 9:

Aby prawidłowo rozwiązać zadanie z prawdopodobieństwa klasycznego należy zatem obliczyć liczbę wszystkich zdarzeń elementarnych oraz liczbę zdarzeń sprzyjających, jak pokazuje następujący przykład.

Treść zadania:

Oblicz prawdopodobieństwo otrzymania trzech liczb podzielnych przez 3 w trzykrotnym rzucie sześcienną kostką do gry.

Rozwiązanie:

Zauważmy, że zakładamy, że kostka jest symetryczna, to znaczy, że szansa otrzymania każdego wyniku jest równa $\frac{1}{6}$. Policzmy najpierw moc zbioru Ω. Korzystając z reguły mnożenia otrzymujemy

$|\Omega| = 6 \cdot 6 \cdot 6 = 216$, ponieważ w każdym rzucie mogliśmy otrzymać jedną z 6 liczb. Zdarzenie A polega na tym, że trzy razy wyrzuciliśmy liczbę podzielną przez 3, czyli za każdym razem była to 3 lub 6 (dwie możliwości na każdym rzucie), stąd: $|A| = 2^3 = 8$. Korzystając ze wzoru na prawdopodobieństwo klasyczne otrzymujemy: $P(A) = \frac{8}{216} = \frac{1}{27}$.

ZADANIE

Zadanie 10:

Treść zadania:

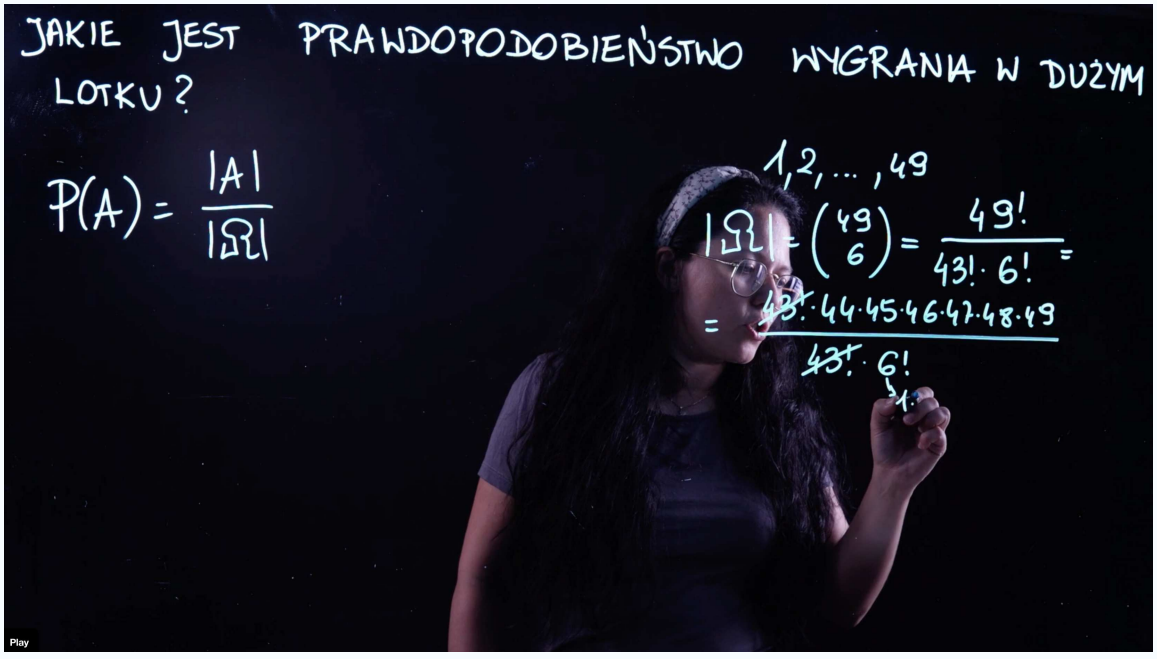
Doświadczenie losowe polega na dwukrotnym rzucie symetryczną sześcienną kostką do gry. Oblicz prawdopodobieństwo otrzymania sumy oczek mniejszej od 4.

Rozwiązanie:

Tak, jak poprzednio zakładamy, że kostka jest symetryczna, to znaczy, że szansa otrzymania każdego wyniku jest równa $\frac{1}{6}$. Korzystając z reguły mnożenia otrzymujemy $|\Omega| = 6 \cdot 636$, ponieważ w każdym rzucie mogliśmy otrzymać jedną z 6 liczb. Wypiszmy wszystkie zdarzenia sprzyjające zdarzeniu $A = \{(1,1),(1,2),(2,1)\}$. Zauważmy, że $|A| = 3$ i korzystając ze wzoru na prawdopodobieństwo klasyczne otrzymujemy: $P(A) = \frac{3}{36} = \frac{1}{12}$.

PRZYKŁAD

Przykład 13: Wygrana w dużym lotku

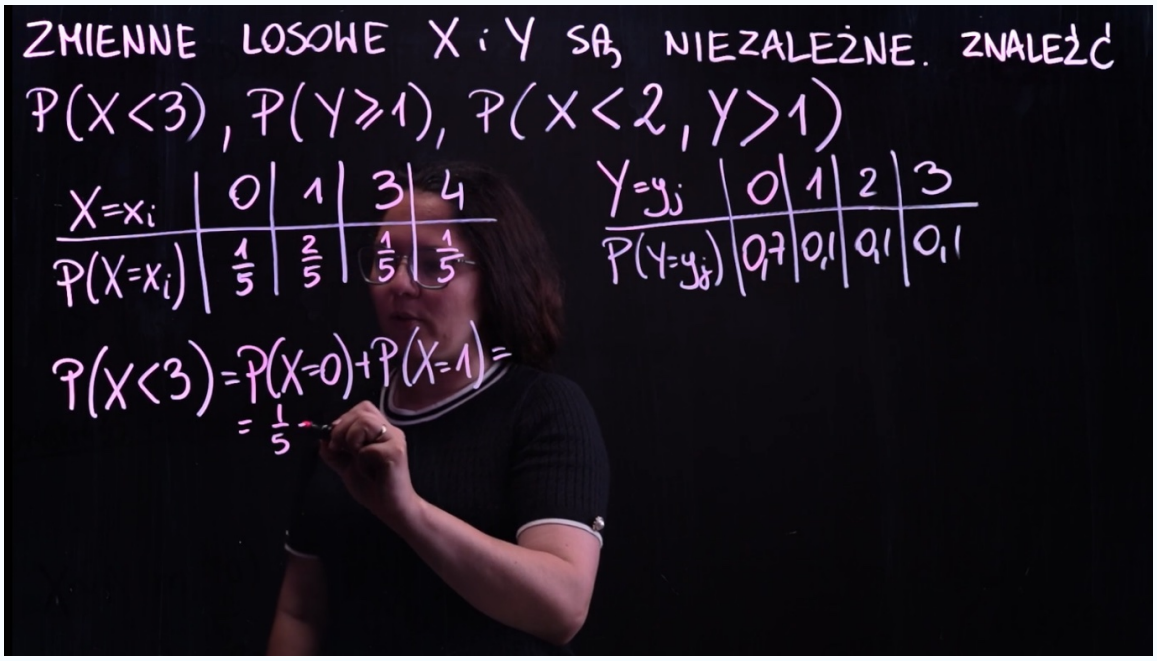


https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2378/reader#openaghmod:2378:openaghhfivep:1

Prawdopodobieństwo klasyczne - przykład z dużym lotkiem

PRZYKŁAD

Przykład 14: Prawdopodobieństwo



https://vimeo.com/1097815116/01066062ad

Prawdopodobieństwo

3.3. Doświadczenie losowe wieloetapowe

Autorzy/Autorki: Anna Serwatka

W rachunku prawdopodobieństwa często mamy do czynienia z sytuacjami, w których wynik końcowy zależy od ciągu niezależnych lub wzajemnie powiązanych zdarzeń. Takie sekwencje nazywamy doświadczeniami wieloetapowymi. Analiza tych doświadczeń wymaga zrozumienia, jak prawdopodobieństwa poszczególnych

etapów łączą się, by wpłynąć na prawdopodobieństwo ostatecznego rezultatu. Na przykład proces produkcyjny składający się z kilku faz, każda z własnym ryzykiem defektu, albo seria rzutów monetą, gdzie interesuje nas liczba orłów, to typowe przykłady doświadczeń wieloetapowych [5].

DEFINICJA

Definicja 15: Doświadczenie wieloetapowe

Doświadczenie wieloetapowe to proces, w którym przebieg zdarzenia składa się z kilku kolejnych etapów, a wynik każdego etapu może zależeć od wcześniejszych etapów i wpływa na dalszy rozwój sytuacji. Każdy etap doświadczenia charakteryzuje się określonymi prawdopodobieństwami przejścia do różnych stanów. Tego rodzaju procesy są często modelowane za pomocą drzewa stochastycznego, które graficznie przedstawia wszystkie możliwe ścieżki rozwoju sytuacji oraz ich prawdopodobieństwa.

DEFINICJA

Definicja 16: Drzewo stochastyczne

Drzewem stochastycznym nazywamy graf ilustrujący przebieg wieloetapowego doświadczenia losowego. Wierzchołkom drzewa stochastycznego przyporządkowane są wyniki poszczególnych etapów doświadczenia, na krawędziach zaznaczane jest prawdopodobieństwo uzyskania tych wyników. Suma prawdopodobieństw przyporządkowanych krawędziom wychodzącym z tego samego wierzchołka jest równa 1. Prawdopodobieństwo zdarzenia reprezentowanego przez jedną gałąź drzewa jest równe iloczynowi prawdopodobieństw przyporządkowanych krawędziom, z których składa się rozważana gałąź.

ZADANIE

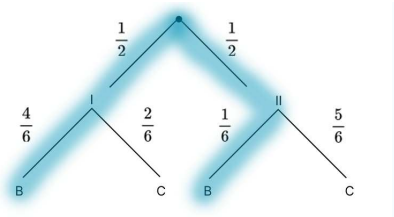
Zadanie 11:

Treść zadania:

Mamy dwie urny: w jednej z nich są 4 kule białe i 2 czarne, a w drugiej urnie jest 1 biała i 5 czarnych. Z urny przypadkowo wybranej wyciągamy losowo kulę. Oblicz prawdopodobieństwo zdarzenia polegającego na wyciągnięciu kuli białej, jeżeli prawdopodobieństwo wyboru każdej z urn jest równe 0,5.

Rozwiązanie:

Doświadczenie składa się z dwóch etapów. W pierwszym etapie losujemy urnę I lub II. Następnie jeśli wylosowaliśmy urnę I to możemy wylosować z niej kulę białą, oznaczmy to jako B (z prawdopodobieństwem $\frac{4}{6}$) lub czarną, co oznaczymy jako C (z prawdopodobieństwem $\frac{2}{6}$). Analogicznie, jeśli wypadła urna II, to możemy z niej wylosować kulę białą (z prawdopodobieństwem $\frac{1}{6}$) lub czarną (z prawdopodobieństwem $\frac{5}{6}$). Zobrazujemy to na drzewie. Interesuje nas ścieżka prowadząca do wylosowania kuli białej. Stąd mamy $P(B) = \frac{1}{2} \cdot \frac{4}{6} + \frac{1}{2} \cdot \frac{1}{6} = \frac{5}{12}$.



Rysunek 1: Drzewo stochastyczne dotyczące doświadczenia losowego z zadania 1. Opracowanie własne

3.4. Prawdopodobieństwo warunkowe, zupełne i niezależność zdarzeń

Autorzy/Autorki: Anna Serwatka

DEFINICJA

Definicja 17: Niezależność dwóch zdarzeń

Mówimy, że zdarzenia $A, B \subset \Omega$ są niezależne, jeżeli $P(A \cap B) = P(A)P(B)$.

DEFINICJA

Definicja 18: Niezależność zespołowa n zdarzeń

Mówimy, że zdarzenia $A_1, A_2, \ldots, A_n \subset \Omega$ są zespołowo niezależne, jeżeli prawdopodobieństwo łącznego zajścia dowolnych m, $m \leq n$, różnych zdarzeń spośród nich jest równe iloczynowi prawdopodobieństw tych zdarzeń: $P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_m})$ dla każdego $m \leq n$ i każdego m-wyrazowego rosnącego ciągu $i_1, i_2, \ldots, i_m$ liczb naturalnych, takich, że $1 \leq i_1, i_2, \ldots, i_m \leq n$.

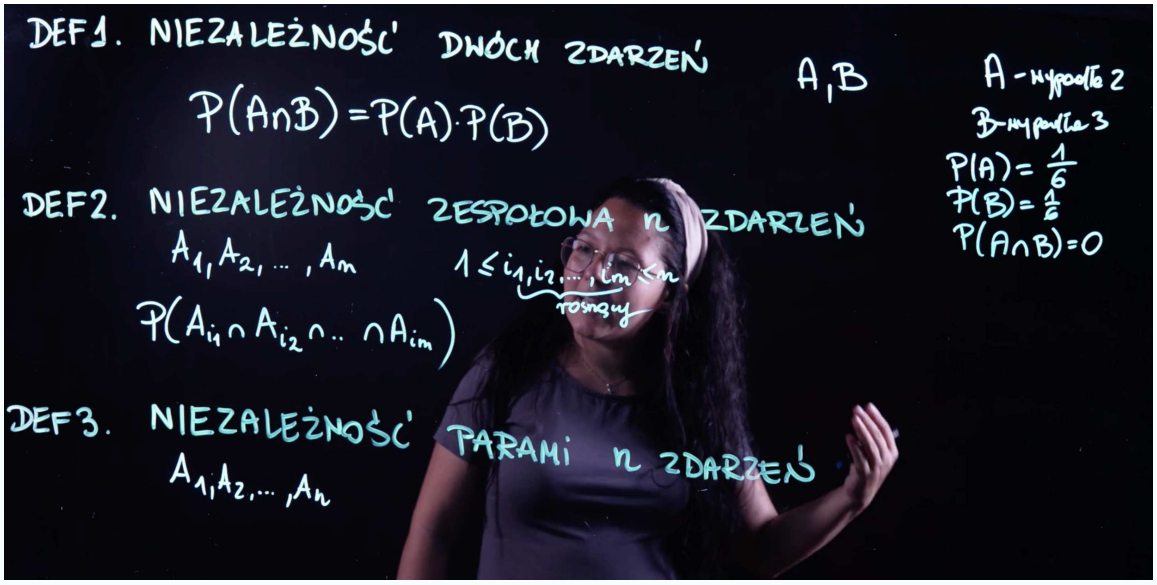
DEFINICJA

Definicja 19: Niezależność parami n zdarzeń

Mówimy, że zdarzenia $A_1, A_2, \ldots, A_n \subset \Omega$ są parami niezależne, jeżeli każde dwa różne zdarzenia spośród nich są niezależne, czyli $P(A_i \cap A_j) = P(A_i)P(A_j)$, dla $i \neq j$, $i, j = 1, 2, \ldots, n$.

PRZYKŁAD

Przykład 15: Różnica między niezależnością dwóch zdarzeń a niezależnością zespołową



https://vimeo.com/1027621416/c540a4836f

Niezależność zdarzeń

TWIERDZENIE

Twierdzenie 6: Twierdzenie o prawdopodobieństwie zupełnym

ZAŁOŻENIE:

Niech $B \subset \Omega$ będzie dowolnym zdarzeniem, niech $A_1, \ldots, A_n \subset \Omega$ będą zdarzeniami wykluczającymi się parami, tzn. dla każdego $i \neq j$ zachodzi $A_i \cap A_j = \emptyset$. Ponadto $A_1 \cup \ldots \cup A_n = \Omega$, dla $i = 1, \ldots, n$ prawdopodobieństwo $P(A_i) > 0$.

TEZA:

Prawdopodobieństwo zdarzenia B wyraża się równością:

$$P(B) = P(A_1)P(B|A_1) + \ldots + P(A_n)P(B|A_n). \tag{3.1}$$

TWIERDZENIE

Twierdzenie 7: Twierdzenie Bayesa

ZAŁOŻENIE:

Niech $B \subset \Omega$ będzie dowolnym zdarzeniem, takim, że $P(B) > 0$, niech $A_1, \ldots, A_n \subset \Omega$ będą zdarzeniami wykluczającymi się parami, tzn. dla każdego $i \neq j$ zachodzi $A_i \cap A_j = \emptyset$. Ponadto $A_1 \cup \ldots \cup A_n = \Omega$, dla $i = 1, \ldots, n$ prawdopodobieństwo $P(A_i) > 0$.

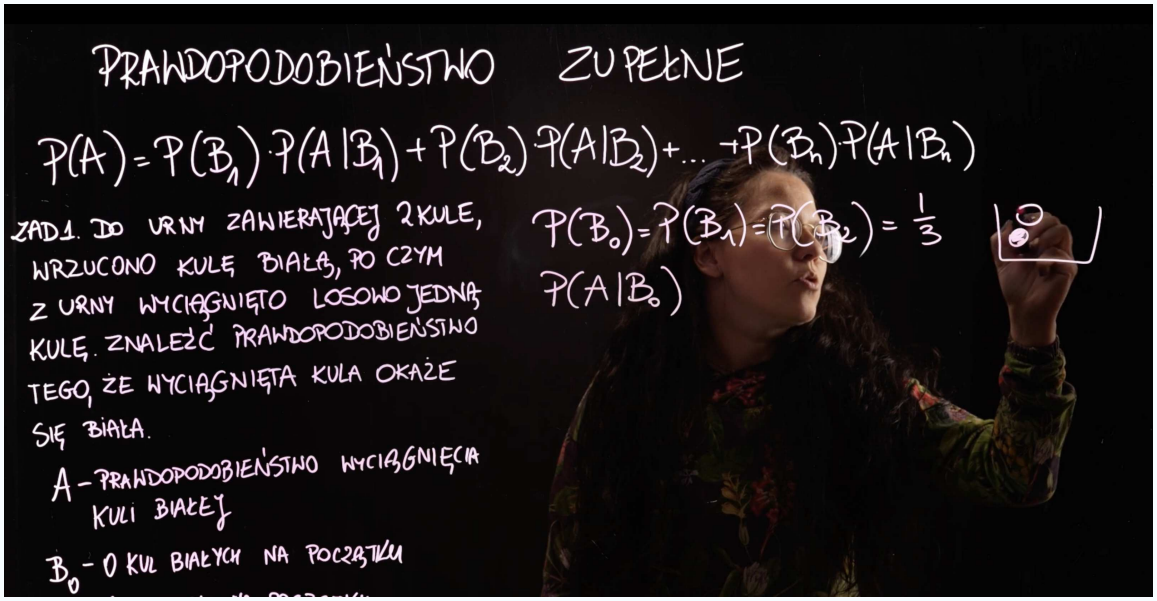
TEZA:

Prawdopodobieństwa warunkowe $P(A_k|B)$ zdarzeń A_k przy warunku B wyrażają się równościami:

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}, \text{dla} k = 1, \ldots, n. \tag{3.2}$$

PRZYKŁAD

Przykład 16: Zadanie z prawdopodobieństwem zupełnym

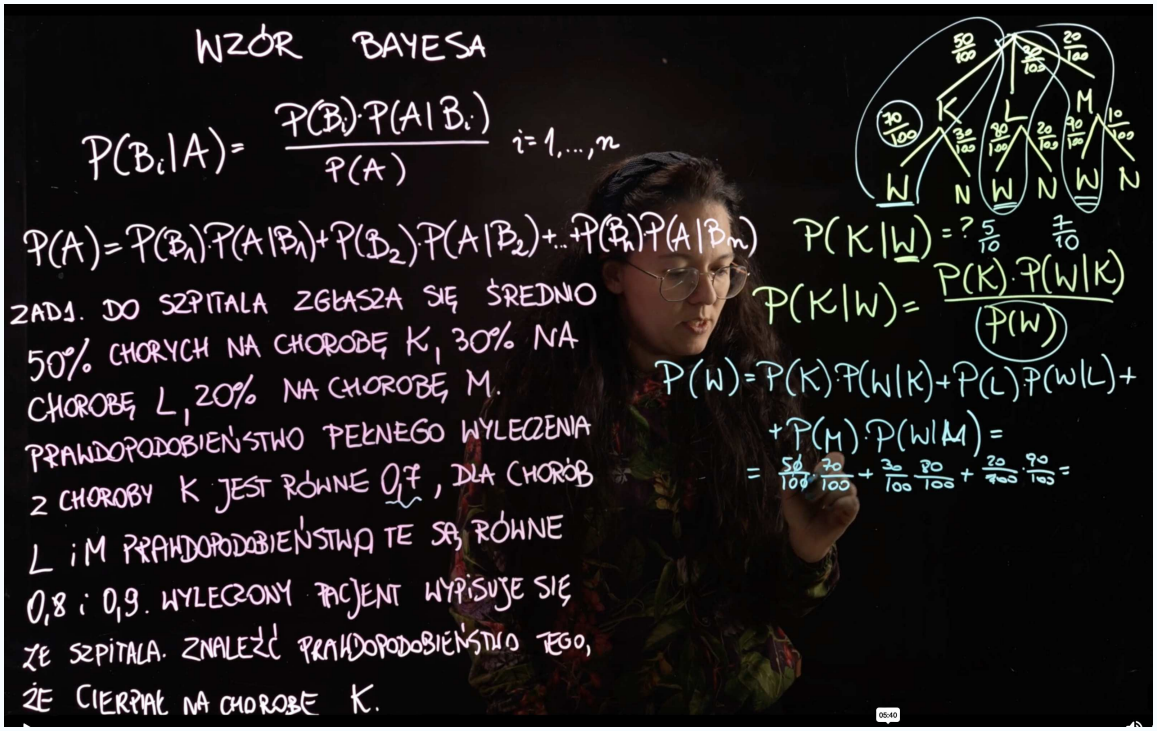


https://vimeo.com/1029557190/e9d0af04ad

Prawdopodobieństwo zupełne

PRZYKŁAD

Przykład 17: Zastosowanie twierdzenia Bayesa



https://vimeo.com/1029564754/58c4ca487b

Wzor Bayesa

3.5. Schemat Bernoulliego

Autorzy/Autorki: Anna Serwatka

Wiele zjawisk losowych, z którymi spotykamy się na co dzień, można opisać jako sekwencję powtórzeń tego samego, prostego eksperymentu. Jeśli każdy z tych eksperymentów ma tylko dwa możliwe wyniki – zazwyczaj określane jako "sukces" lub "porażka" – a prawdopodobieństwo sukcesu pozostaje stałe w każdym powtórzeniu, mówimy wówczas o schemacie Bernoulliego. Jest to konstrukcja w rachunku prawdopodobieństwa, pozwalająca na modelowanie serii niezależnych prób, takich jak rzuty monetą, kontrola jakości produktów na linii produkcyjnej czy określanie liczby osób spełniających dany warunek w losowej próbie [2].

TWIERDZENIE

Twierdzenie 8: Wzór Bernoulliego

ZAŁOŻENIE:

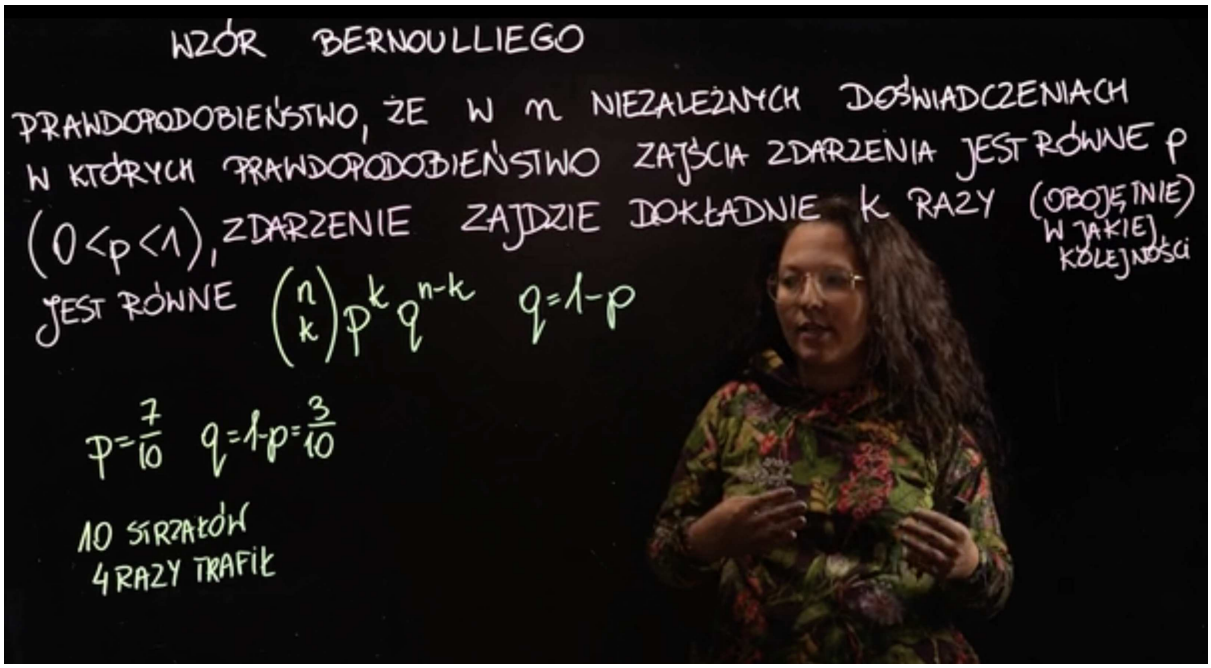
Rozważamy doświadczenie losowe polegające na n niezależnych doświadczeniach. Prawdopodobieństwo zajścia pojedynczego zdarzenia jest równe $p \in [0, 1]$.

TEZA:

Prawdopodobieństwo, że w n niezależnych doświadczeniach, zdarzenie zejdzie dokładnie k razy $(k \leq n)$ jest równe

$$\binom{n}{k} p^k q^{n-k} \tag{3.3}$$

gdzie $q = 1 - p$.



https://vimeo.com/1029565298/2d3487a826

Wzór Bernoulliego

Rozdział 4

Jednowymiarowe zmienne losowe

4.1. Zmienna losowa i jej dystrybuanta

Autorzy/Autorki: Anna Serwatka

W rachunku prawdopodobieństwa i statystyce matematycznej bardzo ważnym pojęciem jest zmienna losowa. Pozwala ona na matematyczne modelowanie zjawisk, w których wynik nie jest z góry znany, a jedynie można określić prawdopodobieństwo jego wystąpienia. Zmienna losowa przypisuje wynikom eksperymentu losowego wartości liczbowe, co umożliwia ich analizę statystyczną. Ściśle związana ze zmienną losową jest jej dystrybuanta, która dostarcza kompleksowej informacji o rozkładzie prawdopodobieństwa tej zmiennej.

Zmienna losowa to funkcja mierzalna [4], która każdemu zdarzeniu elementarnemu z przestrzeni zdarzeń elementarnych Ω przyporządkowuje pewną liczbę rzeczywistą. Aby przedstawić definicję zmiennej losowej, zacznijmy od definicji σ-ciała.

DEFINICJA

Definicja 20: σ-ciało

σ-**ciałem** (lub σ-**algebrą**) $\mathcal{F}$ na przestrzeni Ω nazywamy rodzinę podzbiorów Ω, która spełnia następujące warunki:

1. Cała przestrzeń zdarzeń elementarnych Ω należy do $\mathcal{F}$:
$$\Omega \in \mathcal{F}$$
2. Jeśli zdarzenie A należy do $\mathcal{F}$, to jego dopełnienie A^c (zdarzenie, że A nie zajdzie) również należy do $\mathcal{F}$:
$$\text{Jeśli } A \in \mathcal{F}, \text{ to } A^c \in \mathcal{F}$$
3. Jeśli mamy przeliczalną sekwencję zdarzeń $A_1, A_2, A_3, \ldots$ należących do $\mathcal{F}$, to ich suma (zdarzenie, że zajdzie przynajmniej jedno z nich) również należy do $\mathcal{F}$:
$$\text{Jeśli } A_1, A_2, A_3, \cdots \in \mathcal{F}, \text{ to } \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

DEFINICJA

Definicja 21:

Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną, gdzie Ω to przestrzeń zdarzeń elementarnych, $\mathcal{F}$ to σ-ciało zdarzeń, a P to miara probabilistyczna. Zmienną losową X nazywamy funkcję $X : \Omega \to \mathbb{R}$ taką, że dla każdej liczby rzeczywistej $x \in \mathbb{R}$, zbiór $\{\omega \in \Omega : X(\omega) \leq x\}$ należy do $\mathcal{F}$.

Intuicyjnie oznacza to, że możemy obliczyć prawdopodobieństwo, że zmienna losowa przyjmie wartość mniejszą lub równą x. Przykłady, definicje i interpretację tego zagadnienia można znaleźć w rozdziale Definicja jednowymiarowej zmiennej losowej.

Zmienne losowe można podzielić na dwa główne typy [6]: dyskretne i ciągłe.

Zmienna losowa **dyskretna** to taka, która może przyjmować jedynie skończoną lub przeliczalną liczbę wartości. Wartości te są zazwyczaj liczbami całkowitymi. Dla zmiennej losowej dyskretnej możemy określić funkcję prawdopodobieństwa (masy prawdopodobieństwa), która dla każdej możliwej wartości x_i zmiennej losowej X podaje prawdopodobieństwo $P(X = x_i)$. Suma wszystkich prawdopodobieństw musi być równa 1.

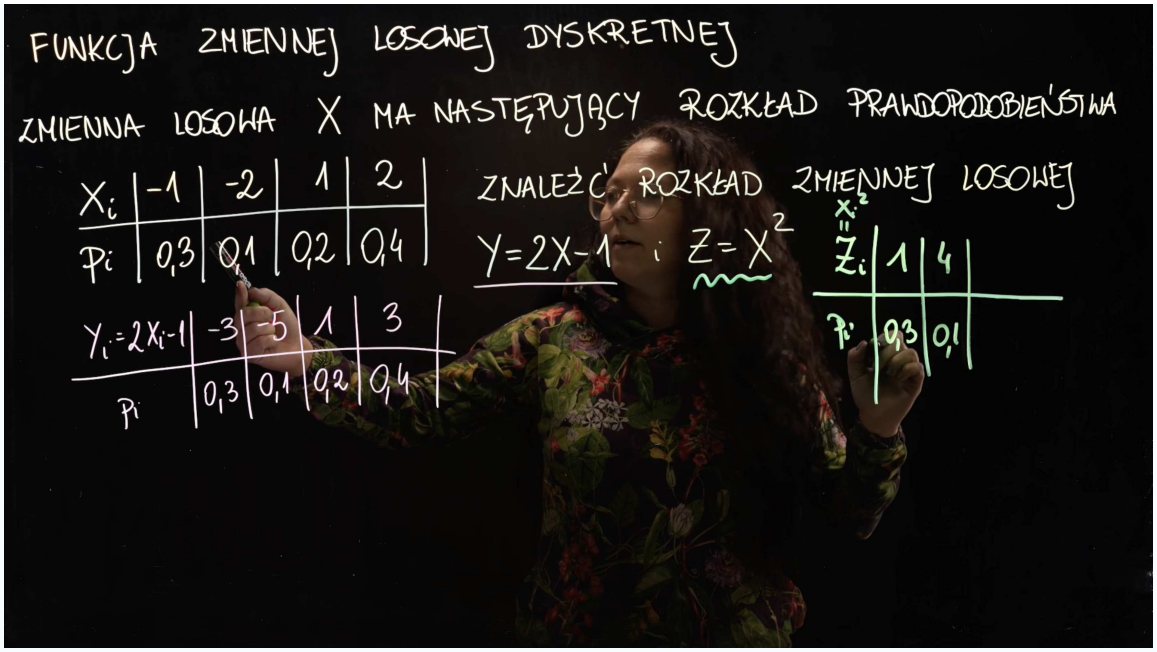
PRZYKŁAD

Przykład 18: Rzut kostką

Rzut kostką do gry. Niech X będzie zmienną losową reprezentującą wynik rzutu sześcienną kostką. Możliwe wartości X to $\{1, 2, 3, 4, 5, 6\}$. Dla każdej z tych wartości prawdopodobieństwo wynosi $1/6$.

PRZYKŁAD

Przykład 19: Funkcje zmiennej losowej skokowej



https://vimeo.com/1027592929/2ddc4692c2

Funkcja zmiennej losowej dyskretnej

Zmienna losowa **ciągła** [2] to taka, która może przyjmować dowolną wartość z pewnego przedziału liczb rzeczywistych. W przeciwieństwie do zmiennych dyskretnych, dla zmiennej losowej ciągłej prawdopodobieństwo przyjęcia konkretnej wartości jest równe zero, tj. $P(X = x) = 0$. Zamiast tego, dla zmiennych ciągłych definiujemy funkcję gęstości prawdopodobieństwa, $f_X(x)$, która spełnia warunki $f_X(x) \geq 0$ dla wszystkich x oraz $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Prawdopodobieństwo, że zmienna losowa przyjmie wartość z danego przedziału $[a, b]$, oblicza się poprzez całkowanie funkcji gęstości w tym przedziale: $P(a \leq X \leq b) = \int_a^b f_X(x)dx$.

PRZYKŁAD

Przykład 20: Oczekiwanie na autobus

Czas oczekiwania na autobus. Niech Y będzie zmienną losową reprezentującą czas oczekiwania na autobus (w minutach). Y może przyjmować dowolną wartość nieujemną. Jeśli zakładamy, że czas oczekiwania jest rozłożony wykładniczo, to jego funkcja gęstości będzie postaci $f_Y(y) = \lambda e^{-\lambda y}$ dla $y \geq 0$, gdzie $\lambda > 0$ jest parametrem.

Dystrybuanta zmiennej losowej X, oznaczana jako $F_X(x)$, jest zdefiniowana dla każdej liczby rzeczywistej x jako prawdopodobieństwo, że zmienna losowa X przyjmie wartość mniejszą lub równą x:

$$F_X(x) = P(X \leq x) \tag{4.1}$$

Można przyjąć w definicji dystrybuanty nierówność ostrą $F_X(x) = P(X < x)$ Dystrybuanta dla zmiennej losowej dyskretnej ma charakter schodkowy. Jej wartość zmienia się skokowo tylko w punktach, w których zmienna losowa może przyjmować wartości, a w pozostałych przedziałach jest stała. Skok w punkcie x_i jest równy $P(X = x_i)$.

PRZYKŁAD

Przykład 21: Rzut kostką - ciąg dalszy

Dystrybuanta dla zmiennej losowej X (wynik rzutu kostką) wygląda następująco: $F_X(x) =$

$$\begin{cases} 0 & \text{dla } x < 1 \\ 1/6 & \text{dla } 1 \leq x < 2 \\ 2/6 & \text{dla } 2 \leq x < 3 \\ 3/6 & \text{dla } 3 \leq x < 4 \\ 4/6 & \text{dla } 4 \leq x < 5 \\ 5/6 & \text{dla } 5 \leq x < 6 \\ 1 & \text{dla } x \geq 6 \end{cases}$$

Dystrybuanta dla zmiennej losowej ciągłej jest funkcją ciągłą. Można ją otrzymać poprzez całkowanie funkcji gęstości prawdopodobieństwa:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{4.2}$$

Z kolei funkcję gęstości prawdopodobieństwa można odzyskać z dystrybuanty poprzez różniczkowanie (tam, gdzie dystrybuanta jest różniczkowalna) (Zmienna losowa i jej dystrybuanta):

$$f_X(x) = \frac{d}{dx}F_X(x) \tag{4.3}$$

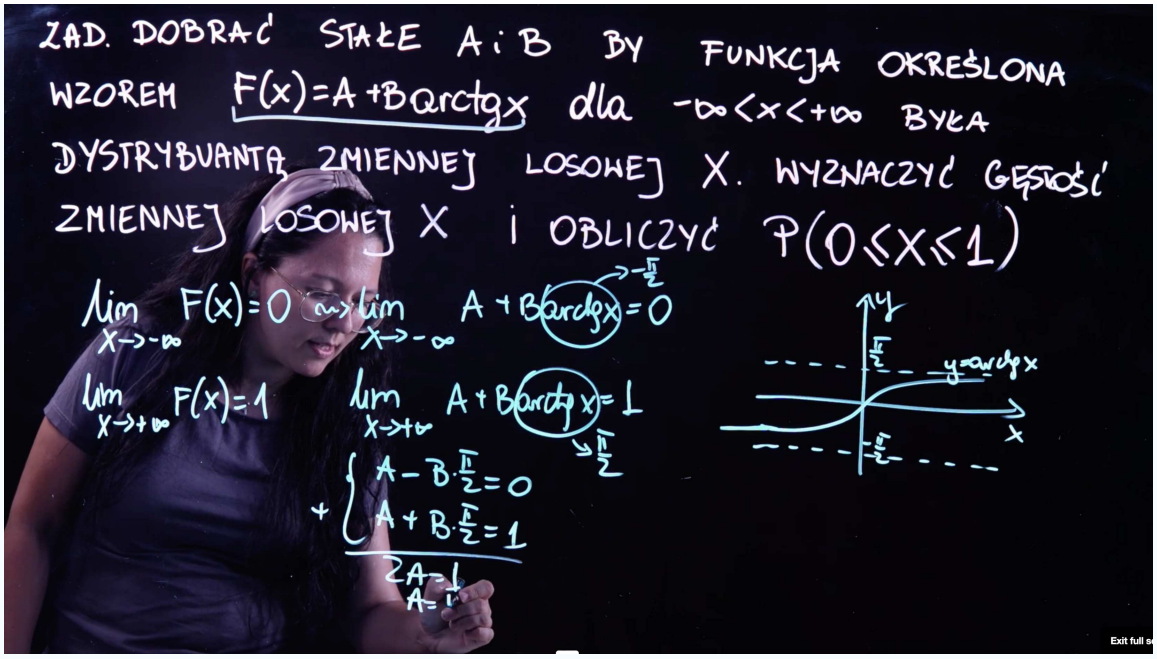
PRZYKŁAD

Przykład 22: Oczekiwanie na autobus - ciąg dalszy

Jeśli $f_Y(y) = \lambda e^{-\lambda y}$ dla $y \geq 0$, to dystrybuanta $F_Y(y)$ dla $y \geq 0$ wynosi: $F_Y(y) = \int_0^y \lambda e^{-\lambda t}dt = [-e^{-\lambda t}]_0^y = -e^{-\lambda y} - (-e^0) = 1 - e^{-\lambda y}$ Dla $y < 0$, $F_Y(y) = 0$.

PRZYKŁAD

Przykład 23: Dystrybuanta i gęstość zmiennej losowej ciągłej



https://vimeo.com/1027598775/7fc454bb0d

Dystrybuanta

4.2. Wartość oczekiwana

Autorzy/Autorki: Anna Serwatka

Wartość oczekiwana zmiennej losowej jest parametrem opisującym rozkład prawdopodobieństwa. Jest to miara położenia rozkładu, intuicyjnie reprezentująca "przeciętną" wartość, jaką zmienna losowa przyjmuje przy wielokrotnym powtarzaniu eksperymentu. Wartość oczekiwana zmiennej losowej X jest oznaczana jako $E[X]$ lub μ [4].

Dla zmiennej losowej dyskretnej X, która przyjmuje wartości $x_1, x_2, \ldots, x_n, \ldots$ z prawdopodobieństwami odpowiednio $P(X = x_1), P(X = x_2), \ldots, P(X = x_n), \ldots$, wartość oczekiwaną definiujemy jako sumę iloczynów każdej możliwej wartości przez jej prawdopodobieństwo (1).

Sumowanie odbywa się po wszystkich możliwych wartościach, jakie zmienna losowa X może przyjąć.

PRZYKŁAD

Przykład 24:

Wartość oczekiwana rzutu sześcienną kostką Wartość oczekiwaną wyniku rzutu sześcienną kostką (X) liczymy następująco: $E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$
Oznacza to, że jeśli rzucamy kostką bardzo wiele razy, średni wynik będzie zbliżał się do 3.5.

Dla zmiennej losowej ciągłej X z funkcją gęstości prawdopodobieństwa $f_X(x)$, wartość oczekiwaną

definiujemy jako całkę z iloczynu wartości zmiennej losowej i jej funkcji gęstości, po całym zakresie wartości, jakie zmienna może przyjąć (2).

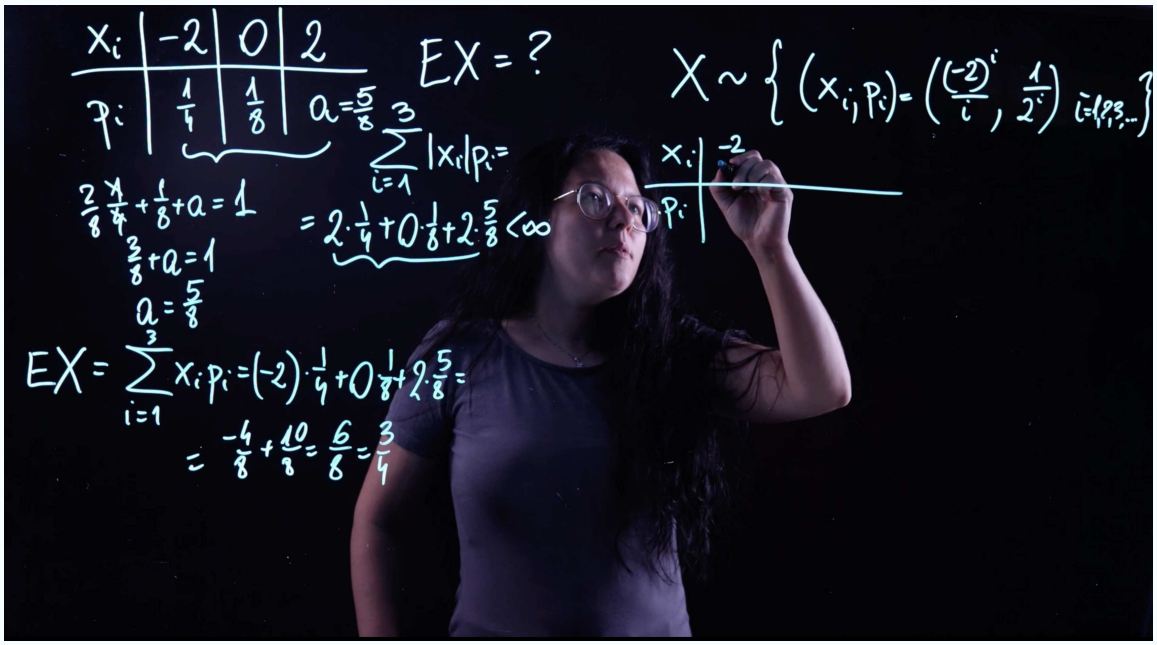
PRZYKŁAD

Przykład 25: Wartość oczekiwana czasu oczekiwania na autobus

Wartość oczekiwana czasu oczekiwania na autobus (Y) z rozkładem wykładniczym o funkcji gęstości $f_Y(y) = \lambda e^{-\lambda y}$ dla $y \geq 0$: $E[Y] = \int_0^\infty y \cdot \lambda e^{-\lambda y} dy$ Całkując przez części ($\int udv = uv - \int vdu$, gdzie $u = y$, $dv = \lambda e^{-\lambda y} dy$, $du = dy$, $v = -e^{-\lambda y}$): $E[Y] = \left[-ye^{-\lambda y}\right]_0^\infty - \int_0^\infty (-e^{-\lambda y})dy = (0-0) + \int_0^\infty e^{-\lambda y} dy$ $E[Y] = \left[-\frac{1}{\lambda} e^{-\lambda y}\right]_0^\infty = (0 - (-\frac{1}{\lambda})) = \frac{1}{\lambda}$ Jeśli np. średnio autobus przyjeżdża co 10 minut ($\lambda = 1/10$), to średni czas oczekiwania wynosi $1/\lambda = 10$ minut.

PRZYKŁAD

Przykład 26:

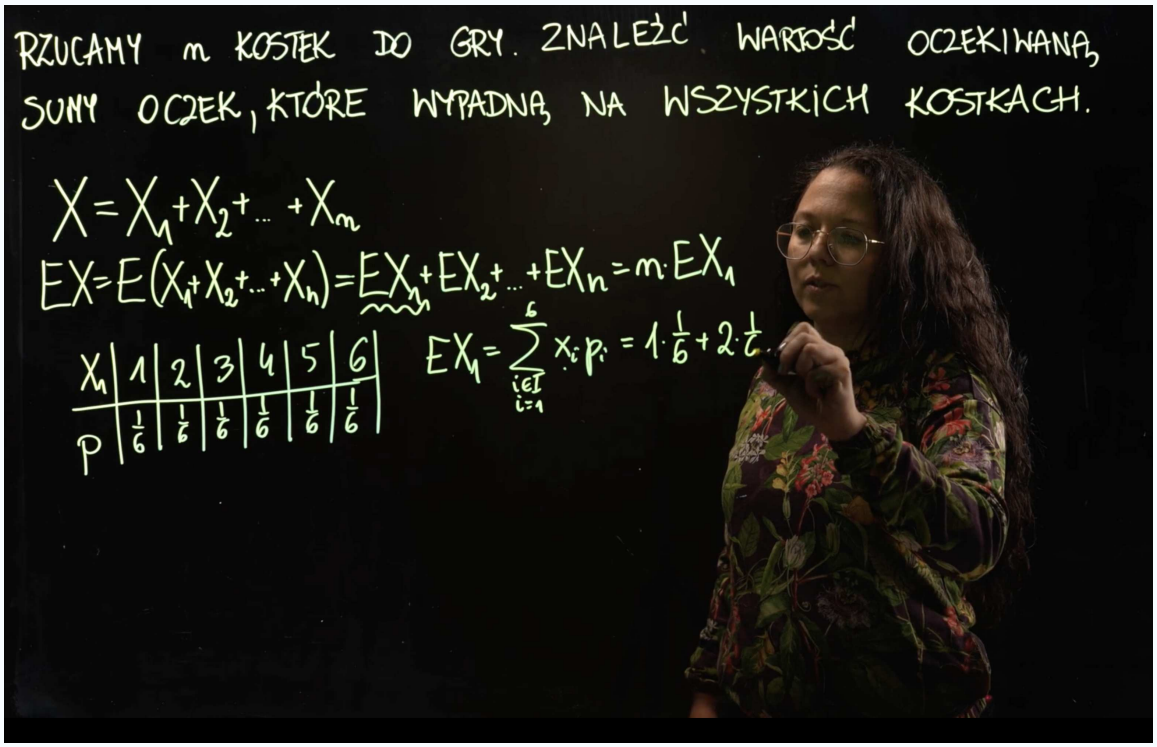


https://vimeo.com/1029557566/598f8c5419

Wartość oczekiwana cz 1.

PRZYKŁAD

Przykład 27:



https://vimeo.com/1029563052/4ee9a55640

Wartość oczekiwana cz 2.

4.3. Momenty

Autorzy/Autorki: Anna Serwatka

W statystyce, **momenty** to ogólne miary, które opisują kształt i charakterystykę rozkładu prawdopodobieństwa zmiennej losowej. Są to uogólnienia wartości oczekiwanej i dostarczają informacji o różnych aspektach rozkładu, takich jak jego położenie, rozproszenie, asymetria i spłaszczenie.

Istnieją dwa główne rodzaje momentów:

Momenty zwykłe to wartości oczekiwane potęg zmiennej losowej. Moment zwykły rzędu k zmiennej losowej X jest zdefiniowany jako $E[X^k]$.

Pierwszy moment zwykły ($k = 1$) to po prostu **wartość oczekiwana** (średnia) zmiennej losowej: $E[X^1] = E[X]$.

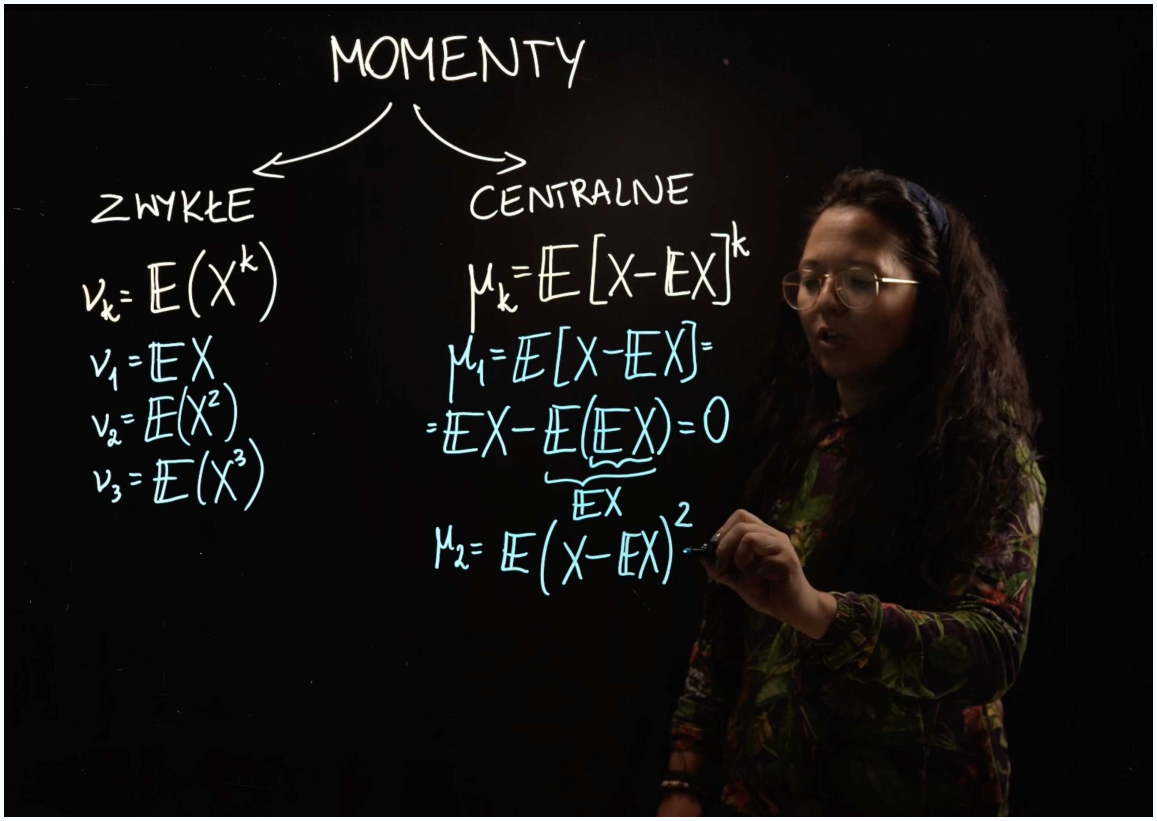
Momenty centralne to wartości oczekiwane potęg odchyleń zmiennej losowej od jej wartości oczekiwanej

(średniej). Moment centralny rzędu k zmiennej losowej X jest zdefiniowany jako $E[(X - E[X])^k]$.

- Pierwszy moment centralny ($k = 1$) zawsze wynosi zero: $E[X - E[X]] = E[X] - E[X] = 0$.
- Drugi moment centralny ($k = 2$) to **wariancja** zmiennej losowej: $E[(X - E[X])^2] = Var(X) = D^2X$. Wariancja mierzy rozproszenie danych wokół średniej [6].
- Trzeci moment centralny ($k = 3$) jest związany ze **skośnością (asymetrią)** rozkładu. Dodatnia wartość oznacza rozkład z „dłuższym ogonem" po prawej stronie (skośność prawostronna), a ujemna po lewej (skośność lewostronna).
- Czwarty moment centralny ($k = 4$) jest związany z **kurtozą (spłaszczeniem)** rozkładu. Mierzy on „szczytowość" rozkładu i „grubość ogonów" w porównaniu do rozkładu normalnego. Wysoka kurtoza oznacza bardziej szpiczasty rozkład z cięższymi ogonami, niska — bardziej płaski rozkład z lżejszymi ogonami.

PRZYKŁAD

Przykład 28: Momenty



https://vimeo.com/1027601740/b69c140a1b

Momenty

4.4. Nierówność Czebyszewa

Autorzy/Autorki: Anna Serwatka

Nierówność Czebyszewa (lub nierówność Czebyszewa-Bienaymé) jest fundamentalnym twierdzeniem w teorii prawdopodobieństwa, które dostarcza górnego oszacowania prawdopodobieństwa, że zmienna losowa oddali się od swojej wartości oczekiwanej o więcej niż pewną ustaloną wartość. Co istotne, nierówność ta jest uniwersalna — obowiązuje dla **dowolnego** rozkładu prawdopodobieństwa, pod warunkiem, że zmienna losowa ma skończoną wartość oczekiwaną i skończoną wariancję. Dzięki swojej ogólności jest niezwykle użyteczna w sytuacjach, gdy nie znamy dokładnego rozkładu zmiennej losowej, a jedynie jej średnią i wariancję.

DEFINICJA

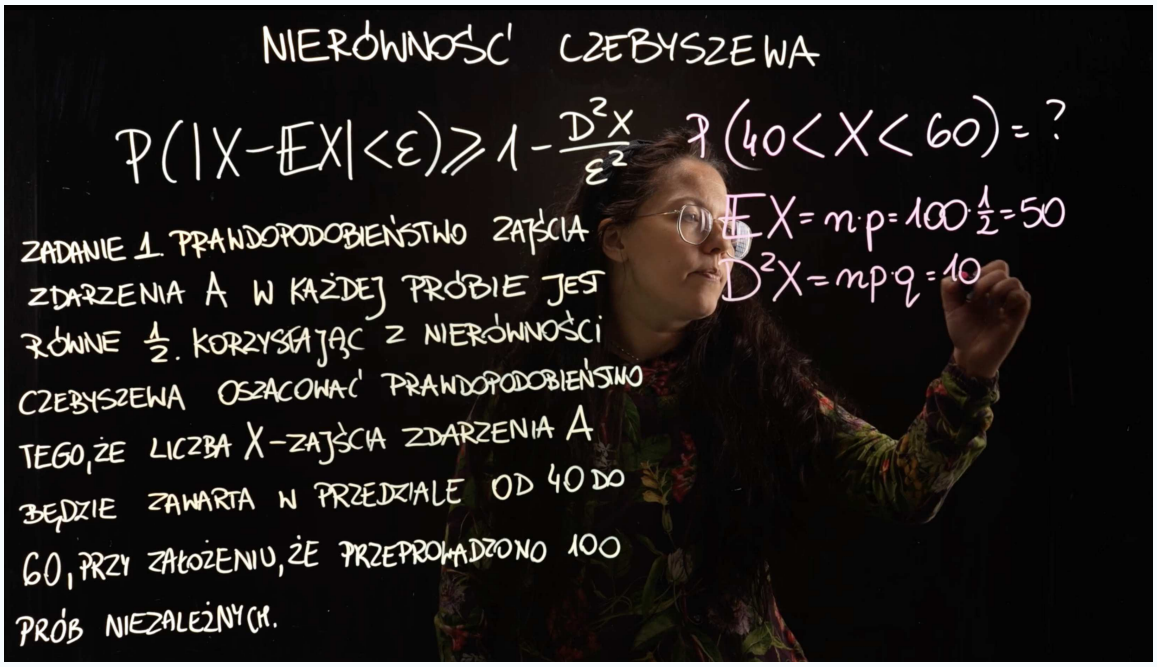
Definicja 22:

Niech X będzie zmienną losową o skończonej wartości oczekiwanej $E[X] = \mu$ i skończonej wariancji $D^2X = \sigma^2$. Wówczas dla dowolnej liczby rzeczywistej $\epsilon > 0$, nierówność Czebyszewa [6] głosi, że:

$$P(|X - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2} \tag{4.4}$$

PRZYKŁAD

Przykład 29:



https://vimeo.com/1027609060/af03a35803

Nierówność Czebyszewa

Rozdział 5

Wielowymiarowe zmienne losowe

5.1. Dwuwymiarowa zmienna losowa

Autorzy/Autorki: Anna Serwatka

W analizie statystycznej i rachunku prawdopodobieństwa często spotykamy się z sytuacjami, w których istotne są nie pojedyncze zmienne losowe, lecz ich zestawy. Przykładem może być jednoczesne badanie wzrostu i wagi, dochodu i wydatków czy temperatury i wilgotności. W takich przypadkach konieczne jest rozszerzenie pojęcia zmiennej losowej na więcej niż jeden wymiar. Rozdział ten wprowadza pojęcie dwuwymiarowej zmiennej losowej, która opisuje zależność między dwiema zmiennymi losowymi rozpatrywanymi równocześnie. Omawiamy tu zarówno przypadki dyskretne, jak i ciągłe, wprowadzając pojęcia takie jak funkcja prawdopodobieństwa wspólnego, funkcja gęstości, funkcje rozkładów brzegowych, funkcje rozkładów warunkowych oraz kowariancja i korelacja, które pozwalają zrozumieć zależności między zmiennymi [7].

DEFINICJA

Definicja 23: Dwuwymiarowa zmienna losowa

Niech X i Y będą zmiennymi losowymi określonymi na pewnych przestrzeniach probabilistycznych, to parę (X, Y) nazywamy dwuwymiarową zmienną losową.

DEFINICJA

Definicja 24: Dystrybuanta dwuwymiarowej zmiennej losowej

Dystrybuantą dwuwymiarowej zmiennej losowej nazywamy funkcję dwóch zmiennych x i y taką, że

$$F(x, y) = P(X < x, Y < y) \tag{5.1}$$

dla $(x, y) \in \mathbb{R}^2$.

TWIERDZENIE

Twierdzenie 9: Własności dystrybuanty dwuwymiarowej zmiennej losowej

TEZA:

Niech F będzie dystrybuantą dwuwymiarowej zmiennej losowej. Wtedy:

1. $\lim_{y\to-\infty}F(x,y)=0$ dla dowolnego $x\in\mathbb{R}$ oraz $\lim_{x\to-\infty}F(x,y)=0$ dla dowolnego $y\in\mathbb{R}$;
2. $\lim_{x\to\infty y\to\infty}F(x,y)=1$;
3. F jest funkcją niemalejącą i lewostronnie ciągłą względem każdego argumentu x i y.

Przejdźmy do definicji dwuwymiarowej zmiennej losowej dyskretnej i ciągłej.

DEFINICJA

Definicja 25: Dwuwymiarowa zmienna losowa dyskretna

Zmienna losowa (X,Y) nazywana jest *dwuwymiarową zmienną losową skokową*, jeśli przyjmuje wartości w skończonym lub przeliczalnie nieskończonym zbiorze punktów (x_i,y_j) i istnieje funkcja prawdopodobieństwa $p(x,y)$ taka, że:

$$p(x_i,y_j)=P(X=x_i,\ Y=y_j). \tag{5.2}$$

Spełnia ona warunki: $p(x_i,y_j)\geq 0$ dla każdego i,j oraz $\sum_i\sum_j p(x_i,y_j)=1$

Dystrybuanta dwuwymiarowej zmiennej losowej dyskretnej określona jest wzorem

$$F(x,y)=P(X<x,\ Y<y)=\sum_{x_i<x}\sum_{y_j<y}p(x_i,y_j). \tag{5.3}$$

DEFINICJA

Definicja 26: Dwuwymiarowa zmienna losowa ciągła

Zmienna losowa (X,Y) nazywana jest *dwuwymiarową zmienną losową ciągłą*, jeśli istnieje funkcja gęstości prawdopodobieństwa $f(x,y)$ taka, że dla dowolnych liczb rzeczywistych $a<b$, $c<d$ zachodzi:

$$P(a<X<b,\ c<Y<d)=\int_a^b\int_c^d f(x,y)\,dy\,dx \tag{5.4}$$

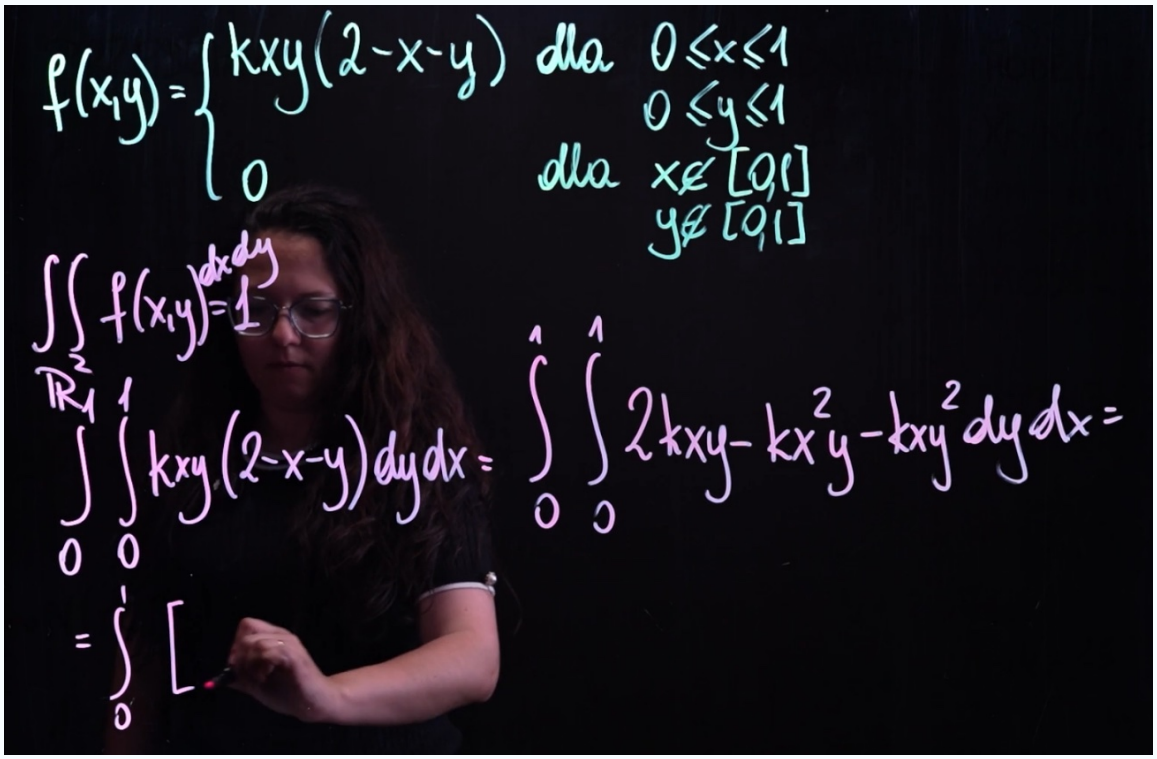
Funkcja $f(x,y)$ spełnia: $f(x,y)\geq 0$ dla każdego (x,y), oraz $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\,dy\,dx=1$

Dystrybuanta dwuwymiarowej zmiennej losowej ciągłej

$$F(x,y) = P(X < x,\ Y < y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,dv\,du \tag{5.5}$$

PRZYKŁAD

Przykład 30: Gęstość rozkładu dwuwymiarowego (zmienna ciągła)

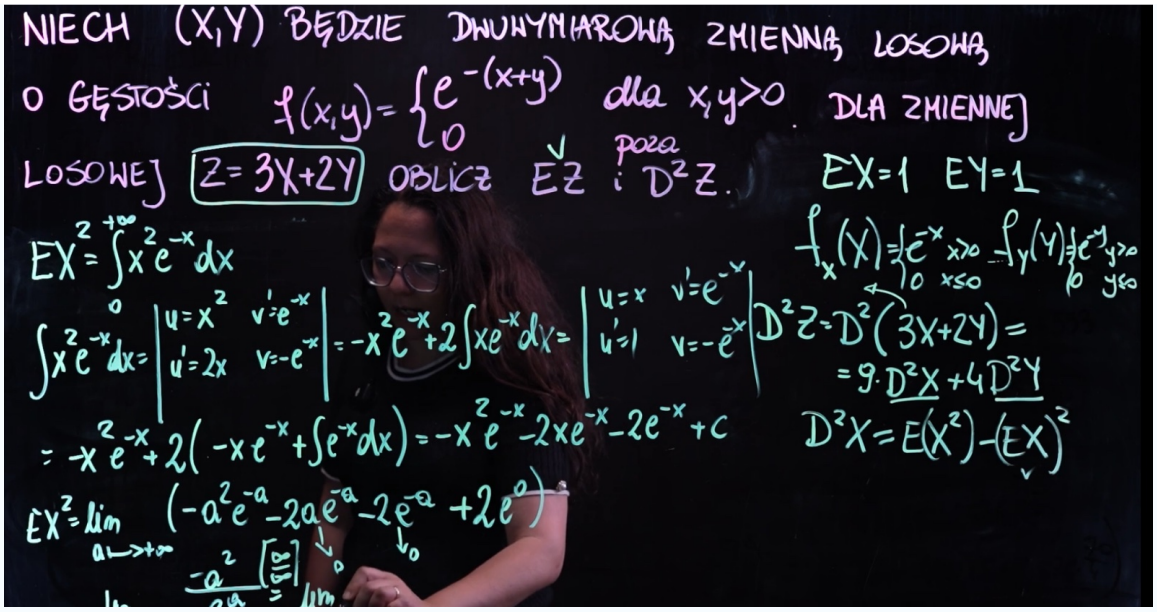


https://vimeo.com/1097814253/94246e0b27

Gęstość rozkładu dwuwymiarowego (zmienna ciągła)

PRZYKŁAD

Przykład 31: Funkcje dwuwymiarowej zmiennej losowej



https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2464/reader#openaghmod:2464:openaghhfivep:1

Funkcje dwuwymiarowej zmiennej losowej

5.2. Rozkład brzegowy i warunkowy

Autorzy/Autorki: Anna Serwatka

W analizie dwuwymiarowych zmiennych losowych często jesteśmy zainteresowani zachowaniem jednej ze zmiennych z pominięciem drugiej, lub też chcemy badać zachowanie jednej zmiennej pod warunkiem, że druga przyjęła konkretną wartość. W tym celu wprowadzamy dwa ważne pojęcia rozkładu brzegowego i rozkładu warunkowego.

Znając rozkład prawdopodobieństwa wektora (X, Y) możemy wyznaczyć rozkłady prawdopodobieństwa zmiennych X, Y. Nazywamy je rozkładami brzegowymi [6].

DEFINICJA

Definicja 27: Rozkład brzegowy (marginalny)

Niech dana będzie dwuwymiarowa zmienna losowa (X, Y). Rozkładem brzegowym nazywamy rozkład

- dla zmiennej losowej **dyskretnej**:

$$P_X(x_i) = \sum_j P(X = x_i,\ Y = y_j), \quad P_Y(y_j) = \sum_i P(X = x_i,\ Y = y_j) \tag{5.6}$$

- dla zmiennej losowej **ciągłej**:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \tag{5.7}$$

Rozkład warunkowy pozwala na badanie zależności między zmiennymi. Określa on prawdopodobieństwo jednej zmiennej pod warunkiem znajomości wartości drugiej [7].

DEFINICJA

Definicja 28: Rozkład warunkowy

Niech dana będzie dwuwymiarowa zmienna losowa (X,Y). Rozkładem brzegowym nazywamy rozkład

- dla zmiennej losowej **dyskretnej**:

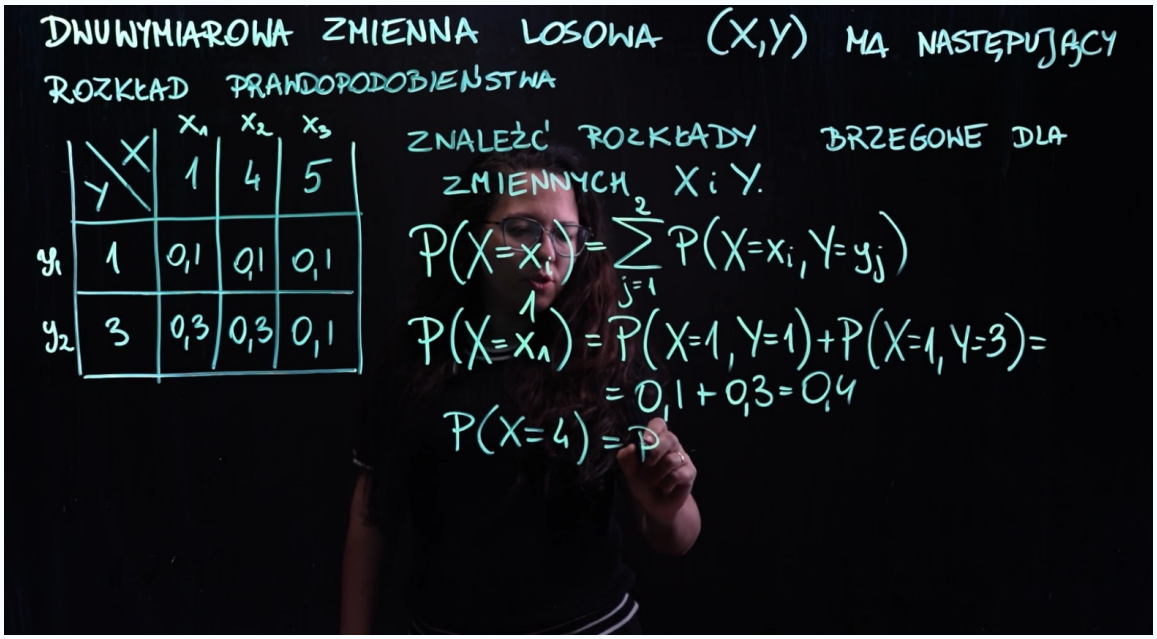
$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i,\ Y = y_j)}{P_Y(y_j)} \quad \text{dla } P_Y(y_j) > 0 \tag{5.8}$$

- dla zmiennej losowej **ciągłej**:

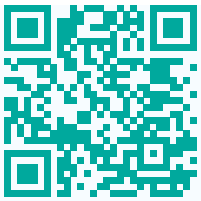
$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)} \quad \text{dla } f_Y(y) > 0 \tag{5.9}$$

PRZYKŁAD

Przykład 32: Rozkład brzegowy i warunkowy (dwuwymiarowa zmienna dyskretna)



Y \ X	x_1	x_2	x_3
y_1 1	0,1	0,1	0,1
y_2 3	0,3	0,3	0,1



https://vimeo.com/1097813890/91b87ee8f1
Rozkład brzegowy i warunkowy (dwuwymiarowa zmienna dyskretna)

5.3. Niezależność zmiennych losowych

Autorzy/Autorki: Anna Serwatka

Intuicyjnie, dwie zmienne losowe są niezależne, jeśli informacja o jednej z nich nie wpływa na wiedzę o drugiej — tzn. znajomość wartości jednej zmiennej nie zmienia rozkładu prawdopodobieństwa drugiej. W ujęciu formalnym niezależność wyraża się warunkiem na wspólny rozkład zmiennych — tak, aby ich wspólne prawdopodobieństwo (lub gęstość) rozkładało się na iloczyn rozkładów brzegowych [6].

TWIERDZENIE

Twierdzenie 10: Niezależność zmiennych losowych

Zmiennie losowe (X, Y) są niezależne wtedy i tylko wtedy, gdy

$$F_{(X,Y)}(x, y) = F_X(x) \cdot F_Y(y). \tag{5.10}$$

W przypadku zmiennych dyskretnych warunek ten równoważny jest warunkowi

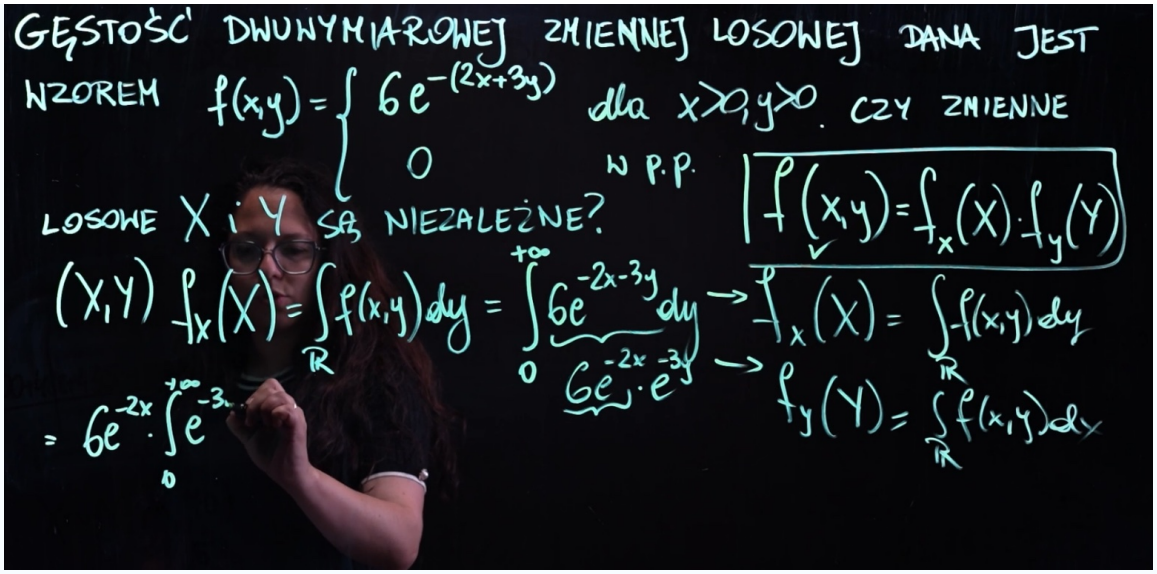
$$p_{ik} = p_i \cdot p_k \quad \text{dla wszystkich } i, k \tag{5.11}$$

a dla zmiennych typu ciągłego – warunkowi

$$f_{(X,Y)}(x, y) = f_X(x) f_Y(y) \quad \text{dla wszystkich } x, y \in \mathbb{R}. \tag{5.12}$$

PRZYKŁAD

Przykład 33: Niezależność zmiennych losowych - przypadek zmiennej ciągłej [7]



https://vimeo.com/1097814555/65212697e4

Niezależność zmiennych losowych

5.4. Parametry rozkładu wektorów losowych

Autorzy/Autorki: Anna Serwatka

W wielu sytuacjach nie interesuje nas ogólne zachowanie zmiennej losowej X, lecz jej zachowanie pod warunkiem, że inna zmienna Y przyjęła konkretną wartość. Takie podejście prowadzi do pojęcia warunkowej wartości oczekiwanej i pozwala lepiej opisać zależność między zmiennymi losowymi.

Warunkowa wartość oczekiwana to średnia wartość zmiennej X, obliczana przy założeniu, że zmienna Y ma ustaloną wartość. W przypadku rozkładu dyskretnego oblicza się ją jako sumę ważoną możliwych wartości X, a w przypadku rozkładu ciągłego — jako wartość oczekiwaną względem warunkowej gęstości $f(x \mid y)$ [6]

DEFINICJA

Definicja 29: Warunkowa wartość oczekiwana

Niech dana będzie dwuwymiarowa zmienna losowa (X, Y). Warunkowa wartość oczekiwana zmiennej X przy warunku Y wyraża się wzorem:

- dla zmiennej losowej **dyskretnej**:

$$\mathbb{E}(X \mid Y = y_k) = \sum_i x_i \cdot P(X = x_i \mid Y = y_k) \tag{5.13}$$

- dla zmiennej losowej **ciągłej**:

$$\mathbb{E}(X \mid Y = y_k) = \int_{-\infty}^{\infty} x \cdot f(x \mid y_k)\, dx \tag{5.14}$$

Jeśli chcemy przewidywać zmienną Y w oparciu o wartość zmiennej X, to interesuje nas tzw. funkcja regresji zmiennej Y względem X. Jest to funkcja przypisująca każdej wartości x wartość średnią zmiennej Y, przy założeniu, że $X = x$. Innymi słowy, funkcja regresji opisuje przeciętną (oczekiwaną) wartość Y dla danej wartości X.

DEFINICJA

Definicja 30: Funkcja regresji

Funkcją regresji zmiennej losowej Y względem zmiennej losowej X nazywamy funkcję:

$$m_Y(x) = \mathbb{E}(Y \mid X = x) \tag{5.15}$$

Funkcja regresji opisuje zależność wartości oczekiwanej zmiennej Y od zmiennej X. Wykres tej funkcji nazywany jest *krzywą regresji*.

LEMAT

Lemat 1:

Niech (X, Y) będzie wektorem losowym. Załóżmy, że istnieje $\mathbb{E}X$. Wtedy $\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}X$.

LEMAT

Lemat 2:

Niech (X, Y) będzie wektorem losowym, takim, że zmienne X i Y są niezależne. Załóżmy, że istnieje $\mathbb{E}X$. Wtedy $\mathbb{E}(X \mid Y) = \mathbb{E}X$.

5.5. Dwuwymiarowy rozkład normalny

Autorzy/Autorki: Anna Serwatka

Rozkład normalny odgrywa ważną rolę w statystyce i teorii prawdopodobieństwa. W przypadku dwóch zmiennych losowych jednocześnie podlegających analizie, naturalnym rozszerzeniem jednowymiarowego rozkładu normalnego jest **dwuwymiarowy rozkład normalny** [7]. Jest on używany do modelowania zjawisk, w których występuje współzależność pomiędzy dwiema zmiennymi losowymi o rozkładzie normalnym.

DEFINICJA

Definicja 31: Dwuwymiarowy rozkład normalny

Para zmiennych losowych (X, Y) ma *dwuwymiarowy rozkład normalny*, jeśli ich wspólna funkcja gęstości prawdopodobieństwa ma postać:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right) \tag{5.16}$$

gdzie:

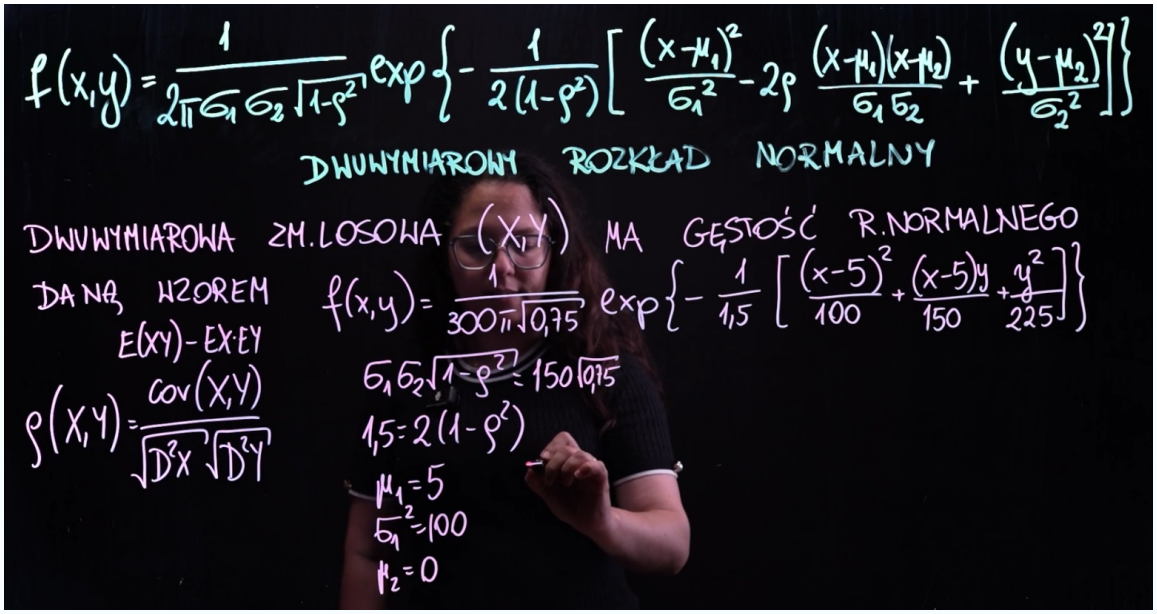
- μ_X, μ_Y — wartości oczekiwane zmiennych X i Y,
- σ_X, σ_Y — odchylenia standardowe zmiennych X i Y,
- ρ — współczynnik korelacji między X i Y, $\rho \in (-1, 1)$,
- $f(x, y)$ — gęstość prawdopodobieństwa w punkcie (x, y).

Rozkład ten oznaczamy symbolem:

$$(X,Y) \sim \mathcal{N}_2\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}\right) \tag{5.17}$$

PRZYKŁAD

Przykład 34: Dwuwymiarowy rozkład normalny



https://vimeo.com/1097818297/ae99b17287

Dwuwymiarowy rozkład normalny

Rozdział 6

Elementy statystyki opisowej

6.1. Średnie, mediana i moda

Autorzy/Autorki: Anna Serwatka

Średnia arytmetyczna, dominanta i mediana to trzy miary tendencji centralnej, z których każda w inny sposób określa punkt "środkowy" zestawu danych [6].

DEFINICJA

Definicja 32: Średnia arytmetyczna

Średnia arytmetyczna to suma wszystkich wartości w zbiorze danych podzielona przez liczbę tych wartości. Dla zbioru liczb $x_1, x_2, \ldots, x_n$ średnia arytmetyczna jest określona wzorem:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{6.1}$$

Zauważmy, że średnia arytmetyczna jest użyteczna do określania przeciętnej wartości w danych, ale jest wrażliwa na wartości ekstremalne, które mogą ją znacząco zniekształcić, co pokazuje następujący przykład:

PRZYKŁAD

Przykład 35:

W pewnej firmie, w której zatrudnionych jest 5 pracowników odnotowano następujące miesięczne zarobki: 4000 zł, 4000 zł, 4000 zł, 4000 zł, 20 000 zł. Jakie są średnie zarobki w tej firmie?

Korzystając ze wzoru 1 otrzymujemy: $\overline{x} = \frac{4000+4000+4000+4000+20000}{5} = 7200$.

Ta wartość nie oddaje struktury danych, gdzie cztery osoby zarabiały poniżej średniej, a tylko jedna osoba zarabiała dużo więcej od pozostałych.

DEFINICJA

Definicja 33: Moda (dominanta)

Moda (dominanta) to wartość, która najczęściej występuje w zbiorze danych.

PRZYKŁAD

Przykład 36:

Oblicz modę liczb:

1. $1, 2, 2, 2, 3, 4, 5, 6, 6, 7$
2. $-1, -1, -1, -2, -2, -2, -3, -4, -5$
3. $1, 2, 3, 4, 5, 6, 7, 8, 9$

1. Najczęściej występuje liczba 2 dlatego moda jest równa 2.
2. Każda z liczb -1 i -2 występuje 3 razy i żadna liczba nie występuje częściej. Mamy tutaj zatem dwie mody: -1 i -2.
3. W tym przypadku nie ma mody, bo każda wartość występuje taką samą liczbę razy.

DEFINICJA

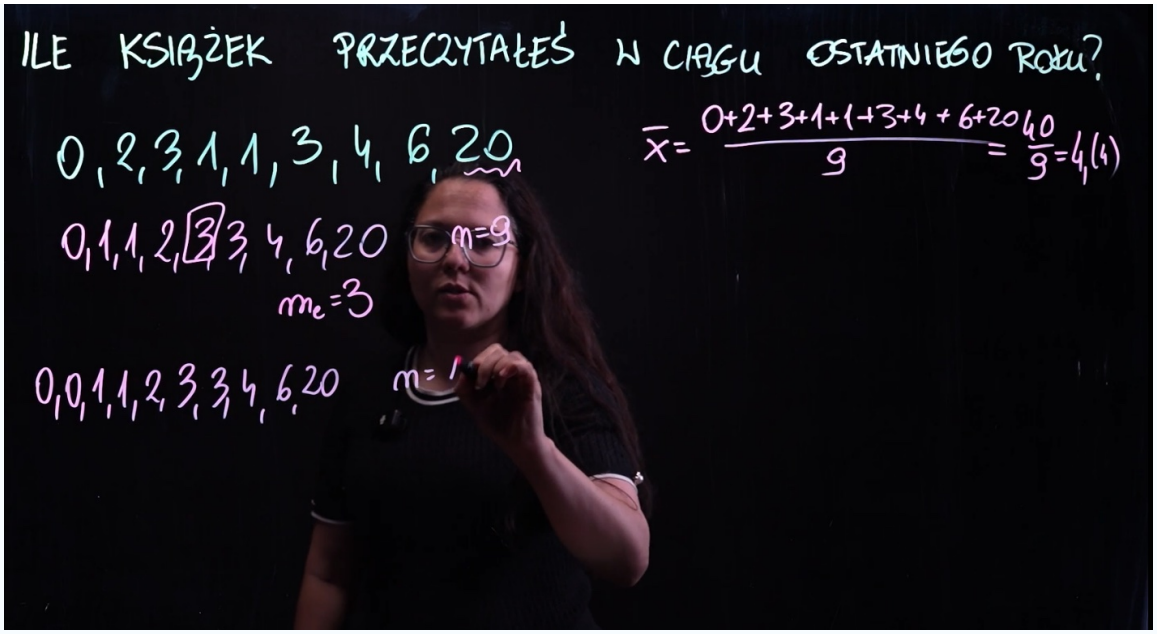
Definicja 34: Mediana

Mediana to wartość środkowa w uporządkowanym (rosnąco lub malejąco) zbiorze danych.

Jeżeli mamy nieparzystą liczbę danych, to istnieje dokładnie jedna wartość środkowa. Jeżeli mamy parzystą liczbę danych, to mediana jest równa średniej arytmetycznej dwóch środkowych wartości. Należy pamiętać, że licząc medianę musimy najpierw uporządkować zbiór od najmniejszej do największej liczby lub odwrotnie.

PRZYKŁAD

Przykład 37: Mediana



https://vimeo.com/1097822864/178974ba70

Mediana

Oprócz średniej arytmetycznej, istnieją inne rodzaje średnich, które mogą być bardziej odpowiednie w zależności od charakteru danych. Poniżej przedstawimy definicję średniej geometrycznej, harmonicznej i kwadratowej. Średnia geometryczna jest używana głównie do analizy wzrostu proporcjonalnego, np. stóp zwrotu inwestycji czy tempa wzrostu populacji.

DEFINICJA

Definicja 35: Średnia geometryczna

Niech $x_i > 0$ dla każdego $i = 1, \ldots n$. Dla zbioru liczb $x_1, x_2, \ldots, x_n$ średnia geometryczna jest określona wzorem:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \tag{6.2}$$

Średnia harmoniczna jest stosowana w przypadku średnich prędkości, wskaźników czy wydajności.

DEFINICJA

Definicja 36: Średnia harmoniczna

Niech $x_i > 0$ dla każdego $i = 1, \ldots n$. Dla zbioru liczb $x_1, x_2, \ldots, x_n$ średnia harmoniczna jest

określona wzorem:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{6.3}$$

Średnia kwadratowa jest często stosowana w statystyce, fizyce i inżynierii, np. w analizie napięć elektrycznych czy sygnałów dźwiękowych.

DEFINICJA

Definicja 37: Średnia kwadratowa

Średnia kwadratowa n liczb, jest to pierwiastek ze średniej arytmetycznej kwadratów tych liczb. Dla zbioru liczb $x_1, x_2 \ldots, x_n$ średnia kwadratowa jest określona wzorem:

$$K = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}} \tag{6.4}$$

Średnia ważona jest szeroko stosowana w różnych dziedzinach, ponieważ pozwala uwzględnić różne znaczenie (wagę) poszczególnych wartości. Na przykład indeksy giełdowe są liczone na podstawie średniej ważonej cen akcji spółek, a także oceny w szkołach i na uczelniach wystawiane są na koniec semestru z wykorzystaniem średniej ważonej.

DEFINICJA

Definicja 38: Średnia ważona

Dla zbioru liczb $x_1, x_2, \ldots, x_n$, z których każda ma przyporządkowaną pewną nieujemną wagę $w_1, w_2, \ldots, w_n$ średnia ważona jest określona wzorem:

$$W = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n} \tag{6.5}$$

6.2. Miary zmienności: odchylenie standardowe, wariancja

Autorzy/Autorki: Anna Serwatka

Miary zmienności to wskaźniki statystyczne opisujące rozproszenie danych wokół wartości centralnej (np. średniej). Do najczęściej używanych miar zmienności należą odchylenie standardowe i wariancja. Wariancja jest średnią arytmetyczną kwadratów odchyleń poszczególnych wartości cechy jednostek zbiorowości od ich wartości średniej (arytmetycznej) – kwadrat odchylenia standardowego.[8].

DEFINICJA

Definicja 39: Wariancja

Dla zbioru liczb $x_1, x_2, \ldots, x_n$ wariancja jest określona wzorem:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{6.6}$$

gdzie $\overline{x}$ oznacza średnią arytmetyczną liczb $x_1, x_2, \ldots, x_n$.

Wariancja pokazuje, jak bardzo wartości są rozproszone wokół średniej. Im większa wariancja, tym większe rozproszenie danych. Wartości w wariancji są podniesione do kwadratu, co utrudnia interpretację w jednostkach oryginalnych i nie ma interpretacji ekonomicznej. Stąd wprowadzamy kolejną miarę zmienności jaką jest odchylenie standardowe. Odchylenie standardowe jest pierwiastkiem kwadratowym z wariancji. Intuicyjnie rzecz ujmując, odchylenie standardowe informuje o tym, jak szeroko wartości danej wielkości są rozrzucone wokół jej średniej. Im mniejsza wartość odchylenia tym obserwacje są bardziej skupione wokół średniej[9].

DEFINICJA

Definicja 40: Odchylenie standardowe

Dla zbioru liczb $x_1, x_2, \ldots, x_n$ odchylenie standardowe jest określone wzorem:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{6.7}$$

gdzie $\overline{x}$ oznacza średnią arytmetyczną liczb $x_1, x_2, \ldots, x_n$.

Wprowadzimy jeszcze jedną miarę zmienności, a mianowicie odchylenie przeciętne, które jest średnią arytmetyczną bezwzględnych wartości odchyleń zmiennej od wartości jej średniej arytmetycznej.

DEFINICJA

Definicja 41: Odchylenie przeciętne

Dla zbioru liczb $x_1, x_2, \ldots, x_n$ odchylenie przeciętne jest określona wzorem:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \tag{6.8}$$

gdzie $\overline{x}$ oznacza średnią arytmetyczną liczb $x_1, x_2, \ldots, x_n$.

Rozdział 7

Przedziały ufności

7.1. Pojęcia wstępne: populacja, próba

Autorzy/Autorki: Anna Serwatka

Statystyka zajmuje się opisywaniem i analizowaniem zjawisk. Aby móc prowadzić analizę statystyczną, trzeba najpierw określić, co jest przedmiotem badania. W tym celu wprowadzamy podstawowe pojęcia, takie jak populacja, jednostka statystyczna i próba [10]. Są to fundamenty każdego badania statystycznego – niezależnie od dziedziny.

DEFINICJA

Definicja 42: Populacja

Populacja ogół elementów, które są przedmiotem badania statystycznego i posiadają określoną cechę lub zespół cech.

DEFINICJA

Definicja 43: Próba statystyczna

Próbą statystyczną nazywamy wybrany podzbiór populacji, który poddawany jest bezpośredniej analizie. Na jej podstawie wnioskujemy o całej populacji.

DEFINICJA

Definicja 44: Jednostka statystyczna

Jednostka statystyczna to pojedynczy element populacji, który podlega obserwacji.

ZADANIE

Zadanie 12:

Treść zadania:

Biblioteka miejska planuje przeprowadzić badanie dotyczące liczby przeczytanych książek wśród mieszkańców miasta powyżej 15. roku życia. W tym celu zamierza zebrać dane od wybranej grupy osób. Na podstawie opisu:

1. Wskaż, co jest populacją, jednostką statystyczną, próbą.
2. Czy poniższe metody wyboru prób są losowe? Uzasadnij.
   (a) Wybrano pierwsze 100 osób, które odwiedziły bibliotekę w poniedziałek.
   (b) Wylosowano 200 osób z listy mieszkańców posiadających Kartę Mieszkańca.

Rozwiązanie:

Populacja to wszyscy mieszkańcy miasta w wieku powyżej 15 lat. Jednostka statystyczna to jeden mieszkaniec miasta. Natomiast próba to grupa osób, od których zostaną zebrane dane (np. 100 lub 200 osób).

Pierwszy sposób wyboru 100 osób nie spełnia kryterium losowości. Ta próba nie jest losowa, ponieważ wybór osób zależy od dnia i miejsca, w którym były dostępne (biblioteka). Oznacza to, że nie każdy mieszkaniec miasta miał jednakową szansę znalezienia się w próbie. Osoby odwiedzające bibliotekę mogą różnić się np. pod względem wieku lub zainteresowań od reszty populacji, co może zniekształcić wyniki badania. Natomiast druga próba ma charakter losowy, ponieważ osoby zostały wybrane poprzez losowanie z listy mieszkańców. Każdy, kto posiada Kartę Mieszkańca, miał taką samą szansę znalezienia się w próbie, co oznacza, że wyniki badania będą bardziej reprezentatywne i obiektywne w stosunku do całej populacji.

7.2. Estymacja punktowa i przedziałowa

Autorzy/Autorki: Anna Serwatka

Często nie mamy dostępu do pełnych danych dla całej populacji. Zamiast tego dysponujemy jedynie danymi z próby. Na ich podstawie chcemy oszacować nieznane parametry populacji, takie jak średnia, odchylenie standardowe czy odsetek. Proces ten nazywamy estymacją. Wyróżniamy estymację punktową i przedziałową.

DEFINICJA

Definicja 45: Estymacja punktowa

Estymacja punktowa polega na podaniu jednej konkretnej liczby jako przybliżenia wartości parametru populacyjnego.

Przykładem estymatora jest średnia arytmetyczna z próby [10], będąca estymatorem punktowym średniej w

populacji. Nie każdy estymator parametru populacji jest równie dobry. Aby móc porównać różne estymatory i wybrać ten najbardziej odpowiedni, statystyka matematyczna wprowadza kilka kluczowych cech, które powinien spełniać tzw. „dobry estymator".

DEFINICJA

Definicja 46: Estymator nieobciążony

Estymator $\hat{\theta}$ parametru θ nazywamy *nieobciążonym*, jeśli spełnia warunek:

$$\mathbb{E}(\hat{\theta}) = \theta \tag{7.1}$$

Cechy estymatorów dotyczą trafności, stabilności oraz wykorzystania informacji zawartej w danych. W tabeli poniżej zestawiono najważniejsze z nich wraz z symbolicznymi zapisami.

Tabela 1: Opis tabeli

Cecha estymatora	Opis	Symbolicznie
Nieobciążoność	Estymator średnio trafia w prawdziwą wartość parametru.	$\mathbb{E}(\hat{\theta}) = \theta$
Zgodność	Estymator zbliża się do prawdziwego parametru wraz ze wzrostem liczby obserwacji.	$\lim_{n\to\infty}\hat{\theta}_n = \theta$
Efektywność	Estymator ma najmniejszą możliwą wariancję spośród estymatorów nieobciążonych.	D^2X minimalna
Mała zmienność	Estymator ma niewielką wariancję — jest stabilny.	D^2X mała
Wystarczalność (opcjonalnie)	Estymator zawiera całą informację o parametrze dostępną w próbie.	–

Tabela 1: Cechy dobrego estymatora

DEFINICJA

Definicja 47: Estymacja przedziałowa

Estymacja przedziałowa polega na podaniu przedziału liczbowego, w którym – z określonym poziomem ufności (np. 95%) – znajduje się szacowany parametr populacji.

Przy estymacji przedziałowej nie mówimy, że z 95% prawdopodobieństwem parametr jest w tym konkretnym przedziale. Prawidłowa interpretacja dotyczy procedury losowania: 95% tak skonstruowanych przedziałów zawiera prawdziwy parametr.

7.3. Przedział ufności dla średniej

Autorzy/Autorki: Anna Serwatka

W tym rozdziale przechodzimy do obliczania przedziałów ufności. Przedział ufności to przedział, w którym z określoną pewnością możemy spodziewać się prawdziwej wartości parametru.

DEFINICJA

Definicja 48: Przedział ufności

Przedziałem ufności nazywamy przedział wyznaczony za pomocą estymatora, mający tę własność, że z dużym, z góry zadanym prawdopodobieństwem $(1 - \alpha)$, pokrywa wartość szacowanego parametru θ. Zachodzi równość:

$$P(a < \theta < b) = 1 - \alpha, \tag{7.2}$$

gdzie a i b noszą nazwę dolnej i górnej granicy przedziału ufności, a prawdopodobieństwo $1 - \alpha$ jest z góry określone.

Jest to narzędzie, które można wykorzystywać do tego, aby na podstawie próby oszacować parametry populacji. Aby rozwiązać zadanie z wyznaczaniem przedziału ufności potrzebujemy:

1. próby losowej, czyli danych, na podstawie których dokonujemy estymacji;
2. statystyki próbkowej, na przykład średniej arytmetycznej lub odchylenia standardowego;
3. rozkładu statystyki próbkowej (zazwyczaj będzie to rozkład normalny lub t-Studenta);
4. krytycznej wartości z rozkładu statystycznego, którą będziemy odczytywać z tablic;
5. poziomu ufności $1 - \alpha$, który oznacza z jakim prawdopodobieństwem parametr jest pokryty przedziałem ufności. Najczęściej wybierane są wartości 0,9, 0,95, 0,99 czy 0,999.

Jednym z najczęściej szacowanych parametrów populacji generalnej jest średnia wartość badanej cechy. Przedziały ufności dla poszczególnych parametrów populacji wyznacza się używając rozkładów odpowiednich statystyk, które są estymatorami tych parametrów. W tym rozdziale analizujemy przedziały ufności dla wartości średniej, a najlepszym uzyskanym metodą największej wiarygodności, estymatorem średniej wartości m populacji generalnej jest średnia arytmetyczna $\overline{x}$ obliczona z próby. Ten estymator jest zgodny, nieobciążony, efektywny i dostateczny. W zależności od modelu, otrzymuje się konkretne wzory na przedział ufności dla średniej, w oparciu o rozkład normalny lub rozkład t-Studenta.[10]

MODEL

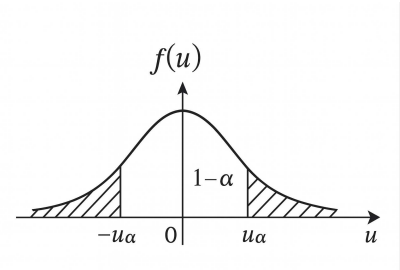
Model 1: Populacja generalna ma rozkład normalny o nieznanej średniej m i znanym odchyleniu standardowym σ

Załóżmy, że mamy daną populację generalną o rozkładzie $N(m, \sigma)$. Znamy odchylenie standardowe tego rozkładu σ i losujemy próbę n elementów z tej populacji w sposób niezależny. Przedział ufności dla średniej m populacji dany jest wzorem:

$$P(\overline{x} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}} < m < \overline{x} + u_\alpha \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha, \tag{7.3}$$

gdzie $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ oznacza średnią arytmetyczną obliczoną z próby, $1 - \alpha$ jest współczynnikiem ufności. Załóżmy, że zmienna losowa U, ma standaryzowany rozkład normalny $U \sim N(0, 1)$. Wtedy u_α jest wartością zmiennej losowej U, wyznaczoną z równania

$$P(-u_\alpha < U < u_\alpha) = 1 - \alpha \tag{7.4}$$



Rysunek 2: Rozkład normalny standaryzowany, źródło: opracowanie własne

MODEL

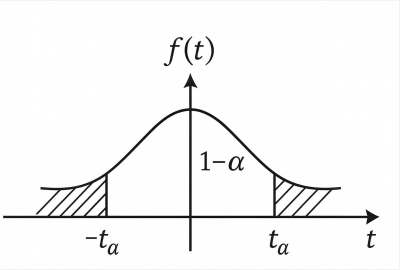
Model 2: Populacja generalna ma rozkład normalny o nieznanej średniej m i nieznanym odchyleniu standardowym σ

Załóżmy, że mamy daną populację generalną o rozkładzie $N(m, \sigma)$. W tym przypadku nie znamy ani odchylenia standardowe tego rozkładu ani wartości średniej. Losujemy próbę n elementów z tej populacji w sposób niezależny. Przedział ufności dla średniej m populacji dany jest wzorem:

$$P(\overline{x} - t_\alpha \cdot \frac{s}{\sqrt{n-1}} < m < \overline{x} + t_\alpha \cdot \frac{s}{\sqrt{n-1}}) = 1 - \alpha, \tag{7.5}$$

gdzie $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ oznacza średnią arytmetyczną obliczoną z próby, $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ jest odchyleniem standardowym z próby, a $1 - \alpha$ jest współczynnikiem ufności. Załóżmy, że zmienna losowa t, ma rozkład t-Studenta. Wtedy t_α jest wartością zmiennej losowej t odczytaną z tablicy tego rozkładu dla $n - 1$ stopni swobody, tak aby zachodziła równość:

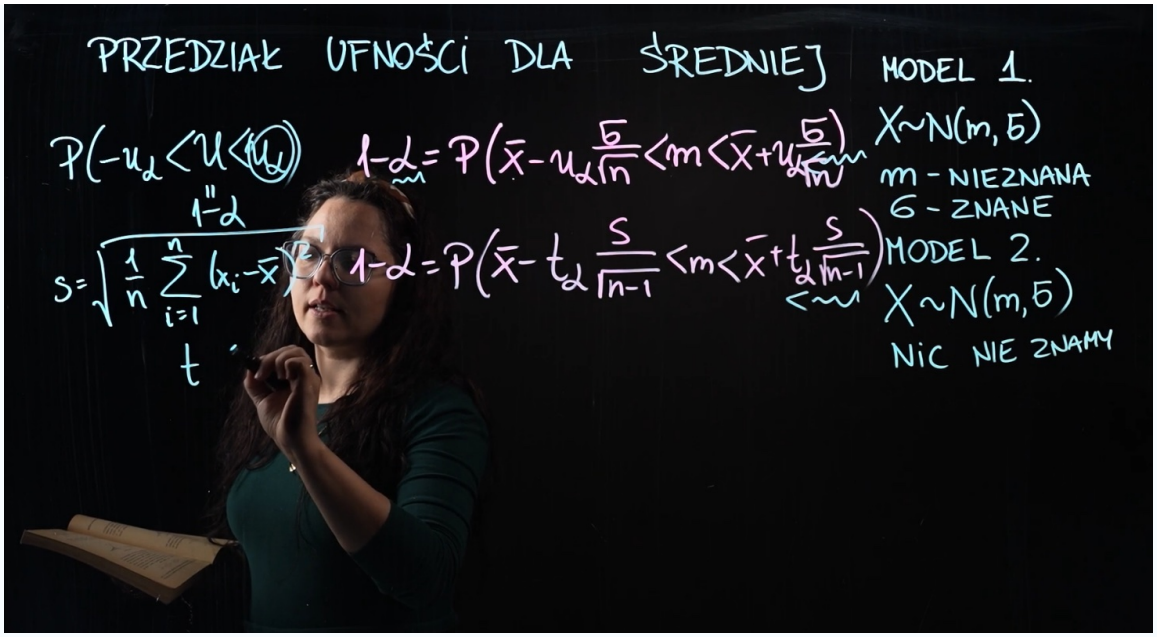
$$P(-t_\alpha < t < t_\alpha) = 1 - \alpha \tag{7.6}$$



Rysunek 3: Rozkład t-Studenta, źródło: opracowanie własne

PRZYKŁAD

Przykład 38: Przedział ufności dla średniej z nieznanym odchyleniem standardowym

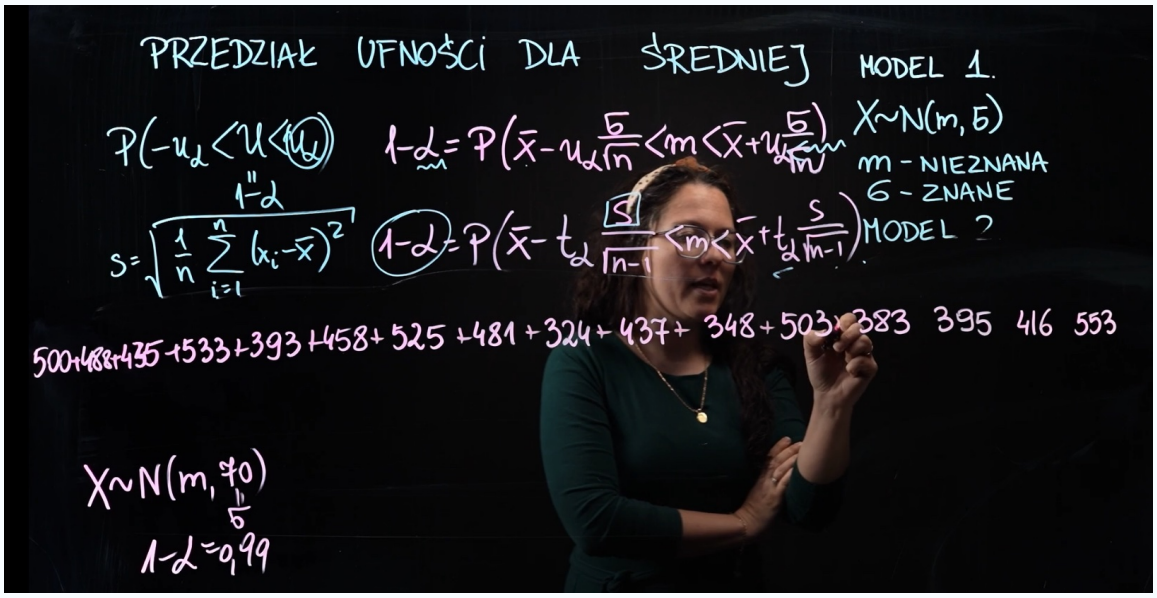


https://vimeo.com/1097828195/1f9df16075

Przedział ufności dla średniej z nieznanym odchyleniem standardowym

PRZYKŁAD

Przykład 39: Przedział ufności dla średniej ze znanym odchyleniem standardowym



https://vimeo.com/1097813522/21162dbe39
Przedział ufności dla średniej ze znanym odchyleniem standardowym

7.4. Przedział ufności dla wariancji

Autorzy/Autorki: Anna Serwatka

Przedział ufności dla wariancji stosujemy wtedy, gdy chcemy z określoną pewnością oszacować, jak bardzo zróżnicowane są dane w populacji — na przykład ocenić stabilność pomiarów, rozrzut wyników testów czy zmienność produkcji. Jest to szczególnie przydatne w analizie jakości, w badaniach społecznych i edukacyjnych, a także w finansach, gdzie interesuje nas ryzyko i odchylenia od przeciętnych wartości [10]. Tego typu przedziały pomagają określić, czy obserwowane różnice są naturalne, czy też wymagają dalszej analizy lub interwencji. Gdy populacja ma rozkład normalny $N(\mu, \sigma)$ (lub zbliżony do normalnego), wtedy można zbudować przedział ufności dla wariancji σ^2 populacji. Najczęściej używanym estymatorem wariancji σ^2 populacji generalnej są dwie statystyki:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{7.7}$$

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{7.8}$$

Estymator $\hat{s}^2$ 7.8 jest nieobciążonym estymatorem wariancji σ^2, natomiast estymator s^2 7.7 jest obciążonym estymatorem wariancji σ^2. Oba te estymatory są obciążonymi estymatorami odchylenia standardowego σ. W zależności od liczebności próby i tego, czy znamy parametry populacji, stosujemy różne modele przedziałów ufności dla wariancji. Gdy badamy zmienność cechy i zakładamy normalność rozkładu, możemy skonstruować przedział ufności dla wariancji z wykorzystaniem rozkładu chi-kwadrat. Co ważne, jeśli obliczymy pierwiastek kwadratowy z obu końców przedziału ufności dla wariancji, uzyskamy przedział ufności dla odchylenia standardowego, który również bywa wykorzystywany w analizie danych do oceny stabilności lub rozproszenia.

MODEL

Model 3: Populacja generalna ma rozkład normalny o nieznanej średniej μ i nieznanym odchyleniu standardowym σ. Model dla małej próby $n < 30$

Załóżmy, że mamy daną populację generalną o rozkładzie $N(\mu, \sigma)$. Nie znamy odchylenia standardowego tego rozkładu i nie znamy wartości oczekiwanej. Losujemy próbę n elementów z tej populacji w sposób niezależny (n jest małe takie, że $n < 30$). Przedział ufności dla wariancji σ^2

populacji dany jest wzorem:

$$P\left(\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{ns^2}{\chi^2_{\frac{\alpha}{2}}}\right) = 1-\alpha, \tag{7.9}$$

gdzie $1-\alpha$ jest współczynnikiem ufności, a $\chi^2_{\frac{\alpha}{2}}$ i $\chi^2_{1-\frac{\alpha}{2}}$ to odpowiednie kwantyle rozkładu chi-kwadrat z $n-1$ stopniami swobody, spełniające następujące warunki:

$$P(\chi^2 < \chi^2_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \tag{7.10}$$

$$P(\chi^2 < \chi^2_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2} \tag{7.11}$$

Aby wyznaczyć przedział ufności dla odchylenia standardowego σ, obliczamy pierwiastek kwadratowy z końców przedziału ufności dla wariancji σ^2 :

$$P\left(\sqrt{\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}}} < \sigma < \sqrt{\frac{ns^2}{\chi^2_{\frac{\alpha}{2}}}}\right) = 1-\alpha, \tag{7.12}$$

MODEL

Model 4: Populacja generalna ma rozkład normalny o nieznanej średniej μ i nieznanym odchyleniu standardowym σ. Model dla dużej próby $n \geq 30$

Załóżmy, że mamy daną populację generalną o rozkładzie $N(\mu, \sigma)$. Nie znamy odchylenia standardowego tego rozkładu i nie znamy wartości oczekiwanej. Losujemy próbę n elementów z tej populacji w sposób niezależny (n jest duże takie, że $n \geq 30$). Obliczamy z tej próby odchylenie standardowe $s = \sqrt{s^2}$. Przedział ufności dla odchylenia standardowego σ populacji dany jest wzorem:

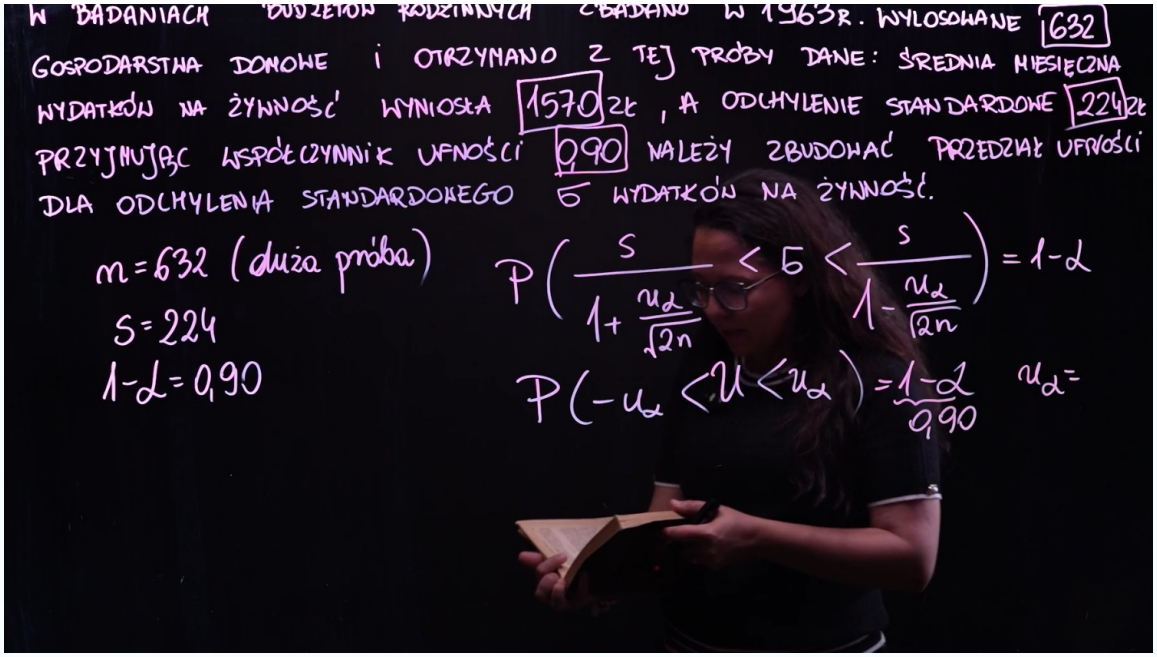
$$P\left(\frac{s}{1+\frac{u_\alpha}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{u_\alpha}{\sqrt{2n}}}\right) = 1-\alpha, \tag{7.13}$$

gdzie $1-\alpha$ jest współczynnikiem ufności. Załóżmy, że zmienna losowa U, ma standaryzowany rozkład normalny $U \sim N(0,1)$. Wtedy u_α jest wartością zmiennej losowej U, spełniającą warunek

$$P(-u_\alpha < U < u_\alpha) = 1-\alpha. \tag{7.14}$$

PRZYKŁAD

Przykład 40: Przedział ufności dla odchylenia standardowego



https://vimeo.com/1097823568/b6d2cc0d2c

Przedział ufności dla odchylenia standardowego

7.5. Przedział ufności dla wskaźnika struktury

Autorzy/Autorki: Anna Serwatka

Wskaźnik struktury (inaczej: udział strukturalny) to miara statystyczna wyrażająca, jaką część całości stanowi dana część zbiorowości. Oblicza się go jako stosunek liczby jednostek posiadających określoną cechę do liczby wszystkich jednostek w populacji lub próbie. Wskaźnik ten pozwala opisać strukturę wewnętrzną badanej zbiorowości, czyli proporcje między poszczególnymi jej elementami. Najczęściej wyraża się go w postaci ułamka dziesiętnego, procenta lub promila.

Aby zbudować przedział ufności dla wskaźnika struktury p w populacji dla dużej próby n, korzystamy z własności, że estymatory uzyskane metodą największej wiarygodności mają rozkład asymptotycznie normalny.

MODEL

Model 5: Populacja generalna ma rozkład dwupunktowy z parametrem p. Model dla licznej próby $n > 100$

Załóżmy, że mamy daną populację generalną o rozkładzie dwupunktowym z parametrem p. Losujemy próbę n elementów z tej populacji w sposób niezależny (n jest duże takie, że $n > 100$). Niech m oznacza liczbę elementów wyróżnionych znalezionych w próbie. Przedział ufności dla wskaźnika

struktury p populacji dany jest wzorem:

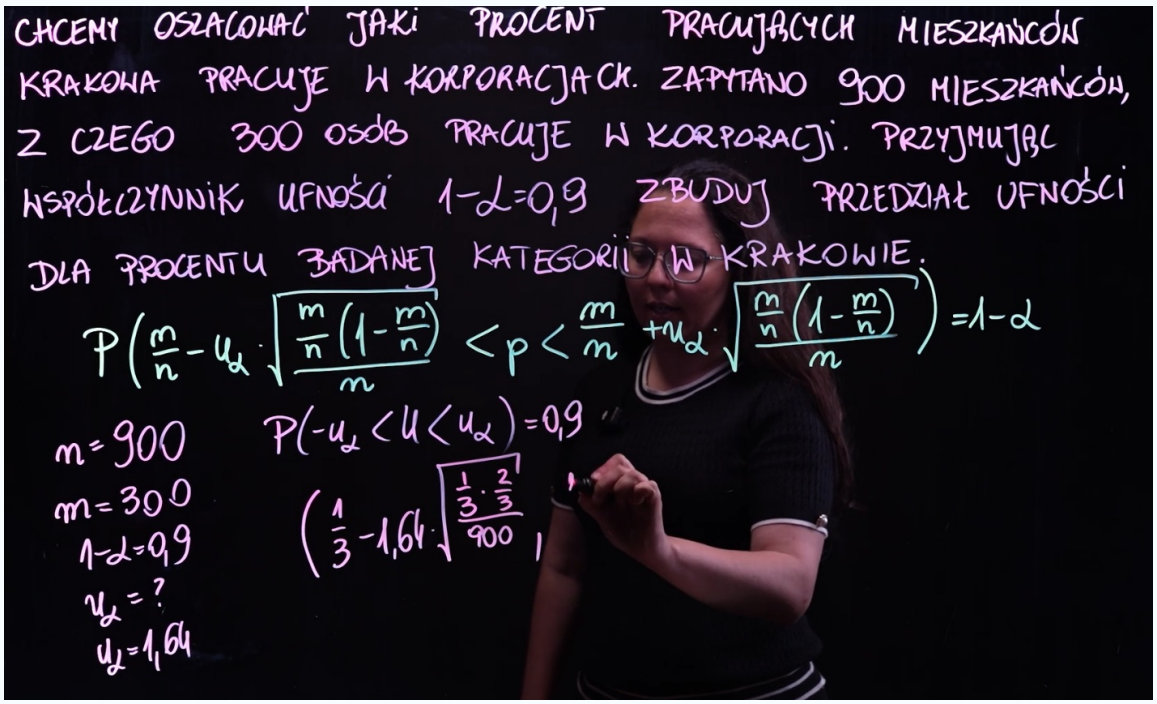
$$P\left(\frac{m}{n} - u_\alpha \cdot \sqrt{\frac{\frac{m}{n}(1-\frac{m}{n})}{n}} < p < \frac{m}{n} + u_\alpha \cdot \sqrt{\frac{\frac{m}{n}(1-\frac{m}{n})}{n}}\right) = 1-\alpha, \tag{7.15}$$

gdzie $1-\alpha$ jest współczynnikiem ufności. Załóżmy, że zmienna losowa U, ma standaryzowany rozkład normalny $U \sim N(0,1)$. Wtedy u_α jest wartością zmiennej losowej U, spełniającą warunek

$$P(-u_\alpha < U < u_\alpha) = 1-\alpha. \tag{7.16}$$

PRZYKŁAD

Przykład 41: Przedział ufności dla wskaźnika struktury/procentu



https://vimeo.com/1097825284/b6e7152bd1

Przedział ufności dla wskaźnika struktury/procentu

7.6. Niezbędna liczba pomiarów dla próby

Autorzy/Autorki: Anna Serwatka

Minimalna liczebność próby to najmniejsza liczba obserwacji, jaką należy zebrać, aby z zadaną precyzją

(błędem maksymalnym) i poziomem ufności oszacować interesujący nas parametr populacji, np. średnią lub proporcję (wskaźnik struktury). Zanim przejdziemy do konkretnych modeli, warto zrozumieć, jakie czynniki wpływają na wymaganą wielkość próby:

1. **Poziom ufności:** Określa prawdopodobieństwo, że przedział ufności, który zbudujemy na podstawie naszej próby, będzie zawierał prawdziwą wartość parametru populacji. Najczęściej stosowane poziomy to 90%, 95% lub 99%. Wyższy poziom ufności wymaga większej próby.
2. **Błąd maksymalny:** To maksymalna dopuszczalna różnica między oszacowaną wartością z próby a prawdziwą wartością w populacji. Im mniejszy błąd chcemy osiągnąć (większa precyzja), tym większa próba jest potrzebna.
3. **Wariancja:** Im większe zróżnicowanie (rozproszenie) danych w populacji, tym większa próba będzie potrzebna, aby uzyskać precyzyjne oszacowania.
4. **Wielkość populacji:** W przypadku bardzo dużych populacji (powyżej około 100 000), wielkość populacji ma mniejsze znaczenie dla minimalnej liczebności próby. Jest jednak ważna w przypadku mniejszych populacji (poniżej 10 000).

Istnieją różne modele w zależności od tego, jaki parametr populacji chcemy oszacować (np. średnią czy proporcję) oraz czy znamy wariancję populacji. Gdy chcemy oszacować średnią wartość jakiejś zmiennej ilościowej w populacji (np. średni wzrost, średnie zarobki) stosujemy następujący model [10]:

MODEL

Model 6: Model szacowania średniej populacji

Minimalną liczebność próby dla oszacowania średniej wyliczamy ze wzoru

$$n = \left(\frac{Z \cdot \sigma}{E} \right)^2, \tag{7.17}$$

gdzie:

- n - minimalna liczebność próby
- Z - wartość krytyczna odpowiadająca wybranemu poziomowi ufności (np. dla 95% ufności, $Z \approx 1,96$; dla 99% ufności, $Z \approx 2,58$)
- σ - odchylenie standardowe populacji (jeśli nie znamy, możemy użyć odchylenia standardowego z badania pilotażowego lub wstępnych szacunków)
- E - dopuszczalny błąd maksymalny (precyzja)

PRZYKŁAD

Przykład 42:

Chcemy oszacować średni tygodniowy czas spędzany na nauce przez studentów w Polsce. Chcemy to zrobić z 95% poziomem ufności ($Z = 1,96$). Chcemy, aby nasz błąd maksymalny nie przekroczył 0.5 godziny ($E = 0,5$). Na podstawie wcześniejszych badań, szacujemy, że odchylenie standardowe czasu

nauki w populacji studentów wynosi 2 godziny ($\sigma = 2$).

Podstawiamy do wzoru: $n = \left(\frac{1,96 \cdot 2}{0.5}\right)^2 = \left(\frac{3,92}{0.5}\right)^2 = (7,84)^2 \approx 61,46$. Musimy zaokrąglić w górę, ponieważ liczba obserwacji musi być liczbą całkowitą. Zatem minimalna liczebność próby to 62 studentów.

Gdy chcemy oszacować proporcję (udział procentowy) cechy w populacji (np. odsetek osób popierających daną partię, odsetek produktów wadliwych) stosujemy następujący model:

MODEL

Model 7: Model szacowania proporcji populacji

Minimalną liczebność próby dla proporcji populacji wyliczamy ze wzoru

$$n = \frac{Z^2 \cdot p(1-p)}{E^2}, \tag{7.18}$$

gdzie:

- n - minimalna liczebność próby
- Z - -wartość krytyczna odpowiadająca wybranemu poziomowi ufności (np. dla 95% ufności, $Z \approx 1,96$)
- p - szacowana proporcja sukcesów w populacji (jeśli nie mamy żadnych wcześniejszych danych, najbezpieczniej jest użyć $p = 0.5$, ponieważ ta wartość maksymalizuje $p(1-p)$ i daje największą wymaganą próbę, co zabezpiecza nas przed niedoszacowaniem)
- E - dopuszczalny błąd maksymalny (precyzja), wyrażony jako ułamek dziesiętny.

PRZYKŁAD

Przykład 43:

Chcemy oszacować odsetek Polaków, którzy popierają wprowadzenie nowej technologii. Chcemy to zrobić z 95% poziomem ufności ($Z = 1,96$). Chcemy, aby nasz błąd maksymalny nie przekroczył 3 punktów procentowych ($E = 0,03$). Nie mamy wcześniejszych danych, więc przyjmujemy, że **szacowana proporcja** $p = 0,5$ (50%).

Podstawiamy do wzoru: $n = \frac{(1,96)^2 \cdot 0,5(1-0,5)}{(0,03)^2} = \frac{3,8416 \cdot 0,25}{0,0009} = \frac{0,9604}{0,0009} \approx 1067,11$. Musimy wynik zaokrąglić w górę. Zatem minimalna liczebność próby to 1068 osób.

Rozdział 8

Parametryczne testy istotności

8.1. Test dla wartości średniej populacji

Autorzy/Autorki: Anna Serwatka

Testowanie hipotez statystycznych jest podstawowym rodzajem wnioskowania statystycznego. Hipotezy statystyczne są to przypuszczenia dotyczące rozkładu populacji. Hipotezy parametryczne, precyzujące wartości parametrów w rozkładzie populacji, należą do najczęściej sprawdzanych hipotez statystycznych.[10] Aby zweryfikować hipotezę statystyczną będziemy wykorzystywać odpowiednio dobrany test statystyczny. Jest to reguła postępowania, która każdej możliwej próbie losowej przyporządkowuje decyzję przyjęcia lub odrzucenia sprawdzanej hipotezy. W zależności od postaci postawionej hipotezy zerowej oraz postaci hipotezy alternatywnej (tzn. konkurencyjnej w stosunku do hipotezy zerowej), sposób budowy testu jest różny. Istota rzeczy przy budowie każdego testu polega jednak na tym, żeby uchronić się zarówno przed popełnieniem błędu pierwszego rodzaju, polegającym na odrzuceniu hipotezy prawdziwej, jak i przed popełnieniem błędu drugiego rodzaju, polegającym na przyjęciu hipotezy fałszywej.[10] Testy istotności to procedura, w której na podstawie rezultatów próby losowej podejmuje się wyłącznie decyzję o odrzuceniu badanej hipotezy lub stwierdza brak podstaw do jej odrzucenia. W takich testach nie formułuje się decyzji o akceptacji sprawdzanej hipotezy, ponieważ uwzględnia się jedynie ryzyko popełnienia błędu pierwszego rodzaju (czyli odrzucenia prawdziwej hipotezy, mierzone poziomem istotności), bez analizowania skutków ewentualnego popełnienia błędu drugiego rodzaju.

Jak powstają statystyczne testy istotności? W zależności od postaci hipotezy zerowej buduje się statystykę Z z wyników n-elementowej próby i wyznacza się jej rozkład przy założeniu prawdziwości hipotezy H_0. W rozkładzie tym wybiera się taki obszar Q wartości statystyki Z, by spełniona była równość:

$$P(Z \in Q) = \alpha \tag{8.1}$$

gdzie α jest ustalonym z góry, dowolnie małym prawdopodobieństwem. Obszar Q nazywa się obszarem krytycznym testu, gdyż ilekroć wartość statystyki Z z próby znajdzie się w nim, to podejmuje się decyzję odrzucenia hipotezy H_0 na korzyść hipotezy alternatywnej H_1. Natomiast, gdy otrzymana z konkretnej próby wartość statystyki Z nie należy do obszaru krytycznego Q, to nie ma podstaw do odrzucenia hipotezy H_0. Co nie jest to równoznaczne z jej przyjęciem. W tym rozdziale przeanalizujemy testy dla wartości średniej populacji, rozważając trzy podstawowe modele populacji generalnej [10].

MODEL

Model 8: Populacja generalna ma rozkład normalny o nieznanej średniej μ i znanym odchyleniu standardowym σ

Na podstawie wyników n-elementowej próby losowej sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

$$H_0 : \mu = \mu_0 \tag{8.2}$$

$$H_1 : \mu \neq \mu_0 \tag{8.3}$$

Obliczamy średnią z próby $\overline{x}$, a następnie wartość zmiennej normalnej U:

$$U = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}. \tag{8.4}$$

Wyznaczamy wartość krytyczną u_α z tablicy standaryzowanego rozkładu normalnego $N(0,1)$, aby dla zadanego poziomu istotności α zachodził warunek $P(|U| \geq u_\alpha) = \alpha$.

Jeśli z próby otrzymamy taką wartość U, że $|U| \geq u_\alpha$, to hipotezę H_0 odrzucamy. Gdy zaś $|U| < u_\alpha$, to nie ma podstaw do odrzucenia hipotezy zerowej. Zbiór U taki, że $|U| \geq u_\alpha$ jest obszarem krytycznym tego testu.

W zadaniach możemy spotkać inną postać hipotezy alternatywnej. Powyższy test, w którym przedstawiliśmy dwustronny obszar krytyczny stosujemy jedynie dla hipotezy alternatywnej $H_1 : \mu \neq \mu_0$. Jak wyglądają obszary jednostronne i w jakich sytuacjach je stosujemy przedstawiono w uwagach.

UWAGA

Uwaga 1: Lewostronny obszar krytyczny

Jeśli hipoteza alternatywna ma postać $H_1 : \mu < \mu_0$ do testu istotności należy wybrać lewostronny obszar krytyczny, tak, aby spełniona była nierówność $U \leq u_\alpha$. W rozkładzie standaryzowanego rozkładu normalnego szukamy u_α takiego, że $P(U \leq u_\alpha) = \alpha$.

UWAGA

Uwaga 2: Prawostronny obszar krytyczny

Jeśli hipoteza alternawtywna ma postać $H_1 : \mu > \mu_0$ do testu istotności należy wybrać prawostronny obszar krytyczny, tak, aby spełniona była nierówność $U \geq u_\alpha$. W rozkładzie standaryzowanego rozkładu normalnego szukamy u_α takiego, że $P(U \geq u_\alpha) = \alpha$. Hipoteza H_0 zostanie odrzucona tylko wtedy, gdy wartość statystyki U spełni nierówność $U \geq u_\alpha$.

MODEL

Model 9: Populacja generalna ma rozkład normalny o nieznanej

średniej μ i nieznanym odchyleniu standardowym σ

Na podstawie wyników n-elementowej próby losowej sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

$$H_0 : \mu = \mu_0 \tag{8.5}$$

$$H_1 : \mu \neq \mu_0 \tag{8.6}$$

Obliczamy średnią z próby $\overline{x}$, a ponieważ nie znamy odchylenia standardowego σ rozkładu obliczamy odchylenie standardowe z próby s oraz wartość statystyki T danej wzorem:

$$T = \frac{\overline{x} - \mu_0}{s}\sqrt{n-1}. \tag{8.7}$$

Statystyka t ma przy założeniu prawdziwości hipotezy H_0 rozkład t-Studenta o $n-1$ stopniach swobody. Z tablicy rozkładu t-Studenta o $n-1$ stopniach swobody, dla poziomu istotności α odczytuje się wartość t_α, taką, że $P(|T| \geq t_\alpha) = \alpha$. W tym teście mamy obustronny obszar krytyczny. Jeśli dla obliczonej statystyki t zachodzi $|t| \geq t_\alpha$, to hipotezę H_0 należy odrzucić na korzyść hipotezy alternatywnej H_1, w przeciwnym razie nie ma podstaw do odrzucenia hipotezy zerowej.

Czy obszar krytyczny w przypadku modelu II zawsze będzie dwustronny? Nie, analogicznie do modelu I, jeśli hipoteza alternatywna H_1 będzie postaci $\mu < \mu_0$ stosujemy lewostronny obszar krytyczny wyznaczony tak, aby był spełniony warunek $P(t \leq t_\alpha) = \alpha$. Przy hipotezie H_1 postaci $m > m_0$ stosujemy lewostronny obszar krytyczny tak, aby był spełniony warunek $P(t \geq t_\alpha) = \alpha$.

MODEL

Model 10: Populacja generalna ma dowolny rozkład o nieznanej średniej μ i o skończonym, ale nieznanym odchyleniu standardowym σ dla dużej próby

Na podstawie wyników n-elementowej próby losowej sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

$$H_0 : \mu = \mu_0 \tag{8.8}$$

$$H_1 : \mu \neq \mu_0 \tag{8.9}$$

Test istotności dla tej hipotezy prowadzimy tak jak w modelu I, wykorzystując do obliczeń zamiast wartości σ obliczoną z próby wartość s.

Warto zwrócić uwagę na dwie uwagi, które pomogą w interpretacji otrzymanych wyników.

UWAGA

Uwaga 3: Błąd pierwszego rodzaju

Możemy pomylić się i odrzucić hipotezę, która w gruncie rzeczy była prawdziwa (błąd pierwszego rodzaju), jednakże prawdopodobieństwo takiej pomyłki jest bardzo małe, równe obranej dowolnie

liczbie α. Jeżeli natomiast wartość statystyki Z z próby n-elementowej znalazła się poza obszarem krytycznym, to prawdopodobieństwo tego zdarzenia, przy prawdziwości hipotezy H_0, jest równe $1-\alpha$, co jest bliskie 1, więc nie ma podstaw do odrzucenia hipotezy H_0.

UWAGA

Uwaga 4: Akceptacja, odrzucenie hipotezy a dowiedzenie jej prawdziwości

Akceptacja lub odrzucenie hipotezy w analizie statystycznej nie jest tożsame z formalnym dowiedzeniem jej prawdziwości bądź fałszywości. Odrzucając w teście statystycznym badaną hipotezę, kierujemy się wyłącznie tym, że uzyskane dane liczbowe z obserwacji rzeczywistości wskazują na niskie prawdopodobieństwo jej prawdziwości i są z nią niezgodne. Może się jednak zdarzyć sytuacja odwrotna, w której to hipoteza jest poprawna, a jedynie zebrane wyniki z próby są błędne lub po prostu mało prawdopodobne przy założeniu tej hipotezy.

8.2. Test dla dwóch średnich

Autorzy/Autorki: Anna Serwatka

Załóżmy, że mamy dwie populacje o rozkładzie normalnym, a naszym zadaniem jest porównanie średnich w tych dwóch populacjach. W zależności od danych, jakie mamy o rozkładach populacji będziemy stosować jeden z trzech modeli, aby zweryfikować hipotezę zerową $H_0 : \mu_1 = \mu_2$ (gdzie μ_1 jest średnią w pierwszej populacji, a μ_2 jest średnią w drugiej populacji). Hipoteza alternatywna H_1 może być postaci [10]:

1. $\mu_1 \neq \mu_2$ - wybieramy obustronny obszar krytyczny
2. $\mu_1 < \mu_2$ - wybieramy lewostronny obszar krytyczny
3. $\mu_1 > \mu_2$ - wybieramy prawostronny obszar krytyczny.

MODEL

Model 11: Obie populacje mają rozkłady normalne o nieznanych średnich μ_1 i μ_2 i znanych odchyleniach standardowych równych σ_1 i σ_2

Załóżmy, że mamy dwie populacje generalne mające rozkłady normalne $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$. Zakładamy ponadto, że znamy odchylenia standardowe σ_1 i σ_2. Losujemy dwie niezależne próby o liczebności n_1 z pierwszego rozkładu i o liczebności n_2 z drugiego rozkładu. Na podstawie tych wyników sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

1. model z obustronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.10}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{8.11}$$

Obliczamy średnią z każdej próby $\overline{x_1}$ i $\overline{x_2}$, a następnie wartość statystyki U:

$$U = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{8.12}$$

Ta statystyka ma rozkład normalny $N(0,1)$. Wyznaczamy wartość krytyczną u_α z tablicy standaryzowanego rozkładu normalnego, aby dla zadanego poziomu istotności α zachodziła równość $P(|U| \geq u_\alpha) = \alpha$.

Jeśli z próby otrzymamy taką wartość u, że $|U| \geq u_\alpha$, to hipotezę H_0 odrzucamy na korzyść hipotezy H_1. Gdy zaś $|U| < u_\alpha$, to nie ma podstaw do odrzucenia hipotezy zerowej. Zbiór U taki, że $|U| \geq u_\alpha$ jest obszarem krytycznym tego testu.

2. model z lewostronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.13}$$

$$H_1 : \mu_1 < \mu_2 \tag{8.14}$$

Obliczamy średnią z każdej próby oraz wartość statystyki U zgodnie z wzorem (8.12). Ustalamy lewostronny obszar krytyczny $U \leq u_\alpha$. Wartość krytyczna u_α spełnia warunek $P(U \leq u_\alpha) = \alpha$.

3. model z prawostronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.15}$$

$$H_1 : \mu_1 > \mu_2 \tag{8.16}$$

Obliczamy średnią z każdej próby oraz wartość statystyki U zgodnie z wzorem (8.12). Ustalamy lewostronny obszar krytyczny $U \geq u_\alpha$. Wartość krytyczna u_α odczytana z tablicy standaryzowanego rozkładu normalnego spełnia warunek $P(U \geq u_\alpha) = \alpha$.

MODEL

Model 12: Obie populacje mają rozkłady normalne o nieznanych średnich m_1 i m_2 i nieznanych równych odchyleniach standardowych $\sigma_1 = \sigma_2$

Załóżmy, że mamy dwie populacje generalne mające rozkłady normalne $N(m_1, \sigma_1)$ i $N(m_2, \sigma_2)$. Zakładamy ponadto, że nie znamy ani średnich ani odchyleń standardowych tych rozkładów. Ponadto zakładamy, że $\sigma_1 = \sigma_2$. Losujemy dwie niezależne próby o liczebności n_1 z pierwszego rozkładu i o liczebności n_2 z drugiego rozkładu (dopuszczamy małe próby). Na podstawie tych wyników sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

1. model z obustronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.17}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{8.18}$$

Obliczamy średnią z każdej próby $\overline{x_1}$ i $\overline{x_2}$, a następnie wariancje s_1^2 i s_2^2. Liczymy następnie

wartość statystyki t:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}. \tag{8.19}$$

Ta statystyka ma rozkład t-Studenta o $n_1 + n_2 - 2$ stopniach swobody (przy założeniu prawdziwości hipotezy H_0). Wyznaczamy wartość krytyczną t_α z tablicy rozkładu t-Studenta o $n_1 + n_2 - 2$ stopniach swobody, aby dla zadanego poziomu istotności α zachodziła równość $P(|t| \geq t_\alpha) = \alpha$.

Jeśli z próby otrzymamy taką wartość t, że $|t| \geq t_\alpha$, to hipotezę H_0 odrzucamy na korzyść hipotezy H_1. Gdy zaś $|t| < t_\alpha$, to nie ma podstaw do odrzucenia hipotezy zerowej.

2. model z lewostronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.20}$$

$$H_1 : \mu_1 < \mu_2 \tag{8.21}$$

Obliczamy średnią z każdej próby oraz wartość statystyki t zgodnie z wzorem (8.19). Ustalamy lewostronny obszar krytyczny $t \leq t_\alpha$. Wartość krytyczna t_α spełnia warunek $P(t \leq t_\alpha) = \alpha$.

3. model z prawostronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.22}$$

$$H_1 : \mu_1 > \mu_2 \tag{8.23}$$

Obliczamy średnią z każdej próby oraz wartość statystyki U zgodnie z wzorem (8.19). Ustalamy prawostronny obszar krytyczny $t \geq t_\alpha$. Wartość krytyczna t_α odczytana z tablicy rozkładu t-Studdenta spełnia warunek $P(t \geq t_\alpha) = \alpha$.

MODEL

Model 13: Obie populacje mają rozkłady o nieznanych, ale skończonych wariancjach σ_1^2 i σ_2^2. Model dla dużych prób

Zauważmy, że po pierwsze w tym modelu nie ma konieczności, aby były to rozkłady normalne, a po drugie wymagamy dużej liczebności prób.

Załóżmy, że mamy dwie populacje generalne mające rozkłady z nieznanymi wariancjami $\sigma_1^2 < \infty$ i $\sigma_2^2 < \infty$ Losujemy dwie niezależne duże próby o liczebności n_1 z pierwszego rozkładu i o liczebności n_2 z drugiego rozkładu. Na podstawie tych wyników sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

1. model z obustronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.24}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{8.25}$$

Obliczamy średnią z każdej próby $\overline{x_1}$ i $\overline{x_2}$, oraz wariancję z każdej próby s_1^2 i s_2^2 a następnie

wartość statystyki U:

$$u = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{8.26}$$

Ta statystyka ma rozkład normalny $N(0,1)$. Wyznaczamy wartość krytyczną u_α z tablicy standaryzowanego rozkładu normalnego, aby dla zadanego poziomu istotności α spełniony był warunek $P(|U| \geq u_\alpha) = \alpha$.

Jeśli z próby otrzymamy taką wartość u, że $|u| \geq u_\alpha$, to hipotezę H_0 odrzucamy na korzyść hipotezy H_1. Gdy zaś $|u| < u_\alpha$, to nie ma podstaw do odrzucenia hipotezy zerowej. Zbiór U taki, że $|U| \geq u_\alpha$ jest obszarem krytycznym tego testu.

2. model z lewostronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.27}$$

$$H_1 : \mu_1 < \mu_2 \tag{8.28}$$

Obliczamy średnią z każdej próby oraz wartość statystyki U zgodnie z wzorem (8.12). Ustalamy lewostronny obszar krytyczny $U \leq u_\alpha$. Wartość krytyczna u_α spełnia warunek $P(U \leq u_\alpha) = \alpha$.

3. model z prawostronnym obszarem krytycznym

$$H_0 : \mu_1 = \mu_2 \tag{8.29}$$

$$H_1 : \mu_1 > \mu_2 \tag{8.30}$$

Obliczamy średnią z każdej próby oraz wartość statystyki U zgodnie z wzorem (8.12). Ustalamy prawostronny obszar krytyczny $U \geq u_\alpha$. Wartość krytyczna u_α odczytana z tablicy standaryzowanego rozkładu normalnego spełnia warunek $P(U \geq u_\alpha) = \alpha$.

8.3. Test dla wskaźnika struktury

Autorzy/Autorki: Anna Serwatka

W analizie statystycznej, szczególne miejsce zajmują testy istotności, które umożliwiają podejmowanie decyzji dotyczących populacji na podstawie danych z próby. Jednym z ważnych przypadków, mających zastosowanie m.in. w badaniach społecznych, rynkowych czy demograficznych, jest ocena, czy udział określonego składnika w strukturze populacji odbiega istotnie od zakładanego poziomu. W takich sytuacjach stosuje się parametryczny test istotności dla wskaźnika struktury, zwany również testem dla jednej proporcji. Test ten pozwala odpowiedzieć na pytanie, czy zaobserwowany w próbie udział danego zjawiska różni się istotnie od wartości teoretycznej (np. prognozowanej, normatywnej, historycznej). Wskaźnik struktury p może przyjąć wartości z przedziału $(0,1)$. Omówimy model parametrycznego testu istotności dla wskaźnika struktury dla dużej próby. Przy $n > 100$ zakładamy, że wskaźnik struktury $\frac{m}{n}$ ma rozkład asymptotycznie normalny, co pozwala na zbudowanie obszaru krytycznego w oparciu o rozkład normalny [10].

MODEL

Model 14: Populacja generalna ma rozkład dwupunktowy z parametrem p. Model dla licznej próby $n > 100$

Załóżmy, że mamy daną populację generalną o rozkładzie dwupunktowym z parametrem p. Losujemy próbę n elementów z tej populacji w sposób niezależny (n jest duże takie, że $n > 100$). Niech m oznacza liczbę elementów wyróżnionych znalezionych w próbie. Na podstawie wyników n-elementowej próby losowej sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

$$H_0 : p = p_0 \tag{8.31}$$

$$H_1 : p \neq p_0 \tag{8.32}$$

gdzie p_0 jest hipotetyczną wartością parametru p. Obliczamy wskaźnik struktury z próby $\frac{m}{n}$ oraz wartość statystyki

$$u = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}, \tag{8.33}$$

gdzie $q_0 = 1 - p_0$.

Zakładając prawdziwość hipotezy H_0 dostajemy informację o tym, że statystyka u ma rozkład asymptotycznie normalny $N(0, 1)$. Idąc dalej, znajdujemy z tablicy standaryzowanego rozkładu normalnego wartość krytyczną u_α, aby zachodziła równość $P(|U| \geq u_\alpha) = \alpha$. Jeżeli dla obliczonej wartości statystyki u zajdzie $u \in (-\infty, -u_\alpha] \cup [u_\alpha, +\infty)$ wówczas hipotezę H_0 odrzucamy na korzyść hipotezy alternatywnej H_1. W przeciwnym wypadku brak podstaw do odrzucenia hipotezy H_0.

UWAGA

Uwaga 5: Jednostronne obszary krytyczne

Warto zauważyć, że opisany powyżej test istotności dotyczy przypadku **dwustronnego**, w którym hipoteza alternatywna zakłada możliwość odchylenia parametru w obu kierunkach – zarówno poniżej, jak i powyżej wartości zakładanej w hipotezie zerowej. Istnieją jednak sytuacje, w których interesuje nas tylko jeden kierunek odchylenia – na przykład, gdy chcemy sprawdzić, czy udział pewnej cechy w populacji **przekracza** wartość oczekiwaną (test prawostronny), albo czy jest **niższy** od tej wartości (test lewostronny).

W takich przypadkach stosujemy odpowiednio **prawostronny** lub **lewostronny obszar krytyczny**. Konstrukcja testu zmienia się wówczas w zakresie wyboru wartości krytycznej oraz interpretacji statystyki testowej:

- **W teście prawostronnym**, hipoteza alternatywna ma postać $H_1 : p > p_0$, a obszar krytyczny obejmuje wartości większe niż u_α, tzn. $u \in [u_\alpha, +\infty)$.
- **W teście lewostronnym**, hipoteza alternatywna ma postać $H_1 : p < p_0$, a obszar krytyczny to $u \in (-\infty, -u_\alpha]$.

W obu przypadkach wartość krytyczna u_α wyznaczana jest z tablicy rozkładu normalnego standaryzowanego tak, aby spełniona była równość $P(U \geq u_\alpha) = \alpha$ dla testu prawostronnego lub $P(U \leq -u_\alpha) = \alpha$ dla testu lewostronnego.

Dzięki takiemu podejściu test można lepiej dostosować do konkretnego problemu badawczego, uwzględniając kierunek potencjalnych odchyleń parametru od wartości oczekiwanej.

8.4. Test dla dwóch wskaźników struktury

Autorzy/Autorki: Anna Serwatka

Celem parametrycznych testów istotności dla dwóch wskaźników struktury jest sprawdzenie, czy różnice zaobserwowane w próbach pochodzących z dwóch populacji mają charakter przypadkowy, czy też są statystycznie istotne. Punktem wyjścia jest tu porównanie dwóch wskaźników struktury (proporcji) p_1 i p_2, oszacowanych na podstawie niezależnych prób, poprzez skonstruowanie odpowiedniej statystyki testowej i określenie jej rozkładu przy założeniu prawdziwości hipotezy zerowej. Rozdział ten przedstawia metodykę przeprowadzania testu dla dwóch proporcji, w tym formułowanie hipotez, wyznaczanie wartości krytycznych, warunki stosowalności testu oraz interpretację wyników — zarówno w wersji dwustronnej, jak i jednostronnej. Dzięki temu narzędziu możliwe jest podejmowanie racjonalnych decyzji na podstawie danych empirycznych w kontekście porównań między grupami. Przedstawiony poniżej model umożliwia weryfikację hipotezy H_0 przy dwóch licznych próbach ($n_1 > 100$, $n_2 > 100$).

MODEL

Model 15: Dwie populacje generalne o rozkładach dwupunktowych z parametrami p_1 i p_2. Model dla licznych prób $n_1, n_2 > 100$

Załóżmy, że mamy dane dwie populacje generalne o rozkładzie dwupunktowym z parametrami odpowiednio p_1 i p_2. Losujemy dwie duże próby (dla pierwszej populacji liczebność $n_1 > 100$, dla drugiej populacji liczebność $n_2 > 100$). Niech m_1 oznacza liczbę elementów wyróżnionych znalezionych w próbie pierwszej, a m_2 oznacza liczbę elementów wyróżnionych znalezionych w próbie drugiej.
Na podstawie wyników z tych prób losowych sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

$$H_0 : p_1 = p_2 \tag{8.34}$$

$$H_1 : p_1 \neq p_2. \tag{8.35}$$

Obliczamy wartość średniego wskaźnika struktury z obu prób $\overline{p}$, wartość pseudoliczebności próby n oraz wartość statystyki u, z następujących wzorów:

$$\overline{p} = \frac{m_1 + m_2}{n_1 + n_2}, \tag{8.36}$$

$$n = \frac{n_1 n_2}{n_1 + n_2}, \tag{8.37}$$

$$u = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{\overline{p} \cdot \overline{q}}{n}}}, \tag{8.38}$$

gdzie $\overline{q} = 1 - \overline{p}$.

Statystyka u ma rozkład asymptotycznie normalny $N(0,1)$ (przy założeniu prawdziwości hipotezy H_0). Znajdujemy z tablicy standaryzowanego rozkładu normalnego wartość krytyczną u_α, aby zachodziła równość $P(|U| \geq u_\alpha) = \alpha$. Jeżeli dla obliczonej wartości statystyki u zajdzie $u \in (-\infty, -u_\alpha] \cup [u_\alpha, +\infty)$ wówczas hipotezę H_0 odrzucamy na korzyść hipotezy alternatywnej H_1. Jeśli $u \in (-u_\alpha, u_\alpha)$ to nie ma podstaw do odrzucenia hipotezy H_0.

Analogicznie możliwe jest tworzenie jednostronnych obszarów krytycznych [10].

8.5. Test dla wariancji

Autorzy/Autorki: Anna Serwatka

Wariancja jest parametrem opisującym zmienność cechy w populacji. Jej znajomość pozwala nie tylko zrozumieć rozproszenie danych wokół wartości średniej, ale również odgrywa istotną rolę w wielu zagadnieniach praktycznych. W związku z tym często pojawia się potrzeba sprawdzenia, czy obserwowana w próbie wariancja odpowiada zakładanej wartości teoretycznej. W takich sytuacjach stosuje się parametryczny test istotności dla wariancji populacji generalnej, który umożliwia weryfikację hipotezy dotyczącej konkretnej wartości wariancji. Podstawą testu jest założenie, że populacja ma rozkład normalny, co pozwala wykorzystać rozkład chi-kwadrat (χ^2) jako rozkład statystyki testowej [10].

MODEL

Model 16: Populacja generalna ma rozkład normalny $N(\mu, \sigma)$ o nieznanych parametrach μ i σ.

Załóżmy, że mamy daną populację generalną o rozkładzie normalnym $N(\mu, \sigma)$ o nieznanych parametrach μ i σ. Losujemy próbę n elementów z tej populacji w sposób niezależny. Ustalmy poziom istotności α. Na podstawie wyników n-elementowej próby losowej sprawdzamy hipotezę zerową H_0, wobec hipotezy alternatywnej H_1, gdzie:

$$H_0 : \sigma^2 = \sigma_0^2 \tag{8.39}$$

$$H_1 : \sigma^2 > \sigma_0^2 \tag{8.40}$$

gdzie σ_0^2 jest hipotetyczną wartością wariancji σ^2. Obliczamy wariancję z próby s^2 oraz wartość statystyki

$$\chi^2 = \frac{ns^2}{\sigma_0^2}, \tag{8.41}$$

która przy założeniu prawdziwości hipotezy H_0 ma rozkład χ^2 z $n-1$ stopniami swobody. Znajdujemy z tablicy rozkładu χ^2 (dla poziomu istotności α i $n-1$ stopni swobody) wartość krytyczną χ_α, aby spełniony był warunek $P(\chi^2 \geq \chi_\alpha^2) = \alpha$. Jest to prawostronny obszar krytyczny. Wartość statystyki χ^2 obliczamy z próby i sprawdzamy czy $\chi^2 \geq \chi_\alpha^2$. Jeśli zachodzi ta nierówność to wówczas hipotezę H_0 odrzucamy na korzyść hipotezy alternatywnej H_1. Jeśli zachodzi nierówność $\chi^2 < \chi_\alpha^2$ to nie ma podstaw do odrzucenia hipotezy zerowej.

UWAGA

Uwaga 6:

W tablicach rozkładu χ^2 najczęściej podawane są wartości dla liczby stopni swobody od 1 do 30. Dla większych wartości nie umieszcza się zazwyczaj szczegółowych danych, ponieważ rozkład chi-kwadrat wraz ze wzrostem liczby stopni swobody upodabnia się do rozkładu normalnego.

Inaczej mówiąc, gdy liczba stopni swobody przekracza 30, rozkład χ^2 można z dobrym przybliżeniem

traktować jako rozkład normalny, co znacznie upraszcza obliczenia i interpretację wyników. Takie przybliżenie jest uzasadnione teoretycznie i praktycznie — zwłaszcza w zastosowaniach, gdzie precyzyjne wartości z tablic nie są konieczne.

Rozdział 9

Związki korelacyjne między zmiennymi różnych typów

9.1. Korelacja i kowariancja

Autorzy/Autorki: Anna Serwatka

Jednymi z narzędzi stosowanych w statystyce i analizie danych, które pozwalają sprawdzić, czy dwie zmienne są ze sobą powiązane czy też nie są kowariancja i korelacja.

Kowariancja to średnia arytmetyczna iloczynu odchyleń wartości zmiennych X i Y od ich wartości średnich. Dodatnia wartość kowariancji oznacza, że przy wzroście wartości X wartości Y na ogół także rosną. Ujemna wartość kowariancji będzie wskazywała, że przy wzroście X wartości Y na ogół maleją.[6].

DEFINICJA

Definicja 49: Kowariancja

Niech dane będą dwie zmienne losowe X i Y. Kowariancją określamy:

$$cov(X,Y) = E((X - EX)(Y - EY)) = E(XY) - EXEY. \tag{9.1}$$

Wartość kowariancji jest ściśle związana z jednostkami, w jakich mierzone są zmienne. To oznacza, że nie możemy bezpośrednio porównywać kowariancji między różnymi zbiorami danych, jeśli zmienne są wyrażone w innych jednostkach. Na przykład, kowariancja między wzrostem mierzonym w centymetrach a wagą w kilogramach będzie miała inną wartość niż ta sama kowariancja, gdy wzrost jest w metrach, a waga w gramach – mimo że sam związek między wzrostem a wagą pozostaje taki sam. Dodatkowo, kowariancja nie jest standaryzowana, co oznacza, że jej zakres wartości jest nieograniczony. Trudno więc ocenić siłę związku, patrząc tylko na wartość kowariancji, bo duża wartość może wynikać po prostu z dużych wartości samych zmiennych, a niekoniecznie z silnego powiązania między nimi. Na koniec, kowariancja mówi nam tylko o kierunku liniowego związku. Dodatnia wartość wskazuje, że zmienne poruszają się w tym samym kierunku (np. wzrost jednej oznacza wzrost drugiej), ujemna – że w przeciwnych (np. wzrost jednej oznacza spadek drugiej), a wartość bliska zeru sugeruje brak liniowej zależności. Nie mówi nam jednak nic o sile tego związku.

Z kolei współczynnik korelacji jest standaryzowaną formą kowariancji, która pozwala ocenić nie tylko kierunek, ale również siłę liniowego związku między zmiennymi. Wartość korelacji mieści się zawsze w

przedziale od -1 do 1, co ułatwia interpretację i porównywanie siły zależności niezależnie od jednostek pomiarowych.

DEFINICJA

Definicja 50: Współczynnik korelacji Pearsona

Niech dane będą dwie zmienne losowe X i Y. Współczynnik korelacji to miara statystyczna określająca siłę i kierunek związku liniowego między dwiema zmiennymi losowymi X i Y.

Najczęściej stosowanym współczynnikiem korelacji jest współczynnik korelacji Pearsona, wyrażony wzorem:

$$\rho = \frac{cov(X,Y)}{\sqrt{D^2X}\sqrt{D^2Y}}. \tag{9.2}$$

Współczynnik korelacji liniowej mierzy siłę i kierunek zależności liniowej między dwiema zmiennymi losowymi X i Y. Jego wartości mieszczą się w przedziale $[-1, 1]$. Poniżej przedstawiono znaczenie różnych wartości tego współczynnika:

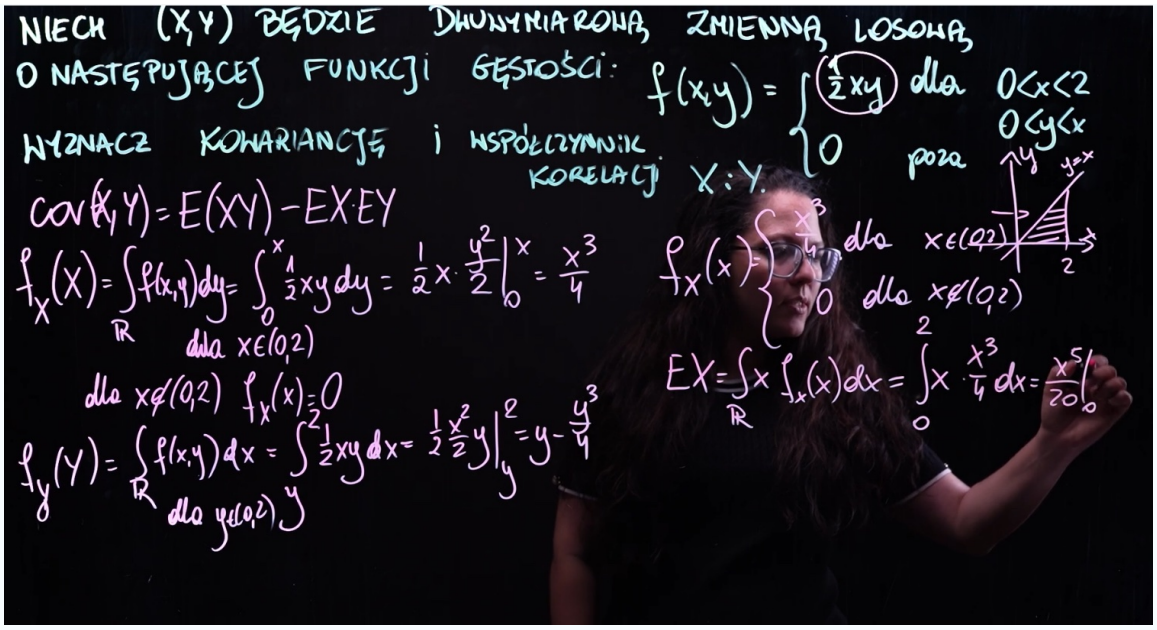
Tabela 2: Opis tabeli

Wartość ρ	**Interpretacja**
-1	Korelacja doskonała ujemna
$(-1,\ -0.9]$	Silna ujemna zależność liniowa
$(-0.9,\ -0.7]$	Silna ujemna korelacja
$(-0.7,\ -0.4]$	Umiarkowana ujemna korelacja
$(-0.4,\ -0.1]$	Słaba ujemna korelacja
$(-0.1,\ 0.1)$	Brak (lub bardzo słaba) zależność liniowa
$[0.1,\ 0.4)$	Słaba dodatnia korelacja
$[0.4,\ 0.7)$	Umiarkowana dodatnia korelacja
$[0.7,\ 0.9)$	Silna dodatnia korelacja
$[0.9,\ 1)$	Silna dodatnia zależność liniowa
1	Korelacja doskonała dodatnia

Silna korelacja między dwiema zmiennymi nigdy nie oznacza, że jedna zmienna jest przyczyną drugiej. Istnieje wiele powodów, dla których możemy zaobserwować korelację bez związku przyczynowego. Mogą to być na przykład zmienne ukryte, czyli inne, niemierzalne czynniki, które wpływają na obie nasze zmienne, tworząc iluzję bezpośredniego powiązania. Pomyślmy o sprzedaży lodów i liczbie utonięć – obie te rzeczy są ze sobą dodatnio skorelowane, ale nie dlatego, że lody powodują utonięcia, lecz dlatego, że zarówno sprzedaż lodów, jak i utonięcia rosną latem, gdy jest ciepło i ludzie więcej pływają.

PRZYKŁAD

Przykład 44: Kowariancja i współczynnik korelacji [7]



https://vimeo.com/1097819169/bef718590a

Kowariancja i współczynnik korelacji

9.2. Prosta regresja liniowa

Autorzy/Autorki: Anna Serwatka

Regresja liniowa to jedna z fundamentalnych i najczęściej używanych technik w statystyce i uczeniu maszynowym. Jej głównym celem jest modelowanie związku między jedną zmienną zależną (wynikową) a jedną lub wieloma zmiennymi niezależnymi (predyktorami). W najprostszej formie, regresja liniowa próbuje znaleźć linię (lub hiperplan w przypadku wielu predyktorów), która najlepiej pasuje do danych, minimalizując odległość między przewidywanymi wartościami a rzeczywistymi wartościami zmiennej zależnej.

Rozróżniamy dwa główne typy regresji liniowej: prostą regresję liniową (gdy mamy tylko jeden predyktor) i wieloraką regresję liniową (gdy mamy wiele predyktorów).

DEFINICJA

Definicja 51: Prosta regresja liniowa

Prosta regresja liniowa to metoda statystyczna, która modeluje liniowy związek między jedną zmienną zależną (oznaczmy ją jako Y) a jedną zmienną niezależną (oznaczmy jako X). Celem jest znalezienie prostej linii (zwanej linią regresji), która najlepiej opisuje ten związek, umożliwiając prognozowanie wartości Y na podstawie X. Model prostej regresji liniowej wyraża równanie:

$$Y = \beta_0 + \beta_1 X, \tag{9.3}$$

gdzie Y to zmienna zależna, którą chcemy przewidzieć lub wyjaśnić, a zmienna X to zmienna niezależna (predyktor, zmienna wyjaśniająca).

Szukamy współczynników β_0 i β_1. Wyraz wolny czyli β_0 (przecięcie z osią OY), to przewidywana wartość Y, gdy X wynosi 0. Współczynnik nachylenia prostej β_1, informuje o kierunku i sile liniowego związku między X a Y. W rzeczywistych danych punkty zazwyczaj nie leżą idealnie na prostej, a do oszacowania parametrów β_0 i β_1 często stosuje się bardziej złożone obliczenia (metoda najmniejszych kwadratów), które minimalizują sumę kwadratów błędów [7].

Główna idea metody najmniejszych kwadratów polega na znalezieniu linii (lub krzywej) najlepiej dopasowanej do zbioru punktów danych, poprzez minimalizowanie sumy kwadratów różnic (reszt) między obserwowanymi wartościami a wartościami przewidywanymi przez model. Należy znaleźć minimum dla sumy kwadratów różnic wartości obserwowanych i wartości teoretycznych. Aby znaleźć wartości β_0 i β_1, które minimalizują tę sumę kwadratów, stosuje się rachunek różniczkowy. Oblicza się pochodne cząstkowe sumy kwadratów reszt względem β_0 i β_1, a następnie przyrównuje je do zera. Rozwiązanie tego układu równań liniowych daje nam wzory na optymalne wartości β_0 i β_1.

Dla prostej regresji liniowej wzory te wyglądają następująco:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{9.4}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}, \tag{9.5}$$

gdzie $\bar{X}, \bar{Y}$ to średnie arytmetyczne odpowiednio X i Y.

W metodzie najmniejszych kwadratów zakładamy, że związek między zmiennymi X i Y jest liniowy, a liczba obserwacji musi być większa niż liczba szacowanych parametrów.

ZADANIE

Zadanie 13: Zadanie z regresji liniowej

Treść zadania:

Załóżmy, że chcemy zbadać, czy istnieje liniowy związek między wydatkami na reklamę X_i (w tysiącach złotych) a sprzedażą produktu Y_i (w tysiącach sztuk) dla pewnej firmy. Dane z ostatnich 5 miesięcy przedstawia tabela:

Tabela 3: Dane

Miesiąc	Wydatki na reklamę (X_i) (w tys. zł)	Sprzedaż (Y_i) (w tys. sztuk)
1	1	10
2	2	12
3	3	15
4	4	17
5	5	21

1. Napisz równanie prostej regresji liniowej $Y = \beta_0 + \beta_1 X$ używając metody najmniejszych kwadratów.
2. Jaka jest przewidywana sprzedaż, jeśli wydatki na reklamę wyniosą 3 500 zł?

Rozwiązanie:

Mamy $n = 5$ obserwacji.

Krok 1: Obliczanie średnich arytmetycznych X **i** Y $\bar{X} = \frac{\sum_{i=1}^{5} X_i}{5} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$ $\bar{Y} = \frac{\sum_{i=1}^{5} Y_i}{5} = \frac{10+12+15+17+21}{5} = \frac{75}{5} = 15$

Krok 2: Obliczanie sum potrzebnych do wzoru na β_1

Tabela 4: Obliczanie sum

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	10	$1 - 3 = -2$	$10 - 15 = -5$	$(-2) \cdot (-5) = 10$	$(-2)^2 = 4$
2	12	$2 - 3 = -1$	$12 - 15 = -3$	$(-1) \cdot (-3) = 3$	$(-1)^2 = 1$
3	15	$3 - 3 = 0$	$15 - 15 = 0$	$0 \cdot 0 = 0$	$0^2 = 0$
4	17	$4 - 3 = 1$	$17 - 15 = 2$	$1 \cdot 2 = 2$	$1^2 = 1$
5	21	$5 - 3 = 2$	$21 - 15 = 6$	$2 \cdot 6 = 12$	$2^2 = 4$
Suma		**0**	**0**	**27**	**10**

Z tabeli odczytujemy sumy: $\sum_{i=1}^{5}(X_i - \bar{X})(Y_i - \bar{Y}) = 27$ $\sum_{i=1}^{5}(X_i - \bar{X})^2 = 10$

Krok 3: Obliczanie współczynnika nachylenia β_1 $\beta_1 = \frac{\sum_{i=1}^{5}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{5}(X_i - \bar{X})^2} = \frac{27}{10} = 2,7$

Krok 4: Obliczanie wyrazu wolnego β_0 $\beta_0 = \bar{Y} - \beta_1 \bar{X} = 15 - (2,7) \cdot 3 = 15 - 8,1 = 6,9$

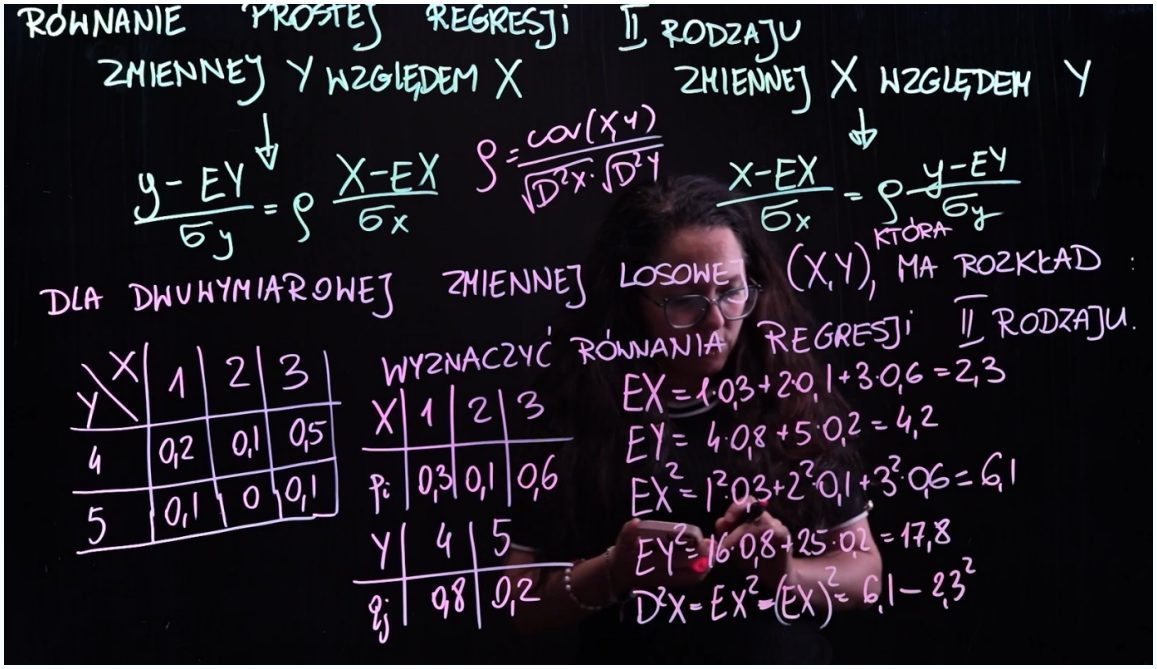
Krok 5: Zapisanie równania prostej regresji Oszacowane równanie prostej regresji liniowej to: $Y = 6,9 + 2,7X$

Jeśli wydatki na reklamę wyniosą 3.5 tys. zł, podstawiamy $X = 3.5$ do naszego równania regresji: $Y = 6,9 + 2,7 \cdot (3,5) = 6,9 + 9,45 = 16,35$.

Przewidywana sprzedaż wyniesie 16,35 tys. sztuk, jeśli wydatki na reklamę wyniosą 3,5 tys. zł.

PRZYKŁAD

Przykład 45: Równanie regresji II rodzaju



https://vimeo.com/1097821578/600f583da4

Równanie regresji II rodzaju

9.3. Wieloraka regresja liniowa

Autorzy/Autorki: Anna Serwatka

Wieloraka regresja liniowa jest rozszerzeniem prostej regresji liniowej.

DEFINICJA

Definicja 52: Wieloraka regresja liniowa

Załóżmy, że dana jest jedna zmienna zależna Y oraz zmienne niezależne $X_1, X_2, \ldots, X_k$. Zakładamy, że związek między tymi zmiennymi jest liniowy. Wtedy model wielorakiej regresji liniowej dla i-tej obserwacji wyraża się wzorem:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i, \tag{9.6}$$

gdzie:

- Y_i to i-ta obserwacja zmiennej zależnej,
- X_{ij} to i-ta obserwacja j-tej zmiennej niezależnej,
- β_0 to wyraz wolny,
- β_j to **cząstkowy współczynnik regresji** dla j-tej zmiennej niezależnej dla $j = 1, \ldots, k$,

- ϵ_i to składnik błędu (reszta) dla i-tej obserwacji.

Parametry są najczęściej szacowane za pomocą **metody najmniejszych kwadratów (MNK)**. Model regresji liniowej wielorakiej dla wszystkich n obserwacji można zapisać w postaci macierzowej:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{9.7}$$

gdzie:

- $\mathbf{Y}$ to wektor kolumnowy zmiennej zależnej (wymiar $n \times 1$): $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$
- $\mathbf{X}$ to macierz wartości zmiennych objaśniających (wymiar $n \times (k+1)$): $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}$
- $\boldsymbol{\beta}$ to wektor kolumnowy nieznanych parametrów regresji (wymiar $(k+1) \times 1$): $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$
- $\boldsymbol{\epsilon}$ to wektor kolumnowy składników błędu (wymiar $n \times 1$): $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$

Wtedy estymator wektora współczynników $\boldsymbol{\beta}$ metodą najmniejszych kwadratów, oznaczany jako $\hat{\boldsymbol{\beta}}$, jest dany wzorem:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{9.8}$$

Po obliczeniu $\hat{\boldsymbol{\beta}}$, możemy zapisać równanie przewidywanego modelu regresji jako:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{9.9}$$

Lub dla pojedynczej obserwacji:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_k X_{ik} \tag{9.10}$$

9.4. Regresja nieliniowa

Autorzy/Autorki: Anna Serwatka

Regresja nieliniowa to statystyczna metoda modelowania związków między zmienną zależną a jedną lub więcej zmiennymi niezależnymi, gdzie ten związek nie jest liniowy. W przeciwieństwie do regresji liniowej, która zakłada, że dane można dopasować do linii prostej (lub płaszczyzny w przypadku wielu zmiennych), regresja nieliniowa wykorzystuje funkcje, które mają kształt krzywej. Te nieliniowe funkcje mogą lepiej odzwierciedlać złożone zależności, które często występują w rzeczywistych danych.

DEFINICJA

Definicja 53: Regresja nieliniowa

Model regresji nieliniowej można zapisać jako: $Y = f(X, \boldsymbol{\beta}) + \epsilon$ Gdzie:

- Y to zmienna zależna.
- X to jedna lub więcej zmiennych niezależnych.
- $f(X, \boldsymbol{\beta})$ to funkcja niezależnych zmiennych losowych X oraz zestawu parametrów nieliniowych $\boldsymbol{\beta}$.
- $\boldsymbol{\beta}$ to wektor nieznanych parametrów nieliniowych, które muszą być oszacowane na podstawie danych.
- ϵ to składnik błędu, reprezentujący niewyjaśnioną zmienność.

Kluczowe różnice od regresji liniowej:

- **Kształt funkcji:** W regresji liniowej funkcja jest zawsze liniowa (np. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$). W regresji nieliniowej funkcja f może przybierać dowolny kształt krzywej, np. wykładniczy ($Y = \beta_0 e^{\beta_1 X}$).
- **Estymacja parametrów:** O ile w regresji liniowej parametry można często oszacować analitycznie za pomocą metody najmniejszych kwadratów, o tyle w regresji nieliniowej rzadko jest to możliwe. Zazwyczaj stosuje się **iteracyjne algorytmy optymalizacyjne**, takie jak algorytm Gaussa-Newtona, Levenberga-Marquardta czy gradient prosty, które minimalizują sumę kwadratów reszt poprzez kolejne przybliżenia. Algorytmy te wymagają podania początkowych wartości parametrów.
- **Interpretacja parametrów:** Interpretacja parametrów w regresji nieliniowej jest często mniej intuicyjna niż w regresji liniowej i zależy od konkretnej postaci funkcji nieliniowej. Współczynnik β_1 w modelu liniowym zawsze oznacza jednostkową zmianę Y na jednostkową zmianę X, natomiast w modelu nieliniowym jego wpływ może być zmienny.

Regresję nieliniową stosuje się, gdy teoretyczna wiedza o zjawisku lub analiza wizualna (np. wykres rozrzutu) sugeruje, że związek między zmiennymi nie jest liniowy. Jest powszechnie używana w wielu dziedzinach, takich jak biologia i medycyna (modelowanie wzrostu populacji), ekonomia (krzywe podaży i popytu, modele wzrostu ekonomicznego), inżynieria (charakterystyki materiałów, dynamika systemów). Wybór odpowiedniego modelu nieliniowego często wymaga głębokiej wiedzy dziedzinowej oraz eksperymentalnego dopasowywania różnych funkcji, a także oceny dopasowania modelu za pomocą wskaźników takich jak R^2 czy analiza reszt.

Rozdział 10

Przykłady paradoksów statystycznych i probabilistycznych

10.1. Paradoks Simpsona

Autorzy/Autorki: Anna Serwatka

Analiza danych stanowi fundament współczesnych decyzji w wielu dziedzinach, od medycyny po biznes. Często opieramy się na intuicji, wyciągając wnioski z pozornych zależności. Istnieje jednak statystyczne zjawisko, znane jako paradoks Simpsona, które potrafi całkowicie odwrócić lub zniwelować zaobserwowane trendy, gdy agregujemy dane z różnych podgrup. To zjawisko uwypukla kluczową zasadę: ogólne statystyki mogą wprowadzać w błąd, jeśli nie uwzględnia się kontekstu i struktury danych. Zrozumienie paradoksu Simpsona jest zatem niezbędne dla każdego, kto dąży do rzetelnej interpretacji informacji i unikania powierzchownych wniosków. Paradoks Simpsona przedstawiają dwa następujące przykłady.

PRZYKŁAD

Przykład 46: Dwa leki na chorobę

Wyobraźmy sobie, że prowadzimy badanie kliniczne, by sprawdzić, który z dwóch nowych leków – **Lek A** czy **Lek B** – jest skuteczniejszy w leczeniu pewnej choroby. Dla uproszczenia podzielimy pacjentów na dwie grupy: **młodych** i **starszych**. Oto wyniki leczenia:

Tabela 5: *Wyniki leczenia*

Grupa pacjentów	Lek A (wyleczeni/leczonych)	Lek B (wyleczeni/leczonych)
Młodzi	90/100 (90% skuteczności)	10/10 (100% skuteczności)
Starsi	10/10 (100% skuteczności)	90/100 (90% skuteczności)

Spójrzmy na poszczególne grupy wiekowe:

- wśród młodych pacjentów: Lek B jest wyraźnie skuteczniejszy (100%) niż Lek A (90%).
- wśród starszych pacjentów: Lek A jest wyraźnie skuteczniejszy (100%) niż Lek B (90%).

W każdej grupie wiekowej jeden lek jest lepszy od drugiego. Logiczne, prawda? Teraz zsumujmy dane dla wszystkich pacjentów (młodych i starszych razem):

- **Lek A ogółem:** $\frac{90(\text{młodzi})+10(\text{starsi})}{100(\text{młodzi})+10(\text{starsi})} = \frac{100}{110} \approx 90.9\%$ **skuteczności**
- **Lek B ogółem:** $\frac{10(\text{młodzi})+90(\text{starsi})}{10(\text{młodzi})+100(\text{starsi})} = \frac{100}{110} \approx 90.9\%$ **skuteczności**

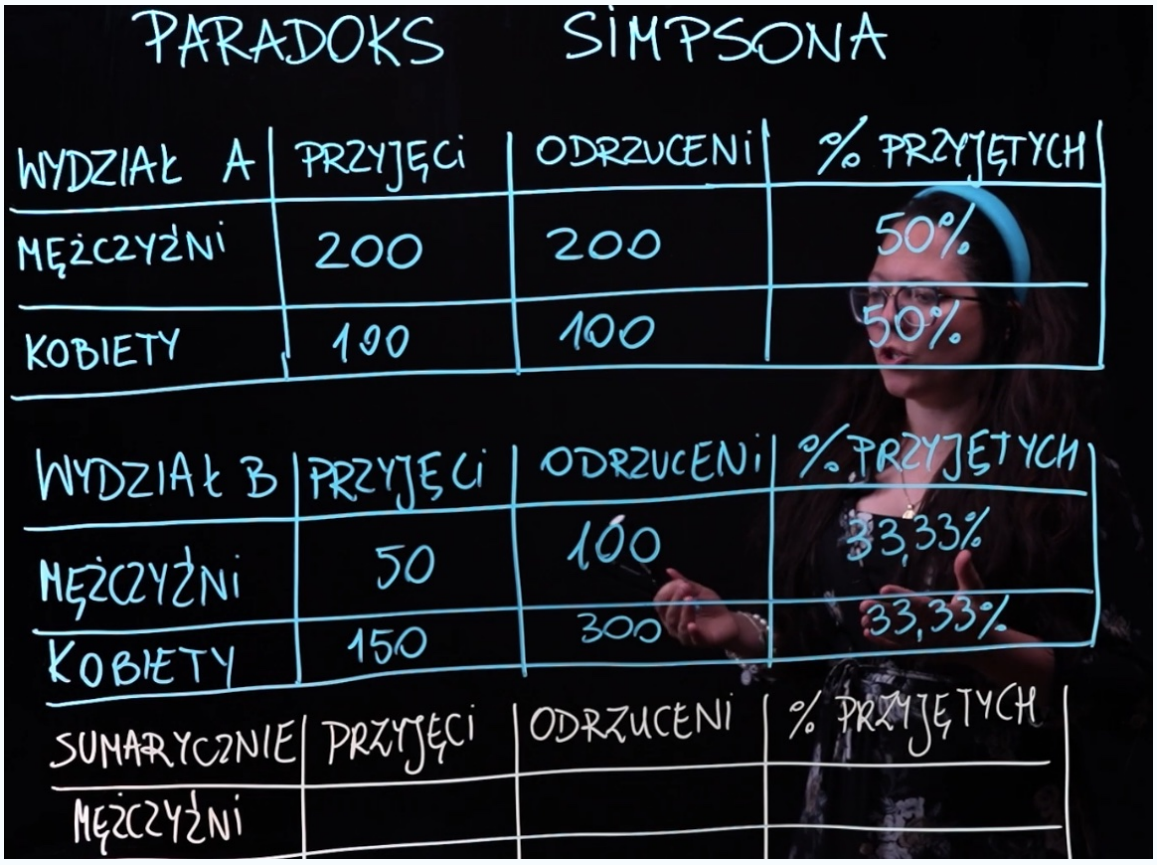
I tu pojawia się paradoks. Patrząc na dane ogólne, okazuje się, że **oba leki mają praktycznie taką samą skuteczność (około 90.9%)**! Ale przecież w każdej z podgrup (młodzi i starsi) jeden lek był wyraźnie lepszy od drugiego. Jak to możliwe, że po połączeniu danych różnice znikają? Sekret tkwi w **nierównym rozkładzie pacjentów w grupach**. W naszym przykładzie:

- Większość **młodych pacjentów** była leczona **Lekiem A**, mimo że Lek B był dla nich lepszy.
- Większość **starszych pacjentów** była leczona **Lekiem B**, mimo że Lek A był dla nich lepszy.

Lek A miał przeważnie gorsze wyniki wśród młodych, ale większość młodych go brała. Lek B miał gorsze wyniki wśród starszych, ale większość starszych go brała. Kiedy łączymy te dane, proporcje w grupach zniekształcają ogólny obraz. Mimo że Lek B był lepszy dla młodych, brało go tylko 10 młodych osób. Lek A był gorszy dla młodych, ale brało go 100 młodych osób. Podobnie jest ze starszymi. Ten **nierówny rozkład** pacjentów między leki w poszczególnych grupach wiekowych jest właśnie tym ukrytym czynnikiem, który prowadzi do paradoksu.

PRZYKŁAD

Przykład 47: Dyskryminacja na studiach



https://vimeo.com/1097826659/fc7eef0c4b
Paradoks Simpsona

10.2. Paradoks Monty'ego Halla

Autorzy/Autorki: Anna Serwatka

Paradoks Monty'ego Halla to klasyczna zagadka prawdopodobieństwa, która od lat fascynuje matematyków, statystyków i laików, często prowadząc do zdziwienia, a nawet niedowierzania. Nazwany na cześć prowadzącego amerykańskiego teleturnieju „Let's Make a Deal", w którym był prezentowany, paradoks ten ilustruje, jak nasze intuicyjne rozumienie szans może być mylące, gdy konfrontujemy się z warunkową naturą prawdopodobieństwa. Choć wydaje się prosty, kryje w sobie głęboką lekcję na temat tego, jak aktualizować nasze przekonania w świetle nowych informacji.

Nazwa paradoksu pochodzi bezpośrednio od Monty'ego Halla, wieloletniego gospodarza popularnego amerykańskiego teleturnieju telewizyjnego „Let's Make a Deal". Gra, która stała się podstawą paradoksu, była stałym elementem tego programu. Uczestnicy faktycznie stawali przed trzema drzwiami, wybierali jedne, a następnie Monty, który znał położenie nagrody (zazwyczaj samochodu), otwierał jedne z niewybranych drzwi, zawsze ujawniając bezwartościową nagrodę (często kozę). Następnie oferował uczestnikowi szansę zmiany początkowego wyboru.

Problem ten zyskał szeroką sławę w 1990 roku, kiedy to został opublikowany w rubryce „Ask Marilyn" w magazynie „Parade", prowadzonej przez Marilyn vos Savant, posiadaczkę jednego z najwyższych zmierzonych IQ. Jej poprawna odpowiedź, że zmiana wyboru podwaja szanse na wygraną, wywołała burzliwą debatę i tysiące listów od czytelników, w tym od wielu naukowców i matematyków, którzy upierali się, że Marilyn się myli. To masowe niezrozumienie i gorąca dyskusja utrwaliły nazwę „Paradoks Monty'ego Halla", czyniąc go jednym z najbardziej znanych i kontrowersyjnych problemów w teorii prawdopodobieństwa.

PRZYKŁAD

Przykład 48: Paradoks Monty'ego Halla i prawdopodobieństwo

Wyobraźmy sobie następującą sytuację, typową dla teleturnieju:

1. Stoimy przed trzema zamkniętymi drzwiami.
2. Za jednymi drzwiami znajduje się samochód (nagroda, którą chcemy wygrać).
3. Za pozostałymi dwoma drzwiami są kozy (brak nagrody).
4. Prowadzący, Monty Hall, wie, gdzie jest samochód.
5. Wybieramy jedne drzwi, na przykład drzwi nr 1.
6. Zanim je otworzymy, Monty (który wie, gdzie jest samochód) otwiera jedne z pozostałych, niewybranych przez nas drzwi, za którymi zawsze jest koza. Powiedzmy, że otwiera drzwi nr 3 i

ukazuje się koza.

7. Teraz Monty zadaje kluczowe pytanie: „Czy chcemy zmienić nasz wybór na drzwi nr 2, czy wolimy pozostać przy drzwiach nr 1?"

Intuicja wielu z nas podpowiada, że po odsłonięciu kozy szanse rozkładają się równomiernie między dwoma pozostałymi drzwiami, czyli wynoszą 50/50. Ale czy na pewno? Aby zrozumieć paradoks, musimy dokładnie przeanalizować prawdopodobieństwa w obu scenariuszach: pozostania przy pierwotnym wyborze i zmiany wyboru.

Scenariusz 1: Pozostajemy przy naszym pierwszym wyborze
Załóżmy, że wybieramy drzwi nr 1 i nie zmieniamy decyzji.

- **Początkowe prawdopodobieństwo:** Na początku mamy $\frac{1}{3}$ szans na wybranie drzwi z samochodem i $\frac{2}{3}$ szans na wybranie drzwi z kozą.
- **Jeśli wybraliśmy samochód (prawd. $\frac{1}{3}$):** Monty otworzy jedne z dwóch pozostałych drzwi z kozą. Pozostając przy swoim wyborze, wygrywamy samochód.
- **Jeśli wybraliśmy kozę (prawd. $\frac{1}{3}$ dla każdej kozy, czyli łącznie $\frac{2}{3}$):** Monty musi otworzyć drugie drzwi z kozą. Pozostając przy swoim wyborze, wygrywamy kozę.

Prawdopodobieństwo wygrania samochodu, pozostając przy pierwszym wyborze, wynosi $\frac{1}{3}$. Nasz początkowy wybór się nie zmienia, więc szanse pozostają takie same.

Scenariusz 2: Zmieniamy nasz wybór

Załóżmy, że wybieramy drzwi nr 1 i zawsze zmieniamy decyzję po tym, jak Monty otworzy drzwi z kozą.

- **Początkowe prawdopodobieństwo:** Nadal mamy $\frac{1}{3}$ szans na wybranie drzwi z samochodem i $\frac{2}{3}$ szans na wybranie drzwi z kozą.
- **Jeśli wybraliśmy samochód (prawd. $\frac{1}{3}$):** Monty otworzy jedne z pozostałych drzwi z kozą. Gdy zmieniamy swój wybór, musimy wybrać drugie drzwi z kozą. Wygrywamy kozę.
- **Jeśli wybraliśmy kozę (prawd. $\frac{2}{3}$):** Monty *musi* otworzyć drugie drzwi z kozą, czyli te, za którymi *nie ma* samochodu. Drzwi, na które zmieniamy wybór, *muszą* być drzwiami z samochodem. Wygrywamy samochód.

Prawdopodobieństwo wygrania samochodu, zmieniając swój wybór, wynosi $\frac{2}{3}$.

Dlaczego zmiana wyboru podwaja nasze szanse?

Kluczem do zrozumienia paradoksu jest to, że Monty Hall nie otwiera drzwi losowo. On zawsze wie, gdzie jest samochód, i zawsze otwiera drzwi z kozą spośród tych, których nie wybraliśmy.

Gdy początkowo wybieramy drzwi:

- Mamy $\frac{1}{3}$ szans, że wybraliśmy samochód.
- Mamy $\frac{2}{3}$ szans, że samochód jest za jednymi z *dwóch pozostałych* drzwi.

Kiedy Monty otwiera jedne z tych pozostałych drzwi, ujawniając kozę, tak naprawdę koncentruje całe $\frac{2}{3}$ prawdopodobieństwa na jednych pozostałych, zamkniętych drzwiach. On nie zmienia początkowych szans naszych wybranych drzwi (które nadal mają $\frac{1}{3}$ szans). Zamiast tego, eliminuje jedną z „kozich" opcji z puli niewybranych drzwi, co sprawia, że pozostałe niewybrane drzwi, które proponuje nam wybrać, stają się znacznie bardziej prawdopodobnym miejscem ukrycia samochodu.

Możemy to zilustrować w następujący sposób:

Wyobraźmy sobie, że zamiast trzech, mamy sto drzwi. Wybieramy jedne. Pozostaje 99 innych. Monty otwiera 98 z nich, wszystkie z kozami. Teraz stoimy przed dylematem: pozostać przy naszych pierwotnych drzwiach (szansa $\frac{1}{100}$) czy zmienić na te jedne pozostałe, które Monty nie otworzył. W tym scenariuszu staje się to znacznie bardziej intuicyjne: te jedne pozostałe drzwi z 99 miały skumulowane prawdopodobieństwo $\frac{99}{100}$ bycia drzwiami z samochodem, a Monty skutecznie wskazał nam, które to są, eliminując wszystkie pomyłki.

Paradoks Monty'ego Halla uczy nas, że prawdopodobieństwo jest dynamiczne i powinno być aktualizowane w miarę pojawiania się nowych informacji. Intuicja, która podpowiada 50/50, ignoruje fakt, że prowadzący dostarcza cenną informację poprzez swój celowy wybór drzwi do otwarcia. Zmiana wyboru w tej grze to zawsze najlepsza strategia, podwajająca nasze szanse na wygraną. To potężna lekcja o tym, jak ważne jest uwzględnianie wszystkich warunków i procesów w analizie statystycznej, aby uniknąć błędnych wniosków. Dla wielu Polaków teleturniej „Idź na całość" to coś więcej niż tylko program telewizyjny – to kawałek dzieciństwa i symbol niedzielnych wieczorów spędzanych przed telewizorem. Emitowany w latach 90. i na początku 2000., program zaskarbił sobie sympatię milionów widzów dzięki swojej nieprzewidywalności, emocjom i charyzmie prowadzącego, Zygmunta Chajzera. Każdy odcinek był niczym loteria, w której uczestnicy musieli wybierać między atrakcyjnymi nagrodami a wszechobecnym symbolem programu – Zonk'iem, czyli kozą ukrytą za jednymi z trzech drzwi. Ta prosta, ale genialna koncepcja, oparta była właśnie na dylematach podobnych do tych, które ilustruje Paradoks Monty'ego Halla. Uczestnicy, stojąc przed wyborem, czy pozostać przy pierwotnych drzwiach, czy zmienić decyzję po ujawnieniu Zonku, nieświadomie mierzyli się z tą zagadką prawdopodobieństwa.

10.3. Paradoks hazardzisty

Autorzy/Autorki: Anna Serwatka

Paradoks hazardzisty, często nazywany złudzeniem Monte Carlo, to powszechne, a zarazem błędne przekonanie, że przeszłe, niezależne zdarzenia wpływają na prawdopodobieństwo przyszłych wyników. Innymi słowy, mamy tendencję wierzyć, że jeśli coś dzieje się rzadziej niż normalnie w krótkim okresie, to wkrótce musi zdarzyć się częściej, aby „wyrównać" średnią. To błąd logiczny, ponieważ zdarzenia losowe nie mają pamięci. Każdy rzut kostką, obrót ruletką czy losowanie lotto to niezależne zdarzenie, a jego wynik nie zależy od tego, co wydarzyło się wcześniej. Podstawą paradoksu jest niezrozumienie **prawa wielkich liczb**. Prawo to mówi, że w bardzo długiej serii prób, wyniki zbliżą się do oczekiwanego prawdopodobieństwa. Nie oznacza to jednak, że krótkie serie odchyleń zostaną „skorygowane" przez przyszłe wyniki. Każda pojedyncza próba ma to samo prawdopodobieństwo, niezależnie od historii.

PRZYKŁAD

Przykład 49: Ruletka

Wyobraźmy sobie, że w kasynie kulka na ruletce wylądowała na czarnym polu dziesięć razy z rzędu. My, jako osoby podatne na paradoks hazardzisty, będziemy przekonani, że następny rzut musi wypaść na czerwone, bo „przecież czarne już tyle razy było". Rzeczywistość jest taka, że prawdopodobieństwo wylądowania na czerwonym (lub czarnym) w kolejnym rzucie nadal wynosi około $48,6\%$ (zakładając europejską ruletkę z jednym zerem), tak samo jak w każdym poprzednim rzucie. Każdy obrót jest nowym, niezależnym zdarzeniem.

PRZYKŁAD

Przykład 50: Rzut monetą

Jeśli rzucimy monetą dziesięć razy i za każdym razem wypadnie orzeł, paradoks hazardzisty skłoni nas do myślenia, że szanse na reszkę w kolejnym rzucie są znacznie wyższe. Jednak prawdopodobieństwo wyrzucenia reszki wciąż wynosi dokładnie 50%, ponieważ moneta nie „pamięta" poprzednich wyników.

PRZYKŁAD

Przykład 51: Inwestowanie na giełdzie

Czasami inwestorzy, widząc serię spadków cen akcji, błędnie zakładają, że „teraz to już musi odbić" i kupują, wierząc, że trend musi się odwrócić. Chociaż analiza fundamentalna i techniczna są kluczowe, ślepe bazowanie na przekonaniu o „wyrównaniu" losowych wahań jest formą paradoksu hazardzisty.

Paradoks hazardzisty wynika z naszej wrodzonej tendencji do poszukiwania wzorców i porządku w chaosie. Nasz mózg próbuje przewidywać, a kiedy widzi „serię", spodziewa się jej zakończenia. Ignorujemy jednak fundamentalną zasadę statystyki: niezależność zdarzeń. W hazardzie, gdzie wyniki są często całkowicie losowe, to złudzenie prowadzi do kosztownych błędów i utraty pieniędzy.

Zrozumienie paradoksu hazardzisty to klucz do racjonalniejszego myślenia o prawdopodobieństwie i uniknięcia pułapek, które nasza intuicja zastawia na drodze do zrozumienia losowych procesów.

10.4. Paradoks urodzinowy

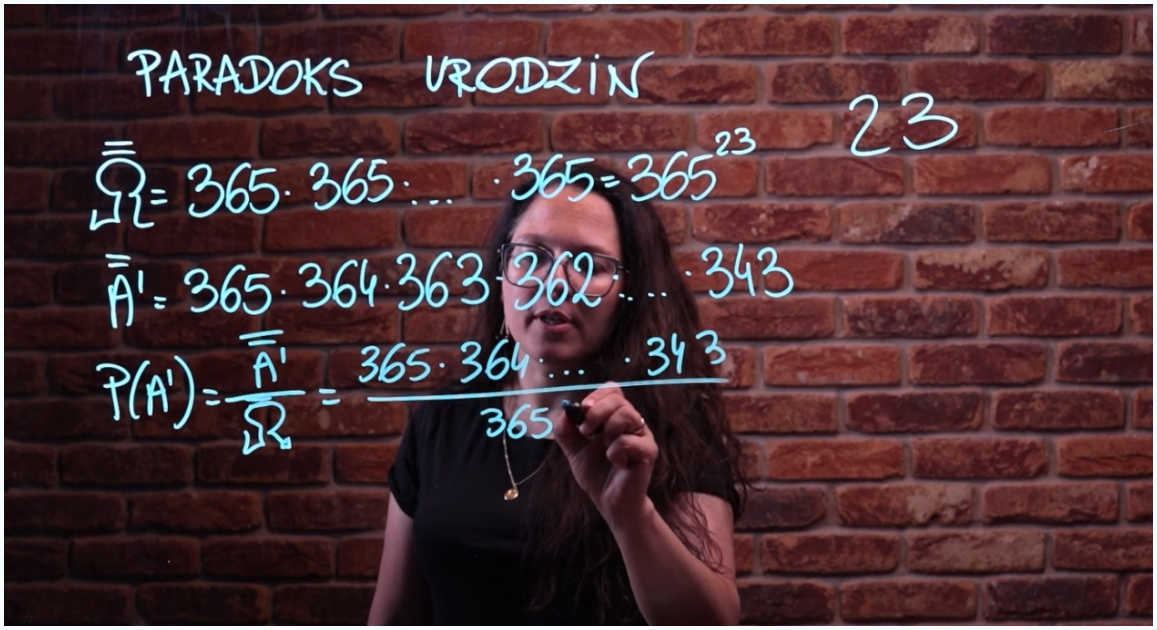
Autorzy/Autorki: Anna Serwatka

Paradoks urodzin (zwany też problemem urodzin) odnosi się do zaskakująco wysokiego prawdopodobieństwa, że w stosunkowo niewielkiej grupie ludzi znajdą się co najmniej dwie osoby, które obchodzą urodziny tego samego dnia. Kiedy po raz pierwszy o tym usłyszeliśmy, pomyśleliśmy, że to musi być błąd. Przecież rok ma 365 dni (pomijamy lata przestępne dla uproszczenia), więc aby dwie osoby miały te same urodziny, musielibyśmy mieć naprawdę dużą grupę, prawda? Okazuje się, że w grupie liczącej zaledwie 23 osoby prawdopodobieństwo, że co najmniej dwie z nich mają urodziny tego samego dnia, wynosi ponad 50%. W grupie wielkości małej klasy szkolnej, szanse na wspólne urodziny są większe niż 50%. A kiedy grupa rośnie do 57 osób, prawdopodobieństwo to przekracza już 99%. Te liczby kompletnie przeczą naszej intuicji, która podpowiadałaby nam, że potrzeba setek ludzi, aby tak się stało.

Sekret tego paradoksu tkwi w sposobie, w jaki obliczamy prawdopodobieństwo. Zazwyczaj, myśląc o urodzinach, skupiamy się na znalezieniu osoby, która ma urodziny w konkretnym dniu, albo na znalezieniu osoby, która ma urodziny tego samego dnia co my. Paradoks urodzin jednak pyta o coś innego: jakie jest prawdopodobieństwo, że jakiekolwiek dwie osoby w grupie dzielą ten sam dzień urodzenia?

PRZYKŁAD

Przykład 52: Paradoks urodzin



https://epodreczniki.open.agh.edu.pl/handbook/2369/module/2488/reader#openaghmod:2488:openaghhfivep:1

Paradoks urodzin

10.5. Paradoks więźnia

Autorzy/Autorki: Anna Serwatka

Paradoks więźnia, będący jednym z najbardziej znanych i fascynujących zagadnień w teorii gier, doskonale ilustruje sytuację, w której decyzja podjęta przez jedną osobę decyduje o losie kogoś innego, z kim nie możemy się porozumieć. Pokazuje, jak racjonalne, egoistyczne decyzje poszczególnych jednostek mogą prowadzić do nieoptymalnego, a nawet gorszego rezultatu dla całej grupy. To problem, który zmusza nas do refleksji nad naturą współpracy i konkurencji.

Paradoks Więźnia to fundamentalny scenariusz w teorii gier, który opisuje sytuację, w której dwie racjonalne jednostki, działające we własnym interesie i nie mające możliwości komunikacji, ostatecznie podejmują decyzje, które prowadzą do gorszego wyniku dla obu stron niż w przypadku, gdyby współpracowały. Pokazuje on dylemat między egoizmem a altruizmem, między maksymalizacją własnych korzyści a potencjalnymi korzyściami wynikającymi ze współpracy.

PRZYKŁAD

Przykład 53: Klasyczny przykład: dwóch podejrzanych

Wyobraźmy sobie następujący scenariusz, który jest klasycznym przykładem Paradoksu więźnia:

Dwaj przestępcy, Ania i Bartek, zostali aresztowani pod zarzutem popełnienia przestępstwa. Policja nie ma wystarczających dowodów, aby skazać ich za główne przestępstwo bez zeznań jednego z nich, ale ma wystarczające dowody, aby skazać ich obu za mniejsze przestępstwo (np. posiadanie nielegalnych przedmiotów).

Policja rozdziela Anię i Bartka do osobnych cel przesłuchań, tak aby nie mogli się ze sobą porozumiewać. Każdemu z nich oferuje następującą umowę:

1. Jeśli ty zeznajesz (zdradzasz) przeciwko drugiemu, a on milczy (współpracuje), to ty wychodzisz na wolność, a on dostaje 10 lat więzienia.
2. Jeśli oboje milczycie (współpracujecie), obaj dostajecie po 1 roku więzienia (za mniejsze przestępstwo).
3. Jeśli oboje zeznajecie (zdradzacie) przeciwko sobie, obaj dostajecie po 5 lat więzienia.
4. Jeśli ty milczysz (współpracujesz), a drugi zeznaje (zdradza), to ty dostajesz 10 lat więzienia, a drugi wychodzi na wolność.

Możemy to przedstawić w postaci macierzy wypłat:

	Bartek milczy	**Bartek zeznaje**
Ania milczy	Ania: 1 rok, Bartek: 1 rok	Ania: 10 lat, Bartek: wolność
Ania zeznaje	Ania: wolność, Bartek: 10 lat	Ania: 5 lat, Bartek: 5 lat

Przeanalizujmy, jak Ania (i analogicznie Bartek) będzie podejmować decyzje, działając racjonalnie i we własnym interesie, nie wiedząc, co zrobi druga strona:

- **Co jeśli Bartek milczy?**
  - Jeśli Ania też milczy, dostaje 1 rok.
  - Jeśli Ania zeznaje, wychodzi na wolność.
  - Dla Ani, lepszą opcją jest zeznawanie.
- **Co jeśli Bartek zeznaje?**
  - Jeśli Ania milczy, dostaje 10 lat.
  - Jeśli Ania zeznaje, dostaje 5 lat.
  - Dla Ani, lepszą opcją jest zeznawanie.

W obu możliwych scenariuszach (niezależnie od tego, co zrobi Bartek), dla Ani racjonalną i optymalną decyzją jest zeznawanie. Ta strategia, która jest optymalna niezależnie od strategii przeciwnika, nazywana jest strategią dominującą.

Bartek, myśląc w ten sam sposób, również dojdzie do wniosku, że dla niego najlepszym wyjściem jest zeznawanie. W rezultacie, zarówno Ania, jak i Bartek, działając racjonalnie i egoistycznie, zeznają przeciwko sobie. Ostatecznie oboje lądują w więzieniu na 5 lat.

Paradoks polega na tym, że gdyby Ania i Bartek współpracowali i oboje milczeli, każdy z nich dostałby tylko 1 rok więzienia. Wynik (5 lat dla każdego) jest gorszy dla obu stron niż wynik, który mogliby osiągnąć poprzez wzajemną współpracę (1 rok dla każdego).

Paradoks Więźnia jest potężnym narzędziem do analizowania wielu rzeczywistych sytuacji, w których brak zaufania i komunikacji prowadzi do suboptymalnych wyników:

- **W ekonomii:** firmy konkurujące na rynku (np. ustalające ceny, produkcję) mogą dążyć do maksymalizacji własnych zysków, co w efekcie prowadzi do wojen cenowych i niższych zysków dla

wszystkich.

- **W polityce międzynarodowej:** kraje podejmujące decyzje dotyczące zbrojeń. Każdy kraj może czuć się bezpieczniej, zwiększając swoje zbrojenia, ale w rezultacie wszyscy stają się mniej bezpieczni z powodu wyścigu zbrojeń.
- **W ekologii:** wspólne zasoby (np. łowiska, lasy). Każdy rybak/drwal może chcieć złowić/wyciąć jak najwięcej, co prowadzi do wyczerpania zasobów dla wszystkich.

Paradoks Więźnia uczy nas, że indywidualna racjonalność nie zawsze prowadzi do kolektywnego dobra. Podkreśla on znaczenie zaufania, komunikacji, i mechanizmów egzekwowania umów (np. kontrakty, prawo) w celu osiągnięcia optymalnych wyników dla społeczeństwa jako całości. Pokazuje, jak trudne bywa osiągnięcie współpracy, nawet gdy jest ona korzystna dla wszystkich zaangażowanych stron.

10.6. Paradoks Bertrand'a

Autorzy/Autorki: Anna Serwatka

Matematyka, ze swoją precyzją i logiką, często wydaje się nie pozostawiać miejsca na dwuznaczności. Jednak historia zna przypadki, gdy nawet najlepsi myśliciele natrafiali na paradoksy, które obnażały subtelne luki w naszym rozumieniu podstawowych pojęć. Jednym z takich fascynujących zagadnień jest Paradoks Bertranda, nazwany na cześć francuskiego matematyka Josepha Louisa François Bertranda. Został on po raz pierwszy przedstawiony w 1889 roku i do dziś stanowi doskonały przykład tego, jak brak ścisłej definicji może prowadzić do wielu poprawnych, lecz wzajemnie sprzecznych rozwiązań tego samego problemu z rachunku prawdopodobieństwa.

Paradoks Bertranda dotyczy prawdopodobieństwa wylosowania cięciwy w okręgu, której długość jest większa niż długość boku trójkąta równobocznego wpisanego w ten okrąg. Brzmi prosto, prawda?
Otóż problem polega na tym, że w zależności od metody, jaką wybierzemy do „losowania" tej cięciwy, otrzymujemy zupełnie różne wyniki prawdopodobieństwa. I co najbardziej intrygujące, każda z tych metod wydaje się równie sensowna i logiczna!

PRZYKŁAD

Przykład 54: Trzy metody losowania cięciwy w Paradoksie Bertrand'a

Wyobraźmy sobie okrąg o promieniu R. W jego wnętrzu rysujemy trójkąt równoboczny. Chcemy znaleźć prawdopodobieństwo, że losowo wybrana cięciwa będzie dłuższa niż bok tego trójkąta. Długość boku trójkąta równobocznego wpisanego w okrąg wynosi $R\sqrt{3}$. Bertrand przedstawił trzy różne, pozornie logiczne metody wyboru cięciwy, z których każda prowadzi do innego wyniku:

Metoda 1: Losowanie końców cięciwy na okręgu

Wybieramy dwa losowe punkty na obwodzie okręgu. Te dwa punkty tworzą cięciwę. Jeśli umieścimy w okręgu trójkąt równoboczny, to jego boki dzielą okrąg na trzy równe łuki. Cięciwa będzie dłuższa od boku trójkąta, jeśli jej końce znajdą się na tym samym łuku.

Jeśli ustalić jeden koniec cięciwy, powiedzmy w wierzchołku trójkąta, to drugi koniec musi wypaść na łuku naprzeciwko, który ma długość $\frac{1}{3}$ obwodu okręgu. Prawdopodobieństwo wynosi $\frac{1}{3}$.

Metoda 2: Losowanie punktu środkowego cięciwy w okręgu

Wybieramy losowy punkt wewnątrz okręgu. Ten punkt będzie środkiem cięciwy. Cięciwa będzie prostopadła do promienia przechodzącego przez ten punkt. Długość boku trójkąta równobocznego wpisanego w okrąg wynosi $R\sqrt{3}$. Cięciwa będzie dłuższa niż bok trójkąta, jeśli jej odległość od środka okręgu jest mniejsza niż połowa boku trójkąta równobocznego o podstawie $R\sqrt{3}$. Wysokość tego trójkąta to $R + \frac{R}{2} = \frac{3R}{2}$. Promień okręgu wpisanego w ten trójkąt ma długość $\frac{R}{2}$. Pole małego okręgu o promieniu $\frac{R}{2}$ do pola dużego okręgu o promieniu R. Jest to $\frac{\pi(R/2)^2}{\pi R^2} = \frac{1}{4}$.

Metoda 3: Losowanie promienia i punktu na promieniu

Jak w metodzie 2, cięciwa będzie dłuższa od boku trójkąta równobocznego, jeśli jej odległość od środka okręgu jest mniejsza niż $\frac{R}{2}$. Promień ma długość R. Punkt środkowy musi leżeć w pierwszej połowie promienia (odległość od środka jest mniejsza bądź równa $\frac{R}{2}$). Prawdopodobieństwo wynosi $\frac{1}{2}$.

Mamy więc trzy „logiczne" podejścia, dające trzy różne wyniki: $\frac{1}{3}$, $\frac{1}{4}$ i $\frac{1}{2}$. Gdzie tkwi błąd? Paradoks Bertranda nie jest błędem w obliczeniach matematycznych - jest to raczej ostrzeżenie dotyczące **nieprecyzyjnego sformułowania problemu**.

Kluczem do zrozumienia paradoksu jest to, że termin „losowo wybrana cięciwa" nie ma jednej, uniwersalnie akceptowanej matematycznej definicji. Każda z przedstawionych metod implikuje inną dystrybucję prawdopodobieństwa dla zbioru możliwych cięciw. Mówiąc inaczej, „losowanie cięciwy" może oznaczać różne rzeczy, a każda z tych interpretacji prowadzi do innego zbioru wyników.

- **Metoda 1** zakłada jednolitą dystrybucję punktów na obwodzie okręgu.
- **Metoda 2** zakłada jednolitą dystrybucję punktów wewnątrz okręgu.
- **Metoda 3** zakłada jednolitą dystrybucję punktów na promieniu.

Każda z tych metod jest matematycznie spójna i poprawna w ramach własnych założeń, ale te założenia są różne.

10.7. Błędy i pułapki w interpretacji danych statystycznych

Autorzy/Autorki: Anna Serwatka

W dzisiejszym świecie jesteśmy bombardowani danymi statystycznymi. Od wiadomości o gospodarce, przez badania medyczne, po analizy marketingowe – statystyka jest wszędzie. Daje nam ona potężne narzędzie do zrozumienia otaczającej nas rzeczywistości, ale jednocześnie kryje w sobie liczne pułapki, które mogą prowadzić do błędnych wniosków, a co za tym idzie – do złych decyzji. Zrozumienie najczęstszych błędów w interpretacji danych statystycznych jest kluczowe dla każdego, kto chce wyciągać rzetelne i wartościowe wnioski.

Jedną z najbardziej podstępnych pułapek jest błąd selekcji, czyli sytuacja, gdy dane, które analizujemy, nie są reprezentatywne dla całej populacji, którą chcemy badać. Odmianą tego jest błąd przeżywalności.

PRZYKŁAD

Przykład 55: Błąd selekcji i błąd przeżywalności

Podczas II wojny światowej matematyk Abraham Wald analizował uszkodzenia samolotów wracających z misji, aby określić, gdzie wzmocnić opancerzenie. Początkowo chciano wzmocnić miejsca z

największą liczbą dziur po kulach. Wald jednak zauważył, że analizują tylko te samoloty, które przeżyły. Samoloty z dziurami w krytycznych miejscach (silnik, kokpit) po prostu nie wracały. Oznaczało to, że to właśnie te miejsca, które były nienaruszone w powracających maszynach, wymagały wzmocnienia, ponieważ ich uszkodzenie prowadziło do zestrzelenia. Skupianie się na „ocalałych" danych (samolotach, które wróciły) doprowadziło do błędnych wniosków.

Zawsze zastanówmy się, czy dane, które analizujemy, pochodzą z pełnego i reprezentatywnego zbioru. Co mogło zostać pominięte i dlaczego? Korelacja oznacza, że dwie zmienne zmieniają się razem. Korelacja nie implikuje przyczynowości! To jedno z najczęstszych i najbardziej szkodliwych błędów. Możemy zaobserwować silny związek między dwoma zjawiskami, ale to nie znaczy, że jedno powoduje drugie.

PRZYKŁAD

Przykład 56: Fałszywa korelacja

Zaobserwowano bardzo silną korelację między liczbą rozwodów w stanie Maine a konsumpcją margaryny. Oczywiście, jedzenie margaryny nie powoduje rozwodów (ani odwrotnie). To czysty przypadek, lub obie zmienne są pod wpływem jakiegoś trzeciego, ukrytego czynnika (np. ogólnych trendów społecznych lub ekonomicznych).

Często na pozór oczywisty związek między dwiema zmiennymi jest w rzeczywistości kontrolowany przez trzecią zmienną, której nie bierzemy pod uwagę.

PRZYKŁAD

Przykład 57: Pominięcie zmiennych ukrytych

Badanie pokazuje, że studenci, którzy piją więcej kawy, osiągają lepsze wyniki w nauce. Czy to oznacza, że kawa sprawia, że jesteśmy mądrzejsi? Niekoniecznie. Zmienną ukrytą może być fakt, że studenci z dobrymi wynikami, aby zdążyć na czas z nauką, często uczą się do późna i potrzebują kawy, żeby utrzymać koncentrację. To nie kawa jest przyczyną, a intensywna nauka.

Zawsze warto zadać sobie pytanie: „Czy jest coś jeszcze, co może wpływać na te wyniki, a czego nie uwzględniam?". Wyciąganie wniosków na podstawie zbyt małej liczby obserwacji jest bardzo ryzykowne. Małe próby są bardziej podatne na przypadkowe fluktuacje, a wyniki uzyskane z nich mogą nie być reprezentatywne dla większej populacji.

PRZYKŁAD

Przykład 58: Błąd zbyt małej próby

Badanie przeprowadzone na 5 osobach wykazało, że nowy produkt jest genialny. Jednak grupa docelowa to miliony konsumentów. Wyniki z tak małej próby są statystycznie bezwartościowe.

Agregacja danych może ukrywać ważne wzorce lub nawet odwracać kierunek zależności. Paradoks Simpsona to zjawisko, w którym trend obserwowany w kilku grupach danych zanika lub odwraca się, gdy te grupy zostaną połączone.

PRZYKŁAD

Przykład 59: Agregacja danych a Paradoks Simpsona

Uczelnia została oskarżona o dyskryminację kobiet, ponieważ analiza ogólnych danych wykazała, że procent przyjętych mężczyzn był wyższy niż kobiet. Jednak gdy dane podzielono według wydziałów, okazało się, że na większości wydziałów procent przyjętych kobiet był wyższy lub równy mężczyzn, a na kilku pozostałych różnica była minimalna. Problem polegał na tym, że kobiety częściej aplikowały na wydziały z niższą ogólną stopą przyjęć.

Ten przykład pokazuje, że czasami warto dekomponować dane i analizować je na różnych poziomach agregacji, aby uniknąć ukrytych pułapek. Interpretacja danych statystycznych to sztuka i nauka. Nie wystarczy umieć liczyć; trzeba także rozumieć kontekst, potencjalne źródła błędów i ograniczenia metodologii. Świadomość tych pułapek pozwala nam na bardziej krytyczne podejście do przedstawianych nam liczb i wyciąganie bardziej rzetelnych wniosków. Liczby same w sobie są obiektywne, ale sposób ich zbierania, analizowania i interpretowania jest już bardzo subiektywny i podatny na błędy.

Bibliografia

[1] M. Kurczab, E. Kurczab, E. Świda. *Podręcznik do liceów i techników. Zakres rozszerzony. Klasa 3*. Warszawa: Oficyna Edukacyjna Krzysztof Pazdro Sp. z o.o., 2023.

[2] P. Drygaś, P. Pusz. *Elementy rachunku prawdopodobieństwa*. Rzeszów: Wyd.Uniwersytetu Rzeszowskiego, 2022.

[3] W. Guzicki. *Kombinatoryka*. Tarnów: Omega, 2024.

[4] R.L. Graham, D.E. Knuth, O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Reading, MA: Addison-Wesley, 1994.

[5] A. Plucińska, E. Pluciński. *Probabilistyka Statystyka matematyczna Procesy stochastyczne Rachunek prawdopodobieństwa*. Warszawa: Wydawnictwo Naukowe PWN, WNT, 2017.

[6] W. Krysicki, J. Bartos, W. Dyczka. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach. Część 1*. Warszawa: Wydawnictwo Szkolne PWN, 2012.

[7] W. Regel. *101 zadań ze statystyki matematycznej z pełnymi rozwiązaniami krok po kroku*. Rzeszów: Bila, 2012.

[8] Innowacji US – Departament. *Pojęcia stosowane w statystyce publicznej*. URL: https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/2950,pojecie.html.

[9] Innowacji US – Departament. *Pojęcia stosowane w statystyce publicznej*. URL: https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/2912,pojecie.html.

[10] J. Greń. *Modele i zadania statystyki matematycznej*. Warszawa: Państwowe Wydawnictwo Naukowe, 1970.