KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

AKADEMIA MORSKA SZCZECIN

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Rachunek prawdopodobieństwa
I
Elementy statystyki

Materiały do zajęć audytoryjnych

Lech KASYK

Spis treści

W wielu zagadnieniach życia codziennego zastanawiamy się nad szansami wystąpienia pewnych zdarzeń. Na przykład: jaka jest szansa wygrania miliona w toto lotka, jakie jest prawdopodobieństwo awarii silnika okrętowego, jakie są szanse zdobycia złotego medalu mistrzostw świata przez drużynę Polski. Szacowaniem tych „szans" zajmuje się właśnie rachunek prawdopodobieństwa.

Rachunek prawdopodobieństwa zajmuje się doświadczeniami losowymi czyli takimi, których wyniku nie da się przewidzieć. Jednym z najprostszych doświadczeń losowych jest rzut kostką. Wynikiem tego doświadczenia może być „jedno oczko" , „dwa oczka", „trzy oczka", „cztery oczka", „pięć oczek" lub „sześć oczek", nigdy jednak z całkowitą pewnością nie możemy powiedzieć, że w danym rzucie wypadnie „jedno oczko". Podobnie jest w przypadku, gdy na punkcie raportowym systemu VTS mierzymy czas pomiędzy przejściami kolejnych jednostek torowych. Nie jesteśmy w stanie przewidzieć jaki będzie czas dla następnych statków.

Rachunek prawdopodobieństwa jest działem matematyki, który na początku kursu wymaga jedynie podstawowych wiadomości, takich jak działania na zbiorach czy działania na liczbach wymiernych. Ponadto cechą charakterystyczną tego działu matematyki jest częste stosowanie opisów słownych na określenie wyników różnych doświadczeń losowych.

Rozdział I. Podstawowe pojęcia rachunku prawdopodobieństwa

Każde doświadczenie losowe kończy się jakimś wynikiem. W rzucie monetą może to być „ wyrzucenie orła", w rzucie kostką może to być „ wyrzucenie jednego oczka", w losowaniu jednej liczby ze zbioru {1, 6, 8, 11} może to być „wylosowanie 1". Każdy pojedynczy wynik doświadczenia losowego nazywamy ***zdarzeniem elementarnym***. Zbiór wszystkich zdarzeń elementarnych nazywamy ***przestrzenią zdarzeń elementarnych*** lub ***przestrzenią wyników***. Oznaczamy ją najczęściej grecką literą Ω.

Przykład 1
Doświadczenie polega na rzucie kostką. Przestrzeń zdarzeń elementarnych składa się z 6 elementów: „jedno oczko" , „dwa oczka", „trzy oczka", „cztery oczka", „pięć oczek", „sześć oczek". Przeważnie korzystamy z jakiegoś krótszego zapisu, np. $\Omega = \{1, 2, 3, 4, 5, 6\}$

Przykład 2
Doświadczenie polega na trzykrotnym rzucie monetą. Zbiór Ω składa się z 8 zdarzeń elementarnych, którymi są 3-elementowe ciągi (x, y, z), gdzie każdy element tego ciągu oznacza wynik rzutu monetą (*o* lub *r*). $\Omega = \{(o,o,o), (o,o,r),(o,r,o), (r,o,o), (o,r,r), (r,o,r), (r,r,o), (r,r,r)\}$

Przykład 3
Doświadczenie polega na ustawieniu trzech liter: A, B, C w dowolnej kolejności. Przestrzenią wyników jest zbiór: $\Omega = \{ABC, ACB, BCA, BAC, CAB, CBA\}$

Każdy podzbiór przestrzeni wyników nazywamy ***zdarzeniem***. O każdym zdarzeniu elementarnym będącym elementem zdarzenia *Z*, mówimy, że sprzyja zdarzeniu *Z*.

Przykład 4
Dla doświadczenia z przykładu 2, niech *A* oznacza zdarzenie polegające na tym, że dokładnie raz wypadnie orzeł. Zdarzeniu *A* odpowiadają 3 zdarzenia elementarne:
$A = \{(o,r,r), (r,o,r), (r,r,o)\}$

Przykład 5
Dla doświadczenia polegającego na dwukrotnym rzucie kostką, niech *B* oznacza zdarzenie polegające na tym, że suma wyrzuconych oczek jest większa od 9:
$B = \{(6,4), (6,5), (6,6), (5,6), (5,5), (4,6)\}$

Na zdarzeniach, jako że są zbiorami, można wykonywać różne działania, takie jak suma, iloczyn czy różnica.

Przykład 6
W trzykrotnym rzucie monetą, niech *A* oznacza zdarzenie: wypadła dwa razy reszka. Stąd
$A=\{(o,r,r), (r,o,r), (r,r,o)\}$.
Niech *B* oznacza zdarzenie: za drugim razem wypadł orzeł. $B=\{(o,o,o), (o,o,r), (r,o,o), (r,o,r)\}$.

Sumą zdarzeń *A* i *B* jest złączenie zbiorów *A* i *B* , czyli zbiór:
$A \cup B = \{(o,o,o), (o,o,r), (r,o,o), (o,r,r), (r,o,r), (r,r,o)\}$.

Natomiast iloczynem zdarzeń *A* i *B* jest zbiór, który jest częścią wspólną zbiorów *A* i *B*:
$A \cap B=\{(o,r,r), (r,o,r), (r,r,o)\} \cap \{(o,o,o), (o,o,r), (r,o,o), (r,o,r)\} =\{ (r,o,r)\}$

Różnicą zdarzeń *A* i *B* są te zdarzenia elementarne, sprzyjające zdarzeniu *A,* które nie sprzyjają zdarzeniu *B* czyli zbiór:
$A-B = \{(o,o,o), (o,o,r), (r,o,o), (o,r,r), (r,o,r), (r,r,o)\}$.

Zdarzeniem przeciwnym do zdarzenia *B* określonego w przykładzie 9, jest zdarzenie
$B' = \Omega - B = \{(o,o,o), (o,o,r),(o,r,o), (r,o,o), (o,r,r), (r,o,r), (r,r,o), (r,r,r)\} - \{(o,o,o), (o,o,r), (r,o,o), (r,o,r)\} = \{(o,r,o), (o,r,r), (r,r,o), (r,r,r)\}$,
tzn. *B*' jest zdarzeniem polegającym na tym, że za drugim razem nie wypadł orzeł (inaczej mówiąc *B*' jest zdarzeniem: za drugim razem wypadła reszka).

Klasyczna definicja prawdopodobieństwa

Jeżeli w przestrzeni zdarzeń elementarnych Ω jest skończenie wiele wyników i wszystkie są jednakowo prawdopodobne, do wyznaczania prawdopodobieństw możemy zastosować tak zwany model klasyczny.
Prawdopodobieństwem danego zdarzenia *A* nazywamy iloraz liczby zdarzeń elementarnych odpowiadających zdarzeniu *A*, przez liczbę wszystkich zdarzeń elementarnych. Mamy więc:

$$P(A) = \frac{n(A)}{n(\Omega)}$$

gdzie $n(A)$ oznacza liczbę zdarzeń sprzyjających zdarzeniu *A*, a $n(\Omega)$ liczbę wszystkich zdarzeń elementarnych.

Przykład 7

Dla zdarzenia *A* z przykładu 4, $n(A)=3$ i $n(\Omega)=8$, stąd $P(A)=\frac{n(A)}{n(\Omega)}=\frac{3}{8}$

Dla zdarzenia *B* z przykładu 5, $n(B)=6$ i $n(\Omega)=36$, stąd $P(B)=\frac{n(B)}{n(\Omega)}=\frac{6}{36}=\frac{1}{6}$

Z powyższej definicji wynika, że prawdopodobieństwo jest liczbą mniejszą bądź równą 1. Wynika to stąd, że liczba zdarzeń sprzyjających danemu zdarzeniu nie może przekroczyć liczby wszystkich zdarzeń elementarnych.

Z drugiej strony, prawdopodobieństwo jest liczbą większą bądź równą 0. Wynika to stąd, że liczba zdarzeń sprzyjających danemu zdarzeniu nie może być liczbą ujemną.

Zdarzenie, którego prawdopodobieństwo jest równe 1, nazywamy ***zdarzeniem pewnym***. A zdarzenie, którego prawdopodobieństwo jest równe 0, nazywamy ***zdarzeniem niemożliwym***.

Zdarzenie *A'*, któremu odpowiadają wszystkie zdarzenia elementarne, które nie odpowiadają zdarzeniu *A* nazywamy ***zdarzeniem przeciwnym*** do *A*.

Elementy kombinatoryki

W związku z tym, że w klasycznym podejściu do prawdopodobieństwa liczymy liczbę zdarzeń elementarnych, niezbędne stają się różne sposoby obliczania możliwych wyników poszczególnych doświadczeń. A więc teraz zajmiemy się elementami kombinatoryki, które posłużą temu celowi.

Jeżeli *n* jest liczba naturalną, to symbol ***n!*** (*n* silnia) określamy w następujący sposób:

$$\begin{cases} 0!=1 \\ n!=1\cdot 2\cdot 3\cdot \ldots \cdot n \quad \text{dla } n\geq 1 \end{cases}$$

Symbolem Newtona nazywamy wyrażenie $\binom{n}{k}$, które czytamy „*n po k*", a obliczamy następująco:

$$\binom{n}{k}=\frac{n!}{k!\cdot(n-k)!}$$

Permutacją zbioru *n* – elementowego nazywamy każdy *n* – wyrazowy ciąg utworzony ze wszystkich elementów tego zbioru. Liczbę permutacji zbioru *n* – elementowego wyrażamy wzorem:

$$P_n = n!$$

Kombinacją *k* – elementową ze zbioru *n* – elementowego nazywamy dowolny podzbiór *k* – elementowy danego zbioru *n* – elementowego. Liczbę takich kombinacji wyrażamy wzorem:

$$C_n^k=\binom{n}{k}=\frac{n!}{k!\cdot(n-k)!}$$

Wariacją *k* – elementową ***bez powtórzeń*** ze zbioru *n* – elementowego nazywamy każdy *k* – wyrazowy ciąg o różnych wyrazach, należących do danego zbioru *n* – elementowego. Liczbę takich wariacji wyrażamy wzorem:

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

Wariacją *k* – elementową ***z powtórzeniami*** ze zbioru *n* – elementowego nazywamy każdy *k* – wyrazowy ciąg o wyrazach należących do danego zbioru *n* – elementowego. Liczbę takich wariacji wyrażamy wzorem:

$$W_n^k = n^k$$

Przykład 8
Doświadczenie polega na pięciokrotnym rzucie monetą. Zbiór Ω składa się ze zdarzeń elementarnych, którymi są 5-elementowe ciągi (x, y, z, u, t), gdzie każdy element tego ciągu oznacza wynik rzutu monetą (*o* lub *r*). Wynika z tego, że liczba wszystkich zdarzeń elementarnych jest równa liczbie pięcioelementowych wariacji z powtórzeniami ze zbioru dwuelementowego {*o, r*}, czyli
$n(\Omega) = W^5_2 = 2^5=32$.

Przykład 9
Doświadczenie polega na wylosowaniu trzech studentów z 20 – osobowej grupy. Zbiór Ω składa się ze zdarzeń elementarnych, którymi są trzyelementowe podzbiory zbioru 20 – elementowego. Stąd liczba wszystkich zdarzeń elementarnych jest równa liczbie kombinacji trzyelementowych ze zbioru 20 – elementowego, czyli

$$n(\Omega) = C_{20}^3 = \binom{20}{3} = 1140$$

Geometryczna definicja prawdopodobieństwa

Nie wszystkie doświadczenia mają skończoną czy przeliczalną liczbę wyników. W takich przypadkach model klasyczny jest nieodpowiedni, bo nie można policzyć zdarzeń elementarnych. Jeżeli przestrzeń wyników jest zbiorem, który można w jakiś sposób zmierzyć, to można wtedy zastosować geometryczny model prawdopodobieństwa.
Jeżeli Ω jest obszarem w R^n o skończonej mierze i prawdopodobieństwo trafienia w obszar $A \subset \Omega$ zależy tylko od miary obszaru A i nie zależy od położenia obszaru A wewnątrz obszaru Ω, to

$$P(A) = \frac{\mathrm{m}A}{\mathrm{m}\Omega}$$

gdzie m*A* oznacza miarę zbioru *A*, a mΩ miarę zbioru Ω.

Przykład10
Losujemy jedną liczbę z przedziału od 0 do 10. Jakie jest prawdopodobieństwo zdarzenia *A*, że będzie to liczba mniejsza od 2?

Wszystkich liczb rzeczywistych w przedziale $\langle 0, 10\rangle$ jest nieskończenie wiele, dlatego klasyczna definicja jest nie do zastosowania. Można jednak zmierzyć długość tego przedziału, wynosi ona 10 jedno-

stek. Natomiast zdarzeniu *A*, odpowiada przedział $\langle 0,2)$, którego długość wynosi 2. Stąd prawdopodobieństwo zdarzenia *A* wynosi

$$P(A)=\frac{\mathrm{m}A}{\mathrm{m}\Omega}=\frac{2}{10}$$

Definicja aksjomatyczna prawdopodobieństwa

Podejście klasyczne czy podejście geometryczne jakkolwiek często stosowane nie wyczerpują wszystkich możliwych sytuacji, w których stosuje się prawdopodobieństwo. Dlatego została stworzona aksjomatyczna definicja prawdopodobieństwa, będąca podstawą całego rachunku prawdopodobieństwa.

Dana jest przestrzeń zdarzeń elementarnych Ω oraz wyróżniona rodzina zdarzeń losowych *M*.
Prawdopodobieństwem nazywamy funkcję określoną na rodzinie *M* o wartościach należących do zbioru liczb rzeczywistych *R*, która spełnia następujące aksjomaty:

Aksjomat 1
Dla dowolnego zdarzenia $A\in M$ prawdopodobieństwo $P(A)$ spełnia nierówność: $0\le P(A)\le 1$

Aksjomat 2
Prawdopodobieństwo *zdarzenia pewnego* Ω jest równe jedności: $P(\Omega)=1$.

Aksjomat 3
Prawdopodobieństwo sumy przeliczalnej liczby zdarzeń wyłączających się parami $(A_i\cap A_j=\varnothing,\ i\neq j)$ jest równe sumie prawdopodobieństw tych zdarzeń:

$$P\left(\bigcup_{k=1}^{\infty}A_k\right)=\sum_{k=1}^{\infty}P(A_k)$$

Własności prawdopodobieństwa

Na podstawie powyższych aksjomatów można wykazać, że prawdopodobieństwo ma następujące własności:

1. Jeżeli $A\subset B$, to $P(A)\le P(B)$
2. $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
3. $P(A')=1-P(A)$

Udowodnimy tu własność 3. Z definicji zdarzenia przeciwnego wiadomo, że $A'\cup A=\Omega$ i zdarzenia A' i A są rozłączne. Stąd $P(A'\cup A)=P(\Omega)=1$. Z aksjomatu 3 wynika, że $P(A'\cup A)=P(A')+P(A)$ czyli
$P(A')+P(A)=P(A'\cup A)=P(\Omega)=1\Rightarrow P(A')=1-P(A)$.

Zdarzenia niezależne

1. Niezależność pary zdarzeń:

Zdarzenia A i B nazywamy ***niezależnymi*,** jeżeli zachodzi równość

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

2. Niezależność n zdarzeń $(n \geq 2)$

Zdarzenia $A_1, A_2, ..., A_n$ nazywamy niezależnymi, jeżeli dla każdej liczby naturalnej $k \leq n$ i dowolnego skończonego ciągu liczb naturalnych $i_1, i_2, ..., i_k$ spełniających nierówności $i_1 \leq i_2 \leq ... \leq i_k \leq n$ zachodzi wzór

$$P\left(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}\right) = P\left(A_{i_1}\right) P\left(A_{i_2}\right) ... P\left(A_{i_k}\right)$$

Przykład 11

Doświadczenie polega na czterokrotnym rzucie monetą. Niech A oznacza zdarzenie: dwa razy wypadła reszka, a zdarzenie B: za drugim razem wypadł orzeł. Przestrzeń zdarzeń elementarnych składa się z 16 elementów:

Ω={*(o,o,o,o), (o,o,o,r),(o,o,r,o), (o,r,o,o),(r,o,o,o), (o,o,r,r), (o,r,o,r), (r,o,o,r),(r,o,r,o), (o,r,r,o), (r,r,o,o), (r,r,r,o), (r,r,o,r), (r,o,r,r), (o,r,r,r), (r,r,r,r)*}.

Natomiast zdarzenia *A* i *B* są następującymi zbiorami:

A = { *(o,o,r,r), (o,r,o,r), (r,o,o,r),(r,o,r,o), (o,r,r,o), (r,r,o,o)*}

B= {*(o,o,o,o), (o,o,o,r),(o,o,r,o),(r,o,o,o), (o,o,r,r), (r,o,o,r),(r,o,r,o), (r,o,r,r)*}

Zdarzenie *A*∩*B* jest częścią wspólna zbiorów *A* i *B* czyli *A*∩*B* = {*(o,o,r,r), (r,o,o,r),(r,o,r,o)*}

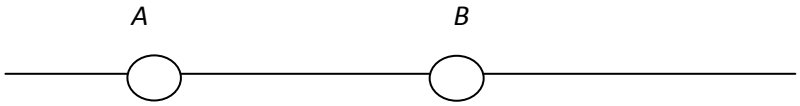
Stąd otrzymujemy prawdopodobieństwa:

P(*A*) = 6/16, P(*B*) = 8/16 i P(*A*∩*B*) = 3/16.

Iloczyn prawdopodobieństw: P(*A*)·P(*B*) = 6/16 · 8/16 = 3/16 jest równy prawdopodobieństwu P(*A*∩*B*), stąd wnioskujemy, że zdarzenia *A* i *B* są niezależne.

Przykład 12

Pewien układ składa się z dwóch elementów połączonych szeregowo. Prawdopodobieństwo awarii pojedynczego elementu wynosi 0,3. Jeżeli założymy, że elementy działają niezależnie od siebie, czyli działanie jednego elementu nie wpływa na działanie drugiego, możemy obliczyć prawdopodobieństwo awarii całego układu. Oznaczając elementy literami *A* i *B*, niech P(*A*) i P(*B*) oznaczają prawdopodobieństwo awarii tych elementów.



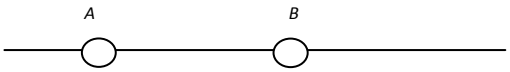
Przy takim połączeniu cały układ przestanie działać, gdy „popsuje się" chociaż jeden z elementów czyli *A* lub *B*. Prawdopodobieństwo awarii całego układu P(*U*) jest więc równe P(*A*∪*B*). Z własności

prawdopodobieństwa wynika, że $P(A\cup B)=P(A)+P(B)-P(A\cap B)$. Ponadto awarie elementów *A* i *B* są zdarzeniami niezależnymi. Stąd

$$P(A\cup B)=P(A)+P(B)-P(A)\cdot P(B)=0,3+0,3-0,3\cdot 0,3=0,51$$

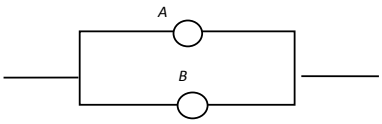
Można odwrócić sytuację i badać prawdopodobieństwo prawidłowego zadziałania układu czyli inaczej mówiąc jego niezawodność (tutaj wspominamy o niezawodności tylko w bardzo wąskim zakresie).

Przykład 13
Oblicz niezawodność układu szeregowego i równoległego dwóch elementów *A* i *B*, przy założeniu, że działają one niezależnie i niezawodność każdego z nich wynosi *q=0,9.*



W przypadku połączenia szeregowego (rysunek powyżej), żeby działał cały układ muszą działać oba elementy, czyli

$$N=P(A\cap B)=P(A)\cdot P(B)=q\cdot q=0,9\cdot 0,9=0,81$$



W przypadku połączenia równoległego(rysunek powyżej), żeby działał cały układ musi działać element *A* lub element *B*, czyli

$$N=P(A\cup B)=P(A)+P(B)-P(A)\cdot P(B)=0,9+0,9-0,9\cdot 0,9=0,99$$

Prawdopodobieństwo warunkowe

Jeżeli $P(A)>0$, to ***prawdopodobieństwem warunkowym*** zajścia zdarzenia *B* pod warunkiem zajścia zdarzenia *A* nazywamy liczbę: $\frac{P(A\cap B)}{P(A)}$, którą oznaczamy symbolem: $P(B|A)$ czyli

$$P(B|A)=\frac{P(A\cap B)}{P(A)}$$

Przykład 14
Prawdopodobieństwo warunkowe można dobrze zilustrować przy pomocy tabeli liczebności zdarzeń.

	B	D
A	200	100
C	250	300

Liczby w powyższej tabeli oznaczają liczby wspólnych zdarzeń elementarnych, czyli 200 to liczba zdarzeń elementarnych odpowiadających zdarzeniu $A\cap B$, a 100 to liczba zdarzeń elementarnych odpowiadających zdarzeniu $A\cap D$. Stąd liczba zdarzeń elementarnych odpowiadających zdarzeniu *A* wynosi

300 (200+100). A liczba zdarzeń elementarnych odpowiadających zdarzeniu *B* wynosi 450 (200 dla $A \cap B$ + 250 dla $B \cap C$). Natomiast wszystkich zdarzeń elementarnych jest w sumie 850. Wykorzystując częstość zdarzenia do oszacowania jego prawdopodobieństwa, można wyznaczyć prawdopodobieństwa poszczególnych zdarzeń. I tak $P(A)=\frac{300}{850}$, $P(A \cap B)=\frac{200}{850}$. Stąd prawdopodobieństwo warunkowe zajścia zdarzenia *B* pod warunkiem zajścia zdarzenia *A* wynosi:

$$P(B|A)=\frac{P(A \cap B)}{P(A)}=\frac{\frac{200}{850}}{\frac{300}{850}}=\frac{200}{300}=\frac{2}{3}$$

Na podstawie powyższej tabeli możemy wyznaczyć inne prawdopodobieństwa warunkowe:

$$P(A|B)=\frac{P(B \cap A)}{P(B)}=\frac{\frac{200}{850}}{\frac{450}{850}}=\frac{200}{450}=\frac{4}{9},\ P(A|D)=\frac{P(D \cap A)}{P(D)}=\frac{\frac{100}{850}}{\frac{400}{850}}=\frac{100}{400}=\frac{1}{4}$$

Prawdopodobieństwo całkowite

Jeżeli zdarzenia $A_1, A_2, \ldots, A_n$ wyłączają się parami i żadne z nich nie jest zdarzeniem niemożliwym oraz suma zdarzeń A_i jest zdarzeniem pewnym (tzn. $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$), to wówczas dla dowolnego zdarzenia *B* zachodzi wzór:

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

Jest to wzór na tak zwane ***prawdopodobieństwo całkowite***.

Stosuje się je często w sytuacjach, w których badane elementy pochodzą z kilku rozłącznych grup (np.: fabryki, klasy, miasta, wydziały), w których wspólna cecha występuje z różnym natężeniem.

Przykład 15
45% floty pewnego armatora stanowią chemikaliowce, 20% tankowce, a reszta to masowce. Spośród chemikaliowców połowa jest wyposażona w system PNS (Pilotowy System Nawigacyjny). Spośród tankowców 60% ma PNS. Natomiast spośród masowców, co piąty ma PNS. Obliczyć prawdopodobieństwo, że losowo wybrany statek tego armatora ma PNS.

Mamy trzy rozłączne grupy statków: chemikaliowce (A_1), tankowce (A_2) i masowce (A_3). Z treści zadania wynika, że $P(A_1)=0,45$, $P(A_2)=0,2$ i $P(A_3)=0,35$. Cechą wspólną jest posiadanie PNS (zdarzenie B). Z treści zadania wynika, że w grupie chemikaliowców prawdopodobieństwo posiadania PNS wynosi 0,5 czyli inaczej mówiąc prawdopodobieństwo, tego że statek ma PNS, pod warunkiem, że jest chemikaliowcem wynosi: $P(B|A_1)=0,5$. Podobnie $P(B|A_2)=0,6$ i $P(B|A_3)=0,2$. Czyli prawdopodobieństwo całkowite tego, że losowo wybrany statek tego armatora ma PNS wynosi:

$$P(B)=\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)=0,45 \cdot 0,5+0,2 \cdot 0,6+0,35 \cdot 0,2=0,415$$

Wzór Bayesa

Zakładamy, że zdarzenia $A_1, A_2, \ldots, A_n$ spełniają założenia podane przy prawdopodobieństwie całkowitym oraz $P(B) > 0$. Wówczas

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

Prawdopodobieństwo to jest czasami nazywane prawdopodobieństwem ***a posteriori*** czyli „po fakcie". Stosuje się je często w sytuacjach, w których wiemy, że element ma jakąś cechę i na podstawie intensywności występowania tej cechy w różnych grupach, obliczamy prawdopodobieństwo, że należy on do danej grupy.

Przykład 16
Kontynuując poprzedni przykład obliczymy prawdopodobieństwo warunkowe tego, że wybrany statek to tankowiec, jeżeli wiadomo, że ma PNS. Czyli jest to prawdopodobieństwo $P(A_1|B)$.

$$P(A_1|B) = \frac{P(A_1) \cdot P(B|A_1)}{P(B)} = \frac{0{,}45 \cdot 0{,}5}{0{,}415} \approx 0{,}54$$

Schemat Bernulliego

Zakładamy, że w wyniku doświadczenia losowego *S* mogą zajść dwa zdarzenia: zdarzenie *A*, które nazywamy *sukcesem* oraz zdarzenie przeciwne *A'*, które nazywamy *porażką***.** Ponadto zakładamy, że prawdopodobieństwo zajścia zdarzenia *A* jest dla każdego doświadczenia *S* stałe i równa się *p*. Prawdopodobieństwo porażki *A'* oznaczamy symbolem $q\ (q = 1 - p)$. Ciąg powtórzeń doświadczenia *S* nazywamy ***schematem Bernoulliego*** (doświadczenia *S* nazywamy próbami Bernoulliego). Prawdopodobieństwo $P(S_n = k)$ otrzymania $k (0 \le k \le n)$ sukcesów w ciągu *n* prób Bernoulliego określone jest wzorem

$$P(S_n = k) = \binom{n}{k} p^k \cdot q^{n-k}, \text{ gdzie } p = P(A), q = P(A')$$

Przykład 17
Rozpatrzmy doświadczenie losowe polegające na dziesięciokrotnym powtórzeniu rzutu monetą. Obliczymy prawdopodobieństwo tego, że 4 razy wypadnie orzeł. W tym wypadku sukcesem jest wyrzucenie orła, a porażką wyrzucenie reszki. Prawdopodobieństwo sukcesu w pojedynczej próbie wynosi $p = \frac{1}{2}$, a prawdopodobieństwo porażki $q = \frac{1}{2}$. Liczba prób to *n*=10, a liczba sukcesów to *k*=4. Z powyższego wzoru wynika więc, że

$$P(S_{10} = 4) = \binom{10}{4} \cdot \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^6 = \frac{210}{1024} \approx 0{,}205$$

Rozdział II. Zmienne losowe

W poprzednim rozdziale rozpatrywaliśmy wiele doświadczeń losowych, których wyniki przedstawialiśmy w sposób opisowy. W niniejszym rozdziale wprowadzimy pojęcie zmiennej losowej, które pozwoli „przerobić" wyniki doświadczeń losowych na liczby. Taki sposób ujęcia da nowe możliwości opisu zjawisk losowych.

Zmienną losową X nazywamy każdą funkcję określoną na przestrzeni wyników Ω, o wartościach rzeczywistych taką, że dla każdej liczby rzeczywistej *x*, zbiór zdarzeń elementarnych $\omega \in \Omega$, dla których $X(\omega) < x$, jest zdarzeniem losowym.

Zmienne losowe oznaczamy najczęściej wielkimi literami: *X, Y, T, W, Z*. Wyróżniamy dwa podstawowe typy zmiennych losowych: skokowe i ciągłe.

Zmienna losowa skokowa

Zmienną losową nazywamy ***skokową*** (dyskretną), jeżeli przyjmuje skończoną lub przeliczalną liczbę wartości.

Rozkładem prawdopodobieństwa zmiennej losowej skokowej *X* nazywamy zbiór par: $\{(x_i, p_i), i = 1,2,...\}$, gdzie x_i oznacza wartość zmiennej *X*, a p_i jest prawdopodobieństwem, z jakim zmienna *X* przyjmuje wartość x_i czyli $p_i = P(X=x_i)$. Często rozkład zmiennej losowej skokowej jest przedstawiany w postaci tabeli:

x_i	x_1	x_2	...	x_n
$P(X = x_i)$	p_1	p_2	...	p_n

Przykład 18
Rzucamy 3 razy monetą. Niech X oznacza liczbę wyrzuconych orłów. Przestrzeń wyników jest następująca: Ω = {(*o,o,o*), (*o,o,r*),(*o,r,o*), (*r,o,o*), (*o,r,r*), (*r,o,r*), (*r,r,o*), (*r,r,r*)}.
Dla poszczególnych zdarzeń elementarnych zmienna losowa X ma następujące wartości:

$X((o,o,o)) = 3$

$X((o,o,r)) = 2$

$X((o,r,o)) = 2$

$X((r,o,o)) = 2$

$X((o,r,r)) = 1$

$X((r,o,r)) = 1$

$X((r,r,o)) = 1$

$X((r,r,r)) = 0$

Wartości 0 odpowiada jedno zdarzenie elementarne, czyli P(X=0)=1/8. Wartości 1 odpowiadają trzy zdarzenia elementarne, czyli P(X=1)=3/8. Wartości 2 odpowiadają trzy zdarzenia elementarne, czyli P(X=2)=3/8. Wartości 3 odpowiada jedno zdarzenie elementarne, czyli P(X=3)=1/8. Stąd rozkład prawdopodobieństwa tej zmiennej jest następujący:

x_i	0	1	2	3
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Dystrybuanta

Dystrybuantą zmiennej losowej X nazywamy funkcję $F(x)$ zmiennej $x \in R$ określoną wzorem $F(x) = P(X < x)$, dla każdego $x \in R$. Inaczej mówiąc jest to prawdopodobieństwo skumulowane.

Dystrybuanta zmiennej losowej typu skokowego określona jest wzorem

$$F(x) = \sum_{\substack{i \\ x_i < x}} p_i$$

gdzie sumowanie rozciąga się na te wskaźniki i, dla których $x_i < x$.

Przykład 19

Kontynuując poprzedni przykład mamy:

dla $x \leq 0$, $F(x) = 0$, (nic nie sumujemy, bo na lewo od jakiegokolwiek x z tego przedziału zmienna nie ma żadnej wartości)

dla $0 < x \leq 1$, $F(x) = P(X=0) = p_1 = 1/8$, (uwzględniamy tu tylko jedną wartość zmiennej losowej, która jest mniejsza od jakiegokolwiek x z tego przedziału)

dla $1 < x \leq 2$, $F(x) = P(X=0) + P(X=1) = p_1 + p_2 = 4/8$

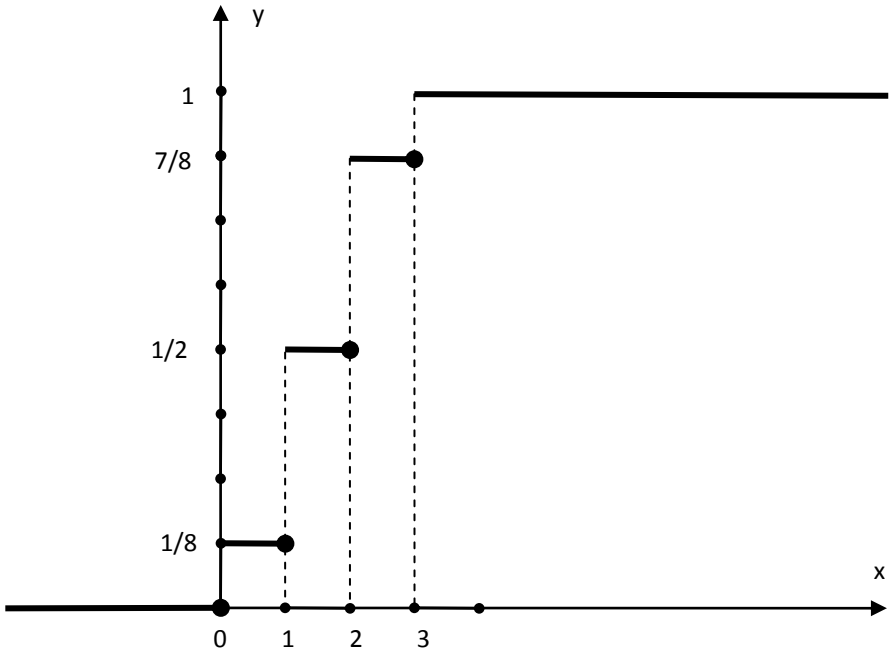
dla $2 < x \leq 3$, $F(x) = P(X=0) + P(X=1) + P(X=2) = p_1 + p_2 + p_3 = 7/8$

dla $x > 3$, $F(x) = P(X=0) + P(X=1) + P(X=2) + P(X=2) = p_1 + p_2 + p_3 + p_4 = 1$

Stąd dystrybuanta ma następującą postać:

$$F(x) = \begin{cases} 0, \text{ dla } x \leq 0 \\ \frac{1}{8}, \text{ dla } x \in (0, 1\rangle \\ \frac{4}{8}, \text{ dla } x \in (1, 2\rangle \\ \frac{7}{8}, \text{ dla } x \in (2, 3\rangle \\ 1, \text{ dla } x > 3 \end{cases}$$

Wykres tej funkcji , przedstawiony poniżej, wyjaśnia nam skąd wzięła się nazwa zmienna skokowa.



Charakterystyki liczbowe rozkładu zmiennej losowej

Do opisu zmiennej losowej, oprócz rozkładu, wykorzystujemy pewne specjalne liczby, zwane parametrami tej zmiennej. Podstawową charakterystyką jest ***wartość oczekiwana*** zwana inaczej *wartością przeciętną*, a dawniejsza jej nazwa to *nadzieja matematyczna*.

Wartością oczekiwaną zmiennej losowej *X* typu skokowego o rozkładzie prawdopodobieństwa $\{(x_i, p_i), i = 1,2,...\}$ nazywamy liczbę *EX* określoną wzorem

$$EX = \sum_i x_i p_i$$

Przykład 20

Niech zmienna losowa X oznacza dzienną liczbę wezwań statku ratowniczego w pewnym porcie. A jej rozkład prawdopodobieństwa niech będzie następujący:

x_i	0	1	2	3	4
$P(X = x_i)$	0,15	0,3	0,4	0,1	0,05

$EX = 0 \cdot 0,15 + 1 \cdot 0,3 + 2 \cdot 0,4 + 3 \cdot 0,1 + 4 \cdot 0,05 = 1,6$

Oznacza to, że przeciętna liczba wezwań statku ratowniczego wynosi 1,6 na dzień. Oczywiście liczba ta nic nie znaczy dla załogi tego statku danego dnia. Jednak chcąc obliczyć orientacyjną liczbę wezwań statku w ciągu miesiąca, mnożymy wartość przeciętną przez 30 i wiemy, że będzie to ok. 48

wezwań. A to już jest cenna informacja, dla obsługi tego statku, bo pozwala oszacować ilość potrzebnego paliwa, liczbę godzin pracy itp.

Innym parametrem jest ***wariancja.*** Jest to liczba charakteryzująca rozrzut wartości zmiennej losowej od jej wartości przeciętnej. Wariancją D^2X zmiennej losowej X typu skokowego o rozkładzie $\{(x_i, p_i), i = 1,2,...\}$ nazywamy liczbę

$$D^2X = \sum_i (x_i - EX)^2 p_i$$

jeśli szereg jest bezwzględnie zbieżny.

Czasami łatwiej obliczyć wariancję ze wzoru

$$D^2X = E(X^2) - (EX)^2$$

gdzie $E(X^2) = \sum_i (x_i)^2 \cdot p_i$

Wariancja jako miara rozrzutu jest o tyle niewygodna, że wyrażona jest w jednostkach kwadratowych. Dlatego zdefiniowano ***odchylenie standardowe***. Odchyleniem standardowym σ (lub DX) zmiennej losowej X nazywamy pierwiastek arytmetyczny drugiego stopnia z wariancji.

$$\sigma = \sqrt{D^2X}$$

Przykład 21
Kontynuując przykład 100 mamy:

$$EX = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1,5$$

$$E(X^2) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} = \frac{24}{8} = 3$$

$$D^2X = E(X^2) - (EX)^2 = 3 - (1,25)^2 = 0,75$$

$$\sigma = \sqrt{D^2X} = \sqrt{0,75} \approx 0,87$$

Wśród nieskończonej ilości zmiennych losowych o rozkładach skokowych wyróżnia się kilka charakterystycznych typów tych zmiennych.

Wybrane rozkłady prawdopodobieństwa zmiennej losowej skokowej i ich parametry

Rozkład zero – jedynkowy to zbiór dwuelementowy: $\{(0,q),(1,p)\}$, gdzie $0 < p < 1, q = 1 - p$
Parametrami tego rozkładu są liczby p i q, dla których mamy: $EX = p$, $D^2X = p \cdot q$

Rozkład dwumianowy *(Bernoulliego)* to zbiór: $\{(k, p_n^k), k = 0,1,2,...,n\}$, gdzie prawdopodobieństwo p_n^k obliczamy, tak jak w schemacie Benoulliego: $p_n^k = P(S_n = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}$, przy danych prawdopodobieństwach p i q takich, że: $0 < p < 1, q = 1 - p$

Parametrami tego rozkładu są liczby *p* i *q*, dla których mamy: $EX = np, D^2X = npq$

Przykład 22
Niech X oznacza liczbę wadliwych elementów wśród pięciu wylosowanych do kontroli. Wiadomo, że wadliwość tych elementów wynosi 0,2. Rozkład zmiennej X jest rozkładem dwumianowym z parametrami p=0,2 i q=0,8. Wartościami tej zmiennej są: 0, 1, 2, 3, 4, 5 (ponieważ wśród pięciu wylosowanych tyle może być elementów wadliwych). Prawdopodobieństwa obliczamy ze wzoru Bernoulliego i otrzymujemy rozkład prawdopodobieństwa zmiennej X:

x_i	0	1	2	3	4	5
$P(X = x_i)$	0,32768	0,4096	0,2048	0,0512	0,0064	0,00032

Rozkład Poissona to nieskończony, przeliczalny zbiór: $\left\{\left(k, p_k\right), k = 0,1,2,\ldots\right\}$, gdzie prawdopodobieństwo p_k obliczamy następująco:

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!}$$

Liczba λ jest parametrem tego rozkładu i jednocześnie wartością oczekiwaną i wariancją zmiennej o tym rozkładzie:

$$EX = \lambda, D^2X = \lambda$$

Przykład 23
Dzienna liczba rozładowanych statków w porcie G, jest zmienną losową o rozkładzie Poissona. Przeciętnie dziennie rozładowuje się 5 statków Jakie jest prawdopodobieństwo, że w ciągu dnia zostanie rozładowanych mniej niż 4 statki.

Średnia czyli parametr λ jest równa 5. Mamy policzyć P(*X*<4), czyli $p_0 + p_1 + p_2 + p_3$.

$$p_0 = \frac{5^0 e^{-5}}{0!} \approx 0,0067$$

$$p_1 = \frac{5^1 e^{-5}}{1!} \approx 0,0337$$

$$p_2 = \frac{5^2 e^{-5}}{2!} \approx 0,0842$$

$$p_3 = \frac{5^3 e^{-5}}{3!} \approx 0,1404$$

Czyli $P\left(X < 5\right) = p_0 + p_1 + p_2 + p_3 = 0,265$.

Zmienna losowa ciągła

Zmienną losową nazywamy ***ciągłą,*** jeżeli może ona przyjmować każdą wartość z pewnego skończonego lub nieskończonego przedziału.

Zmienna losowa ciągła jest określona poprzez ***funkcję gęstości prawdopodobieństwa*** *f(x)*, która spełnia dwa warunki:

1) $f(x) \geq 0$

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$

Pierwszy warunek związany jest z tym, że prawdopodobieństwo jest liczbą nieujemną, natomiast drugi warunek wiąże się z tym, że suma wszystkich prawdopodobieństw jest równa 1. Drugi warunek geometrycznie oznacza, że pole pod wykresem funkcji gęstości jest równe 1. Ogólnie rzecz biorąc, prawdopodobieństwo tego, że zmienna losowa mieści się w przedziale od *a* do *b*, obliczymy całkując funkcję gęstości w granicach od *a* do *b*, czyli

$$P(a < X < b) = \int_a^b f(x)dx$$

Przykład 24

Żeby poniższa funkcja była gęstością pewnej zmiennej losowej X, musi spełniać oba powyższe warunki.

$$f(x) = \begin{cases} \frac{a}{x^3} \text{ dla } x \in \langle 1,2 \rangle \\ 0 \quad \text{dla } x \notin \langle 1,2 \rangle \end{cases}$$

Z pierwszego warunku wynika, że liczba *a* musi być dodatnia. Z drugiego warunku wynika, że $\int_1^2 \frac{a}{x^3}dx = 1$. Rozpatrujemy tylko przedział od 1 do 2, ponieważ poza tym przedziałem funkcja przyjmuje wartość 0, czyli pole pod wykresem jest zerowe. Obliczając tę całkę, otrzymujemy:

$$\int_1^2 \frac{a}{x^3}dx = 1 \Rightarrow a\int_1^2 x^{-3}dx = 1 \Rightarrow \frac{-a}{2x^2}\Big|_1^2 = 1 \Rightarrow \frac{a}{2} - \frac{a}{8} = 1 \Rightarrow a = \frac{8}{3}$$

Przykład 25

Obliczymy teraz, dla zmiennej z poprzedniego przykładu, prawdopodobieństwo P(X < 1,5). Jako, że funkcja gęstości jest różna od zera tylko w przedziale od 1 do 2, to

$$P(X < 1,5) = \int_1^{1,5} \frac{\frac{8}{3}}{x^3}dx = \frac{-4}{3x^2}\Big|_1^{1,5} = \frac{4}{3} - \frac{4}{6,75} \approx 0,74$$

Natomiast prawdopodobieństwo P(X > 1,7) wynosi:

$$P(X > 1,7) = \int_{1,7}^2 \frac{\frac{8}{3}}{x^3}dx = \frac{-4}{3x^2}\Big|_{1,7}^2 \approx 0,128$$

Prawdopodobieństwo P(X < −1) jest równe 0, gdyż na tym przedziale, czyli $(-\infty, 1)$ funkcja gęstości jest zerowa.
Prawdopodobieństwo P(X < 5) jest równe 1, gdyż ten przedział, czyli $(-\infty, 5)$ obejmuje całe pole pod wykresem funkcji gęstości.

Dystrybuanta

Dystrybuanta zmiennej losowej typu ciągłego o funkcji gęstości $f(x)$ określona jest wzorem

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

Przykład 26

Dla funkcji gęstości $f(x)=\begin{cases}\frac{8}{3x^3}\text{dla } x\in\langle 1,2\rangle\\ 0 \quad \text{dla } x\notin\langle 1,2\rangle\end{cases}$ trzeba rozważyć trzy przedziały:

1) dla $x\in(-\infty,1\rangle$ mamy $F(x)=\int_{-\infty}^{x} 0dt=0$

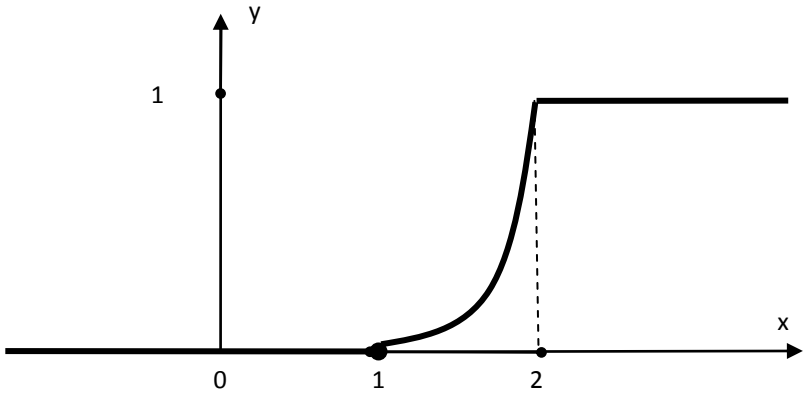
2) dla $x\in(1,2\rangle$ mamy $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{1} 0dt+\int_{1}^{x}\frac{8}{3t^3}dt=0-\frac{4}{3t^2}\Big|_1^x=\frac{4}{3}-\frac{4}{3x^2}$

3) dla $x\in(2,\infty)$ mamy $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{-\infty}^{1} 0dt+\int_{1}^{2}\frac{8}{3t^3}dt+\int_{2}^{\infty}0dt=0+1+0=1$

Stąd dystrybuanta ma następujący wzór:

$$F(x)=\begin{cases}0 \text{ dla } x\in(-\infty,1\rangle\\ \frac{4}{3}-\frac{4}{3x^2} \text{ dla } x\in(1,2\rangle\\ 1 \text{ dla } x\in(2,\infty)\end{cases}$$

A jej wykres jest linią ciągłą, jak widać to poniżej.



Parametry rozkładu zmiennej losowej ciągłej

Wartością oczekiwaną zmiennej losowej *X* typu ciągłego o gęstości $f(x)$ nazywamy liczbę

$EX = \int_{-\infty}^{+\infty} x \cdot f(x)dx$, jeżeli całka jest bezwzględnie zbieżna.

Wartością oczekiwaną zmiennej losowej $Y = g(X)$ nazywamy liczbę

$EY = \int_{-\infty}^{+\infty} g(x) \cdot f(x)dx$,

gdzie $f(x)$ jest gęstością zmiennej losowej *X*.

W szczególnym przypadku $E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx$,

Wariancją D^2X zmiennej losowej *X* typu ciągłego o gęstości $f(x)$ nazywamy liczbę

$D^2X = \int_{-\infty}^{+\infty} (x - EX)^2 f(x)dx$, jeżeli całka jest bezwzględnie zbieżna.

Przykład 27

Dla funkcji gęstości $f(x) = \begin{cases} \frac{5}{4x^2} \text{ dla } x \in \langle 1,5 \rangle \\ 0 \quad \text{dla } x \notin \langle 1,5 \rangle \end{cases}$ wartość oczekiwana jest następująca:

$$EX = \int_{-\infty}^{+\infty} x \cdot f(x)dx = \int_1^5 x \cdot \frac{5}{4x^2} dx = \frac{5}{4} \ln x \Big|_1^5 = \frac{5}{4} \ln 5 \approx 2,01$$

Żeby obliczyć wariancję obliczymy najpierw $E(X^2)$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx = \int_1^5 x^2 \cdot \frac{5}{4x^2} dx = \frac{5}{4} x \Big|_1^5 = 6$$

Stąd $D^2X = E(X^2) - (EX)^2 = 6 - (2,01)^2 \approx 1,96$

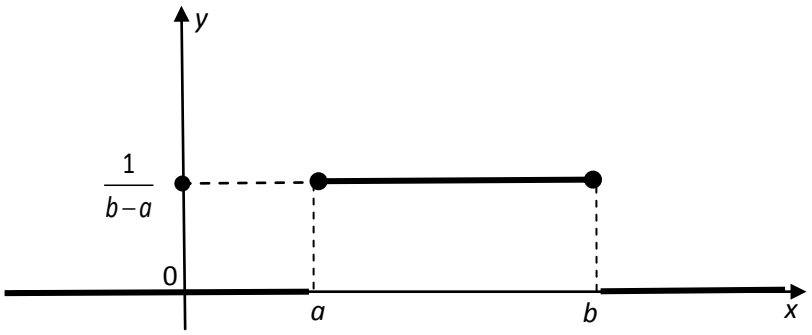
Wybrane rozkłady prawdopodobieństwa zmiennej losowej ciągłej i ich parametry

Spośród mnóstwa różnych rozkładów prawdopodobieństwa zmiennych losowych ciągłych, poniżej omówimy trzy: rozkład jednostajny, rozkład wykładniczy i rozkład normalny. Na koniec wspomnimy o dwóch rozkładach często wykorzystywanych w statystyce: o rozkładzie chi – kwadrat i o rozkładzie t-Studenta.

Rozkład jednostajny (równomierny lub prostokątny)
Zmienna losowa *X* ma rozkład jednostajny w przedziale (a,b), jeżeli ma gęstość $f(x)$ określoną wzorem

$$f(x)=\begin{cases}0, & x<a\\ \dfrac{1}{b-a}, & a\le x\le b\\ 0, & x>b\end{cases}$$

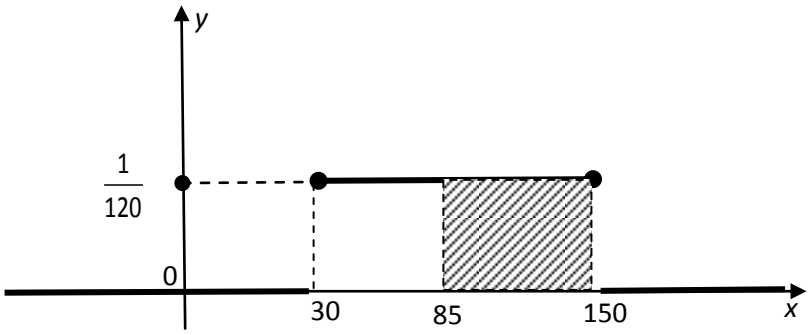
Wykres funkcji gęstości rozkładu jednostajnego przedstawiono poniżej.



Wartość oczekiwana i wariancja tego rozkładu wyrażają się następującymi wzorami:

$$EX=\frac{a+b}{2},\quad D^2X=\frac{(b-a)^2}{12}$$

Przykład 28
Czas oczekiwania na wejście do śluzy jest zmienną losową o rozkładzie jednostajnym. Maksymalny czas oczekiwania wynosi 150 minut, a minimalny 30 minut. Jakie jest prawdopodobieństwo, że na wejście do śluzy trzeba będzie czekać ponad 85 minut?



$P(X>85)$ geometrycznie oznacza to pole obszaru zakreskowanego na powyższym rysunku, czyli:

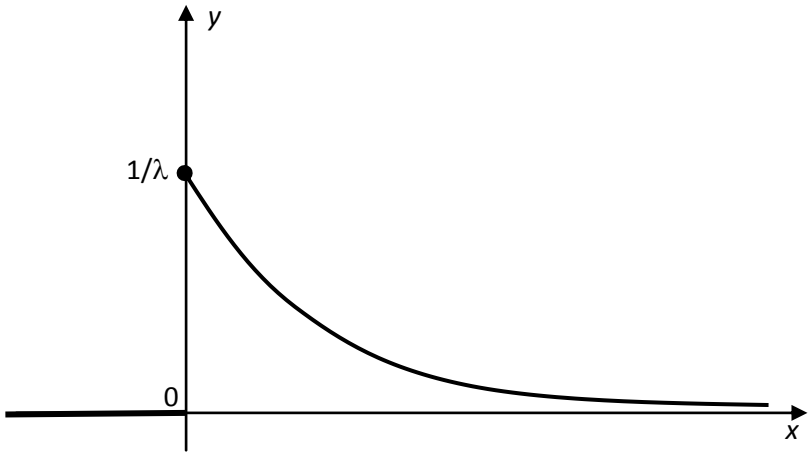
$$P(X>85)=\frac{1}{120}\cdot(150-85)=\frac{65}{120}\approx 0,542$$

Rozkład wykładniczy

Zmienna losowa *X* ma rozkład wykładniczy o parametrze $\lambda > 0$, jeżeli ma gęstość $f(x)$ określoną wzorem

$$f(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, & x \geq 0 \end{cases}$$

Wykres funkcji gęstości rozkładu wykładniczego przedstawiono poniżej.



Wartość oczekiwana i wariancja tego rozkładu wyrażają się następującymi wzorami:

$$EX = \lambda, \quad D^2X = \lambda^2$$

Przykład 29

Zmienna losowa *X* ma rozkład wykładniczy z parametrem λ= 5. Oblicz prawdopodobieństwa $P(X < 3), P(X > 2)$.

$$P(X < 3) = \int_0^3 \frac{1}{5} e^{\frac{-x}{5}} dx \approx 0{,}45$$

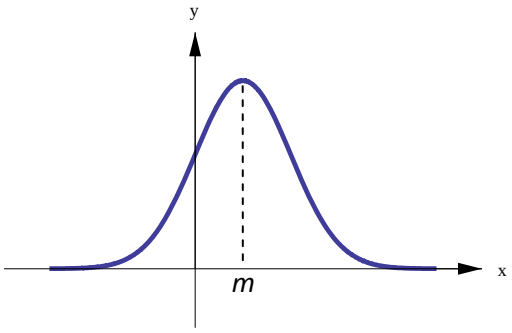
$$P(X > 2) = 1 - P(X < 2) = 1 - \int_0^2 \frac{1}{5} e^{\frac{-x}{5}} dx \approx 1 - 0{,}33 = 0{,}67$$

Rozkład normalny (Gaussa)

Zmienna losowa *X* ma rozkład normalny z parametrami *m* i σ, jeżeli ma gęstość $f(x)$ określoną wzorem

$$f(x) = \frac{1}{\sqrt{2}\,\sigma} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}} \quad \text{gdzie: } m \in R, \sigma > 0$$

Taki rozkład normalny oznaczamy symbolem: $N(m, \sigma)$. Wykres funkcji gęstości rozkładu normalnego, czyli tzw. krzywą Gaussa, przedstawiono poniżej.



Wartość oczekiwana w tym rozkładzie to parametr *m*, a wariancja to kwadrat parametru σ, czyli

$$EX = m,\ D^2X = \sigma^2$$

Tak jak w poprzednich rozkładach ciągłych, prawdopodobieństwa wyznaczamy licząc całki oznaczone z funkcji gęstości. Jednak w tym przypadku, żeby obliczyć te całki trzeba zastosować metody, których student I roku nie zna. Dlatego na końcu *Materiałów do ćwiczeń* znajduje się tabela wartości funkcji $\Phi(x)$, która jest dystrybuantą rozkładu normalnego $N(0, 1)$. Jest to funkcja określona następująco:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2}} \cdot e^{-\frac{t^2}{2}} dt$$

Funkcja ta pozwala obliczać prawdopodobieństwa dla wszystkich normalnych zmiennych losowych. Stosuje się w tym celu standaryzację zmiennych.

Mówimy, że zmienna losowa $U = \frac{X-m}{\sigma}$, gdzie *m* jest wartością oczekiwaną zmiennej *X* i σ jest odchyleniem standardowym zmiennej *X*, jest zmienną losową ***standaryzowaną*** (lub unormowaną). Wartością oczekiwaną zmiennej *U* jest 0, a jej odchylenie standardowe wynosi 1. Stąd dzięki standaryzacji, możemy dowolną zmienną losową o rozkładzie $N(m, \sigma)$ „przerobić" na zmienną $N(0, 1)$, której dystrybuanta jest stablicowana.

Przykład 30
Zmienna losowa *X* ma rozkład normalny N(70; 5). Wyznacz prawdopodobieństwa P(*X* < 78) oraz P(*X* > 65) .

$$P(X<78) = P\left(\frac{X-70}{5} < \frac{78-70}{5}\right) = P(U<1{,}4) = \Phi(1{,}44) \approx 0{,}919$$ (wartość odczytana z tablic)

$$P(X>65) = 1 - P(X<65) = 1 - P\left(\frac{X-70}{5} < \frac{65-70}{5}\right) = P(U<1) = \Phi(1) \approx 0{,}841$$ (wartość odczytana z tablic).

Rozkład χ^2 (chi-kwadrat)

Rozkładem χ^2 (chi-kwadrat) z k stopniami swobody nazywamy rozkład następującej sumy:

$$\chi^2 = X_1^2 + X_2^2 + ... + X_k^2 = \sum_{i=1}^{k} X_i^2$$

gdzie X_1, X_2,..., X_k są niezależnymi zmiennymi losowymi o rozkładzie normalnym z wartością oczekiwaną równą 0 i odchyleniem standardowym równym 1. Funkcja gęstości prawdopodobieństwa tej zmiennej ma postać:

$$f\left(\chi^2\right) = \begin{cases} \dfrac{1}{\sqrt{2^k}\cdot\Gamma\left(k/2\right)}\cdot\left(\chi^2\right)^{\frac{k}{2}-1}\cdot e^{\frac{-\chi^2}{2}} & \text{dla } \chi^2 > 0 \\ 0 & \text{dla } \chi^2 \le 0 \end{cases}$$

gdzie Γ(x) jest specjalna funkcją zwaną ***funkcją gamma.***
Tablica z odpowiednimi wartościami rozkładu χ^2 została umieszczona na końcu *Materiałów do ćwiczeń*.

Rozkład t - Studenta

Rozkładem *t*-Studenta z *k* stopniami jest to rozkład prawdopodobieństwa zmiennej losowej t_k określonej następująco:

$$t_k = \frac{t}{\sqrt{\chi_k^2}}\cdot\sqrt{k}$$

gdzie *t* i χ_k^2 są to niezależne zmienne losowe, *t* ma rozkład *N*(0,1), natomiast χ_k^2 ma rozkład chi-kwadrat z *k* stopniami swobody. Funkcja gęstości prawdopodobieństwa tej zmiennej ma postać:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\cdot\Gamma\left(\frac{k}{2}\right)\cdot\sqrt{k}}\cdot\left(\frac{t^2}{k}+1\right)^{\frac{k+1}{2}}$$

dla $-\infty < t < +\infty$

Rozkład *t*-Studenta jest stablicowany. Tablica z odpowiednimi wartościami tego rozkładu została umieszczona na końcu *Materiałów do ćwiczeń*.

Rozdział III. Zmienne losowe dwuwymiarowe

Uporządkowaną parę (X,Y) zmiennych losowych *X, Y* nazywamy ***dwuwymiarową zmienną losową*** lub ***dwuwymiarowym wektorem losowym.***

Dystrybuantą zmiennej losowej (X,Y) nazywamy funkcję rzeczywistą *F* określoną wzorem

$$F(x,y) = P(X < x, Y < y), x, y \in R$$

Podobnie jak wśród zmiennych losowych jednowymiarowych wyróżnia się zmienne skokowe i ciągłe. Poniżej krótko omówimy zmienną losowa skokową i jej parametry.

Zmienna losowa typu skokowego

Zmienna losowa (X,Y) jest ***typu skokowego*** (dyskretna), jeżeli istnieje skończony lub przeliczalny zbiór par wartości $(x_i,y_j)(i=1,2,...,j=1,2,...)$ taki, że $P(X=x_i,Y=y_j)=p_{ij}$ dla każdej pary wskaźników *i*, *j*, gdzie $p_{ij}>0$ oraz $\sum_{i,j} p_{ij}=1$. ***Rozkładem prawdopodobieństwa*** zmiennej losowej (X,Y) typu skokowego nazywamy zbiór $\{((x_i,y_j),p_{ij});i=1,2,...,j=1,2,...\}$.

Dystrybuanta *F* zmiennej losowej (X,Y) typu skokowego jest postaci $F(x,y)=\sum_{\substack{i,j \\ x_i<x,y_j<y}} p_{ij}$.

Wartością oczekiwaną (wartością przeciętną) zmiennej losowej (X,Y) typu skokowego nazywamy liczbę $E(X,Y)=\sum_{i,j} x_i y_j p_{ij}$ przy założeniu, że szereg jest bezwzględnie zbieżny.

Oznaczając odpowiednie sumy następującymi symbolami: $p_{i\square}=\sum_j p_{ij}$, $p_{\square j}=\sum_i p_{ij}$, możemy zdefiniować ***rozkłady brzegowe***.
Zbiory $\{(x_i,p_{i\square});i=1,2,...\},\{(y_j,p_{\square j});j=1,2,...\}$ nazywamy odpowiednio rozkładem brzegowym zmiennej losowej X oraz rozkładem brzegowym zmiennej losowej Y.

Zmienne losowe *X*, *Y* typu skokowego są ***niezależne***, jeżeli $p_{ij}=p_{i\square}\cdot p_{\square j}$ dla wszystkich i, j.
Dystrybuantą F_1 rozkładu brzegowego zmiennej losowej X nazywamy funkcję

$$F_1(x)=\sum_{\substack{i \\ x_i<x}} p_{i\square}\ ,\quad x\in R$$

Dystrybuantą F_2 rozkładu brzegowego zmiennej losowej *Y* nazywamy funkcję

$$F_2(y)=\sum_{\substack{j \\ y_j<y}} p_{\square j}\ ,\quad x\in R$$

Kowariancją zmiennych losowych *X, Y* nazywamy liczbę: $\operatorname{cov}(X,Y)=E\left[(X-EX)(Y-EY)\right]$

Można wykazać, że $\operatorname{cov}(X,Y)=E(X\cdot Y)-EX\cdot EY$.

Jeżeli $\operatorname{cov}(X,Y)=0$, to *X, Y* nazywamy zmiennymi losowymi ***nieskorelowanymi***.

Współczynnikiem korelacji ρ zmiennej losowej (X,Y) nazywamy liczbę

$$\rho=\frac{\operatorname{cov}(X,Y)}{\sqrt{D^2X}\cdot\sqrt{D^2Y}}$$

Przykład 31
Zmienna (X,Y) ma następujący rozkład prawdopodobieństwa

X \ Y	1	3	5	$p_{i\square}=\sum_j p_{ij}$
5	0,2	0	0,1	0,3
10	0	0,3	0,4	0,7
$p_{\square j}=\sum_i p_{ij}$	0,2	0,3	0,5	1

Rozkładem brzegowym zmiennej *X* jest zbiór {{5; 0,3}, {10; 0,7}}.

Rozkładem brzegowym zmiennej *Y* jest zbiór {{1; 0,2}, {3; 0,3}, {5; 0,5}}.

Zmienne *X* i *Y* nie są niezależne, gdyż nie zachodzi równość: $p_{ij}=p_{i\square}\cdot p_{\square j}$ dla wszystkich i, j . Na przykład p_{11}=0,2 jest różne od iloczynu: $p_{1\square}\cdot p_{\square 1}=0,3\cdot 0,2=0,06$.
$E(X\cdot Y)=5\cdot 1\cdot 0,2+5\cdot 3\cdot 0+5\cdot 5\cdot 0,1+10\cdot 1\cdot 0,2+10\cdot 3\cdot 0+10\cdot 5\cdot 0,1=10,5$
EX = 8,5 i *EY* = 3,6. Stąd $\operatorname{cov}(X,Y)=E(X\cdot Y)-EX\cdot EY=10,5-8,5\cdot 3,6=-20,1$

Zmienna losowa typu ciągłego

Teraz kilka podstawowych informacji na temat zmiennej losowej ciągłej. Dwuwymiarowa zmienna losowa (X,Y) jest ***typu ciągłego***, jeżeli istnieje nieujemna funkcja *f* (***funkcja gęstości***, gęstość) taka, że dla każdej pary liczb rzeczywistych (x,y) zachodzi wzór

$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy$, gdzie *F* jest dystrybuantą zmiennej losowej (X,Y).

Wartością oczekiwaną zmiennej losowej (X,Y) typu ciągłego o gęstości *f* nazywamy liczbę

$E(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy$, przy założeniu, że całka podwójna jest bezwzględna zbieżna.

Funkcję f_1 określoną wzorem $f_1(x)=\int_{-\infty}^{+\infty} f(x,y)dy$ nazywamy ***funkcją gęstości rozkładu brzegowego*** zmiennej losowej *X* w dwuwymiarowym rozkładzie zmiennej losowej (X,Y) typu ciągłego. Podobnie funkcję f_2 określoną wzorem $f_2(y)=\int_{-\infty}^{+\infty} f(x,y)dx$ nazywamy ***funkcją gęstości rozkładu brzegowego*** zmiennej losowej *Y*.

Funkcję F_1 określoną wzorem $F_1(x)=\int_{-\infty}^{x} f_1(x)dx, x\in R$ nazywamy dystrybuantą rozkładu brzegowego zmiennej losowej *X*. I podobnie funkcję F_2 określoną wzorem $F_2(y)=\int_{-\infty}^{y} f_2(y)dy, y\in R$ nazywamy dystrybuantą rozkładu brzegowego zmiennej losowej *Y*.

Zmienne losowe *X, Y* typu ciągłego są **niezależne**, jeżeli $f(x,y)=f_1(x)\cdot f_2(y)$ dla wszystkich $x,y\in R$.

Jeżeli zmienne losowe *X, Y* są niezależne, wówczas $E(X\cdot Y)=EX\cdot EY$

Prawa wielkich liczb

Na koniec tego rozdziału podamy dwa twierdzenia, które niejako łączą *rachunek prawdopodobieństwa* ze *statystyką*. Są to twierdzenia dotyczące „wielkiej" liczby zmiennych losowych. Okazuje się, że ta „wielka liczba" nie musi być wcale taka wielka. Wielu autorów przyjmuje że tą liczbą jest 30. Poniższe twierdzenia pokazują, że wielka liczba zmiennych losowych zachowuje się w sposób „nielosowy".

Twierdzenie Bernoulliego

Jeżeli w każdym z *n* doświadczeń niezależnych prawdopodobieństwo zajścia zdarzenia *A* jest stałe i równe *p*, to przy dostatecznie dużej liczbie doświadczeń, wartość bezwzględna różnicy między częstością względną zdarzenia *A*, a prawdopodobieństwem *p*, jest mniejsza od dowolnie małej liczby $\varepsilon > 0$ z prawdopodobieństwem bardzo bliskim jedności, czyli

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1$$

Centralne twierdzenie graniczne (Lindeberga-Levy'ego)

Jeżeli $X_1, X_2, \ldots$ jest ciągiem zmiennych losowych o jednakowych rozkładach ($EX = m$, $D^2X = \sigma$), to zmienna $Y_n=\frac{\sum_{k=1}^{n} X_k - m}{\sigma\sqrt{n}}$ ma rozkład asymptotycznie normalny, czyli

$$\lim_{n\to\infty} P(Y_n<y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-0,5x^2}dx=\Phi(y)$$

gdzie $\Phi(y)$ oznacza wartość dystrybuanty rozkładu normalnego $N(0, 1)$.

Rozdział IV. Elementy statystyki

Przedmiotem badań statystyki matematycznej są zbiory (zbiorowości), które nazywamy ***populacjami generalnymi***. Mogą to być mieszkańcy jakiegoś miasta, statki danego armatora, silniki okrętowe określonego typu itp. Własności elementów populacji generalnej, które podlegają badaniom statystycznym nazywamy ***cechami***.

Każdy podzbiór elementów wylosowanych z populacji generalnej nazywamy ***próbą losową***. Próbę losową traktujemy jako *n* elementową zmienną losową $(X_1, X_2, \ldots, X_n)$, której wartościami są *n* elementowe ciągi $(x_1, x_2, \ldots, x_n)$. Jeżeli zmienne losowe $X_1, X_2, \ldots, X_n$ są niezależne, to próbę losową nazywamy ***prostą***.

Zmienną losową $U = U(X_1, X_2, \ldots, X_n)$, która jest funkcją zmiennych losowych $X_1, X_2, \ldots, X_n$ nazywamy ***statystyką***.

Parametry i ich estymatory

Parametrami populacji nazywamy charakterystyki liczbowe całej populacji. ***Estymatorem*** parametru populacji jest statystyka z próby używana do oszacowania tego parametru. Oceną lub szacunkiem parametru jest konkretna wartość liczbowa estymatora z danej próby. Jeżeli jako ocenę podajemy jedną wartość liczbową, nazywamy ją oceną punktową (***estymacją punktową***) parametru populacji.

Dany parametr możemy szacować przy pomocy różnych estymatorów. Żeby wybrać ten najbardziej odpowiedni w danej sytuacji należy przyjrzeć się ich własnościom. W mniejszym skrypcie ograniczymy się do podania tylko tych podstawowych.

Estymator nazywamy ***nieobciążonym***, jeżeli jego wartość oczekiwana jest równa parametrowi populacji, do oszacowania, którego służy. Systematyczne odchylanie się wartości estymatora od szacowanego parametru nazywa się ***obciążonością*** estymatora. Estymator nazywamy ***efektywnym***, jeżeli ma niewielką wariancję. ***Zgodny*** estymator to taki, dla którego prawdopodobieństwo, że jego wartość będzie blisko wartości szacowanego parametru, wzrasta ze wzrostem liczebności próby.

Frakcją (częstością) w populacji (*p*) jest iloraz liczby elementów populacji należących do pewnej kategorii (N_p), którą się interesujemy, przez liczbę wszystkich elementów populacji (*N*), czyli $p = \frac{N_p}{N}$.

Jej estymatorem jest frakcja w próbie czyli $\bar{p} = \frac{n_p}{n}$, gdzie n_p jest liczbą elementów próby należących do interesującej nas kategorii, a n licznością próby.

Średnia w populacji to liczba

$$m = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

gdzie x_i to elementy populacji, *N* to liczność populacji.

Jej estymatorem jest średnia w próbie czyli liczba

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

gdzie x_i to elementy próby, *n* to liczność próby.

Wariancja w populacji to liczba: $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(X_k - m)^2$

A wariancja w próbie to liczba: $S^2 = \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2$

Estymatorem odchylenia standardowego w populacji jest odchylenie standardowe w próbie, które można liczyć dwoma wzorami:

$s = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2}$ - estymator obciążony

$\hat{s} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2}$ - estymator nieobciążony

Mimo, że drugi estymator jest nieobciążony (czyli „lepszy"), to w wielu poniższych wzorach wykorzystano pierwszy estymator.

Estymacja przedziałowa

Przedziałem ufności dla parametru Θ nazywamy przedział losowy $\langle u_1, u_2 \rangle$, o którym z danym prawdopodobieństwem $1-\alpha$ (***poziom ufności***) możemy twierdzić, że zawiera nieznany parametr Θ:

$$P\left(u_1 \le \Theta \le u_2\right) = 1-\alpha$$

Z centralnego twierdzenia granicznego wynika, że jeżeli pobieramy próbę z populacji o średniej *m* i skończonym odchyleniu standardowym σ i gdy liczebność próby wzrasta nieograniczenie, to rozkład średniej z próby $\bar{X}$, dąży do rozkładu normalnego o średniej *m* i odchyleniu standardowym $\frac{\sigma}{\sqrt{n}}$, czyli „dla dostatecznie dużych n " (n >30) mamy: $\bar{X} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

Przedział ufności dla wartości oczekiwanej m w populacji generalnej o rozkładzie normalnym $N(m, \sigma)$

Niech $(X_1, X_2, \ldots, X_n)$ będzie próbą losową prostą pobraną z populacji generalnej o rozkładzie normalnym $N(m, \sigma)$, gdzie wartość oczekiwana jest nieznana, a odchylenie standardowe jest znane. Estymatorem wartości oczekiwanej *m* jest średnia z próby $\bar{X}$ o rozkładzie $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$. Standaryzując zmienną losową $\bar{X}$ otrzymujemy statystykę $Y = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - m}{\sigma}\sqrt{n}$, która jest zmienną losową o rozkładzie *N*(0, 1).

Dla ustalonego poziomu prawdopodobieństwa $1-\alpha$, z tablic rozkładu normalnego *N*(0, 1), odczytujemy taką liczbę *u*, dla której mamy: $P\left(-u < Y < u\right) = 1-\alpha$. Stąd otrzymujemy:

$$P\left(-u<\frac{\overline{X}-m}{\sigma}\sqrt{n}<u\right)=1-\alpha,$$

a po przekształceniu wewnętrznej nierówności, tak by wyznaczyć parametr *m* , mamy:

$$P\left(\overline{X}-u\frac{\sigma}{\sqrt{n}}<m<\overline{X}+u\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

I ten właśnie wzór określa przedział ufności dla wartości oczekiwanej *m* w populacji generalnej o rozkładzie normalnym $N(m,\sigma)$, gdy znamy parametr σ.

Mając konkretną próbę losową $(x_1,x_2,...,x_n)$ można w skrócie powiedzieć, że przedziałem ufności dla średniej, gdy znane jest σ, jest przedział liczbowy:

$$\left(\overline{x}-u\cdot\frac{\sigma}{\sqrt{n}},\ \overline{x}+u\cdot\frac{\sigma}{\sqrt{n}}\right)$$

gdzie *u* jest liczbą, którą odczytujemy z tablic dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u)=1-\frac{\alpha}{2}$.

Przykład 32
W pewnym punkcie akwenu dokonano 7 pomiarów głębokości i otrzymano następujące wyniki [m]: 24, 25, 26, 25, 26, 26, 27. Na poziomie ufności 0.95 wyznacz przedział ufności dla średniej głębokości akwenu. (Przy założeniu, że rozkład pomiarów jest normalny, a odchylenie standardowe wynosi 0,5m).
Średnia z próby jest równa: $\overline{x}=25,4286$, a liczba *u* dla α = 0,05 wynosi: 1,96. Stąd przedział ufności dla średniej głębokości akwenu ma postać: (25,06; 25,8).

Podobnie konstruuje się pozostałe przedziały ufności, z tym, że wykorzystuje się odpowiednie dla danego estymatora parametru rozkłady prawdopodobieństwa.

Przedział ufności dla wartości oczekiwanej m w populacji generalnej o rozkładzie normalnym $N(m,\sigma)$***, w którym*** σ ***nie jest znane***:

$$\left(-t_\alpha\cdot\frac{S}{\sqrt{n-1}}+\overline{X},\ t_\alpha\cdot\frac{S}{\sqrt{n-1}}+\overline{X}\right)$$

gdzie t_α jest liczbą, którą odczytujemy z tablic rozkładu *t* – Studenta dla $n-1$ stopni swobody i danego poziomu ufności $1-\alpha$ taką, że $P(|t|>t_\alpha)=\alpha$.

Przedział ufności dla wartości oczekiwanej m w dowolnym rozkładzie dla dużej próby $(n>120)$***:***

$$\left(-u\cdot\frac{S}{\sqrt{n}}+\overline{X},\ u\cdot\frac{S}{\sqrt{n}}+\overline{X}\right)$$

gdzie liczbę *u* odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$ dla danego poziomu ufności $1-\alpha$, $\Phi(u)=1-\frac{\alpha}{2}$.

Przedział ufności dla wskaźnika struktury (prawdopodobieństwa p zmiennej losowej o rozkładzie Bernoulliego)

$$\left(\frac{2k+u^2-u\sqrt{4k\left(1-\frac{k}{n}\right)+u^2}}{2\left(n+u^2\right)},\ \frac{2k+u^2+u\sqrt{4k\left(1-\frac{k}{n}\right)+u^2}}{2\left(n+u^2\right)}\right)$$

gdzie *k* jest ilością sukcesów wśród *n* niezależnych doświadczeń, *u* jest liczbą, odczytaną z tablic dystrybuanty rozkładu $N(0,1)$, spełniającą warunek $\Phi(u)=1-\frac{\alpha}{2}$

Przedział ufności dla wariancji σ^2 ***w populacji generalnej o rozkładzie normalnym*** $N(m,\sigma)$***, w którym m i*** σ ***są nieznane***

$$\left\langle\frac{n\cdot S^2}{\chi_2^2},\ \frac{n\cdot S^2}{\chi_1^2}\right\rangle$$

gdzie liczby χ_1^2, χ_2^2 odczytane z tablic rozkładu χ^2 (chi kwadrat) dla $n-1$ stopni swobody i danego poziomu ufności $1-\alpha$, spełniają warunki

$$P\left(\chi^2>\chi_1^2\right)=1-\frac{\alpha}{2},\ P\left(\chi^2>\chi_2^2\right)>\frac{\alpha}{2}$$

Weryfikacja hipotez statystycznych

Badanie całych populacji (wszystkich studentów, wszystkich dorosłych Polaków, czy wszystkich mieszkańców Szczecina) jest bardzo kłopotliwe i kosztowne. Z tego powodu rozwinęły się metody wnioskowania statystycznego, polegające między innymi na tym, że na podstawie próby losowej stawia się hipotezę dotyczącą całej populacji. Hipotezę taką nazywamy ***hipotezą zerową*** (H_0). Hipotezy mogą dotyczyć parametrów rozkładu (hipotezy parametryczne) lub rodzaju samego rozkładu (hipotezy nieparametryczne). Do weryfikacji hipotezy służy odpowiednio skonstruowana

statystyka zwana ***testem statystycznym*** lub ***statystyką testową***. Istotą weryfikacji hipotez jest działanie prowadzące do stwierdzenia, że daną hipotezę trzeba odrzucić. Hipotezę zerową odrzucamy, jeżeli wynik testu należy do ***obszaru krytycznego*** i wówczas przyjmujemy hipotezę alternatywną (H_A). W przeciwnym wypadku, nie ma podstaw do odrzucenia H_0. Wielkość obszaru krytycznego jest uzależniona od ***poziomu istotności*** α. Najczęściej stosowanym poziomem istotności jest 5% czyli α=0,05.

Hipoteza alternatywna może mieć jedną z następujących postaci: $H_A : \Theta \neq \Theta_0$ lub $H_A : \Theta < \Theta_0$ lub $H_A : \Theta > \Theta_0$. Dwie ostatnie, w przypadku odrzucenia hipotezy zerowej, niosą więcej informacji niż pierwsza.

Weryfikacja hipotezy o wartości oczekiwanej m w populacji generalnej o rozkładzie normalnym $N(m, \sigma)$***, jeżeli odchylenie standardowe*** σ ***jest znane.***

Hipotezą zerową jest hipoteza orzekająca, że wartość oczekiwana *m* jest równa liczbie m_0, czyli

$$H_0 : m = m_0$$

Dla hipotezy alternatywnej: $H_A : m \neq m_0$ testem jest statystyka: $\frac{\bar{X} - m_0}{\sigma} \cdot \sqrt{n}$, obszarem krytycznym jest suma przedziałów: $(-\infty, -u) \cup (u, +\infty)$, a obszarem przyjęcia hipotezy jest przedział: $\langle -u, u \rangle$, gdzie *u* jest liczbą, którą odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u) = 1 - \frac{\alpha}{2}$ dla danego poziomu istotności α.

Dla hipotezy alternatywnej: H_A: $m > m_0$, testem jest statystyka: $\frac{\bar{X} - m_0}{\sigma} \cdot \sqrt{n}$, obszarem krytycznym jest przedział: $(u, +\infty)$, a obszarem przyjęcia hipotezy jest przedział: $(-\infty, u\rangle$, gdzie *u* jest liczbą, którą odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u) = 1 - \alpha$ dla danego poziomu istotności α.

Dla hipotezy alternatywnej: H_A: $m < m_0$, testem jest statystyka: $\frac{\bar{X} - m_0}{\sigma} \cdot \sqrt{n}$, obszarem krytycznym jest przedział: $(-\infty, u)$, a obszarem przyjęcia hipotezy jest przedział: $\langle u, +\infty)$, gdzie *u* jest liczbą, którą odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u) = 1 - \alpha$ dla danego poziomu istotności α.

Przykład 33
Liczba statków u wylosowanych armatorów kształtowała się następująco: 43, 28, 22, 53, 69, 70. Na poziomie ufności 0.9 sprawdzić hipotezę, że średnia liczba statków jest równa 50, wobec hipotezy alternatywnej $H_A : m \neq 50$ przy założeniu że rozkład liczby statków jest normalny, ze znanym odchyleniem standardowym σ = 2.

Średnia z próby wynosi 47,5. Stąd statystyka testowa $\frac{\bar{X}-m_0}{\sigma}\cdot\sqrt{n}$ ma wartość: −3,06. Natomiast $\Phi(u)=1-\frac{\alpha}{2}=1-\frac{0,1}{2}=0,95$ i z tablic rozkładu normalnego odczytujemy, że liczba *u* wynosi 2,33. Stąd obszar krytyczny jest następujący: $(-\infty,-2,33)\cup(2,33,+\infty)$ i wartość statystyki testowej należy do tego obszaru. Czyli na poziomie ufności 0,9, należy odrzucić hipotezę orzekającą, że średnia liczba statków u badanych armatorów jest równa 50, na korzyść hipotezy, że średnia liczba statków u badanych armatorów nie jest równa 50.

Weryfikacja hipotezy o wartości oczekiwanej m w populacji generalnej o rozkładzie normalnym $N(m,\sigma)$***, jeżeli odchylenie standardowe*** σ ***nie jest znane (duża próba: n>120)***

W tym przypadku stosujemy te same wzory, co dla poprzedniej hipotezy, tylko parametr σ zastępujemy odchyleniem standardowym z próby *S*.

Weryfikacja hipotezy o wartości oczekiwanej m w populacji generalnej o rozkładzie normalnym $N(m,\sigma)$***, jeżeli odchylenie standardowe*** σ ***nie jest znane (n<120)***

Hipotezą zerową jest hipoteza orzekająca, że wartość oczekiwana *m* jest równa liczbie m_0, czyli

$$H_0: m = m_0$$

Dla hipotezy alternatywnej: $H_A: m \neq m_0$ testem jest statystyka: $\frac{\bar{X}-m_0}{S}\cdot\sqrt{n-1}$, obszarem krytycznym jest suma przedziałów: $(-\infty,-t_\alpha)\cup(t_\alpha,+\infty)$, a obszarem przyjęcia hipotezy jest przedział: $\langle -t_\alpha, t_\alpha\rangle$, gdzie t_α jest liczbą, którą odczytujemy z tablic rozkładu *t* – Studenta dla *n* – 1 stopni swobody i danego poziomu istotności α taką, że $P(|t|>t_\alpha)=\alpha$.

Dla hipotezy alternatywnej: H_A: $m > m_0$, testem jest statystyka: $\frac{\bar{X}-m_0}{S}\cdot\sqrt{n-1}$, obszarem krytycznym jest przedział: $(t_{2\alpha},+\infty)$, a obszarem przyjęcia hipotezy jest przedział: $(-\infty, t_{2\alpha}\rangle$, gdzie $t_{2\alpha}$ jest liczbą, którą odczytujemy z tablic rozkładu *t* – Studenta dla $n-1$ stopni swobody i danego poziomu istotności α taką, że $P(|t|>t_{2\alpha})=2\alpha$.

Dla hipotezy alternatywnej: H_A: $m < m_0$, testem jest statystyka: $\frac{\bar{X}-m_0}{S}\cdot\sqrt{n-1}$, obszarem krytycznym jest przedział: $(-\infty,-t_{2\alpha})$, a obszarem przyjęcia hipotezy jest przedział: $\langle t_{2\alpha},+\infty)$, gdzie $t_{2\alpha}$ jest liczbą, którą odczytujemy z tablic rozkładu *t* – Studenta dla $n-1$ stopni swobody i danego poziomu istotności α taką, że $P(|t|>t_{2\alpha})=2\alpha$.

Weryfikacja hipotezy dla wskaźnika struktury (prawdopodobieństwa p zmiennej losowej o rozkładzie Bernoulliego)

Hipotezą zerową jest w tym wypadku hipoteza orzekająca, że wskaźnik struktury *p* jest równy liczbie p_0 czyli

$$H_0 : p = p_0$$

W przypadku tej hipotezy w poniższych wzorach na statystykę testową występuje liczba gdzie k. Jest to ilość elementów w próbie n elementowej, posiadających badaną cechę. W innej interpretacji jest to ilość sukcesów. Dla hipotezy alternatywnej: $H_A : p \neq p_0$ testem jest statystyka: $\frac{k - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$, obszarem krytycznym jest suma przedziałów: $(-\infty, -u) \cup (u, +\infty)$, a obszarem przyjęcia hipotezy jest przedział: $\langle -u, u \rangle$, gdzie *u* jest liczbą, którą odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u) = 1 - \frac{\alpha}{2}$ dla danego poziomu istotności α.

Dla hipotezy alternatywnej: $H_A : p > p_0$, testem jest statystyka: $\frac{k - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$, obszarem krytycznym jest przedział: $(u, +\infty)$, a obszarem przyjęcia hipotezy jest przedział: $(-\infty, u\rangle$, gdzie *u* jest liczbą, którą odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u) = 1 - \alpha$ dla danego poziomu istotności α.

Dla hipotezy alternatywnej: $H_A : p < p_0$, testem jest statystyka: $\frac{k - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}}$, obszarem krytycznym jest przedział: $(-\infty, u)$, a obszarem przyjęcia hipotezy jest przedział: $\langle u, +\infty)$, gdzie *u* jest liczbą, którą odczytujemy z tablicy dystrybuanty $\Phi(x)$ rozkładu $N(0,1)$, spełniającą warunek $\Phi(u) = 1 - \alpha$ dla danego poziomu istotności α.

Weryfikacja hipotezy o wariancji σ^2 w populacji generalnej o rozkładzie normalnym $N(m, \sigma)$, w którym m i σ nie są znane.

Weryfikowaną hipotezą jest hipoteza mówiąca, że wariancja jest równa liczbie σ_0^2, czyli

$$H_0 : \sigma^2 = \sigma_0^2$$

Dla hipotezy alternatywnej: $H_A : \sigma^2 \neq \sigma_0^2$ testem jest statystyka: $\frac{nS^2}{\sigma_0^2}$, obszarem krytycznym jest suma przedziałów: $(0, \chi_1) \cup (\chi_2, +\infty)$, a obszarem przyjęcia hipotezy jest przedział: $\langle \chi_1, \chi_2 \rangle$, gdzie

liczby χ_1^2, χ_2^2 odczytane z tablic rozkładu chi kwadrat dla $n-1$ stopni swobody i danego poziomu ufności $1-\alpha$, spełniają warunki: $P\left(\chi^2 > \chi_1^2\right) = 1-\frac{\alpha}{2}$, $P\left(\chi^2 > \chi_2^2\right) > \frac{\alpha}{2}$.

Dla hipotezy alternatywnej: $H_A : \sigma^2 > \sigma_0^2$, testem jest statystyka: $\frac{nS^2}{\sigma_0^2}$, obszarem krytycznym jest przedział: $\left(\chi_2, +\infty\right)$, a obszarem przyjęcia hipotezy jest przedział: $\left\langle 0, \chi_2 \right\rangle$, gdzie χ_2^2 jest liczbą odczytaną z tablic rozkładu χ^2 dla $n-1$ stopni swobody i danego poziomu istotności α, spełniającą warunek $P\left(\chi^2 > \chi_2^2\right) = \alpha$.

Dla hipotezy alternatywnej: $H_A : \sigma^2 < \sigma_0^2$, testem jest statystyka: $\frac{nS^2}{\sigma_0^2}$, obszarem krytycznym jest przedział: $\left(0, \chi_1\right)$, a obszarem przyjęcia hipotezy jest przedział: $\left\langle \chi_1, +\infty \right\rangle$, gdzie χ_1^2 jest liczbą odczytaną z tablic rozkładu χ^2 dla $n-1$ stopni swobody i danego poziomu istotności α, spełniającą warunek $P\left(\chi^2 > \chi_1^2\right) = 1-\alpha$.

Powyższe hipotezy dotyczyły parametrów rozkładu danej cechy w jednej populacji. Ale można również testować hipotezy o równości parametrów danej cechy w dwóch populacjach. W poniższym omówieniu hipotez dotyczących dwóch populacji podamy tylko założenia i statystyki testowe, natomiast konstrukcja obszarów krytycznych i obszarów przyjęcia hipotezy jest analogiczna do tych przedstawionych powyżej.

Weryfikacja hipotezy o równości wartości oczekiwanych w dwóch populacjach o rozkładach normalnych $N\left(m_1, \sigma_1\right)$ ***i*** $N\left(m_2, \sigma_2\right)$***, jeżeli odchylenia standardowe*** σ_1 ***i*** σ_2 ***są znane***

Hipotezą zerową jest hipoteza orzekająca, że wartość oczekiwana pierwszej populacji m_1 jest równa wartości oczekiwanej drugiej populacji m_1, czyli

$$H_0 : m_1 = m_2$$

Statystyką testową jest wyrażenie: $\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, gdzie

n_1 – liczność próby pobranej z pierwszej populacji

n_2 – liczność próby pobranej z drugiej populacji

$\overline{X_1}$ - średnia z próby pobranej z pierwszej populacji

$\overline{X_2}$ - średnia z próby pobranej z drugiej populacji

Weryfikacja hipotezy o równości wartości oczekiwanych w dwóch populacjach o rozkładach normalnych $N(m_1,\sigma_1)$***, i*** $N(m_2,\sigma_2)$***, jeżeli odchylenia standardowe*** σ_1 ***i*** σ_2 ***są znane (*** $n_1+n_2>122$***)***

W tym przypadku stosujemy te same wzory, co dla poprzedniej hipotezy, tylko parametry σ_1 i σ_2 zastępujemy odchyleniami standardowymi z obu prób S_1 i S_2.

Weryfikacja hipotezy o równości wartości oczekiwanych w dwóch populacjach o rozkładach normalnych $N(m_1,\sigma_1)$***, i*** $N(m_2,\sigma_2)$***, jeżeli odchylenia standardowe*** σ_1 ***i*** σ_2 ***są znane (*** $n_1+n_2\leq 122$***)***

Hipotezą zerową jest hipoteza orzekająca, że wartość oczekiwana pierwszej populacji m_1 jest równa wartości oczekiwanej drugiej populacji m_1, czyli

$$H_0 : m_1 = m_2$$

Statystyką testową jest wyrażenie: $\dfrac{\overline{X_1}-\overline{X_2}}{\sqrt{\dfrac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$

Przy konstrukcji obszaru krytycznego i obszaru przyjęcia hipotezy korzystamy z rozkładu t – Studenta z (n_1+n_2-2) stopniami swobody.

Weryfikacja hipotezy o równości wariancji w dwóch populacjach o rozkładach normalnych

Hipotezą zerową jest hipoteza orzekająca, że wariancja pierwszej populacji σ_1^2 jest równa wartości oczekiwanej drugiej populacji σ_2^2, czyli

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Wobec hipotezy alternatywnej: $H_0 : \sigma_1^2 > \sigma_2^2$

Statystyką testową jest wyrażenie: $\dfrac{\widehat{S}_1^2}{\widehat{S}_2^2}$, gdzie $\widehat{S}_1^2$ i $\widehat{S}_2^2$ to nieobciążone estymatory odchylenia standardowego, obliczone dla prób pobranych z różnych populacji. Ze względu na postać hipotezy alternatywnej, należy oznaczyć jako pierwszą tę populację i próbę z niej, żeby zachodziło: $\widehat{S}_1^2 > \widehat{S}_2^2$

Obszarem krytycznym jest przedział (F_α,∞), a obszarem przyjęcia hipotezy jest przedział $(-\infty,F_\alpha\rangle$, gdzie F_α jest wartością odczytaną z tablic rozkładu *F* Snedecora dla odpowiedniego poziomu istotności α oraz dla n_1-1 i n_2-1 stopni swobody.

Test zgodności chi kwadrat (Pearsona)

Wszystkie omówione powyżej hipotezy dotyczyły parametrów danego rozkładu. Natomiast teraz przedstawimy weryfikację hipotezy dotyczącej postaci rozkładu prawdopodobieństwa zmiennej losowej *X*. Hipoteza H_0 jest hipotezą orzekającą, że dystrybuanta zmiennej losowej *X* ma postać $F(x)$, a hipotezą alternatywną jest hipoteza, która stwierdza, że rozkład zmiennej *X* ma dystrybuantę różną od $F(x)$.

Zakładamy, że zmienna losowa *X* ma rozkład o nieznanej dystrybuancie $F(x)$. Dysponujemy *n* elementową próbą losową o wartościach $x_1, x_2, ..., x_n$. Zbiór możliwych wartości zmiennej losowej *X* dzielimy na *r* rozłącznych podzbiorów J_k, $k = 1, 2, ..., r$ za pomocą liczb

$-\infty = a_0 < a_1 < ... < a_r = \infty$.

Niech $p_k\,(p_k > 0)$ oznacza prawdopodobieństwo, że zmienna losowa *X* przyjmuje wartość z przedziału J_k, tzn.

$$p_k = P(X \in J_k) = F(a_k) - F(a_{k-1}),\ k = 1, 2, ..., r$$

gdzie $F(x)$ jest hipotetyczną dystrybuantą.

Tworzymy zmienną losową χ^2 (statystykę χ^2 Pearsona)

$$\chi^2 = \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k}$$

Liczba np_k jest oczekiwaną liczbą obserwacji *n* elementowej próbki według założonego rozkładu, które powinny znaleźć się w przedziale J_k, natomiast N_k jest zmienną losową o wartościach n_k będących liczbą obserwacji, które znalazły się w przedziale J_k.

Dla zadanego poziomu istotności α, z tablic rozkładu χ^2 odczytujemy liczbę χ^2_α taką, że

$$P(\chi^2 \geq \chi^2_\alpha) = \alpha$$

Jeżeli $\chi^2_z < \chi^2_\alpha$, to nie mamy podstaw do odrzucenia hipotezy H_0, natomiast gdy $\chi^2_z > \chi^2_\alpha$, to hipotezę H_0 odrzucamy. χ^2_z oznacza wartości statystyki testowej zaobserwowaną w próbie.

Test niezależności chi kwadrat

Jeżeli przedmiotem badania jest populacja ze względu na występowanie dwóch cech *X* i *Y*, to w celu stwierdzenia niezależności tych cech stosujemy test niezależności chi kwadrat. Jest on oparty o tak zwaną tablicę niezależności. Tablica ta zawiera tyle wierszy ile jest wariantów cechy *X* i tyle kolumn

ile jest wariantów cechy *Y*. Niech *k* oznacza liczbę wariantów cechy *X*, a *r* liczbę wariantów cechy *Y*. Wtedy tablica niezależności wygląda następująco:

X \ *Y*	y_1	y_2	...	y_r	
x_1	n_{11}	n_{12}	...	n_{1r}	$\sum_{j=1}^{r} n_{1j}$
x_2	n_{21}	n_{22}	...	n_{2r}	$\sum_{j=1}^{r} n_{2j}$
...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kr}	$\sum_{j=1}^{r} n_{kj}$
	$\sum_{i=1}^{k} n_{i1}$	$\sum_{i=1}^{k} n_{i2}$	...	$\sum_{i=1}^{k} n_{i2}$	

Testem jest w tym wypadku statystyka: $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}\left(\frac{n_{ij}^2}{\widehat{n}_{ij}}\right) - n$, gdzie n to liczność próby , n_{ij} to zaobserwowane liczności z tabeli niezależności, natomiast $\widehat{n}_{ij}$ to teoretyczne liczności wystąpienia odpowiednich wariantów, gdyby zmienne *X* i *Y* były niezależne. Teoretyczne liczności oblicza się według wzoru: $\widehat{n}_{ij} = \sum_{i=1}^{k} n_{ij} \cdot \sum_{j=1}^{r} n_{ij}$

Dla zadanego poziomu istotności α, z tablic rozkładu χ^2 z $(r-1)(k-1)$ stopniami swobody odczytujemy liczbę χ_α^2 taką, że

$$P\left(\chi^2 \geq \chi_\alpha^2\right) = \alpha$$

Jeżeli $\chi_z^2 < \chi_\alpha^2$, to nie mamy podstaw do odrzucenia hipotezy H_0, natomiast gdy $\chi_z^2 > \chi_\alpha^2$, to hipotezę H_0 odrzucamy. χ_z^2 oznacza wartości statystyki testowej zaobserwowaną w próbie.

Empiryczny współczynnik korelacji i regresja liniowa

Niech (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) będą realizacjami zmiennej losowej dwuwymiarowej (X Y).
Empirycznym współczynnikiem korelacji nazywamy liczbę:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{X} \cdot \overline{Y}}{n \cdot S_X \cdot S_Y}$$

gdzie

$$S_X = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2} \quad \text{i} \quad S_Y = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y}\right)^2}$$

Powyższy współczynnik jest miernikiem siły związku prostoliniowego między dwoma cechami mierzalnymi X i Y.

Bezpośrednio z pojęciem korelacji wiąże się zagadnienie ***regresji.*** Polega ono na znalezieniu takiej linii o równaniu $y = f(x)$, aby suma kwadratów różnic pomiędzy wartościami zaobserwowanymi y_i i obliczonymi $f(x_i)$ była najmniejsza (***metoda najmniejszych kwadratów***). Najprostszą i najczęściej używaną funkcją w regresji jest funkcja liniowa. Mówimy wtedy o regresji linowej. Wtedy zależność między zmiennymi *X* i *Y* jest opisana funkcją liniową: $y = a \cdot x + b$, gdzie

$$a = r \cdot \frac{S_Y}{S_X} \quad \text{i} \quad b = \overline{Y} - a \cdot \overline{X}$$