# Statystyka, zadania

## Junior Data Scientist

Wojciech Artichowicz

# 1. Operacje na zbiorach i zdarzeniach

W rachunku prawdopodobieństwa operuje się zdarzeniami. Z punktu widzenia matematyki, zdarzenia są elementami zbiorów, podzbiorami lub całymi zbiorami. Jest mało prawdopodobne, że pracując jako data scientist będziesz rozważać zdarzenia i zbiory z czysto probabilistycznej perspektywy. Jednak częste będzie określanie szansy na wystąpienie pewnych przypadków mających znaczenie strategiczne. W praktyce dane zdarzenia, lub ich układy zwykle określa się na podstawie obserwacji istniejących w bazie danych (np. przy użyciu instrukcji SQL SELECT DISTINCT).

Niekiedy zachodzi konieczność określenia wszystkich możliwych zdarzeń, czyli przestrzeni zdarzeń elementarnych. W rachunku prawdopodobieństwa zazwyczaj interesująca jest tylko ich ilość, jednak w praktyce może być potrzebne wypisanie wszystkich elementów przestrzeni zdarzeń elementarnych.

### 1.1. Zadanie

Dane są następujące zbiory: $A = \{a, b, c, d\}$, $B = \{c, d, e\}$, $C = \{a, d\}$. Znajdź zbiory: $A \cup B$, $A \cap B$, $A - C$.

*Rozwiązanie:* Określenie sumy zbiorów $\cup$ polega na znalezieniu zbioru, który zawiera wszystkie elementy z jednego **lub** drugiego zbioru (lub większej liczby zbiorów). Ponadto każdy element zbioru może wystąpić w nim tylko raz.

$$A \cup B = \{a, b, c, d, e\}$$

Uwaga: Operację sumy zbiorów wykonuje m.in. polecenie UNION języka SQL, które łączy dwie kolumny wybierając unikatowe wartości.

W wyniku określenia iloczynu zbiorów otrzymuje się zbiór zawierający tylko elementy wspólne dla obu zbiorów. Jeśli takie elementy nie występują to wynikiem operacji jest zbiór pusty $\emptyset$.

$$A \cap B = \{c, d\}$$

Uwaga: Operację iloczynu zbiorów wykonuje m.in. polecenie INTERSECT języka SQL. W bazach, które nie oferują tej instrukcji można użyć INNER JOIN w połączeniu z SELECT DISTINCT, które łączy dwie tabele poprzez wartości jednocześnie występujące w obu tabelach. Gdyby `A` było kolumną tabeli `Table1`, natomiast `B` kolumną tabeli `Table2`, to zapytanie realizujące operację iloczynu zbiorów mogłoby wyglądać następująco:

```
(SELECT A FROM Table1)
INTERSECT
(SELECT B FROM Table2);
```

lub

```
SELECT DISTINCT(Table1.A) FROM Table1
INNER JOIN Table2
ON Table1.A = Table2.B;
```

| Table1 |
|---|
| **A** |
| a |
| b |
| c |
| d |

| Table2 |
|---|
| **B** |
| c |
| d |
| e |

| Table3 |
|---|
| **C** |
| a |
| d |

Różnicę zbiorów otrzymuje się tworząc zbiór zawierający wszystkie elementy zawarte w zbiorze pierwszym, ale nie obecne w zbiorze drugim.

$$A - C = \{b, c\}$$

Uwaga: Operację różnicy zbiorów wykonuje m.in. polecenie EXCEPT języka SQL. W bazach, które nie oferują tej instrukcji można użyć instrukcji SELECT DISTINCT, w połączeniu z warunkiem (WHERE NOT IN).

```
(SELECT A FROM Table1)
EXCEPT
(SELECT C FROM Table3);
```

lub

```
SELECT DISTINCT(A) FROM Table1
WHERE A NOT IN (SELECT C FROM Table3);
```

### 1.2. Zadanie

Pewna firma rozważa dołączanie do zakupu pakietu próbek kosmetyków. Do wyboru jest 10 różnych produktów, przy czym pakiet testowy ma się składać z próbek 3 różnych produktów. Ile jest możliwych pakietów próbek? Wypisz wszystkie rodzaje pakietów gratisowych.

*Rozwiązanie:* Zgodnie z założeniami produkty w pakiecie gratisów nie mogą się powtarzać. (Kolejność umieszczenia produktów w pakiecie nie ma znaczenia.) Liczbę możliwych gratisów można określić jako kombinację $k$ elementów z $N$ elementowego zbioru.

$$C_{k,N} = \binom{N}{k} = \frac{N!}{k!\,(N-k)!}$$

$$C_{3,10} = \binom{10}{3} = \frac{10!}{3!\,(10-3)!} = 120$$

Uwaga: Obliczenia tego typu można wykonać przy użyciu funkcji kombinatorycznych dostępnych w arkuszach kalkulacyjnych (np. `KOMBINACJE(N,k)` w Excelu) czy przy pomocy biblioteki scipy języka Python (`scipy.special.binom(n,k)`).

Wszystkie możliwe wyniki to (wygenerowano przy pomocy biblioteki `itertools` języka Python)

```
{(1, 2, 3), (1, 2, 4), (1, 2, 5), (1, 2, 6), (1, 2, 7), (1, 2, 8), (1, 2, 9),(1, 2, 10),
(1, 3, 4), (1, 3, 5), (1, 3, 6), (1, 3, 7), (1, 3, 8), (1, 3, 9),(1, 3, 10), (1, 4, 5),
(1, 4, 6), (1, 4, 7), (1, 4, 8), (1, 4, 9), (1, 4, 10),(1, 5, 6), (1, 5, 7), (1, 5, 8),
(1, 5, 9), (1, 5, 10), (1, 6, 7), (1, 6, 8),(1, 6, 9), (1, 6, 10), (1, 7, 8), (1, 7, 9),
(1, 7, 10), (1, 8, 9), (1, 8, 10),(1, 9, 10), (2, 3, 4), (2, 3, 5), (2, 3, 6), (2, 3, 7),
(2, 3, 8), (2, 3, 9),(2, 3, 10), (2, 4, 5), (2, 4, 6), (2, 4, 7), (2, 4, 8), (2, 4, 9),
(2, 4, 10),(2, 5, 6), (2, 5, 7), (2, 5, 8), (2, 5, 9), (2, 5, 10), (2, 6, 7), (2, 6, 8),
(2, 6, 9), (2, 6, 10), (2, 7, 8), (2, 7, 9), (2, 7, 10), (2, 8, 9), (2, 8, 10),(2, 9, 10),
(3, 4, 5), (3, 4, 6), (3, 4, 7), (3, 4, 8), (3, 4, 9), (3, 4, 10),(3, 5, 6), (3, 5, 7),
(3, 5, 8), (3, 5, 9), (3, 5, 10), (3, 6, 7), (3, 6, 8),(3, 6, 9), (3, 6, 10), (3, 7, 8),
(3, 7, 9), (3, 7, 10), (3, 8, 9), (3, 8, 10),(3, 9, 10), (4, 5, 6), (4, 5, 7), (4, 5, 8),
(4, 5, 9), (4, 5, 10), (4, 6, 7),(4, 6, 8), (4, 6, 9), (4, 6, 10), (4, 7, 8), (4, 7, 9),
(4, 7, 10), (4, 8, 9),(4, 8, 10), (4, 9, 10), (5, 6, 7), (5, 6, 8), (5, 6, 9), (5, 6, 10),
(5, 7, 8),(5, 7, 9), (5, 7, 10), (5, 8, 9), (5, 8, 10), (5, 9, 10), (6, 7, 8), (6, 7, 9),
(6, 7, 10), (6, 8, 9), (6, 8, 10), (6, 9, 10), (7, 8, 9), (7, 8, 10), (7, 9, 10),(8, 9, 10)}
```

## 2. Prawdopodobieństwo

### 2.1. Zadanie

Eksperyment polega na rzucie kostką sześcienną. Jakie jest prawdopodobieństwo otrzymania ściany z sześcioma oczkami (⁝⁝)?

*Rozwiązanie:* Zadanie jest trywialne, lecz dobrze ilustruje istotę prawdopodobieństwa. Jest to liczba zdarzeń sprzyjających do liczby wszystkich możliwych zdarzeń. Zakładając, że rozważamy zwykłą kostkę sześcienną to prawdopodobieństwo interesującego zdarzenia wyniesie

$$P(⁝⁝) = \frac{1}{6}.$$

### 2.2. Zadanie

Student umie odpowiedzieć na dwadzieścia pięć spośród trzydziestu pytań egzaminacyjnych. Znaleźć prawdopodobieństwo tego, że student odpowie w pełni poprawnie na zadane mu trzy pytania.

*Rozwiązanie:* Wszystkich pytań jest 30, zatem losując trzy razy student pomniejsza liczbę pytań możliwych do wylosowania za każdym razem o jeden (zatem wszystkich możliwości wylosowania pytania będzie $30 \cdot 29 \cdot 28$). Z kolei, jeśli student ma zdać to musi trzykrotnie wylosować pytanie ze zbioru pytań, na które zna odpowiedź ($25 \cdot 24 \cdot 23$). Zatem dzieląc liczbę zdarzeń sprzyjających przez liczbę wszystkich możliwości otrzymuje się

prawdopodobieństwo zakończenia egzaminu sukcesem. W celu określenia liczby zdarzeń stosuje się tu twierdzenie o mnożeniu.

$$P(zdania\ egzaminu) = \frac{25 \cdot 24 \cdot 23}{30 \cdot 29 \cdot 28} = \frac{115}{203}$$

Uwaga: Do obliczeń tego typu wygodnie jest stosować typ ułamkowy `Fraction` języka Python.

**2.3. Zadanie**

Prawdopodobieństwa zajścia każdego z dwóch zdarzeń niezależnych $A_1$ i $A_2$ są odpowiednio równe $p_1$ i $p_2$. Znaleźć prawdopodobieństwo zajścia tylko jednego z tych zdarzeń.

*Rozwiązanie:* Szukamy prawdopodobieństwa tego że zajdzie zdarzenie $A_1$ i nie zajdzie zdarzenie $A_2$ lub nie zajdzie zdarzenie $A_1$ i zajdzie zdarzenie $A_2$. Wyraźmy to przy pomocy rachunku zbiorów:

$$p = P((A_1 \cap A_2') \cup (A_1' \cap A_2)).$$

Na podstawie własności prawdopodobieństwa należy przekształcić powyższe wyrażenie tak, aby możliwe było wykorzystanie informacji, że $P(A_1)=p_1$ i $P(A_2)=p_2$.

$$p = P((A_1 \cap A_2') \cup (A_1' \cap A_2)) = P(A_1 \cap A_2') + P(A_1' \cap A_2) - P((A_1 \cap A_2') \cap (A_1' \cap A_2))$$

Z własności rachunku zbiorów wynika, że $(A_1 \cap A_2') \cap (A_1' \cap A_2) = \emptyset$ więc $P((A_1 \cap A_2') \cap (A_1' \cap A_2)) = 0$. Jeśli zdarzenia $A_1$ i $A_2$ są niezależne, to znaczy, że $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$. Analogiczne zależności są spełnione dla zdarzeń przeciwnych więc:

$$p = P(A_1) \cdot P(A_2') + P(A_1') \cdot P(A_2)$$

Wiadomo, że dla dowolnego zdarzenia zachodzi $P(\bar{A}) + P(A) = 1$ więc:

$$p = p_1 \cdot (1 - p_2) + (1 - p_1) \cdot p_2 = (p_1 - p_1 \cdot p_2) + (p_2 - p_1 \cdot p_2) = p_1 + p_2 - 2 \cdot p_1 \cdot p_2.$$

**2.4. Zadanie**

W komorze reakcji konieczne jest zainstalowanie czujnika wykrywającego niepoprawny przebieg reakcji chemicznej. Na rynku dostępne są dwa typu urządzeń realizujących to zadanie. Jedno wykrywa nieprawidłowy przebieg reakcji z prawdopodobieństwem $p_1$=0,98 drugie z prawdopodobieństwem $p_2$=0,95. Zakładając, że w komorze zamontowane zostaną dwa czujniki oblicz:

a) prawdopodobieństwo, że w przypadku wystąpienia awarii tylko jeden z czujników ją wykryje;

b) prawdopodobieństwo, że w przypadku wystąpienia awarii, zostanie ona wykryta.

*Rozwiązanie*:

a) Jeśli czujniki są różnego typu lub pochodzą od dwóch różnych producentów to można założyć, że działają niezależnie. Wówczas prawdopodobieństwo, że tylko jeden czujnik wykryje awarię można obliczyć ze wzoru wyprowadzonego w zadaniu 2.3.

$$p = p_1 + p_2 - 2p_1 \cdot p_2 = 0{,}068$$

b) Natomiast prawdopodobieństwo tego, że awaria zostanie w ogóle wykryta wyniesie

$$p = p_1 + p_2 - p_1 \cdot p_2 = 0{,}999$$

Uwaga: Należy zwrócić uwagę na to, że gdyby zamontować dwa identyczne urządzenia to mogłyby one się zachować identycznie w takiej samej sytuacji (tu: wystąpienia awarii). Wówczas należałoby uznać, że zdarzenia wykrycia awarii przez jedno i drugie urządzenie są w pełni zależne, a co za tym idzie: pomimo tego, że zamontowane zostały dwa urządzenia pierwszego typu to prawdopodobieństwo wykrycia awarii byłoby równe w

dalszym ciągu 0,98, a gdyby oba urządzenia były drugiego typu to 0,95. Zatem można wnioskować, że zestawienie dwóch urządzeń pracujących niezależnie daje znacznie lepsze wyniki. Takie proste rachunki często można wykorzystać w celu optymalizacji relacji koszt/niezawodność.

# 3. Zmienne losowe

## 3.1. Dyskretna zmienna losowa

### 3.1.1. Zadanie

Narysować wykres rozkładu prawdopodobieństwa i dystrybuanty zmiennej losowej $X$, dla której: $P(X=0)=\frac{1}{10}$, $P(X=1)=\frac{9}{10}$. Obliczyć jej wartość oczekiwaną, wariancję i odchylenie standardowe.

*Rozwiązanie*: Wykres rozkładu prawdopodobieństwa zmiennej losowej dyskretnej tworzy się zaznaczając na osi poziomej punkty (tu $X$=0 oraz $X$=1), w których określono wartości prawdopodobieństwa (tu 1/10 i 9/10).

Wartość oczekiwana dana jest wzorem:

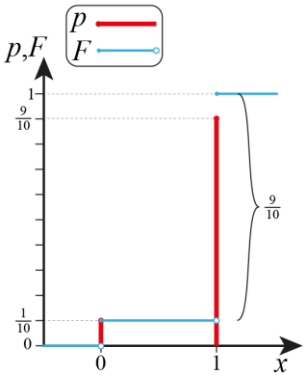
$$E(X)=\sum_{x\in D} x\cdot P(x)$$

czyli

$$E(X)=0\cdot\frac{1}{10}+1\cdot\frac{9}{10}=\frac{9}{10}$$

Wariancja opisana jest wzorem:

$$V(X)=E\left((X-E(X))^2\right)=\sum_{x\in D}(x-E(X))^2\cdot p(x)$$




$$V(X)=\left(0-\frac{9}{10}\right)^2\cdot\frac{1}{10}+\left(1-\frac{9}{10}\right)^2\cdot\frac{9}{10}=\frac{90}{1000}=\frac{9}{100}$$

Odchylenie standardowe jest pierwiastkiem z wariancji:

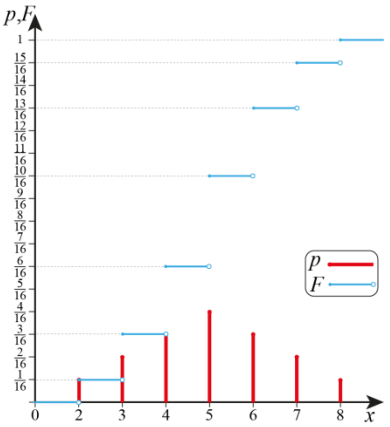
$$s(X)=\sqrt{V(X)}=\sqrt{\frac{9}{100}}=\frac{3}{10}$$

### 3.1.2. Zadanie

Znaleźć rozkład prawdopodobieństwa, dystrybuantę, wartość oczekiwaną i wariancję zmiennej losowej $X$ – sumy oczek otrzymanej w wyniku rzutu dwiema czworościennymi kostkami do gry.

*Rozwiązanie*: Najłatwiej określić prawdopodobieństwa przynależące danym wartościom zmiennej losowej tworząc tabelkę i zapisując w niej zdarzenia sprzyjające.

| $x$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| $\omega$ | 1+1 | 1+2,<br>2+1 | 1+3,<br>2+2,<br>3+1 | 1+4,<br>2+3,<br>3+2,<br>4+1 | 2+4,<br>3+3,<br>4+2 | 3+4,<br>4+3 | 4+4 |
| $p$ | 1/16 | 2/16 | 3/16 | 4/16 | 3/16 | 2/16 | 1/16 |
| $F$ | 1/16 | 3/16 | 6/16 | 10/16 | 13/16 | 15/16 | 16/16 |




Obliczenie wartości oczekiwanej:

$$E(X)=\sum_{x\in D} x\cdot p(x)=2\cdot\tfrac{1}{16}+3\cdot\tfrac{2}{16}+\ldots+8\cdot\tfrac{1}{16}=5\cdot$$

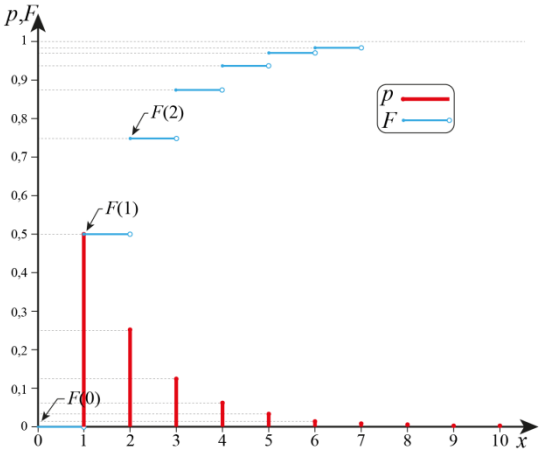
Obliczenie wariancji:

$$V(X)=\sum_{x\in D}(x-E(X))^2\cdot p(x)=(2-5)^2\cdot\tfrac{1}{16}+(3-5)^2\cdot\tfrac{2}{16}+\ldots+(8-5)^2\cdot\tfrac{1}{16}=\tfrac{5}{2}$$

### 3.1.3. Zadanie

Narysować wykres rozkładu prawdopodobieństwa i dystrybuanty $F(x)$ zmiennej losowej $X$, której funkcja rozkładu prawdopodobieństwa jest dana następującym wzorem $P(X=k)=\frac{1}{2^k}$, $k$ = 1, 2, 3, ... Oblicz wartość oczekiwaną oraz wariancję.

*Rozwiązanie*: Dystrybuanta sumuje wszystkie prawdopodobieństwa do $x$ włącznie zatem będzie określona wzorem: $F(x)=\sum_{i=1}^{x}\frac{1}{2^x}$.




*Uwaga: Na wykres nie naniesiono całej dystrybuanty, a prawdopodobieństwo narysowano tylko dla x=0,1 ,2, ..., 10.*

Wartość oczekiwana dana jest wzorem $E(X)=\sum_{x\in D} x\cdot P(x)\cdot$. Po wstawieniu wzoru opisującego prawdopodobieństwo otrzymuje się:

$$E(X)=\sum_{x=1}^{\infty} x\cdot\frac{1}{2^x}\cdot$$

Jak widać powyżej, wartością oczekiwaną jest suma szeregu liczbowego równa 2 więc

$$E(X)=\sum_{x=1}^{\infty}\frac{x}{2^x}=2\cdot$$

*Wynik obliczenia powyższej sumy również można otrzymać korzystając np. z portalu Wolfram Alpha.*

Obliczając wariancję korzystamy ze wzoru $V(X)=\sum_{x\in D}(x-E(X))^2\cdot p(x)$, zatem:

$$V(X)=\sum_{x=1}^{\infty}\frac{(x-2)^2}{2^x}=2\cdot$$

*Wynik z portalu Wolfram Alpha.*

## 3.2. Ciągła zmienna losowa

### 3.2.1. Zadanie

Dana jest dystrybuanta zmiennej losowej $F(X)$ [Rozkład Cauchy]. Znaleźć prawdopodobieństwo, że w wyniku próby $X$ przyjmie wartość należącą do przedziału (-1,1). Znajdź gęstość rozkładu $f(x)$ tej zmiennej losowej. Otrzymane rozwiązanie zaznacz na wykresie rozkładu i dystrybuanty.

$$F(x)=\begin{cases}0 & x\le -2\\ \frac{1}{2}+\frac{1}{\pi}\arcsin\left(\frac{x}{2}\right) & -2<x\le 2\\ 1 & x>2\end{cases}$$

*Rozwiązanie*: Obliczenie prawdopodobieństwa zdarzenia polegającego na tym, że zmienna losowa ciągła przyjmie wartość z przedziału $(a,b)$ umożliwia relacja:

$$P(a<X<b)=F(b)-F(a),$$

zatem prawdopodobieństwo, że zmienna losowa $X$ osiągnie wartość z przedziału (-1,1) wynosi:

$$P(-1<X<1)=F(1)-F(-1)$$

$$P(-1<X<1)=\left(\tfrac{1}{2}+\tfrac{1}{\pi}\arcsin\left(\tfrac{1}{2}\right)\right)-\left(\tfrac{1}{2}+\tfrac{1}{\pi}\arcsin\left(\tfrac{-1}{2}\right)\right)=$$

$$=\tfrac{1}{\pi}\left(\arcsin\left(\tfrac{1}{2}\right)-\arcsin\left(-\tfrac{1}{2}\right)\right)=\tfrac{1}{\pi}\left(\tfrac{\pi}{6}-\left(-\tfrac{\pi}{6}\right)\right)=\tfrac{2}{6}=\tfrac{1}{3}\cdot$$

Dystrybuanta i gęstość rozkładu są związane zależnością: $F'(x)=f(x)$, czyli gęstość rozkładu jest pochodną dystrybuanty. W związku z tym konieczne jest obliczenie pochodnej dystrybuanty. Funkcja ta składa się z trzech elementów, w związku z tym konieczne jest obliczenie pochodnej w każdym przypadku:

$$F'(x)=\begin{cases}0' & x\le -2\\ \left(\frac{1}{2}+\frac{1}{\pi}\arcsin\left(\frac{x}{2}\right)\right)' & -2<x\le 2\\ 1' & x>2\end{cases}$$

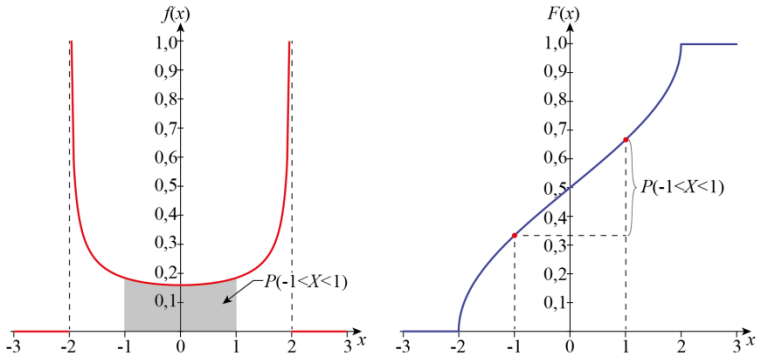
Pochodną ze stałej jest zero więc $0'$ oraz $1'$ są równe zeru. Pochodna funkcji opisującej dystrybuantę w zakresie od -2 do 2 jest następująca:

$$\left(\tfrac{1}{2}+\tfrac{1}{\pi}\arcsin\left(\tfrac{x}{2}\right)\right)'=\tfrac{1}{\pi}\frac{1}{\sqrt{1-\left(\frac{x}{2}\right)^2}}\cdot\tfrac{1}{2}=\frac{1}{2\pi\sqrt{1-\left(\frac{x}{2}\right)^2}}\cdot$$

Ostatecznie otrzymuje się wzór opisujący gęstość prawdopodobieństwa:

$$F'(x)=f(x)=\begin{cases}0 & x\le -2\\ \dfrac{1}{2\pi\sqrt{1-\left(\frac{x}{2}\right)^2}} & -2<x\le 2\\ 0 & x>2\end{cases}$$

Poniżej przedstawiono wykresy funkcji gęstości prawdopodobieństwa i dystrybuanty zmiennej losowej $X$.




### 3.2.2. Zadanie

Dystrybuanta zmiennej losowej ciągłej $X$ – czasu bezawaryjnej serwera komputerowego, jest opisana wzorem: $F(x)=1-e^{-x/T}$ [Rozkład ekspotencjalny]. Znaleźć prawdopodobieństwo bezawaryjnej pracy urządzenia w czasie dłuższym niż $T$.

*Rozwiązanie*: Dystrybuanta $F(x)$ (na podstawie definicji) określa prawdopodobieństwo tego, że zmienna losowa osiągnie wartość mniejszą lub równą $x$. Czyli $F(T)$ oznacza prawdopodobieństwo zdarzenia $A$ polegającego na tym, że serwer będzie pracował przez czas $T$ lub krócej. Zdarzenie polegające na bezawaryjnej pracy urządzenia w czasie dłuższym niż $T$, jest zdarzeniem przeciwnym do $A$, czyli $A'$. Zatem szukane prawdopodobieństwo można obliczyć jako:

$$P(A')=1-P(A)=1-F(T)$$

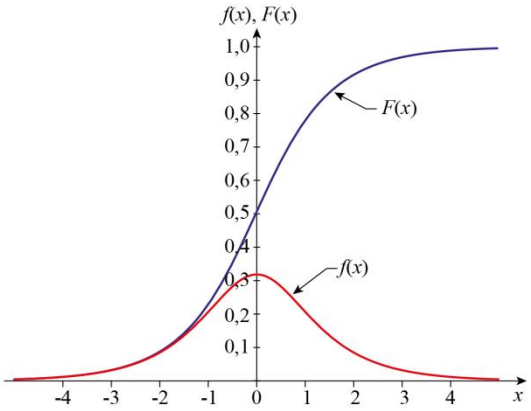
$$P(A')=1-\left(1-e^{-T/T}\right)=1-1+e^{-1}=\frac{1}{e}$$

### 3.2.3. Zadanie

Gęstość prawdopodobieństwa zmiennej losowej ciągłej $X$ określona jest wzorem $f(x)=\frac{2}{\pi\cdot\left(e^{x}+e^{-x}\right)}$. Znaleźć dystrybuantę tej zmiennej losowej oraz kwantyle $Q_{0,1}$, $Q_{0,5}$, $Q_{0,9}$. Zaznacz je na wykresie gęstości oraz dystrybuanty.

*Rozwiązanie*: Dystrybuanta zmiennej losowej wyrażona jest wzorem $F(x)=\int_{-\infty}^{x} f(t)\,dt$. Zatem można zapisać:

$$F(x)=\frac{2}{\pi}\int_{-\infty}^{x}\frac{1}{e^{t}+e^{-t}}\,dt=\frac{2}{\pi}\left[\arctan(e^{t})\right]_{-\infty}^{x}=\frac{2}{\pi}\left(\arctan(e^{x})-0\right)=\frac{2}{\pi}\arctan(e^{x})\cdot$$




Kwantyl $Q_p$ jest wartością dzielącą pole pod wykresem gęstości prawdopodobieństwa w stosunku $p$:1-$p$. Z definicji kwantyl opisany jest zależnością:

$$\int_{-\infty}^{Q_p} f(x)\,dx = p$$

Zauważmy, że znając dystrybuantę, nie jest konieczne obliczanie całki, gdyż można zapisać:

$$F(Q_p) = \int_{-\infty}^{Q_p} f(x)\,dx = p,$$

$$F(Q_p) = p.$$

Zależność tę można interpretować jako znaną rzędną ($y$) na wykresie dystrybuanty, dla której należy znaleźć odciętą ($x$=$Q_p$). Dystrybuanta zmiennej losowej $X$ opisana jest wzorem $F(x) = \frac{2}{\pi}\arctan(e^x)$, zatem można zapisać, że

$$\frac{2}{\pi}\arctan(e^{Q_p}) = p.$$

Po rozwiązaniu powyższego równania otrzymamy wzór na kwantyl rzędu $p$.

$$\arctan\left(e^{Q_p}\right) = \frac{\pi \cdot p}{2}$$

$$e^{Q_p} = \tan\left(\frac{\pi \cdot p}{2}\right)$$

$$\ln(e^{Q_p}) = \ln\left(\tan\left(\frac{\pi \cdot p}{2}\right)\right)$$

$$Q_p \cdot \ln(e) = \ln\left(\tan\left(\frac{\pi \cdot p}{2}\right)\right)$$

$$Q_p = \ln\left(\tan\left(\frac{\pi \cdot p}{2}\right)\right).$$

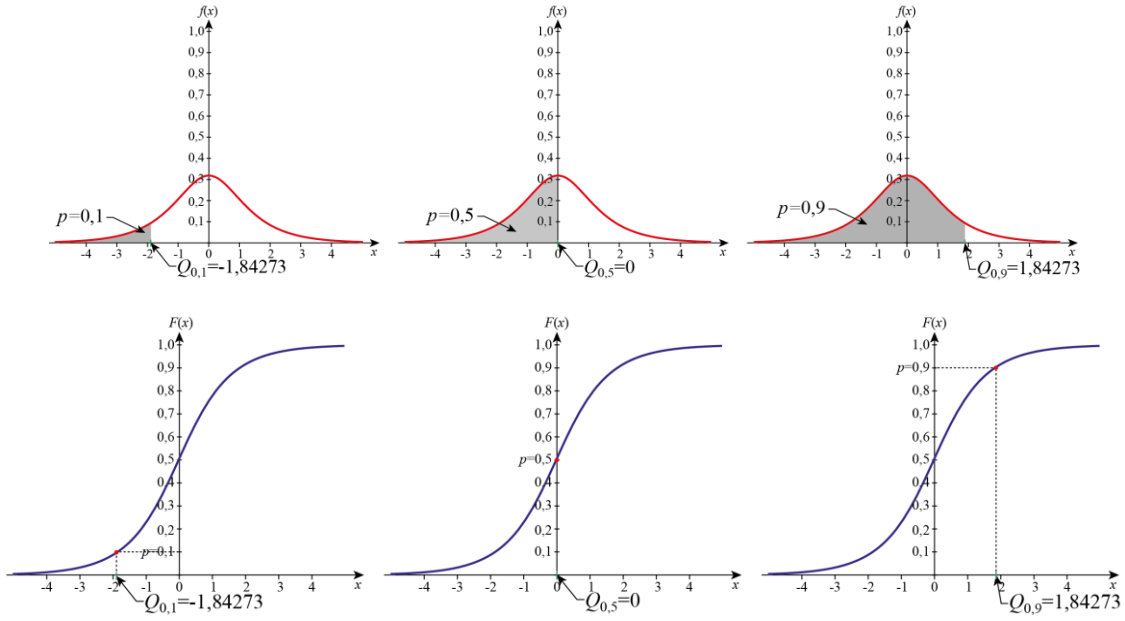
Żeby obliczyć wartości $Q_{0,1}$, $Q_{0,5}$, $Q_{0,9}$ konieczne jest podstawienie do powyższego wzoru odpowiednich wartości $p$:

$$Q_{0,1} = \ln\left(\tan\left(\frac{\pi \cdot 0{,}1}{2}\right)\right) = -1{,}84273;$$

$$Q_{0,5} = \ln\left(\tan\left(\frac{\pi \cdot 0{,}5}{2}\right)\right) = 0;$$

$$Q_{0,9} = \ln\left(\tan\left(\frac{\pi \cdot 0{,}9}{2}\right)\right) = 1{,}84273.$$

*Uwaga: W związku z tym, że funkcja gęstości prawdopodobieństwa f(x) jest symetryczna względem zera, wartości kwantyli $Q_{0,1}$ i $Q_{0,9}$ (czyli $Q_p$ oraz $Q_{1-p}$) są takie same, ale mają przeciwne znaki.*




# 4. Statystyka opisowa

### 4.1. Zadanie

W wyniku eksperymentu otrzymano następującą realizację zmiennej losowej *X*:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3,8 | 2,0 | 1,7 | 5,6 | 6,9 | 8,9 | 2,2 | 4,5 | 0,8 | 0,3 |

Na podstawie próby oblicz: $\bar{X}$ - średnią arytmetyczną, *s* – odchylenie standardowe, *d* -odchylenie przeciętne, *Me* – medianę, *R* - rozstęp. Określ dystrybuantę empiryczną.

*Rozwiązanie*: Średnią arytmetyczną oblicza się na podstawie wzoru $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, w którym *n* oznacza liczebność próby, a $x_i$ jest *i*-tą wartością. Próba liczy dziesięć elementów więc *n*=10:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{3{,}8+2{,}0+1{,}7+5{,}6+6{,}9+8{,}9+2{,}2+4{,}5+0{,}8+0{,}3}{10} = 3{,}67.$$

Odchylenie standardowe jest pierwiastkiem z wariancji, która dana jest wzorem

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}.$$

$$s^2 = \frac{1}{10-1}\sum_{i=1}^{10}(x_i-\bar{x})^2 =$$

$$= \frac{(3{,}8-3{,}67)^2+(2{,}0-3{,}67)^2+(1{,}7-3{,}67)^2+(5{,}6-3{,}67)^2+(6{,}9-3{,}67)^2+(8{,}9-3{,}67)^2+(2{,}2-3{,}67)^2+(4{,}5-3{,}67)^2+(0{,}8-3{,}67)^2+(0{,}3-3{,}67)^2}{9} = 7{,}849$$

Wynika z tego, że $s = \sqrt{s^2} = \sqrt{7{,}849} = 2{,}80161$.

Odchylenie przeciętne dane jest wzorem

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}.$$

$$d = \frac{1}{10}\sum_{i=1}^{10}|x_i - \bar{x}| = \frac{|3{,}8 - 3{,}67| + |2{,}0 - 3{,}67| + \ldots + |0{,}3 - 3{,}67|}{10} = 2{,}27$$

Aby uprościć dalsze obliczenia wygodnie jest uporządkować elementy próby w porządku rosnącym:

| 0,3 | 0,8 | 1,7 | 2,0 | 2,2 | 3,8 | 4,5 | 5,6 | 6,9 | 8,9 |
|---|---|---|---|---|---|---|---|---|---|

Mediana jest wartością środkową z uporządkowanej próby. Jeśli liczebność próby jest parzysta to mediana jest średnią arytmetyczną z dwóch środkowych wartości. Jeśli liczebność próby jest nieparzysta, to medianę stanowi element środkowy. Badana próba liczy dziesięć elementów więc medianę należy obliczyć jako średnią elementów piątego i szóstego.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 0,3 | 0,8 | 1,7 | 2,0 | **2,2** | **3,8** | 4,5 | 5,6 | 6,9 | 8,9 |

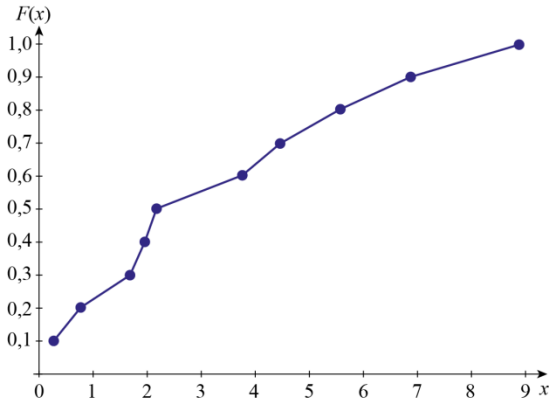
$$Me = \frac{x_{\lceil n/2 \rceil} + x_{\lceil n/2 \rceil + 1}}{2} = \frac{2{,}2 + 3{,}8}{2} = 3$$

Rozstęp jest różnicą wartości największej i najmniejszej. W próbie uporządkowanej rosnąco element pierwszy ma wartość najmniejszą, a element ostatni największą: $x_{min}$=0,3, $x_{max}$=8,9.

$$R = x_{\max} - x_{\min} = 8{,}9 - 0{,}3 = 8{,}6$$

Dystrybuantę empiryczną tworzy się sumując częstości względne osiągnięcia przez zmienną losową wartości z danej klasy. W tym przypadku próba jest tak mała, że grupowanie spostrzeżeń w klasy nie ma sensu. W związku z tym skumulowane częstości względne (czyli dystrybuantę empiryczną $F(x_i)$) należy zaznaczyć w punktach zdefiniowanych przez spostrzeżenia (czyli wartości z próby - $x_i$).

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $F(x_i)$ | 1/10 | 2/10 | 3/10 | 4/10 | 5/10 | 6/10 | 7/10 | 8/10 | 9/10 | 1 |
| $x_i$ | 0,3 | 0,8 | 1,7 | 2,0 | **2,2** | **3,8** | 4,5 | 5,6 | 6,9 | 8,9 |



**4.2. Zadanie**

W wyniku eksperymentu otrzymano próbę o liczności 1000 elementów (surowe dane: plik `X.txt`) będącą realizacją zmiennej losowej $X$ o gęstości rozkładu prawdopodobieństwa opisanej wzorem $f(x)=\frac{2}{\pi\cdot\left(e^{x}+e^{-x}\right)}$ i dystrybuancie $F(x)=\frac{2}{\pi}\arctan(e^{x})$. Dane przedstaw w postaci rozkładu częstości, częstotliwości (czyli prawdopodobieństwa empirycznego), gęstości prawdopodobieństwa i dystrybuanty empirycznej. Oblicz średnią arytmetyczną z próby, i odchylenie standardowe dla danych surowych oraz dla danych wstępnie przetworzonych – dokonaj interpretacji wyników. Porównaj otrzymane wartości z odpowiadającymi im parametrami rozkładu zmiennej losowej $X$. Na podstawie próby oraz na podstawie modelu teoretycznego oblicz prawdopodobieństwo zdarzenia polegającego na tym, że zmienna losowa przyjmie wartość z przedziału [0,2].

*Rozwiązanie*: Aby określić rozkład częstości konieczne jest odnalezienie wartości najmniejszej, największej oraz rozstępu w próbie:

$$x_{\min}=-6{,}29231;$$

$$x_{\max}=8{,}3632;$$

$$R=x_{\max}-x_{\min}=14{,}6555.$$

Kolejnym krokiem jest przyjęcie sposobu określenia ilości i długości klas. Jednym ze sposobów wstępnego obliczenia szerokości klasy jest wykorzystanie wzoru:

$$h^{*}=\frac{2{,}64\cdot(Q_{0{,}75}-Q_{0{,}25})}{\sqrt[3]{N}}.$$

Aby obliczyć szerokość klasy konieczne jest znalezienie kwantyli $Q_{0,25}$ oraz $Q_{0,75}$. W tym celu należy uszeregować elementy próby rosnąco i znaleźć elementy dzielące próbę w stosunku 0,25:0,75 (dla $Q_{0,25}$) oraz 0,75:0,25 (dla $Q_{0,75}$). Liczebność próby jest parzysta i wynosi $N$=1000 elementów więc kwantyl 0,25 będzie średnią arytmetyczną z elementów dwieście pięćdziesiątego oraz dwieście pięćdziesiątego pierwszego, kwantyl 0,75 będzie średnią arytmetyczną z elementów siedemset pięćdziesiątego oraz siedemset pięćdziesiątego pierwszego:

| $i$ | $x_i$ |
|---|---|
| ... | ... |
| 250 | -0,891831 |
| 251 | -0,890837 |
| ... | ... |
| 750 | 0,888402 |
| 751 | 0,891416 |
| ... | ... |

$$Q_{0{,}25}=\frac{-0{,}891831+(-0{,}890837)}{2}=-0{,}891334,$$

$$Q_{0{,}75}=\frac{0{,}888402+0{,}891416}{2}=0{,}889909.$$

Zatem różnica wyniesie $Q_{0,75}$- $Q_{0,25}$= 1,78124, a szerokość klasy będzie równa:

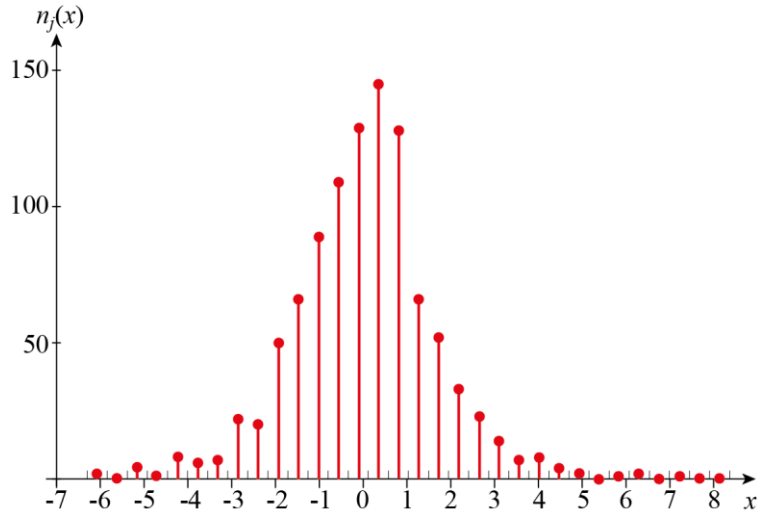
$$h^{*}=\frac{2{,}64\cdot(Q_{0{,}75}-Q_{0{,}25})}{\sqrt[3]{N}}=\frac{2{,}64\cdot 1{,}78124}{\sqrt[3]{1000}}=0{,}470248$$

Aby znaleźć wstępną liczbę klas, należy rozstęp podzielić przez szerokość klasy: $k^{*}=R/h$=31,1655. Wynika z tego, że rozkład należy podzielić na $k$=32 klasy więc ostatecznie szerokość klasy będzie równa: $h=R/k$=0,457985. Granice pierwszej klasy określa się dodając do wartości minimalnej szerokość klasy. Otrzymana wartość, jest końcem klasy pierwszej i zarazem początkiem klasy drugiej. Do tej wartości dodaje się szerokość klasy otrzymując koniec klasy drugiej. Itd. Wyniki są przedstawione w tabeli w kolumnie $x_{p,j}$-$x_{k,j}$.

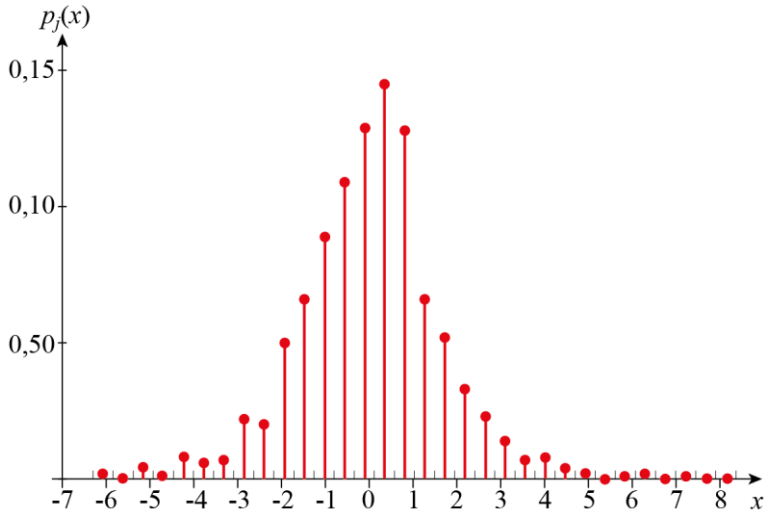
Następnym krokiem jest zliczenie liczby elementów próby, które znalazły się w danej klasie. (Należy zwrócić uwagę na to, że ostatnia klasa jest przedziałem domkniętym obustronnie, a wcześniejsze są przedziałami domkniętymi lewostronnie.) Liczbę elementów znajdujących się w danej klasie wpisano do tabeli przedstawionej poniżej w kolumnie z nagłówkiem $n_j$. Prawdopodobieństwa zdarzenia polegającego na tym, że zmienna losowa przyjmie wartość należącą do danej klasy oblicza się dzieląc częstość $n_j$ przez liczbę elementów w próbie $N$. Gęstość prawdopodobieństwa empirycznego w danej klasie uzyskuje się dzieląc $p_j$ przez szerokość danej klasy $h$ (tu wszystkie klasy mają taką samą szerokość). Dystrybuanta empiryczna $F_j$ z kolei jest sumą wszystkich prawdopodobieństw dla $i\leq j$. Wyniki obliczeń przedstawiono w poniższej tabeli.

| $j$ | $x_{p,j}$-$x_{k,j}$ | $x_{sj}$ | $n_j$ | $p_j=n_j/N$ | $f_j=p_j/h$ | $F_j=\sum_{i\le j} p_i$ |
|---|---|---|---|---|---|---|
| 1 | [-6,29231 ; -5,83432) | -6,06331 | 2 | 0,002 | 0,00436696 | 0,002 |
| 2 | [-5,83432 ; -5,37634) | -5,60533 | 0 | 0,0 | 0,0 | 0,002 |
| 3 | [-5,37634 ; -4,91835) | -5,14734 | 4 | 0.004 | 0,00873392 | 0,006 |
| 4 | [-4,91835 ; -4,46037) | -4,68936 | 1 | 0,001 | 0,00218348 | 0,007 |
| 5 | [-4,46037 ; -4,00238) | -4,23137 | 8 | 0,008 | 0,0174678 | 0,015 |
| 6 | [-4,00238 ; -3,5444) | -3,77339 | 6 | 0,006 | 0,0131009 | 0,021 |
| 7 | [-3,5444 ; -3,08641) | -3,3154 | 7 | 0,007 | 0,0152844 | 0,028 |
| 8 | [-3,08641 ; -2,62843) | -2,85742 | 22 | 0,022 | 0,0480365 | 0,050 |
| 9 | [-2,62843 ; -2,17044) | -2,39944 | 20 | 0,02 | 0,0436696 | 0,070 |
| 10 | [-2,17044 ; -1,71246) | -1,94145 | 50 | 0,05 | 0,109174 | 0,120 |
| 11 | [-1,71246 ; -1,25447) | -1,48347 | 66 | 0,066 | 0,14411 | 0,186 |
| 12 | [-1,25447 ; -0,796489) | -1,02548 | 89 | 0,089 | 0,19433 | 0,275 |
| 13 | [-0,796489; -0,338504) | -0,567497 | 109 | 0,109 | 0,237999 | 0,384 |
| 14 | [-0,338504; 0,11948) | -0,109512 | 129 | 0,129 | 0,281669 | 0,513 |
| 15 | [0,11948 ; 0,577465) | 0,348473 | 145 | 0,145 | 0,316604 | 0,658 |
| 16 | [0,577465; 1,03545) | 0,806457 | 128 | 0,128 | 0,279485 | 0,786 |
| 17 | [1,03545 ; 1,49343) | 1,26444 | 66 | 0,066 | 0,14411 | 0,852 |
| 18 | [1,49343 ; 1,95142) | 1,72243 | 52 | 0,052 | 0,113541 | 0,904 |
| 19 | [1,95142 ; 2,4094) | 2,18041 | 33 | 0,033 | 0,0720548 | 0,937 |
| 20 | [2,4094 ; 2,86739) | 2,6384 | 23 | 0,023 | 0,05022 | 0,960 |
| 21 | [2,86739 ; 3,32537) | 3,09638 | 14 | 0,014 | 0,0305687 | 0,974 |
| 22 | [3,32537 ; 3,78336) | 3,55437 | 7 | 0,007 | 0,0152844 | 0,981 |
| 23 | [3,78336 ; 4,24134) | 4,01235 | 8 | 0,008 | 0,0174678 | 0,989 |
| 24 | [4,24134 ; 4,69933) | 4,47033 | 4 | 0,004 | 0,00873392 | 0,993 |
| 25 | [4,69933 ; 5,15731) | 4,92832 | 2 | 0,002 | 0,00436696 | 0,995 |
| 26 | [5,15731 ; 5,6153) | 5,3863 | 0 | 0,0 | 0,0 | 0,995 |
| 27 | [5,6153 ; 6,07328) | 5,84429 | 1 | 0,001 | 0,00218348 | 0,996 |
| 28 | [6,07328 ; 6,53127) | 6,30227 | 2 | 0,002 | 0,00436696 | 0,998 |
| 29 | [6,53127 ; 6,98925) | 6,76026 | 0 | 0,0 | 0,0 | 0,998 |
| 30 | [6,98925 ; 7,44724) | 7,21824 | 1 | 0,001 | 0,00218348 | 0,999 |
| 31 | [7,44724 ; 7,90522) | 7,67623 | 0 | 0,0 | 0,0 | 0,999 |
| 32 | [7,90522 ; 8,3632] | 8,13421 | 1 | 0,001 | 0,00218348 | 1,000 |

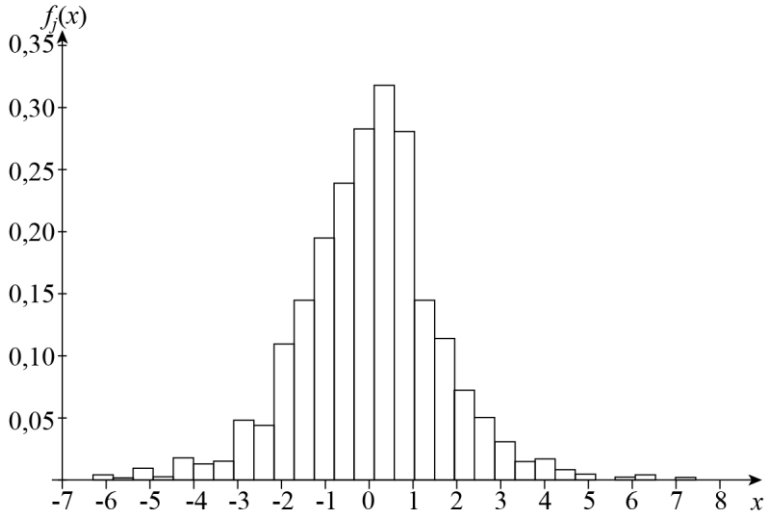
Wykres częstości konstruuje się odnosząc na osi $y$ liczbę elementów próby zaliczonych do danego przedziału w wartości odpowiadającej środkowi przedziału $x_{s,j}$. Np. (-6,06331; 2), (-5,60533; 0), (-5,14734; 4) itd.



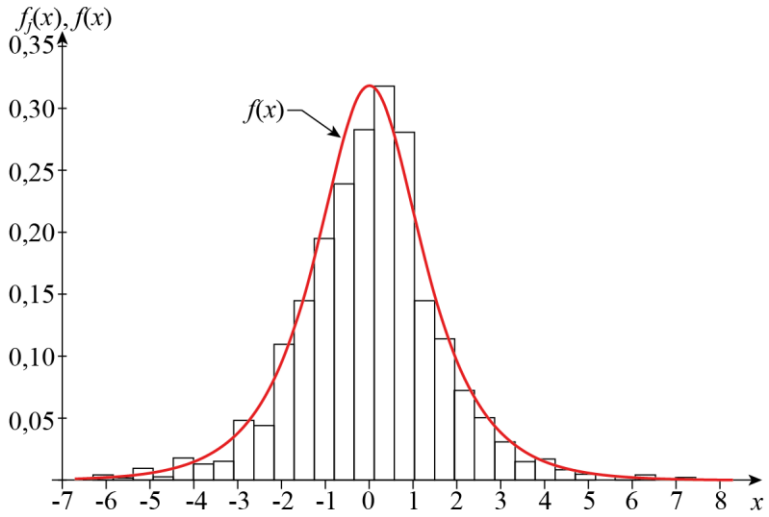
Wykres prawdopodobieństwa empirycznego powstaje poprzez odniesienie na osi $y$ wartości $n_j$ podzielonych przez liczebność próby $N$.



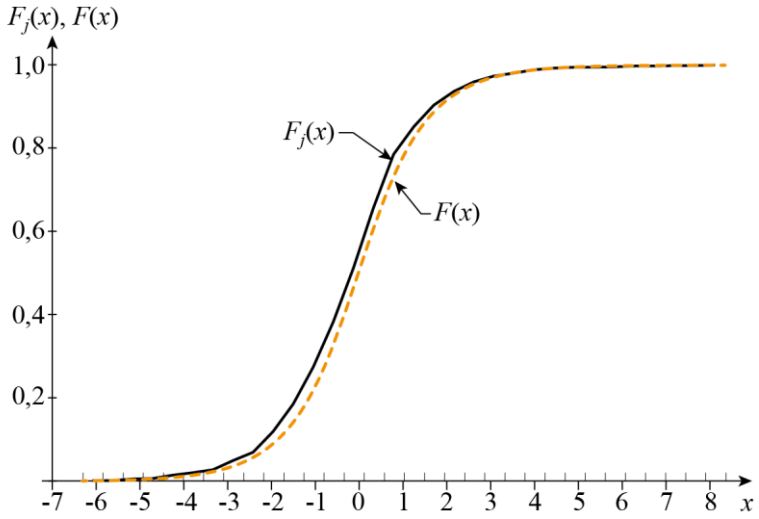

Wykres gęstości prawdopodobieństwa tworzy się rysując słupki o szerokości przedziałów i wysokości określonej przez wartości $f_j$.




Nanosząc na jednym wykresie empiryczną i teoretyczną gęstość prawdopodobieństwa można wizualnie porównać gęstość rozkładu z próby z gęstością rozkładu teoretycznego.




Poniższy wykres przedstawia dystrybuantę empiryczną oraz dystrybuantę rozkładu teoretycznego.




Średnią arytmetyczną i odchylenie standardowe z próby surowej oblicza się na podstawie wzorów $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ oraz $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ :

$$\bar{x} = 0{,}0328656$$

$$s = 1{,}6244$$

Parametry z próby przetworzonej oblicza się jako wartości ważone, gdzie wagami są częstości $n_j$, a wartości $x_i$ zastąpione są wartościami określającymi środki przedziałów $x_{s,j}$.

$$\bar{x} = \frac{\sum_{j=1}^{k} x_{s_j} \cdot n_j}{\sum_{j=1}^{k} n_j} = 0{,}0310892$$

$$s = \sqrt{\frac{\sum_{j=1}^{k}(x_{s_j} - \bar{x})^2 \cdot n_j}{\sum_{j=1}^{k} n_j - 1}} = 1{,}63003$$

Parametry z rozkładu teoretycznego, z którego pochodzi próba to wartość oczekiwana oraz odchylenie standardowe:

$$E(X) = 0$$

$$S(X) = \frac{\pi}{2}$$

Wartości średniej arytmetycznej i odchylenia standardowego obliczone na podstawie próby surowej oraz na podstawie próby przetworzonej różnią się. W wyniku uogólniania danych (np. tworzenie histogramu częstości) część informacji zostaje utracona. Dzieje się tak, ponieważ zamiast konkretnych wartości realizacji eksperymentu ($x_i$) używa się środków klas ($x_{s,j}$) i częstości wystąpienia wartości z danego przedziału ($n_j$).

Wiadomo, że wartość oczekiwana z rozkładu odpowiada wartości oczekiwanej z próby, która pochodzi z tego rozkładu. W związku z tym wartości otrzymane na podstawie analizy próby powinny być zbliżone do wartości teoretycznych z rozkładu. Identyczny wniosek dotyczy pozostałych momentów, np. odchylenia standardowego.

Aby obliczyć prawdopodobieństwo tego, że dana zmienna losowa ciągła przyjmie wartość z pewnego przedziału $[a,b]$ najwygodniej jest wykorzystać własności jej dystrybuanty $P(a < X < b) = F(b) - F(a)$. Dla rozważanego przykładu:

$$P(0 < X < 2) = F(2) - F(0) = \tfrac{2}{\pi}\arctan(e^2) - \tfrac{2}{\pi}\arctan(e^0) = 0{,}414363.$$

Podobnie można postąpić z dystrybuantą empiryczną: tyle, że jest ona określona jako zbiór punktów $x_{s,j}$ o $F_j$. W związku z tym należy wykonać interpolację między punktami najbliższymi wartości a oraz b, w celu odczytania wartości $F_e(a)$ i $F_e(b)$:

$$P_e(0 < X < 2) = F_e(2) - F_e(0) = 0{,}925693 - 0{,}547456 = 0{,}378237.$$

**4.3. Zadanie**

Wykonano 1000 pomiarów natężenia pewnej wielkości fizycznej (plik `X1.txt`). Określ rozkład prawdopodobieństwa. Parametry rozkładu wyznacz metodą największej wiarygodności.

*Rozwiązanie*:

Aby określić typ rozkładu należy narysować wykres gęstości prawdopodobieństwa (histogram) i na tej podstawie określić możliwy rozkład populacji generalnej. Do obliczenia szerokości klasy wykorzystany zostanie następujący wzór:

$$h^* = \frac{2{,}64 \cdot (Q_{0,75} - Q_{0,25})}{\sqrt[3]{N}}.$$

Kwantyle są równe $Q_{0,25} = 2{,}19594$ i $Q_{0,75} = 3{,}87521$.

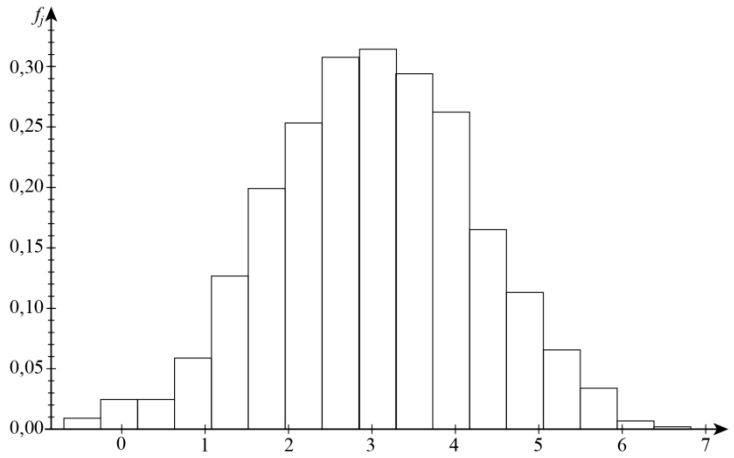
$$h^* = \frac{2{,}64 \cdot (Q_{0,75} - Q_{0,25})}{\sqrt[3]{N}} = 0{,}443325.$$

Wartości minimalna i maksymalna z próby wynoszą $x_{min}$=-0,693279 i $x_{max}$=6,82309. Rozstęp jest równy

$$R = x_{\max} - x_{\min} = 7{,}51637.$$

Oszacowana liczba klas to $k = R/h^* = \lceil 16{,}9545 \rceil = 17$ więc ostatecznie szerokość klasy wyniesie $h=R/17=0{,}4421392783529411$.

| $j$ | $x_{p,j}$-$x_{k,j}$ | $x_{sj}$ | $n_j$ | $p_j=n_j/N$ | $f_j=p_j/h$ |
|---|---|---|---|---|---|
| 1 | [-0,693279 ; -0,25114) | -0,47221 | 4 | 0,004 | 0,00904692 |
| 2 | [-0,25114 ; 0,190999) | -0,0300705 | 11 | 0,011 | 0,024879 |
| 3 | [0,190999 ; 0,633138) | 0,412069 | 11 | 0,011 | 0,024879 |
| 4 | [0,633138 ; 1,07528) | 0,854208 | 26 | 0,026 | 0,058805 |
| 5 | [1,07528 ; 1,51742) | 1,29635 | 56 | 0,056 | 0,126657 |
| 6 | [1,51742 ; 1,95956) | 1,73849 | 88 | 0,088 | 0,199032 |
| 7 | [1,95956 ; 2,4017) | 2,18063 | 112 | 0,112 | 0,253314 |
| 8 | [2,4017 ; 2,84383) | 2,62277 | 136 | 0,136 | 0,307595 |
| 9 | [2,84383 ; 3,28597) | 3,0649 | 139 | 0,139 | 0,314381 |
| 10 | [3,28597 ; 3,72811) | 3,50704 | 130 | 0,13 | 0,294025 |
| 11 | [3,72811 ; 4,17025) | 3,94918 | 116 | 0,116 | 0,262361 |
| 12 | [4,17025 ; 4,61239) | 4,39132 | 73 | 0,073 | 0,165106 |
| 13 | [4,61239 ; 5,05453) | 4,83346 | 50 | 0,05 | 0,113087 |
| 14 | [5,05453 ; 5,49667) | 5,2756 | 29 | 0,029 | 0,0655902 |
| 15 | [5,49667 ; 5,93881) | 5,71774 | 15 | 0,015 | 0,033926 |
| 16 | [5,93881 ; 6,38095) | 6,15988 | 3 | 0,003 | 0,00678519 |
| 17 | [6,38095 ; 6,82309] | 6,60202 | 1 | 0,001 | 0,00226173 |




Można zauważyć, że w przybliżeniu rozkład gęstości jest normalny. Zatem należy użyć estymatorów znalezionych metodą największej wiarygodności dla rozkładu normalnego w zadaniu 7:

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ;$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} .$$

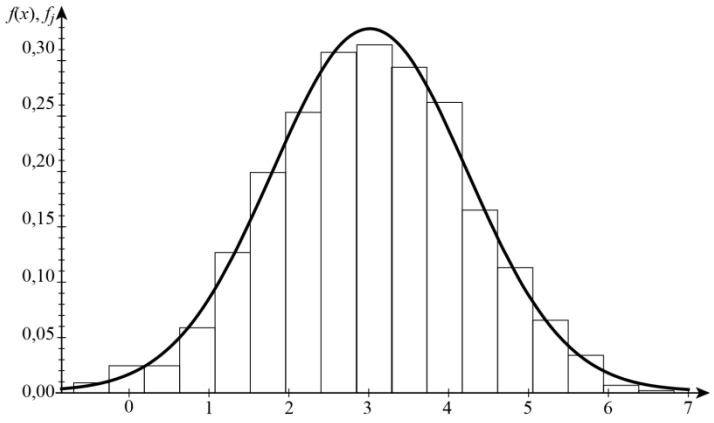
Więc oszacowania estymatorów na podstawie próby będą równe:

$$m = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3{,}02179$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1{,}21866$$

*Uwaga: Przytoczony tu estymator $\sigma$ jest estymatorem obciążonym. W praktyce, należy wykorzystywać estymator nieobciążony* $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ *, w związku z czym s=1,21927.*

Poniżej przedstawiony został wykres gęstości próby i wykres gęstości rozkładu normalnego $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.




# 5. Estymacja statystyczna (parametryczna)

Pojęcie estymacja oznacza szacowanie wartości nieznanego parametru. W praktyce chodzi o to, żeby umożliwić najlepsze oszacowanie parametrów rozkładu na podstawie zebranej próby. Wówczas można skonstruować najlepszy (w sensie błędu estymacji) model populacji generalnej (czyli wszystkich możliwych interesujących nas obiektów).

## 5.1. Podstawowe metody znajdywania estymatorów*

Znajdywanie estymatorów nie należy do zadań osoby pracującej na stanowisku data scientist. Pomimo tego, osoba używająca statystyki powinna mieć świadomość skąd biorą się wzory pozwalające określić parametry rozkładów na podstawie danych zawartych w próbie.

Mówiąc potocznie estymatory są wzorami pozwalającymi określić parametry rozkładu na podstawie próby tak, żeby to oszacowanie było jak najlepsze. Różne metody estymacji mogą prowadzić do różnych wzorów na obliczenie wartości tego samego parametru. Dodatkowo estymatory są niekiedy modyfikowane ze względu na swoje własności (np. obciążenie). W związku z tym w różnych podręcznikach można znaleźć wzory służące do wyznaczania wartości tego samego parametru, które różnią się od siebie. Prawdopodobnie najbardziej popularnym parametrem, którego to dotyczy jest odchylenie standardowe:

- estymator obciążony $s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- estymator nieobciążony $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

Różne rozkłady mają różne zestawy parametrów, które określają położenie oraz kształt danego rozkładu.

### 5.1.1. Metoda największej wiarygodności

Uwaga: obiekty reprezentujące rozkłady prawdopodobieństwa w scipy posiadają metodę `fit`, która umożliwia estymację parametrów danego rozkładu z użyciem metody największej wiarygodności.

Metoda największej wiarygodności polega na znalezieniu maksimum funkcji wiarygodności. W przypadku rozkładu ciągłego funkcja wiarygodności opisana jest następującym wzorem:

$$L = \prod_{i=1}^{n} f(x_i; \Theta_1, \Theta_2, ..., \Theta_k),$$

a przypadku rozkładu dyskretnego funkcja wiarygodności jest następująca:

$$L = \prod_{i=1}^{n} p(x_i; \Theta_1, \Theta_2, ..., \Theta_k).$$

Ze względu na możliwość uproszczenia obliczeń, często korzysta się z logarytmu funkcji wiarygodności:

$$\ln(L) = \sum_{i=1}^{n} \ln(f(x_i; \Theta_1, \Theta_2, ..., \Theta_k)),$$

lub

$$\ln(L) = \sum_{i=1}^{n} \ln(p(x_i; \Theta_1, \Theta_2, ..., \Theta_k)).$$

Następnie oblicza się pochodną funkcji $L$ lub jej logarytmu po każdym z parametrów i przyrównuje się do zera, gdyż zerowanie się pochodnej jest warunkiem koniecznym istnienia ekstremum (w tym maksimum). Z obliczonych pochodnych formuje się układ równań:

$$\begin{cases} \frac{\partial L(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_1} = 0 \\ \frac{\partial L(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_2} = 0 \\ \vdots \\ \frac{\partial L(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_k} = 0 \end{cases} \quad \text{lub} \quad \begin{cases} \frac{\partial \ln(L)(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_1} = 0 \\ \frac{\partial \ln(L)(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_2} = 0 \\ \vdots \\ \frac{\partial \ln(L)(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_k} = 0 \end{cases}$$

Po rozwiązaniu tego układu równań otrzymuje się estymatory parametrów. Następnie konieczne jest sprawdzenie czy drugie pochodne funkcji wiarygodności lub jej logarytmu po danym parametrze mają wartość ujemną w otrzymanych punktach.

#### 5.1.1.1. Zadanie

Metodą największej wiarygodności znaleźć estymatory parametrów $\mu$ i $\sigma$ zmiennej losowej podlegającej rozkładowi normalnemu $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

*Rozwiązanie:* Pierwszym krokiem jest zapisanie funkcji wiarygodności

$$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

i zlogarytmowanie jej:

$$\ln(L) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n}\left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right),$$

a następnie przekształcenie:

$$\ln(L) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n}\left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) =$$

$$\sum_{i=1}^{n}\left(\ln\left(\sqrt{2\pi}\sigma\right)^{-1} - \frac{(x_i-\mu)^2}{2\sigma^2}\right) = \sum_{i=1}^{n}\left(-\ln\left(\sqrt{2\pi}\sigma\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) = \sum_{i=1}^{n}\left(-\ln\left(\sqrt{2\pi}\right) - \ln(\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}\right) =$$

$$-\sum_{i=1}^{n}\ln\left(\sqrt{2\pi}\right) - \sum_{i=1}^{n}\ln(\sigma) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} = -n\cdot\ln\left(\sqrt{2\pi}\right) - n\cdot\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \cdot$$

Ostatecznie otrzymuje się:

$$\ln(L) = -n\cdot\ln\left(\sqrt{2\pi}\right) - n\cdot\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Obliczając pochodne po parametrach $\mu$ i $\sigma$ otrzymuje się:

$$\frac{d\ln(L)}{d\mu} = \frac{d}{d\mu}\left(-n\cdot\ln\left(\sqrt{2\pi}\right) - n\cdot\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\frac{d\ln(L)}{d\sigma} = \frac{d}{d\sigma}\left(-n\cdot\ln\left(\sqrt{2\pi}\right) - n\cdot\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = \frac{-n}{\sigma} - \frac{-2}{2\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 =$$

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{\sigma}$$

W związku z tym, że poszukiwane są dwa parametry należy uformować układ dwóch równań przyrównując obliczone pochodne do zera:

$$\begin{cases} \frac{d\ln(L)}{d\mu} = 0 \\ \frac{d\ln(L)}{d\sigma} = 0 \end{cases}$$

Z pierwszego równania wyznacza się wartość $\mu$:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mu = 0$$

$$\sum_{i=1}^{n}x_i = n\cdot\mu$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$$

Wstawiając powyższe rozwiązanie do równania drugiego wyznacza się parametr $\sigma$:

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\bar{x})^2-\frac{n}{\sigma}=0$$

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n}{\sigma}$$

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

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

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

Oszacowaniem parametru $\mu$ jest średnia arytmetyczna z próby $\mu=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$, a parametru $\sigma$ odchylenie standardowe z próby $\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$. Otrzymany estymator odchylenia standardowego jest obciążony, a zatem nie jest najlepszym możliwym. Obciążenie można łatwo wyeliminować poprzez modyfikacje tego wzoru i odjęcie od liczebności próby w mianowniku wartości 1.

### 5.1.1.2. Zadanie

Dany jest rozkład Borela. Metodą największej wiarygodności znaleźć estymator parametru $\mu$.

*Rozwiązanie:*

$$p(x)=\frac{e^{-\mu\cdot x}(\mu\cdot x)^{x-1}}{x!}$$

$$L=\prod_{i=1}^{n}p(x)=\prod_{i=1}^{n}\frac{e^{-\mu\cdot x_i}(\mu\cdot x)^{x_i-1}}{x_i!}$$

$$\ln L=\sum_{i=1}^{n}\ln\left[\frac{e^{-\mu\cdot x_i}(\mu\cdot x)^{x_i-1}}{x_i!}\right]=\sum_{i=1}^{n}\ln\left[e^{-\mu\cdot x_i}\cdot\mu^{x_i-1}\cdot\frac{x^{x_i-1}}{x_i!}\right]=\sum_{i=1}^{n}\left[\ln e^{-\mu\cdot x_i}+\ln\mu^{x_i-1}+\ln\frac{x^{x_i-1}}{x_i!}\right]=$$

$$=\sum_{i=1}^{n}\ln e^{-\mu\cdot x_i}+\sum_{i=1}^{n}\ln\mu^{x_i-1}+\sum_{i=1}^{n}\ln\frac{x^{x_i-1}}{x_i!}=-\mu\sum_{i=1}^{n}x_i+\ln\mu\sum_{i=1}^{n}(x_i-1)+\sum_{i=1}^{n}\ln\frac{x^{x_i-1}}{x_i!}$$

Utworzenie układu równań (estymacji podlega tylko jeden parametr zatem w układzie występuje tylko jedno równanie):

$$\frac{\partial\ln L}{\partial\mu}=0$$

$$-\sum_{i=1}^{n}x_i+\frac{1}{\mu}\sum_{i=1}^{n}(x_i-1)=0$$

$$\mu = \frac{\sum_{i=1}^{n}(x_i - 1)}{\sum_{i=1}^{n} x_i}$$

### *5.1.2. Metoda najmniejszych kwadratów*

Przedstaw algorytm metody najmniejszych kwadratów znajdywania estymatorów.

*Rozwiązanie*: Metoda najmniejszych kwadratów opiera się na funkcji błędu opisanej wzorem:

$$E(\Theta_1, \Theta_2, ..., \Theta_k) = \sum_{i=1}^{n} \left(x_i - h(\Theta_1, \Theta_2, ..., \Theta_k)\right)^2,$$

której minimum jej poszukiwane względem parametrów $\Theta_1, \Theta_2, ..., \Theta_k$. Zatem, poszukuje się takiego zestawu parametrów $\Theta_1, \Theta_2, ..., \Theta_k$, przy którym sumaryczny błąd między wartościami obserwowanymi, a obliczonymi przy użyciu funkcji *h* będzie najmniejszy.

$$E(\Theta_1, \Theta_2, ..., \Theta_k) = \sum_{i=1}^{n} \left(x_i - h(\Theta_1, \Theta_2, ..., \Theta_k)\right)^2 \to \min.$$

Aby znaleźć minimum funkcji należy obliczyć jej pochodne po każdym z parametrów i przyrównać do zera. W ten sposób otrzymuje się układ *k* równań, po rozwiązaniu którego otrzymane zostaną oszacowania wartości szukanych parametrów $\Theta_1, \Theta_2, ..., \Theta_k$.

$$\begin{cases} \frac{\partial E(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_1} = 0 \\ \frac{\partial E(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_2} = 0 \\ \vdots \\ \frac{\partial E(\Theta_1, \Theta_2, ..., \Theta_k)}{\partial \Theta_k} = 0 \end{cases}$$

#### 5.1.2.1 Zadanie

Metodą najmniejszych kwadratów określ estymatory parametrów $\alpha$ i $\beta$ równania prostej $y = \alpha \cdot x + \beta$ odzwierciedlającej możliwie najlepiej zależność (liniową) między zmiennymi $x$ oraz $y$.

*Rozwiązanie*: Pierwszym etapem rozwiązania jest znalezienie metodą najmniejszych kwadratów estymatorów *a* i *b* parametrów $\alpha$ i $\beta$ równania $y = \alpha \cdot x + \beta$. W tym celu konieczne jest przyjęcie funkcji błędu kwadratowego $\varepsilon(a,b) = \sum_{i=1}^{n} (y_i - (a \cdot x_i + b))^2$ jako kryterium. Znajdując minimum tej funkcji otrzymuje się estymatory szukanych parametrów:

$$\varepsilon(a,b) = \sum_{i=1}^{n} (y_i - (a \cdot x_i + b))^2 \to \min,$$

gdzie *n* jest liczbą punktów. Minimum tej funkcji znajduje się w miejscu zerowania się pierwszych pochodnych. Rozwiązując układ równań

$$\begin{cases} \frac{\partial \varepsilon(a,b)}{\partial a} = 0 \\ \frac{\partial \varepsilon(a,b)}{\partial b} = 0 \end{cases}$$

otrzymuje się szukane parametry.

$$\begin{cases} \frac{\partial}{\partial a}\left(\sum_{i=1}^{n}(y_i-(a\cdot x_i+b))^2\right)=0 \\ \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}(y_i-(a\cdot x_i+b))^2\right)=0 \end{cases}$$

$$\begin{cases} 2\sum_{i=1}^{n}\left((y_i-(a\cdot x_i+b))\cdot(-x_i)\right)=0 \\ 2\sum_{i=1}^{n}\left((y_i-(a\cdot x_i+b))\cdot(-1)\right)=0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n}(-x_i\cdot y_i+a\cdot x_i^2+b\cdot x_i)=0 \\ \sum_{i=1}^{n}(-y_i+a\cdot x_i+b)=0 \end{cases}$$

Następnie rozwija się wyrażenia względem operatora sumy.

$$\begin{cases} -\sum_{i=1}^{n}x_i\cdot y_i+a\cdot\sum_{i=1}^{n}x_i^2+b\cdot\sum_{i=1}^{n}x_i=0 \\ -\sum_{i=1}^{n}y_i+a\cdot\sum_{i=1}^{n}x_i+b\cdot n=0 \end{cases}$$

$$\begin{cases} a\cdot\sum_{i=1}^{n}x_i^2+b\cdot\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_i\cdot y_i \\ a\cdot\sum_{i=1}^{n}x_i+b\cdot n=\sum_{i=1}^{n}y_i \end{cases}$$

Otrzymany układ rozwiązuje się ze względu na szukane parametry *a* i *b*. W tym celu z równania drugiego wyznacza się *b*

$$b=\frac{\sum_{i=1}^{n}y_i-a\cdot\sum_{i=1}^{n}x_i}{n}=\bar{y}-a\cdot\bar{x}$$

po wstawieniu do równania pierwszego

$$a\cdot\sum_{i=1}^{n}x_i^2+\left(\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{a}{n}\sum_{i=1}^{n}x_i\right)\cdot\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_i\cdot y_i$$

$$a\cdot\sum_{i=1}^{n}x_i^2+\frac{1}{n}\sum_{i=1}^{n}y_i\cdot\sum_{i=1}^{n}x_i-\frac{a}{n}\left(\sum_{i=1}^{n}x_i\right)^2=\sum_{i=1}^{n}x_i\cdot y_i$$

$$a\cdot\sum_{i=1}^{n}x_i^2-\frac{a}{n}\left(\sum_{i=1}^{n}x_i\right)^2=\sum_{i=1}^{n}x_i\cdot y_i-\frac{1}{n}\sum_{i=1}^{n}y_i\cdot\sum_{i=1}^{n}x_i$$

$$a\cdot\left(\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right)=\sum_{i=1}^{n}x_i\cdot y_i-\frac{1}{n}\sum_{i=1}^{n}y_i\cdot\sum_{i=1}^{n}x_i$$

i uporządkowaniu otrzymuje się wzór umożliwiający obliczenie estymatora *a*

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}.$$

Ostatecznie otrzymuje się:

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \qquad b = \frac{\sum_{i=1}^{n} y_i - a \cdot \sum_{i=1}^{n} x_i}{n} = \bar{y} - a \cdot \bar{x}$$

## 5.2. Przedziały ufności

### *5.2.1. Przedziały ufności dla średniej*

**5.2.1.1. Zadanie**

W eksperymencie bada się ilość wydzielonego produktu w wyniku reakcji chemicznej. Wykonano $n$=60 niezależnych doświadczeń i otrzymano z nich średnią $\bar{x}$ =46 g oraz odchylenie $s$=13 g. Przyjmując współczynnik ufności 1-$\alpha$ =0,99 oszacować metodą przedziałową średnią ilość wydzielonej substancji.

*Rozwiązanie*: Próba jest duża, zatem na mocy centralnego twierdzenia granicznego do obliczeń wykorzystuje się wzór:

$$P\left(\bar{x} + u_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + u_{1-\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Rozkład średniej arytmetycznej z próby jest w tym przypadku rozkładem normalnym. Kwantyle należy odczytać dla wartości $\frac{\alpha}{2}$ oraz $1 - \frac{\alpha}{2}$. Wstawiając do wzoru otrzymuje się:

$$P\left\{46 - 2{,}58\frac{13}{\sqrt{60}} \le \mu \le 46 + 2{,}58\frac{13}{\sqrt{60}}\right\} = 0{,}99$$

$$P\{46 - 4{,}33 \le \mu \le 46 + 4{,}33\} = 0{,}99$$

Przedział pokrywający z prawdopodobieństwem 0,99 średnią ilość substancji wydzieloną w trakcie reakcji to$\mu = 46 \pm 4{,}33$.

### *5.2.2. Przedziały ufności dla odchylenia standardowego*

**5.2.2.1. Zadanie**

Dla danych z zadania 4.2.1.1. zbuduj przedział ufności dla odchylenia standardowego ($n$=60, $s$=13 s, 1-$\alpha$ =0.99).

*Rozwiązanie*:

Wzór określający przedział ufności dla odchylenia standardowego ma następującą postać:

$$P\left\{\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}}\right\} = 1 - \alpha$$

W celu otrzymania przedziałów ufności konieczne jest określenie kwantyli rzędu $1-\frac{\alpha}{2}$ oraz $\frac{\alpha}{2}$ rozkładu $\chi^2$ z $n-1$ stopniami swobody.

$$\chi^2_{1-\frac{\alpha}{2},n-1} = 90{,}72, \quad \chi^2_{\frac{\alpha}{2},n-1} = 34{,}77$$

$$P\{10{,}48 \leq \sigma \leq 16{,}93\} = 0{,}99$$

### *5.2.3. Przedziały ufności dla proporcji*

**5.2.3.1. Zadanie**

W ankiecie wzięło udział $n = 1000$ respondentów. Pytano ich czy palą regularnie. 20% z nich odpowiedziało „tak". Znajdź przedziały ufności dla proporcji osób palących w populacji generalnej. przyjmij poziom ufności 1-$\alpha$ = 0,99.

*Rozwiązanie*: Wzór określający przedziały ufności ma następująca postać:

$$P\left\{\hat{p} + u_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} = 1-\alpha$$

gdzie $\hat{p}$ określa proporcję otrzymaną na podstawie próby.

$$P\{0{,}167 \leq p \leq 0{,}232\} = 0{,}99$$

### *5.2.4. Krzywe ufności dla funkcji regresji liniowej*

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $y_i$ | 8 | 13 | 14 | 17 | 18 | 20 | 22 |

**5.2.4.1. Zadanie**

Określić współczynniki prostej regresji i jej obszar ufności na poziomie 1-α=0,95 dla zbioru *n*=7 punktów danego w tabeli.

*Rozwiązanie*: Do obliczenia parametrów $a$ i $b$ modelu można wykorzystać metodę najmniejszych kwadratów. Zatem współczynniki te będą dane wzorami:

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \qquad b = \frac{\sum_{i=1}^{n} y_i - a \cdot \sum_{i=1}^{n} x_i}{n} = \bar{y} - a \cdot \bar{x}$$

Obliczenie wymaganych sum

$$\sum_{i=1}^{7} x_i = 28 \qquad \sum_{i=1}^{7} y_i = 112 \qquad \sum_{i=1}^{7} x_i \cdot y_i = 508 \qquad \sum_{i=1}^{7} x_i^2 = 140$$

$$\bar{x} = \frac{1}{7}\sum_{i=1}^{7} x_i = 4 \qquad \bar{y} = \frac{1}{7}\sum_{i=1}^{7} y_i = 16$$

$$a = \frac{\sum_{i=1}^{7} x_i \cdot y_i - \frac{1}{n}\sum_{i=1}^{7} y_i \cdot \sum_{i=1}^{7} x_i}{\sum_{i=1}^{7} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = \frac{508 - \frac{1}{7}\cdot 112 \cdot 28}{140 - \frac{1}{7}28^2} = 2{,}14$$

$$b = \bar{y} - a \cdot \bar{x} = 16 - 2{,}14 \cdot 4 = 7{,}44$$

W celu określenia krzywych ufności korzysta się z następujących wzorów:

$$P\left(\hat{y}_i + t_\alpha \cdot s_{\hat{y}_i} < \tilde{y}_i < \hat{y}_i + t_{1-\alpha} \cdot s_{\hat{y}_i}\right) = 1-\alpha$$

$\tilde{y}_i$ – $i$-ta wartość w populacji generalnej;

$\hat{y}_i$ – $i$-ta wartość obliczona na podstawie estymowanej prostej regresji;

$y_i$ – $i$-ta wartość zaobserwowana w próbie.

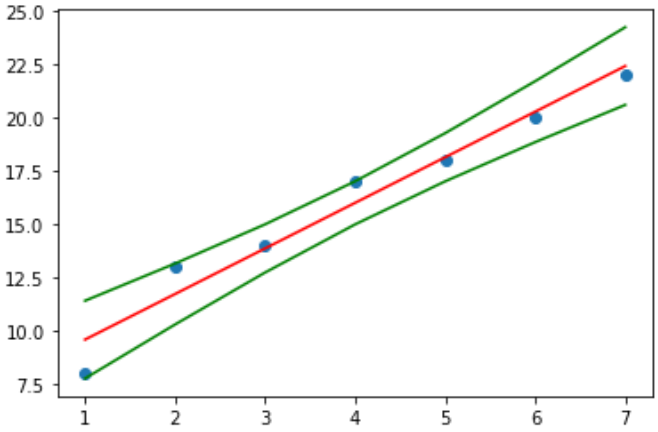
$t_\alpha, t_{1-\alpha}$ – wartości kwantyli rzędu $\alpha$ i $1-\alpha$ rozkładu T-Studenta o $n-2$ stopniach swobody.

$$s_{\hat{y}_i} = \sqrt{\frac{1}{n-2}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}$$

$$\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 = 5{,}43 \qquad \sum_{j=1}^{n}(x_j - \bar{x})^2 = 28 \qquad t_\alpha = -2{,}57 \qquad t_{1-\alpha} = 2{,}57$$

Wartość $\alpha = 0{,}05$.

| $i$ | $\hat{y}_i$ | $\hat{y}_i + t_\alpha \cdot s_{\hat{y}_i}$ | $\hat{y}_i + t_{1-\alpha} \cdot s_{\hat{y}_i}$ |
|---|---|---|---|
| 1 | 9,571 | 7,746 | 11,397 |
| 2 | 11,714 | 10,283 | 13,146 |
| 3 | 13,857 | 12,726 | 14,990 |
| 4 | 16,000 | 14,988 | 17,012 |
| 5 | 18,143 | 17,011 | 19,275 |
| 6 | 20,286 | 18,854 | 21,717 |
| 7 | 22,429 | 20,604 | 24,254 |




## 5.3. Testy statystyczne

### *5.3.1. Testy parametryczne*

#### 5.3.1.1. Zadanie

Norma techniczna przewiduje średnio 55 sekund na wykonanie pewnej operacji technicznej przez robotników. Ponieważ robotnicy skarżyli się, że norma ta jest zła, dokonano pomiarów czasu wykonania tej czynności dla $n$=60 wylosowanych pracowników. Otrzymano z tej próby średnią 72 sekundy i odchylenie standardowe 20 sekund. Czy można na poziomie istotności $\alpha$=0,01 stwierdzić, że rzeczywisty średni czas wykonania tej operacji jest zgodny z normą?

*Rozwiązanie*: Pierwszym krokiem jest określenie hipotezy zerowej i hipotezy alternatywnej. Hipoteza zerowa mówi, że pracownicy, średnio potrzebują $\mu_0$=55 sekund na wykonanie rozważanej czynności. W związku z tym, że pracownicy narzekają, iż norma jest zła, można domniemywać, że średnio czynność ta zajmuje im więcej czasu. Stąd wynika, że hipoteza alternatywna mówi, że pracownicy średnio potrzebują więcej niż $\mu_0$=55 sekund

na wykonanie tej czynności technicznej (wymusza to wykorzystanie prawostronnego obszaru krytycznego). Zatem formalnie można zapisać:

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Próba jest duża więc na mocy centralnego twierdzenia granicznego należy wykorzystać statystykę *U*. Wartość statystyki testowej oblicza się na podstawie wzoru:

$$u = \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \frac{72-55}{20}\sqrt{60} = 6{,}58\cdot$$

Wartość krytyczną $u_\alpha$ odczytuje się z tablic dystrybuanty rozkładu normalnego standaryzowanego. W związku z tym, że obszar krytyczny jest prawostronny $u_\alpha$ jest kwantylem $q_{1-\alpha}$. Dla 1-$\alpha$=0,99 otrzymuje się $u_\alpha$ =2,33 (odczyt dla wartości najbliższej 0,99: $F(u)$=0,990097).

Wartość statystyki testowej *u* jest większa od wartości krytycznej, zatem znajduje się w prawostronnym obszarze krytycznym. Oznacza to, że hipotezę zerową należy odrzucić i przyjąć hipotezę alternatywną. Odpowiadając na pytanie postawione w treści zadania („Czy można na poziomie istotności $\alpha$=0,01 stwierdzić, że rzeczywisty średni czas wykonania tej operacji jest zgodny z normą?") należy stwierdzić, że: **Średni czas wykonania tej czynności technicznej przez pracowników jest większy niż przewiduje to norma**.

**5.3.1.2. Zadanie**

W pewnym biochemicznym doświadczeniu bada się czas życia bakterii w pewnym środowisku. Rozkład czasu przeżycia można uważać za normalny. Dokonano 8 pomiarów i otrzymano następujące czasy życia bakterii: 4,7; 5,3; 4,0; 3,8; 6,2; 5,5; 4,5; 6,0. Przyjmując poziom istotności 0,05 sprawdzić hipotezę, że średni czas życia tych bakterii w tym środowisku wynosi więcej niż 4 godziny.

*Rozwiązanie*: Hipoteza zerowa zakłada, ze średni czas przeżycia bakterii w danym środowisku wynosi $\mu_0$=4 godziny. W związku z domniemaniem, że czas życia bakterii wynosi więcej niż $\mu_0$=4 godziny, hipoteza alternatywna zakłada, że czas przeżycia jest większy niż 4 godziny. Zatem w przypadku odrzucenia hipotezy zerowej, przyjęta zostanie hipoteza alternatywna.

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Próba jest mała i ma rozkład normalny więc należy wykorzystać statystykę *T*. Wartość statystyki testowej oblicza się na podstawie wzoru:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n-1}$$

Obliczenie średniego czasu życia bakterii: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 5$

oraz odchylenia standardowego

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0{,}8912271\cdot$$

Wartość statystyki testowej wyniesie:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n-1} = \frac{5-4}{0{,}8912271}\sqrt{7} = 2{,}9686613$$

Wartość krytyczną $t_\alpha$ odczytuje się z tablic rozkładu T-Studenta. W związku z tym, że obszar krytyczny jest prawostronny $t_\alpha$ jest kwantylem $q_{1-\alpha}$. Zatem należy odczytać wartość $t_\alpha$ dla $\alpha$=0,05 jak dla testu jednostronnego i liczby stopni swobody $df$=$n$–1=7: $t_\alpha$ =1,894579.

Wartość statystyki testowej $t$ jest większa od wartości krytycznej $t_\alpha$ zatem hipotezę zerową należy odrzucić. Oznacza to, że na poziomie istotności 0,05 możemy twierdzić, że czas przeżycia bakterii w badanym środowisku jest większy niż 4 godziny.

**5.3.1.3. Zadanie**

Zakupiono bardzo drogie urządzenie do pomiaru stężenia tlenu rozpuszczonego w wodzie. Producent zapewnia, że odchylenie standardowe pomiarów pewnej stałej wartości nie przekracza wartości $\sigma_0 = 0{,}01$ mg/l. Wykonano $n = 10$ pomiarów stężenia tlenu i otrzymano następujące wyniki:

$$\bar{x} = \{11{,}9845;\ 11{,}987;\ 12{,}0025;\ 11{,}9991;\ 12{,}0007;\ 11{,}9877;\ 11{,}9907;\ 12{,}0075;\ 11{,}9932;\ 12{,}0003\}$$

Na poziomie istotności $\alpha = 0{,}025$ sprawdź hipotezę, że odchylenie standardowe nie przekracza podanej wartości.

*Rozwiązanie*: Celem jest weryfikacja hipotezy, że odchylenie standardowe pomiarów, które ma zapewniać zakupione urządzenie nie przekracza podanej wartości. W związku z tym hipoteza zerowa będzie zakładała równość wartości otrzymanej na podstawie próby i założonej. Natomiast hipoteza alternatywna będzie miała prawostronny obszar krytyczny. W przypadku odrzucenia hipotezy zerowej wykazane zostanie przekroczenie wartości odchylenia standardowego.

$$H_0:\ \sigma = \sigma_0$$

$$H_1:\ \sigma > \sigma_0$$

Statystyka testowa dana jest następującym wzorem:

$$\chi^2 = (n-1)\left(\frac{s}{\sigma_0}\right)^2 = 5{,}38.$$

Zmienna $\chi^2$ ma rozkład chi-kwadrat z $df = n - 1$ stopniami swobody. W związku z prawostronnym obszarem krytycznym, dla danej wartości współczynnika istotności wartość krytyczna (kwantyl rozkładu chi kwadrat) wynosi $\chi_{crit}{}^2_{1-\alpha,n-1} = 19{,}023$.

Wartość statystyki nie znajduje się w obszarze krytycznym, zatem nie ma podstaw do odrzucenia hipotezy zerowej, mówiącej, że prawdziwa wartość odchylenia standardowego pomiarów wynosi $\sigma_0 = 0{,}01$ mg/l.

**5.3.1.4. Zadanie**

Badany jest nowy lek na nadciśnienie. W badaniu uczestniczy grupa $n = 10$ pacjentów. Pierwszy pomiar wartości ciśnienia krwi przeprowadzono przed wprowadzeniem nowego leku, drugi po rozpoczęciu terapii nowym lekiem. Na poziomie istotności $\alpha = 0{,}01$ zweryfikuj hipotezę, że nowy medykament jest lepszy od poprzedniego.

$$x_1 = [135, 122, 134, 126, 134, 136, 125, 135, 133, 132]$$

$$x_2 = [127, 136, 135, 129, 131, 131, 136, 133, 131, 129]$$

*Rozwiązanie*: W związku z tym, że pomiarów dokonywano u tych samych pacjentów (przed i po) należy przeprowadzić tzw. test dla par wiązanych. Statystyka testowa ma postać:

$$t = \frac{\overline{\Delta} - \Delta_0}{s_\Delta / \sqrt{n}}$$

I podlega rozkładowi T-Studenta z $df = n - 1$ stopniami swobody. W powyższym wzorze wartość

$$\Delta = x_1 - x_2.$$

określa różnice w obserwacjach natomiast symbol $s_\Delta$ jest odchyleniem standardowym Δ.Symbol $\Delta_0$ oznacza założoną różnicę między średnimi wartościami obserwacji.

Hipotezą zerową w rozważanym teście jest

$$H_0: \quad \mu_1 = \mu_2$$

Natomiast hipotezą alternatywną

$$H_1: \quad \mu_1 > \mu_2.$$

Powyższe hipotezy są identyczne jak $H_0: \quad \Delta = \Delta_0$ oraz $H_1: \quad \Delta > \Delta_0$, gdzie $\Delta_0 = 0$. Oznacza to, że jeśli nowy lek jest bardziej efektywny w obniżaniu ciśnienia krwi odrzucona zostanie hipoteza zerowa, mówiąca o braku różnic między średnimi wartościami ciśnienia krwi. Jeśli lek nie działa, to znaczy, że hipoteza zerowa nie zostanie odrzucona, tj. wyniki przed wdrożeniem leku i po jego wdrożeniu nie różnią się.

$$\Delta = x_1 - x_2 = [\ 8, -14, -1, -3, 3, 5, -11, 2, 2, 3]$$

$$\overline{\Delta} = -0{,}6$$

$$s_\Delta = 6{,}979$$

Statystyka testowa przyjmuje wartość

$$t = \frac{-0{,}6 - 0}{6{,}979/\sqrt{10}} = -0{,}27,$$

natomiast statystyka krytyczna wynosi 2,82. Brak jest podstaw do odrzucenia hipotezy zerowej mówiącej o braku różnic między próbami.

**5.3.1.5. Zadanie**

Badanie polega na analizie dwóch oczyszczalni ścieków. Każda z rozważanych oczyszczalni używa innej technologii. Przez miesiąc (n=31) prowadzono pomiary $BZT_5$ otrzymując wyniki przedstawione w tabeli.

| | Średnia $\bar{x}$ | Odchylenie standardowe $s$ |
|---|---|---|
| WWTP 1 | 6,25 | 5,26 |
| WWTP 2 | 6,18 | 5,91 |

Zakładając, że obie próby pochodzą z rozkładów normalnych, na poziomie istotności sprawdź $\alpha$=0,01 czy średnie wartości $BZT_5$ między oczyszczalniami różnią się.

*Rozwiązanie*: Aby przeprowadzić T-test dla dwóch prób, należy określić czy odchylenia standardowe w tych próbach są sobie równe, czy są różne. W tym celu konieczne jest przeprowadzenie testu dla dwóch wariancji. Test ten ma następujące hipotezy

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

$$H_1: \ \sigma_1^2 \neq \sigma_2^2$$

A statystyka testowa

$$F = \frac{s_1^2}{s_2^2} = 0{,}7921$$

podlega rozkładowi F-Snedecora z $df_1 = n_1 - 1$ oraz $df_2 = n_2 - 1$ stopniami swobody. W rozważanym tu przypadku $n_1 = n_2 = n$.

Aby sprawdzić hipotezę o równości wariancji należy przyjąć poziom istotności dla tego postępowania. W związku z brakiem określenia tej wartości dla sprawdzenia równości wariancji, przyjmijmy $\alpha_v = 0{,}05$. Wówczas przyjętym obszarem krytycznym będzie $C = [-\infty, 0.4821] \cup [2.074, +\infty]$. Otrzymana wartość statystyki testowej nie znajduje się w obszarze krytycznym, zatem brak jest podstaw do odrzucenia hipotezy zerowej, mówiącej o tym, że odchylenia standardowe są sobie równe.

W związku z powyższym wnioskiem należy wykorzystać wariant testu dla dwóch średnich, w którym zakłada się równość wariancji w populacjach, z których pochodzą próby. Jednak wartości zaobserwowane w próbach różnią się, zatem w teście dla dwóch średnich użyta zostanie średnia wartość z odchyleń standardowych.

$$s = \frac{s_1 + s_2}{2} = 11{,}17$$

Statystyka testowa

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ma rozkład T-Studenta z $df = n_1 + n_2 - 2$ stopniami swobody.

Otrzymana wartość statystyki testowej to $t = 0{,}0247$, natomiast obszar krytyczny $C = [-\infty - 2.66] \cup [2.66, +\infty]$. Statystyka testowa nie leży w obszarze krytycznym więc brak jest podstaw do odrzucenia hipotezy zerowej, mówiącej o równości średnich w obu populacjach generalnych, z których pobrano próby. Wnioskiem jest stwierdzenie, że średnie wartości $BZT_5$ otrzymywane w wyniku procesu oczyszczania ścieków w tych oczyszczalniach nie różnią się.

**5.3.1.6. Zadanie**

W trakcie pewnego procesu biologicznego $p_0$=8% bakterii powinno uzyskać możliwość produkcji energii z tłuszczu. Zbadano $n = 200$ próbek i otrzymano wynik mówiący o tym, że $p$=6% bakterii dokonało adaptacji do używania tłuszczu jako paliwa. Na poziomie istotności $\alpha = 0{,}01$ sprawdź czy mniejsza niż oczekiwana ilość bakterii dokonała adaptacji do nowego środowiska.

*Rozwiązanie*: W teście dla proporcji hipoteza zerowa i alternatywna przyjmą następującą formę:

$$H_0: \ p = p_0$$

$$H_1: \ p < p_0.$$

Statystyka testowa

$$u = \frac{p - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Ma rozkład normalny standaryzowany. Otrzymana wartość statystyki testowej to $u = -1{,}04$. Wartość krytyczna statystyki to $u_{kryt} = -2{,}32$. Statystyka testowa nie znalazła się w obszarze krytycznym zatem brak jest podstaw do odrzucenia hipotezy zerowej, mówiącej, że $p$=8%.

Przy przyjętym poziomie istotności brak jest podstaw do stwierdzenia, że mniej niż 8% bakterii adaptowało się do nowego środowiska.

## *5.3.2. Testy nieparametryczne*

**5.3.2.1. Zadanie**

Dla następującej realizacji próby

| *i* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *x* | 0,28 | 0,13 | 0,11 | 0,21 | 0,63 | 0,48 | 0,04 | 0,41 | 0,92 | 0,44 | 0,27 | 0,50 | 0,26 | 0,78 | 0,41 |

na poziomie istotności $\alpha$=0,05 przy pomocy testu Kołmogorowa zweryfikuj hipotezę, że zmienna losowa, której realizacją jest ta próba ma rozkład jednostajny na odcinku [0;1]. Podaj interpretację graficzną testu Kołmogorowa.

*Rozwiązanie*: W celu przeprowadzenia testu zgodności Kołmogorowa, próbę należy posortować w porządku rosnącym. Następnie, zgodnie ze wzorem:

$$F_N(x)=\begin{cases}0 & x<x_1\\ \dfrac{i}{N} & x_i\le x<x_{i+1}\\ 1 & x>x_N\end{cases}$$

należy obliczyć wartości dystrybuanty empirycznej. Następnie należy obliczyć wartości dystrybuanty teoretycznego rozkładu. Tu jest to rozkład jednostajny unormowany więc jego dystrybuanta dana jest wzorem:

$$F(x)=\begin{cases}0 & x<0\\ x & 0\le x\le 1\\ 1 & x>1\end{cases}$$

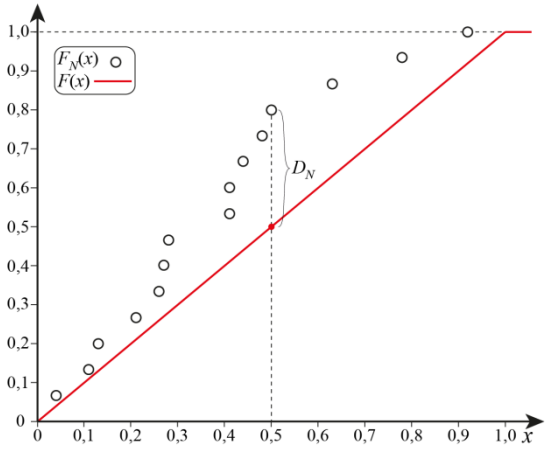
Kolejnym krokiem jest obliczenie różnic między dystrybuantą empiryczną oraz teoretyczną | $F_N(x_i)$ - $F(x_i)$ | w punktach $x_i$. Do obliczenia statystyki testowej konieczne jest znalezienie największej wartości bezwzględnej z różnic między dystrybuantą empiryczną i teoretyczną $D_N=\sup_x|F_N(x_i)-F(x_i)|=0{,}300$. Wartość statystyki testowej $\sqrt{N}\cdot D_N=\sqrt{15}\cdot 0{,}300=1{,}1619$ należy porównać z kwantylem $\lambda_{1-\alpha}$ dla $Q(\lambda_{1-\alpha})$=1-$\alpha$ rozkładu Kołmogorowa-Smirnowa, którego dystrybuanta dana jest wzorem:

| *i* | $x_i$ | $F(x)$ | $F_N(x)$ | $\lvert F_N(x) - F(x)\rvert$ |
|---|---|---|---|---|
| 1 | 0,04 | 0,04 | 0,067 | 0,027 |
| 2 | 0,11 | 0,11 | 0,133 | 0,023 |
| 3 | 0,13 | 0,13 | 0,200 | 0,070 |
| 4 | 0,21 | 0,21 | 0,267 | 0,057 |
| 5 | 0,26 | 0,26 | 0,333 | 0,073 |
| 6 | 0,27 | 0,27 | 0,400 | 0,130 |
| 7 | 0,28 | 0,28 | 0,467 | 0,187 |
| 8 | 0,41 | 0,41 | 0,533 | 0,123 |
| 9 | 0,41 | 0,41 | 0,600 | 0,190 |
| 10 | 0,44 | 0,44 | 0,667 | 0,227 |
| 11 | 0,48 | 0,48 | 0,733 | 0,253 |
| 12 | 0,50 | 0,50 | 0,800 | **0,300** |
| 13 | 0,63 | 0,63 | 0,867 | 0,237 |
| 14 | 0,78 | 0,78 | 0,933 | 0,153 |
| 15 | 0,92 | 0,92 | 1,000 | 0,080 |

$$Q(x)=\begin{cases}0 & x\le 0\\ \sum_{k=-\infty}^{+\infty}(-1)^k e^{-2k^2x^2} & x>0\end{cases}$$

*Uwaga: Wartości kwantyli można odczytać z tablic rozkładu Q.*

Przy $Q(\lambda_{1-\alpha})$=1-$\alpha$ =1-0,05=0,95 wartość kwantyla wyniesie $\lambda_{1-\alpha}\approx 1{,}36$. Obszar krytyczny testu Kołmogorowa jest prawostronny, czyli hipotezę zerową odrzuca się, gdy $\sqrt{N}\cdot D_N>\lambda_{1-\alpha}$. W rozważanym przypadku $\sqrt{N}\cdot D_N$ = 1,1619 < 1,36 = $\lambda_{1-\alpha}$, więc brak jest podstaw do odrzucenia hipotezy zerowej, mówiącej, że próba pochodzi z populacji generalnej mającej rozkład jednostajny unormowany. Graficzna interpretacja do zadania przedstawiona została na poniższym rysunku.




**5.3.2.2. Zadanie**

Zbadano ilość bakterii tworzących kolonie w 1ml ścieków oczyszczonych pobranych na odpływie z dwóch różnych oczyszczalni ścieków. Na poziomie istotności $\alpha$=0,05 zbadaj, czy obie próby pochodzą z populacji o takim samym rozkładzie.

| Liczba bakterii | | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|
| Oczyszczalnia | **A** | 2 | 3 | 8 | 10 | 11 | 11 | 4 | 2 | 1 |
| | **B** | 2 | 8 | 11 | 9 | 7 | 6 | 2 | 1 | 1 |

*Rozwiązanie*: W celu przeprowadzenia testu należy określić dystrybuanty empiryczne dla obu prób.

$$F_N(x)=\begin{cases}0 & x<1\\ \dfrac{\sum_{i=1}^{j} n_i}{N} & j\le x<j+1\\ 1 & x\ge k\end{cases}$$

dla $j$=1, 2, ..., $k$-1, gdzie $k$ oznacza liczbę kategorii, $N=\sum_{i=1}^{k} n_i$ liczebnością próby. Oznacza to, że aby obliczyć wartość dystrybuanty empirycznej dla danej klasy lub kategorii konieczne jest zsumowanie częstości z danej klasy i częstości wszystkich poprzedzających ją klas, i podzielenie tej wartości przez liczebność próby. Aby obliczyć wartość statystyki testowej konieczne jest znalezienie maksymalnej wartości bezwzględnej różnic między obiema dystrybuantami $D_N=\sup_x |F_{A,N}(x_i)-F_{B,N}(x_i)|$. Obliczając liczebności prób otrzymuje się: $N_A=\sum_{i=1}^{k} n_{A,i}=52$, $N_B=\sum_{i=1}^{k} n_{B,i}=47$. Dalsze obliczenia przedstawiono w poniższej tabeli:

| $x$ | Liczba bakterii | A | B | $\sum_{i=1}^{j} n_{A,i}$ | $\sum_{i=1}^{j} n_{B,i}$ | $F_{A,N}(x)$ | $F_{B,N}(x)$ | $\lvert F_{A,N}(x) - F_{B,N}(x) \rvert$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 2 | 2 | 2 | 2 | 0,038 | 0,043 | 0,004 |
| 2 | 7 | 3 | 8 | 5 | 10 | 0,096 | 0,213 | 0,117 |
| 3 | 8 | 8 | 11 | 13 | 21 | 0,250 | 0,447 | **0,197** |
| 4 | 9 | 10 | 9 | 23 | 30 | 0,442 | 0,638 | 0,196 |
| 5 | 10 | 11 | 7 | 34 | 37 | 0,654 | 0,787 | 0,133 |
| 6 | 11 | 11 | 6 | 45 | 43 | 0,865 | 0,915 | 0,05 |
| 7 | 12 | 4 | 2 | 49 | 45 | 0,942 | 0,957 | 0,015 |
| 8 | 13 | 2 | 1 | 51 | 46 | 0,981 | 0,979 | 0,002 |
| 9 | 14 | 1 | 1 | 52 | 47 | 1 | 1 | 0 |

Maksymalna różnica między dystrybuantami wynosi $D_N = 0{,}197$. Wartość statystyki testowej oblicza się na podstawie wzoru $\sqrt{N_{AB}} \cdot D_N$, gdzie $N_{AB} = \frac{N_A \cdot N_B}{N_A + N_B}$. Kwantyl $\lambda_{1-\alpha}$ rozkładu $Q$ jest równy $\lambda_{1-\alpha}$=1,36. Otrzymuje się $\sqrt{N_{AB}} \cdot D_N = 0{,}979 < 1{,}36 = \lambda_{1-\alpha}$, zatem brak jest podstaw do odrzucenia hipotezy zerowej, mówiącej, że obie próby pochodzą z populacji o takim samym rozkładzie.

*Uwaga: Wartości kwantyli można odczytać z tablic rozkładu Q.*

Wykres dystrybuant przedstawiono na poniższym rysunku